Titre: Title:	Human-Centric, Causal and Physics-Guided AI for Designing and Optimizing Sustainable Industrial Processes and Materials
Auteur: Author:	Eslam Ibrahim
Date:	2026
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Ibrahim, E. (2026). Human-Centric, Causal and Physics-Guided AI for Designing and Optimizing Sustainable Industrial Processes and Materials [Thèse de doctorat, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/76289/

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Human-Centric, Causal and Physics-Guided AI for Designing and Optimizing Sustainable Industrial Processes and Materials

ESLAM IBRAHIM

Département de mathématiques et de génie industriel

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*

Génie industriel

Avril 2026

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Cette thèse intitulée :

Human-Centric, Causal and Physics-Guided AI for Designing and Optimizing Sustainable Industrial Processes and Materials

présentée par **Eslam IBRAHIM**

en vue de l'obtention du diplôme de *Philosophiæ Doctor*

a été dûment acceptée par le jury d'examen constitué de :

Thibaut VIDAL, président
Hanane DAGDOUGUI, membre et directrice de recherche
Ahmed RAGAB, membre et codirecteur de recherche
Daria Camilla BOFFITO, membre et codirectrice de recherche
Christopher J. PAL, membre
Faïçal LARACHI, membre externe

DEDICATION

To my beloved Mother and Brother. They are my true heroes.

ACKNOWLEDGEMENTS

I am deeply grateful to my mother for her unwavering love, strength, and sacrifices, which have been the foundation of everything I have achieved. I also thank my brother for his constant support, encouragement, and belief in me throughout this journey.

I would like to extend my sincere and heartfelt gratitude to my PhD supervisor, Prof. Hanane Dagdougoui, for her exceptional mentorship, constant support, and generous guidance throughout my doctoral journey. Her deep knowledge, thoughtful insights, and constructive feedback have been fundamental in shaping both this research and my development as a scholar. I am profoundly grateful for her patience, encouragement, and unwavering dedication to my growth, which continuously motivated me to push beyond my limits and strive for excellence. Her guidance has left an indelible mark on the quality and direction of this thesis.

I am deeply and sincerely grateful to my PhD co-supervisor and co-director of research, Prof. Ahmed Ragab, for his invaluable contributions, generous guidance, and unwavering support throughout this journey. His expertise in artificial intelligence, industrial engineering, and related fields played a pivotal role in shaping the direction and depth of this research. Through his insightful suggestions, critical perspective, and remarkable attention to technical detail, he continuously challenged me to improve and to give my very best. His words of encouragement provided me with the confidence to overcome difficult moments, and his belief in my abilities often carried me forward when the path felt uncertain. Beyond the scientific realm, he also imparted life lessons that strengthened my resilience and shaped my character, teaching me the importance of integrity, humility, and perseverance. His mentorship felt both guiding and deeply caring, and I will always be grateful for the opportunity to have worked with him, to have learned from his vast experience, and to have discovered my weaknesses before my strengths, a gift that has profoundly contributed to my growth as both a researcher and an individual.

I feel very fortunate to have had Prof. Daria Boffito as my co-supervisor during my doctoral studies. Her openness, clarity, and genuine enthusiasm for research brought a refreshing perspective to this work. She consistently challenged me to think more critically, refine my ideas, and approach complex problems with creativity and confidence. Her encouragement, thoughtful questions, and careful attention to details were invaluable at every stage of this journey. Beyond her academic contributions, I deeply appreciate her human approach, her patience during difficult moments, and

her willingness to listen and support. Working under her guidance has been both a privilege and an inspiration, and her impact on my growth as a researcher will stay with me long after this thesis is completed.

I would like to express my sincere appreciation to Dr. Mouloud Amazouz, Manager of the AI team at CanmetENERGY Varennes, for his guidance and generous support throughout this work. His trust, encouragement, and thoughtful advice created an environment in which I felt motivated to grow, explore new ideas, and push my limits. I am truly grateful for his confidence in my abilities and for the positive and inspiring leadership he consistently demonstrated during our collaboration. I would also like to extend my sincere and heartfelt gratitude to Prof. Marzouk Benali, Manager of the Hydrogen & Biorefinery team at CanmetENERGY in Varennes, for his generous guidance, trust, and continuous support throughout this work specifically the work on ensemble and generative learning to help accelerate properties prediction and materials discovery presented by Chapter 4 and Chapter 5. His encouragement and openness created a truly motivating environment that made this collaboration both meaningful and enriching. I am especially thankful to Prof. Mohamed Ali at CanmetENERGY in Devon, whose expertise and dedication were indispensable in guiding me to construct high-quality datasets and realistic simulations that closely reflected real-world case studies presented in Chapter 6 to validate the developed CIRL methodology. His willingness to share his time, knowledge, and technical insight played a crucial role in strengthening and validating the approaches developed in this thesis. I am deeply grateful to both for their confidence in my work and for the invaluable contribution they made to this research journey.

I would also like to express my deepest appreciation to all my friends in Montreal, whom I truly consider my second family. Being far from home, their presence, kindness, and unwavering support became a source of comfort, strength, and belonging throughout this long and demanding journey. They stood by me during moments of doubt, celebrated my small and big victories, and reminded me to keep going when the road felt overwhelming. Their laughter, encouragement, and genuine care turned challenging days into bearable ones and joyful moments into unforgettable memories. This journey would not have been the same without them, and I will forever carry their friendship with gratitude in my heart.

RÉSUMÉ

Le secteur industriel contribue à environ 30 % des émissions mondiales de CO_2, principalement en raison de procédés énergivores et à forte intensité carbone tels que la sidérurgie, la production de ciment et la fabrication de produits chimiques. Les options de décarbonation industrielle, notamment le captage du dioxyde de carbone et l'utilisation de matières premières biosourcées, peuvent jouer un rôle essentiel dans la réduction de ces émissions, mais la mise en œuvre de ces options demeure limitée par des raisons de coûts d'investissement et d'exploitation élevés, un niveau de maturité technologique souvent limitée à 6-7, autrement dit les technologies doivent encore répondre à des exigences opérationnelles réelles avec un déploiement réussi, et des exigences de conception complexes. Une décarbonation industrielle économiquement viable exige de repenser les cadres traditionnels de conception des procédés afin d'intégrer, notamment, les énergies renouvelables, le captage et le stockage du dioxyde de carbone, ainsi que les flux circulaires de ressources, au sein de systèmes adaptatifs pour lesquels la numérisation pourrait jouer un rôle clé dans l'optimisation des opérations. Cependant, l'optimisation de la conception des procédés demeure une tâche complexe et multidisciplinaire, limitée par la rigidité et les exigences computationnelles des approches conventionnelles fondées sur la physique ou sur des modèles empiriques. Cette recherche doctorale répond à ces limitations en développant méthodologie reposant sur une approche hybride fondée sur l'intelligence artificielle, visant à accélérer la conception de procédés et de matériaux industriels à faibles émissions, tout en conservant une supervision complète par des experts humains.

La méthodologie d'optimisation et de conception proposée, centrée sur l'expertise humaine, intègre l'apprentissage automatique interprétable (*communément appelé en anglais, Interpretable Machine Learning, IML*), l'intelligence artificielle explicable (*communément appelé en anglais, Explainable Artificial Intelligence, XAI*) ainsi que des approches génératives de l'IA et d'apprentissage par renforcement (*RL*) assistées par la physique, afin d'articuler l'intelligence issue des données avec la modélisation physique. La méthodologie exploite les capacités prédictives, descriptives et génératives de l'IA/ML tout en les combinant à une analyse de causalité reposant sur l'analyse formelle de concepts et les réseaux bayésiens causaux pour accélérer la conception. Trois méthodes fondamentales ont été élaborées : (1) la première repose sur des modèles d'ensemble IML/XAI assurant une prédiction fiable et une interprétation transparente pour la conception de procédés et de matériaux ; (2) la deuxième est fondée sur l'IA générative tenant

compte des principes physiques et chimiques pour la découverte de solvants et matériaux durables ; et (3) la troisième porte sur l'optimisation par apprentissage par renforcement causal incrémental (CIRL) assurant une prise de décision adaptative intégrée aux simulateurs. Ces méthodes ont été validées à travers des études de cas industriels liées à la décarbonation, notamment le captage du dioxyde de carbone et la valorisation de la lignine. Les résultats obtenus confirment la performance supérieure des méthodes proposées par rapport à celles existantes dans la littérature, en atteignant des niveaux de précision prédictive nettement plus élevés, tout en préservant l'interprétabilité et la transparence des modèles. La méthode générative a démontré une cohérence entre les matériaux synthétisés et les indicateurs de performance ciblés des solvants, notamment la solubilité et le caractère écologique. De même, les méthodes d'optimisation développées ont permis de concevoir des procédés présentant des réductions significatives des coûts et des émissions de CO_2 associés à leur captage. Ces avancées ont été obtenues avec une accélération et une efficacité computationnelles notables, principalement attribuées à l'intégration de l'analyse et du raisonnement causals comme composantes fondamentales des méthodes proposées.

L'originalité de cette recherche réside dans l'unification des connaissances fondées sur la physique et de l'IA causale afin de développer des méthodes d'optimisation interprétables, évolutives et généralisables. Au-delà des avancées méthodologiques - telles que la modélisation par substituts d'ensemble et l'apprentissage causal- ce travail propose des outils concrets, notamment une base de données publique de matériaux, un système structuré d'évaluation du caractère écologique des matériaux, ainsi que et des règles de conception réutilisables/généralisables. L'ensemble de ces contributions offre une base robuste pour une conception industrielle numérique, durable et axée sur les données, favorisant une transition accélérée vers les objectifs de carboneutralité.

ABSTRACT

The industrial sector accounts for approximately 30% of global CO_2 emissions, primarily due to energy-intensive and carbon-intensive processes such as steelmaking, cement production, and chemical manufacturing. Industrial decarbonization options, including carbon capture and the use of bio-based feedstocks, can play a critical role in reducing these emissions; however, their implementation remains constrained by high capital and operating costs, technological maturity levels often limited to TRL 6-7, meaning that technologies must still demonstrate reliable performance under real operational conditions, and complex design requirements. Economically viable industrial decarbonization requires rethinking traditional process design frameworks to integrate renewable energy sources, carbon capture and storage, and circular resource flows within adaptive systems, where digitalization can play a key role in operational optimization. Nevertheless, process design optimization remains a complex and multidisciplinary task, limited by the rigidity and computational demands of conventional physics-based or empirical modeling approaches.

This doctoral research addresses these limitations by developing a hybrid artificial intelligence–based methodology aimed at accelerating the design of low-emission industrial processes and materials while maintaining full expert human oversight. The proposed human-centered optimization and design framework integrates Interpretable Machine Learning (IML), Explainable Artificial Intelligence (XAI), generative AI approaches, and physics-assisted Reinforcement Learning (RL) to combine data-driven intelligence with physical modeling. The methodology leverages the predictive, descriptive, and generative capabilities of AI/ML while incorporating causal analysis based on Formal Concept Analysis and causal Bayesian networks to accelerate design.

Three core methods were developed: (1) an ensemble-based IML/XAI framework ensuring reliable prediction and transparent interpretation for process and material design; (2) a physics- and chemistry-informed generative AI approach for the discovery of sustainable solvents and materials; and (3) an incremental Causal Reinforcement Learning (CIRL) optimization framework enabling adaptive decision-making integrated with process simulators. These methods were validated through industrial decarbonization case studies, including carbon capture and lignin valorization. Results demonstrate superior performance compared to existing methods in the literature,

achieving significantly higher predictive accuracy while preserving interpretability and model transparency. The generative approach showed strong consistency between synthesized materials and targeted solvent performance indicators, particularly solubility and environmental greenness. Likewise, the developed optimization methods enabled the design of processes achieving significant reductions in both costs and CO_2 capture-related emissions. These advancements were achieved with notable computational acceleration and efficiency, primarily attributable to the integration of causal analysis and reasoning as foundational components of the proposed methods.

The originality of this research lies in the unification of physics-based knowledge and causal AI to develop interpretable, scalable, and generalizable optimization methods. Beyond methodological contributions, such as ensemble surrogate modeling and causal learning, this work delivers practical tools, including a public materials database, a structured system for evaluating the environmental greenness of materials, and reusable/generalizable design rules. Together, these contributions provide a robust foundation for digital, sustainable, and data-driven industrial design, supporting an accelerated transition toward carbon neutrality.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LISTE OF SYMBOLS AND ABBREVIATIONS

AI	Artificial intelligence
ANFIS	Adaptive neuro fuzzy inference system
ANN	Artificial neural network
BLR	Bayesian linear regression
BN	Bayesian network
BPNN	Back propagation neural network
CCS	Carbon capture and storage/sequestration
CCU	Carbon capture and utilization
CCUS	Carbon capture, utilization and sequestration
CFD	Computational fluid dynamics
CIRL	Causal incremental reinforcement learning
CNN	Convolutional neural network
DAC	Direct air capture
DAG	Directed acyclic graph
DEPG	Dimethyl ethers of polyethylene glycol
DDPG	Deep deterministic policy gradient
DFT	Density functional theory
DNN	Deep neural network
DQN	Deep-Q network
DT	Decision tree
EE	Energy efficiency
EOR	Enhanced oil recovery
ESA	Electrical swing adsorption

FEM	Finite element method
GA	Genetic algorithm
GAN	Generative adversarial network
GBDT	Gradient boosted decision tree
GIS	Geographic information system
GPR	Gaussian process regression
GPT	Generative pretrained transformer
GRU	Gated recurrent unit
HSP	Hansen solubility parameter
IEA	International energy agency
IGCC	Integrated coal gasification combined cycle
IML	Interpretable machine learning
IoT	Internet of things
IPC	Intelligent predictive controller
KPI	Key performance indicator
LAD	Logical analysis of data
LCA	Life cycle assessment (analysis)
LCIA	Life cycle impact assessment
LSTM	Long short-term memory
MARS	Multivariate adaptive regression splines
MC	Monte Carlo
MCDA	Multi-criteria decision analysis
MDEA	Methyl diethanolamine
MEA	Mono-ethanol amine

MILP	Mixed integer linear programming
ML	Machine learning
MM	Molecular mechanics
MPC	Model predictive controller
NETL	National energy technology laboratory
PCC	Post combustion capture
PSA	Pressure swing adsorption
PSO	Particle swarm optimization
PZ	Piperazine
QM	Quantum mechanics
RBF	Radial basis function
RF	Random forest
RL	Reinforcement learning
RSM	Response surface methodology
RST	Rough set theory
SVM	Support vector machine
SVR	Support vector regression
TRL	Technology readiness level
TSA	Thermal swing adsorption
VAE	Variational autoencoder
VSA	Vacuum swing adsorption
XAI	Explainable artificial intelligence
XGB	Extreme gradient boosting

LIST OF APPENDICES

CHAPTER 1 INTRODUCTION

Industrial decarbonization is among the most vital strategies for achieving *Net-Zero* goals in the global fight against climate change. The industrial sector contributes approximately 30% of global CO_2 emissions, primarily from high-temperature processes such as steelmaking, cement production, and chemical manufacturing [1]. These activities are inherently carbon-intensive due to their reliance on fossil fuels for both energy and feedstocks. To address this, several decarbonization approaches are being pursued. Common pathways include industrial electrification using renewable energy, carbon capture, utilization and storage (CCUS), green hydrogen fuel substitution, bio-based feedstocks, and advanced energy efficiency measures [2]. Each pathway offers distinct advantages and challenges, with feasibility depending on sector-specific needs, regional energy infrastructure, and technological maturity. For instance, while green hydrogen is a promising solution for decarbonizing steel and ammonia production, its deployment is currently constrained by high costs and limited infrastructure [3].

According to many stakeholders, researches and global research and advisory firms, e.g. International Energy Agency (IEA), Gartner, McKinsey Global Institute etc., the importance of these pathways extends beyond environmental benefits [4], [5]. They are also vital for economic resilience, regulatory compliance, and technological innovation [6]. Deep industrial decarbonization is indispensable to achieving net-zero emissions and meeting the targets set by international climate agreements [7]. At the same time, the practical deployment of these strategies must contend with feasibility concerns, such as integration into existing plants, capital intensity, and energy requirements [8].

To make industrial decarbonization truly feasible and cost-effective, the development of new and optimized industrial systems is imperative. Most existing industrial facilities were designed in an era where carbon emissions were not a central concern. Modern decarbonization, however, requires rethinking of process architectures to integrate low-emission technologies, advanced digital control systems, and circular resource flows [9]. Technologies such as digital twins, AI-driven optimization, and modular reactors/equipment's are helping engineers design processes that are not only more technically efficient and cost effective but also dramatically lower in carbon intensity [10]. Furthermore, process integration, including coupling operations with distributed energy resources, leveraging waste heat, or co-locating industries for resource sharing, can unlock

synergies that enhance overall system performance while increasing its energy efficiency and reducing emissions [11].

Process design plays a crucial role in this industrial transition. Redesigning industrial processes to accommodate alternative energy sources like electricity or hydrogen, integrating carbon capture loops, and shifting toward closed-loop manufacturing systems can significantly cut emissions per unit of output [12]. By decreasing the carbon intensity of industrial production systems through innovations in materials handling, energy usage, and system layout, industries cannot only align with net-zero emissions goals but also enhance their long-term competitiveness [13]. The joint efforts of process engineers and system designers, in collaboration with data and AI scientists, play a pivotal role in developing scalable and economically viable pathways with reduced carbon footprint. While advances in physics-based and AI -driven methods as modeling frameworks have accelerated process and material design optimization, the associated tasks remain complex and computationally demanding. Persistent challenges in model generalization, interpretability, and integration limit the widespread adoption of current AI-based approaches, highlighting the need for hybrid modeling frameworks. These frameworks combine the capabilities of both physics-based and AI-driven modeling and optimization techniques ensuring acceptable accuracy and physical realism while being transparent while optimizing industrial systems. Accordingly, process experts will have the ability to make new valid decisions to continually enhance feasibility, energy efficiency and improve carbon footprint of industrial production systems. The following subsection outlines the problem statement, highlighting the limitations of existing modeling and optimization methods and the emerging opportunities for AI and hybrid modeling to accelerate process and material design optimization.

1.1 Problem Statement

Achieving large-scale industrial decarbonization remains a tough challenge due to the limited maturity and readiness of current and emerging decarbonization technologies. Despite growing research and policy efforts, the transition toward low-carbon industrial systems is constrained by interconnected technological, methodological, and data-related barriers that hinder scalability, cost-effectiveness, and integration across value chains. The technological challenges appear from the inherent complexity of industrial systems, which operate as interconnected networks of energy- and carbon-intensive subsystems. For instance, CCUS technologies face high costs, significant

energy demands, slow kinetics, and reduced efficiency in the presence of flue gas contaminants such as moisture, and SOx, which also accelerate capture/utilization materials degradation [14]–[16]. In CCU systems, slow kinetics refers to limitations in mass transfer and/or reaction rates that govern CO_2 capture and conversion. In absorption processes, this may involve slow chemical reactions between CO_2 and solvents, while in adsorption systems it can reflect diffusion limitations within porous adsorbent materials; in utilization pathways, catalytic reaction rates may also become limiting factors. Moisture in flue gas further complicates operation by competing with CO_2 for active sites, diluting solvents, increasing regeneration energy demand, and accelerating material degradation through hydrolysis or oxidation. These effects lower process efficiency, increase operating costs, and add complexity to system design. These CCU materials encounter the challenges of costly synthesis, and energy-intensive regeneration. Additional complexity arises from time-varying operating conditions and energy price uncertainty, necessitating explicit design under uncertainty. Hence, CCUS design becomes a large-scale, multi-objective combinatorial optimization problem typical of industrial decarbonization systems. CCUS system design can be formulated as a multi-objective combinatorial optimization problem, where conflicting objectives such as cost, emissions, efficiency must be optimized simultaneously over both discrete and continuous decision variables. The combinatorial nature arises from selecting and configuring system components (e.g., materials, process structures, equipment), while continuous variables include operating conditions such as temperature, pressure, and flow rates. This results in a mixed-integer, nonlinear optimization problem (MINLP) with a large search space, reflecting the need to jointly optimize system structure and operating conditions under multiple performance criteria. The methodological challenges arise from the excessive computational resources and domain expertise required to model, simulate, and optimize complex processes. Traditional mechanistic and data-driven models are often limited by scalability, long computation times, and the difficulty of constructing accurate objective functions and constraints. Moreover, conventional optimizers lack adaptability and learnability, making it difficult to address system non-linearities, uncertainties, and multi-objective trade-offs in an automated, generalizable manner. It is worth nothing that data serves as a key enabler to help overcome methodological challenges. However, there are still different data-related challenges. For example, industrial data are often heterogeneous, incomplete, or inaccessible due to privacy and siloed operations, limiting the potential of data-driven modeling. Poor data quality and inconsistent formats undermine model accuracy and transferability. Effective

data fusion, cleaning, and integration are, therefore, essential to enable robust predictive and optimization frameworks.

AI, as one of the advanced data-driven techniques, can provide flexible solutions to these challenges in an accelerated way. Yet, they may lack physical consistency, causal interpretability, and reliable generalization. This underscores the need for gray-box (hybrid) frameworks [17] that integrate physics-based simulations, historical data, AI techniques, and human expertise. Such frameworks reduce computational cost, handle nonlinearity, uncertainty, and multi-objective trade-offs more effectively, and enhance transparency and scalability. They also remain constrained by physical laws and data fidelity. Therefore they augment, rather than replace, engineering judgment.

1.2 Research Objectives

The general objective of this doctoral research is to develop a human-centric, causal and physics-guided AI generalized methodologies integrating interpretable ensemble learning, generative modeling, and simulation-based optimization to accelerate the design of low-emission industrial processes and materials for industrial decarbonization. All AI capabilities are leveraged while preserving physical realism, thereby enabling sustainable industrial production value chains and the achievement of net-zero targets. Figure 1.1 illustrate a simplified schematic for the approach followed to build these AI-assisted methodologies. This research adopts a human-centric AI approach, where expert supervision guides every stage from data collection to knowledge extraction and optimized design generation. Human expertise ensures high data quality through informed cleaning, imputation, fusion, and feature engineering to produce analysis-ready data that fuels accurate modeling and optimization. AI techniques are employed to reduce expert workload and accelerate decision-making, while experts oversee model selection, development, training, and validation. Below is a brief description of the specific objectives of this doctoral research, presenting the links between them in a clear and coherent manner.

Specific Objective (S.O.) #1: *Development of interpretable ensemble AI methodologies for reliable classification and selection of industrial processes and materials for decarbonization applications.*

Specific Objective (S.O.) #2: *Development of a physics- and chemistry-guided generative AI methodology for the discovery of novel green materials and processes to support industrial decarbonization.*

Specific Objective (S.O.) #3: *Development of a novel physics- guided causal incremental reinforcement learning methodology for accelerated simulation-based optimization of eco-friendly industrial processes design*

The three specific objectives form a coherent, interconnected and cumulative research trajectory addressing the full AI-driven decision-making pipeline for industrial decarbonization as illustrated in Figure 1.2. The first establishes an interpretable decision-support foundation using explainable ensemble AI for accurate classification, prediction, and selection of materials and processes. The second builds on this knowledge to enable discovery through physics- and chemistry-guided generative AI that produces novel, feasible, and sustainable candidates. The third translates discovery into action via physics-guided, causality-aware simulation-based optimization using incremental reinforcement learning, enabling rapid and physically consistent design of both existing and newly generated systems. Together, these objectives form an integrated framework that ensures interpretability, feasibility, and optimality for industrial decarbonization applications.

It is worth noting that the third objective (**S.O. #3**) implements the capabilities developed in **S.O. #1** and **S.O. #2** within a unified decision-making framework. While **S.O. #1** provides interpretable predictive models for classification and knowledge extraction, and **S.O. #2** enables the generation of physically and chemically consistent candidates, **S.O. #3** translates these outputs into optimized process designs through simulation-based learning and optimization. The proposed causal incremental reinforcement learning (CIRL) framework leverages insights from **S.O. #1** to guide state-space reduction and policy learning, while integrating candidates generated in **S.O. #2** as feasible inputs and/or constraints. This establishes a closed-loop pipeline linking prediction, discovery, and optimization, where **S.O. #3** acts as the decision-making module that consolidates and exploits prior knowledge to produce physically consistent and cost-effective designs.

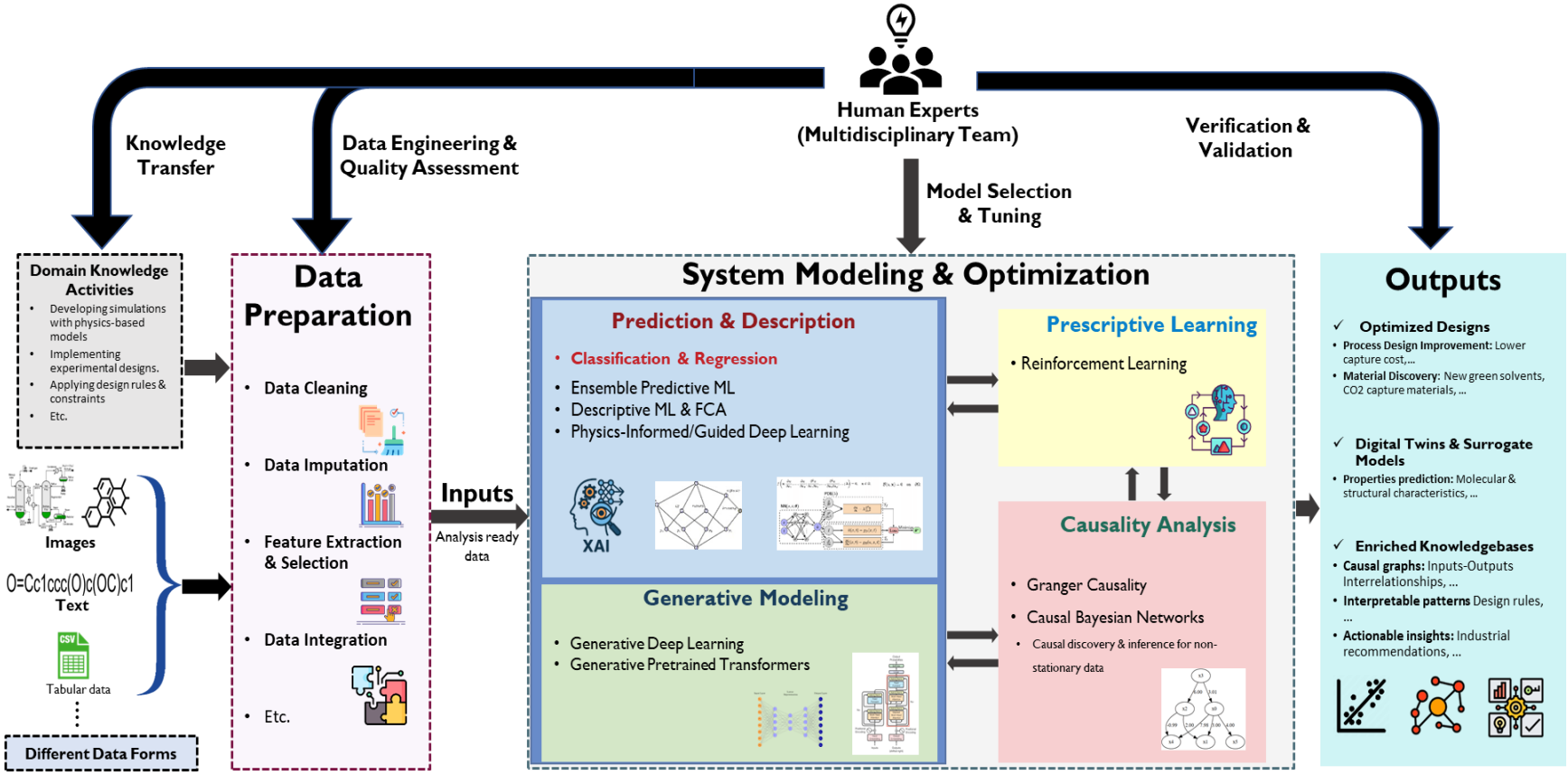


Figure 1.1 AI-assisted generalized methodologies for accelerating sustainable processes and materials design optimization

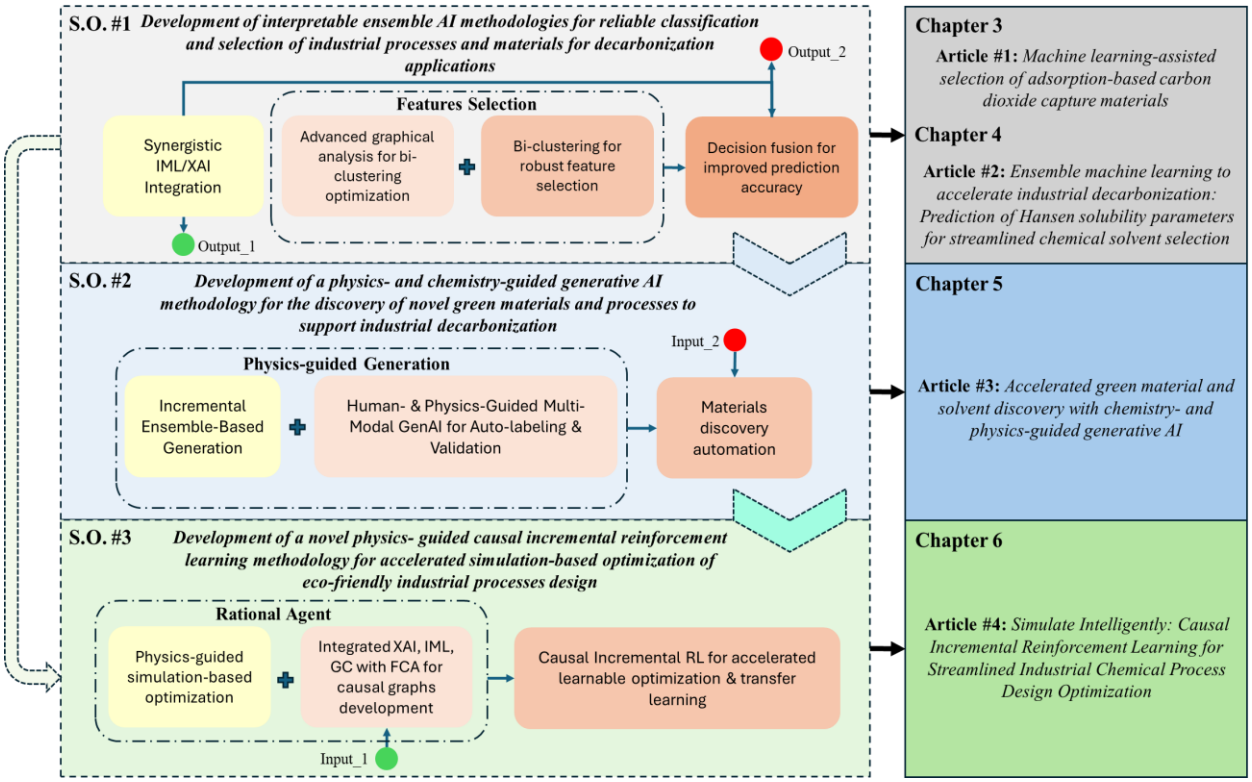


Figure 1.2 Illustration of linked objectives and contributions of the doctoral research

1.3 Research contributions, originality and outcomes

The originality and contributions of this doctoral research cover both methodological and technological aspects. The following is a summary of the methodological contributions:

- **A hybrid ensemble interpretable ML-explainable AI (IML-XAI) methodology** that combines high predictive accuracy with transparency by integrating intrinsic interpretability and post-hoc explainability. The methodology captures both global and local data patterns, extracts actionable design rules, and delivers fast, robust, and explainable classification and regression models validated on industrial decarbonization case studies.
- **A bi-clustering-assisted decision-fusion ensemble learning methodology** that integrates robust feature selection and dimensionality reduction with shallow and deep learners to achieve high predictive accuracy, interpretability, and computational efficiency. A novel non-linear graphical optimization strategy is introduced to tune biclustering parameters and support scalable multi-output prediction.

- **An incremental, physics- and chemistry-guided ensemble generative AI methodology** that integrates generative adversarial networks (GANs), variational autoencoders (VAEs), and large language models (LLMs) within a constrained generation framework. The methodology incorporates human preferences and introduces a novel auto-labeling and validation strategy combining physics-based and data-driven models to ensure chemical validity, physical feasibility, and alignment with sustainability-oriented key performance indicators (KPIs).
- **A novel simulation-based optimization methodology based on causal incremental reinforcement learning (CIRL)**, in which a reinforcement learning agent directly interacts with physics-based simulators. The approach integrates IML, XAI, Granger causality, formal concept analysis, and causal Bayesian networks to construct informed causal models that reduce search space, accelerate convergence, and enable transfer learning across industrial process design scenarios.

These contributions advance the state of the art in AI for industrial systems where causality analysis was central to uncovering relationships between design variables and KPIs across system levels enhancing interpretability and decision-making. Through multi-level causal reasoning, spanning IML/XAI, LLM-based reasoning, and causal Bayesian networks, the framework delivered physically consistent, explainable, and transferable insights for sustainable process and material design optimization enabling efficient decarbonization solutions. As a result, this doctoral research produced diverse outcomes, summarized below:

- Four main peer-reviewed journal articles documenting the contributions and original methodologies.
- Thesis compiling all the research findings and the articles
- Computational tools developed throughout the research based on open-source packages for AI-assisted process and material design optimization

The **first objective** resulted in two peer-reviewed publications in *Journal of Environmental Chemical Engineering*, *Elsevier* and *Digital Chemical Engineering, Elsevier*, detailing the development and validation of combined IML-XAI and ensemble learning methodologies. The **second objective** yielded an article published in *Artificial Intelligence Chemistry, Elsevier*, presenting the proposed physics- and chemistry-aware ensemble generative AI framework for

sustainable material discovery. The **third objective** led to a journal publication in *Journal of Environmental Chemical Engineering*, *Elsevier* and a full conference paper presented at FOCAPD 2024 and published in *Systems and Control Transactions Journal* (Appendix E), demonstrating the causal incremental reinforcement learning (CIRL) framework for accelerated simulation-based optimization. Additionally, a comprehensive review on AI/ML applications in carbon capture, utilization, and storage (CCUS), as a representative of industrial decarbonization systems, was published in *Science of the Total Environment, Elsevier*. It identifies research gaps that guided the methodologies developed in this thesis. The complete list of related publications and presentations is provided in Appendix A. Several reusable tools were developed, including a classification and a regression-based tools for material selection helping process optimization, a generative AI tool for novel green materials discovery, and an accelerated causality-assisted process design optimization tool. These tools were validated on carbon capture systems and designed to be transferable to related applications such as carbon utilization and biomass valorization with minimal adaptation.

1.4 Thesis organization

This thesis comprises seven main chapters. Chapter 1 introduces the research background, problem statement, objectives, contributions, originality, and key outcomes. Chapter 2 reviews prior AI/ML applications in modeling, simulation, and optimization of industrial systems, emphasizing industrial decarbonization. Chapter 3 presents the first published article on the IML-XAI methodology for accurate and transparent data classification, applied to selecting efficient carbon capture adsorbents. Chapter 4 details the second published article introducing ensemble-based ML modeling and biclustering-based feature selection, applied to predicting solubility parameters from SMILES representations and selecting optimal solvents for CO_2 capture and lignin dissolution. Chapter 5 addresses the second objective through a physics/chemistry-aware ensemble GenAI methodology for reliable data generation and a hybrid framework for validating generated materials, demonstrated in generating novel green solvents for carbon capture and lignin valorization. Chapter 6 presents the full CIRL methodology for simulation-based optimization, integrating formal concept analysis, IML/XAI, and Granger causality to guide causal Bayesian networks. The methodology is validated on real flue gas carbon capture processes through continuous interaction between the RL agent and physics-based simulators, ensuring reliable optimization and transferable causal insights. Chapter 7 concludes the thesis with a general discussion, conclusions, and directions for future research.

CHAPTER 2 LITERATURE REVIEW

Currently, climate change and global warming are considered the most critical global challenges, drawing the urgent attention of researchers, decision-makers, and leaders worldwide. The rising threats of droughts, heatwaves, and sea level rise, caused by excessive greenhouse gas (GHG) emissions, are driving nations to seek reliable and scalable solutions to prevent catastrophic environmental, social, and economic consequences. This urgency has been the focus of major international efforts, such as the *Kyoto Protocol*, the *Paris Agreement*, and global summits like COP26, which collectively emphasize the need to limit global temperature rise to well below 2 °C, ideally 1.5 °C, above pre-industrial levels [18]. Achieving this ambitious goal, particularly within the 2025–2030 timeframe, requires rapid and large-scale transformation of emissions-intensive systems, chief among them being the industrial sector, through the adoption of net-zero and sustainable systems rooted in the principles of a circular economy [19]. Industrial decarbonization represents one of the most critical pathways toward achieving Net-Zero emissions and advancing the global effort against climate change

Table 2.1 introduces a brief overview on the technology readiness level of different industrial decarbonization pathways, the preferred field of application for each pathway and technological challenges facing their wide adoption in the current time. These data were collected from various sources including research articles, white papers and technical reports of the global firms in addition to several meetings with industrial decarbonization experts [20]–[23].

Waste heat recovery and process optimization, offer near-term emissions reductions at relatively low cost, while others, such as carbon-negative cement or electrochemical conversion of CO_2, remain in earlier stages of development [24]. Consequently, a multi-pathway, sector-specific approach, backed by policy incentives and cross-industry collaboration, is essential to accelerate industrial transformation and deliver lasting climate benefits [25]. On the technological side, one way for catalyzing industrial decarbonization wide adoption globally is to optimize its associated value chain in order to be cost-effective [20].

Table 2.1 Common Industrial Decarbonization Pathways

Pathway	**Main Sectoral Relevance**	**TRL**	**Feasibility Challenges**
Electrification with renewable energy	Low and medium temperature processes	7-9	Grid capacity Renewables availability Limited to processes under 1k ºC
Green hydrogen utilization	Steel and Petrochemicals industries	5-7	High costs Low infrastructure availability Energy intensive
CCUS	Cement, Steel and (Petro)Chemical industries as well as Refineries	6-8	High energy requirement High costs Expensive and complex storage logistics
Bio-based feedstock and fuels utilization	Power generation plants, Chemical industries and Refineries	5-7	Limited availability Variable biomass properties Land use issues
Process optimization (energy efficiency)	All sectors	8-9	Limited by system constraints
Material substitution	Cement industry	5-7	Regularity barriers Performance uncertainty Alternatives availability
Waste heat recovery	Cement, Steel, (Petro)Chemicals, Pulp & paper industries	8-9	High integration complexity May require co-location or industrial clustering

Process and material design optimization is central to industrial decarbonization, enabling the integration of alternative energy sources, carbon capture, and circular manufacturing systems to reduce emissions and enhance competitiveness. The existing methods for process and material design and optimization are classified into experimental and computational approaches, with the latter being either physics-based or data-driven. Experimental methods rely on iterative work at lab and pilot scales to enhance performance but are often limited by high costs, time consumption, and

their inability to capture interactions among all input variables. To address these issues, statistical techniques are often integrated to design systematic experiments and develop polynomial models, linking input variables to key performance indicators. However, these data-driven methods still face limitations when applied to complex industrial production systems due to simplified assumptions restricting their ability to fully represent real data distributions and capture the entire optimum solutions search space. Besides these experimental-based optimization methods, computational methods play key roles in developing scalable, economically viable, and environmentally sustainable industrial decarbonization solutions. The mechanistic or physics-based computational methods are robust and sophisticated. They differ according to the model details, system scope and the industrial application [26]. These different modeling scales are considered to serve multi-criteria decision making for different industrial processes. For instance, on the molecular and micro levels, more complex models and techniques considering multi-physics modeling/simulation are employed to investigate transport phenomena and molecular interactions. In the macro level modeling case, the techniques and tools used are more focused on the high-level processes and lumped where high level mass/energy balances, equilibrium and kinetics are investigated/considered. Figure 2.1 illustrates common mechanistic methods and tools used for both process and material design optimization based on the level of modeling problem. These modeling and simulation techniques or tools are then associated with traditional optimizers to obtain optimal KPIs or designs. The optimizers commonly used in these cases are different known operation research (OR) techniques including linear programming (LP) and mixed integer linear programming (MILP) [27]. The objectives and objective functions are case specific depending on the problem itself. For instance, in the case of optimizing a process on the macro level, the objectives include energy consumption and overall cost minimization while maximizing the process performance, e.g. CO_2 absorption rate in a carbon capture facility [28]. As a result, human expertise is inevitable while constructing the objective function and defining the problem constraints.

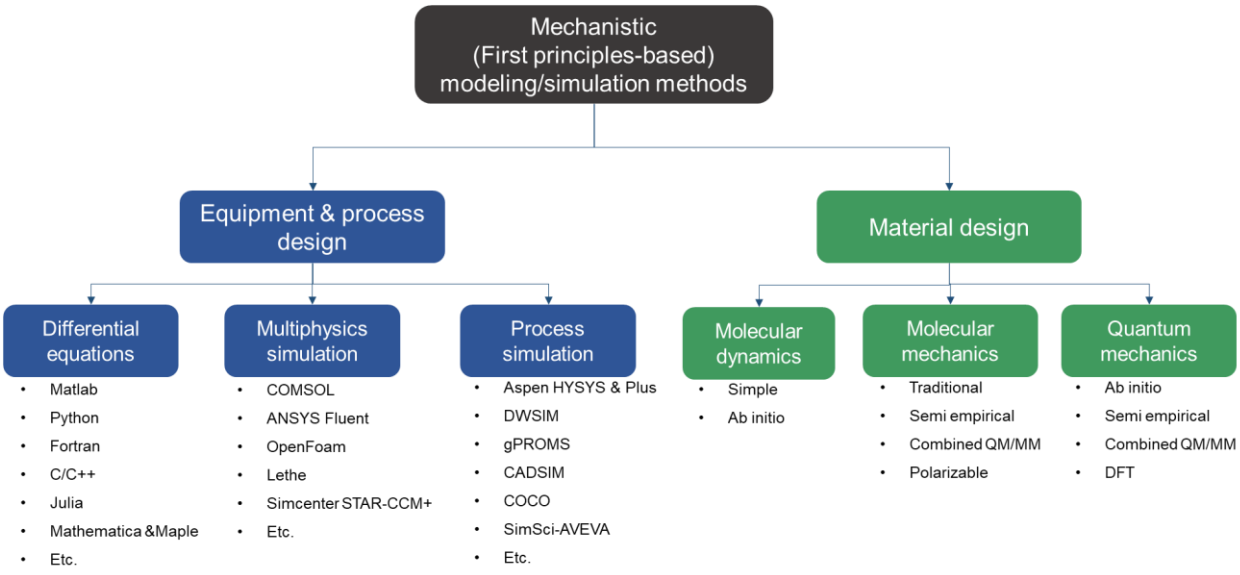


Figure 2.1 Mechanistic methods for process/material modeling and simulation

Despite being robust, these mechanistic techniques suffer from several limitations. For instance, they are time consuming during both implementation and calculation phases. They also need expensive computational resources which consume a lot of energy as well. These two limitations arise mainly in the case of molecular and micro level modeling. In the case of macro level modeling, technologically complex processes can face the obstacle of limited expertise which may hinder its optimization. This is due to being large combinatorial optimization problem requiring complex objective function(s) construction and the definition of thousands or even hundreds of thousands of constraints. Thus, longer computational times and expensive tools are also required in this case to perform the optimization.

One of the emerging data-driven techniques in the field of process/material design optimization is the artificial intelligence (AI) and its related machine learning (ML) technologies. It is worth noting that these techniques have been widely used in the case of processes operations optimization especially reinforcement learning (RL) [29]. The attempts of building AI/ML-based surrogate models to accelerate process/material design optimization will be discussed in the upcoming chapter. This chapter introduces an overview on the applications of AI/ML in various fields with special emphasize on industrial decarbonization pathways. The following subsections present some of the successful case studies where different AI techniques were used for process monitoring, control and optimization in net-zero emissions systems, i.e. industrial decarbonization pathways.

This quick survey covers the applications at the macroscale, microscale, e.g. computational fluid dynamics (CFD) modeling, and molecular scale including chemical reactions/properties prediction and chemical space exploration for the synthesis/discovery of new materials.

2.1 Energy efficiency and biorefinery pathways

The use of AI has been exploded over the past few years in different energy efficiency (EE)-related problems, driven by advanced and relatively inexpensive computing infrastructures and the availability of datasets [30]–[46]. Machine learning (ML) and deep learning (DL) have become at the heart of AI, and have been proven as promising technologies for EE improvement [17], [47]. Many industries start to leverage the capabilities of ML and DL techniques to optimize the use of energy across various sectors and give real opportunities to address the EE-related challenges [48]. These techniques are capable of accepting heterogeneous sources of data such as numbers, text, images, and videos to build predictive and descriptive models to monitor, evaluate and manage energy consumption [36], [44], [49]–[53]. Moreover, these models can help the process operators make the right decisions in situations that affect the process' EE [54]. Commercial and open-source software packages based on ML and DL tools are used in the industry to reduce energy usage during peak hours and detect equipment failures before they occur [55].

AI technologies have a large potential to assist biorefineries to achieve disruptive transformations that strengthen their global competitiveness as well as contribute to a carbon-neutral economy and to circular-economy solutions. It allows process experts to learn faster from feedback, deal more effectively with complexity, and make better sense of abundant data available. A growing number of initiatives are exploring how AI can create new opportunities to address some of the world's most important challenges [49]. However, this sector is currently lagging behind in some fields of AI utilization. AI technologies are developing very fast while adoption in industrial scale biorefineries is often challenging.

Regarding the utilization of AI/ML techniques in biorefineries, on lab scale or through simulation, these techniques were almost applied in all the different parts of the facility. These parts include, but not limited to, the conversion process and materials/products supply chains. For instance, researchers applied these techniques for the prediction and optimization of the products yield from various biochemical and thermo-chemical process. In addition, selection of the best feedstock for the process, profitability maximization and cost minimization were investigated using different

techniques including ANN, CNN and *k*-NN. All of these research studies were summarized/mentioned in many surveys previously done to evaluate the application of the different types of machine learning in biorefinery systems for bioenergy production as well as the prediction of the properties of biofuels, e.g. biodiesel [56]–[60]. Additionally, the utilization of machine learning in biological wastewater treatment as a combined approach for bioenergy production and wastewater treatment was also covered previously in [61]. As it can be observed from these surveys, the vast majority of the studies on the application of machine learning in biorefinery systems are concerned about bioenergy/biofuels production. Hence, there is a promising potential for machine learning to applied in biorefineries producing other value-added materials. Besides, the research was focused on the predictive ability of the techniques while the focus on the descriptive and prescriptive capabilities was relatively poor.

In this work, the terms descriptive, predictive, and prescriptive are used to denote complementary modeling capabilities rather than a strictly sequential hierarchy. Descriptive models aim to characterize and interpret system behavior (e.g., identifying key variables, correlations, readable patterns, or causal relationships). Predictive models estimate system outputs for unseen conditions, while prescriptive models leverage these predictions to recommend optimal decisions given specific objectives and constraints. While predictive performance has been the primary focus of many existing studies, the descriptive and prescriptive dimensions are often underdeveloped in practice. In particular, high predictive accuracy does not necessarily imply interpretability or actionable decision support. Therefore, enhancing descriptive insight (e.g., via interpretable ML and causal analysis) and strengthening prescriptive capabilities (e.g., via optimization or reinforcement learning) are essential to transform predictive models into reliable tools for industrial decision-making.

2.2 Renewable (green) hydrogen utilization pathway

Many research studies refers to hydrogen as the future fuel due to being clean and its combustion emission is just water vapor [62]. It is described by a color code depending on the production method. Five main color codes have been attributed; namely, black/brown, grey, blue, turquoise and green. Black/brown hydrogen is the one produced through coal gasification where a lot of carbon dioxide emissions are released without proper capture [63]. Whereas, blue hydrogen is obtained through the famous natural gas steam reforming method followed by carbon capture [64].

On the other hand, hydrogen obtained through water electrolysis is considered as the green one owing to the clean energy utilization and the absence of harmful carbon dioxide emissions [65]. The sources of energy or electricity for this process are ideally solar, wind and hydroelectric energy [66]–[68]. As a new research trend for the acceleration of hydrogen production, especially green renewable hydrogen, machine learning concepts/techniques are being applied to model systems of hydrogen production and storage [69], [70]. However, regarding the use of ML to investigate the transportation of hydrogen is not mature yet due to the sacristy of real data in this field [71]. Hence, other sources or types of data such as the simulated ones and the aid of ML techniques for data generation should be taken into consideration to advance in this important branch in the hydrogen supply chain.

As an example of using ML for the modeling of hydrogen production, in [72], multi-layer feed forward neural network-based surrogate model was developed to investigate/model the performance of solar-aided steam methane reforming employing molten salts. In another study, methanol was used instead of methane for hydrogen production through steam reforming [73]. Aspen simulation combined with Matlab were used to generate a relatively large data set for the training and validation of different ML techniques to predict three important key performance indicators (KPIs). The developed techniques included SVM, decision tree (DT) and gaussian process regression (GPR). Whereas, the outputs or KPIs were hydrogen production rate and cost as well as carbon dioxide emissions. Biomass and plastic wastes gasification for hydrogen production were also investigated by the aid of various machine learning techniques including kNN, SVM and ANN [74]–[78]. With respect to the hydrogen production through water splitting/electrolysis, i.e. hydrogen evolution reaction, the majority of the studies are focusing on the development of an optimum catalyst [79]–[81]. This catalyst should have high catalytic activity besides being stable and cost-effective. For instance, different catalysts for the hydrogen evolution reaction were screened and designed from various MBenes materials through a novel approach combining machine learning and density function theory [82]. Throughout this study, four machine learning techniques, i.e. support vector regression, kernel ridge, random forest and least absolute shrinkage and selection operator were employed. Besides, DFT calculations were used to generate the essential data used for models training, validation and testing. The usage of ML techniques was extended to include the prediction of utilization performance of hydrogen in fuel cells [83]–[86]. For instance, the authors in [87] proposed a combined approach where ANN in conjunction with

genetic algorithm (GA) were used for modeling and optimization of fuel cell performance. It is worth noting that the research on the utilization of ANNs for the prediction and control of solid oxide fuel cell systems started earlier, i.e. almost in 2002 and 2003 [88], [89].

On the other hand, focusing on the hydrogen storage and adsorption, Rahimi M. et al., used SVM to predict hydrogen adsorption on bio-derived carbon materials based on their microstructure [90]. Whereas, metal organic frameworks (MOFs) were investigated, in another study, as efficient hydrogen storage materials by combining multi-scale calculations with machine learning [91]. Besides, MOFs were also screened upon their attainable hydrogen loadings based on storage pressure and temperature through a combined approach of molecular simulations and machine learning [92]. There are other research studies focusing on micro-level, molecular level and material design in the field of renewable hydrogen production and storage [93]–[98]. They almost follow similar methodologies such as the abovementioned ones.

As in the cases of thermochemical and electrochemical hydrogen production, the biological production process was also investigated and optimized with the aid of machine learning [99]–[104]. For instance, simultaneous wastewater treatment and biohydrogen production via dark fermentation was modeled by various ML techniques [99]. Other studies done to model hydrogen production through different processes using machine learning between 2007 and 2017 were reviewed in [105]. This article covered the employment of other techniques such as genetic programing (GP) and fault sematic network (FSN). Nonetheless, as it may be observed, most of the studies are still in the of lab or simulation phases except for few studies such as [106]. In this study, a well established full scale hydrogen production process were modeled and simulated then the data was used to train the ANN model. However, for a serious adoption of machine learning in the process industry, the models should be trained on real data to reflect the reality. This means that a lot of developments are ahead to reach a full scale green/renewable hydrogen production facilities with considerable feasibility and reasonable machine learning inclusion.

2.3 CCUS Pathway

Recently, there is a growing attention about the capturing and sequestration of carbon dioxide as a solution to the global warming problem [107]. Researchers also use the AI/ML techniques to predict the operation of well established technologies such as post combustion capture (PCC) and help in optimizing them [108]–[112]. Adsorption is another technique that is studied extensively

for the purpose of carbon dioxide capture [113]. In this method. Carbon dioxide is trapped, i.e. adsorbed, physically or chemically on the surface and inside the porous structure of an efficient solid adsorbent [114]. Machine learning techniques including ANNs were used to describe this process as in the case of solvent-based carbon capture one [115]–[119]. Besides, there are some research trials where the abilities of these techniques are being leveraged for producing new capture/utilization materials including solvents, adsorbents and catalysts [120]. The range of applications is wide enough to include the estimation/prediction of molecular properties inside the carbon capture system such as the solubility of carbon dioxide in different solvents/absorbents [121].

Sequestration is one of the options to be applied on the carbon captured from any source. In this process, captured carbon is injected and stored in geological wells as a way to achieve the target of carbon negative world [122]. Machine learning was also utilized for the optimization and anomaly detection of this later process [123], [124]. Anomaly or abnormal events in sequestration process have different figures where the leakage of sequestrated carbon is one of them [125]. For example, in [146] a workflow combining machine learning technique, i.e. multivariate adaptive regression splines, and uncertainty quantification was proposed in order to suggest the best monitoring scheme design. The study confirmed that the pressure data is more important than that of temperature and carbon dioxide saturation. This result indicates that the well pressure is more effective variable on carbon dioxide storage and consequently on choosing/designing its monitoring scheme. Besides, the study recommended the utilization of multiple types of monitoring signals in order to reduce the uncertainty. In another study [147], different machine learning techniques were tested for the detection of carbon dioxide leakage from geological sequestration wells. CNNs, LSTM and multi-layer feedforward networks proved their ability to detect the leakage and give early warnings. Thus, corrective actions can be done to avoid the leakage and mitigate its harmful environmental impacts. In general, ML were found very effective in the modeling of sequestrated/stored carbon dioxide behaviour in depleted reservoirs through various trapping mechanisms [126]–[129]. Moreover, ML techniques were also leveraged to model and predict the carbon dioxide behaviour during its transportation between capture and storage units [130]–[134]. The high accuracy of the prediction of different phenomena and flow metering of dense carbon dioxide during pipe transportation is crucial to have a safe and cost effective CCUS system [135], [136]. The modeling/simulation

results will help the expert to prescribe a prevention actions or even a remediation plan in the worst case.

With respect to enhanced oil recovery (EOR) as a well-known route of captured carbon utilization, machine learning took a role in the modeling and optimization of EOR-based processes in several research studies [137]–[142]. In the case of carbon utilization for the production of chemicals or fuels, the majority of the research studies is still on the lab scale. This includes mainly the design of an optimum catalyst for the conversion of carbon dioxide to the aforementioned value-added materials [143]. In this regard, ML techniques were employed to facilitate the search process for this optimum catalyst [144]–[147].

The physical, thermal and other thermodynamic/equilibrium, e.g. solubility, properties of carbon dioxide in different CCUS systems were modeled/predicted using machine learning techniques such as ANN, SVM, BPNN and ANFIS [148]–[152]. The model training during the majority of the studies in this field was done on experimental data [153]–[155].

Regarding combined AI/ML-optimization approach, Figure 2.2 summarizes the methodology of combining AI/ML and traditional optimization techniques in order to obtain the best desired process/material in most of the studies done on applications related to CCUS. This methodology may suffer from the limitations of time consumption and unreliable optimization results in some cases. In fact, this figure can be generalized to include most of the applications mentioned throughout this chapter. It should be noted that one of the limitations of existing methods is that most of the utilized traditional optimizers (OR techniques) lack learnability and flexibility [156]. This leads to longer computational times to cover the search space where eventually global optima may not be reached.

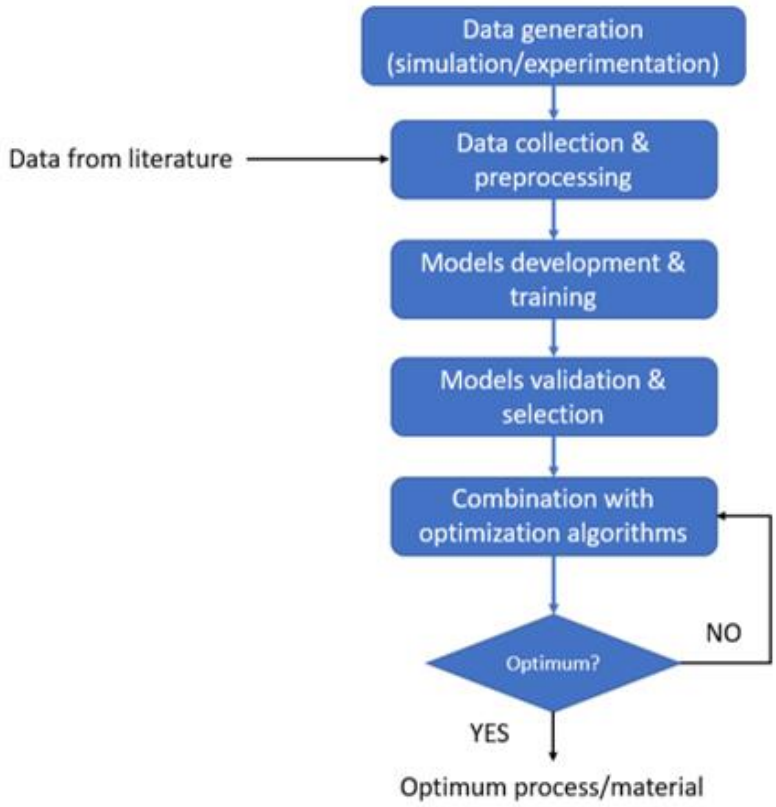


Figure 2.2 General combined ML and optimization algorithms methodology

2.4 Remarks on AI/ML-assisted industrial processes control

This subsection introduces some of the attempts related to the utilization of AI/ML in the field of process control. It surveys the applications made with real data coming from industrial plants as well as the applications that considered simulated processes such as Tennessee Eastman process (TEp) and its associated data.

Pulp and paper mills are among the most abundant plants serving an important industry. These plants also among the large final emitters where different types of pollutants/wastes are produced. Black liquor produced from Kraft's process for pulp and paper production is considered as of these hazardous wastes [157]. This means that it must be managed in a perfect way to avoid its environmental harmful implications. One method of management is to burn it in a controlled combustors, after being concentrated through evaporation, as a source of heat and energy [158]. Chemicals recovery is also an essential step in this controlled management [159]. Black liquor recovery boiler system inside the pulping plant is the part responsible for implementing these tasks

[160]. Accordingly, it needs accurate monitoring and control to perform the operation effectively with the minimum or even zero harmful emissions. On the other hand, thermomechanical pulping (TMP) process also may encounter some energy and environmental problems that can be avoided through optimized process control [161]. In order to achieve this state, AI and other data-driven techniques were developed and applied [162], [163]. For example, in [162], logical analysis of data (LAD) was employed for fault detection and diagnosis in a black liquor recovery boiler system using real data. Fortunately, it showed excellent performance represented by *F1*-score and accuracy reached 100% for the extraction of interpretable patterns of normal and abnormal events. Besides, its training time was comparable with those of other techniques utilized in this study, i.e. support vector machine (SVM) and *k*- nearest neighbour (*k*NN). Yet, it should be noted that this time was relatively high in comparison with other techniques such as random forest (RF) and quadratic discriminant analysis (QDA). On the other hand, regarding the fault detection and diagnosis in TMP process, Ragab A., et al., developed a methodology based on the interaction and combination between LAD and human expertise [164]. It is aiming at maximizing the accuracy and domain of fault detection through discovering hidden reasons for abnormal events that lead to severe top events. Indeed, the novel methodology proved its superiority by enriching the main fault trees of two industrial case studies inside the TMP plant with new abnormal events with their reasons (basic events). This result was obtained due to the excellent ability of LAD to extract hidden interpretable patterns through digging deep into the historical data. However, as illustrated by the authors, this method has a second essential pillar which is the human expert who checks and validates the extracted knowledge and define its meaningfulness.

Tennessee Eastman process (TEp) is a famous simulation-based process consisting of reactor, stripper, condenser, compressor and a vapor/liquid separator, i.e. flash drum [165]. It is widely used to evaluate the proposed monitoring, fault detection and control schemes [166]. AI-based schemes/algorithms were proposed and evaluated by this process in several studies to show their superiority compared to the traditional ones [167]–[169]. For instance, *k*-NN reconstruction on maximize reduce index (MRI) was employed to derive a new and novel fault detection scheme [169]. It overcomes the problem of traditional contribution plot method that only considers/determines the most severe fault in a system of multiple faults [169]. Hence, this new method will identify all the out-of-control variables. After method derivation, TEp was used as a standard evaluation process, where the proposed scheme proved its ability to identify all the

variables responsible for multiple sensors fault. An important advantage of this method was that it can deal with any data distribution.

In another recent study, different types of deep learning neural networks including LSTM, RNN and CNN were used to design a novel scheme suitable for automated early fault detection in industrial process [167]. Upon the evaluation by TEp model, temporal CNN1D2D was the best among the developed schemes with the highest prediction/detection accuracy. Whereas all the developed ones were more accurate than the traditional sensor readings multivariate statistical analysis. Besides, networks training was enhanced through the implementation of GAN to extend the available data extracted from TEp. Moreover, auto-associative ANN deep learning technique was used to fix the problem of traditional statistical monitoring of processes [168]. It can be trained through both supervised and unsupervised methods. The advantage of this technique is that it deals with a lot of data sets that guarantees good training and more accurate fault detection. This advantage was proved by the application of the developed method on TEp model. However, the authors recommended the development of new training algorithm for this neural network in a future study for the extraction of key components relevant to the tasks of process monitoring.

2.5 Remarks on microscale and molecular scale AI/ML-assisted modeling

The application was not limited only for the predictive analysis for product/design/process optimization on macro-scale. However, recently, some researchers went beyond and used AI-related methods as data-driven micro-scale modeling tools [170]–[186]. The computations were faster in comparison with conventional CFD and multi-physics tools where the AI/ML-assisted methods showed stability due to its meshless behaviour. Besides, this approach was utilized as a less expensive predictive modeling tool. The modeling of momentum transfer/transport has the biggest share among these previously mentioned micro-scale data-driven modeling trials. The studied systems included multiphase flows and those involving turbulence where the prediction of velocity profile is the main target [187]. In addition, there are several other studies with the same concept and approach for fluid, i.e. liquid and gas, flow modeling including aerodynamic modeling [188]–[193]. According to [194], the common techniques used in the case aerodynamic data-driven modeling are LSTM, ANN and SVM. Moreover, as it can be observed, the majority of the attempts done for the utilization of combined modeling approach were performed on relatively simple systems which do not include multi-physics or multiscale problems like those in the real-world

systems. Hence, shifting towards the investigation of the capability of the combined AI-CFD approach on multi-physics/multiscale systems should gain more attention. This shift will make the modeling and behaviour prediction easier in such complex systems. This is owing to the fast-learning abilities and high accuracy of existing rigorous AI techniques.

On the molecular scale, data-driven modeling with AI/ML techniques proved their ability to predict the properties of compounds and the result of various reaction as well as designing new compounds/materials [195]–[203]. Besides, catalysis, catalyst design and prediction of materials synthesis routes or reaction pathways/mechanism are famous fields where AI/ML-assisted data-driven optimization models generally have obvious fingerprints as well [204]–[210]. AI/ML also used to enhance the analysis of new structures not only designing them [211]. On a side note, AI/ML is accelerating in the molecular design field especially in applications related to drug design [195], [197], [212]–[217]. The predicted atomic, molecular and ionic properties included solvation free energies [218], [219], dipole moments [220], global charge distribution [221] and chemical reactivity [222]. Reaction properties such as activation energies, rate constants and reaction yield were also among the modeled properties via AI/ML-based models [198]. Most of the studies follow the combined approach of modeling where AI/ML models are trained on experimental and/or simulated data obtained from physics-based computational tools. At the microscale and molecular scale, hybrid AI-physics modeling typically follows a structured surrogate modeling workflow. First, high-fidelity datasets are generated using physics-based simulations (e.g., CFD, DFT, or molecular dynamics) or experimental measurements. These data are then used to train machine learning models, such as neural networks, Gaussian processes, or kernel methods, that approximate input-output relationships of the underlying physical system. In practice, two main integration strategies are employed: (i) black-box surrogates, where the ML model directly maps inputs to outputs, and (ii) physics-informed or constrained models, where physical laws (e.g., conservation equations, thermodynamic constraints) are embedded into the learning process through custom loss functions, feature engineering, or hybrid architectures. Once trained, these surrogate models enable rapid predictions, sensitivity analysis, and optimization, reducing reliance on computationally expensive simulations while preserving acceptable physical consistency within the domain covered by the training data. The advantage of this approach is that it is a time saver besides having high predictive accuracy. Hybrid AI-physics models primarily aim to improve computational efficiency while preserving an acceptable level of physical fidelity within a defined domain of applicability.

Rather than universally surpassing the accuracy of first-principles methods such as ab initio calculations, these models act as surrogate approximations trained on high-fidelity data. Their apparent accuracy advantage in some studies typically arises in specific contexts, such as noisy experimental settings, reduced-order representations, or narrow operating regimes, where data-driven models can capture effective system behavior, including unmodeled phenomena or measurement trends. However, this does not imply that surrogate models exceed the fundamental physical resolution of first-principles methods; instead, they provide a practical trade-off between accuracy, generalization, and computational cost, enabling scalable modeling and optimization. Despite the high accuracy of prediction offered by these studies, the black-box calculations problem and chemically unaware models' issues were inevitable in most of them. Moreover, limited generalization is still an issue where the quality of training data sets and their size determine the degree of generalization of the developed models [223]. It is worth noting that additional details on the coupling of AI technologies with physics-based simulations are presented in the introductory parts in the upcoming sections, as well as the supplementary materials available online for each article. The reference aggregation was performed here to capture recent developments and to summarize the key limitations shared across these studies.

2.6 Path forward towards our AI-assisted framework

This chapter presented a quick overview of the mechanistic or physics-based computational methods used for optimizing process/material design which are powerful and versatile, varying in complexity based on the modeling scale. These methods are typically integrated with traditional optimization algorithms to minimize energy consumption and cost while maximizing performance metrics. Such approaches are robust but require significant human expertise to define objectives and constraints and are often limited by long computation times, high resource demands, and complexity of large-scale optimization problems. It also gave a brief overview on the recent applications of AI/ML for process/material optimization in different fields including the emerging industrial decarbonization pathways. It covered the different modeling scales/levels, i.e. macro, micro and molecular scales. It acts as a critical review where the limitations of these studies are clearly mentioned. These limitations are summarized in the following points:

- Black-box modeling and limited consideration of descriptive abilities of AI/ML
- Limited prediction accuracy owing to the dependence on simple/single models

- Generation or prediction of physically unreliable solutions, e.g. new materials or suggested process conditions.
- High dependence on non-learnable optimizers.
- Lack of causality analysis hindering true understanding of systems studied leading to deficiency of transfer learning
- Time consumption in some cases and lack of investigating representative search spaces of design variables

More examples and related works are presented in the upcoming chapters where specific literature review was done and customized for each paper. In addition, a detailed literature review was done separately to cover the application of AI/ML in the field of CCUS as a promising industrial decarbonization pathway. In this review article, published in *Science of the Total Environment* (Elsevier), we also presented some of the AI/ML opportunities where the majority of them were translated to the methodologies that will be presented in the upcoming chapters. The full citation of the paper is in Appendix A. It highlights the growing need for hybrid, learnable optimization frameworks that integrate AI with physics-based principles to enhance efficiency, accuracy, and interpretability. The general approach of hybrid modeling is illustrated in Figure 2.3.

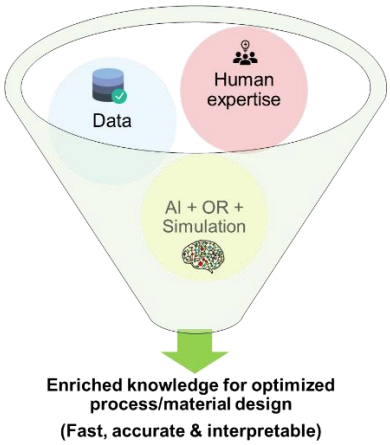


Figure 2.3 Hybrid modelling general approach

This approach combines the merits of all computational pillars, i.e. historical data, physics-based simulations and AI as a data-driven approach, with human expertise for guidance and validation. This combination ensures overcoming the limitations of each approach or technique to achieve fast, accurate and interpretable knowledge leading to optimal designs. The opportunities of AI

include the predictive, descriptive, prescriptive and generative abilities of old and more advanced techniques including interpretable ML (IML), explainable AI (XAI), RL, generative adversarial networks (GANs) and large language models (LLMs). Inclusion of these techniques in an ensemble-based modeling approach will overcome the black-box modeling, deal with uncertainties/non-linearities and consider the true distributions of historical data. In addition to descriptive, predictive, and prescriptive capabilities, recent advances in AI have introduced generative modeling as a complementary functionality. Generative AI refers to models that learn the underlying data distribution and can produce new, plausible samples (e.g., molecules, materials, or process configurations). In this context, generative models extend predictive capabilities by enabling exploration of the design space beyond observed data. They can support descriptive analysis through latent space representations and contribute to prescriptive tasks when coupled with optimization or constraint-based filtering. Therefore, generative AI is not a separate level in the modeling hierarchy, but rather a functional extension that interacts with and enhances descriptive, predictive, and prescriptive modeling.

Accordingly, under the umbrella of the proposed AI-assisted framework, which follows the philosophy of hybrid modeling, we introduce the combination of XAI and IML in Chapter 3 to overcome the black-box modeling. Although black-box machine learning models have been widely applied in prior studies, their limited interpretability and lack of transparency remain critical barriers to adoption in industrial settings. Chapter 3 addresses this gap by developing an integrated interpretable and explainable modeling framework that combines intrinsic interpretability with post-hoc explanation techniques to extract actionable knowledge from data. This contribution is not limited to a specific application domain but rather establishes a generalizable methodological foundation that supports subsequent developments in later chapters, including generative modeling (Chapter 5) and reinforcement learning-based optimization (Chapter 6). In this sense, Chapter 3 provides the interpretability backbone upon which the broader AI-assisted framework is built. In addition, ensemble-based learning combined with explainability is introduced in Chapter 4 to tackle the problem of low prediction accuracies while considering modeling transparency. In Chapter 5 we introduce a generic methodology for physics-guided/aware generative deep learning as a solution to the problem of physically unreliable generations including discovered materials and new processes. Chapter 6 introduces a methodology for simulation-based optimization where causal incremental RL is at its core to mainly tackle both problems of non-learnable optimization

and lack of causality analysis. The combination of these methodologies in this doctoral research represents a physics-guided hybrid learning framework ensuring reliable predictions, data generation, and optimization, while addressing the limitations of black-box ML. Emphasizing multi-level causality analysis, the study progresses from predictive insights (XAI/IML) to deeper reasoning (LLMs) and comprehensive causal mapping (Causal BNs). The resulting interpretable patterns and causal graphs enhance transfer learning and enrich expert knowledge. The industrial applications considered throughout this doctoral research are mainly in the design phase to further make the industrial decarbonization value chains more economic and eco-friendlier to achieve sustainability and *Net-Zero* goals.

CHAPTER 3 ARTICLE 1: MACHINE LEARNING-ASSISTED SELECTION OF ADSORPTION-BASED CARBON DIOXIDE CAPTURE MATERIALS

Eslam G. Al-Sakkari (Eslam Ibrahim), Ahmed Ragab, Terry M.Y. So, Marzieh Shokrollahi, Hanane Dagdougui, Philippe Navarri, Ali Elkamel and Mouloud Amazouz

Article Submitted on April 15, 2023 and Published on August 10, 2023 in Journal of Environmental Chemical Engineering, Elsevier (IF 7.2), Volume 11 , Issue 5, 2023

Online Publication date: August 2023

DOI: https://doi.org/10.1016/j.jece.2023.110732

3.1 Abstract

Recently, carbon capture has gained increased attention as a sustainable way for mitigating global warming. One of the promising technologies for carbon capture is the adsorption-based process. Accordingly, many researchers are focusing on the development of novel adsorbents with desired characteristics, such as high capacities and selectivities, to accelerate the deployment/design of adsorption-based CO_2 capture processes. This paper aims at identifying a group of optimum adsorbents with high capture capacity and selectivity based on the previously published data on different adsorbents. It provides different stakeholders (including experts in process design) with comprehensible and clear human interpretable patterns/rules. This facilitates the selection of optimum adsorbents for carbon capture at the simulation, lab-scale, and pilot-scale levels. Moreover, this work benefits other stakeholders such as researchers in the field of material design by providing the adsorbent characteristics needed to have high capacities and selectivities. Hence, data was collected from different sources including research papers, literature reviews and publicly available datasets. Additionally, an ensemble of diversified (explainable and interpretable) machine learning techniques was utilized to classify these collected data with a classification accuracy above 98%. The extracted rules for achieving the desired characteristics were revised by experts and were found to be appropriate. For instance, it is recommended to design a chemically modified adsorbent with high surface area, large pore volume and small pore diameter to be more selective. The preferred operating conditions are low temperature and low pressure, i.e., near atmospheric. Regarding capacity, high pressures are preferred. Explainable and interpretable machine learning were selected to overcome the black-box prediction problems of commonly used techniques.

3.2 Introduction

There is a growing concern about climate change problems and its resultant crisis, i.e., global warming. In particular, the temperature change over the past 50 years exceeded 3.5 ºC [224][225] which leads to the increase in sea levels [226] and desertification [227][228]. Thus, researchers and other stakeholders are trying to find feasible solutions to mitigate this problem. The solutions include renewable energy production [229] and industrial systems decarbonization [230]. It is expected that these actions will contribute to decreasing global warming to the limit of 1.5 ºC as recommended by the *Paris Climate Agreement* [231] and the *COP-26 Conference* [232].

In this regard, carbon capture emerges as a reasonable and promising solution for systems decarbonization [233]. It involves capturing carbon dioxide from its different sources, particularly heavy industries which are considered as the largest final emitters in current times. These emitters include iron & steel plants, petroleum refineries, petrochemical factories, cement production plants, pulp & paper mills as well as power generation facilities. These facilities are responsible for over 46% of global greenhouse gases (GHG) emissions, which is almost 36.4 billion metric tons of CO_2/year [234]. Where, this captured carbon can be then fixed through different pathways such as the enzymatic [235] and chemical ones [236]. The most common approach is called post-combustion carbon capture where carbon dioxide is separated from flue/waste gas streams from power plants and other facilities after combustion of fossil fuels and other carbon-based materials [237]. There are other approaches including pre-combustion carbon capture where carbon dioxide is removed from the fuels before utilization [238] and the oxy-fuels approach which involves the use of oxygen instead of air for the combustion [239]. In this case, the concentration of carbon dioxide in the flue gas stream is very high which facilitates its capture. There is another approach for carbon capture which is the direct capture from ambient air, commonly known as direct air capture (DAC), that can lead in the future to carbon negativity not only neutrality [240].

All these approaches rely on various capture technologies that involve chemical and physical capture. The most abundant technology is the absorption or solvent-based carbon capture where a liquid solvent is used to absorb carbon dioxide from the flue/waste gas stream [241]. Most of the currently used absorption materials feature chemical absorption where this chemical bond is then broken via heating in a desorption step aiming at solvent regeneration. This step is essential to decrease the operating cost of the process by recycling the regenerated solvent to be reused several times. Only a small amount of fresh solvent is added as make-up to compensate for any losses. However, this desorption step is very energy intensive; thus, researchers are conducting new studies to develop novel solvents with lower energy requirements. These novel solvents include potassium glycinate-based solvents [242] and a novel blend of amines called HNC-5 [243].

Cryogenic capture is another technology that involves very low temperature fractionation to separate carbon dioxide from the flue gases [244], whereas chemical looping-based carbon capture is a one of three main technologies of chemical looping where an earth metal oxide is used for the exothermic chemical capture of carbon dioxide [245]. Typical metal oxides used are calcium and magnesium oxides. The regeneration step is endothermic where heat is supplied to perform the

thermal decomposition to, ideally, have pure carbon dioxide. This purity of carbon dioxide in the outlet gas stream should be considered if air or nitrogen are used for continuous gas purge. On the other hand, there is a growing attention about the utilization of membrane filtration as an alternative way for carbon capture [246]. In this regard, different membrane materials and configurations are developed and examined to have an economic capture process. However, this process is still developing, and more research is ongoing to achieve this challenging goal.

Lastly, the fifth main capture process is the adsorption-based capture that involves the utilization of solid sorbents to adsorb carbon dioxide on its surface and inside its pores [247]. After being saturated, these adsorbents are then regenerated using different techniques. The regeneration process is an essential step in the capture process and determines the process efficiency and cost. Hence, an efficient regeneration process that guarantees efficient adsorbent reutilization will increase capture viability. The main four regeneration/desorption processes are pressure swing [248], thermal swing [249], vacuum swing [250] and electrical swing processes [251]. It is worth noting that the DAC approach rely mainly in the current time on solvent and adsorption-based carbon capture technologies [252][253]. In this regard, various adsorbents were developed to withstand the different conditions of moisture presence and low carbon dioxide concentration in the ambient air [254][255].

Recently, adsorption is getting more attention as a promising technology that can overcome the drawbacks of other technologies, especially the high-energy consumption from solvent-based capture. In this regard, a lot of materials were and still are being developed with the aim of producing sorbents with high capacity, high selectivity, high stability, and low cost. For instance, several zeolites [256], metal organic frameworks (MOFs) [257][258], carbon-based materials [259][260] and porous organic polymers (POPs) [261][262] were synthesized and their application for carbon capture were investigated experimentally. Where, there is a new trend for the adsorbents design/production from waste materials as a way of achieving cost effectiveness [263][264]. During the experimental work, capture capacity and carbon dioxide selectivity over other gases, mainly nitrogen and methane, are the main variables investigated to differentiate between various adsorbents. Regenerability and recyclability of these adsorbents were also explored as important characteristics of solid sorbents as aforementioned. On the other hand, a few studies also considered the cost of adsorption materials as a vital aspect which influences the selection of the adsorbent and the process at a large scale. Adsorbent performance can be measured with several variables,

each with their own target values. For instance, a study reported a capture capacity of over 2 mmol/g and a CO_2/other gas selectivity of 100 as satisfactory values for an adsorbent with an acceptable performance [265]. Other targeted performances are being usable for over 1000 successive cycles and an adsorbent cost of less than USD 10 /kg.

The CO_2 adsorption performance of different materials depends on their textural properties, including the *Brunauer-Emmett-Teller* (BET) specific surface area, pore volume, micro-pore volume and pore diameter. Preparation procedures and conditions greatly influence these textural properties. These procedures include co-precipitation [266], hydrothermal treatments [267][268] and surface functionalization through chemical activations [269][270]. Additionally, operating conditions including temperature, pressure, concentration (partial pressure) of carbon dioxide in the flue gases, and the presence of impurities in the flue gas are important variables that influence the performance of CO_2 adsorbents.

The selection of optimum adsorbents is a challenging task. This is due to the huge number of trials for diverse adsorbents synthesis without a generalized procedure based on high level rules where the data about the previous trials are scattered in the literature. There is a great need to develop such rules for the preparation of an efficient adsorbent, but the search space is too large. Besides, the experimental work as well as the first principles-based computational work for designing and simulating new adsorption materials are expensive and time consuming [271].

Given these limitations, challenges and needs, it is necessary to create a new efficient methodology based on data-driven modeling techniques employing the scattered data in literature to accelerate the design of new adsorbents for carbon capture. Data-driven modeling techniques, i.e. artificial intelligence (AI) and machine learning (ML) related technologies, arise as promising techniques to overcome the abovementioned challenges and satisfy the need for accelerated material design due to their outstanding abilities and high accuracies [204], [272], [273]. The success of these techniques in other applications, including industrial operations control [162], thermodynamic and equilibrium properties prediction [274][155] and optimizing processes performances [275][276], support the new direction towards their application in material design and discovery [277]–[279].

Recently, AI and ML technologies are being used to help optimizing the adsorption-based carbon capture on the process and material levels [115], [280]–[283]. For instance, the Random Forest (RF) algorithm was used to explore the critical factors affecting carbon dioxide adsorption capacity

of porous carbon materials [284]. Several carbon-based adsorbents were investigated considering their textural properties and chemical compositions at different pressures from 0 to 1 bar. As a result, the textural properties were found to have higher effect on the capacity than the chemical compositions. In addition, another study focused on the prediction of carbon dioxide adsorption capacity and rate on carbon-based materials but in this case a deep convolutional neural network (D-CNN) was used [285]. The study concluded that the most important variable affecting adsorption rate and capacity is adsorption pressure followed by adsorption temperature, adsorbent surface area and adsorbent pore volume. Deep neural networks (DNNs) were also employed successfully to predict carbon dioxide adsorption capacity of carbon-based materials based on their textural properties (i.e. BET surface area, micropore and mesopore volumes) [154][286]. Besides, other tree-based models such as gradient boosting decision tree (GBDT) were used to model the adsorption performance of biomass waste-derived porous carbons towards carbon dioxide [287]. The study revealed that the effect of adsorption conditions is more significant than that of textural properties on adsorption capacity. In a more recent study, DNN proved again its ability for predicting the carbon dioxide adsorption capacity of graphene oxides (GOs) [288]. The chosen inputs were adsorption conditions, such as temperature and pressure, in addition to GOs' textural properties, i.e., BET surface area and pore volume. In an earlier study [289], a data driven quantitative structure-property relationship (QSPR) was developed for MOFs based on gradient boosted tree regression. The datasets used for training, testing and validation consisted of several hypothetical MOFs. The predicted variables in that study were working adsorption capacity and CO_2/H_2 selectivity. These models were developed based on six geometric descriptors in combination with three atomic property-weighted descriptors. Furthermore, six machine learning-based regression techniques including the RF, conditional inference decision tree (DT) and gradient boosting machine (GBM) were used to explore the effect of different descriptors on carbon capture working capacity and CO_2/N_2 & CO_2/H_2 selectivity [290]. The DT was responsible for predicting the change in parent MOFs capture metrics because of surface functionalization, whereas the other five models mainly focused on the prediction of working capacity and selectivity based on the chemical-structural descriptors. In another study, a support vector machine (SVM) was employed to classify a set of simulated MOFs based on their capture performance as a preliminary step for accelerating adsorbents design [291]. The acceleration was achieved through fast pre-screening and reduction of search space from 1.5 million adsorbents to only about 150,000 adsorbents.

Although these previously mentioned techniques are highly accurate, they act as black-boxes that lack interpretability and explainability [292]. Hence, there is an ongoing interest to add explainability to these accurate models in order of having better understanding about the effect of process and material properties on the variables of interest [293][294]. In this regard, model agnostic explainable AI technique such as LIME [295] or SHAP [296] is added after performing the black-box modeling task to reveal the effect of each variable on the measured variable of interest [297]. In a recent study, SHAP coupled with Random Forest was used as a figure of explainable machine learning for the prediction of carbon dioxide adsorption on porous carbons [298]. It was concluded that the textural properties (BET surface area, micropore and mesopore volumes) have more significant influence on the adsorption capacity than the chemical composition of the studied carbon-based adsorbents.

Given the abovementioned attempts and limitations, there is still a need for a more generalized methodology considering the wide range of existing adsorbents. This general methodology should give clear patterns for the adsorbent design properties and adsorption conditions to have higher selectivity and capacity. These patterns also should be combined with explanation of the effect of each variable and the relative importance of these variables/effects on selectivity and capacity. This general methodology, when successfully developed and applied, results in reducing the search space and accelerates the adsorbents design in the longer term.

This paper aims at collecting the data in the literature on different adsorption materials developed over time and classifying them using diversified AI/ML methods including interpretable and explainable algorithms. During this study, the existing adsorbents are classified to high/low capacity and high/low selectivity adsorbents. The final goal is to find a group of optimum cost-effective adsorbents having desired specifications (e.g., high adsorption capacity and selectivity). Finally, a database of adsorbents and knowledge base of how to obtain a good adsorbent are created. These databases and extracted knowledge can be used by different stakeholders including experts in the field of adsorption materials design/discovery and carbon capture process designers. Researchers will gain new insights to help design new adsorbents having the desired characteristics with the minimal efforts in an accelerated manner, whereas process designers will depend on the results to select the optimum adsorbents among the existing ones to design the capture process based on them. This will also accelerate the process design step by minimizing the search space.

Following are the main scientific contributions of the current study that define its novelty:

1- Constructing a consolidated database based on data scattered in the literature about different adsorbents developed for carbon capture. This database was constructed and made ready for analysis after several preprocessing steps including, cleaning, outlier removal and imputation. The database summarizes the adsorbents' morphological/textural properties, i.e., surface area, pore volume and pore diameter, adsorbents' structure, and operating conditions, i.e., pressure and temperature, as input (controlled) variables. In addition, it contains the adsorption capacity and selectivity of these different adsorbents as the outputs. This database will be made available publicly to help save time for different stakeholders. In addition, this database will be updated continuously by adding new variables, new points and revising the available data for further knowledge enrichment and decision enhancements.

2- Combining interpretable ML (IML) with explainable AI (XAI) having high prediction accuracy. This IML/XAI combination gives a complete picture of how data variables are impacting classification accuracy both on a global and local scales. The diversified IML/XAI techniques capture different types of patterns within the data thus getting a more comprehensive understanding of the underlying relationships in the data. When used together, IML and XAI potentially lead to better prediction accuracy and guard against overfitting. This approach has the advantages of giving clear design rules, fast classification with high accuracy and explanation of black-box modeling to provide weights of the importance/effect of each predictor, i.e., controlled variable, on the selected responses, i.e., measured variables of interest.

3- Extracting design rules to build a knowledge base that can be used afterwards to design new efficient capture materials. This will lead to accelerated material design based on the interaction between expertise, machine learning and simulation. These rules are summarized in tables for both low and high temperature adsorbents. In addition, illustrations on the different relations between each manipulated variable and the target ones (capacities and selectivities), are provided depicting their relative significance as well. The combination of these tables along with these illustrations represent the concrete knowledgebase that will be provided to the designers and operators afterwards to facilitate/accelerate the adoption of feasible adsorption-based carbon capture processes.

To the best of the authors' knowledge, the combination between these rules, relations and significance analysis based on the available different adsorption materials has not been provided in the literature so far. It is worth mentioning that this knowledgebase is updatable as to be mentioned and discussed in detail in the future work part/section.

It is worth mentioning that the methodology developed in the current study is transferable and can be generalized to other applications of material discovery such as solvents, adsorbents for direct carbon capture, membranes, and catalyst screening. It can also be applied to optimize both process operation and design. Thus, the codes are made available publicly as the constructed database.

This paper is organized to have five sections. **Section 3.3** introduces the methodology developed to screen different adsorbents and the ML and AI techniques employed. In addition, the prepared database and the data collection, organization, and assembly efforts are also discussed here. **Section 3.4** presents the results of the study including the extracted rules for adsorbents classification and selection. It also discusses these results from the point of view of both data analytics and carbon capture experts. **Section 3.5** analyzes the findings of the current study and introduces opportunities for future work. Lastly, **Section 3.6** concludes the work done throughout the paper.

3.3 Datasets and Methods

This section introduces the general research methodology, data collection/preprocessing methodology and the developed combined interpretable/explainable ML methodology for enhanced classification. In addition, it presents the adsorption materials datasets collected and constructed in the current study for validating the proposed methods and the selection of the group of optimum adsorbents.

3.3.1 General methodology

The general methodology of the current research is summarized in Figure 3.1. The experimental and simulated data were collected from different sources including original research papers, literature reviews and official databases. Two datasets were considered in the current study. The first one gathered various types of adsorbents including metal organic frameworks (MOFs), zeolites and carbon-based materials, while the second dataset was compiled from an open-source dataset for MOFs [299][300]. The variables of interest are the adsorption capacity (mmol/g) and IAST CO_2/N_2 selectivity. To screen the adsorbents based on these variables, the input variables

chosen are BET specific surface area (BET), pore volume (VP), micro-pore volume (VMP) and pore diameter (DP). As well, the reported capture temperatures (T), pressures (P) and carbon dioxide partial pressure (CPP) in the feed are also considered to account for the effects of capture process conditions. As previously mentioned, these textural properties represent physical factors affecting adsorption performance.

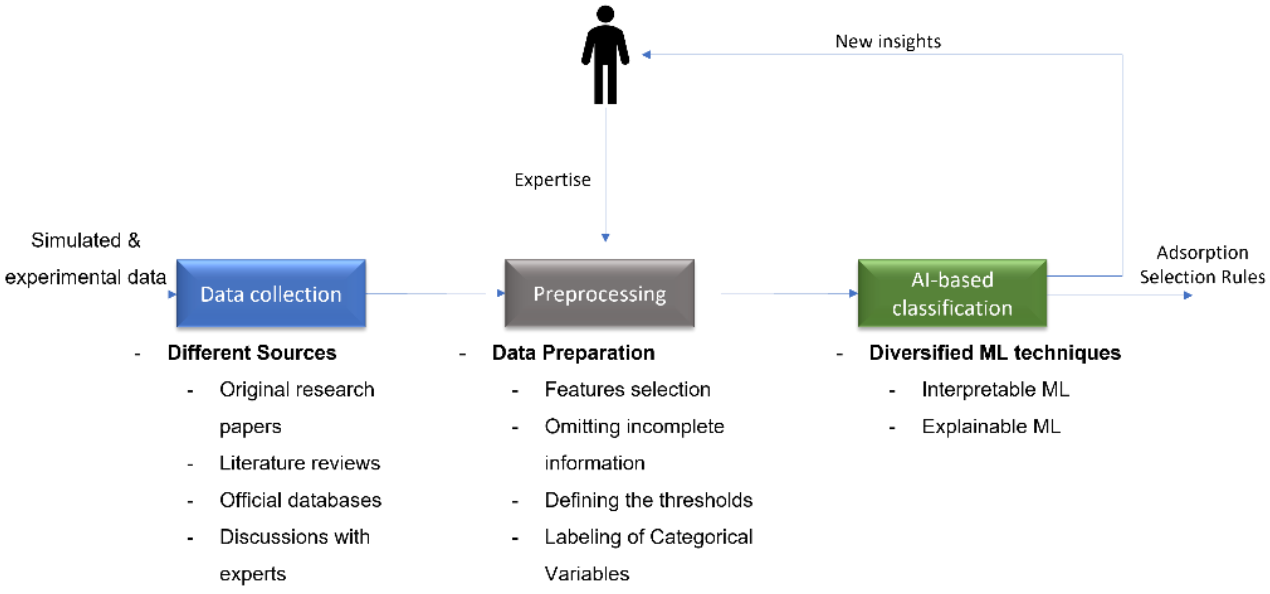


Figure 3.1 General research methodology

The adsorbents are also categorized based on broader characteristics considering the chemical composition. In particular, the adsorbent categories according to the main base material are: carbon-based materials (CAR) including carbon nanotubes [301], carbon nano-fibers [302][303] and activated carbons [304]; ceramics (CER) which include lithium containing materials [305][306], geopolymers [307] and metasilicates [308]; metal oxide (MO) which include commonly-used oxides such as magnesium and calcium oxide [309]–[312]; metal organic frameworks (MOF); porous organic polymers (POP) including hyper-crosslinked polymers and conjugated microporous polymers which have high surface area and are easily functionalized [313]; silica-based materials (SIL); and zeolites (ZEO).

3.3.2 Datasets

As mentioned above, two sets of data are used for the training and testing steps required for the development of screening models. Dataset I is manually collected from diverse review articles and

experimental original research papers measuring the physical/textural characteristics of several adsorbents and mapping them to the carbon capture capacity and selectivity. All the articles are peer-reviewed and cited in the bibliography section (sheet) in the supplementary materials. The analytical and experimental methods used for the calculation of different textural properties considered in this current study are summarized and clarified in [314], [315] and these reviewed publications [316]. While variables other than those abovementioned are recorded; a review of the generated dataset demonstrates that the abovementioned variables are most reported between the different papers in literature. The total data collected is provided in supplementary materials. In addition, the data about adsorption selectivity and capacity was acquired from the literature where each research and review article mentioned their calculation methods. According to these articles, the common method used is the *Ideal Adsorbed Solution Theory* (IAST) [317]. As mentioned previously, these articles were cited in the supplementary material file named "Dataset I (raw data with labeling)" in the bibliography section/sheet. There are additional studies that mentioned the use of IAST for selectivity calculation based on other adsorbents and systems [318]–[320].

To better capture the temperature and pressure effects, i.e., process conditions, different data points for each adsorbent are taken for deep evaluation of the adsorption capacity/selectivity. Ultimately, Dataset I present 261 unique adsorbents with 499 data points after data cleaning and preprocessing (580 raw datapoints). The summary of Dataset I is presented in Table 3.1 including the number of data points for each variable related to each main category as well as the total number of data points in this dataset.

Table 3.1 Dataset I summary (in relation to the available main categories)

Variable	Category*							Total number of points
	MOF	POP	SIL	ZEO	MO	CAR	CER	
Category	145	174	20	60	83	92	6	580
Adsorbent name	145	174	20	60	83	92	6	580
P (bar)	139	174	20	60	81	92	1	567
T (°C)	139	174	20	60	83	92	6	574
CPP (bar)	139	174	20	60	83	92	6	574
BET (m^2/g)	131	172	20	45	68	91	1	528
VMP (cc/g)	23	40	12	16	27	70	-----	188
VP (cc/g)	67	108	12	27	37	67	-----	318
DP (nm)	35	62	12	47	36	23	-----	215
Capacity (mmol/g)	129	145	20	54	45	91	-----	493
Cyclability	1	-----	-----	-----	-----	-----	-----	1
Selectivity (CO_2/N_2)	37	77	4	6	-----	36	-----	160

Where, P is adsorption pressure, T is adsorption temperature, (BET) is BET specific surface area, (VP) is pore volume, (VMP) is micro-pore volume and (DP) is pore diameter, (T) is capture temperature, (P) is capture pressure and (CPP) is the feed carbon dioxide partial pressure.

* Categories: (MOF) metal organic frameworks; (POP) porous organic polymers; (SIL) silica-based materials; (ZEO) zeolites; (MO) metal oxide; (CAR) carbon-based materials; (CER) ceramics.

The collection of these data was challenging and the raw form of the data after collection is difficult to be used for models' development. Thus, several data preprocessing and preparation steps were performed including elimination of outliers using the *EXPLORE* software developed by CanmetENERGY-*Natural Resources Canada* [321], labeling of categorical variables, i.e. adsorbents categories and adsorbents names, and elimination of variables with more than 70 %

missing data. First, the variables are divided into categorical and numerical ones before labeling. Figure 3.2 summarizes the steps of preparing the data and converting raw data to analysis-ready data (ARD). The process of data collection and pretreatment is a dynamic one. In addition, the *EXPLORE* software was also used for preliminary clustering of the available adsorbents. Moreover, different methods were used and tested for the imputation including expectation maximization (EM) [322], averaging, neural network modeling [323], eXtreme Gradient Boosting (XGB), and Random Forest (RF) [324] and the use of generative adversarial networks (GANs) that capture the true data distribution and give high quality imputations [325]. They are selected in this work to handle different types of missing data (e.g., missing completely at random, missing at random, missing not at random). This helps account for uncertainty about the correct imputation model to use. More specifically, we have used the GAIN technique, a deep learning-based method that leverages GANs, to impute missing values. It uses a generator network to produce imputations and a discriminator network to distinguish between observed and imputed values (For more detail, see this [325]). The EM is used in this work as a statistical method that iteratively estimates missing values given the observed data, then updates the model parameters based on these imputations, until convergence [326]. We have used a type of ANN called an autoencoder, trained to reproduce its input, which forces it to learn compact representations and dependencies in the data [327]. The XGBoost is used here as it has a built-in capability to handle missing data [328][329]. When XGB encounters a missing value at a node, it tries both the left and right child nodes and learns the direction that leads to higher gain for the specific missing value. This allows the model to create a default direction for missing values in future encounters. The RF is a type of ensemble learning method, where the prediction is based on the average of a large number of decision trees, each built from a slightly different bootstrap sample of the data [330][331]. The RF is used as a predictive model where the variable with missing data is the target variable (i.e., capacity and selectivity). The RF model is then trained on the observed values and used to predict the missing values. The output of this imputation process was the average of the results obtained from the ANN, GAIN, EM, XGB, and RF. The results obtained from this combination were validated on a disjoint dataset, and this process was repeated multiple times to get multiple imputed datasets, which provides a measure of the uncertainty of the imputations. This assures more accurate results compared to applying each single imputation technique separately.

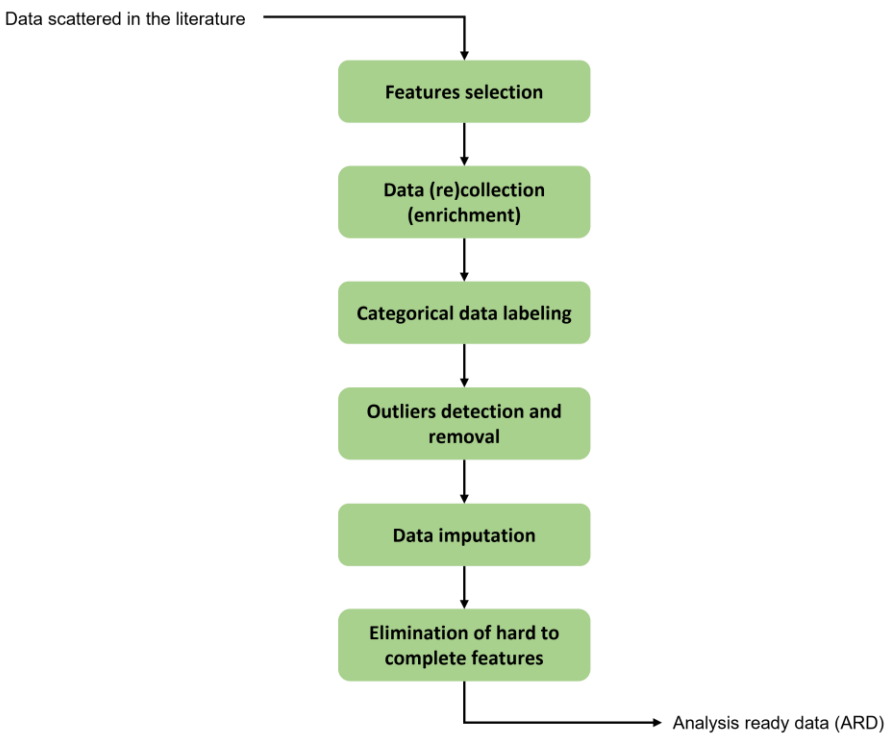


Figure 3.2 Data preprocessing/pretreatment methodology

Table 3.2 summarizes the categories and their corresponding labels while Table 3.3 represents a sample of the adsorbents' names along with their labels.

Table 3.2 High level adsorbents categories and their corresponding labels

Category	**Label**	**ID**
Carbon-Based	CAR	1
Ceramic	CER	2
Metal Oxide	MO	3
Metal Organic Framework	MOF	4
Porous Organic Polymer	POP	5
Silica	SIL	6
Zeolite	ZEO	7

Table 3.3 Adsorbents labeling sample.

Adsorbent Name	**ID (Label)**
Cu2(abtc)3_(SNU-5)	17
Cu2(bdcppi)(DMF)2_SNU-50	18
Dy(BTC)(H2O)__DMF	22
Al4(OH)2(OCH3)4_(BDC-NH2)3	5
Amine-MIL-101_(Cr)	6
Amine-MIL-101_(Cr)-50%_PEI-amine	8
Amine-MIL-101_(Cr)-75%_PEI-amine	9
Amine-MIL-101_(Cr)-100%_PEI-amine	7
CNF-2	230

In addition, to facilitate the classification, the adsorption capacity was divided into two classes, low capacity taking the label (0) and high capacity taking the label (1). Similar classification was done in the case of selectivity. The thresholds for capacity and selectivity were chosen as 2 mmol CO_2/g adsorbent and 20 CO_2/other gas, respectively, to agree with the literature and to have a balanced classification problem. Further, a combined problem to achieve adsorbents featuring high capacities and high selectivities was formulated as well. To formulate the multi-class classification problem, the adsorption capacity and selectivity are chosen and thresholds based on previous studies are established as mentioned above. The conditions are defined as the following:

Condition I: The CO_2 adsorption capacity is greater than 2 mmol CO_2/g adsorbent.

Condition II: The CO_2/N_2 adsorption selectivity is greater than 20.

The combination of these conditions creates different classifications. Adsorbents that fail both ***Conditions I*** and ***II*** are labeled as (0). Low capacity and highly selective adsorbents that fail ***Condition I*** and pass ***Condition II*** are labeled as (1). Conversely, high capacity and low selective

sorbents that pass ***Condition I*** and fail ***Condition II*** are labeled as (2). If the adsorbent passes both ***Conditions I*** and ***II***, then it should be labeled as (3).

Given the difficulty in acquiring data along with data inconsistency between different papers, a second dataset is also used to develop the machine learning-based classification model. Dataset II is based on the dataset by COSMOS, made publicly available on the COSMOS website [299][300]. The dataset uses molecular structures of numerous MOFs in the Cambridge Structural Database (CSD) to compute the pore-limiting diameter, largest cavity diameter and accessible surface area using Zeo++. The structural data from Zeo++ is originally presented by COSMOS and remains unchanged; the adsorption performance score originally reported by COSMOS has been adapted to the adsorption capacity, per definition. The adsorption capacity is determined at 25°C and 1 bar, at 15% CO_2 and balance N_2. The adsorbents' six-letter identifiers originally tied with the CSD have also been linked to the full name given in the CSD. Dataset II adds an additional 3654 adsorbents culminating with 3816 data points. The total data points used are provided in the supplementary materials. The same procedure of capacity and selectivity classifications in the case of Dataset I was followed here in Dataset II as well.

3.3.3 Statistical analysis

As an essential initial step for understanding/explaining the available data during data-driven modeling, basic statistical analysis including the determination of data distribution should be done. Accordingly, the frequency of data/values related to each numerical and categorical variable was determined. This was done to have a visual/graphical representation of the data distribution for each variable. Besides, other simple statistical analysis including the determination of means, minimums and maximums was done for better data description. It should be noted that the determination of these values is an essential step for data normalization, i.e., data range is 0 to 1, or standardization, i.e., data range is -1 to 1, if needed. In addition, a correlation matrix based on Pearson's correlation coefficient was constructed using built-in libraries in Python. This gives an initial representation of the negative/positive correlations between different variables and most importantly the correlations between the selected predictors and the capacity/selectivity classes.

3.3.4 Machine learning models employed in this study

As previously mentioned in the introduction section, interpretable and explainable machine learning techniques were employed to perform the data-driven modeling. Firstly, the interpretable techniques are employed to explore the data and then extract clear patterns (IF-THEN rules) to reach a specified class [332], or high capacity/selectivity in the current study. These rules can be then given to the experts for verification as well as to gain new insights which in consequence reduce the search space in future material discovery applications. In the current study, decision tree (DT) was used as an interpretable ML technique. DT is a supervised machine learning and flowchart-like technique as illustrated in Figure G.1 in supplementary materials (Appendix G).

Despite having the advantage of interpretability, some DTs can suffer from overfitting where the prediction accuracy for the testing is significantly lower than that of the training. Accordingly, it was proposed to couple the interpretable ML-based modeling with other deep learning-based techniques that possess higher prediction accuracies. The suggested techniques are the artificial neural network (ANN) and extreme gradient boosting (XGBoost). The results of these two techniques were compared from the perspective of prediction accuracy and training time to select the most efficient technique, where the optimum technique is the one giving higher accuracy within shorter training/prediction time. Simple graphical illustrations of ANN and XGBoost are presented in Figure G.2 in the supplementary materials (Appendix G).

ANN consists generally of several layers, i.e., input layer, output layer and hidden layer(s) between these two layers. Each layer consists of small units called neurons. The number of neurons of the input layers is the same as the number of predictors (X) and, for binary classifications, the number of neurons in the output layer is two. The number of neurons per hidden layer as well as the number of hidden layers are hyperparameters that should be optimized for higher accuracies.

The XGBoost is an ensemble supervised learning algorithm that consists of multiple simple/weak trees applying the concept of boosting for error, bias and underfitting minimization. The trees in this model are constructed in parallel unlike the sequential trees building in the case of gradient boosting decision trees (GBDT). It gained a huge attention in recent times due to scalability, high accuracy, and fast predictions. In addition, it is used on a wide range of applications including regression, classification, and ranking.

However, these techniques are still black-box ones; hence, an explainable technique is added to these high prediction accuracy techniques. The explainable technique used in the current study is the SHAP values calculation to give a numerical representation of the weights of the effect of each controlled variable on the measured ones. The interactions between different variables were also considered.

The combination of interpretability and explainability enhances the classification and combines the advantages of both techniques while avoiding their limitations. Figure 3.3 illustrates the proposed methodology for combining interpretable and explainable machine learning techniques with high prediction accuracy. It is worth noting that there is an essential step that was taken after modeling the data using interpretable machine learning which is the pattern selection as illustrated in Figure 3.3. In this step, the patterns are selected based on their readability and selectivity. A prime pattern is more readable whereas a spanned one is more selective.

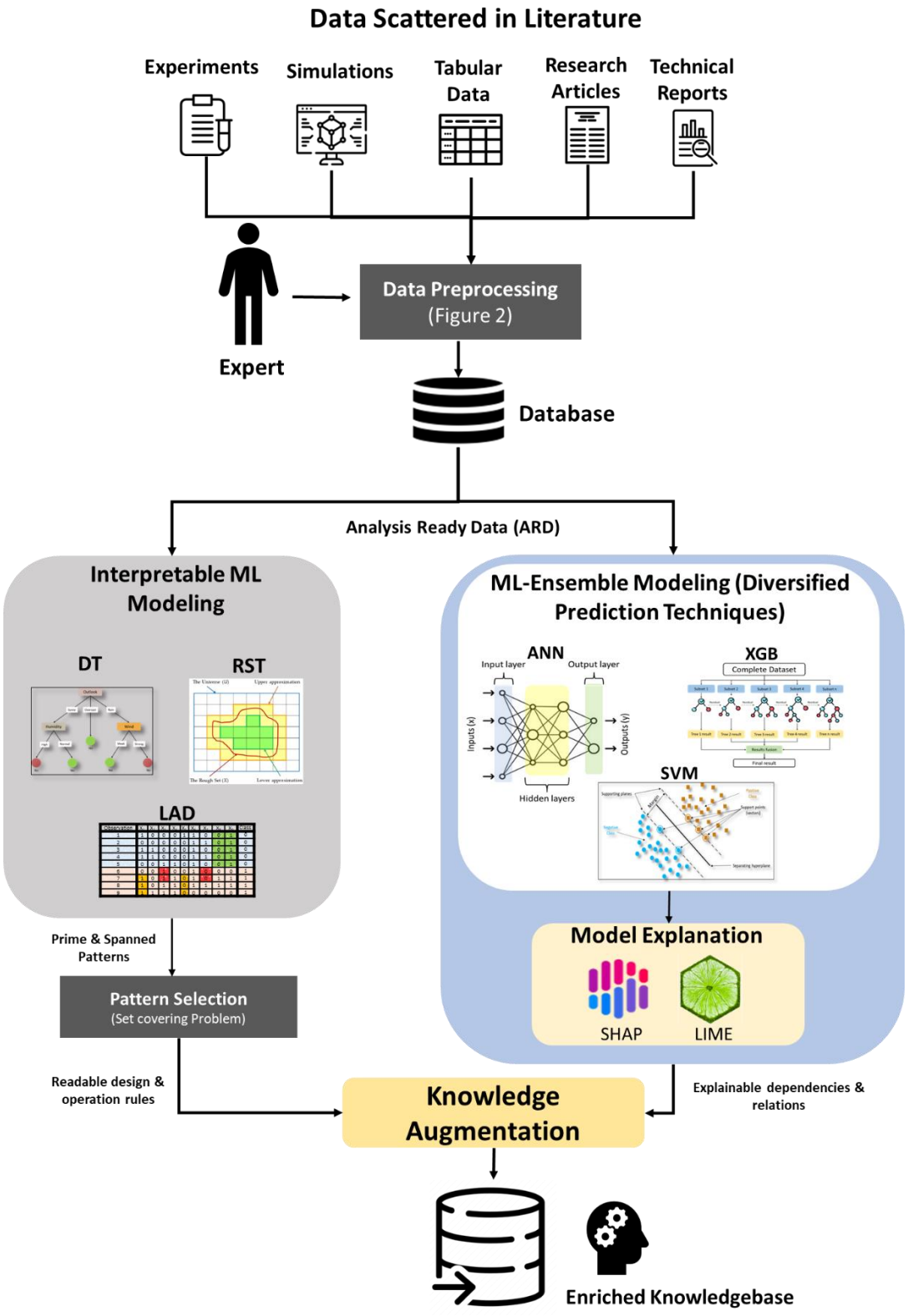


Figure 3.3 Combined interpretable/explainable ML approach

After building the models, different hyperparameters optimization approaches based on trial and error were followed to achieve the best available model ready for the classification task. In this regard, manual search as well as Grid search using the *Optuna* package (hyperparameters optimization package for *Python*) were employed to obtain these optimum hyperparameters. The hyperparameters and their ranges are maximum depth of 2-10 for decision tree; number of estimators of 50-200 for XGBoost; learning rate of 0.0001-0.01, number of hidden layers of 2-5, number of neurons of 10-200 and batch size of 1-32 for ANN.

The evaluation methods and metrics utilized during the current study are confusion matrix, accuracy, and *F1*-score. The graphical representation of confusion matrix is depicted by Figure S3 in supplementary materials. This matrix combines the different possibilities of the classification problem. For instance, in the case of negative/positive binary classification, the correctly classified negative instances take the label true negative (*TN*). The negative observations misclassified as positive ones are false positive (*FP*). The positive observations misclassified as negative ones are false negative (*FN*). Finally, the correctly classified positive instances take the label true positive (*TP*).

The mathematical formula of *F1*-score is represented by Equation 3.1.

$$F1 - score = \frac{2\,(Precision \times Recall)}{Precision + Recall} \tag{3.1}$$

Where the precision and recall are determined by the following equations.

$$Precision = \frac{TP}{TP + FN} \tag{3.2}$$

$$Recall = \frac{TP}{TP + FP} \tag{3.3}$$

F1-score as a function of TP, FP and FN is presented by equation 4 [163].

$$F1 - score = \frac{2\,TP}{2TP + FP + FN} \tag{3.4}$$

3.4 Results and discussions

3.4.1 Analysis of Collected Data

The frequency/distribution of variables collected in Dataset I can be observed in Figure 3.4 and the distributions of selected variables related to Dataset II are illustrated in Figure 3.5.

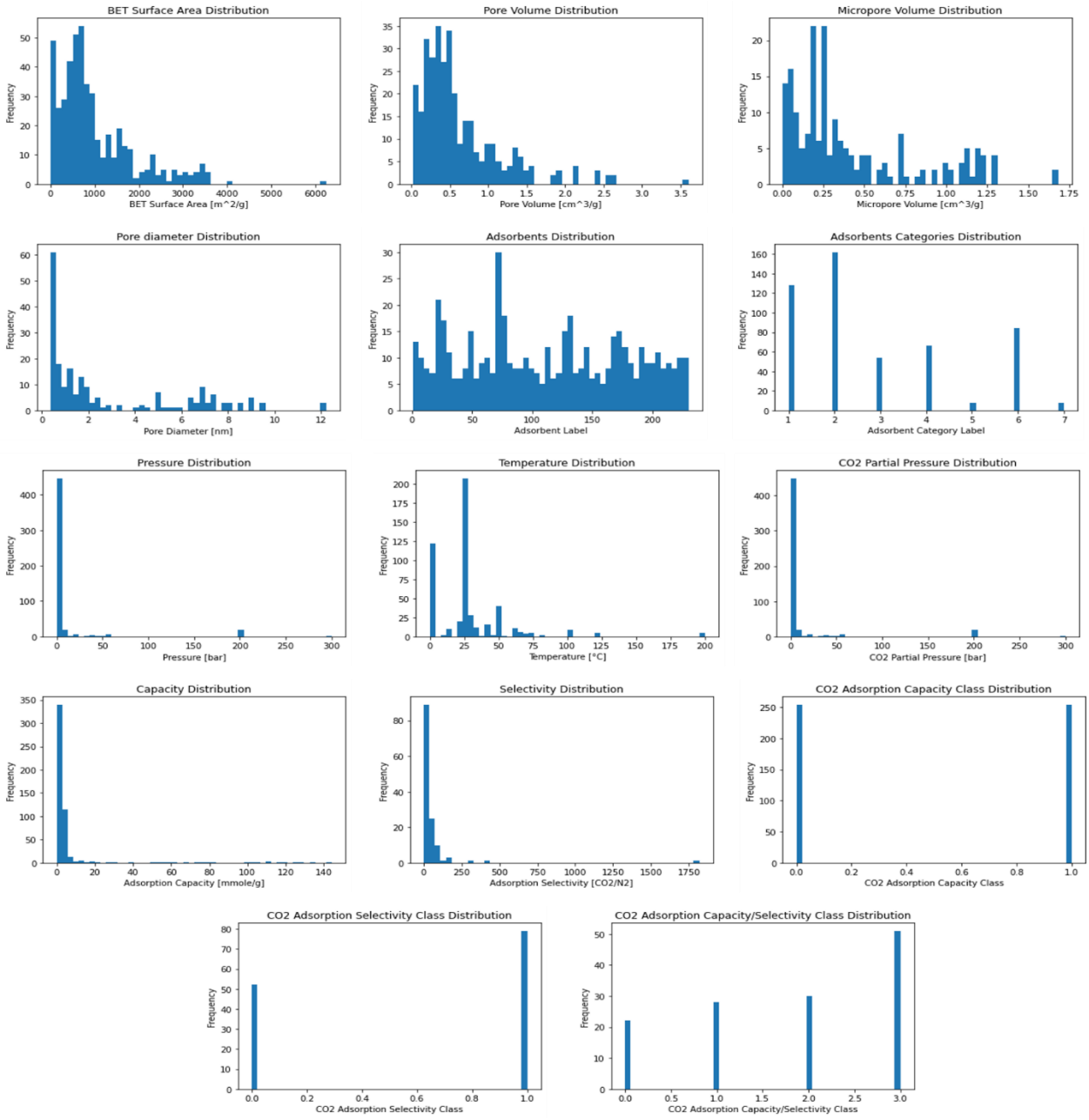


Figure 3.4 Dataset I Variables distributions

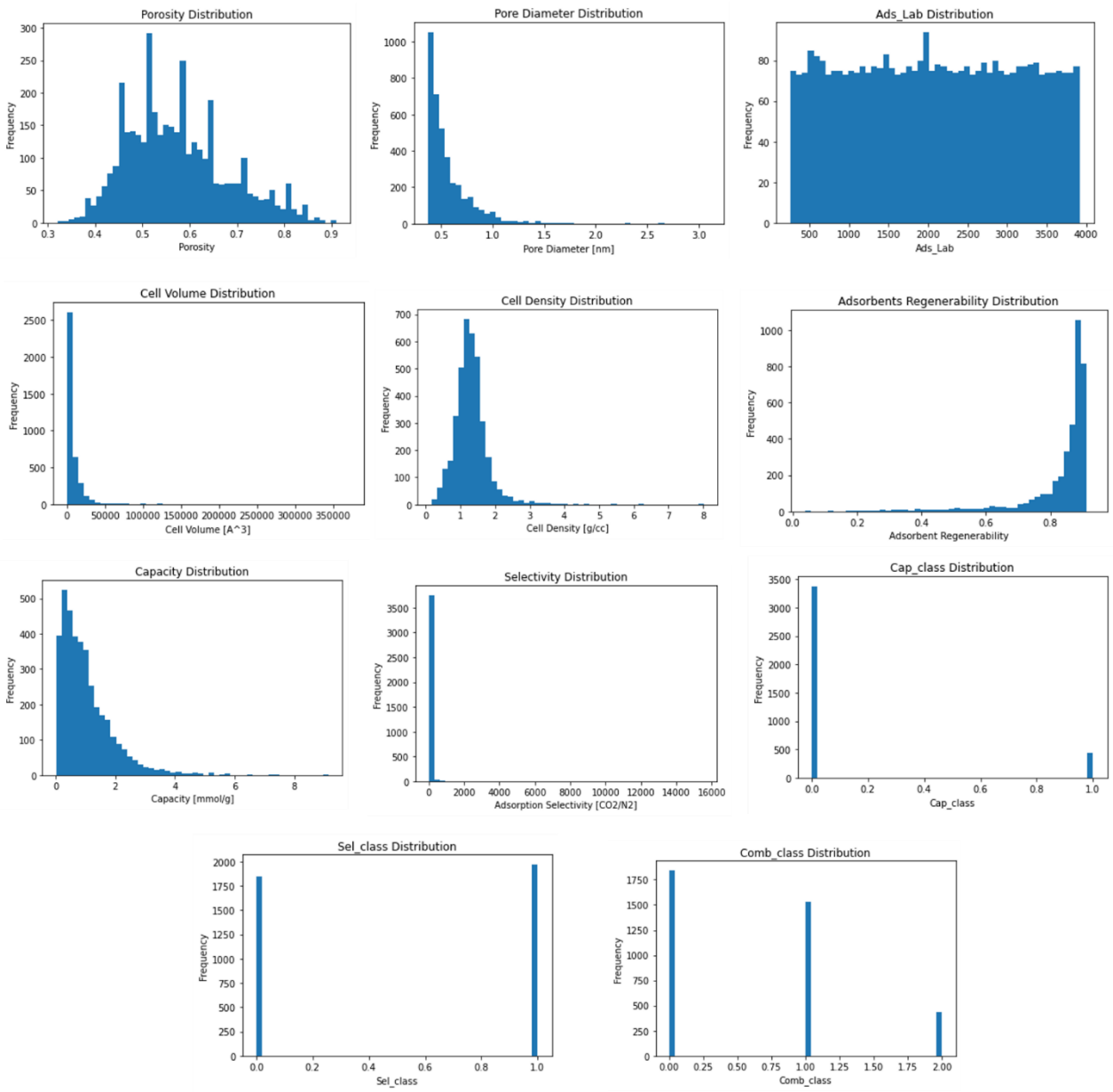


Figure 3.5 Dataset II selected variables distributions

From these two figures, the distributions of all the variables are random and do not follow any known ones such as Gaussian distribution. In addition, the maximum, minimum, and average values of each variable of Dataset I and Dataset II are tabulated in Table 3.4 and Table 3.5, respectively.

Table 3.4 Dataset I simple description

Value	Variables								
	BET [m^2/g]	**VP** [cc/g]	**DP** [nm]	**VMP** [cc/g]	**CPP** [bar]	**P** [bar]	**T** [^oC]	**Capacity** [mmol/g]	**Selectivity** [CO_2/N_2]
Maximum	6240	3.6	31.23	1.68	300	300	750	144	1818
Minimum	2.4	0.01	0.38	0	0.002	0.002	-78	0.004	1.6
Average	951	0.61	3.86	0.42	9.77	9.97	57	7.1	52

Where, (BET) is BET specific surface area, (VP) is pore volume, (VMP) is micro-pore volume and (DP) is pore diameter, (T) is capture temperature, (P) is capture pressure and (CPP) is the feed's carbon dioxide partial pressure.

Table 3.5 Dataset II simple description

Value	Variables						
	Cell density [g/cc]	**Cell volume** [A^3]	**DP** [nm]	**Porosity**	**Regenerability**	**Capacity** [mmol/g]	**Selectivity** [CO_2/N_2]
Maximum	8.02	371231.44	3.1	0.91	0.931	9.151	15488.43
Minimum	0.16	369.68	0.375	0.32	0.039	0.009	1.65
Average	1.3	9052.24	0.576	0.574	0.83	1.02	48.65

Where, (DP) is pore diameter.

As aforementioned, to get a preliminary relation between each two variables available in the datasets, several correlation matrices based on Pearson's correlation factor were generated. Figure 3.6 and Figure 3.7 illustrate the different correlation matrices for Dataset I and Dataset II.

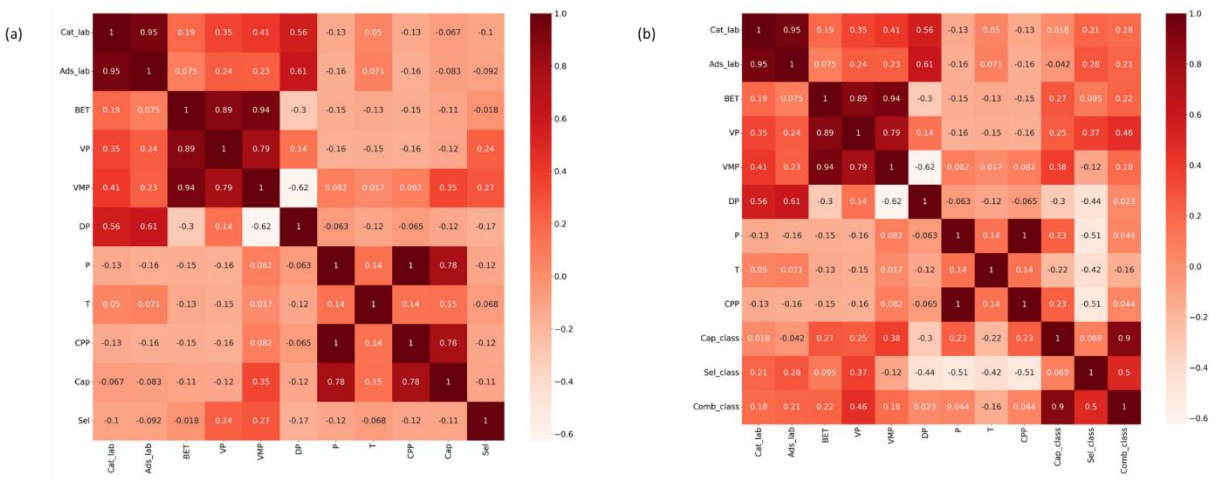


Figure 3.6 a. Correlation matrix using real values of capacities and selectivities (Dataset I), b. Correlation matrix based on capacities and selectivities categorization (Dataset I)

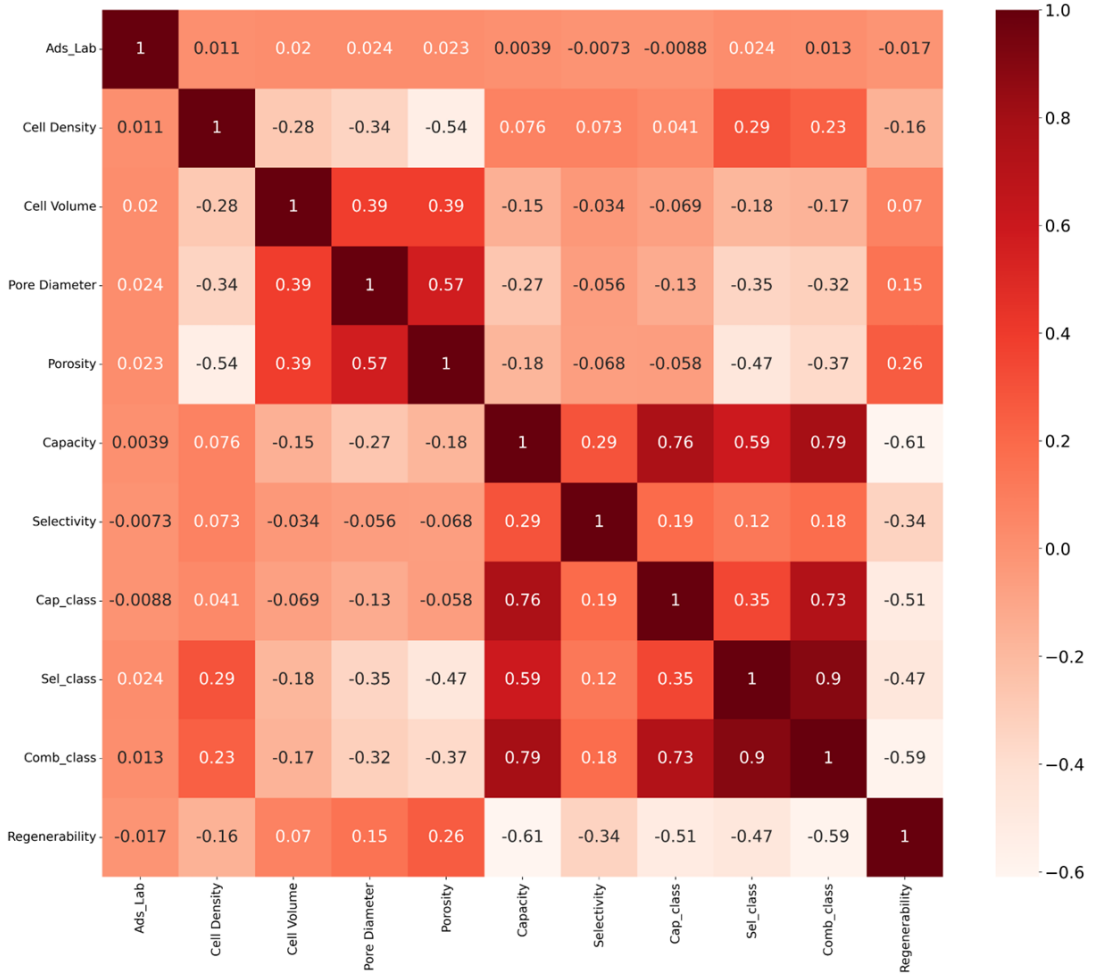


Figure 3.7 Correlation matrix based on capacities and selectivities categorization (Dataset II)

As a general observation, there are significant correlations between process conditions, adsorbents' textural properties and capacity as well as selectivity. For instance, there is a negative correlation between adsorbents' pore diameter and both capacity and selectivity. In contrast, pore volume has positive correlation with capacity and selectivity. On the process condition level, temperature has negative correlation with capacity and selectivity. However, adsorption pressure is positively correlated with capacity and negatively correlated with selectivity. This means that the process pressure should be optimized to have acceptable levels of capacity and selectivity. On the variables level, BET surface area is strongly correlated with pore volume and micropore volume where the correlation is positive. However, the correlation between BET surface area and pore diameter is a significant negative one. In addition, micropore volume has a strong negative correlation with pore diameter, whereas the correlation between pore volume and micropore volume is a strong positive one which is reasonable as the micropore volume is a part of the total pore volume.

In the case of MOFs in Dataset II, cell volume, pore diameter and adsorbent porosity have negative correlations with both selectivity and capacity where those related to capacity are more significant. As it can be observed, pore diameter is highly correlated with MOF porosity and this correlation is positive. This means that increasing the porosity increases the mesopores (in the range of 2-50 nm) and possibly decreases the micropores (less than 2 nm) and micropore volume which is the part of pore volume responsible for enhancing carbon dioxide adsorption selectivity and capacity. With respect to cell volume, this volume is the unit cell volume of the MOF crystal, and it is positively correlated with the pore diameter where the larger the cell volume the larger the pore diameter and vice versa. Contrary, cell volume, pore diameter and adsorbent porosity have positive correlations with MOF regenerability. This suggests that it may be preferred to design MOFs with larger cell volumes, pore diameters and porosities to ease the regeneration process. However, care should be taken to avoid low selectivities and capacities. This means that robust optimization should be performed to have an adsorbent with reasonable capacity, and selectivity while possessing easy regeneration to enhance the feasibility of capture process.

3.4.2 Results of data-driven models

This section summarizes the most important results obtained from ML modeling of the available datasets using DTs, ANNs, and XGBoost.

Decision tree (DT) modeling

The main results presented here are the patterns, confusion matrix and other evaluation metrics besides the average variables importance graphs as an additional preliminary way of representing the dependence of capacity/selectivity on the available variables considering the relative importance.

The optimum decision tree developed for capacity classification was the one having a maximum depth of 8. Table 3.6 summarizes the patterns extracted upon employing this optimized decision tree to classify the available adsorption materials in Dataset I according to their capacities' classes, i.e., high capacity (passing ***Condition I***) and low capacity (failing ***Condition I***). There are 14 patterns, out of 27 patterns, that can be followed to achieve high adsorption capacities.

Table 3.6 Decision tree patterns for capacity (Dataset I)

*This table was moved to Appendix B (Table B.1) in (APPENDIX)

Where, Ads_Lab is the adsorbent label, BET is the specific surface area, DP is pore diameter, VP is pore volume and CPP is the carbon dioxide partial pressure.

As it can be depicted, on the process conditions level, the rules in general favor low temperatures and high pressures as common operating conditions to have high capture capacity. On the material design side (textural properties), adsorbents with large pore volumes, small pore diameters and large specific surface areas are the ones that feature high capture capacities, i.e., ***Condition I***. Carbon dioxide partial pressure has a significant effect like the effect of total pressure. The adsorbent label refers to the adsorbent type and functionalization/treatment method. In general, the functionalized materials possess higher capacities. In this regard, one of the patterns to have high capacity is P1 where the process should be operated at temperatures around 24ºC or lower. On the adsorbents textural properties side, the adsorbents should have a surface area of about 854.5 m^2/g or lower and pore volume higher than 0.067 cc/g. Besides, the pore diameter of this case equals to or is less than 1.145 nm. In addition, according to pattern P14, to have a high adsorption capacity when working at temperatures above 45ºC it is recommended to employ an adsorbent having surface area higher than 1655 m^2/g and pore volume higher than 1.415 cc/g where the partial pressure of carbon dioxide is higher than 0.75 bar. The full list of adsorbents having high adsorption capacity is tabulated in the supplementary materials. SNS2-20 (porous carbon), COF-5 (COF) and

COF-JLU2 (COF) are among those materials having high capture capacity covered by patterns P14, P9 and P1, respectively. Their corresponding adsorption capacities are 2.92, 31 and 5 mmol CO_2/g adsorbent.

The average variables importance related to adsorption capacity is presented in Figure 3.8 (a). It suggests that BET surface area textural property has the biggest influence on the capacity. Then, pore diameter came second, followed by temperature, carbon dioxide partial pressure, adsorption pressure and pore volume, respectively.

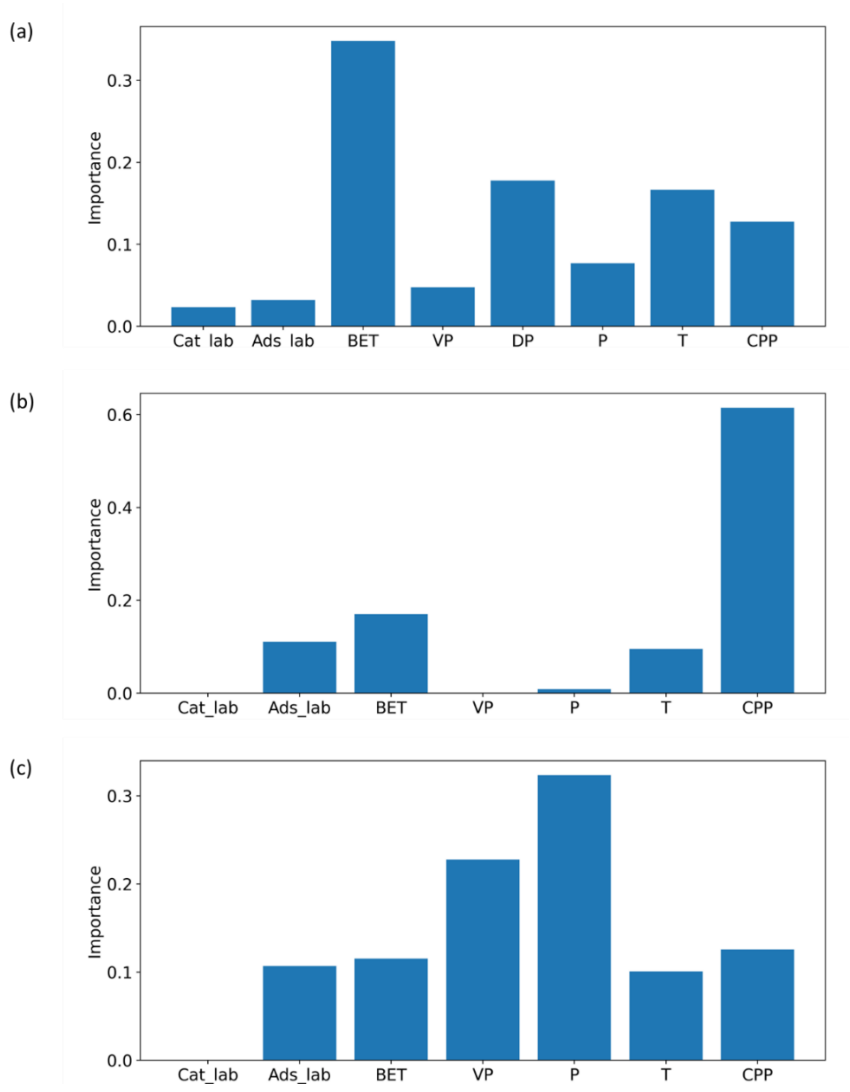


Figure 3.8 Variables importance related to (a) adsorption capacity, (b) adsorption selectivity, and (c) capacity/selectivity combined classification (Dataset I). Where, Cat-lab is the high-level adsorbent category, Ads_Lab is the adsorbent label, BET is the specific surface area, DP is pore diameter, VP is pore volume and CPP is the carbon dioxide partial pressure.

Similarly, in the case of selectivity-based classification, decision tree was able to generate patterns to identify the adsorbents having low selectivity (failing ***Condition II***) or high selectivity (passing ***Condition II***). The optimum tree in this case has a maximum depth of 7. Table 3.7 introduces those patterns related to Dataset I where there are 5 patterns, out of 12, that lead to satisfying ***Condition II***. In this case, lower adsorption temperatures and pressures are favorable. Yet, the adsorbents are preferred to have larger BET surface areas for better adsorption selectivity. According to pattern P1, high adsorption selectivities can be achieved by operating at pressures under 3.5 bar and temperatures below 62.5ºC using an adsorbent having BET surface area higher than 22.25 m^2/g. The full list of adsorbents having high adsorption selectivity is tabulated in the supplementary materials. Cu-BTTri-mmen (MOF), PECONF-1 (POP) and HCM-DAH-1 (activated carbon) are samples of the high selectivity adsorbents covered by the patterns P1, P3 and P2, respectively. Their corresponding adsorption selectivities are 327, 110 and 28 CO_2/N_2.

About variable importance, as illustrated in Figure 3.8 (b), carbon dioxide partial pressure has the most significant impact on adsorption selectivity followed by BET surface area, adsorbent type, and adsorption temperature. To recall, carbon dioxide partial pressure is highly correlated to adsorption pressure where the correlation is positive. It is worth mentioning that pore diameter was not included in this case as the available data was too scarce which affected the prediction accuracy. This was the motivation to explore the COSMOS dataset to extract additional knowledge about the effect of pore diameter on adsorption selectivity.

Table 3.7 Decision tree patterns for selectivity (Dataset I)

*This table was moved to Appendix B (Table B.2) in (APPENDIX)

Where, Ads_Lab is the adsorbent label, BET is the specific surface area, and CPP is the carbon dioxide partial pressure.

The optimum decision tree developed for the combined capacity/selectivity classification was the one having a maximum depth of 8. As it can be observed in Table 3.8, there are four main patterns that can be followed to have an adsorption process featuring high capacity and high selectivity, i.e., satisfying both ***Condition I*** and ***II***. According to pattern P3, it is better to conduct the adsorption process at a pressure equal to or lower than 12.5 bar where the partial pressure of carbon dioxide is between 1.02 and 0.575 bar. To achieve high capacity and selectivity at these conditions, a pore volume larger than 0.391 cc/g is required. To avoid encountering low capacity & low selectivity in

adsorption process, according to pattern P21, it is recommended to avoid a pressure higher than 12.5 bar in the presence of an adsorbent having a pore volume lower than 0.031 cc/g. The full list of adsorbents having high adsorption selectivity and capacity is available in the supplementary materials.PPN-6-SO3H (POP), Cs@RWY (zeolite) and PFPOP-3 (POP) are samples of the high capacity-high selectivity adsorbents covered by the patterns P3, P1 and P2, respectively. Their corresponding adsorption (capacities & selectivities) are (3.6 & 150), (3.21 & 180) and (4.74 & 56.5) [mmol CO_2/g adsorbent & CO_2/N_2]. In addition, besides the cases used for the training, some other cases/points from the literature were used to validate the proposed methodology for design/operation rules extraction. These cases are summarized in the supplementary materials (Appendix G).

Table 3.8 Decision tree patterns for combined capacity-selectivity classification (Dataset I)

*This table was moved to Appendix B (Table B.3) in (APPENDIX)

Where, Ads_Lab is the adsorbent label, BET is the specific surface area, VP is pore volume and CPP is the carbon dioxide partial pressure.

Pressure followed by adsorbent pore volume and BET surface area have the most significant impact on combined capacity/selectivity (Figure 3.8 (c)). The variables importance graphs give other preliminary useful results about the relative importance/impact of the controlled variables on the responses or measured variables. However, they do not give the complete effect including the magnitude and sign. Hence, other methods, i.e., SHAP values, were developed and employed to overcome this limitation.

Various metrics were used to evaluate the prediction ability of the developed decision tree. The first one is the confusion matrix which is then used to calculate the *F1*-score. Testing confusion matrices for capacity, selectivity and capacity/selectivity combined classification are presented in Figure 3.9. In addition, precisions, recalls and *F1*-scores are summarized in Table 3.9. The green diagonals represent the percentage of correct classifications while the red zones represent the percentage of misclassified adsorbents. From Figure 3.9, it can be concluded that the developed models have high classification accuracy and that explains the high values of *F1*-scores of these models as illustrated by Table 3.9.

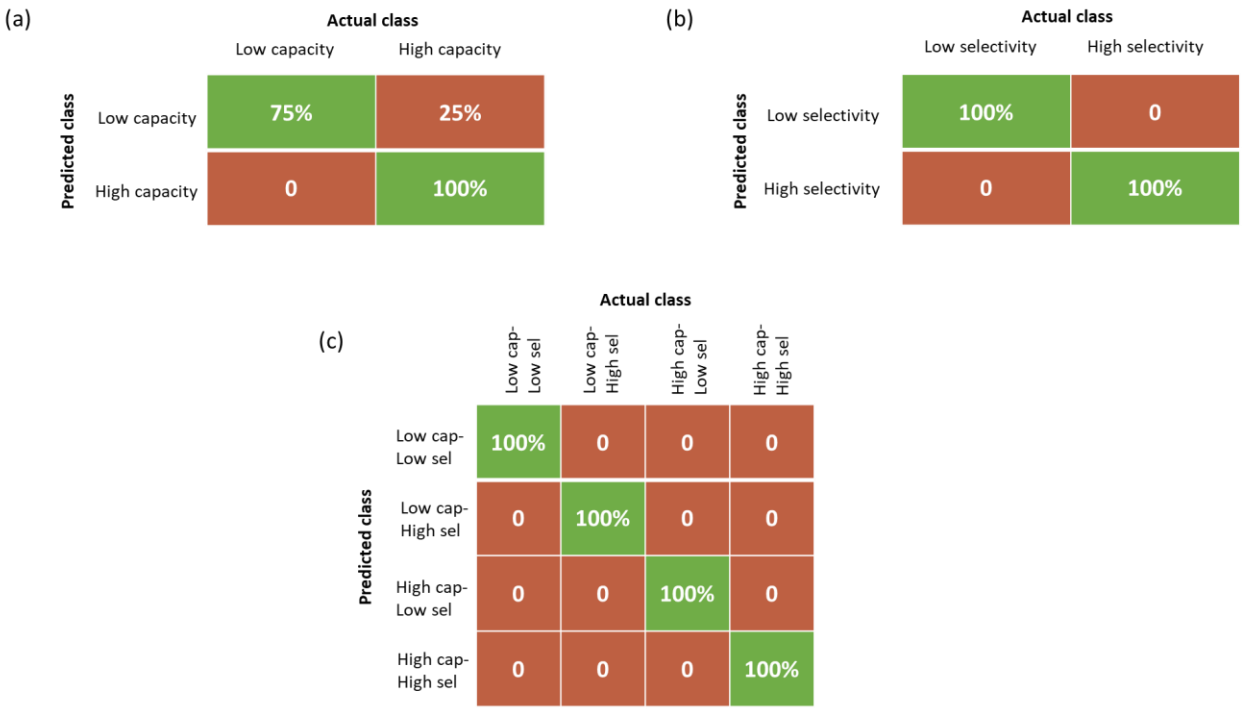


Figure 3.9 Confusion matrix for: a) capacity, b) selectivity and c) combined capacity/selectivity DT classification (Dataset I)

Table 3.9 Testing precisions, recalls and *F1*-scores of capacity, selectivity and capacity/selectivity combined DT classification (Dataset I)

Metric	**Classification Case**		
	Capacity	**Selectivity**	**Capacity-Selectivity**
Precision	0.571	1	1
Recall	1	1	1
***F1*-score**	0.73	1	1

The same procedure was followed in the case of Dataset II where classification prime patterns based on adsorbents labels and pore diameters were obtained. Selected patterns are presented in supplementary materials. The accuracy in the case of Dataset II was lower and this is due to the

absence of other important variables that were not reported clearly in detail by the creators of this MOFs-based dataset (COSMOS). Among these missing variables are BET surface area and pore volume. However, working on this dataset gave additional knowledge about the effect of adsorbents porosity and pore diameter in addition to the specific knowledge related to MOFs structure such as the cell density and cell volume. For instance, according to the available data in this dataset, it was found that increasing pore diameter, adsorbent porosity and its unit cell volume hinders both selectivity and capacity of carbon dioxide adsorption.

The training and testing accuracies for the classification of capacity, selectivity and combined capacity/selectivity are (100% & 83%), (100% & 71%) and (100% & 61%), respectively. The optimum developed trees with spanned patterns have maximum depth of 10. The other trees with prime patterns to classify capacity, selectivity and combined capacity/selectivity have maximum depths of 1, 3 and 4, respectively. One of the important prime patterns is to have an adsorbent with pore diameter lower than 0.442 nm to ensure high adsorption selectivity.

As mentioned above, DT was used as an excellent interpretable machine learning technique to give clear patterns/rules for the design/synthesis of a good adsorbent based on the textural properties. These rules considered the operation side as well based on the partial pressure of carbon dioxide, temperatures, and pressures of the adsorption process. However, to have higher prediction/classification accuracies, other methods/techniques, such as ANN and XGBoost, were used.

Results of ANN modeling and explainable ML modeling

The optimum ANN structure contained 4 hidden layers with 70, 10, 20 and 30 neurons in each hidden layer, respectively. Additionally, the optimum learning rate was found as 0.001. Upon employing this optimized ANN for the present classification problem, the accuracies increased in the case of capacity classification to 96% compared to only 81% in the case of using DT. However, this accuracy decreased in the selectivity classification case compared to the results obtained from the DT. The same problem was encountered in the case of selectivity/capacity combined classification. Hence, XGBoost was used to overcome this problem.

Results of XGBoost and explainable ML modeling

The optimum XGBoost model after hyperparameters optimization was determined as the one having 100 estimators. Fortunately, this model outperformed ANN and DT with respect to accuracy and training time. In particular, the accuracy exceeded 99% and the model was 10 times faster for all cases of classification related to Dataset I. Confusion matrices as well as SHAP values figures resulted from the developed model are shown in Figure 3.10.

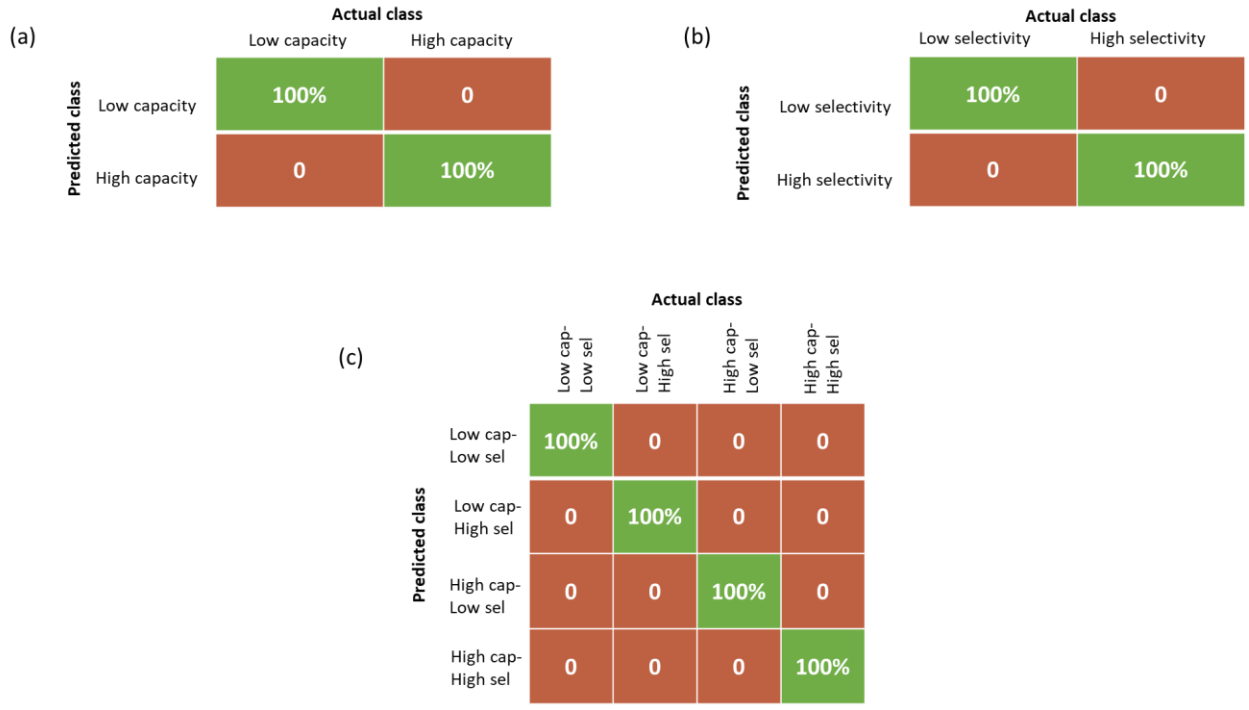


Figure 3.10 Confusion matrices for (a) capacity class prediction, (b) selectivity class prediction, and (c) combined capacity/selectivity classification by the XGBoost (Dataset I)

As shown, the developed model was accurate enough to precisely classify/screen the high and low-capacity adsorbents in the test set without any misclassification. In particular, the low-capacity adsorbents in this set are all accurately classified. Similarly, this was achieved in the case of the high-capacity adsorbents. The same high accuracy was achieved in the training step based on the training set.

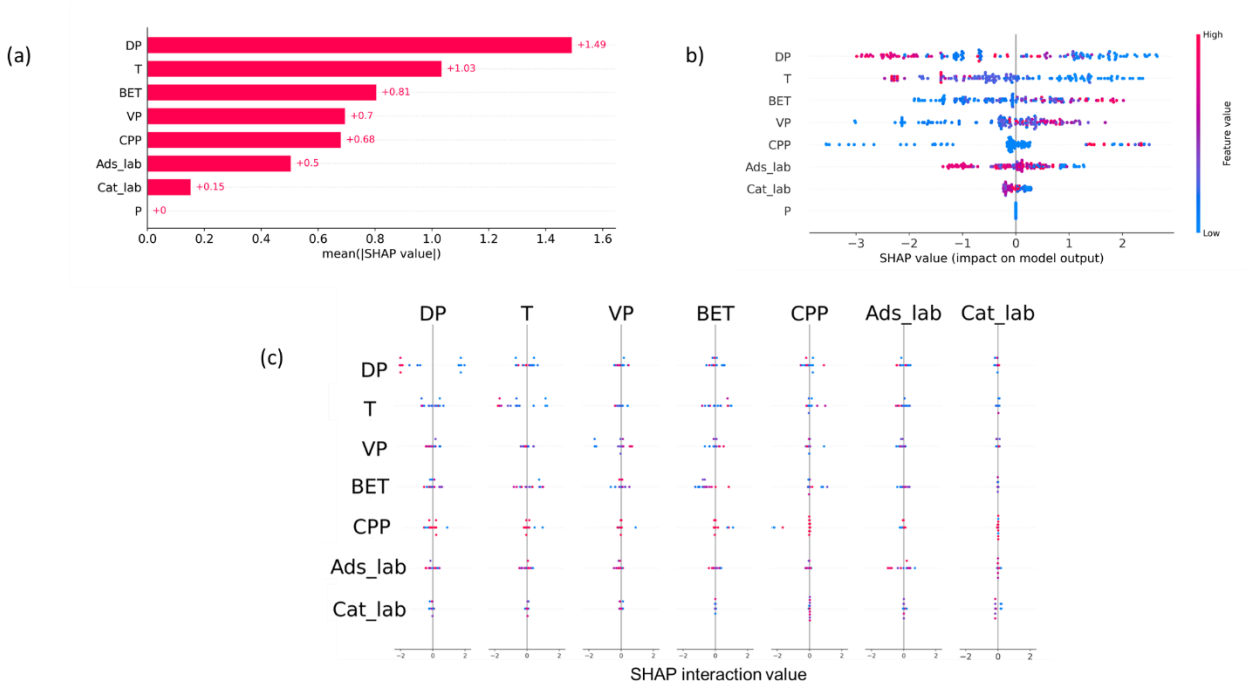


Figure 3.11 **(a)** SHAP values for capacity class prediction by XGBoost, (b) impact of individual variables on capacity, and (c) impact of variables interactions on capacity (interactions values scale is between -2 & 2)

Based on this figure and according to the available data, it can be concluded that pore diameter has the highest impact on the adsorbent's capacity followed by process temperature, BET surface area, pore volume and carbon dioxide partial pressure which is positively correlated with the process pressure. The effect of pore diameter is negative which suggests that the lower the pore diameter the higher the carbon capture capacity. The process temperature also affects the capacity negatively. On the other hand, BET surface area, pore volume and carbon dioxide partial pressure have positive impacts on the capacity. The magnitudes of relative impacts of these variables are shown in Figure 3.11 (a).

According to Figure 3.11 (c), the effects of temperature (T)-temperature (T) interaction as well as the pore diameter (DP)-pore diameter (DP) one on the capacity are negative. Besides, due to the negative impact of both DP (textural property) and T (adsorption condition), their interaction also has a negative effect on capacity. On the other hand, the interaction between BET surface area and pore volume (VP) has a positive impact on the adsorption capacity due to the individual positive impact of both BET and VP variables.

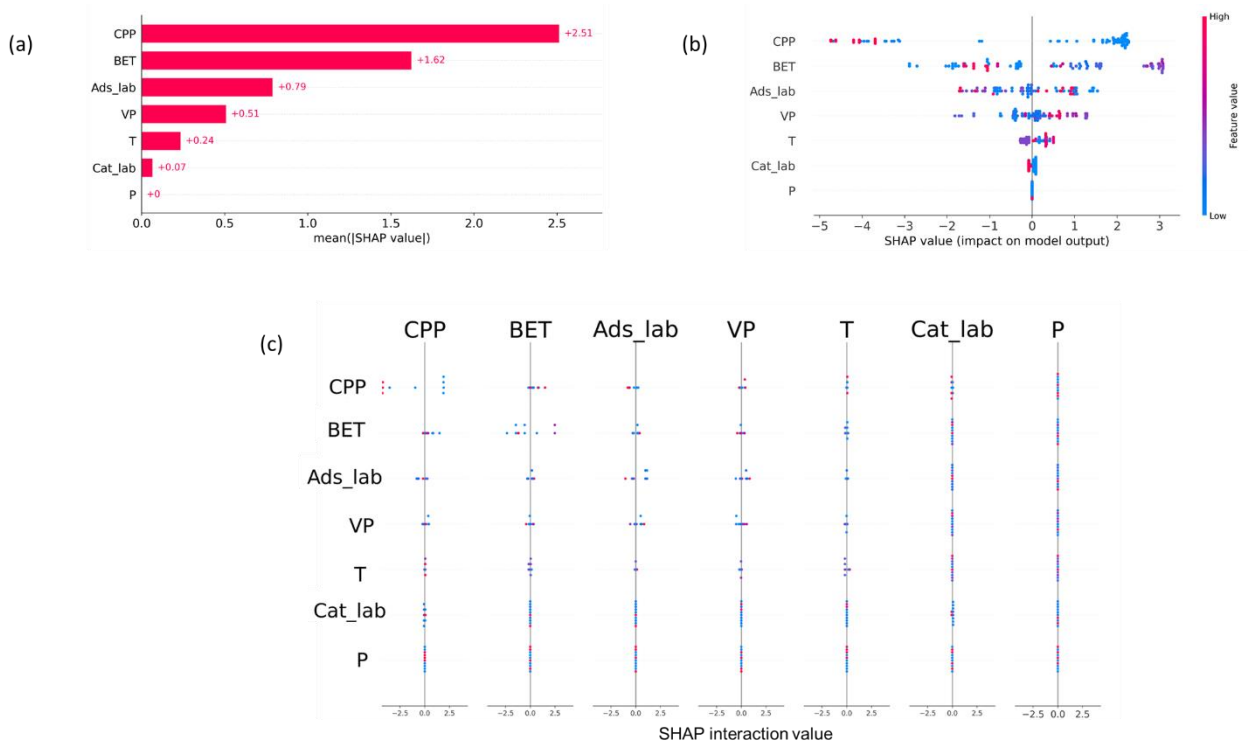


Figure 3.12 SHAP values for selectivity class prediction by the XGBoost, (b) impact of individual conditions on selectivity, and (c) Impact of variables interactions on selectivity (interactions values scale is between -2.5 & 2.5)

From Figure 3.12, it can be concluded that carbon dioxide partial pressure has the most significant impact on adsorption selectivity followed by BET surface area, adsorbent type, pore volume and adsorption temperature. Carbon dioxide partial pressure is correlated to adsorption pressure where the impact of the partial pressure on the selectivity is negative. Temperature also has a negative impact. However, the surface area and pore volume have positive impact. With respect to the variables interactions, the most significant ones in the case of selectivity classification are the interaction between the effect of BET surface area and carbon dioxide partial pressure (CPP), BET-BET interaction and CPP-CPP interaction. Where the BET-CPP interaction has a positive impact on selectivity like the BET-BET interaction. Nonetheless, the CPP-CPP interaction is a negative one that hinders adsorption selectivity.

Other attempts were made to consider the pore diameter and investigate its effect on the selectivity. However, the number of data points were relatively small in this case but fortunately they were representative to capture the effect of pore diameter. In addition, the prediction accuracies for training and testing in this case are 100% and 90%, respectively. Figure 3.13 shows the results of

SHAP values determination in this case. As can be deduced, the increase in pore diameter negatively affects selectivity. This is presumably due to the fact that the kinematic/molecular diameters of carbon dioxide are smaller than those of nitrogen [333]. Where the kinematic diameter of carbon dioxide and nitrogen are 0.33 nm [334] and 0.364 nm [335], respectively. This is one of the reasons making the adsorbents having smaller pore diameters more selective towards carbon dioxide.

In general, the reduction of adsorbents' pore diameters can lead to size-exclusion effects [336], [337]. Smaller pores restrict the access of nitrogen's larger molecules while allowing the smaller molecules of CO_2 to enter and absorb onto the adsorbent surface. This size selectivity phenomenon is known as "Molecular Sieving" [338]–[340]. In more detail, the kinematic diameters of nitrogen and carbon dioxide are 0.364 and 0.33 nm, respectively. Hence, according to the aforementioned "Molecular Sieving Mechanism", an adsorbent with pore diameter below 0.44 nm gives higher opportunity to carbon dioxide molecules to diffuse in its pores and then be adsorbed/caged inside this porous structure, unlike the case of nitrogen. Having bigger kinematic diameter makes nitrogen molecules in the flue gas stream face increased diffusion barriers/ resistance and consequently will not be adsorbed by the solid-based capture material.

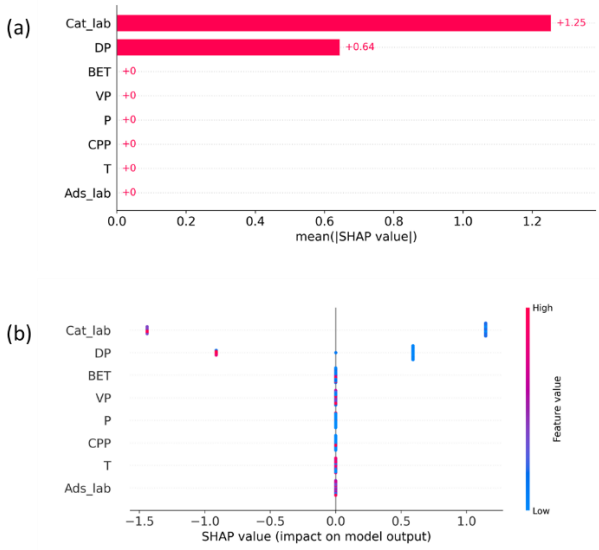


Figure 3.13 a) SHAP values for selectivity class prediction by XGBoost and b) impact of individual conditions on selectivity (pore diameter considered) Dataset I

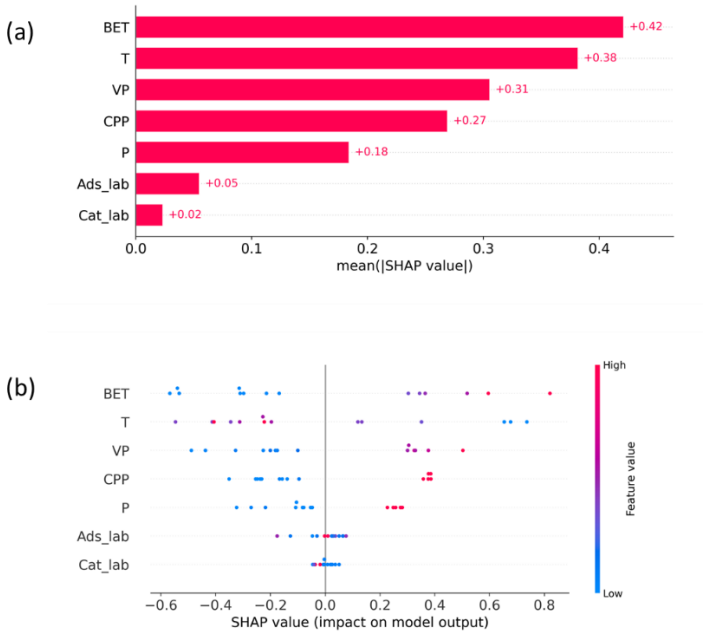


Figure 3.14 (a) SHAP values for combined capacity/selectivity classification and (b) impact of individual conditions on combined capacity/selectivity classification

The SHAP values in Figure 3.14 suggest that BET surface area is the most significant variable followed by temperature and pore volume when considering the combined capacity/selectivity. The effects of BET, temperature and pore volume are positive, negative, and positive, respectively. In addition, Figure 3.10 illustrates the high precision of prediction and *F1*-scores of the developed models. They are equal to 1 in all the three classification cases.

Upon working on Dataset II, as illustrated in Figure 3.15, Figure 3.16 and Figure 3.17, it was found that pore diameter side by side with adsorbent porosity have the highest impact on adsorption capacity and selectivity. These impacts are negative, which agrees with the previous results of Dataset I. In addition, the DP-DP and porosity-porosity interactions have significant impacts on both adsorption capacity and selectivity.

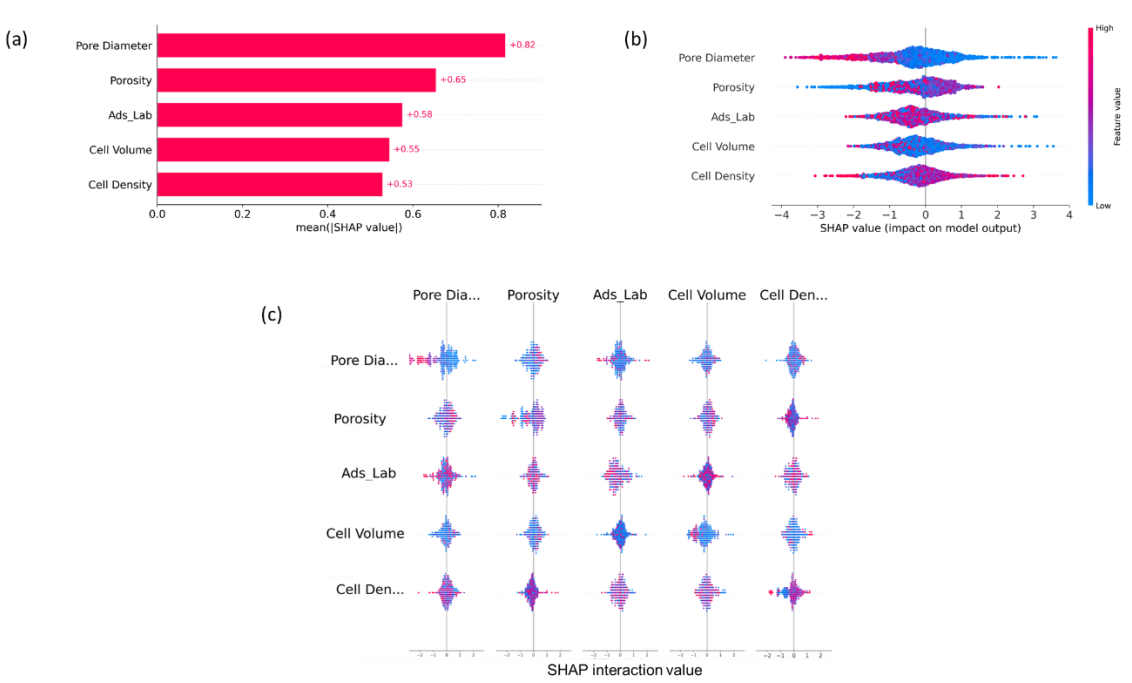


Figure 3.15 SHAP results for capacity classification (Dataset II) (interactions values scale is between -2 & 2)

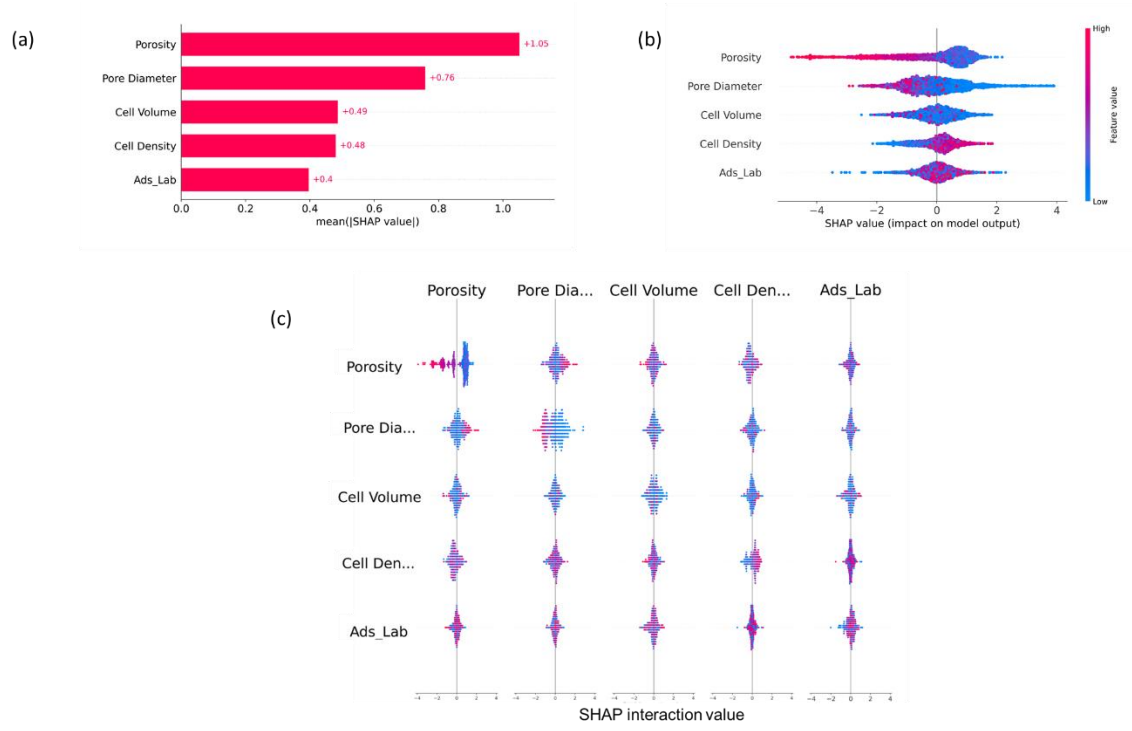


Figure 3.16 SHAP results for selectivity classification (Dataset II) (interactions values scale is between -4 & 4)

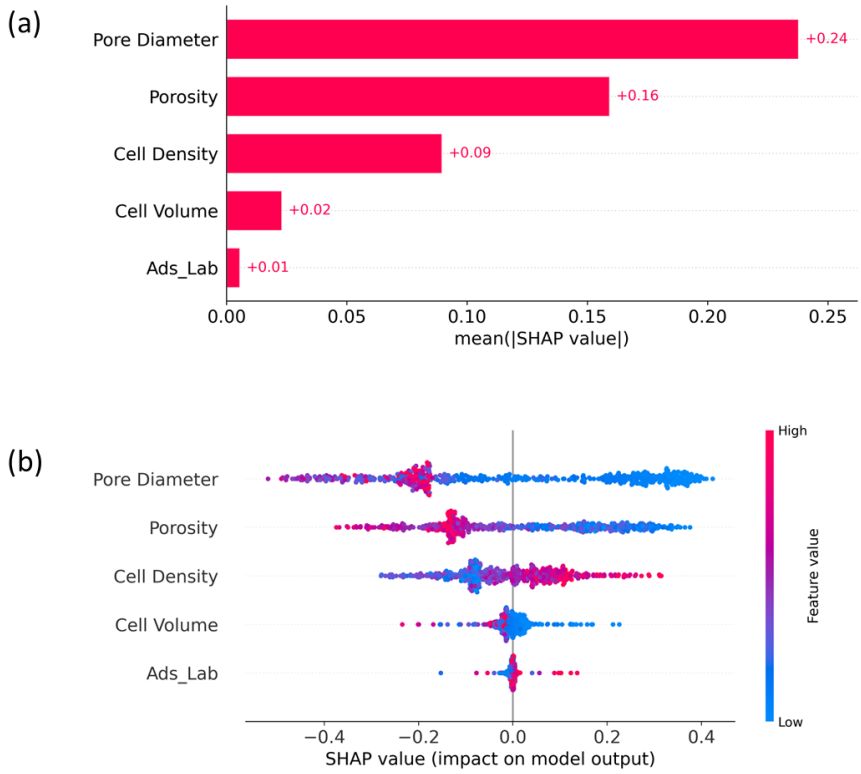


Figure 3.17 SHAP results for combined capacity/selectivity classification (Dataset II)

High temperature adsorbents classification

As previously mentioned in the methods section, *EXPLORE* software was used to firstly cluster the available data about adsorbents. In this regard, two main groups/clusters were obtained. The first one is the adsorbents working at relatively lower temperatures which does not exceed 200 ºC. Whilst the second one corresponds to the adsorbents working at elevated temperatures above 200 ºC which can reach 750 ºC. These two clusters are shown in Figure 3.18 (a).

Upon applying the developed methodology on the cluster of high temperature adsorbents, an optimized tree with maximum depth of 4 was generated to give 4 distinct patterns/rules to differentiate between low-capacity and high-capacity adsorbents accurately. These patterns are summarized in Table 3.10. Besides, Figure 3.18 (b) depicts the preliminary quantification of the relative importance of different variables. Where the precision, recall and *F-1* score of this optimized tree equal 1.

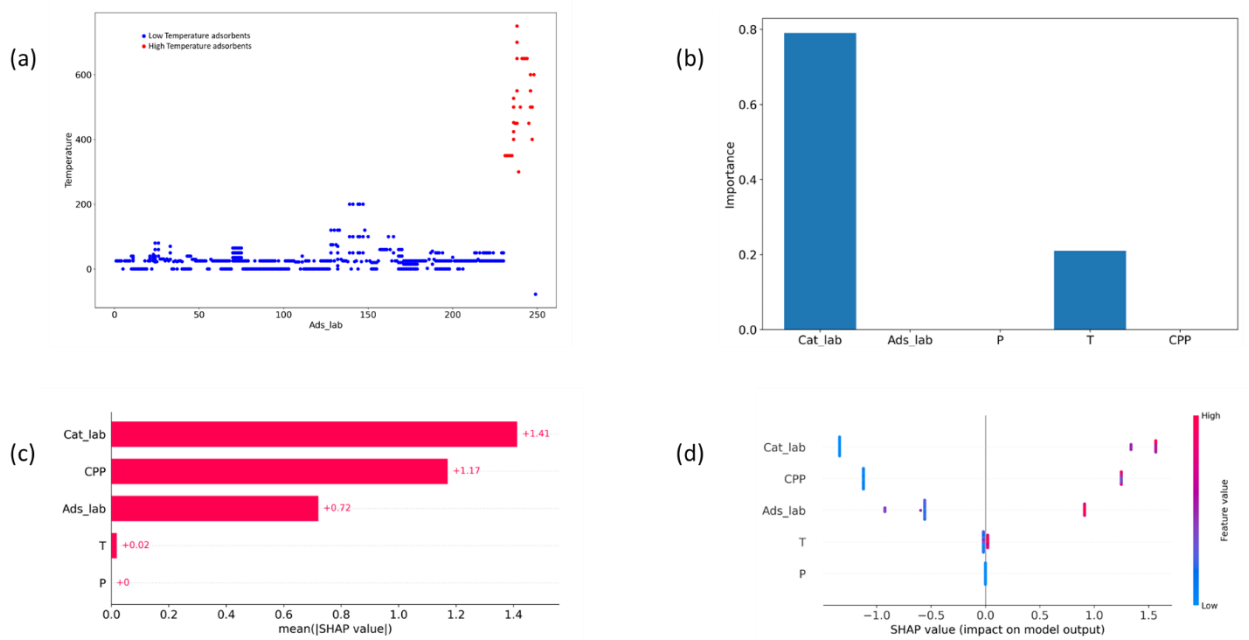


Figure 3.18 (a) Low and high temperature adsorbents, (b) Variables importance related to capacity classification of high temperature adsorbents, (c) SHAP values for capacity classification of high temperature adsorbents, and (d) impact of individual variables on high temperature adsorption (Dataset I)

Table 3.10 Patterns for capacity classification of high temperature adsorbents (Dataset I)

Pattern ID	**Category Label**	**Temperature (°C)**	**Pressure (bar)**	**Ads_Lab***	**CPP* (bar)**	**Capacity class**
P1	> 8	350 < T <= 700				High (above 2)
P2	<= 8					
P3	> 8	<= 350				Low (below 2)
P4	> 8	> 700				

* Ads_Lab is the adsorbent label, and CPP is the carbon dioxide partial pressure.

To have a high adsorption capacity it is recommended that the category/type of the adsorbent to be calcium oxide and its derivatives or lithium zirconate and its derivatives. In addition, the

recommended range of operating temperatures is from 350 ºC to 700 ºC where the optimal region is around 650 ºC -700 ºC.

Moreover, the prediction of capacity class using XGBoost with 100 estimators resulted in a high accuracy for both training and testing phases, i.e., exceeded 99%. Where the precision, recall and *F-1* score 1 for this optimized model as well. The explainability part gave more thorough insights about the impact of different variables on the adsorption capacity of high temperature adsorbents. These insights are summarised in Figure 3.18 (c and d). For instance, the most impactful variables on adsorption capacity are the category of the high temperature adsorbent followed by carbon dioxide partial pressure, chemical composition represented by adsorbent label, and operating temperature. The best performing adsorbent related to this cluster is *CaO/$Ca_{12}Al_{14}O_{33}$ (75% CaO)* with an adsorption capacity of 11.6 mmol CO_2/ g adsorbent. This adsorbent follow pattern P1 where its category is the calcium oxide-based adsorbents and the operating temperature to have this capacity is 650 ºC.

3.5 Remarks, Recommendations and Future Work

It is well known among CCUS stakeholders that economic feasibility is one of the most important performance indicators in the CCUS value chain. The capture process is considered the most expensive part of this chain. Accordingly, to achieve such goal of CCUS economic feasibility, the capture process/material design acceleration is inevitable. This study represents a good step towards having a cost-effective adsorption-based carbon capture process. It proves the applicability of machine learning to accelerate this task. One of the important lessons learned is that no single method can work perfectly in all applications. This means that it is recommended to diversify the ML methods used. For instance, in the current study IML was utilized to give interpretable patterns, with no need to know the training process and its details. Besides, explainable AI (XAI) was used to overcome the problem of black-box modeling, where sensitivity analysis is done and the average weights/effects of different controlled variables on the outputs/predicted variables are determined. This achieves a reasonable balance in the context of Accuracy – Explainability tradeoff.

Another crucial lesson learned is that data is the fuel of AI, and the role of experts is essential. Data collection and processing need careful attention from the expert to select the essential variables to be studied and prepare analysis ready data. This role is also crucial in the modeling phase where the human expert approves and validates the obtained results.

The presented classification/selection work will accelerate adsorption materials design through providing the rules to be followed to have an optimum adsorbent. Besides, the results of this work will be the input to process design phase where the selected optimum adsorbent(s) and the optimum operating conditions will be considered during the design. This will indeed lead to the enhancement of carbon capture process feasibility.

However, there is still a need for extensive research work in the future to achieve higher design acceleration and process feasibility. There is still a lack of information in the literature related to cost and cyclability (number of regeneration cycles). These are important characteristics that should be investigated/evaluated and reported in future studies. Where the future work related to machine learning-assisted carbon capture materials selection is to update the database with this important data continuously. It is believed that the addition of more representative data will lead to more insights about adsorbents design/discovery. This will be done under the umbrella of a unified and open platform that will serve different CCUS stakeholders for better decision making where data sharing and interdisciplinary collaboration on many levels will be encouraged. Our current study is a step towards the implementation of this open platform which will pave the road to large-scale CCUS value chain adoption. Additionally, adsorbents used for direct carbon capture from atmosphere can be included where the developed methodology can be extended to include these types of adsorbents and classify them. Besides, the utilization of ensemble of models including language models and image recognition techniques combined with generative deep learning will enhance the prediction accuracy and accelerate the design. The future utilization of state-of-the-art natural language processing (SOTA-NLP) models to dig into the diverse data sources including research articles, review papers, technical reports and white papers for information retrieval is essential. The extraction of more valuable data/information will enhance the database continuously and hence more insights will be provided to the experts. The combination of more representative data and utilization of ensemble of models has the advantages of ensuring/improving the results reproducibility and creating enhanced and unified decisions as summarized in Figure 3.19.

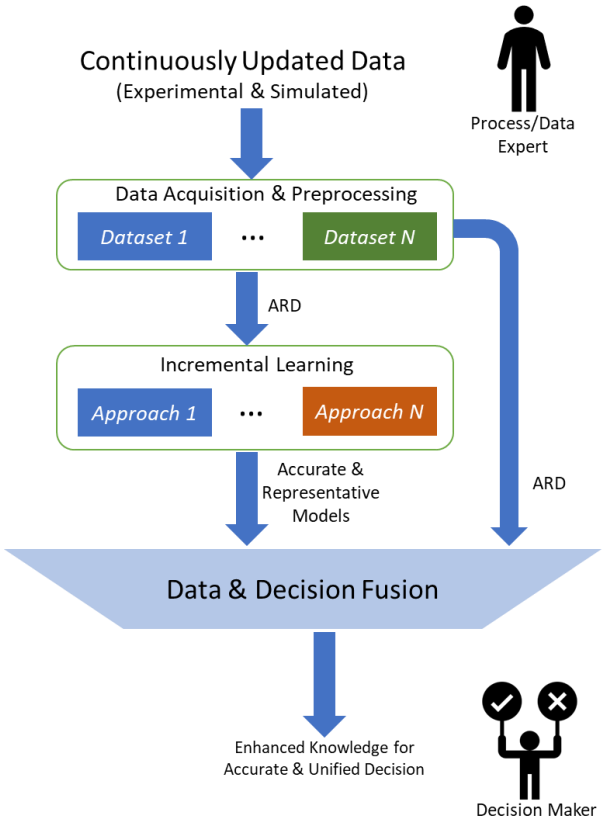


Figure 3.19 General methodology for decision fusion and enhancement

3.6 Conclusion

The combined interpretable-explainable ML methodology developed in this study proved its ability to extract augmented knowledge from data scattered in the literature about adsorbents design/discovery. This knowledge is in the form of patterns/rules and representations about the effect of each variable on both adsorption capacity and selectivity. In this regard, two datasets were prepared and employed for knowledge extract where the prediction accuracy exceeded 98% in three classification cases and the calculated *F1*-scores further proved high predictability. According to the rules, it is recommended to design a chemically modified adsorbent with large pore volumes (e.g., between 0.5 and 1.45 cc/g or higher), high surface area (e.g., 450 m^2/g or higher) and a pore diameter lower than 0.442 nm, to achieve high adsorption capacity and selectivity. On the process conditions side, it is preferred to operate at low temperatures at around 20 ^oC, and pressures higher than 1 bar to guarantee high adsorption capacity. Yet, lower pressures are preferred for high selectivities. This is a multi-objective optimization problem where the objectives are competing. Hence, an extensive optimization step should be performed after

selecting the optimum adsorbent to achieve the highest possible capacities and selectivities. In future work, the constructed datasets can also be updated with data about adsorbents for direct carbon capture. The developed methodology is generalizable and could cover this type of adsorbents as well. Where, as aforementioned, it can be applied to optimize both process operation and design besides the optimization of material design/discovery.[1]

[1] All supplementary materials related to this chapter are available in (Appendix G)

CHAPTER 4 ARTICLE 2: ENSEMBLE MACHINE LEARNING TO ACCELERATE INDUSTRIAL DECARBONIZATION: PREDICTION OF HANSEN SOLUBILITY PARAMETERS FOR STREAMLINED CHEMICAL SOLVENT SELECTION

Eslam G. Al-Sakkari (Eslam Ibrahim), Ahmed Ragab, Mostafa Amer, Olumoye Ajao, Marzouk Benali, Daria C. Boffito, Hanane Dagdougui and Mouloud Amazouz

Article Submitted on September 23, 2024 Published on December 13, 2024 in Digital Chemical Engineering, Elsevier (IF 4.1), Volume 14, 2025, Article number 100207

Online Publication date: December 2024

DOI: https://doi.org/10.1016/j.dche.2024.100207

4.1 Abstract

Several processes and strategies have been developed to promote the utilization of lignin and to facilitate its market adoption across a broad spectrum of applications within the expanding lignin bioeconomy. However, the inherent variability in lignin properties, resulting from diverse feedstock sources and varied recovery and downstream processing methods, remains a significant challenge. This highlights the critical need to investigate lignin's miscibility and reactivity with polymers and solvents, as most lignin valorization pathways involve mixing, blending, or solubilization. Accurate estimation of Hansen solubility parameters (HSP) is crucial for solvent selection in several fields such as polymer science, coatings, adhesives, lignin-based biorefineries and solvent-based carbon capture. Traditional methods for predicting HSP are time-consuming and involve complex experiments, especially in applications dealing with carbon dioxide and lignin solubility. This paper introduces a novel ensemble modeling methodology based on machine learning (ML) techniques for accurate HSP prediction using Simplified Molecular Input Line Entry System (SMILES) codes as entries. The methodology integrates different ML approaches, including deep and shallow learning, to enhance prediction accuracy. The decision fusion of individual ML models is achieved through a hybrid approach combining non-learnable and learnable methods, resulting in reduced errors and enhanced accuracy. The results highlight the effectiveness of the ensemble-based methodology, which achieved 99% accuracy in predicting dispersion solubility parameters, outperforming other individual ML techniques. The proposed generic methodology, from data preprocessing to decision fusion through diverse ML algorithms, can be applied to various chemical analytics beyond HSP prediction.

4.2 Introduction

The integration of artificial intelligence (AI), machine learning (ML), and AI-based predictive analytics is revolutionizing the fields of chemistry and process system engineering including biorefineries [341]–[351]. Recent studies in prominent journals highlight the transformative potential of these digital technologies [352]–[357]. By harnessing the power of AI, researchers are advancing the understanding and prediction of molecular interactions that govern separation processes and the behavior of complex molecules and materials, enabling more accurate predictions and deeper insights into molecular structures [358]–[362]. Additionally, AI-driven models are being utilized to predict environmental impacts/properties and process emissions

[363][364][365][366], enabling more sustainable and environmentally friendly industrial practices, as well as predicting catalyst selectivity [367]. Moreover, the acceleration of reaction condition identification through AI is significantly reducing the time and cost associated with experimental trials [368], [369].

It is in this context that accurate prediction of lignin properties and their functionality is crucial for the efficient design and operation of large-scale lignin-biorefineries [370]–[373]. The complexity and variability of lignin make it challenging to process, but understanding its properties can lead to significant advancements in producing high-value lignin derivatives [374]. AI and ML methods are essential in accelerating the development and evolution of such biorefineries [375], [376]. AI and ML can analyze vast amounts of data from various sources, including experimental results, literature, and operational data [377]–[380]. This analysis can uncover patterns and relationships that are not immediately apparent, leading to a deeper understanding of lignin properties and processing methods. ML algorithms can create predictive models that accurately forecast lignin behavior under different conditions, which requires advanced AI approach combining ensemble-based learning algorithms and fusion of data from diverse sources.

The reasons for the creation and use of an ensemble-based machine learning for decision fusion include statistical considerations, data-related limitations, and models/techniques-related constraints [381][382]. From the statistical point of view and taking the neural network modeling for classification and regression as an example, it is common that the high accuracy of results during model training does not guarantee the same results when developed models are tested [383]. This is well-known as the problem of AI/ML generalization to model unseen data where their ability to adapt properly to this new and previously invisible information can be hindered due to various aspects [384], [385]. A possible cause is that the testing datasets may not be representative of the same distribution/range as that of the training dataset. However, employing different models complemented by results averaging or majority voting can decrease generalization risk [386]. The data-related limitations are the variation between the processing/modeling of very large or limited available data [387]–[389]. In the case of very large datasets, it is better to split this dataset into smaller arrays to make the model training step more efficient. On the other hand, in the case of limited data, it is difficult to construct a representative dataset where the model can be successfully trained to catch the underlying distribution of the whole population. In this regard, data resampling and constructing other different data inventories from the parent dataset with limited or imbalanced

inputs can be a good choice [390]. These datasets can be modeled using different ML techniques and their results can be subsequently combined/fused to obtain more accurate predictions. With respect to the models/technique-related constraints, no single method or model performs well enough for all problems; instead, each has its merits and demerits as well as regions of best performance. This is well-known as the *no-free-lunch* theorem [391][392]. Hence, the diversification of methods and techniques is a key solution that can play an inevitable role in reaching the right decision [393].

Given the above-mentioned needs and challenges, several methods have been proposed in literature for implementing ensemble-based learners, especially in the classification problems. These methods of ensemble creation include bagging [394][395], boosting [396][397], AdaBoost [398] and stacked generalization [399][400] mechanisms. AdaBoost refers to a particular method of training a boosted classifier. Combining the results of the ensembles takes several forms such as majority voting [401][402], weighted majority voting (weighted averaging) [403][404] and behavior knowledge space [393] for categorical output variables. Some combination methods are specific to continuous outputs, and they include averaging [405], weighted averaging [406] and *Dempster-Shafer* based combination [407].

As a response to these needs and technical challenges of single ML-modeling, a new ensemble-based methodology is proposed for a decision fusion in regression problems seeking high predictability and generalization during testing and validation. This is inspired by the bagging and boosting methods/techniques that have been successfully used in classification problems as well as other ensemble-based modeling time series cases [408]–[410]. To evaluate the performance and effectiveness of the proposed methodology, the prediction of the solubility of various solvents/substances was selected as a case study.

Hansen solubility parameters (HSPs) are well known as a unique predictive approach to quantify solubility considering the molecular permanent dipole and molecular hydrogen bonding interactions. The HSPs are firstly presented by Prof. *Charles M. Hansen* in his PhD thesis in 1967 [411], [412]. They represent the combination of dispersion (δ_D), polarization (δ_P) and hydrogen bonding (δ_H) energies to explain the mechanism of solvation for single solvent or solvents mixture systems [413][414]. Unlike the *Hildebrand* solubility parameter [415] which focuses mainly on the non-polar solvents, *Hansen* solubility parameters are suitable for describing both polar and non-

polar solvents systems [416]. This is because they consider polarization and hydrogen bonding energies as previously mentioned [416].

There are three main traditional approaches/methods for the calculation of different solubility parameters including *Hansen* solubility parameters [417]. These approaches/methods include the functional group contribution method [418][419], the intrinsic viscosity method [420] and the properties correlations or regression method [421][422]. However, these methods have some limitations, including limited access to reliable data, reliance on simplified assumptions (e.g., ideal solution behavior), and the need for approximations. Additionally, they are time consuming as they primarily depend on several lab experiments [417][423]. There is also a new trend to calculate these parameters using computational methods such as molecular dynamics, density functional theory (DFT) and other quantum mechanics-based calculations [424], which are also computationally resource intensive and time consuming despite their lack of robustness. Therefore, new faster methods are needed and in response several data-driven methods have been proposed recently to overcome these limitations [425]. For instance, AI/ML-based approaches are garnering attention to calculate/predict *Hansen* solubility parameters as well as other molecular and thermodynamic properties in general [426]. These approaches have been investigated as a faster way to estimate different properties of interest with acceptable accuracies [427]. Based on their nonlinearity, these approaches create new QSPR/QSAR (Quantitative structure–property relationship/Quantitative structure–activity relationship) models with the above-mentioned merits [428]. Examples of the ML-based models used in predicting *Hansen* solubility parameters and their space are neural networks [429][430], combined Gaussian processes & Bayesian ML approach [431], random forest [432] and support vector machine/regression [433]. It is worth noting that the support vector machine was also used for the prediction of *Hildebrand* solubility parameter [434]. The selection of an appropriate solvent is critical for applications involving polymers, where identifying suitable solvents and non-solvents is crucial. This necessity has driven the development of quantitative models of polymer-solvent compatibility based on the principle that "like dissolves like." Furthermore, Gaussian Process Regression (GPR) has been employed to predict both the Hildebrand and Hansen solubility parameters for various polymers. These predictions facilitate the classification of materials as solvents or non-solvents for specific polymers [416]. The accuracy of the models used in most of these studies ranged from around 50% to above 90%. As observed, some cases possess low prediction accuracy, which can affect the calculations done by experts in

future applications such as solvent-polymer compatibility determination. In addition, in some cases the dataset utilized was relatively small, i.e. below 100 points, which affects model/method generalization. All these limitations/drawbacks along with some suggested solutions will be discussed in detail in the upcoming paragraphs.

To facilitate the processing of different chemical molecules and their related molecular properties historical data used for training, a new form for molecular structure representation, i.e. SMILES codes, is introduced and increasingly used as an open standard in chemistry for formula representation since 2007 [435]. The SMILES expression stands for Simplified Molecular Input Line Entry System [436]. These codes introduce a simplified way for representing chemical formulas and structures easy to be used by computers that can be then converted to ML-readable descriptors/features [437].

The utilization of data-driven computational models presents a significant advantage in the prediction of solvent and material properties from their SMILES representations, alongside pertinent geometric and molecular descriptors [277][438]. This approach notably reduces the duration traditionally required for experimental determination through labor-intensive laboratory procedures. Moreover, these models can be trained using historical datasets obtained from alternative computational methodologies, thereby serving as expedient surrogate models that can potentially supplant conventional techniques [430]. Ensemble or combination of different techniques/methods in parallel and sequential ways can be a novel approach to overcome the problem of low accuracies of single models. This is the core of the proposed methodology that will be discussed in detail in the subsequent sections. Besides, the black box-based calculations by some techniques represent a significant obstacle to understanding the impact of each feature/descriptor on the output variables of interest, i.e., *Hansen* solubility parameters. Fortunately, automatic features selection of tree-based techniques including the ensemble ones can help in solving this problem. In addition, the employment of explainable AI (XAI) techniques will give a complete description of these impacts qualitatively and quantitively [439]. Hence, it is recommended to include these tree-based models in the new ensemble method followed by the employment of XAI.

The following points summarize the challenges, gaps, and needs that prompted this study and the development of the proposed methodology:

- Existing machine learning techniques exhibit suboptimal prediction accuracies.

- Current feature selection methods are often inefficient, relying on manual selection as observed in some literature.
- There is a prevalent issue of black-box modeling, leading to a lack of interpretability and explanation.
- There is an imperative for a precise tool to predict Hansen solubility parameters for various materials, including solvents, which is crucial for optimizing industrial processes such as lignin valorization, carbon capture, and polymer manufacturing.

This work aims to propose a novel framework for data preparation, feature selection, and decision fusion, enhancing prediction accuracy. Furthermore, we extend the application of various AI and ML techniques to achieve precise predictions of *Hansen solubility* parameters, serving as a practical case study. The scientific and technical contributions of this study are summarized as follows:

1. Investigate the use of bi-clustering as a powerful data mining approach (inspired by the successful work done in the field of biology and bioinformatics) to perform robust features selection and use advanced non-linear clustering techniques to optimize the parameters of bi-clustering graphically.
2. Develop a decision fusion method to enhance the prediction accuracy based on gathering different machine learning models, each having its own advantages, through different mechanisms.
3. Validate the developed methodology to perform highly accurate prediction of Hansen solubility parameters using SMILES codes as inputs. This will help in the accurate definition of the *Hansen* sphere to select the best solvents to dissolve a certain material, e.g. diverse lignins and carbon dioxide (CO_2).
4. Interpret the black-box models by incorporating XAI techniques to extract knowledge and actionable insights that help the users, e.g., material and process designers accelerate the solvent selection and design optimization procedures afterwards.

The proposed AI-powered tool will help discover new solvents for the materials of interest by exploring the predicted *Hansen* solubility parameters. In addition, it is the first step towards the acceleration of AI-assisted solvents material design, where there is no unique chemical structure (e.g., lignin chemical structure is feedstock and separation process dependent) or the new non-

existing solvents will be generated/discovered and designed based on the desired *Hansen* solubility parameters as desired input properties. It can also be used for the identification of novel lignin-based products and potential derivatives.

4.3 Methodology

The general approach to this work is summarized in Figure 4.1. As depicted in this figure, the raw data is first collected and preprocessed to ensure its reliability and make it ready to perform the analysis. After that, the cleaned/preprocessed data is partitioned to training and validation sets based on the different case studies and human expertise. The training dataset is then utilized to train an ensemble of different ML techniques. The results are then fused to enhance the overall prediction accuracy. As is clearly stated in the figure, all the work is done under the supervision of human experts, where their expertise is applied in each step. This will ensure rationality and maximum possible modeling efficiency based on the data-human expertise-AI/simulation combined approach [156], [440]. In return, the results after fusion give new insights helping in increasing knowledge/expertise to tackle process improvements. The work should be interdisciplinary where the expertise in different fields will be exchanged and augmented to achieve high prediction accuracies.

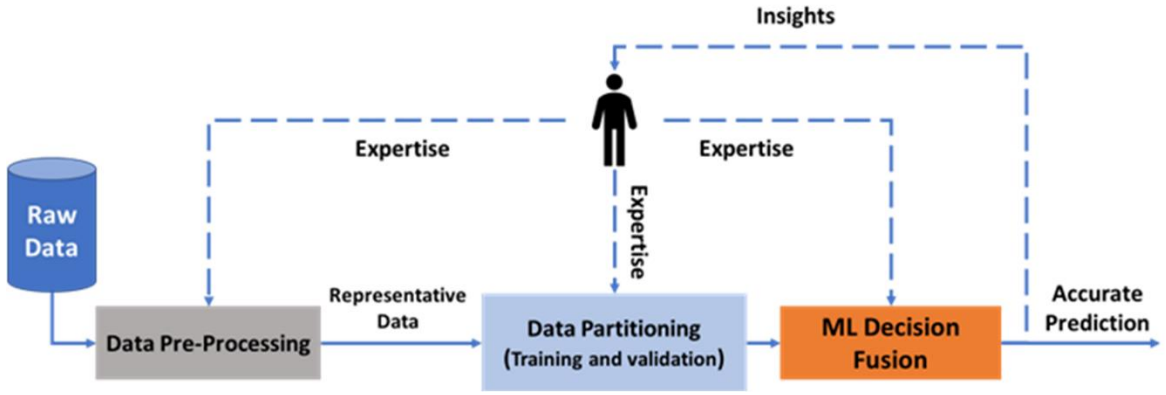


Figure 4.1 General approach.

4.3.1 Data cleaning, preprocessing and validation.

It is well known that data is the foundation of AI modeling. Thus, careful preprocessing and preparation of the data is essential to ensure high prediction accuracy. The raw data can be in various forms such as texts/strings, tabular data, images, etc. As illustrated in Figure 4.2, the first

step is to clean the raw data. This can be done by removing any empty, missing, incomplete or erroneous values. Additionally, different validation techniques should be applied to further clean the data. Efficient computational tools should be utilized to test the different forms of available raw data to remove any invalid points, e.g., incorrect texts, invalid symbols, and any incorrect mix between numerical & categorical/alphabetical values. This step is key to ensure high accuracy and avoid confusion. After invalid data elimination, the valid ones are gathered in a new refined dataset ready for further processing. The cleaned data is then further processed to extract the most representative and informative features to be fed to the analysis and ML modeling. These features can be in the form of various descriptors, eigenvalues, eigenvectors, and fingerprints. Afterwards, they are gathered and combined with other variables to start assigning the inputs and outputs. Commonly, the features and data points in the available datasets are in large number, which can be computationally expensive and time consuming. Besides, some of the features in the raw data can be confusing and misleading thereby hindering the prediction accuracy upon ML modeling. Therefore, a dimensionality reduction step is essential at this stage for proper features selection. In this regard, employing biclustering is proposed as a powerful data mining and dimensionality reduction approach to perform this task. However, it is key to determine the optimum number of components where the data will be reduced to. This step can be time-consuming and computationally expensive, especially with larger datasets. Hence, the utilization of other clustering techniques for data visualization, in the initial clustering step, is suggested to determine the optimum number of components (latent dimensions) graphically. The models used for clustering is further discussed in **Section 4.3.3**.

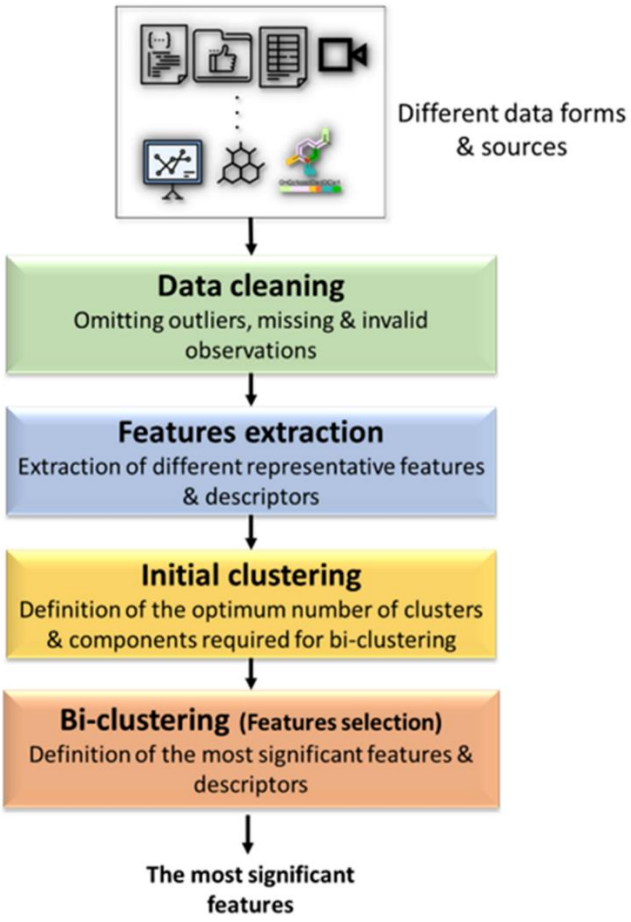


Figure 4.2 Data preprocessing methodology.

4.3.2 Biclustering

Biclustering is a useful technique for dimensionality reduction and data mining that gathers both rows and columns of the input data simultaneously by identifying other submatrices with rational similarity patterns [441], [442]. Within the context of ML, biclustering can be utilized for feature selection where the goal is to identify subsets of the most relevant features depending on the case study or the task [443]–[445]. There are several methods to perform bi-clustering [446] such as spectral bi-clustering [447], *Plaid* models [448], *Bayesian* bi-clustering [449], [450], sparse singular value decomposition (SSVD) [451] and non-negative matrix factorization (NNMF) [452], [453]. In the proposed methodology, the use of NNMF is recommended due to its easy implementation, high accuracy, clear visualization of results, and its successful application in the biomedical/bioinformatics fields. Figure 4.3 presents a simple schematic illustrating the general concept behind NNMF biclustering (adapted from [454]).

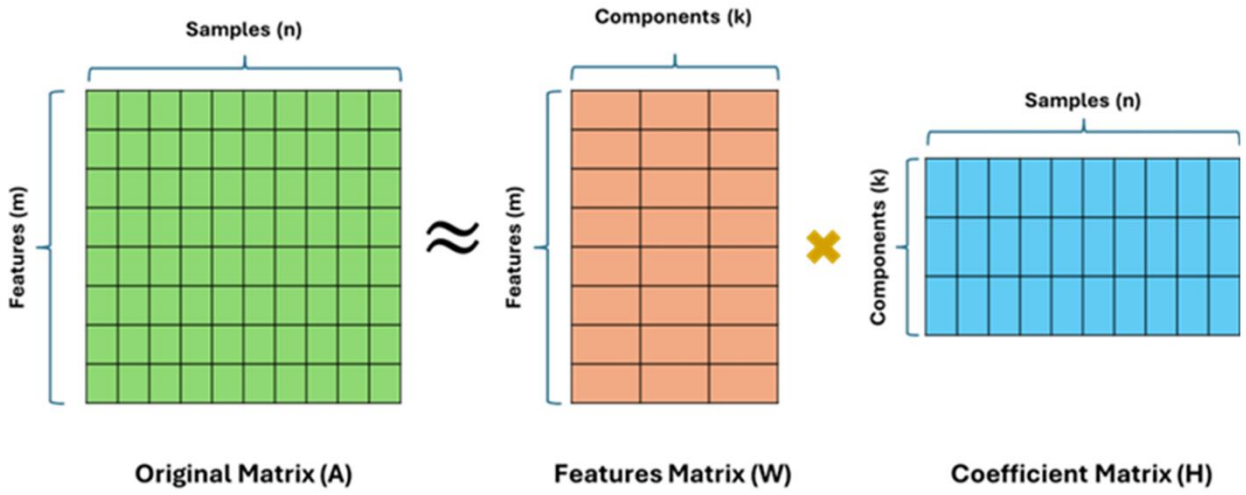


Figure 4.3 NNMF concept simple schematic

In the presented schematic, which is the most common one, the original data matrix (A) is decomposed into two new matrices. The first one is the feature matrix (W) and the second is the coefficient matrix (H), where m, n and k are the number of features, points/instances, and components of reduced data. For the mathematical background, types and algorithms of this method, readers can refer to [455]–[457]. The different methods of NNMF performance evaluation are discussed in **Section 2.5**. It is important to note that NNMF can be computationally expensive and complex, especially when optimizing its key hyperparameter, the number of components, and when working with very large datasets. The complexity of NNMF is primarily influenced by the dimensions of the original matrix ($m \times n$) and the number of components (k). Therefore, performing data clustering to reduce the problem's dimensionality before biclustering can offer a promising solution. Besides, in this study, we propose a promising and effective solution using clustering techniques to determine the optimal number of components, thereby improving the scalability of NNMF for larger datasets.

4.3.3 Clustering

In the proposed methodology, uniform manifold approximation and projection (UMAP) [458], [459] and t-distributed stochastic neighbor embedding (t-SNE) [460], [461] combined with k-means algorithm [462] were chosen as efficient dimensionality reduction, clustering and data visualization techniques. There are many other dimensionality reduction techniques being widely used in literature including principal component analysis (PCA) [463], [464]. However, these two

techniques, i.e., UMAP and t-SNE, were selected because of their ease of handling non-linearity, preservation of data structure and flexible visualization [465], [466]. For instance, one of the limitations of PCA is its linearity where, in contrast, UMAP and t-SNE are nonlinear techniques [467]. In addition, UMAP can preserve the local and global data structures [468]. They are also flexible and have better visualization. Moreover, UMAP is relatively computationally inexpensive where its processing time is relatively short and can handle large datasets. For more information about the mathematical background and assumptions considered during the implementation of UMAP and t-SNE, readers can refer to [458], [467], [469]. The different methods of UMAP and t-SNE performances evaluation are discussed in **Section 4.3.5**.

4.3.4 ML modeling and Decision Fusion

The ML modeling and decision fusion methodology is illustrated in Figure 4.4. As shown, after preparing the data, these data containing the combinations of the most representative features is fed to the selected ML algorithms. Before modeling, the data is divided into two parts, i.e., one used for training the models whereas the second part is used for validation. After modeling, the results of each optimized model are fused through different ways, i.e., learnable, and non-learnable fusion. The upcoming sections elaborate more on the data partitioning, ML models considered, and the different fusion methods proposed.

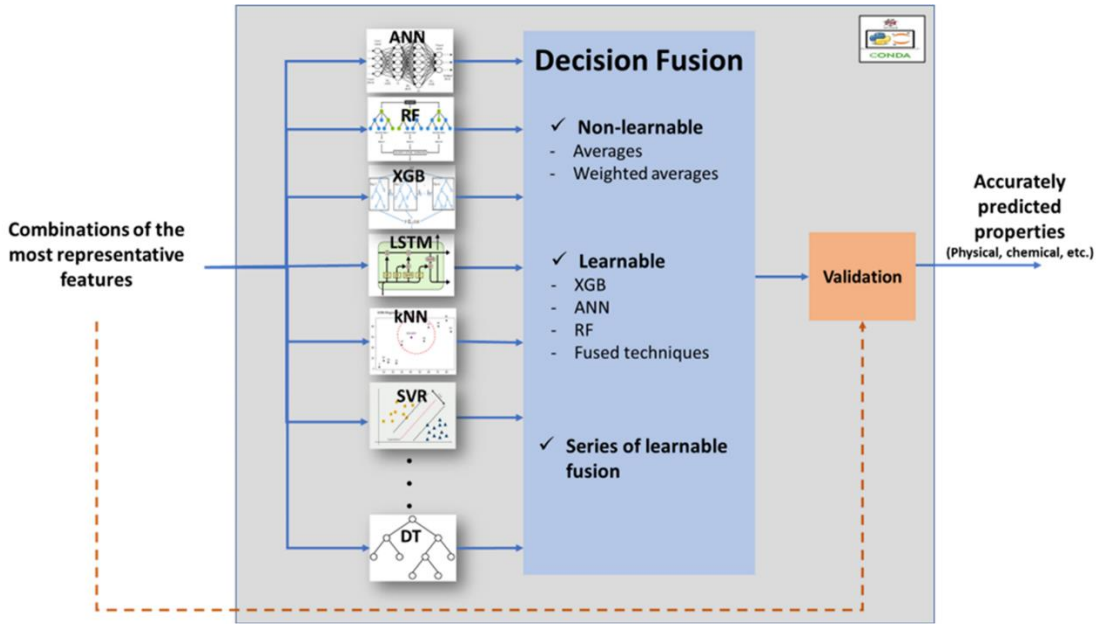


Figure 4.4 Decision fusion methodology

Data partitioning for model training

As previously mentioned, the analysis-ready data will be divided into two sets, one for training and the other one for validation. There are several factors that control the choice of the percentages of both training and validation subsets. These factors include but are not limited to the available dataset size, which determines the amount of information available for training, the complexity of the models to be developed and the models' hyperparameters and their number. Yet, based on the knowledge of the human experts, there are common values of these percentages typically employed for several machine learning modeling case studies. The related common values are as follows:

- **70 to 30 split ratio:** this case considers 70% of the input data to train the developed model, and the remaining 30% is used for model validation. This ratio is usually employed for relatively large datasets and relatively simple models.
- **80 to 20 split ratio:** This is a common split ratio in the case of reasonably large datasets and significantly complex models. Where 80% of the data is dedicated for models training and the other 20% is devoted to validating the developed model.
- **90 to 10 split ratio:** when the input dataset is small and the model is simple, 90% of the input data is used for model training and only 10% of it is reserved for model validation.

In addition, validation datasets are usually employed to facilitate hyperparameters tuning. Thus, based on the model complexity and the number of their hyperparameters to be tuned, the validation dataset size will vary significantly. Moreover, an additional set may be utilized in specific cases to evaluate/test the models' performance after their training and hyperparameters tuning. It is worth noting that the splitting of the data is done randomly to ensure generalization and avoid any bias during the training and validation.

ML techniques under consideration

Several ML techniques (shallow and deep learning) are proposed for being employed at the first stages of modeling; however, the most relevant ones are selected based on the determination coefficient (R^2) and mean squared error (MSE). These techniques are the random forest (RF), artificial neural network (ANN), support vector regression (SVR), optimized decision tree (DT), *k*-nearest neighbour (*k*-NN) regression, extreme gradient boost (XGB), long short-term memory (LSTM), Gaussian process regression (GPR) and convolutional neural network (CNN). As can be deduced, this is a wide spectrum of techniques where shallow, deep learning, non-parametric and

ensemble methods are employed. Each one of these techniques has its own merits and that is why the utilization of all these different techniques is proposed. It is worth mentioning that these models were selected based on an extensive literature review and the widely recognized best practices in the AI/ML community [156], [343], [357], [362], [470]. Table 4.1 summarizes the key merits of each individual model, which justify their selection and motivate the development of the *Ensemble-of-Ensembles* technique in this study. For an in-depth explanation and the mathematical foundation of each technique, readers are encouraged to consult the references listed in Table 4.1.

Table 4.1 Overview of the different ML techniques and their key merits.

*This table was moved to Appendix C (Table C.1) in (APPENDIX)

Hyperparameter optimization of ML techniques

Each technique (except the non-parametric *k*-NN lazy ML model) has hyperparameters that should be optimized to obtain the best possible prediction accuracy. In the current study, the utilization of *Grid Search* and *Bayesian Optimization* techniques are suggested to perform this hyperparameter optimization. Table 4.2 summarizes the different hyperparameters of each technique along with their corresponding proposed ranges. It is worth noting that the determination of optimum *k*, i.e., number of neighbours, in the case of *k*-NN is also essential. Hence, according to literature, the selected range for *k* is from 2 to 7 neighbours to avoid high bias as well as overfitting.

Table 4.2 Hyperparameters and their corresponding ranges

Model	**Hyperparameters**	**Ranges**
DT	Depth	8-64
RF	Number of estimators (trees)	10-300
XGB	Number of estimators (trees)	10-300
SVR	C penalty of misclassification Gamma decision boundary range Kernel function	1-500 *Scale, Float & Auto* *RBF, Sigmoid & Poly*
ANN	Learning rate Batch size Number of hidden layers Number of neurons per hidden layer Activation function	0.0005-0.1 1-128 1-5 20-200 *Tanh* & *ReLU*
LSTM	Number of LSTM layers Number of LSTM units in each layer Number of units in dense layers Batch size Number of epochs	1-3 10-100 1-3 8-64 20-200
CNN	Number of convolutional layers Number of filters per convolutional layers Number of dense layers Number of units per dense layers Kernel size Activation function Pooling size	1-3 16-256 1-3 16-256 1-5 *Tanh* & *ReLU* 1-5
GPR	Kernels	*RBF, Constant & White noise*

Decision Fusion techniques

The techniques proposed for fusion in this study are divided into two main categories. The first one is the non-learnable fusion where the decision (the final prediction) is taken based on the average and weighted average of all the results of the ML techniques. In the average-based ensemble modeling, the final prediction output is the average of the output obtained from single models where all the values of each model have the same weight. In the case of weighted average, the weights are put based on the determination coefficient, mean squared error and a combination of both. The mathematical representations of the final output predictions based on the proposed average weights are introduced by the following formulas:

$$P_{if} = \frac{\sum P_{ij} \times R_j^2}{\sum R_j^2} \tag{4.1}$$

$$P_{if} = \frac{\sum P_{ij} \times \frac{1}{MSE_j}}{\sum \frac{1}{MSE_j}} \tag{4.2}$$

$$P_{if} = \frac{\sum P_{ij} \times \frac{R_j^2}{MSE_j}}{\sum \frac{R_j^2}{MSE_j}} \tag{4.3}$$

Where P_{if} is the final predicted value of instance/observation (i) based on weighted average of models (j) results, P_{ij} is the predicted value of instance (i) based on model (j), R^2_j is the determination coefficient of model (j) and MSE_j is the mean squared error of model (j) *for all i, j*.

The second fusion method is the learnable fusion where all the data points obtained from training and validating the ML techniques are gathered and used to train some other ML techniques inspired by staking-based ensemble modeling. In this case the inputs are the predicted values obtained from various single models, average-based ensemble and weighted average-based ensemble models whereas the outputs are the actual values (experimental or computational raw data of variable of interest). In addition, the techniques used are ANN, XGB, RF and their fusion. A new variant of the learnable fusion is the series of learnable fusions to maximize the accuracy of prediction as much as possible. This is done through continuous learning of the decision fusion results on multiple stages. In particular, the results from the first learnable fusion step are fed to another learnable fusion agent/model which in turn gives new results that are introduced to another fusion

stage and so on. In the methodology in the current study, it is recommended to perform this series on three stages. Figure 4.5 shows a simple schematic of the proposed series of learnable decision fusions.

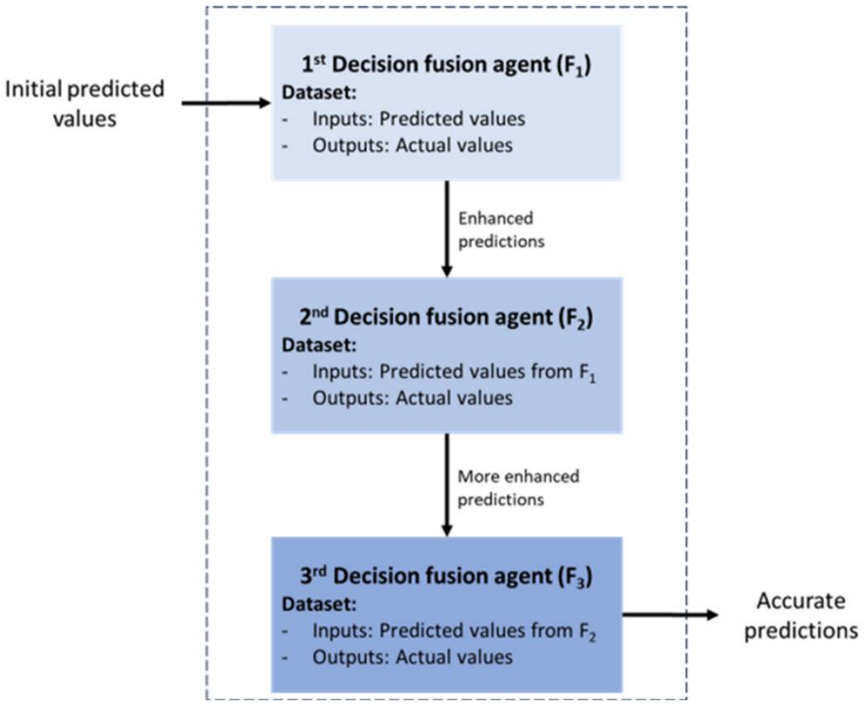


Figure 4.5 Series of learnable decision fusions.

It is worth noting that non-learnable fusion techniques are proposed and implemented as an intermediate stage, with their results used as inputs for the learnable fusion technique. The objective is to develop an ensemble-based approach that combines simple statistical methods (i.e. non-learnable techniques) with those based on ML concepts comprising shallow and deep learning. This multi-stage fusion of non-learnable and learnable techniques helps reduce error and improve prediction accuracy. However, we intentionally avoided relying heavily on other statistical techniques to prevent introducing pre-assumptions about the data. Our learnable fusion techniques are based on both shallow ML and deep learning concepts, aiming to capture the actual distribution without the need for approximations.

4.3.5 Results evaluation and validation

In the case of clustering, silhouette score and the silhouette plot analysis method are employed to enable the selection of optimum components and clusters numbers. Whereas, in the case of bi-

clustering other additional metrics, i.e., variance (EV) and reconstruction error (RE), are evaluated. The explained variance can be referred to as the determination coefficient, but it is multiplied by the total variance. In addition, the results of machine learning modeling are evaluated mainly based on the model accuracy represented by determination coefficient (R^2) and mean squared error (MSE). The mathematical representations of these indicators are presented with the following equations:

$$EV = R^2 \times Total\ Varriance = \frac{SS_{exp.}}{n} \tag{4.4}$$

$$RE = \frac{1}{n} \times \sum_{i=1}^{n} \|x_i - \widehat{x_i}\|^2 \tag{4.5}$$

$$R^2 = \frac{\sum_{i=1}^{n}(y_{i_{Predicted}} - \bar{y})^2}{\sum_{i=1}^{n}(y_{i_{Actual}} - \bar{y})^2} = \frac{SS_{exp.}}{SS_{tot.}} \tag{4.6}$$

$$MSE = \frac{1}{n} \times \sum (y_{i_{Actual}} - y_{i_{Predected}})^2 \tag{4.7}$$

$SS_{exp.}$ is the explained sum of squares, n is the number of observations, $SS_{tot.}$ is the total sum of squares, x_i is the i-th original data point and $\widehat{x_i}$ is the i-th reconstructed data point.

4.3.6 Models explainability

To avoid black-box modeling, an explainable AI (XAI) method was employed with the optimized ML models. This method is the calculation of SHAP values. By employing this method, the developed ensemble-based regression model can be considered a white box as it gives the weights related to the different input variables that define their relative impacts on the output variables. The weights are given in the form of relative magnitudes with their corresponding signs, i.e., positive, or negative, to totally define the impacts. Readers interested in more details about this method and its other possible applications can refer to this recently published study [471].

4.4 Case Study: Hansen solubility parameters estimation

In the present study, we focus on estimating the Hansen solubility parameters through a proposed case study. Initially, we introduce the data sources and the nature of the variables. Subsequently, we provide a detailed description of the preprocessing methodology.

4.4.1 Data source and type

The dataset in the current study is gathered from the literature and *HSPiP* software [472]–[475]. The data essentially included solvents names, their SMILES codes, and thermodynamic properties. The properties covered are Antoine parameters, boiling points, melting points, flash points, vapor pressures, water solubility and Hansen solubility parameters. The variables initially selected for the HSP prediction are the SMILES codes and the three *Hansen* solubility parameters. Table 4.3 illustrates an excerpt of the data and variables used for ML modeling along with the corresponding compound's name of each SMILES code. SMILES codes were selected as the inputs instead of only the solvents names as they can be easily used to extract more features to well define the solvents. Features extraction will be discussed in more detail in the upcoming section. As can be observed, the data in the current state is heterogeneous, i.e., SMILES are alphabetical/categorical and hence the features extraction step is essential to convert SMILES code to numerical representations that are homogeneous with the HSP and easy to process for the prediction purpose.

HSP prediction/calculation for diverse materials and solvents is an essential step for different applications. For example, lignins fractionation and blinding with diverse polymers at high compatibility depend on the accurate calculation of their HSP [370]. However, as discussed in [370], the common methods for identifying suitable solvents for different lignins are based on arbitrary selection of solvents and performing empirical laboratory procedures to assess their performance [476]. Additionally, compatibility with polymers is assessed through an *Ad hoc* way that requires prior expertise. Therefore, this study [370] suggested lignin' HSP calculation as a robust and systematic way to overcome the drawbacks of the other common methods. On the other hand, the traditional HSP calculation methods including experimentation and computational work suffer from being resource intensive and time consuming. Hence, the fast and easy predictions of these important parameters is crucial to accelerate the prediction of lignin computability with diverse polymers and technological adoption on large scale.

The compounds in this dataset represent a broad selection of known solvents used in several industrial applications, including lignin valorization and carbon capture. The dataset was curated from the literature and *HSPiP* software, incorporating a wide range of chemical classes and solvent types to ensure diversity and relevance to the target applications of this study, particularly solvent compatibility assessments, which are crucial in lignin valorization and carbon capture. As shown

in the simple data analysis performed on this constructed dataset, it includes heterocyclic compounds, amines, alcohols, and other chemical classes representing both polar and non-polar solvents, essential for capturing the full spectrum of solubility interactions. Furthermore, compounds were selected based on the availability and completeness of data, ensuring that each entry included corresponding dispersion, polarity, and hydrogen bonding solubility parameters, as well as a valid SMILES code.

Table 4.3 Dataset excerpt - Solvents with their SMILES codes and Hansen Solubility Parameters

*This table was moved to Appendix C (Table C.2) in (APPENDIX)

4.4.2 SMILES-HSP data preprocessing

Preprocessing and preparation of the SMILES-HSP data were done to ensure a high prediction accuracy. In this study, the raw molecular representations used are the SMILES codes of different solvents in the dataset. Initially, the SMILES codes were cleaned by removing any empty or missing values. They were then validated using *RDKit*, a computational tool commonly used by experts, to filter out invalid codes from the raw dataset [477]. *RDKit* was utilized through *Python* programing language for this essential task. After eliminating the invalid codes, the remaining valid ones were compiled into a refined dataset for further processing. Additionally, *RDKit* tool was used to extract more representative features, including molecular descriptors and Morgan fingerprints. These descriptors and fingerprints were then combined with the Hansen solubility parameters into a single dataset, where the features serve as the input values and the parameters as the output values. It was revealed that there are too many descriptors and using all of them or the ones having low significance can hinder the accuracy of the modeling afterwards. Accordingly, a features selection step is essential. Where, the biclustering data mining approach, particularly the non-negative matrix factorization algorithm (NNMF), was used to select the most significant descriptors. The NNMF requires determination of the number of components as an essential hyperparameter. Thus, other clustering techniques, namely UMAP and tSNE, were employed to help in choosing the best number of components through data visualization. After features selection, the decoded SMILES (most significant descriptors and fingerprints) along with the Hansen solubility parameters were randomly divided into two sets, one for training and the other one for validation where their

percentages are 85% and 15%, respectively. These percentages were selected upon trial and error until they reached satisfactory accuracy in both cases of training and testing. They are between the 80/20 and 90/10 common split ratios as the proposed ensemble method contains simple and complex models at the same time where the dataset is relatively large. 85% of the data was also used for the training to give the chance to the models to capture the true distribution of the available data.

4.4.3 ML modeling and decision fusion

The results of the optimized ensemble-based technique were gathered, evaluated, and benchmarked for the seven individual ML techniques selected and employed to predict HSP based on the key features extracted from the raw SMILES codes. In addition, all the fusion techniques were applied and compared to each other as well. Details of ML modeling and building the ensemble-based model are presented in **Section 4.3.4**.

4.4.4 Models explainability in selected cases

The SHAP values as an informative XAI item were calculated for the developed ML ensemble-based regression model for decision fusion that calculates the different solubility parameters. The impacts of each descriptor on each solubility parameter were calculated in detail. In this case, the problem is a regression one where both input and output variables of interest are continuous. In addition, three other distinct cases were selected to show the importance of adding an XAI part to the developed high prediction accuracy ensemble-based model for decision fusion. These cases of materials that need to be dissolved are softwood-based kraft lignin, sugar cane bagasse-based lignin and CO_2. For validation purposes and to demonstrate the generalizability of the developed white-box methodology across various materials and systems, CO_2 was considered alongside different lignins. As a crucial material, selecting the appropriate solvents for CO_2 can significantly accelerate climate change mitigation efforts. After the calculation of each relative energy difference (RED) for each case, the solvents were classified as good and bad solvents taking the categorical labels of 1 and 0, respectively. The classification was based on the *Hansen* sphere concept; where, solvents scoring an RED equal to or higher than 1 are categorized as bad solvents, and the solvents with REDs lower than 1 are categorized as good solvents. The solubility parameters of the materials related to the three cases are summarized in Table 4.4. RED is calculated according to Equation 4.8 [478].

$$RED = {}^{R_a}/_{R_o} \tag{4.8}$$

Where R_o is the experimental sphere radius of the material to be dissolved (m) and R_a is the radius of interaction between solvents and the material to be dissolved. Ra is calculated according to Equation 4.9 [423].

$$R_a = \sqrt{4 \times \left(\delta_D^m - \delta_D^{solvent}\right)^2 + \left(\delta_P^m - \delta_P^{solvent}\right)^2 + \left(\delta_H^m - \delta_H^{solvent}\right)^2} \tag{4.9}$$

Table 4.4 Hansen solubility parameters of selected materials

Material (Solute)	**δ_D**	**δ_P**	**δ_H**	**R_o**	**Reference**
Softwood-based kraft lignin	20.93	16.98	15.71	15	[370]
Sugar Cane bagasse-based lignin	21.42	8.57	21.80	13.56	[479]
CO_2	15.6	5.2	5.8	4	[480][481]

It is worth noting that these two lignins have a wide range of applications. Our study focuses on the solvents' properties that make them suitable for dissolving these key materials, facilitating their efficient valorization in the future. Below are some specific applications that can be covered by these lignins:

- **Compatibility and blending with polymers:** Precise prediction of *Hansen* solubility parameters (HSP) enables the accurate selection of solvents for blending lignin with polymers, which can then be used to create composite materials.
- **Lignin valorization through fractionation:** Understanding the solubility behavior and the key solvent properties aids in the efficient fractionation of lignin into its simpler components/units. This will facilitate their utilization as precursors to other value-added products, including carbon fibers and bio-based resins as well as other intermediates or specialty chemicals.
- **Carbon capture and other environmental applications:** Solubility studies of sugarcane bagasse lignin (and even softwood-based lignin), can support its subsequent utilization in the synthesis of bio-based sorbents for carbon capture purposes, aligning with global Net-

Zero emission goals. Identifying the most efficient solvents allows for the extraction of high purity lignins, which can be used to produce activated carbon-based adsorbents with tailored and desired capture capacities and selectivities [482], [483]. These lignins can also be utilized in other applications, such as the development of effective adsorbents for water and wastewater treatment [484]–[487] or as catalysts in biofuels production [488]–[490].

4.4.5 Further case to validate model generalizability

To ensure that we built a generalizable ensemble model that avoids overfitting, it was further validated using a separate dataset. This new dataset contains over 1200 compounds with different distribution than the big dataset used for training. It includes experimental HSP data of these solvents, collected from various sources in the literature [491]–[493]. It is worth noting that cross validation was also considered during the training phase to overcome any kind of overfitting besides the dependence on models that avoid this problem from the beginning such as the Random Forest and XGB. From 5 to 10 *k*-fold cross validation were considered and tested in the current study. The *k*-fold (*k=5*) was chosen and employed to avoid high consumption of time and computational resources.

4.5 Results and discussion

This section presents the results obtained by applying the developed methodology on the HSP-SMILES dataset to achieve high quality predictions of the different solubility parameters. It gives an overview about the distribution of the types of the solvents in this dataset where the most abundant solvents in the available repository are the heterocyclic ones (>3500). It also presents the results of clustering techniques and biclustering through NNMF to identify the key descriptors where they were reduced from 208 to 70. Afterwards, the results of ML modeling, through individual and ensemble models/techniques, are introduced and compared. Based on the developed ensemble modeling, the average prediction accuracy reached 99%.

4.5.1 Data analysis of HSP-SMILES dataset

Firstly, *RDKit* was used to validate the SMILES codes in the available dataset to clean data. As a result, about 170 invalid codes/data points were removed out of about 12000 due to different errors, including syntax errors and other technical ones such as the atom bond kekulization [494]. After

obtaining the valid SMILES codes dataset, simple descriptive statistical analysis was done to determine the distribution of the available molecules of solvents/compounds. Figure 4.6 presents the distribution of the different types of solvents included in the available data based on their SMILES codes. As can be observed, the most frequent type of solvent is the heterocyclic one with more than 3500 solvents. The second most common type is the amines followed by alcohol at the numbers of around 1800 and 1200 solvents, respectively.

Figure 4.7 depicts the distribution of data for each of the three Hansen solubility parameters within the parent dataset, illustrating the frequency intervals for each parameter. Additionally, Table 4.5 provides a straightforward statistical summary of each variable in the HSP dataset, including the maximum, minimum, and average values of each parameter. The data for the dispersion solubility parameter appears to follow a nearly normal distribution, whereas the polarization and hydrogen bonding parameters exhibit noticeable skewness to the left.

On the other hand, *RDKit* was used to extract different descriptors based on the available SMILES codes. 208 descriptors in total were extracted to represent the SMILES code of each compound. However, as a first step, 18 common descriptors were chosen besides the *Morgan* fingerprints to represent each compound. These descriptors are molecular weight (MolWt), topological polar surface area (TPSA), number of heavy atoms (HeavyAtomCount), heavy atom molecular weight (HAMWt), molecular logarithm of the partition coefficient (LogP), number of rotatable bonds (NRB), number of hydrogen donors (NHD), number of hydrogen acceptor (NHA), *Hall-Kier* alpha shape index (HKA), Chi squared n descriptor (Chi2n), *Labute's* approximate surface area (LASA), number of radical electrons (NRE), 2D autocorrelation descriptor capturing spatial patterns and relationships within the molecule (Autoc), number of valence electrons (NVE), fingerprint density descriptor based on the *Morgan* algorithm with radius of 2 (FDFDM2), exact molecular weight (ExMwt), maximum absolute partial charges on the atoms in the molecule (MaxAPC) and minimum absolute partial charges on the atoms in the molecule (MinAPC). As previously mentioned, the most significant variables/descriptors will be selected to represent each compound. Hence, the correlation between each descriptor based on the *Pearson's* method was performed and presented in Figure 4.8 to prove visually the need for reducing the number of descriptors to the most important and uncorrelated ones. This is a preliminary step before performing the biclustering for rigorous variables selection.

The numbers mentioned above correspond to different case studies or scenarios presented in the manuscript. In the first instance, the 18 components were selected as part of an initial example to demonstrate the effectiveness of NNMF and our combined feature selection approach in identifying impactful features or variables influencing the prediction of the desired outputs. These 18 variables, chosen randomly from a larger pool, were all highly relevant. However, to provide a more comprehensive evaluation, we extended the analysis to consider all available descriptors. In this extended example, the data was optimally grouped into three components, with each component representing a cluster of significant descriptors. Thus, for the initial case study, 18 components best described the data, while for the extended case study, three components offered a more balanced and efficient representation of the dataset. The final number of components selected for the study is three, as it provides an optimal trade-off between complexity and performance.

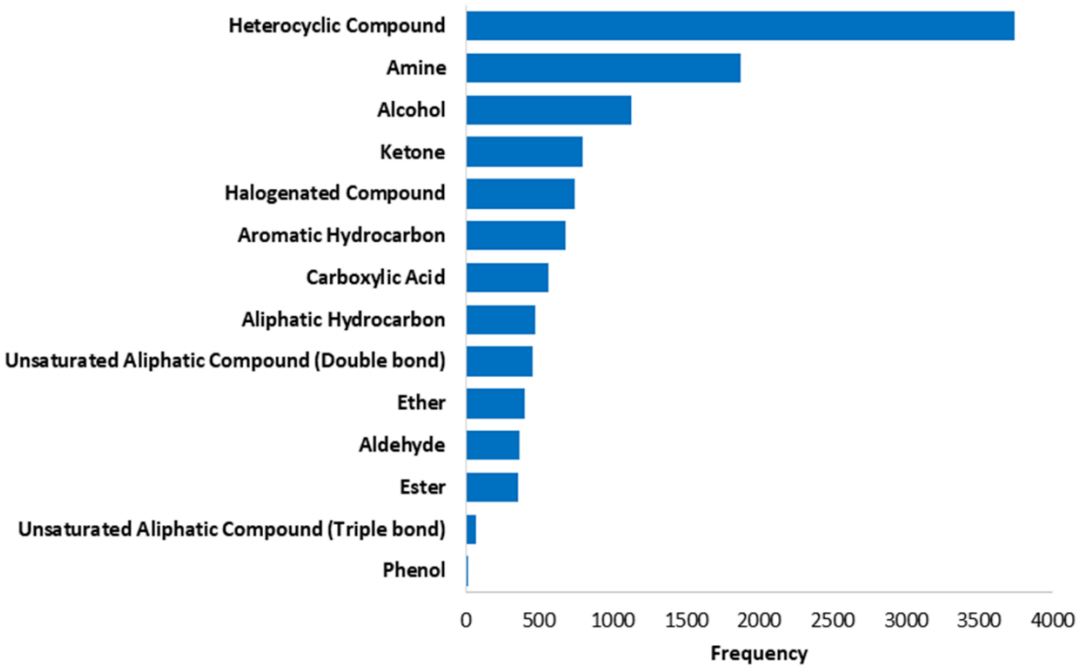


Figure 4.6 Solvent type distribution

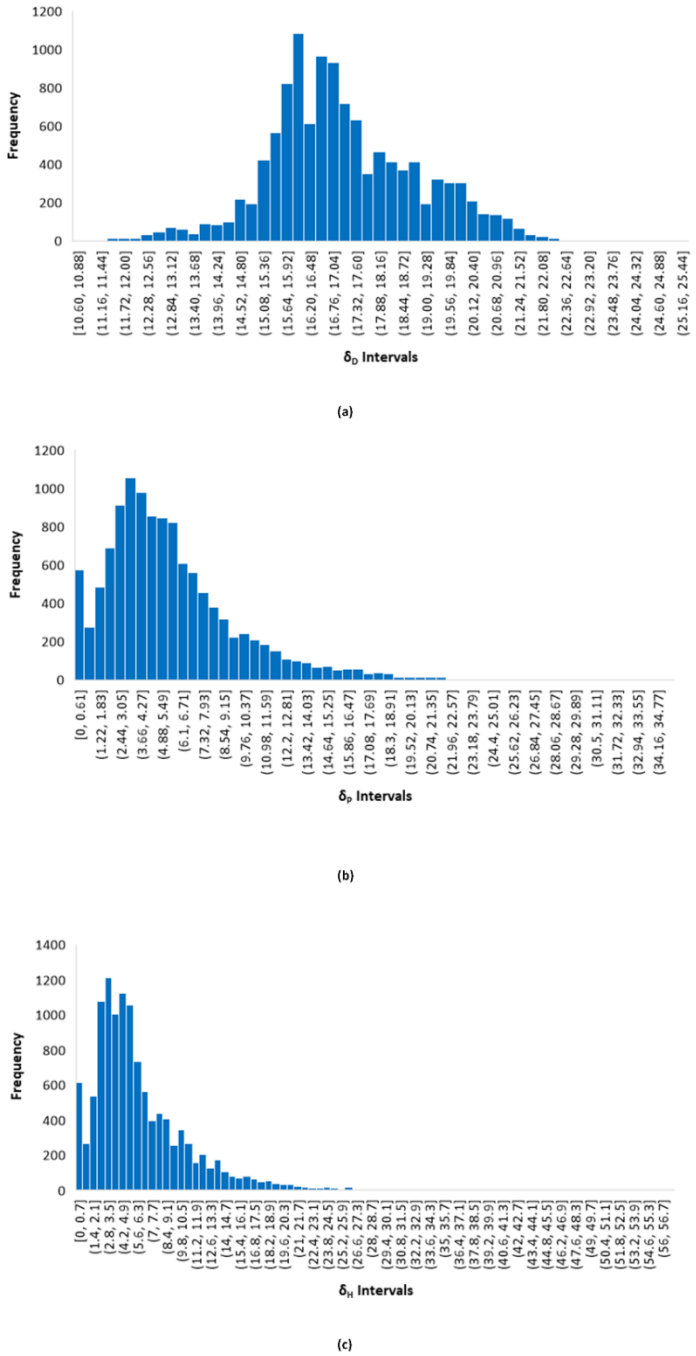


Figure 4.7 Hansen solubility parameters (a) dispersion, (b) polarization and (c) hydrogen bonding distributions

Table 4.5 Simple statistical description of *Hansen* solubility parameters data

Statistical parameter	Variables		
	δ_D [MPa$^{0.5}$]	δ_P [MPa$^{0.5}$]	δ_H [MPa$^{0.5}$]
Maximum	25.3	35.2	57
Minimum	10.6	0	0
Average	17.14	5.69	6.05
Median	16.9	4.8	4.8
Standard Deviation	1.80	3.98	4.54
Mode	16	0.1	0.1
Skewness	0.26	1.49	1.83
Kurtosis	0.20	3.68	6.18

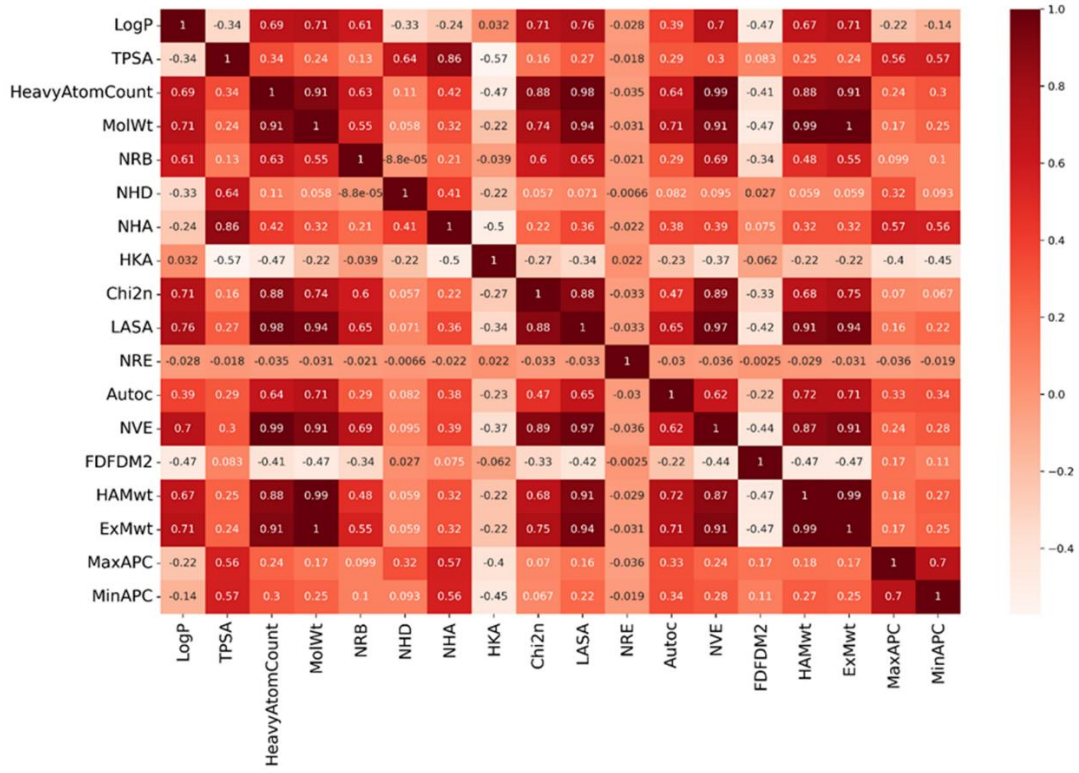


Figure 4.8 Descriptors correlation matrix

As it is shown in Figure 4.8, there are some highly correlated variables/descriptors such as the high positive correlation of heavy atom molecular weight (HAMwt) with both exact molecular weight (ExMwt) and molecular weight (MolWt). The quality of this correlation is very close to the unit, which means that only one descriptor can be used. This will lead to dimensionality reduction that facilitates computational work. To achieve this goal, a robust feature selection method should be employed and that is why the combined clustering/bi-clustering methodology is proposed to select the most significant descriptors.

4.5.2 Clustering results

The clustering methods were employed to give a preliminary recommendation about the best number of components graphically. This number is essential for performing the NNMF analysis to complete the features selection. Firstly, the best number of data clusters was determined for the solvents' available data according to silhouette test which determines the degree of data clusters separation. The clustering was performed based on *k*-means clustering method. In the current study, the optimal number of data clusters is 2 (Figure 4.9(a)) where the separation of the clusters is significant and visible as indicated by the plot on the right side of the figure. In this case, the silhouette score is 0.2. Dividing the data into more than 2 clusters results in lower silhouette score values which means having inseparable clusters. This is further confirmed by the plot on the right side of Figure 4.9(b). This result was consistent for both cases considered in this study: the first involving the randomly selected 18 common descriptors, and the second including all 208 physio-chemical descriptors.

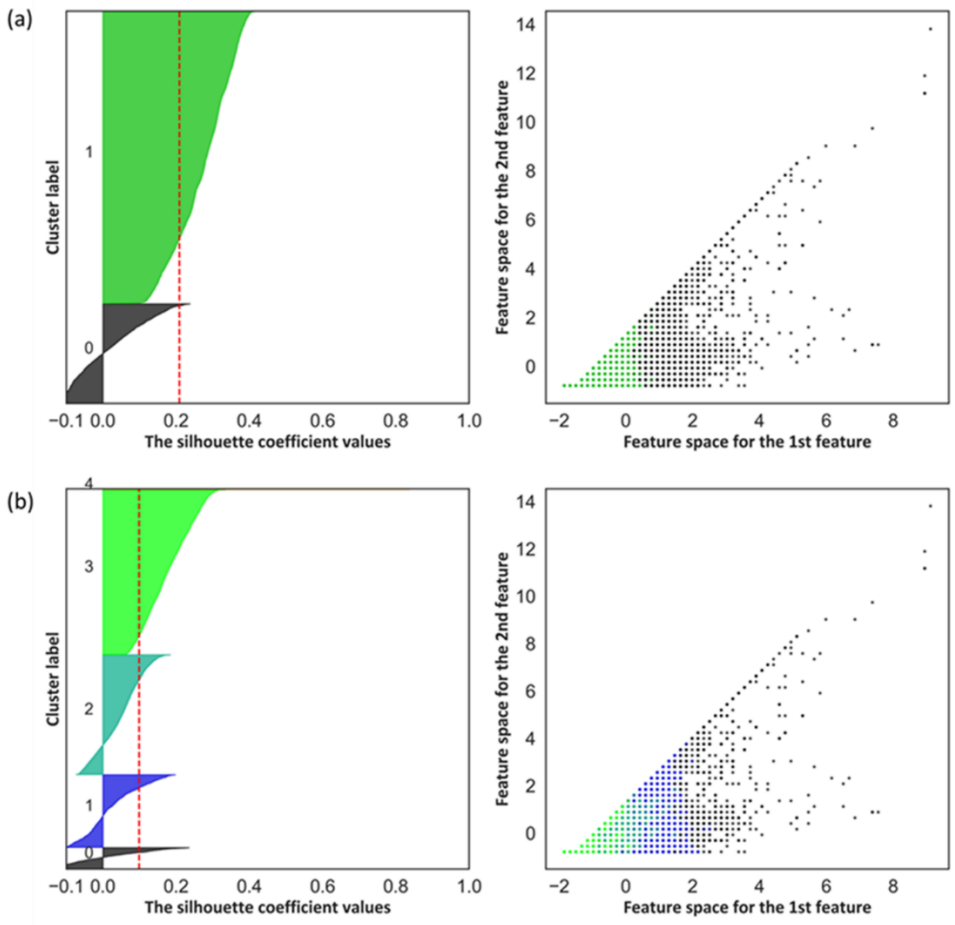


Figure 4.9 Silhouette analysis plots for (a) the optimum number of clusters ($n_{descriptors}$ = 208 & $n_{clusters}$=2) (b) example of inseparable clusters ($n_{descriptors}$ = 208 & $n_{clusters}$=5).

After choosing the best number of clusters, the work was oriented to the determination of the best number of components (the input variables of the data after performing the dimensionality reduction). Both UMAP and tSNE combined with *k*-means were employed for this task. It should be noted that in this study, the selected number of components is used to guide the NNMF for subsequent features selection. This approach helps mitigate the issue of interpretability, as the NNMF determines the actual number of important variables, and the variables used as inputs afterwards are the most significant original variables/features extracted from the SMILES codes. To further reduce the computational complexity of NNMF, both UMAP and tSNE were incorporated into our novel methodology. UMAP outperformed tSNE in terms of both the computational time and graphical representation. In particular, the UMAP completed the dimensionality reduction in only 5 seconds while tSNE completed it in 112 seconds.

In the first case study, where the 18 common descriptors were randomly chosen to represent the available data, the number of components that best represents the data is 18 according to silhouette score test and the graphical representations obtained by UMAP (Figure 4.10 (a & c)). The 3D graph (Figure 4.10 (c)) was added for more clarification of the obvious separation of the two clusters upon selecting the optimum number of components, i.e. 18. In the second case, upon extracting all the possible descriptors of each SMILES code using *RDKit*, the best number of components that the data can be reduced to was surprisingly found as 3. The graphical representation of 2D UMAP clustering in the case of extracting all the descriptors is provided by Figure 4.10 (b). This step of extracting all the descriptors led to having more representative data. In conclusion, based on this advisory step, the data can be clustered into two main clusters, and its variables can be reduced to three primary components. These results were then used as inputs to NNMF, enabling robust features/variables selection without sacrificing the interpretability or the integrity of the original data.

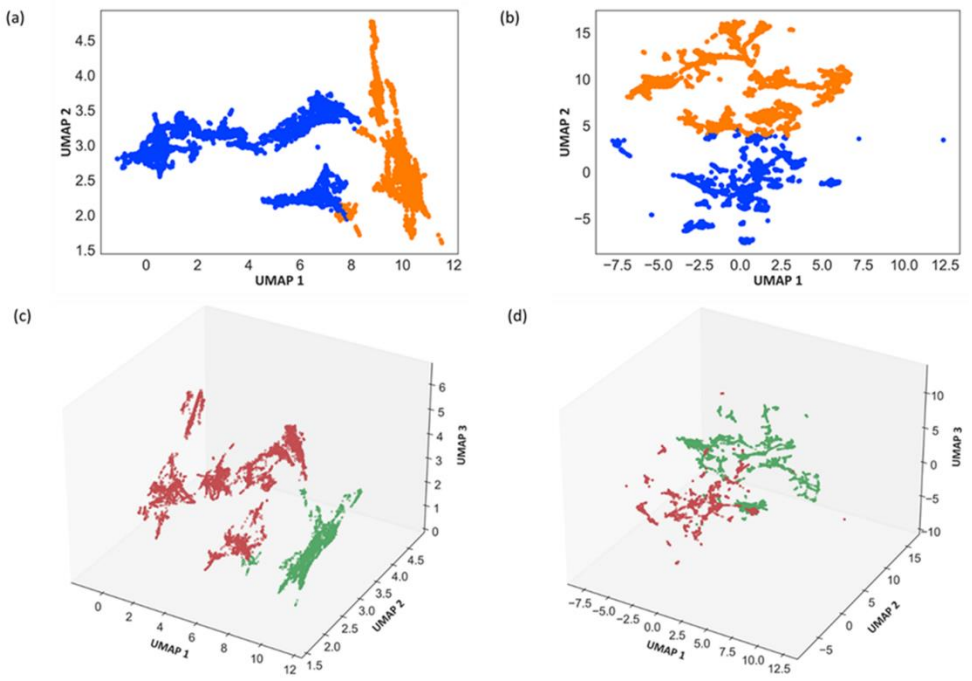


Figure 4.10 Graphical representations of data using UMAP dimensionality reduction method for (a) ($n_{descriptors}$ = 18 & $n_{components}$ = 18) and (b) ($n_{descriptors}$ = 208 & $n_{components}$ = 3); (c) and (d) are the 3D graphs for both cases, respectively.

4.5.3 Bi-clustering (NNMF) results

In the first case study, the best number of clustering components are 18 according to EV and RE tests in agreement with the results obtained in the proposed graphical method in the previous section. Where the EV and RE values were found as > 0.99 and 2.28E-5, respectively. EV and RE tests were conducted to validate the results obtained from the initial clustering step, which guided the NNMF modeling and features/variables selection process.

The relative importance of each descriptor in representing the data is determined according to its score in each component. These scores are illustrated in the small squares (cells) depicted in the plot on the left side of Figure 4.11. In particular, the higher the score the higher the relative significance of the descriptor in a specific data component. According to Figure 4.11, the most significant descriptor for the first component is the number of heavy atoms (HeavyAtomCount) with a score of 6.74 followed by number of valence electrons (NVE) with the score of 6.64. The expert has an additional essential role in the interpretation of the NNMF results and the determination of the best number of significant descriptors per each component. For instance, only the highest score descriptor can be chosen or in other cases the most significant descriptors can be chosen as the best two descriptors with the highest scores. Trial and error can also assist the expert in determining this number. Through this trial-and-error analysis and due to the random selection of these 18 descriptors in the first case, all of them were found to be significant. This prompted us to extend our investigation to include all the possible physio-chemical descriptors, allowing us to select the most significant ones in a more robust manner.

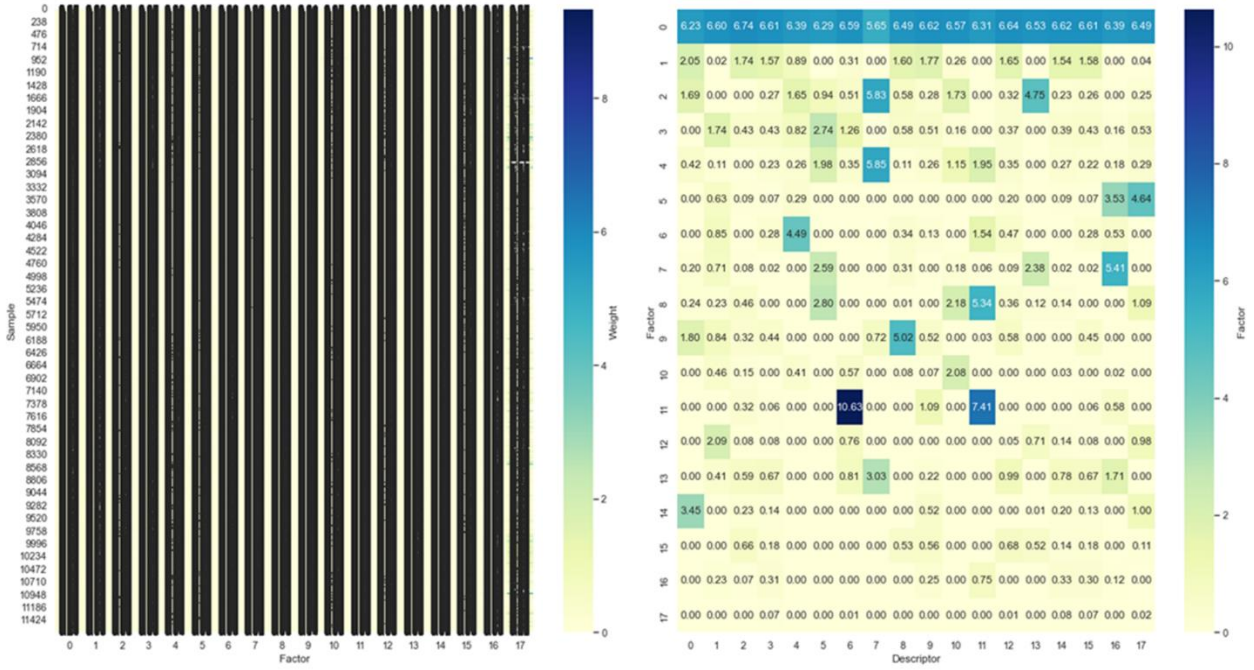


Figure 4.11 NNMF results for ($n_{descriptors}$ = 18 & $n_{components}$ = 18)

In the case of extracting all the 208 descriptors, the best number of components was also 3 as the obtained preliminary graphical recommendation from the clustering method (UMAP). The EV and RE values at this number of components were > 0.99 and 0.0006, respectively. The figure of NNMF results in this case is provided in the supplementary files due to its large size. By trial-and-error approach, among the 208 descriptors and the 2048 bit-strings fingerprints, it was found that the first 25 descriptors per component with the highest scores are the most important ones to represent the available data. This led to having 70 unique descriptors due to the presence of 5 replicated descriptors that are significant in more than single components. The list of these 70 most significant descriptors is provided in the supplementary materials. According to these two case studies it can be observed that the use of graphical approach for determining the number of components is comparable with the quantitative analysis approach based on EV and RE. This also reinforces our intuition of using the initial clustering to guide the NNMF afterwards with the optimal number of data components, helping to mitigate its computational complexity, particularly with larger datasets.

4.5.4 Comparison of different ML modeling methods

As previously mentioned, the hyperparameters of each model were optimized to obtain the highest possible accuracy. In this regard, the optimum SVR model was obtained at the hyperparameters of C = 300 and a kernel type of radial basis function (RBF); while, the optimum number of estimators, i.e., trees, of the optimized XGB model was 200. In the case of ANN model, the optimum hyperparameters are learning rate of 0.001, *ReLU* activation function, 3 hidden layers and a batch size of 16. The number of neurons of the first, second and third layers are 150, 100 and 50, respectively. Whereas the optimum RF model had 100 trees. In addition, the optimum number of neighbours (k) for *k*-NN model was 3. The optimum max depth of the optimized decision trees is 16. The optimum LSTM model contains one LSTM layer with 32 units and a dense layer with a single unit for the output. The optimum number of epochs and batch size were 200 and 16, respectively, where the "*Adam*" optimizer was adopted. It should be noted that both CNN and GPR techniques were also used. However, they gave lower accuracies, i.e. < 80%, and took relatively longer training times. This is due to the relatively large dataset used, which increases the computational complexity especially in the case of GPR modeling. Hence, their results were not included.

Table 4.6 summarizes the results obtained after performing the ML modeling using the optimized models. As demonstrated, SVR and XGB models have the highest accuracy and mean square error. This observation is extended to the cases of all three solubility parameters. The third best model was RF, as shown in Table 4, and this is due to the ability of RF to handle tabular numerical data. Being an ensemble-based technique, this enabled RF to outperform the optimized decision tree as illustrated in the table as well. Another important observation is that the LSTM model was a bit better than the ANN one. This is presumably due to the additional ability in its layers to have the seize sequential data and sequences in the available data.

After obtaining all the results from different models, they are fused to enhance the prediction accuracy. As previously mentioned, the fusion methods developed are categorized into two main categories, i.e., learnable, and non-learnable. The results of the fusion attempts are presented in Table 4.7. In the case of non-learnable fusion, both techniques based on MSE and combined R^2/MSE weighted averaging were better than those considering only simple averaging and R^2-based weighted averaging. In the case of learnable fusion, in general, it possessed higher

accuracies; however, ANN did not show high accuracy such as XGB and SVR in this stage. This is the reason why it was not considered in the second and third fusion steps. As evident from the data of Table 4.7, the non-learnable decision based on simple and weighted averaging was inspired by the bagging mechanism. While the series learnable decision fusion was inspired by the boosting mechanism to minimize both variance and bias.

In general, SVR is the best choice when the number of features is near to or higher than the number of observations. In the current case, when fusion goes on to the second and third stages, the ratio between number of observations and variables decreases. This made SVR the best candidate for completing the third learnable fusion and the results in Table 4.6 are in good agreement with this conclusion.

Figure 4.12, Figure 4.13 and Figure 4.14 were designed to visualize the comparison among the different methods. From these figures it can be observed that the decision fusion has a very significant effect on reducing the errors of individual techniques. The proposed decision fusion strategy/methodology combines the advantages of each individual regression technique which is excellent prediction model at certain regions. As a result, the decision fusion-based regression model avoids the high error of some techniques at the regions of their poor performance and follows the high predictability of the models with excellent performance at these regions. This is the reason why different shallow, deep, parametric, and non-parametric machine learning models were selected in the current study. The remaining results related to the other two solubility parameters are provided in the supplementary materials for more information. Figure 4.12 represents a sample of the data and the results. As shown, individual ML models are compared to each other and to our proposed ensemble-based methodology. For example, in Figure 4.12 (b) some of these models were selected and added for the sake of comparison, demonstrating the capability of our developed fusion (ensemble) technique to adapt and deliver high performance in predicting the desired output, i.e. HSPs in this case. However, Figure 4.13 depicts the whole experimental data available versus the predicted values of the dispersion solubility parameter after applying all the decision fusion steps. As shown, the developed model based on the proposed methodology possesses high prediction accuracy of 99.54% compared to almost 96% obtained by both individual SVR and XGB models.

LSTM was included alongside XGB and ANN for a comprehensive comparison and to demonstrate the versatility and robustness of our developed fusion technique. The inclusion of LSTM highlights the ensemble method's ability to integrate models with varying individual accuracies while still achieving high overall performance in predicting the desired outputs, such as HSPs in this study. This approach underscores the adaptability of the ensemble method in leveraging diverse model characteristics to enhance predictive accuracy.

Table 4.6 ML modeling results

Table 4.7 Decision fusion results (Solubility_1 "δ_D")

*These two tables were moved to Appendix C (Tables C.3 and C.4) in (APPENDIX)

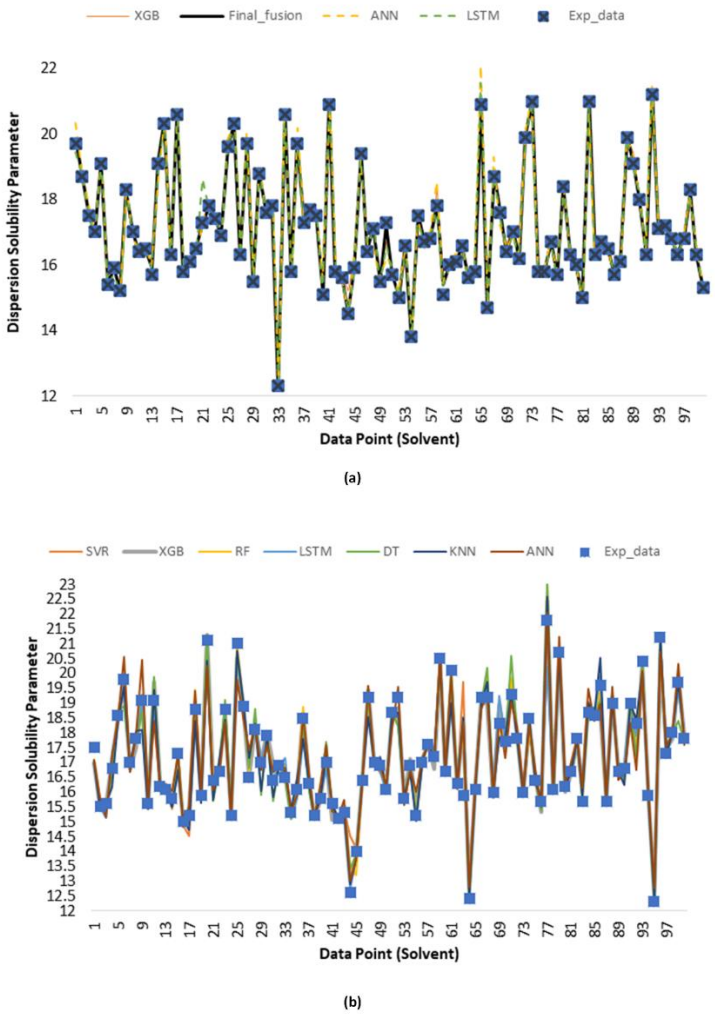


Figure 4.12 Results of (a) final decision fusion vs. selected different individual techniques, (b) different individual techniques.

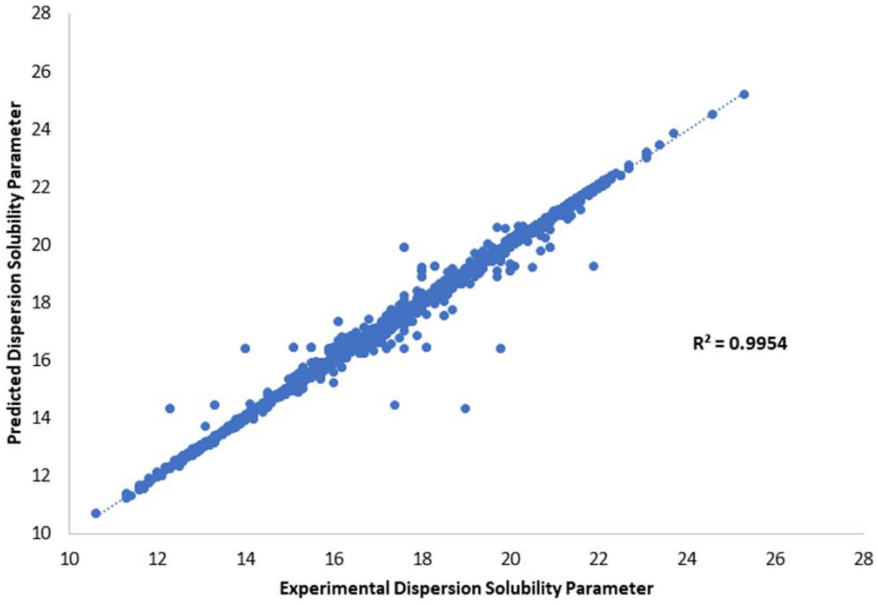


Figure 4.13 Predicted *vs.* experimental values of dispersion solubility parameter (final decision fusion results).

4.5.5 Effect of dataset size on prediction accuracy

This section discusses the effect of training dataset size on the prediction accuracy of the ML models. Figure 4.14 presents the illustration of this effect for the dispersion solubility as a relevant parameter. As shown, there is a general trend of accuracy increase and mean squared error decrease upon increasing the training dataset size. This is presumably due to giving more chances to the developed model to capture and learn all the possible combinations and data/variables distribution. However, it can be deduced that sizes above 70% to 90% are sufficient and give high prediction accuracy. This also gives an indication about the amount of data from the same distribution required to train the developed ML models. For instance, in the current case of predicting HSP based on SMILES codes, it is recommended to work on representative datasets containing more than 8500 data points to have prediction accuracies higher than 90%. These results are in good agreement with those stated in previous research studies [495].

As previously mentioned, the results in Figure 4.14 are related to the prediction of the dispersion solubility parameter. The same trends were observed in the case of the other two solubility parameters, i.e., polarity and hydrogen bonding solubility parameters. These results are documented in the supplementary materials.

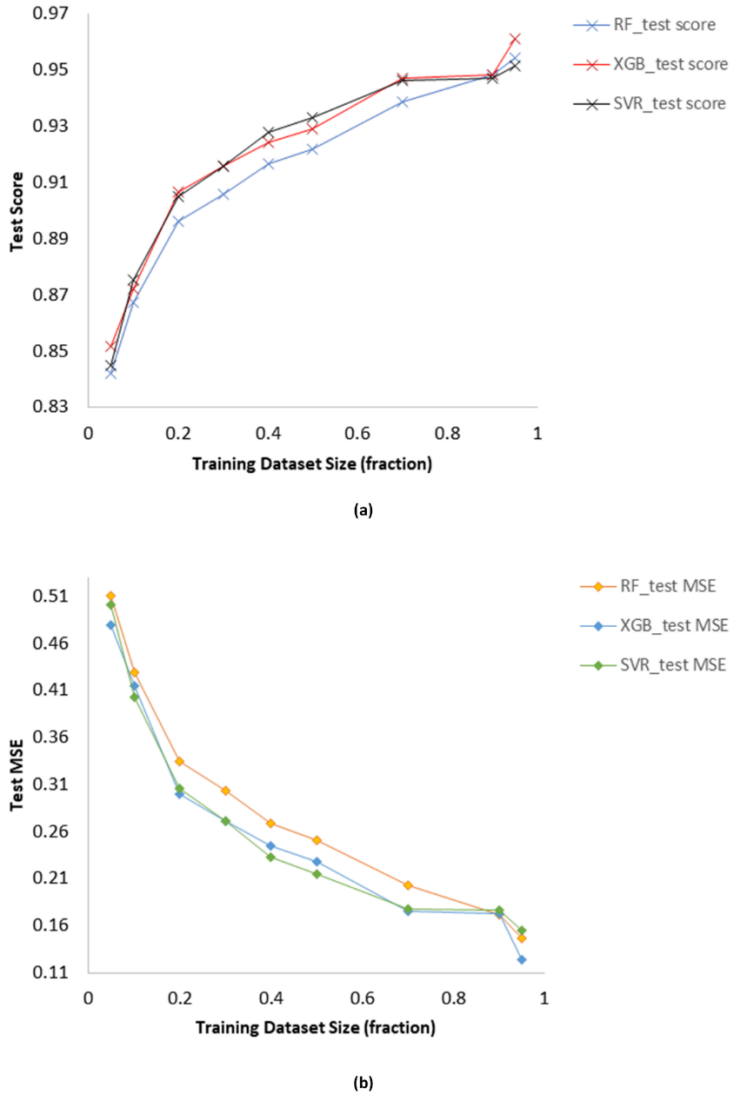


Figure 4.14 Effect of dataset size on (a) the accuracy/score and (b) mean squared error of ML modeling using RF, XGB and SVR for Dispersion solubility parameter.

4.5.6 Model explanation

The results after the addition of the XAI part are summarized in the following figures (Figure 4.15, Figure 4.16 and Figure 4.17). In these figures, the magnitude of significance of each descriptor is presented on the left-hand side, whereas the right one gives the information about the sign of this magnitude, i.e., positive, or negative. With respect to the sign, for a certain input variable, if there is a high concentration of red dots on the right side of the SHAP value scale (right of the zero-value locus), then this means that the input variable positively affects the output variable. The vice versa is true. A red dot means that the value of the output variable is high at this point and SHAP value.

In the left-hand side of the graphs, the average value of the SHAP values for all the studied points for each variable is presented to indicate the global effect of the input variables on the output ones. The following are examples of the explanation of these graphs to breakdown the extracted knowledge from them as a powerful XAI technique to overcome the problems of black-box modeling. For instance, according to Figure 4.15, the most significant factor/descriptor influencing the first solubility parameter related to the dispersion of energy is the number of aromatic rings in the molecule, whereby this impact is a positive one. Elsewise the most significant factor in the case of polarization solubility parameter is the molecular logarithm of the partition coefficient (MolLogP) and it has a negative impact (Figure 4.16). The negative impact of MolLogP extends to both dispersion and hydrogen bonding solubility parameters. As seen, each parameter has its own set of significant descriptors. This means that all different relations should be considered. Accordingly, the designer should take care about any conflict that can lead to competing multi-objective optimization problems. This information will help the designer or process operator in selecting or producing the optimum solvent(s). Specifically, if operators aim to select a solvent or group of solvents that possess high dispersion solubility parameter, then, based on the findings of this explainability study, they should choose the solvents with higher number of aromatic rings in their structure. Besides, if it is desired to have a specific solvent with a lower hydrogen bond or polarization solubility parameter, then the operators should select the group of solvents with higher MolLogP values. These guidelines can also be followed in the case of designing new solvents tailored to a specific application.

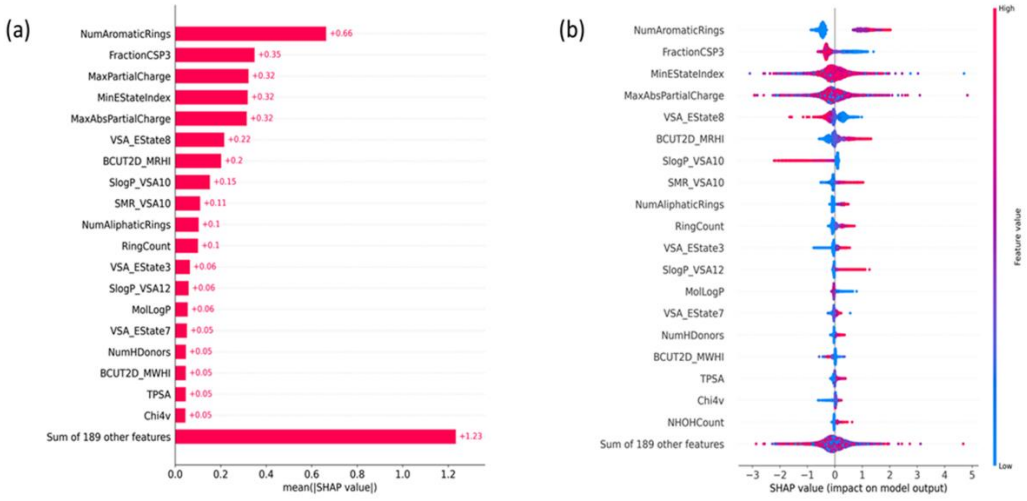


Figure 4.15 SHAP values for the Dispersion solubility parameter

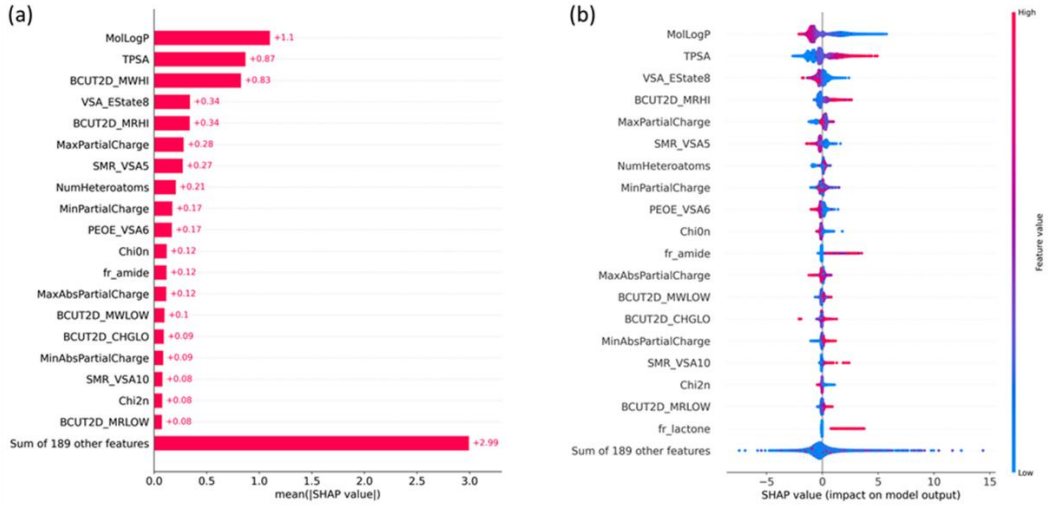


Figure 4.16 SHAP values for the Polarization solubility parameter

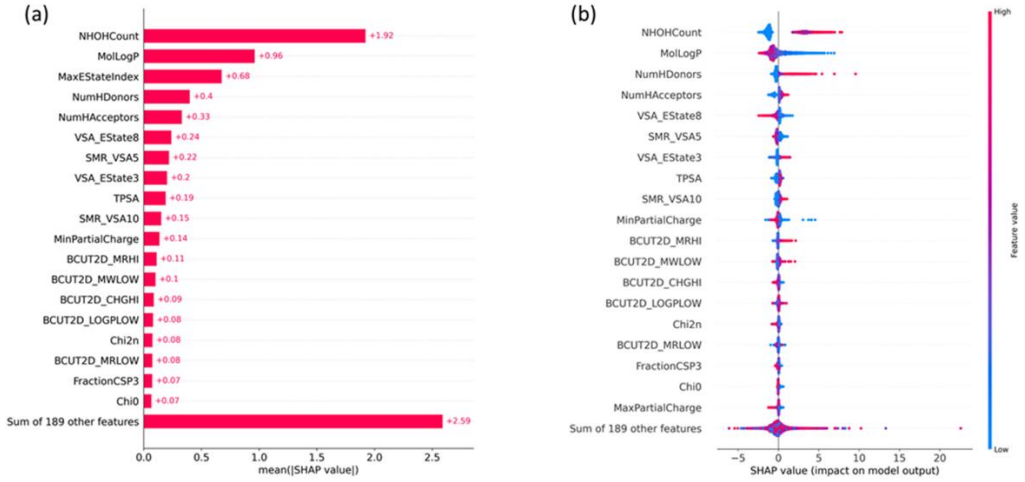


Figure 4.17 SHAP values for the Hydrogen Bonding solubility parameter

As previously stated, other specific case studies related to the selection of suitable solvents based on the calculations of *Hansen* sphere are also selected. These cases include different lignin materials, i.e., kraft lignin from a softwood feedstock and sugar cane bagasse lignin, besides CO_2.

For softwood-based kraft lignin, the polarization energy parameter exerts the greatest influence, with all three parameters positively affecting lignin dissolution (Figure 4.19). In this case, a robust interpretable machine learning (IML) model was developed to capture the key patterns that determine desired solvents based on their descriptors. The IML model was optimized to achieve the highest possible classification accuracy, with an average testing accuracy of 95%, a recall of

0.91, and an *F1*-score of 0.95. The training accuracy was nearly 100%. At the descriptor level (Figure 4.18), TPSA emerges as the most significant factor positively influencing the dissolution of this type of lignin. Conversely, for bagasse-based lignin, the hydrogen bonding parameter is the most influential, followed by the positive effect of the dispersion energy parameter. This aligns well with Figure H.11 in the supplementary materials, which highlights the number of hydrogen donors as a significant descriptor positively impacting bagasse-based lignin dissolution. These findings indicate that different lignin materials exhibit varied behaviors, necessitating more comprehensive future studies to explore the interactions between various solvents and the distinct functional groups and monomeric units of lignin materials. From these examples, it is evident that, generally, higher hydrogen bonding and dispersion solubility parameters are required for a solvent to be effective for lignin materials, regardless of their composition. Based on these findings, researchers, designers and process operators can narrow their search when selecting a group of optimum solvents for these lignins, focusing on those with higher hydrogen bonding and dispersion solubility parameters. In addition, drawing from the results of the first part of this explainability study, these solvents can be tailored to serve this application through selecting/designing these solvents with lower MolLogP values and a higher number of aromatic rings.

According to Figure H.12, the 6^{th} order electro-topological state of a fragment within a molecule (VSA_EState6) is the most significant feature that affects the solubility of CO_2 in a certain solvent, whereby this impact is mainly a negative one. The number of hetero atoms is, too, a very significant feature/descriptor compared to the others where this impact is also a negative one. This means that considering decreasing those features upon designing a new solvent will give a better performance to dissolve and capture CO_2. On the solubility parameters level, Figure S13 suggests that the dispersion solubility parameter/energy is the most significant one among the three parameters, whereby this impact is also negative, which means that selecting or designing a solvent having lower dispersion can be beneficial in the case of CO_2 dissolution/capture. Upon comparing these selected cases, the lignin problem/case is more complex and different compared to the CO_2 case study.

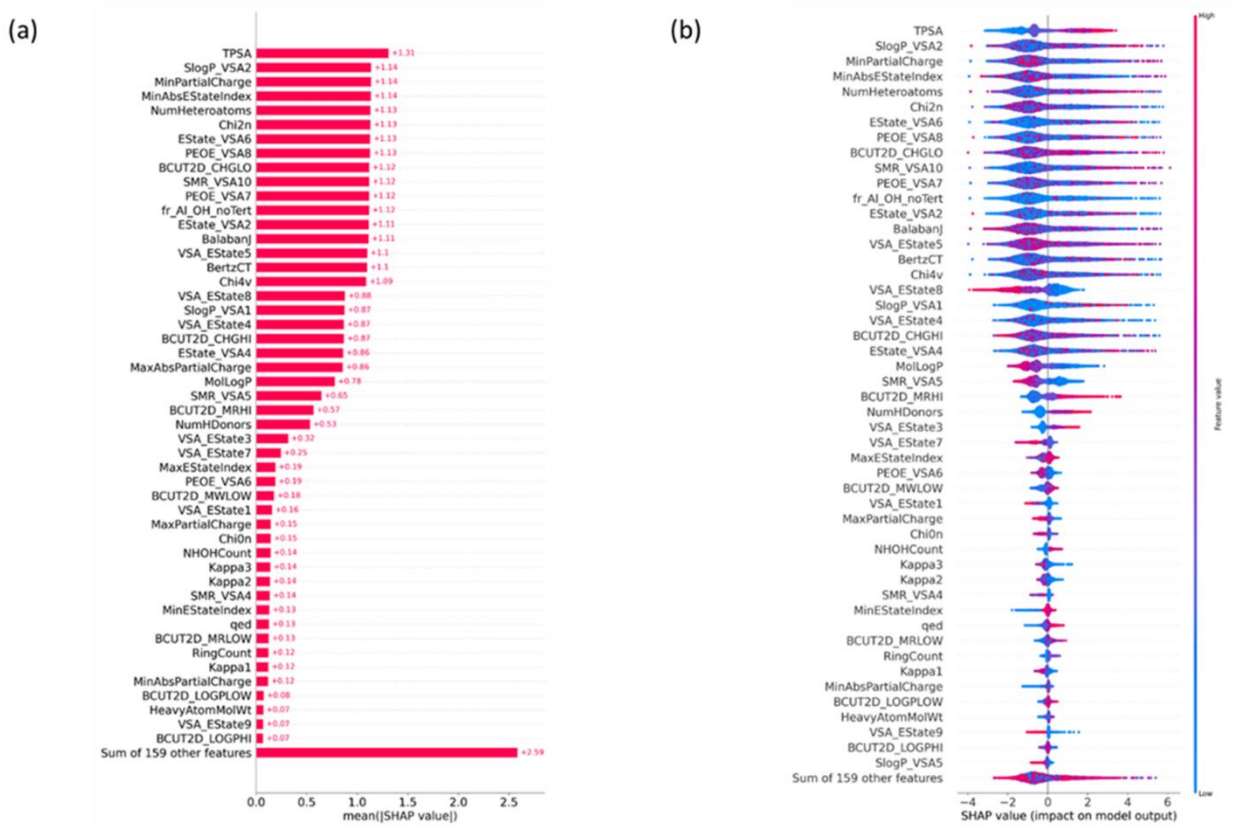


Figure 4.18 SHAP values of softwood-based kraft lignin solvents classification based on RED (**Descriptors**)

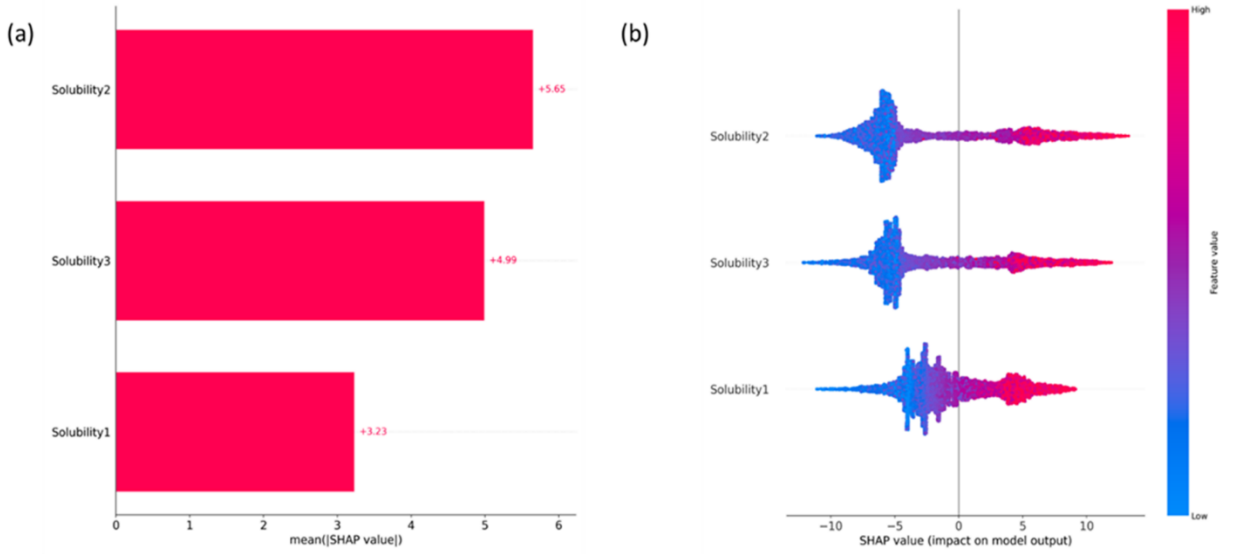
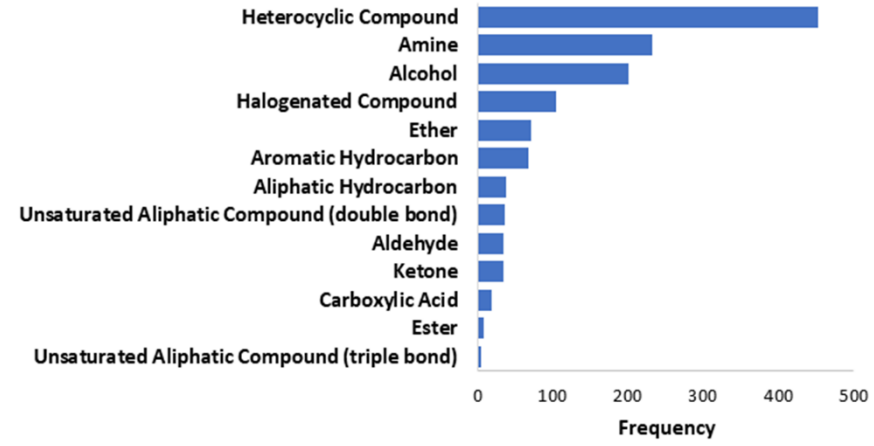


Figure 4.19 SHAP values of softwood-based kraft lignin solvents classification based on RED (**Hansen solubility parameters**)

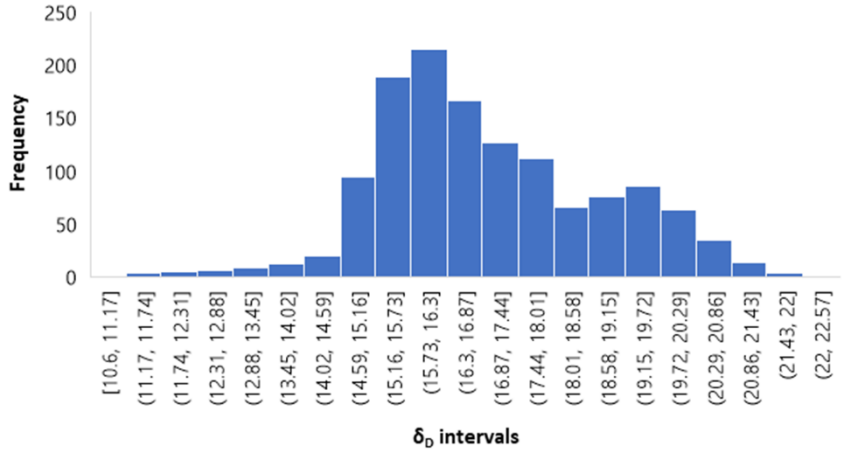
4.6 Model validation and generalization

The dataset employed in this validation case exhibits the chemical types of distribution illustrated in Figure 4.20(a), based on the available SMILES codes. As shown, this distribution differs somewhat from the larger dataset used for training the developed ensemble model. For example, this dataset lacks phenols, and both ketones and carboxylic acids are significantly underrepresented compared to the training dataset. Additionally, the numerical distribution of the Hansen solubility parameters (HSPs) varies. Figure 4.20(b) displays the frequency of dispersion parameter values (intervals) in the validation dataset, where the maximum value is 22 and the average is 16.9.

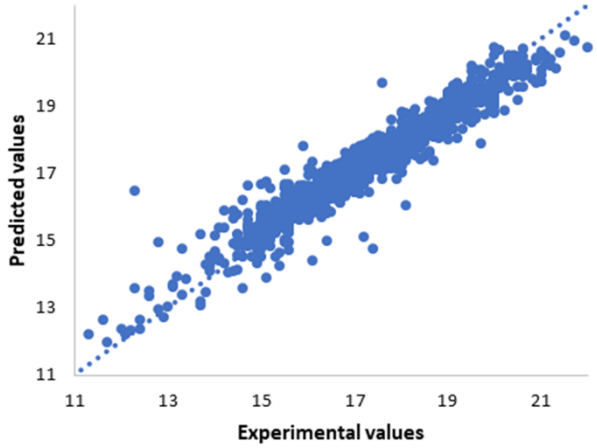
Regarding ensemble-based modeling, the developed model demonstrated its efficiency in predicting HSPs for the validation dataset, achieving high prediction accuracy across all three parameters. For instance, Figure 4.20(c) compares the predicted and experimental values of the dispersion Hansen solubility parameter, showing a determination coefficient of 93% and a mean squared error below 0.24. These results confirm that the developed ensemble-modeling technique can be effectively generalized to other cases, enabling accurate determination of HSPs.



(a)



(b)



(c)

Figure 4.20 (a & b) Distributions and (c) results of validation dataset

4.7 Comparison with literature, impacts and remarks

Recently, there is a considerable number of studies that investigate the application of AI/ML to predict various solubility parameters/representations including HSPs, molar equilibrium solubilities and *Hildebrand* solubility parameter [496]–[498]. Table H.4 (Appendix H) summarizes different selected attempts in the literature related to the AI-assisted solubility parameters prediction to present their performances and compare them with the current study. Some of these studies focus on relatively small datasets or specialized datasets with a very narrow and focused spectrum of chemical classes, which may limit the model's ability to generalize. These limited datasets do not capture the full spectrum of chemical classes or structures. In addition, as mentioned in the introduction section, most of these studies rely on black-box ML modeling, which hinders the understanding of the impact of various input features/variables on the target outputs (solubility parameters). For more details on the data types, number of observations and models used in each study, readers can refer to Table S4 in the supplementary materials. According to the summary in the table, our methodology has several advantages such as high predictability and explainability. The use of larger datasets enhances the model generalization where larger true distribution is captured during training. The high predictability comes from the combination of several different ML techniques and architectures through various learnable and non-learnable fusion techniques as well. The dependence on the molecular structure and the corresponding molecular descriptors besides the use of large dataset of diverse solvents/materials ensure generalization and make the developed method suitable for a wide range of materials including ionic liquids and organic solvents. With respect to features selection, our methodology introduces a robust way depending on NNMF technique instead of relying on manual selection as most of the studies in literature. On the technological and managerial sides, the developed methodology and the accompanying tool for HSPs prediction will offer the following:

- Prediction of HSPs of new materials or solvents based on their molecular structure, represented by their SMILES codes, and the corresponding molecular descriptors with high accuracy in fractions of second. This will waive the burden of experimental procedures needed to evaluate these important parameters for new materials/solvents
- Streamlined assessment and screening of solvents based on their HSPs to be used for dissolving materials of interest such as various lignins and CO_2 which is a current hot real-

world application related to *Net-Zero* strategies. This can lead to the suggestion of new solvents to perform specific environmental tasks that were not their primary field of application.

- Easy evaluation of compatibility between different materials, e.g. lignins and other polymers, during mixing based on the selection of the group of optimum (good) solvent for both polymeric materials.
- Accelerated decision making can be performed by different stakeholders based on the fast and accurate predictions offered by the developed methodology and tool to select the best solvent among a group of optimum solvents from a technical point of view which can be associated with economic one in future research.

It is worth noting that there are other studies that used HSPs as inputs to predict different important variables such as gelation properties [433], and polymers' cloud point (as an indicator of phase behavior of polymer solution system) [495]. Another application is the prediction of power conversion efficiency of solar cells system (specifically non-halogenated green solvent-processed organic solar cells) [499]. Accordingly, our developed methodology and tool can play a key role in future research in these fields through providing accurate predictions of HSPs of new materials. These predicted HSPs can be then fed as inputs to have a reasonable estimation of the abovementioned variables of interest.

Combining high prediction accuracy and XAI has an additional important advantage where its results will be provided to the material designers to accelerate the solvents design/discovery for specific cases such as those presented in the current study. The results can also be used by the process designers to accelerate the selection of the most suitable solvents to make their process more feasible and greener. According to the presented case studies, accelerating the solvent design as well as the process design will accelerate the adoption of *Net-Zero* processes including carbon capture and lignin valorization biorefineries. Therefore, the future work will focus on the generative deep leaning to discover new materials, e.g., lignin and CO_2 solvents, based on their desired properties. The input desired properties will be mainly the Hansen solubility parameters sought. Language models including the generative pre-trained transformers (GPTs) will play a key role in this future research. Another direction for future research is the data augmentation for different types of SMILES codes for various chemical compounds and the corresponding properties for further enhancement of data quality. Furthermore, advanced data imputation using

generative deep learning methods also serves in data quality enhancement and this will increase prediction accuracy. Besides, considering data/models uncertainty quantification will increase the confidence of the obtained results. As a side note, SMILES codes alone have some limitations in fully representing the solvents/compounds of interest. To address this, we proposed using both physio-chemical descriptors and *Morgan* fingerprints for a more comprehensive representation of each compound or solvent. As mentioned earlier, these descriptors and fingerprints were extracted using *RDKit*. However, other physical/chemical properties or descriptors such as phase change points (melting or boiling points), refractive index and heat of vaporization could be incorporated as inputs in future research to further enhance the prediction accuracy and confidence. Additionally, other software packages can be used to extract more informative descriptors, avoiding the limitations that arise when relying on a single library. For future work, we also plan to conduct an extensive comparison between our developed methodologies and other traditional methods, such as functional group contribution methods. This is to provide a robust quantitative assessment in terms of both time efficiency and accuracy along with reasonable interpretability. As we believe, this methodology is generalizable, we plan to extend its use to other applications such as the characterization of hydro-char and pyrolysis products, e.g. biochar, and predicting their yields. Preliminary results have been obtained from this extension, and they are promising.

4.8 Conclusions and prospects

This research work is a contribution to the development of Green and Sustainable Science and Engineering, where green chemistry and analytical chemistry combined with artificial intelligence methods play an increasingly important role in accurately characterizing the properties of biobased materials for their valorization or conversion into high value-added products. This paper introduces a new AI-methodology to enhance chemical compound identification and to predict accurately the properties of lignin solubility and solvent selection using ensemble machine learning methods for decision fusion. This methodology relies on the creation of an ensemble of different shallow and deep machine learning models having various architectures and algorithms followed by decision fusion through weighted averaging and other learnable pathways. As a result, upon applying the methodology for *Hansen* solubility parameters prediction, the prediction accuracy reached 99% with a mean squared error of 0.02. With respect to the feature selection step, NNMF was successfully used to reduce the number of input variables from 208 descriptors and 2048-bit strings fingerprints to only 70 most significant variables/descriptors. This helped to largely reduce the

computational requirements for the training-step. This also increased the prediction accuracy afterwards after eliminating unnecessary input variables that can deteriorate modeling efficiency. The addition of the XAI part to the developed models gave a clearer picture of the impacts of different variables on the predicted outputs that overcame the obstacle of black-box predictions. Clearly, the utilization of an ensemble-based ML approach emerges as a critical strategy in accurately predicting the functional properties of diverse materials including lignin and its derivatives, as well as CO_2 liquid sorbents. This methodology proves reliable to assess lignin compatibility with diverse solvents and polymers, addressing the challenges faced by traditional models and individual ML techniques. By integrating various algorithms and decision fusion methods, our approach significantly enhances prediction accuracy, providing a robust tool for researchers and practitioners in fields such as polymer science, coatings, and biorefineries. The versatility of this ensemble-based ML methodology extends its applicability to a wide range of chemical analytics, reinforcing its importance in advancing the understanding and utilization of lignins in various industrial applications, as well as discovering new efficient solvents for carbon capture. The techniques proposed herein, whereby CO_2 and lignin dissolution represent the two case-studies, have the potential to accelerate the design of green process, thus contributing to *net-zero* emissions targets.[2]

[2] All supplementary materials related to this chapter are available in (Appendix H)

CHAPTER 5 ARTICLE 3: ACCELERATED GREEN MATERIAL AND SOLVENT DISCOVERY WITH CHEMISTRY- AND PHYSICS-GUIDED GENERATIVE AI

Eslam G. Al-Sakkari (Eslam Ibrahim), Ahmed Ragab, Marzouk Benali, Olumoye Ajao, Daria C, Boffito, and Hanane Dagdougui

Article Submitted on October 8, 2025 and Published (online) on January 2, 2026 in Artificial Intelligence Chemistry, Elsevier, Volume 4, Issue 1, 2026, Article number 100106

Online Publication Date: January 2026

DOI: https://doi.org/10.1016/j.aichem.2025.100106

5.1 Abstract

Carbon capture, utilization and storage (CCUS), along with lignocellulosic biomass valorization (e.g., lignin, cellulose), are promising decarbonization strategies for hard-to-abate industries. Green solvents, such as deep eutectic solvents and ionic liquids, enable efficient CO_2 capture and selective lignin extraction, enhancing lignin depolymerization into high-value products. However, current molecular design tools are slow and computationally expensive, limiting green material innovation. This study introduces a novel data-driven framework for green material discovery using generative AI, including transformers, generative adversarial networks, and variational autoencoders. The generation process was guided by rule-based and physics- and chemistry-informed models for automatic labeling, with feedback loops to reduce invalid SMILES strings. The approach achieved 70% molecular validity and 94% novelty in generating new solvents for CO_2 capture and lignin applications. Model training averaged under one hour, and molecule generation took only seconds, significantly faster than traditional methods. Ensemble machine learning models assessed the environmental sustainability of candidates, and retrosynthesis analysis identified feasible, green synthesis pathways. This flexible, scalable methodology extends beyond solvent discovery to broader applications in process design and optimization, enabling the rapid generation of novel and cost-effective process configurations.

5.2 Introduction

Achieving net-zero emissions goals requires rapid deployment of effective industrial decarbonization strategies reducing production systems' carbon intensity. Key pathways include carbon capture, utilization, and sequestration (CCUS), biomass and biowaste valorization through standalone or integrated biorefineries, and fuel switching to low-carbon alternatives. Ensuring these pathways are sustainable and cost-effective is essential for enabling their broad adoption across diverse industrial sectors.

Solvent-based CO_2 capture remains the most mature CCUS technology [156], [500]. Yet, conventional solvents such as monoethanolamine (MEA) and methyldiethanolamine (MDEA) face major limitations, including high regeneration energy, thermal degradation, and poor environmental compatibility [501]. This has intensified interest in greener alternatives, e.g. non-toxic, non-flammable solvents that reduce cumulative energy demand [502] and improve process performance, including emerging candidates such as superbases and ionic liquids. A similar

challenge appears in lignin valorization, where solvent choice critically influences lignin solubility, extraction efficiency, and compatibility with downstream processes. Despite lignin's potential for producing fuels, chemicals, and advanced materials [503]–[506], optimal solvent identification is hindered by its structural variability [507], [508], the need for solvent blends, and the technical and environmental shortcomings of many existing solvents [509]. Issues such as limited solubility, undesired reactions, high energy requirements, toxicity, and poor biodegradability highlight the need for new, sustainable solvent systems across both CCUS and biorefinery applications [510]–[513].

Despite these limitations, developing new greener, sustainable and techno-economically efficient industrial materials remains imperative to catalyze adopting industrial decarbonization systems [514]. Yet, discovering new "green" solvents remains difficult because experimental screening [501], [515], [516] and physics-based evaluations [517]–[525] are slow, expensive, resource-intensive, and can be fundamentally limited in their ability to explore large chemical spaces [526]–[528]. For instance, *Hansen* solubility parameters (HSPs) [411], [412], [416] provide a physically grounded framework to estimate solvent-solute affinity and have been widely used to assess solvent suitability in CO_2 capture and biomass/lignin processing. However, only a small subset of potentially useful organic solvents has been experimentally characterized, and existing datasets are sparse, noisy, and chemically biased. This makes solvent discovery dependent on manual heuristics or incremental modifications of known molecules, frequently overlooking unexplored regions of chemical space [370].

Recent advances in rule-based methods, data-driven modeling and generative artificial intelligence (GenAI) offer powerful tools for molecular discovery/design. They enable the automated discovery of novel chemical structures and the rapid exploration of broad design spaces [529]–[537]. The common GenAI models used are long short-term memory (LSTM), generative adversarial networks (GANs), variational autoencoders (VAEs) and generative pre-trained transformers (GPTs) [197], [538]–[540]. More cases in the literature are introduced in the supplementary materials file (**Table I.1** and **Figure I.1**). Yet, purely data-driven models may produce invalid or chemically implausible molecules and rarely incorporate the combined physical/chemical principles underpinning CO_2 solubility, lignin affinity, or green-chemistry constraints. Besides, rule-based methods face some limitations such as high reliance on human expertise and long computational time. These limitations also appear due to lack of ensemble and incremental

learning. Integrating ensemble generative modeling with physics-based evaluation under the umbrella of a human-centric approach is therefore essential for producing novel molecules. These molecules are not only syntactically valid but also aligned with real-world thermodynamic, environmental, and process requirements.

This paper proposes a new physics-guided ensemble-based generative AI methodology for the discovery of greener materials (solvents) based on desired physio-chemical properties. GPT-2, GAN and VAE are ensembled in this methodology. To make it physically and chemically aware [541], an additional external loss loop is added to examine the produced molecular structures' validity. This makes it physics-guided *GenAI* with human preference approach. Here, the early proper datasets preparation based on human preferences plays an analogous role as a direct preference optimization (DPO) mechanism [542]. This accelerates generative models training, especially GPTs and enhances their efficiency. In the case of solvent generation, the dataset used for training will contain only the good solvents that can dissolve the materials of interest. This will be done through ensuring that all the materials in the dataset have specified HSPs and relative energy differences (REDs) that make them suitable solvents for the material of interest. The proper training dataset distribution guides the generative models to create novel solvents having primary properties (KPIs) similar to or near the ones in the training dataset. Afterwards, another optimization steps are performed to improve the additional KPIs, e.g. greenness indicators, e.g. toxicity and flammability indicators. This action accelerates the overall generation process. The main scientific and technical contributions of this study addressing key gaps in green solvent design and related data-driven modeling are:

1. Develop an incremental ensemble-based generative modeling methodology considering human preferences while deploying the generation. Image processing and natural language processing (NLP) concepts are combined to build this new methodology to leverage their different capabilities on various levels for a more robust generation.
2. Build a novel methodology comprising physics-based and data-driven models for auto-labeling to validate the generation results making the generative model physically and chemically aware, i.e. physics-guided GenAI.
3. Automate generation evaluation through AI-assisted KPI(s) determination. Therefore, ensemble-based models are developed and deployed to perform the evaluation. Various

machine learning (ML) models with different architectures are combined, and their results are fused.

4. Apply the developed methodology to generate green solvents for crucial materials such as CO_2 and lignins. A new metric defining the greenness of generated molecules is constructed where a score is given to the generated material based on its properties to facilitate ranking and selection.

5.3 Methodology

5.3.1 General approach

This study builds upon a previous work that focused on developing an ensemble-based ML model for solubility prediction and solvent selection [543]. Figure 5.1 illustrates the general approach of the current study. Raw data, including all the possible types, go through a series of pre-processing steps to ensure generating new valid texts/images. After obtaining the analysis of ready data, each type of data is fed to the corresponding generative modeling technique in the proposed ensemble-based generation mechanism. The main generative techniques adopted in the current study are GPTs, i.e. GPT-2, advanced GAN and modified VAE. Validation of the generated texts and images is performed through a combined approach comprising data-driven auto-labeling and physics-based evaluation. The data-driven auto-labeling step is done by adopting transformers-based text classification. After validation, the generated texts/images are evaluated through a comparative analysis to ensure they have the desired primary KPIs. Additional evaluations of the generated designs (texts/images) are performed to confirm meeting other desired qualities. The proposed approach follows the incremental learning philosophy to further minimize models' hallucination and generation inefficiency. To ensure broad applicability, the proposed method supports raw inputs in the form of text, images, and text–image pairs.

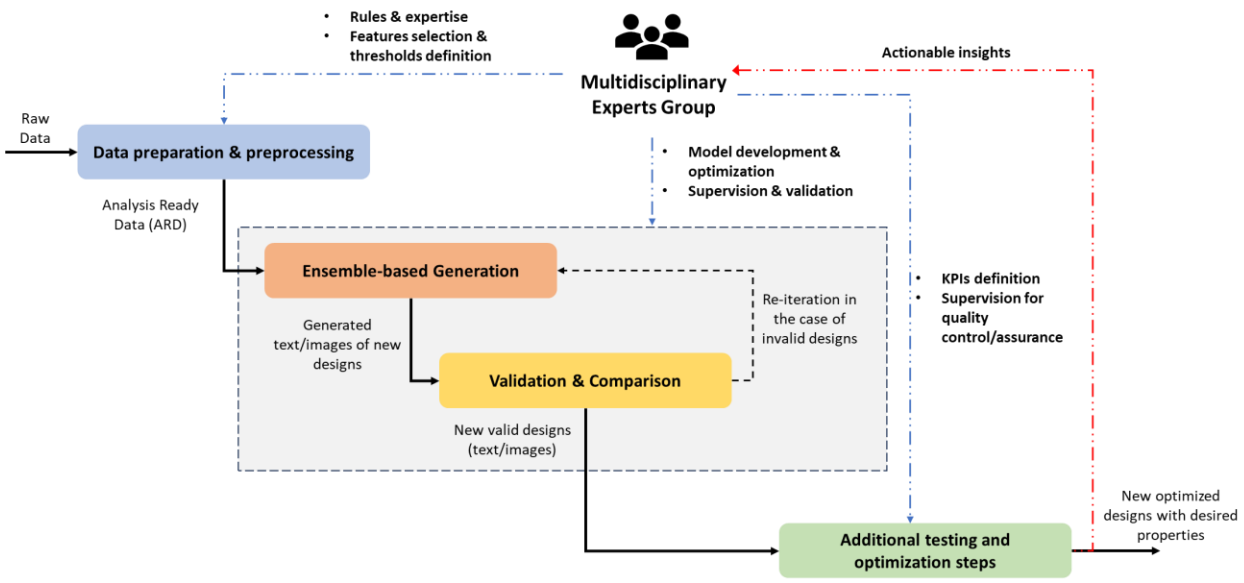


Figure 5.1 General Gen-AI approach

5.3.2 Data preprocessing and preparation

Available data types include tabular values, images, and text formats. Each type needs different preprocessing and modelling approaches. This study focuses on text and image generation. In the case of GPTs, the text data undergoes tokenization and positional encoding to produce character/word embeddings for training and generation purposes. **Figure I.2** illustrates general tokenization and positional encoding procedures for the GPT-2 model. *GloVe* [544], *Spacy* [545] and *Word2Vec* [546] embeddings generation libraries were among the preliminary suggested strategies to generate the embeddings. Yet, the best one was using the byte pair encoding (BPE)-based tokenizer and positional embedding generator already built in the GPT-2 model itself to deal with character/string sequences. The other libraries are more effective for whole words representation.

If the data is in image form, these images should be first converted to the textual form to be processed efficiently by the transformers-based model. On the other hand, the textual data should be converted to image-like forms to be compatible with GANs, convolutional VAEs or hybrid GAN-VAEs. Texts and string sequences are used to build adjacency and feature matrices to facilitate GANs and VAEs training. In the case of image-based raw data, the images should undergo a vectorization step before being introduced to GANs and VAEs. The full vectorization procedure for image-based raw data is illustrated in supplementary materials (**Figure I.3**).

Before executing the steps described above, human experts should define the threshold values of the desired KPIs and the preferred data range or distribution. This human preference-based generation approach helps accelerate the generation process and improves its precision. A data cleaning step is also essential to ensure the validity of the processed data and to remove unnecessary variables. Figure 5.2 illustrates these preprocessing steps.

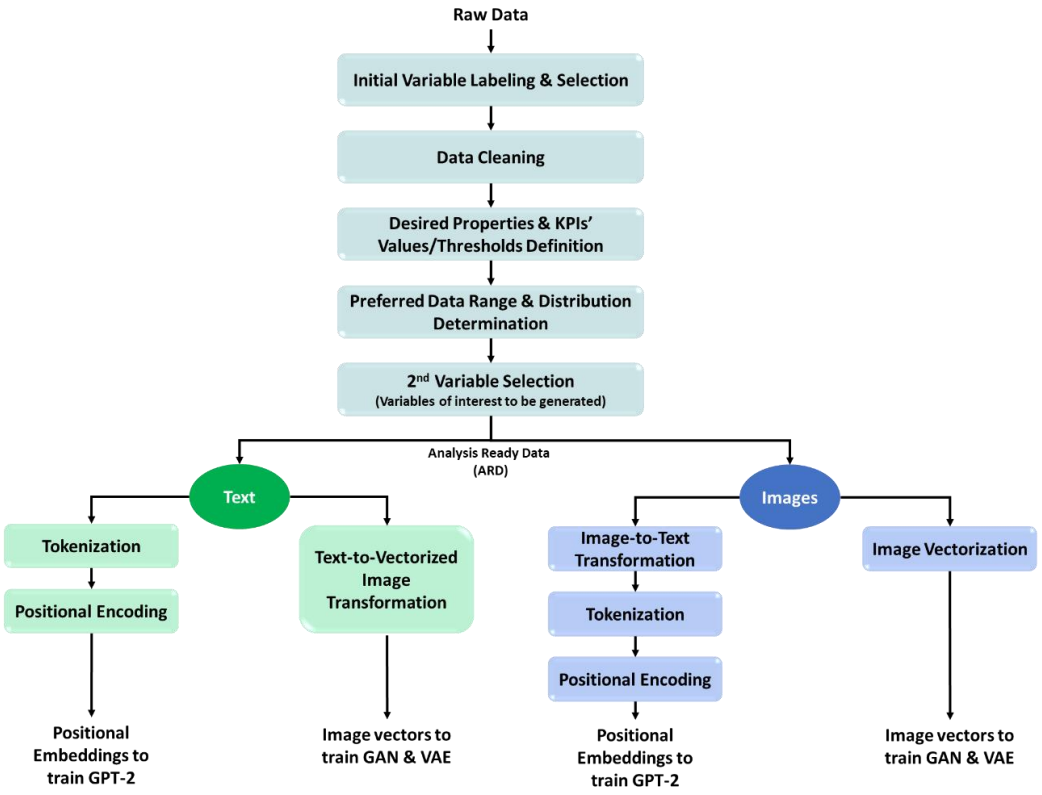


Figure 5.2 General data preprocessing and preparation methodology

5.3.3 Ensemble-based Generation

The main strategy of creating the ensemble is to concatenate the outputs of the three models, GPT-2, *Wasserstein* GAN (W-GAN) and convolutional VAE, each assigned equal weight. The equal weighting of GPT-2, W-GAN, and VAE was used only as an initial baseline to ensure that each generative model contributed equally at the start of the incremental learning process. This allowed us to evaluate the relative performance of each model in terms of novelty, validity, and overall molecular quality without introducing bias at the outset. After the first-generation round, we applied adaptive reweighting based on the observed performance of each model. Below is a brief

description of each model, including the rationale of its adoption of the relevant hyperparameters and their ranges.

Generative pretrained transformers (GPTs)

GPTs rely on self-supervised learning [547] where data labels are generated directly from input data, eliminating the need for manual or any weak labeling [548]. This is an advantage of GPTs where training time, expenses and the dependence on domain expertise in specialized tasks required during supervised learning are minimized [549]. It also avoids errors associated with weakly supervised labeling. Given its substantially smaller size relative to GPT-3, GPT-3.5, and GPT-4, GPT-2 was selected for the generation step. Its architecture is presented in Figure S4 of the supplementary materials During training, the model learns both token and positional embeddings, enabling it to predict each subsequent token based on the preceding context. The self-attention mechanism enables transformers to assign importance of weights to words or characters within a sequence during data processing. These weights capture the contextual relations between words in the sentence or understand the sequence of the characters with acceptable reasoning. Further, the mathematical foundation and architecture of GPT-2 can be found in the *OpenAI GPT2* documentation on the *Hugging Face* website [550] and in Radford *et al.* study [551].

For this work, the previously fine-tuned GPT-2 model [552] was adopted and modified significantly to fit this study objectives and generation preferences. The pre-training step is performed on a broad distribution of the available data, followed by re-training on targeted data subsets aligned with predefined human preferences. Further examples and illustrations will be provided in **Section 3** through case studies. Several hyperparameters of the fine-tuned GPT-2 are considered in the current study including learning rate, decay weight, number of epochs, the optimizer type, and batch size. Additional hyperparameters related to tokenization and positional encoding steps include the vocabulary size and maximum sequence length. The optimized values of these hyperparameters based on random search strategies are presented in the **Results and Discussion** section.

GAN

An improved version of GANs, the *Wasserstein* GAN (W-GAN), was adopted in the current study. First introduced in 2017 [553], W-GAN addresses limitations of tradition of GANS, including training instability [554] and mode collapse [555]. It replaces the *KL* divergence and *JS* divergence

with the *Earth Mover's* distance in the cost function, which improves training stability [556]. This improvement prevents vanishing gradient problems through the discriminator enforcement to become *Lipschitz* continuous [553], [556]. W-GAN is also capable of generating realistic samples from complex image distributions [557], making it suited for our case studies introduced below. Nevertheless, W-GAN's ability to generate novel and diverse samples is yet limited by the size of the training dataset. As such, in this study, W-GAN is used to generate images of medium complexity or those derived from sequences of moderate length or intricacy. In the case of generating new SMILES codes representing molecular structures, W-GAN is used for the generating small to medium-sized molecule structures, while GPT-2 is employed for producing larger and more complex molecular structures with higher novelty and diversity. The W-GAN hyperparameters comprise the optimizer type, learning rate, batch size, and number of epochs. Their actual values considered in the current study are provided in **Section 4**. The implementation of the W-GAN model was adopted from *Keras* website [558], with several modifications tailored to our methodology.

Convolutional VAE

The architecture of the developed convolutional VAE is based on the encoder-decoder structure, in which both encoder and decoder are using convolutional neural networks. The open-source implementation of convolutional VAE available on the *Keras* website [559] was adopted to develop our model. Several modifications were made to align with our methodology and related case studies. The encoder converts the discrete image inputs (or the images derived from strings or text) into a continuous latent vector. While the decoder reconstructs these latent representations back into the input original discrete form, like traditional autoencoders (AEs), VAEs are trained to minimize reconstruction error. However, VAEs introduce a regularization term in the loss function to avoid overfitting and ensure a structured and continuous latent space. Unlike AEs, the encoder of VAEs converts the input data into a distribution representing the latent space instead of encoding each observation as a single point. Accordingly, the produced latent space is guaranteed to be structured and continuous. This makes VAEs applicable and useful for image generation. For further details on the mathematical background and basic concepts of VAE or convolutional VAE (CVAE) with examples, readers are referred to [560]–[562]. The hyperparameters of the adopted CVAE include learning rate, optimizer type, batch size, and epochs number. The final values of these hyperparameters are detailed in **Section 4**.

5.3.4 GenAI results evaluation

Accuracy, determination coefficient (R^2) and mean squared error are the common evaluation metrics of predictive and descriptive supervised ML techniques. However, specialized metrics are required for evaluating generative AI/DL models. *BLEU*, *ROUGE*, *Perplexity,* and *F1*-Score are among the main metrics commonly used for language models, including GPT-2 evaluation. For GANs, evaluation is performed using *Fréchet* inception distance (FID), inception score, and visual inspection. Additionally, reconstruction loss, *KL* Divergence, and mean average precision are suitable for evaluating VAEs.

On the other hand, there are other human evaluation and validation methods/metrics when dealing with other forms of data, e.g. unstructured string/character-based data. Upon considering new material/molecule generation, the metrics mainly evaluate the validity and uniqueness of the generated text/image. The selection of the optimum metric(s) is case-specific. Thus, we relied mainly on human-related evaluation and validation methods.

General validation strategies

To ensure validity of generated outputs, several methods were developed and adopted including physics-based and rule-based validation methods, depending on the data type. For instance, when generating new SMILES, it is essential to ensure physical and chemical likelihood. To this end, fundamental rules must be respected. For example, non-aromatic atoms are referred to by upper-case letters, while the aromatic atoms are represented by lowercase letters. If the atoms are represented by two letters, then, the second letter must be lowercase. The chain branches are represented by placing their symbols between parentheses. The location of the branch symbol is exactly after the symbolic representation of the connected atom in the chain. If the branch connection is not through a single bond, the bond symbol is written directly after the left parenthesis [436]. Violating these rules leads to chemically invalid molecules. Therefore, we implemented a "*Human in the loop*" approach to manually validate the generated molecules. For instance, the list of generated molecules is subjected to a validity test using physics- and chemistry-based libraries such as *RDKit* to filter out any invalid SMILES codes (molecules). In addition, human experts further examine the generated molecules according to their chemical knowledge and writing rules of the SMILES code. This validation step is essential, as it enables differentiation between chemically valid and invalid molecule structures, which would otherwise be impossible.

In parallel, a data-driven-based auto labeling validation strategy is implemented using language models such as Llama, BERT, Roberta, and GPT 3.5 language models. These language models are fine-tuned and trained on a specialized dataset to accurately validate the generated texts or images. These models are trained in domain-specific question-and-answer datasets. For example, in the case of SMILES codes classification using Llama, each instance is evaluated via questions like "*Is this SMILES code valid?*". The expected output is a binary output "*Yes, it is valid.*" or "*No, it is invalid.*". In the case of BERT, the dataset is provided in the form of two columns, the first one is the SMILES code the other one contains the respective label, i.e. (1: valid code) and (0: invalid code). In more advanced settings, models are trained to provide reasons for being valid or invalid. This gives some reasoning to the developed auto-labeling Llama-based model. These questions and answers are case-specific and customizable upon the human expert's preferences. In the current study, the dataset used for training the auto-labeling models was constructed from famous databases such as ChEMBL. Besides, some invalid SMILES codes with different errors to capture all the invalidity reasons. This dataset contains 100k instances of both valid and invalid SMILES codes for accurate training. While training the models, an 80/20 split ratio was considered.

Evaluation via numerical comparison with desired KPIs

Once validated, generated texts and images are processed to extract the relevant features to determine their properties and compare them to the desired ones. Image processing and NLP techniques are implemented to ease the accomplishment of this goal. For SMILES codes, meaningful numerical descriptors and fingerprints are derived to assess the KPIs associated with physio-chemical properties. The numerical descriptors are fed to a physics-based or an accurate surrogate model to predict/evaluate the corresponding KPIs values. Figure 5.3 illustrates the overall methodology for ensemble-based generation, validation, and KPIs comparison. **Section 3** provides further details on the implementation of these AI-assisted steps, the definition of KPIs, and their desired values based on the available case studies. This is an iterative process under the full supervision of human experts who define the desired KPIs and ensure the validation of outputs, which are concatenated in a single list of outputs.

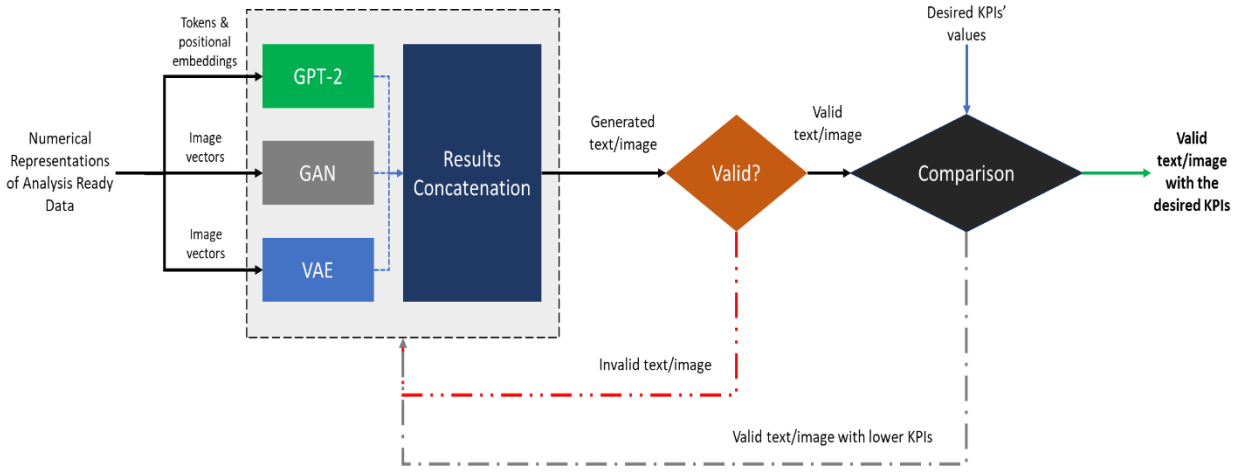


Figure 5.3 Proposed overall methodology for ensemble-based generation, results validation and KPI(s) comparison

5.4 Case studies

Molecular structures can be represented in 3D, 2D image formats, or as text forms, like SMILES (*"Simplified Molecular Input Line Entry System")* codes [435], [436]. Following a previous work on machine learning-based modeling for accurate prediction of lignin solubility [543], SMILES are the preferred data format in this study due to their ability to represent chemical formulas and easy molecular characteristics extraction using ML models [437]. They are used here to facilitate the discovery of new solvents (Table 5.1 and Table 5.2).

As a validation of our approach, molecular structures were represented by SMILES notation. The raw dataset used in our study consists solely of SMILES codes in the textual form. The image-processing component was included earlier in the manuscript to demonstrate the generality of the methodology and to clarify how image-based data could be prepared for analysis. For the GAN and VAE models, the SMILES codes were converted to graph-based representations to leverage the capability of these representations to generate new molecular graph structures.

Although alternative representations such as SELFIES have been proposed to guarantee syntactic validity in certain generative modeling contexts [563], SMILES was deliberately selected due to its seamless integration with the RDKit cheminformatics toolkit, which was used throughout this work for molecular validation and descriptor generation. All generated SMILES strings were rigorously filtered using RDKit prior to descriptor extraction, and any invalid structures were

discarded. This step effectively mitigates the primary advantage of SELFIES in terms of guaranteed validity while preserving compatibility with well-established descriptor calculation workflows.

The generative models developed here, GPT-2, W-GAN, and VAE, were pretrained on large SMILES-based datasets and leverage the syntactic patterns and structural regularities inherent in this representation [564]. Using SMILES therefore ensured compatibility with existing pretrained models, facilitated fine-tuning, and allowed direct use of models developed in our earlier work to evaluate and compare the desired primary KPIs, i.e. HSPs, which also relied on SMILES as the base molecular representation. Furthermore, other machine learning models developed and used to evaluate the greenness of the generated molecules/solvents in this study rely exclusively on graph-based and physicochemical descriptors derived from validated molecular structures, rather than on the linear string representation itself. In addition, SMILES remains the dominant representation in major chemical databases and cheminformatics platforms, which enhances the interoperability, reproducibility, and practical applicability of the generated solvent candidates. We fully acknowledge that SELFIES provide enhanced robustness during generation because every SELFIES string corresponds to a valid molecule. However, our incremental ensemble-learning strategy incorporates multiple validity checks (canonicalization, VAE and GAN filtering, auto-labeling, and expert review) that effectively mitigate invalid-generation issues even when using SMILES.

Nevertheless, the potential advantages of SELFIES in large-scale, unconstrained molecular generation tasks are recognized. In future work, a direct comparison between SMILES- and SELFIES-based generative pipelines will be conducted to assess their relative impact on exploration efficiency, structural diversity, and optimization performance. The methodology presented here will be compatible with SELFIES, with some modifications. Therefore, the integration of SELFIES-based generation is indeed planned for future extensions of this framework. However, for the descriptor-driven and validation-controlled framework developed and employed in the present study, SMILES represents the most appropriate and effective choice.

5.4.1 Case study 1: Carbon dioxide CO_2 solvents

The first case study demonstrates the proposed methodology in the context of discovering new solvents for CO_2 capture. A dataset of suitable solvents for CO_2 capture applications was constructed based on the *Hansen* solubility sphere concept [565], [566]. This involves a larger

dataset of about 12,000 unique solvent entries compiling several types of solvents along with their *Hansen* solubility parameters. This parent dataset was assembled from peer-reviewed literature sources and the *Hansen* Solubility Parameters in Practice (HSPiP) database. It was used in our previous work to build an accurate ensemble ML model for predicting HSPs based on the physio-chemical descriptors extracted from SMILES codes [543]. Each entry includes the solvent name, its canonical SMILES representation, and an associated set of thermodynamic and physicochemical properties. These properties comprise Antoine equation coefficients, normal boiling and melting points, flash point, vapor pressure, aqueous solubility, and *Hansen* solubility parameters (δ_D, δ_P, δ_H). These properties were used to support both the generative modeling process and the subsequent physics-based screening and validation of candidate solvents. This work benefited also from the previous experimental work published by two co-authors on the solubility quantification of diverse ligins and performance prediction [370]. To guarantee molecular uniqueness, we applied strict preprocessing procedures to the parent dataset prior to any analysis or model training. These steps included:

- Removal of invalid SMILES strings
- Canonicalization of all SMILES representations to ensure a unique structural encoding for each molecule; and
- Systematic detection and removal of duplicate entries based on canonical SMILES comparison

These procedures ensured that each molecule in the dataset was unique before model development. The values of R_a and relative energy difference (RED) were calculated according to the following mathematical relationships:

$$R_a = \sqrt{4 \times \left(\delta_D^m - \delta_D^{solvent}\right)^2 + \left(\delta_P^m - \delta_P^{solvent}\right)^2 + \left(\delta_H^m - \delta_H^{solvent}\right)^2} \tag{5.1}$$

$$RED = {}^{R_a}/_{R_o} \tag{5.2}$$

R_a is the radius of interaction between solvents and the material to be dissolved. δ_D, δ_P and δ_H are the three *Hansen* solubility parameters related to dispersion, polarization, and hydrogen bonding, respectively. R_o is the experimental sphere radius of the material (m) to be dissolved. The good or suitable solvents for the material of interest(m) must have an RED value of 1 or below.

The HSP values and R_0 of CO_2 used to calculate the R_a and RED are **15.6, 5.2, 5.8,** and **4**, respectively [481]. The resulting dataset was reduced to 3,180 solvents that are suited to dissolve CO_2. An excerpt of the cleaned dataset of the CO_2 solvents is presented in Table 5.1. This step was essential to provide clean and relevant data to the generative models, ensuring that new molecules are obtained from the same distribution while minimizing generation inefficiencies.

Table 5.1 CO_2 solvents dataset (SMILES codes)

SMILES Codes
CC(=NOC)C
CC(=O)OC=CC=C
CCOC(=O)CC(CC(=O)OCC)(C(=O)OCC)OC(=O)C
CC(OCC=C)=O
CC(=O)CC(=O)OCC=C
C=CCBr
CCOCC=C
C=CCOC=O
CC(C)OCC=C
C=CCS
CC(=C)C(=O)OCC=C
COCC=C
CC(CC1=CC=CC=C1)N
O=C(OCCCCC)C
CC(=C)C(=O)OCC1=CC=CC=C1
BrC(C)=C
CC=CBr
BrCC=C=C
C=CCCBr

5.4.2 Case study 2: Lignin solvents

The second case study focuses on the discovery of novel solvents for diverse lignins following the same methodology as the CO_2 case. Two lignin types were used:

- Softwood-based kraft lignin: HSPs and R_0 are **20.93, 16.98, 15.71, and 15** [370].
- Sugarcane bagasse-based lignin: HSPs and R_0 are **21.42, 8.57, 21.8, and 13.56** [479].

Excerpts of the final SMILES datasets of suitable solvents for both lignin types are presented in Table 5.2. The number of suitable solvents was 1,400 for bagasse-based lignin and 3,240 for softwood-based lignin.

Table 5.2 Softwood-based kraft lignin and sugarcane bagasse-based lignin solvents dataset (SMILES codes)

SMILES Codes	
Softwood-based kraft lignin solvents	**Sugarcane bagasse-based lignin solvents**
CC(N)=O	C/C=N/O
CC(=O)NC1=CC=CC=C1	CC(N)=O
CC(OC(C)=O)=O	CC(=O)NC1=CC=CC=C1
CC(C)(O)C#N	CC(C)(O)C#N
CC#N	CC(=O)OC1=CC=CC=C1C(=O)O
CC(C1=CC=CC=C1)=O	NC1=NC=NC2=C1N=CN2
CC(N(CCCCC1)C1=O)=O	OC(CCCCC(O)=O)=O
CC(=O)N1CCCCC1	CNCC(C1=CC(=C(C=C1)O)O)O
CC(=O)N1CCCC1=O	C=CCO
CC(=O)OC1=CC=CC=C1C(=O)O	NC1=CC=CC=N1
CC(C1=CC=CS1)=O	C(O)(=O)C1(CC1)(N)
CC(Br)=O	NC1=CC=NC=C1
CC(Cl)=O	NC1=CC=CC=C1
CC(=O)N1CCOCC1	COC(C=CC=C1)=C1N
C=CC=O	COC1=CC=C(N)C=C1
C=CC[N+]#[C-]	O=C1C(O)=C(O)[C@@H]([C@@H](O)CO)O1
C=CC(O)=O	NC(C1=CC=CC=C1)=O
C=CC#N	OC1=CC(O)=CC=C1
C=CC(=O)Cl	C1=CC=C2C(=C1)NC=N2
NC1=NC=NC2=C1N=CN2	C12=CC=CC=C1ON=C2

5.4.3 Evaluation procedures

Validation and novelty evaluation

To validate and assess the novelty of generated solvents, two methods were developed and used. A physics-based validation is performed using *RDKit* package [567] and *Chemical Identifier Search* web-based tool related to the *ChemSpider* open-access online database [568]. In parallel, data-based validation is conducted using the proposed LLM-based auto-labeling strategy. In this study, the auto-labeling models were trained on a 100,000-instance dataset composed of valid

SMILES from ChEMBL and synthetically generated invalid SMILES covering common error types. An 80/20 train-test split with 5-fold cross validation was used for model development and evaluation. Molecular validity and novelty rates were calculated using the mathematical formulas presented in **Eq. 5.3-5.4**.

$$Molecular\ Validity\ (\%) = \frac{Number\ of\ valid\ generated\ data}{Total\ number\ of\ generated\ data} \times 100 \quad (5.3)$$

$$Molecular\ Novelty\ (\%) = \frac{Number\ of\ unique\ generated\ data}{Total\ number\ of\ valid\ generated\ data} \times 100 \quad (5.4)$$

It is worth noting that the novelty was assessed primarily through exact-match verification based on canonical *InChIKey* comparison and *ChemSpider* lookup ensuring database-level uniqueness. This was complemented by a structural similarity analysis based on Morgan fingerprints and Tanimoto coefficients to quantify the degree of similarity between generated molecules and those in the training dataset.

KPIs numerical comparison

The valid SMILES codes were analyzed to extract key features, determine their properties, i.e. *Hansen* solubility parameters and RED, and compare them to the desired values. An ensemble-based ML model [543] was used to predict the HSPs of the new molecules based on *RDKit*-derived descriptors. The molecule selection criteria are based on a relative error of ≤ 5% between desired and the predicted HSPs. The RED must be below 1.

Synthesis routes prediction

Synthesis routes were predicted using multiple algorithms, including *Monte Carlo* simulation and Transformers-based modeling/simulation. *RetroTRAE* [569], *ASKCOS* [570], [571] and *AIZynthFinder* [572] were employed. These tools with different algorithms were adopted to provide insights for expert validation, and removal of invalid results, while enriching the knowledge.

Toxicity testing

The toxicity of the generated solvent was determined as an indicator of environmental safety. Toxicity indicators include *LC50* (lethal concentration for vapors/gases) and *LD50* (median lethal dose for dermal/oral exposure). Material toxicity is reversibly proportional to the values of *LC50*

and *LD50*, which means that solvents with higher *LC50* and *LD50* are safer. The *TEST* toxicity assessment tool was used to perform these calculations. Furthermore, Developmental Toxicity (D.T.) was evaluated. To automate this step, an ensemble-based model was trained on 11,000+ SMILES codes using *TEST* toxicity data, mapping molecular descriptors to *LC50 (continuous variable), LD50 (continuous variable),* and D.T. (binary variable). Regression models were used for *LC50* and *LD50*, while a binary classifier was used for D.T.

Flammability testing

Flammability is another indicator that characterizes the solvent safety and was assessed via Flash point (*FPt*). A high *FPt* correlated with lower flammability. Hence, an ensemble-based ML modeling method was developed and utilized to predict the flash point of each solvent based on its molecular descriptors. The results are compared and ranked to provide the solvent safety status. Flash points were determined from normal boiling points of the corresponding compounds and solvents using nonlinear equations recommended in [573], [574], [575] and [576]. The predicted flash points were then mapped to the molecular descriptors by tree-based ML models and vector regression (SVR) to achieve higher prediction accuracy of flash points.

Combined greenness score

To facilitate the selection and ranking of new solvents in each case study, a combined greenness score for individual solvents was created. This score is denoted as G_s^{Score} combining arbitrary scores for each item of toxicity and flammability tests. The subscript (s) refers to the solvent of interest. Scoring criteria are based on: *International Labour Organization* (ILO), the toxicity is categorized based on their *LC50* and *LD50* values with specified ranges; and *Occupational Safety and Health Administration* (OSHA) for which the flammability is categorized based on specified flash points ranges. Table 5.3 provides different toxicity-flammability categories and the scores given to each category. Scoring is application-dependent and determined by domain experts.

The following equation represents the mathematical formulation of the suggested greenness score:

$$G_s^{Score} = LC50_s^{Score} + LD50_s^{Score} + D.T._s^{Score} + FPt_s^{Score} \quad (5.5)$$

Both individual and combined scores will guide the AI-assisted generation of new molecules where thresholds will be set to ensure the training data compliance with the desired KPI(s). Consequently, the generative models will produce further new valid molecules achieving expert-specific KPI(s).

Table 5.3 Toxicity-flammability categories and their corresponding scores with average color ranking

*This table was Moved to Appendix D (Table D.1) in APPENDIX

5.5 Results and discussion

5.5.1 Data analysis

The parent dataset of all available solvents comprised about 12,000 solvents, with different types/classes including heterocyclic compounds, amines and alcohols. In particular, in this parent dataset, heterocyclic compounds constitute the most abundant solvent class, with over 3,500 entries. Amines represent the second most common class (approximately 1,800 solvents), followed by alcohols with around 1,200 entries. The full data analysis of the parent dataset which was presented in our previous work [543] is also available in supplementary materials (Appendix I).

In the first case study on CO_2 solvents, the dataset extracted from the parent dataset consisted of approximately 3,180 solvents. For existing CO_2 solvents in this dataset, most of them are heterocyclic compounds, alcohols, and carboxylic acids (Figure 5.4).

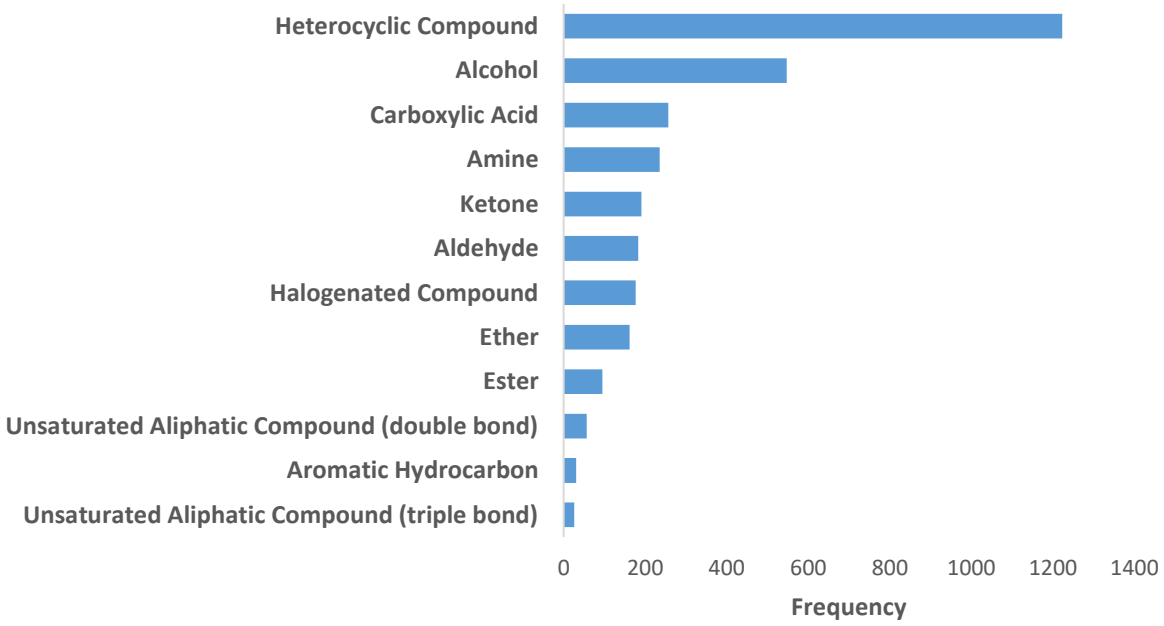


Figure 5.4 Distribution of CO_2 solvents

For softwood-based lignin, the solvent distribution is different from that of CO_2 solvents. Amine-based solvents were the most frequent, followed by heterocyclic compounds, alcohols, and ketones. Complete distribution data of selected lignin solvents are provided in the supplementary materials.

In terms of SMILES string specifications, the maximum and minimum character counts in the case of CO_2 solvents dataset were 83 and 3, respectively. For the bagasse-based lignin dataset, the maximum character length was 135 and the minimum remains at 3. These differences in distribution and string lengths across dataset motivated the adoption of case-specific model training (or re-training in the case of GPT-2) rather than using the whole parent dataset. This makes the model(s) more rational and generates molecules achieving the desired KPI(s) specific to each case. Each subset of solvents relevant to carbon capture and lignin valorization exhibits common characteristics, while differing at both broader and granular levels. Hence, fine-tuning or retraining the generative model on solvent subsets specific to a given application increases the likelihood of generating new molecules from the same distribution. This, in turn, enhances the probability of generating new molecules that satisfy the desired values of predefined KPIs. Conversely, fine-tuning on the whole parent dataset results in a broader and more heterogeneous distribution of generated molecules, covering wide ranges of solubility parameters and other KPIs.

High-level differences between datasets were identified through the abovementioned simple statistical analysis, while other low-level differences were elucidated using explainable AI (XAI). XAI provides interpretable insights into the relationships between molecular descriptors and the desired KPIs. For instance, molecules with higher MolLogP and VSA-EState8 and lower number of aromatic rings and TPSA are preferred as CO_2 solvents. Integrating this information along with human expertise incorporation during model training adds more rationality to the developed generative ensemble model.

5.5.2 Ensemble-based generation results

GPT-2 results

The developed and adopted GPT-2 model was configured with a learning rate of $5x10^{-4}$, maximum of 30 epochs, minimum of 15 epochs, 6 transformer layers, 12 attentions, zero weight decay, and a batch size of 16. The model utilized a sequence length of 512, a vocabulary size of 1,072, and an embedding dimension of 576 (12x48). *Adam* optimizer was used.

The model was pre-trained on a larger dataset including various complex molecules (from *PubChem* database) with larger characters count per string (data point) to understand the relevant factors for SMILES construction rules. This step provided the model with essential reasoning capabilities via a robust tokenizer. Then, case-specific datasets, i.e. datasets containing good solvents for materials of interest, were used for fine-tuning and training.

Table 5.4 summarizes the results, including the validity and novelty rates, alongside examples of generated SMILES codes from the first iteration of generation. The total number of molecules generated in each case is 100. The validity was completed using the combined human-/physics-based and data-driven auto-labeling methodology presented in **Sections 5.3 and 5.4**. The auto-labeling data-driven model was optimized, and its testing accuracy reached 99.5% where the F1-score reached 0.98. In addition, a further validation step for the developed auto-labeling model was performed using a dataset of about 12,000 labeled SMILES. The validation accuracy of this classification model achieved 99% accuracy. For lignin-related solvents, generated molecules were more structurally complex, reflecting the complex nature of their respective datasets. Most of the generated molecules considered non-original were not present in the input dataset for each case, resulting in near 100% originality from a generative AI standpoint. This is an industrial advantage to identify novel solvents with potentially improved solubility and reduced toxicity.

Table 5.4 GPT-2 generation results

*This table was Moved to Appendix D (Table D.2) in APPENDIX

W-GAN results

Each SMILES code in the available datasets is converted into a graph format using adjacency and feature matrices (vectorized image form) to train the adopted W-GAN model. The adjacency matrix captured the atomic connectivity and bond types. The feature matrix represents the number of atoms with their types for each code. The dataset involved 12 unique atom types and 4 bond types. The W-GAN model was configured with a learning rate of $5x10^{-4}$, a batch size of 16, 10 epochs, and a dropout rate of 0.2 for both generator and discriminator. The generator used dense layers of 128, 256, and 512 units with *Tanh* activation function. The discriminator used ReLU activation function with a dense layer of 512 units. The graph convolution layers were uniformly set at 128units.

For CO_2 solvents, the validity rate was above 90%, but the novelty rate was below 5%. As observed, the rate of validity is higher than that of GPT-2; whereas the rate of unique molecules is significantly lower. This is because GANs need a large amount of training data to perform well, which was not available in this case due to the initial positive discrimination of the dataset. Moreover, the molecules generated by the W-GAN model tend to be simpler (having shorter SMILES sequences and lower complexity) compared to those generated by GPT-2. Similar trends were observed for bagasse lignin and sugarcane solvent datasets.

Convolutional VAE results

Like the W-GAN, SMILES codes were converted from the string to the graph format to train the VAE model. Hyperparameters of the adopted convolutional VAE included a learning rate of $5x10^{-4}$, a batch size of 100, and 10 epochs. Encoder and decoder activation functions were *ReLU* and *Tanh*, respectively, with dropout rates of 0 and 0.2. The *Adam* optimizer was used. The average validity rate was 60%, and the novelty rate was 5%, i.e. the W-GAN model generated molecules with significantly shorter SMILES sequences and reduced structural complexity. Among the three models (W-GAN, VAE, and GPT-2), GPT-2 emerged as the most effective option in achieving balanced validity, complexity, and uniqueness, while requiring moderate computational infrastructure. As described in the methodology, the results of the three generative models were concatenated into a unified list for further testing and evaluation, each weighted equally.

5.5.3 KPIs numerical comparison

Previous studies reported an ensemble ML model capable of predicting the HSPs of various solvents from their most significant molecular descriptors, achieving nearly 99% accuracy [543]. These HSPs compute the Relative Energy Difference (RED) values to differentiate between “good” and “bad” solvents (considered as solvent labels). Table 5.5 presents the generated molecules and their corresponding HSPs, REDs and labels. Among valid molecules generated by GPT-2 model, approximately 70% were labelled as “good” for CO_2 solvents. The rate was 57.5% and 35% for softwood lignin and bagasse lignin, respectively. For CO_2 solvents, the W-GAN and convolutional VAE generated 35% and 5% “good” solvents, respectively, achieving the desired KPIs) among the valid molecules. It is important to note that the generative models were first pretrained on large and diverse datasets covering a wide chemical space, enabling them to learn the syntax of valid SMILES codes and related chemical patterns. After this pretraining step, the models were fine-

tuned on the datasets containing good solvents in each case. This fine-tuning strategy focuses the model on the region of chemical space associated with high performance for the chosen application. When the parent dataset of approximately 12,000 various solvents that combined both good and poor solvents was used, the fraction of "good" solvents identified for each case or material of interest did not exceed 5% because the model learned from a mixed-quality target distribution. This confirmed the advantage of using only high-performing examples during the fine-tuning stage which significantly improves the ability to discover solvents tailored to the requirements of a given system, while still preserving chemical generality through the diverse pretraining phase.

In our previous work [543] we discussed the key chemistry-based aspects that influence the solubility and determine why certain solvents are more effective for specific solute. Building on those findings, we were able to select the datasets of good solvents used in the present case studies(CO_2 capture and lignin dissolution). To strengthen the interpretation of our results, we further evaluated molecular descriptors for the newly generated SMILES structures. These analyses confirm trends consistent with our earlier study [543]. Where molecules with lower values of 6^{th}-order electro-topological state of a fragment within a molecule (VSA_EState6) and lower number of heteroatoms in the solvent's molecular structure, both associated with reduced dispersion solubility parameters, tend to exhibit higher CO_2 solubility (Table 5.5). In contrast, molecules with higher TPSA and a greater number of aromatic rings, combined with lower MolLogP (resulting in higher hydrogen-bonding and dispersion solubility parameters as shown in **Table 5**), are more favorable for dissolving lignins, especially softwood kraft lignin.

Table 5.5 Predicted HSPs, REDs and labels of generated molecules sample

*This table was Moved to Appendix D (Table D.3) in APPENDIX

The generation can be guided toward better solvents by adjusting RED thresholds to be lower than one (e.g. the threshold can be set at 0.8). This may reduce training dataset size. The threshold should be carefully optimized to balance having enough training data by effectively guiding the generative models toward better solvents.

5.5.4 Retrosynthesis results

Retrosynthesis analysis was conducted for three selected different generated molecules (SMILES code), each representing a case study (i.e. CO_2, softwood lignin, and bagasse lignin). The

precursors or reactants were identified using *ASKCOS* tool. The first molecule is related to CO_2 solvents generation and has the SMILES code of "CCOCCC(CC)OCC". The second molecule was "O=C(O)CC(O)C(COC)O", which was considered a "good" solvent for softwood lignin. Finally, a generated molecule that can dissolve bagasse-derived lignin was "OCCNNCO". Figure 5.5 shows the suggested synthesis routes for those three molecules with the highest average plausibility scores (≥ 0.95).

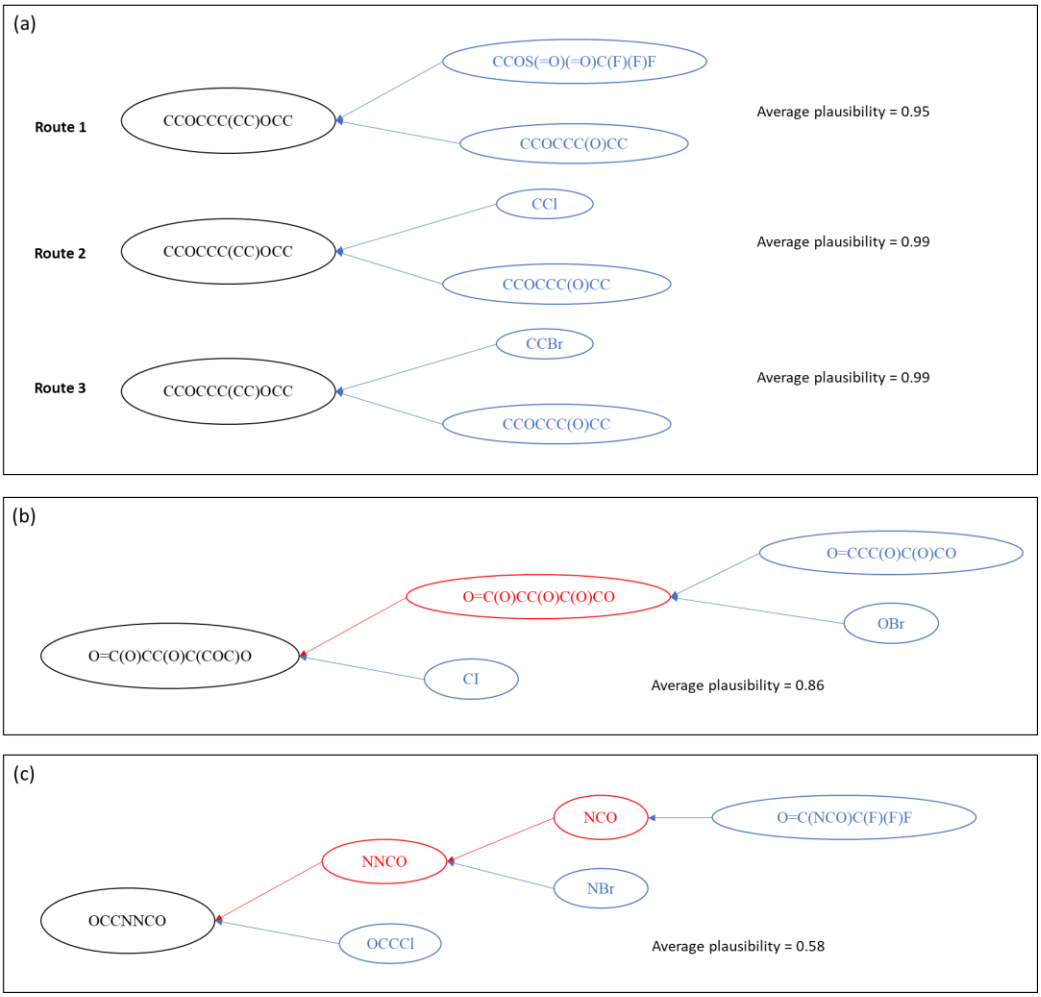


Figure 5.5 Synthesis routes for selected molecules of (a) CO_2 solvent, (b) softwood lignin solvent and (c) bagasse lignin solvent

These results provide experts with actionable insights to inform lab-scale and industrial-scale chemical synthesis.

5.5.5 Solvents greenness testing results

The developed ensemble-based ML model for estimating the flash point (solvent flammability) reached a testing determination coefficient (R^2) of 0.98, using the molecular descriptors determined by an optimized non-negative matrix factorization technique [543]. The ensemble model included XGBoost (with 200 trees), Random Forest (with 100 trees), and SVR (with RBF kernel, C-parameter = 300).

Toxicity metrics (*LC50* and *LD50*) R^2 values of 0.945 and 0.9, respectively. Toxicity classification reached 92% accuracy with an F1-score of 0.915. Table 5.6 summarizes the results of toxicity and flammability testing of the generated molecules ranked using a color-code system. *Light green* indicates that the material is potentially safe, with values well below the toxicity threshold. *Dark orange* signals that the material is potentially unsafe, exceeding the threshold, as detailed in Table 3. The higher greenness score corresponds to a darker green color, indicating a more environmentally favorable molecule. The yellow color means that the material is potentially safe and crossed on the threshold but with relatively low value. The lightest green color reflects materials that are within the safe range, with values well below the threshold. The lighter orange color suggests lower toxicity or flammability, although some level of risk still exists.

Table 5.6 Toxicity and flammability results (Color-based Ranking)

*This table was Moved to Appendix D (Table D.4) in APPENDIX

5.5.6 Role of incremental learning

In the early stages of modeling, generative models may face "***hallucination phenomenon***", meaning that molecules are potentially unstable or chemically impossible in real conditions or their structures do not align with established chemical rules. In addition, the models can produce some new data/designs with undesired properties. To address this, our implemented incremental learning-based approach offers a solution to this problem and guides the generation to preferable results. The models are retrained after updating the datasets with the new solvents that achieved validity, novelty, and the desired KPIs. Following the second retraining and fine-tuning trial, we observed significant improvements: molecular validity rate reached 71.30% and molecular novelty rate enhanced to 94.39%. Another trial of fine-tuning was conducted based on the results of the first generation runs where the weights of the ensemble approach were changed. GPT-2 displayed

the strongest balance between novelty and structural coherence, whereas the W-GAN and VAE models contributed to more conservative but chemically realistic structures. As a result, GPT-2 was assigned a higher weight (0.9) in this subsequent round, while W-GAN and VAE were down-weighted. The fine-tuning dataset was updated by the generated SMILES codes following the new weights. With these adjusted weights, the later generation round achieved substantially improved molecular validity reached 94% and molecular novelty rate enhanced to 98%. This adaptive weighting strategy therefore supports the effectiveness of the incremental ensemble learning approach and demonstrates that the model combination was indeed optimized beyond the initial equal-weight assumption.

5.6 Conclusion

This study presented a novel incremental, ensemble-based, physics-informed generative deep learning methodology, guided by human preferences and expertise. It demonstrated strong capabilities in accelerating the discovery of greener solvents for CO_2 capture and lignin valorization. In the first iteration, the ensemble-based model achieved a molecular validity of 60% and a molecular novelty of 90%. These indicators improved to 71.3% and 94.4%, respectively, in the second iteration, underlying the effectiveness of incremental learning. The model training required less than one hour, while inference ranged from 5 to 10 seconds. The greenness of generated solvents was enhanced by minimizing their toxicity and flammability risks. Particularly, the solvents generated exhibited *LC50* values exceeding 1,500 mg/L and flash points above 150 ^{o}C. Plausible synthesis routes were estimated by physics-based and data-driven combined techniques, guiding future experimental validation. These results confirmed that the proposed methodology significantly accelerates material discovery while minimizing experimental and computational costs.

Future work will further advance this methodology by incorporating end-to-end active learning frameworks to streamline molecule design cycles and accelerate convergence toward optimal solvents. More advanced generative architectures will be investigated to improve validity, originality, and robustness of both molecule generation and auto-labeling. Promising directions include hybrid models such as Generative Adversarial Network (GAN)-based Variational Autoencoders (VAEs), GPT-GAN architectures that combine Generative Pre-trained Transformers (GPTs) with adversarial training, and Retrieval-Augmented Generation (RAG) enhanced by

Mixture of Experts (MoE). These approaches are expected to improve exploration of chemical space, reduce bias from limited training datasets, and enable more reliable discovery of molecules with properties aligned to target application requirements. Furthermore, coupling these generative AI models with chemistry- and physics-based simulations and high-throughput experimental validation will be ensuring that the generated solvents are not only novel and chemically valid but also experimentally viable for applications such as CO_2 capture and lignin valorization.

Ensuring a continuous supply of high-quality representative datasets, including both valid and invalid molecules, will support sustained incremental learning. Human expert supervision will remain essential for dataset curation, model fine-tuning, and validation. The integration of reinforcement learning (RL) in future research is anticipated to further optimize space exploration and adapt the discovery process based on lab-scale feedback. In parallel, quantum-inspired optimization techniques will be explored to improve hyperparameter tuning efficiency while minimizing the energy and emissions footprint of training and inference.

On the application side, the proposed methodology will be extended to support the generation of new sustainable conceptual designs for industrial processes.[3]

[3] Supplementary materials of this chapter are available in (Appendix I)

CHAPTER 6 ARTICLE 4: SIMULATE INTELLIGENTLY: CAUSAL INCREMENTAL REINFORCEMENT LEARNING FOR STREAMLINED INDUSTRIAL CHEMICAL PROCESS DESIGN OPTIMIZATION

Eslam G. Al-Sakkari (Eslam Ibrahim), Ahmed Ragab, Mohamed Ali, Hanane Dagdougui, and Daria C. Boffito

Article Submitted on August 13, 2025 and Published (online) on November 6, 2025 in Journal of Environmental Chemical Engineering, Elsevier, Volume 13, Issue 6, 2025, Article Number 120167

Online Publication Date: November 2025

DOI: https://doi.org/10.1016/j.jece.2025.120167

6.1 Abstract

Achieving efficient industrial decarbonization and net-zero emissions requires the widespread adoption of cost-effective and eco-friendly carbon capture technologies. This study presents a hybrid simulation-based optimization methodology integrating physics-based models with data-driven learning, under expert supervision, to optimize carbon capture process design. A causal incremental reinforcement learning (RL) agent is developed to search for optimal design variables within a predefined space. To enhance decision-making, formal concept analysis, Granger causality, and causal Bayesian networks are used for informed causality analysis. The RL agent is guided by this analysis to improve key performance indicators (KPIs), i.e. capture cost and total relative emissions. The proposed method is applied to two case studies: monoethanolamine (MEA)-based and dimethyl ethers of polyethylene glycol (DEPG)-based carbon capture processes. For the MEA process, the methodology achieves a capture cost of 41.3 USD/ton CO_2 representing an 8.14% reduction from the 44.96 USD/ton baseline and total relative emissions of 0.28 kg CO_2/kg CO_2 captured, with optimal design variables including 3 bar absorption pressure, 25.5 °C flue gas temperature, 31 °C lean MEA temperature, and a boil-up ratio of 0.087. In the DEPG case, optimal values are 7.5 bar pressure, 4.5 solvent-to-feed molar ratio and 9 absorption stages, yielding a capture cost of 44.5 USD/ton CO_2 and emissions of 0.00071 kg CO_2/kg CO_2 captured. All extracted causal graphs, dependencies, and design rules are stored in a unified, updatable knowledgebase to enable transfer learning and generalization to other industrial processes.

6.2 Introduction

Finding cost-effective and eco-friendly carbon capture technologies as one of the promising industrial decarbonization solutions is crucial to achieve *Net-Zero* goals [577]. One way to achieve this goal is to efficiently design the capture process. It is worth noting that the optimization of control and monitoring schemes of the existing industrial processes has a significant contribution to improving their performance in terms of total cost and emissions. However, there is a new trend to focus on the design or the re-design of the industrial chemical processes in order to improve their techno-economic performances and to achieve their sustainability [578]. The proper optimization during the first stages of conceptual design (process synthesis) and process engineering (process development, feasibility analysis and optimization) can lead to above 30% cost reductions. Various optimization approaches based on mechanistic models, often combined with traditional operations

research techniques such as linear and nonlinear programming, are commonly used to enhance carbon capture processes. While these methods are robust, they have notable limitations, including high computational costs, significant time requirements, and a lack of learnability. These constraints can impede the rapid advancement of carbon capture technologies, which is crucial given the urgent need to mitigate global warming and climate change.

To address the limitations of traditional process optimization methods, there is a pressing need for innovative approaches. Artificial intelligence (AI) and machine learning (ML), particularly reinforcement learning (RL), offer promising solutions due to their ability to function as learnable optimizers. RL has already demonstrated success in various domains, including process control [29], [579] highlighting its potential for accelerating chemical process design. RL algorithms can be categorized into model-free and model-based approaches, as well as policy-based and value-based agents. The policy-based agents/algorithms can be further categorized to on-policy and off-policy agents/algorithms. Notable examples include deep Q-networks (DQN) [580] and proximal policy optimization (PPO) [581] as policy-based methods, while deep deterministic policy gradient (DDPG) [582] combines both policy-based and value-based strategies.

Recent research has explored the use of reinforcement learning (RL) in chemical engineering to optimize process sequences and generate process flow diagrams for relatively simple systems [583]–[589]. For example, a hierarchical deep RL agent was used to automate flowsheet synthesis for ethyl tert-butyl ether (ETBE) production [587] using the *SynGameZero* approach [590]. Similarly, the Soft Actor-Critic (SAC) algorithm was applied to design a distillation train for separating benzene, toluene, and p-xylene by interacting with the Distillation Gym environment, which interfaces with the COCO open-source simulator [586]. Other studies have surveyed RL-assisted flowsheet synthesis, highlighting interactions with simulators like DWSIM [591] and the potential of transfer learning for process design [585].

Despite these advancements, many RL-based approaches remain in early development stages. A key challenge is the generation of technically infeasible flowsheets due to the lack of reasoning in RL agents. Additionally, existing studies largely overlook causality, which is crucial for process optimization. By analyzing historical optimization data, process designers can identify root causes affecting KPIs, such as cost. Integrating causality analysis enhances result generalizability, uncovers hidden relationships, and enables knowledge transfer to improve other processes.

This paper is an extension of our previous work [592]. Here, we give full details of our developed methodology about causal incremental RL, while focusing on the improvements added to the previous methodology to enhance the causality analysis framework. The main goal and novelty of this current study, as well as the previous one [592], is the introduction of causal incremental deep RL (CIRL) and its adoption for industrial chemical processes design acceleration, with special emphasize on carbon capture processes. Accordingly, two common carbon capture systems were considered for illustration and validation: 1) Selexol process as a representative of physical absorption processes and 2) mono ethanol amine (MEA)-based carbon capture process as a representative of chemical absorption. In this extension, we discuss the application of our method for multi-objective optimization, where these objectives can compete to achieve the optimum scenario/process. Particularly, here we add the total CO_2 emissions per kg of captured CO_2 as an objective to be minimized alongside the total cost per kg of captured CO_2. This addition plays a key role in the advancement of process design procedure, where the expert can evaluate the process from technical, economic and environmental point of views simultaneously. This indeed will accelerate the design process to achieve process sustainability. In this regard, we designed new reward formulations for more acceleration and adaptation to achieve the desired KPIs. All the changes and improvements in the reward formulations are presented in detail in ***Section 3***.

As previously mentioned in our first paper [592], introducing the causality analysis concept to the simulation-based optimization process design field increases the intelligent agent's rationality/reasoning. It also upgrades the process design expert knowledge with extra useful insights. In this essence, we focus on making our developed methodology generalizable to be applied for other industrial processes under the full supervision of human experts, including data scientists and process engineers, for design process acceleration as well. The detailed causality analysis methodology, which is added to the RL agent to have reasoning during the optimization process, and its novelty here is summarized in the following points:

- Combination between formal concept analysis (FCA) and *Granger* causality (GC) for preliminary extraction of cause-effect relations
- Construction of well-informed directed acyclic graphs (DAGs) with the aid of the abovementioned combination to build informative causal *Bayesian* networks (BNs) for robust causal discovery and inference

- Conversion of the obtained results (causal graphs) to a more readable forms by the process experts such as empirical equations or easier to use such as causality-informed ML models.

In here, compared to our first paper, we introduce FCA as a key component along with IML+XAI and GC to correctly construct the DAGs and remove the unnecessary connections between different variables where they do not affect each other. These DAGs will be then used to develop the causal BNs to quantify the effects of each variable on the target output or even the cause-effect relationships between input variables themselves.

The rest of the paper is structured into 5 other main sections. ***Section 6.3*** summarizes selected related works previously published in literature with a gap analysis showing the importance of our new methodology and investigation. ***Section 6.4*** presents the details of our methodology about the development of causal incremental RL (CIRL) for accelerated process design. Carbon capture-based case studies employed for validating our methodology are summarized in ***Section 6.5***. The results of applying our CIRL methodology on the two case studies are presented and discussed in ***Section 6.6***. Both ***Section 6.7*** and ***Conclusion*** sections give an overview of the most important findings, general remarks/recommendations and proposed future works.

6.3 Related Work

This section discusses the previous work to optimize the carbon capture process which is an essential industrial decarbonization chemical process. As previously mentioned, the modeling and optimization on different levels is done via different mechanistic and data-driven techniques [366], [593]–[596]. The majority of these efforts were reviewed previously in recent studies [121], [156], [597]–[600]. Particularly, in our previous perspective paper [156], we mentioned several key studies where we discussed the modeling, simulation and optimization methodologies including derivative-based and derivative-free/black-box optimization applied for case studies related to carbon capture process operation and design. We also introduced our perspectives on the best pathways to optimize process design using AI/ML, mainly three pathways. One of them was the development of surrogate models for fast decision making and accelerated hyperparameters optimization of RL agent when dealing with very complex processes. The final phase of our perspective is to link Aspen HYSYS (or any physics-based simulation environment) directly with the RL agent motivated by physics-guided optimization, acceleration and ease of transfer learning. Fortunately, some recent studies successfully developed surrogate models representing various

processes including carbon capture and utilization and coupled them with different optimization frameworks [601]–[605]. These optimization techniques includes both derivative-based and derivative-free ones [606]–[608] where some of the studies tried successfully to link genetic algorithm (GA) directly with Aspen HYSYS simulation [609]. While depending on the surrogate models offers a good way for the acceleration, the accuracy of these models must be very high to give reliable solutions and to be representative of the simulation. They are also case-specific, where each model represents a very specific simulation case, any change should be represented by another model. In the case of direct interaction, it will be easier to make any changes in the simulation case, and the trained RL agent will work on the optimization smoothly. It is worth noting that a validation step should be performed after having optimal design variable via surrogate models where there is a probability that these values are not physically reliable. Table 6.1 summarizes selected studies considering the modeling and optimization of solvent-based carbon capture with special emphasize on the data-driven-techniques, i.e. statistical-based and AI/ML-based. Most of the AI/ML-based studies focus on the predictive abilities of different techniques without investigating the other descriptive or prescriptive abilities. The modeling and optimization in this case are done in two separate steps where the data-driven model is firstly built then an operation research (OR)-based method is employed for optimization. The investigation of the causal relationships between various independent and dependant variables in these studies is not considered. In addition, in the case of employing traditional operation research-based optimization techniques, the optimizer is not learnable unlike the case of reinforcement learning (RL)-based optimization [610]–[613]. One of the advantages of RL-based optimization is that both modeling and optimization is done simultaneously, which accelerates the design process.

Causality analysis is a novel and crucial part in our methodology where we provide the design experts with different knowledge forms, e.g. interpretable rules and knowledge causal graphs. This will accelerate the process design and enrich the experts' knowledge, where they can transfer this knowledge to other processes/use cases. The development of RL agent to interact with the simulator directly accelerates the design process, where the search space is narrowed, unlike the case of grid search-based optimization, where all the defined search space should be fully investigated then optimization is done through various methods. Furthermore, during the interactive simulation-based optimization using RL, the agent learns the policy to reach the optimal design in shorter times compared to the traditional optimization routes. The following section

presents in detail the proposed novel methodology for simulation-based optimization of process design through causal incremental RL.

Table 6.1 Selected studies on data-driven-assisted solvent-based carbon capture process design/control optimization

*This table was moved to Appendix E (Table E.1) in (APPENDIX)

6.4 Methodology

Figure 6.1 illustrates the general approach of the current study, where causal incremental reinforcement learning is developed and employed for process design optimization. The industrial process design optimization is done through an interactive hybrid modeling approach under the full supervision of human experts. In particular, the experts feed the digital twining environment (industrial process digital twin) with their expertise in simulation, models' development & optimization and data engineering. Then, their knowledge is updated continuously with the results extracted from this environment in the form of optimal designs, database and knowledgebase. One advantage of this interactive method is the transferability of this knowledge to other industries. Besides, the causality analysis is a core item in this digital twin, where it gives the Simulation-RL interaction part the insights and updates needed for further improvement in the optimization process. The following subsections give more details on the different parts of the digital twining environment.

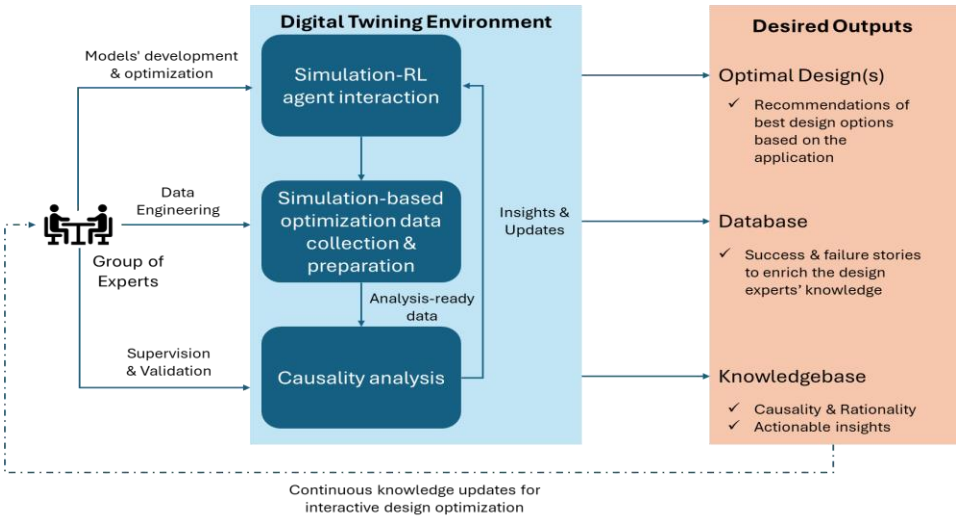


Figure 6.1 General approach of causal incremental reinforcement learning for process design optimization

6.4.1 Simulator-RL Agent Interaction

Figure 6.2 introduces a simple and informative illustration of the interaction between RL agent and the process simulator where possible actions, observations, and rewards are clearly mentioned. For instance, in the case of carbon capture process, the possible actions are the operating conditions (temperature, pressure and flowrate) and design variables including equipment sizing. The possible observations that define the state of the process include product streams purity and energy consumption. Rewards can take several forms as functions of different KPIs, such as capture cost or net emissions or combined KPIs, such as combining capture cost with product purity.

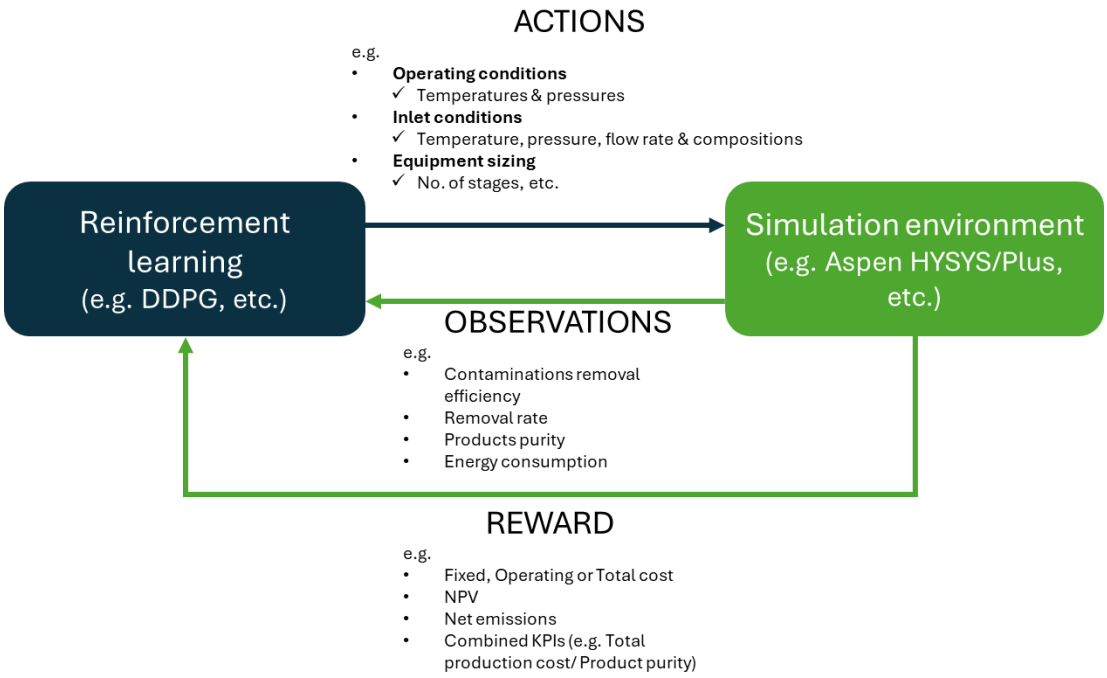


Figure 6.2 Industrial process simulation and RL agent interaction

Selecting the RL algorithm is one of the important steps and it depends on the problem in hand [614]. For instance, in the case of stochastic process optimization, soft actor-critic (SAC) is preferred; whilst DDPG is more applicable in the case of deterministic process optimization. The mathematical background of SAC can be found in [615], [616] whereas the details of DDPG is in [582], [617], [618]. It is worth noting that RL is closely related to Markov Decision Process (MDP), in particular, MDP provides the mathematical framework for RL for decision-making/optimization problem modeling. Hence, it is important to present the process design optimization problem formulation as MDP for better understanding of the RL optimization procedure and results

afterwards. MDP is commonly defined by a tuple **(S, A, P, R, γ)** where *S* is the set of states, and *A* is the actions' set, *P* is the transition probability function that defines the probability of the system's transition from state s_t to state s_{t+1} after taking the action a_t, *R* is the immediate reward received after performing the action a_t, and γ is the discount factor that defines the importance of the future rewards and takes values between 0 and 1. MDP is essentially designed to formalize the sequential decision-making problems. This is obviously applicable in the case of control (operation) problems, where the systems are in the dynamic mode. However, in the early stages of industrial chemical process design, the system is typically treated as steady-state, ignoring process dynamics and temporal variations (i.e., single-step optimization). It should be noted that the state of the industrial system in the case of process design is deterministic and can be mapped to a certain set of actions (each actions' set leads to a unique state regardless the initial state). It is worth noting that these states are very important as their evolution reflects the progress of the agent and quality of solutions/actions taken gradually. Therefore, to consider MDP to formalize this bandit-like problem and to introduce RL for its optimization, adaptations should be done while taking into consideration some essential assumptions or approximations. One of the important assumptions taken into consideration in this study (design optimization of steady state industrial systems) is that the current state of the process is only affected by or related to the previous state and affects the next state; where the current state guides the RL agent (that learns a direct mapping from states to actions) to the next state that is achieved through taking the right actions' set towards system's optimality. This is perfectly aligned with the memoryless property of MDPs. This assumption is quite valid and is completely achieved when the RL agent learns the deterministic policy to reach the optimal set of actions. As the processes considered here are fully deterministic and the actions' space is continuous, then the selected RL agent is DDPG. Any discreate actions can be adapted and converted to a special continuous form to be optimized through DDPG. In other cases, multi agents can be considered, where DDPG is considered for continuous variables and DQN is adopted for discrete variables optimization. Following is the summary of the adaptations made to formulate the problem as MDP and solve it via DDPG RL agent; where being learnable is the motivation of the adoption of this RL agent for the optimization of industrial process design.

In the case of optimizing the design of steady state industrial chemical process, the immediate rewards are more important to be considered instead of the future rewards. Accordingly, the discount factor **γ** will take the value of 0. The future reward term and *Bellman* expectation equation

for State-Action value function (***Q-function***) can be simplified to be described as **Equation 6.1**. The objective is to find the optimal policy that maximizes this immediate deterministic reward ***R(s, a)*** in the ***Q-function***.

$$Q(s,a) = R(s,a) \tag{6.1}$$

The actor network of DDPG in this case will learn a deterministic policy to identify the optimal action for each state through mapping the states to the specific actions that maximizes the desired reward. On the other hand, the critic network evaluates the state-action pair's ***Q-value*** to provide the feedback on the quality of actor network's suggestions of new actions. The critic network is continuously updated with these results using the simplified ***Bellman*** equation (**Equation 6.1**). After updating the critic, the actor network is updated through the feedback of the critic to adjust its policy and to give better actions gradually in order of achieving optimality. It is worth noting that, the exploration is performed through adding noise to the agent actions in order to escape local optima. *Ornstein-Uhlenbeck* (OU) noise was considered in this study, where it samples the noise from a correlated normal distribution.

The transition probability function ***P*** in this case, where each set of actions leads to a unique state regardless the starting state, is deterministic with no randomness and will take the following form:

$$P(s'|s,a) = \begin{cases} 1 & if\ s' = f(a) \\ 0 & otherwise \end{cases} \tag{6.2}$$

This means that the probability of transitioning from state *s* to the next ***s'*** depends on giving a unique set of actions ***a*** that leads to this unique ***s'***. In this case the probability is **1**. On the other hand, the probability is **0** to transition to any other state other than the specified unique state ***s'***.

Regarding the reward definition, it can take several forms. For instance, in our previous study [592], we followed the philosophy of all or nothing. If the taken action led the system to a state where it achieves a certain KPI value satisfying a predefined stopping criterion, then the agent takes a constant reward. If not, the agent takes no reward. In this study, we are proposing new philosophy, where we include penalties and rewards proportional to the distance between the achieved KPI and predefined value in the stopping criterion.

More details on the case studies considered, actions, states, the problem dimensionality in each case, and the reward definition are presented in **Section 4.** The full description of the industrial processes and their implementation using the selected process simulator, i.e. *Aspen HYSYS*, are introduced as well in the same section. It is worth mentioning that in this study the most needed resource for the RL agent training is CPU where we did not apply GPU acceleration. In addition, the framework used to build the DDPG RL agent is TensorFlow (2.9.1) where we worked on Python (3.9.7) and interfaced with Aspen HYSYS V14 via COM automation.

To ensure reproducibility, two different PCs with different specs were used to perform the causal RL-based optimization where the random seed was fixed. No GPU acceleration was adopted in both cases. The following are the specs of the two PCs:

- The first PC has the specs of Processor: Intel(R) Core (TM) i7-10750H CPU @ 2.60GHz 2.59 GHz, Installed RAM: 16.0 GB, Graphics Card: NVIDIA GeForce GTX 1650 with Max-Q Design (4 GB), Intel(R) UHD Graphics (128 MB) and System Type: 64-bit operating system, x64-based processor.
- The second PC has the specs of Processor: 11th Gen Intel(R) Core (TM) i7-1185G7 @ 3.00GHz 1.80 GHz, Installed RAM: 32.0 GB (31.7 GB usable), Graphics Card: Intel(R) Iris(R) Xe Graphics (128 MB) and System type: 64-bit operating system, x64-based processor

In the harshest cases the maximum CPU utilization was 50% and the memory usage was 30%.

6.4.2 Data collection and preparation

During the optimization, the data produced are stored in spreadsheets to be further investigated to extract the causal relationships and provide them to the expert and the agent for further possible optimization. Figure 6.3 illustrates the essential steps followed to collect these data and to make them ready for analysis. As shown, the collection step is automated, and the database is continuously updated with all the successful and failed optimization attempts during the exploration-exploitation. Data scaling through standardization or normalization is essential to have meaningful causal graphs afterwards. Human experts' role is inevitable also in this step, where features selection/engineering heavily rely on their expertise and rules.

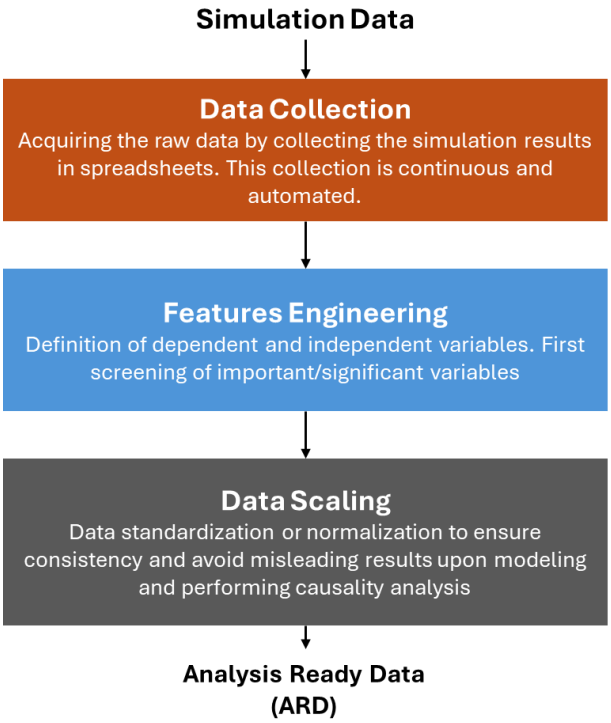


Figure 6.3 Data collection and preparation for causality analysis

6.4.3 Causality analysis

The general methodology to extract causal knowledge to help enhancing the reasoning of the developed RL agent is depicted in Figure 6.4. As shown, it consists mainly of three phases. The first one is a preliminary investigation of the relations between dependant and independent variables. This can give a guidance to the causal discovery and inference step. In the second phase we explore more about the causal relations through the development of more specialized techniques in causal discovery fed with the information extracted in the first phase. The third phase focuses on developing some empirical equations or relations that can be more readable/usable by process experts to assess the variables' dependencies and interrelationships. Following is the detailed explanation of each technique in the three phases.

First phase: Preliminary analysis & guidance

Four distinct approaches are developed and utilized in this phase. These approaches are interpretable ML (IML), explainable AI (XAI), formal concept analysis (FCA), and Granger Causality (GC).

For instance, IML, e.g. decision trees (DTs), gives interpretable rules and variables importance. To perform this analysis, a data preparation step is done to facilitate the classification and obtaining the patterns (rules). In particular, the output KPIs of interest are given certain labels, where the classification is done based on them. For example, if we have the total cost of the process as the single output KPI, then a threshold can be assigned to differentiate between two classes representing the data. These two classes will be "**low cost**" that will take the label of **1** and the other class will be "**high cost**" having the label of **0**. Based on this preparation, the IML technique can classify the data correctly and give the patterns or rules in the form of tables easy to be read by the process experts. Besides, IML gives an indication of the relative importance of different variables through its embedded filtering and features selection mechanism.

In addition, XAI is responsible for enhancing the results related to the variables importance. It is performed through adding explainability technique, e.g. SHAP or LIME, to the predictive ML models, which help avoiding black-box analysis. It gives the magnitude and sign of the importance of each variable not only the relative importance scores as in the case of IML. A key aspect that should be taken into consideration is the high accuracy of the developed predictive ML model. This should be done to ensure representative XAI results and to guarantee obtaining the real-world relations between the input variables and the targets (KPIs). The combination between IML and XAI was previously introduced for knowledge augmentation in one of our previous works [471]. These two approaches are giving preliminary indications about the interrelations between dependant and independent variables based mainly on their correlations and predictive power of the independent variables. Hence, they can be considered as weak yet informative guides to causal discovery but cannot be used solely for the causality analysis. That is the reason why we considered the development of other techniques illustrated below.

GC is employed to confirm the preliminary causal dependence of targets on the independent variable besides the effect of independent variables on each other. It gives insights about the direction of these dependencies by testing the dependence hypothesis through evaluating the ***p*-value** in both directions. The variables can be affecting each other if the ***p*-value** is lower than 0.05 while testing the hypothesis in both directions. If the value is lower than 0.05 in one direction but higher than this value in the other direction, this means that only one variable is affecting the other one while being totally independent. GC is more preferred for analyzing time series or stationary data [619]. In this regard, the data obtained from the optimization trials is assumed as a simple time

series. This is due to the dependence of consecutive states on each other (according to the *Markov* memoryless property assumption mentioned in **Section 3.1**) and the same applies for the actions. The *Augmented Dickey-Fuller* statistical test was performed to support this assumption/approximation. Besides investigating the raw dataset obtained from the simulation runs history, this dataset was transformed to a sort of time-series form through applying *Differencing* methodology to help perform the GC testing. The optimal lag selection was performed through *Vector Autoregressive* (VAR) modeling based on both *Akaike Information Criterion* (AIC) and *Bayesian Information Criterion* (BIC).

Regarding FCA, it is basically used to find hidden structures in complex datasets through uncovering new patterns and relationships in these datasets. This can lead to the discovery of new insights helpful for human experts to improve their knowledge and consequently the processes they optimize. In this context, it is more directed to qualitatively define the hierarchical relations between different variables in each dataset. It is more preferred to model categorical data, however, with proper data preparation, FCA can be used to analyze numerical data effectively. In particular, the continuous numerical data is discretized, and a new dataset is constructed based on these discretized data. This dataset is called the ***Formal Context,*** which is among the three key components of FCA. The other two are ***Formal Concepts*** and ***Concept Lattice***. *Formal Concepts* are responsible for capturing the patterns in the data through clustering different items sharing similar characteristics. A formal concept is a pair ***(E, I)***, where ***E*** "extent" is a set of objects that share common attributes and ***I*** "intent" is a set of attributes shared by all the objects in ***E***. Additionally, ***Concept Lattice*** represents the hierarchical structure of ***Formal Concepts*** that illustrates the important patterns as well as dependencies within the available data in a visual way. For more information about the fundamentals and mathematical background of FCA, readers can refer to these key publications [620]–[622]. It is worth noting that FCA was performed in two complementary directions to extract the maximum useful information to be fed afterwards to ***Phase #2***. The first one is considering the low-level observations while the second one considers the patterns generated by the developed IML model.

The results of GC and FCA are then used to feed and guide the causal discovery and inference models to be developed in ***Phase #2***. These results and information are essential to build well-informed DAGs and suggesting the correct equations in SEM (combined with human expertise).

Second phase: Causal discovery & inference

Our methodology for causality analysis in this phase relies mainly on ***Graphical Causal Models*** (GCMs). Two approaches are proposed to generally extract the causal relationships based on the case in hand and its complexity. The first one is ***Structural Equations Modeling*** (SEM) [623] and the second approach is ***Causal Bayesian Networks*** (CBNs), which consist mainly of ***Directed Acyclic Graphs*** (DAGs) [624]–[630]. SEM is mainly based on factor analysis and regression of observed and latent variables to obtain the causal relationships. The latent variables are sort of unobservable variables defined by the process experts combining other observed variables in a certain way to further define the process's performance. SEM can be considered as a sort of GCMs as the results are presented in the form of causal graphs. Additionally, it relies heavily on the prior expertise/knowledge of process expert and the empirical relations put by the experts based on their own prior research. Hence, it is not using the observed data alone without the dependence on theoretical assumptions and experts' empirical research. It should be noted that SEM is an excellent choice for small well-established systems, however, for larger and complex systems, its ability is limited [631]. One solution to avoid this problem is the process modularity which means that we can divide the complex system to smaller process modules and then investigate each one separately through SEM.

On the other hand, we can go for the development of robust CBNs based on well-informed/constructed DAGs, which is more suitable route for complex larger systems. In addition, this novel combined approach relies mainly on the observed data with minimal dependence on theoretical background of process experts or empirical relations put by them. Yet, a careful development of the DAGs is required to build the correct and logical structural causal model. That is the reason why we proposed the novel approach of combining GC with FCA and use their results, besides some insights from IML+XAI, to well-inform DAGs to consider the correct causal relationships amongst the given dependent and independent variables. An additional advantage of CBNs over SEM is the consideration of uncertainties in the data. The steps performed to have representative CBNs are:

(1) Data preparation through normalization.
(2) Building and learning the structural model (DAG) using *NOTEARS* algorithm.

(3) Modification of the built DAG based on some domain expertise (traditional approach) besides the results/insights obtained from ***Phase#1*** of causality analysis with special emphasize on GC and FCA (our novel proposed approach).
(4) Normalized data discretization through different methods including Decision Tree-based discretization.
(5) Fitting the conditional distribution of the *Bayesian* network based on the discretized data.
(6) Performing interventions and counterfactual analysis through *Do-Calculus* approach to determine the true causal effects.

Third phase: Clear representations for process expert

In this phase we are building empirical models or equations based on the data/information retrieved during the whole steps of optimization and causality analysis. These models target mainly the process experts to represent the causal relations in a simplified way for fast decision making. The collected data is fitted to models with some aid of the causality analysis results to define the significant variables and their interactions. The developed models and equations can be updated continuously in the future to cover other ranges of the design variables and KPIs of interest when needed. The ML model selected to map the inputs with causality-informed modifications to the desired outputs (KPIs) is the tree-based ensemble *CatBoost* model. It was selected due to several merits including accurate modeling results while overcoming the problem of overfitting and excellent management of tabular data. This model is different from the one constructed based on CBNs as it focuses on predicting single value of KPI(s) at certain values of input variables unlike CBNs, which give conditional probability distribution. Another difference between them is that *CatBoost* will work with continuous data directly while CBNs depends on the discretized form of this continuous data. The developed equations will be presented in the results section.

6.4.4 Remarks on incremental learning

One of the key features of our methodology is the gradual update of the RL agent in an incremental way. Firstly, wide ranges are assigned for the actions based on literature and common practices. After initial runs/trials, the data is collected from the simulator to test the importance of the variables and their causal effects on the target KPIs according to ***Phase #1*** and ***Phase #2***. This gives insights about the most effective design variables and their causal effects on the KPIs of interest and how they presumably affect each other. These results, i.e. causal graphs, clear

relationships and interpretable patterns, give a clear vision about the direction of changing variables' ranges to narrow down the search space efficiently. The extracted knowledge guides the human expert to narrow down or even change the range of actions in the agent in the new trial to explore new solutions to reach the desired optimality. This exploration is guided also by gradually changing the baselines of KPIs' values while benefitting from the learned policy in the first trials as previously mentioned. Wide and narrow ranges of the selected process/design variables will be mentioned clearly in **Sections 4.1** and **4.2**.

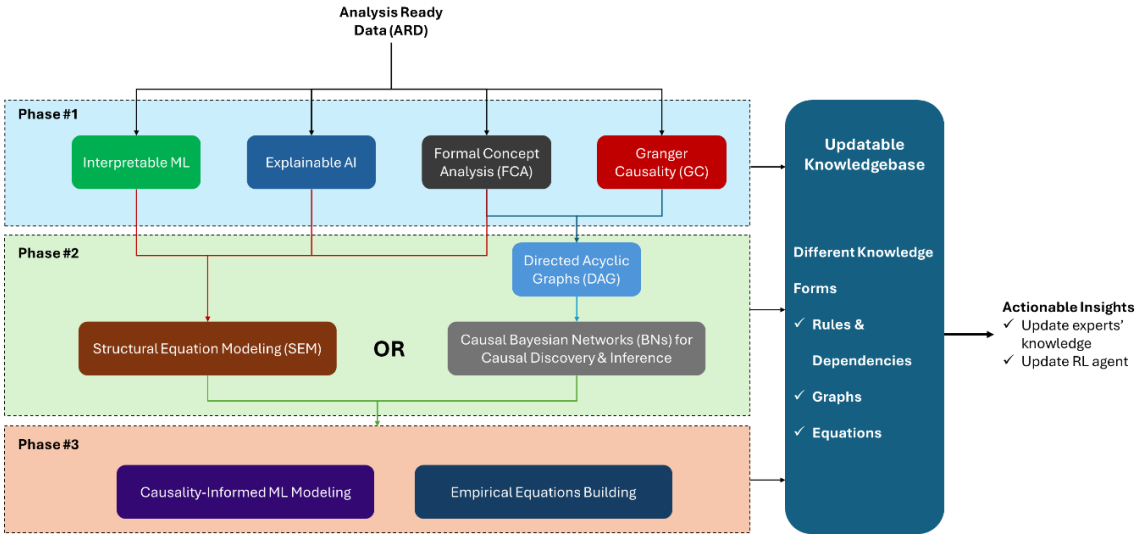


Figure 6.4 General methodology of causality analysis

6.5 Case studies

In the current study we considered two case studies, i.e. physical and chemical absorption for carbon capture, to validate our developed methodology. The following two subsections (**6.5.1** and **6.5.2**) give the details of each case study.

6.5.1 Chemical Absorption of CO_2

The first case study is focused on the optimization of amine-based carbon capture process, where MEA is used to absorb CO_2 chemically from flue gases exiting an actual cement manufacturing plant. The average composition and characteristics of this flue gas stream, provided by another research team under the umbrella of interdisciplinary work in Natural Resources Canada (NRCan), are summarized in Table 6.2 which is different from the flue gas considered in our first study [592].

Table 6.2 Flue gas composition and characteristics

Property [Unit]	**Value**
Temperature [^{o}C]	38
Pressure [bar]	1.5
Flowrate [kmole/h]	7977
Nitrogen composition [mole%]	0.68921
Oxygen composition [mole%]	0.02328
Carbon dioxide composition [mole%]	0.22653
Water composition [mole%]	0.06099

This industrial process consists of absorption and re-boiled stripping (desorption) columns as shown in the simplified process flow diagram illustrated by Figure 6.5. The first contact between the raw flue gases and the lean solvent takes place in the absorption column. The purified flue gas stream exits the top of the absorption column, while the CO_2-rich solvent leaves at the bottom to be regenerated in the stripper. As a matter of heat integration, this CO_2-rich stream is warmed in a heat exchanger by the hot lean solvent stream exiting the desorption column. The initial temperature of the reboiler is 120 ^{o}C and the initial boil-up ratio is 0.2 (molar). Boil-up ratio is defined by default in *Aspen HYSYS* simulator in molar units. A purge stream is taken out of the stream of regenerated solvent exiting the stripping column. This is done to avoid CO_2 accumulation in the system and to guarantee optimal regeneration without excessive solvent degradation. The recycled stream is further cooled to meet the assigned absorption temperature. The cooled recycle stream is then mixed with the fresh make-up solvent stream before being introduced to the absorption column. It should be noted that the *Murphree* efficiency for all the columns during simulation in this study is 100%. Yet, the overall efficiencies of absorption column and the stripping column are considered as 50% and 65%, respectively, during costing to have an actual

representation of capital and total capture costs. In addition, the temperature approach in all heat exchangers is 5 ºC.

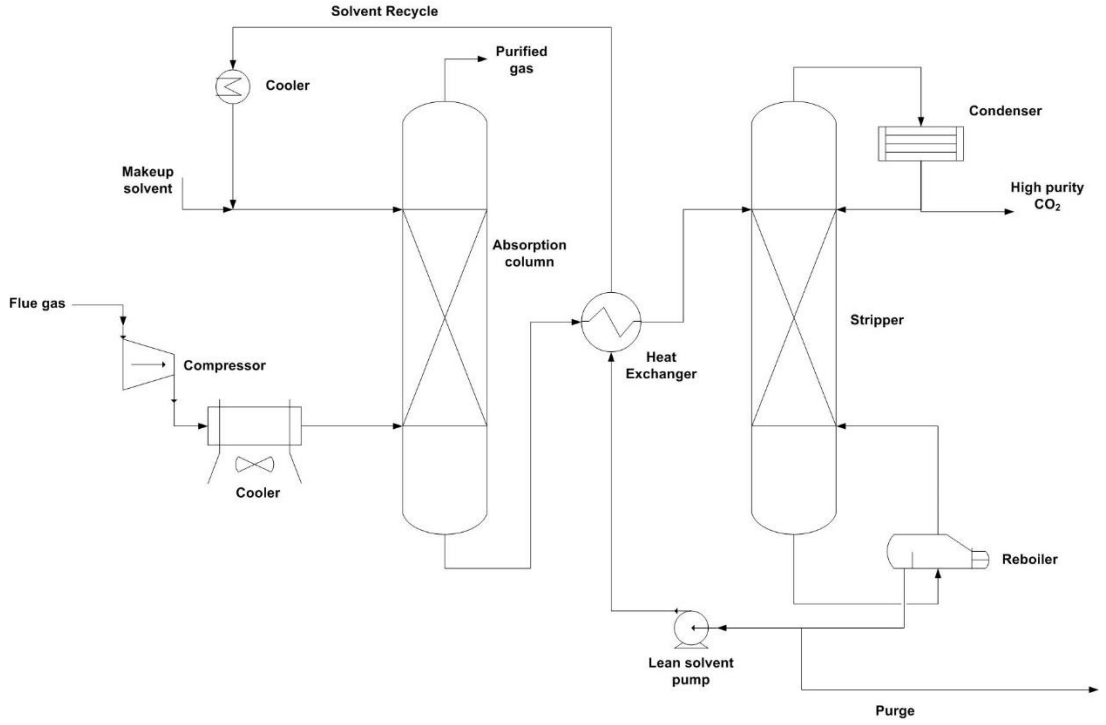


Figure 6.5 MEA-based carbon capture simplified process flow diagram

The conditions and characteristics of the lean solvent entering the absorption column in the first trial (before applying the recycle) are a temperature of 30 ºC, a pressure of 1.5 bar, and CO_2 concentration of 5% (mass). Besides, the flowrate is 16000 kmole/h and the MEA concentration is 20% (mass) and the remaining 75% (mass) is water. It is worth noting that the fluid package utilized in this case is the *Chemical Solvents* package embedded in the *Aspen HYSYS (V:14.0)* industrial process simulator. Besides, the reaction kinetics including the kinetic parameters were not manually defined to avoid any misleading results as the mechanism and reaction kinetics of MEA-based carbon capture are still in the research phase and may possess uncertainty. The actions considered in this case study are four actions, namely, lean solvent temperature (T_1), absorption pressure (P_1), flue gases temperature (T_2) and boil-up ratio inside the stripping column (BUR). As a way of transfer learning, the effect of both recycle stream percentage ($R\%$) and temperature of stream entering stripping column (T_{str}) were considered based on the results obtained through the earlier study [592]. In particular, the highest possible recycle percentage of 98% was considered. T_{str} in the range of 64-68 ºC was also considered to be high enough to decrease the steam

consumption in the stripping column (lower operating costs) while maintaining optimal values of heat exchange area (lower capital costs). It should be noted that solvent-to-flue gas ratio (S/F) was studied in our previous study [592] and we did transfer learning and applied the results in the current study to get the current ratio. The optimization of recycle and make-up streams flowrates in this previous study was essential to obtain the optimal S/F ratio used in the current study. Additionally, temperatures and pressures were the most influential variables based on the causality analysis in our previous study; hence, we considered them here besides the boil-up ratio representing the steam consumption as one of the key variables affecting the total capture cost as well as the total process emissions. *SI* system was considered to represent all the units of pressures, temperatures and other conditions/properties in the studied system. During the coding of the agent and simulation environment, the limits of temperatures and pressures such as the pinch points in heat exchangers were defined and considered. The ranges of the design variables were also defined inside the simulation environment and the RL agent. In addition, the direct link between Aspen HYSYS simulator and the DDPG RL agent make the optimization process physics-guided where all the equilibrium constraints and mass/energy balances are taken into consideration to avoid any physically unreliable solutions. Rules of design, e.g. equipment sizing rules, and human expertise, for continuous supervision of the results' quality, were considered and played key role in this as well.

The state of the process is defined by a vector containing 11 main elements representing important measured/calculated variables reflecting the techno-economic state of the industrial process. These variables include carbon dioxide purity, steam consumption, electricity consumption and carbon dioxide recovery. Reward definition is put based on both total capture cost and the process environmental impact (CO_2 emissions due to electricity and steam consumptions). Following is the breakdown of the rewards and penalties proposed to be given to the agent in each case of meeting or not meeting the predefined value(s) of the desired KPI(s).

$$R_{MEA} = \begin{cases} -1 \times ((1 - Tot\_cost) + (1 - Tot\_emissions)), & Tot\_cost > Goal_C, Tot\_emissions > Goal_{EM} \\ 0, & Tot\_cost = Goal_C, Tot\_emissions = Goal_{EM} \\ 2 \times (1 - Tot\_cost) - (1 - Tot\_emissions), & Tot\_cost < Goal_C, Tot\_emissions > Goal_{EM} \\ -2 \times (1 - Tot\_cost) + 1.5 \times (1 - Tot\_emissions), & Tot\_cost > Goal_C, Tot\_emissions < Goal_{EM} \\ 0 + 2 \times (1 - Tot\_emissions), & Tot\_cost = Goal_C, Tot\_emissions < Goal_{EM} \\ 0 - 1 \times (1 - Tot\_emissions), & Tot\_cost = Goal_C, Tot\_emissions > Goal_{EM} \\ -1 \times (1 - Tot\_cost) + 0, & Tot\_cost > Goal_C, Tot\_emissions = Goal_{EM} \\ 3 \times (1 - Tot\_cost) + 0, & Tot\_cost < Goal_C, Tot\_emissions = Goal_{EM} \\ 3 \times (1 - Tot\_cost) + 2 \times (1 - Tot\_emissions), & Tot\_cost < Goal_C, Tot\_emissions < Goal_{EM} \end{cases} \quad (6.3)$$

Where, *Tot_cost* is the relative total capture cost (USD/kg captured CO_2) and *Tot_emissions* is the relative total calculated emissions (kg CO_2 produced/kg CO_2 captured). These are the main KPIs considered in this case study. Additionally, $Goal_C$ is the predefined threshold (baseline) of total relative capture cost and $Goal_{EM}$ is the predefined threshold (baseline) of total relative emissions. As shown in this formulation, if both capture cost and total emissions are higher than the specified threshold, then the agent will be penalized twice (full penalty). If both capture cost and total emissions are equal to the thresholds then the agent will take zero reward. If both capture cost and total emissions are lower than the thresholds then the agent will be rewarded twice (full reward). If one of the KPIs achieved lower value than the threshold and the other one still higher than its threshold, then the agent will take a reward for the good achievement related to the first KPI and a penalty for not achieving the goal for the second KPI. Following points are important remarks taken into consideration while constructing the reward formulation and the selection process of baselines' values:

- The baselines/thresholds are gradually updated as our proposed optimization methodology is an incremental-based one, which allows more flexible and causality-informed/enhanced optimization.
- The coefficients in each case in the abovementioned reward formulation are based on rigours optimization. This step is performed to ensure that the rewards and penalties for both capture cost and total emissions are within the same scale. It is an essential step so that the agent could minimize both costs and emissions simultaneously without encountering any unnecessary excessive values of rewards or penalties during any step/episode that leads to the negligence of one of the objectives and focusing on only one of them in this multi-objective optimization problem.
- The values of rewards and penalties are in a comparable range such as the one of the standardized actions to facilitate the learning of total emissions and capture costs minimization policy and ensure its stability.
- The values of the relative total emissions and total capture costs are subtracted from 1 in each case to give higher reward values at lower values of both relative total emissions and total capture costs. This formulation along with adding the causality results to the agent build some sort of reasoning and accelerate the optimization process.

6.5.2 Physical Absorption of CO_2

In this case study, we considered the ***Selexol*** process as a representative of physical absorption-based carbon capture. In this process ***DEPG*** is used as the solvent that absorbs CO_2 physically. Thus, the fluid package utilized in this case is the *Physical Solvents* package embedded in the *Aspen HYSYS (V:14.0)* industrial process simulator. The simplified process flow diagram of this process is depicted by Figure 6.6. The solvent regeneration or desorption in this case is different from the one in MEA-based carbon capture one. Pressure change plays the key role to separate the dissolved CO_2 in the stream exiting the high-pressure absorption column. A train of reducers and flash drums was used to perform this duty. The vapor streams exiting the flash drums are then compressed and cooled before being recycled to the absorption column. The final drum in the train produces CO_2 with high purity that is then compressed and cooled to be then stored for further specialized applications.

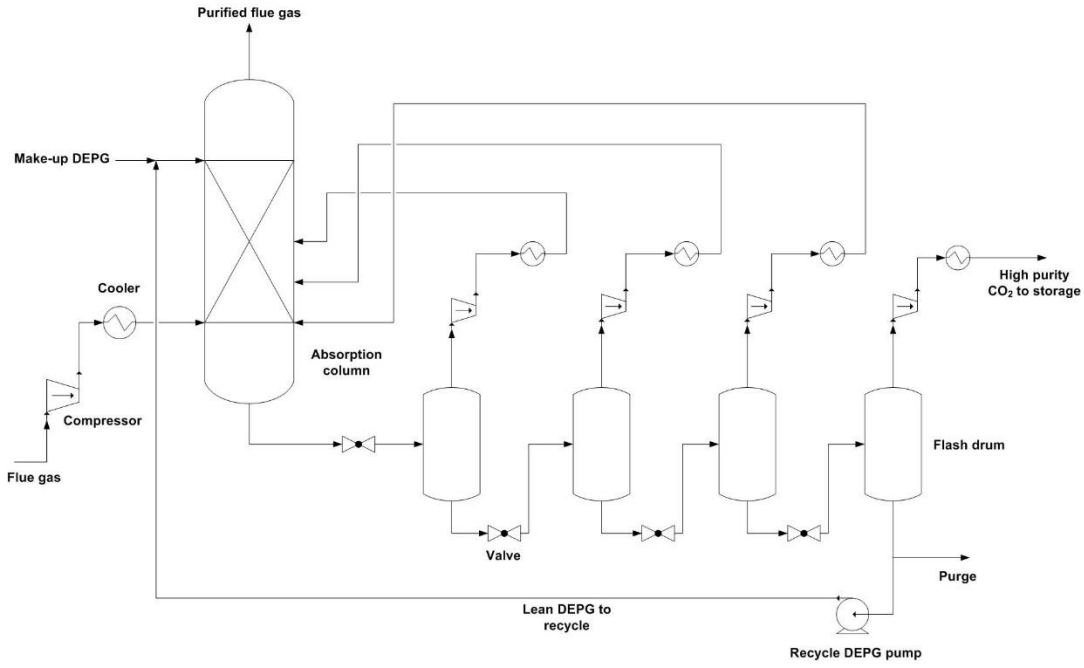


Figure 6.6 DEPG-based carbon capture process flow diagram

Similar state and reward definitions were used here as the ones developed in the case of MEA-based capture. However, the actions here are a bit different due to the nature of the process. In particular, the actions are absorption pressure (P_1), solvent to flue gases mole ratio (S/F) and number of stages of absorption column (N_{abs}). The two KPIs in this case are combined ones, where

they combine capture cost, CO_2 recovery, CO_2 purity and the relative emissions of the capture process. The first new KPI suggested to evaluate the process is given the name *Inverse Capture Effectiveness Score* and calculated through dividing the capture cost by both CO_2 recovery and CO_2 purity (**Equation 6.4**). The second one is given the name *Inverse Capture Eco-Friendliness Score* and calculated through dividing the total relative emissions by both CO_2 recovery and CO_2 purity (**Equation 6.5**).

$$Inverse\ Capture\ Effectiveness\ Score = \frac{Capture\ Cost}{CO_2\ Recovery \times CO_2\ Purity} \tag{6.4}$$

$$Inverse\ Capture\ Eco-Friendliness\ Score = \frac{Relative\ Total\ Emissions}{CO_2\ Recovery \times CO_2\ Purity} \tag{6.5}$$

These new KPIs with these combinations make the optimization process more flexible and ensure separable data classes when working afterwards for extracting the possible causal relationships. These two KPIs were developed in this form to keep consistent comparison as possible where we pursue process optimization through these minimizing scores as in the case of using total cost and relative total emissions. Yet, in future studies for capture processes evaluation, stakeholders and process experts can use the inverse of these two developed scores, where the goal of the optimization is maximizing both capture effectiveness and eco-friendliness. Following is the mathematical representation of the reward formulation in the DEPG-based carbon capture case study.

$$R_{DEPG} = \begin{cases} -1 \times ((1 - Tot\_cost) + (1 - Tot\_emissions)), & Cap\_eff > Goal_{CE}, Cap\_eco > Goal_{CEF} \\ 0, & Cap\_eff = Goal_{CE}, Cap\_eco = Goal_{CEF} \\ 2 \times (1 - Tot\_cost) - (1 - Tot\_emissions), & Cap\_eff < Goal_{CE}, Cap\_eco > Goal_{CEF} \\ -2 \times (1 - Tot\_cost) + 1.5 \times (1 - Tot\_emissions), & Cap\_eff > Goal_{CE}, Cap\_eco < Goal_{CEF} \\ 0 + 2 \times (1 - Tot\_emissions), & Cap\_eff = Goal_{CE}, Cap\_eco < Goal_{CEF} \\ 0 - 1 \times (1 - Tot\_emissions), & Cap\_eff = Goal_{CE}, Cap\_eco > Goal_{CEF} \\ -1 \times (1 - Tot\_cost) + 0, & Cap\_eff > Goal_{CE}, Cap\_eco = Goal_{CEF} \\ 3 \times (1 - Tot\_cost) + 0, & Cap\_eff < Goal_{CE}, Cap\_eco = Goal_{CEF} \\ 3 \times (1 - Tot\_cost) + 2 \times (1 - Tot\_emissions), & Cap\_eff < Goal_{CE}, Cap\_eco < Goal_{CEF} \end{cases} \tag{6.6}$$

Where, *Cap_eff* is the inverse capture effectiveness score and *Cap_eco* is the inverse capture eco-friendliness score. $Goal_{CE}$ is the predefined threshold (baseline) of the inverse capture effectiveness score and $Goal_{CEF}$ is the predefined threshold (baseline) of inverse capture eco-friendliness score. Same remarks considered previously while formulating the reward function in the MEA-based capture process optimization case study are still applicable in this case of optimizing DEPG-based capture process.

6.5.3 Remarks on costing and environmental impact estimation procedures

The full costing calculation procedure followed in the current study to evaluate the capture cost is a typical chemical plant costing procedure, which is summarized in this textbook [632] and these previously published studies [489], [595], [633]. The purge stream was included in the capture process to consider the treatment of regenerated solvent and to avoid degradation. Based on this, costs of make-up solvents were considered and included in the total capture cost. The costs of the start-up solvents/chemicals were also considered and included in the calculations of capture costs according to the abovementioned references. The prices of different solvents are provided in Table S1 in the supplementary materials file. Besides, the need for net-zero emissions is the motivation to include the processes' environmental impact as an important KPI to be optimized simultaneously with capture cost. In this regard, a simple environmental impact calculation procedure based on total emissions evaluation was employed inspired by [489]. The following equations represent this procedure, where the main sources of emissions are the electricity and steam production as well as fuel (natural gas) combustion for other applications such as driving gas turbines to compress flue gas streams.

$$QE_{CO_2} = Total\ Electricity\ Consumption \times EF_{CO_2} \tag{6.7}$$

$$TE_{Heating\ Steam} = Energy\ Required\ for\ Heating \times \frac{1}{\eta_{boiler}} \times F_{CO_2} \tag{6.8}$$

$$TE_{Compression\ Steam} = Energy\ Required\ for\ Compression \times \frac{1}{\eta_{turbine}} \times \frac{1}{\eta_{boiler}} \times F_{CO_2} \tag{6.9}$$

$$TE_{Compression\ NG} = Energy\ Required\ for\ Compression \times \frac{1}{\eta_{turbine}} \times F_{CO_2} \tag{6.10}$$

Equation 6.7 calculate the total emissions due to electricity consumption (QE_{CO2}) based on the emissions factor of certain region (EF_{CO2}). Total emissions related to heating steam consumption ($TE_{Heating\ Steam}$) is calculated through **Equation 6.8** where F_{CO2} (kg CO_2 / GJ of energy produced) is the emissions factor of the used energy source. η_{boiler} represents the efficiency of the boiler used for steam production, which is a function of energy source, e.g. natural gas or electricity. The term ($\eta_{turbine}$) is the efficiency of turbines used to drive the industrial compressors whether they are operated using steam (**Equation 6.9**) or natural gas directly (**Equation 6.10**). Different values of efficiencies and emission factors are in the supplementary materials file.

The baselines for both capture costs and total emissions are put following the common practices, industry standards, literature and the recommendations given by international entities to have feasible carbon capture processes. As this approach is an incremental optimization approach, different baselines for both capture costs and total relative emissions were put sequentially, i.e. each optimization trial has specific baseline value of capture cost and total relative emissions. The trials are done in a sequence so that better results and better learning can be achieved by the RL agent. To have a starting point for the optimization the first baseline was put at relatively higher values of capture costs and relative total emissions as a sort of constraints relaxation to facilitate the RL agent duties, i.e. exploration and exploitation for better learning of the optimization policies. The values of the starting points are arbitrary and high enough to give the agent the ability to understand/learn the initial optimization policy. The second baseline is put based on the literature and common values known by experts to challenge the agent to update its policy and explore better values of design variables and KPIs. The third baseline in this incremental optimization procedure is put based on the expectations of the process experts and their desired optimal values of KPIs based on experimental and theoretical studies. This step is done to see if the learned policy and the agent can go for further improvement after incorporating the knowledge extracted from the causal graphs and relationships obtained based on the results of first two trials.

The whole data including values of heat transfer coefficients, utilities prices and emissions factors in different areas, i.e. the provinces of *Quebec* and *Alberta* in *Canada*, used to evaluate the capture cost and total emissions are summarized in the supplementary materials (Appendix J). *Quebec* and *Alberta* were selected to represent the use of different sources of energy and how this variability can affect the calculations of important process KPIs.

6.6 Results and discussions

This section summarizes and discusses the results obtained after validating our causal incremental RL-based methodology on the carbon capture-related case studies.

6.6.1 Optimized RL agent

The hyperparameters optimization was done through different approaches, i.e. random search and *Bayesian* optimization (BO) approaches. Combining BO for hyperparameters optimization with DDPG for process optimization leverages the different capabilities of optimization algorithms and

benefits from the learnability of DDPG RL algorithms for accelerated process design optimization. After doing this, the developed optimum DDPG agent has the hyperparameters presented in Table 6.3. It is worth noting that these values are obtained through working on the two case studies. The exploration part by the actor network in this study was performed through adding an *Ornstein-Uhlenbeck* (OU) process for noise generation. It samples noise from a correlated normal distribution.

Table 6.3 Optimal hyperparameters of the developed DDPG agent

Hyperparameter	**Value**
Number of episodes	25
Number of steps per episode	100
Discount factor (γ)	0
Replay buffer	50000
Batch size	64
Critic learning rate	0.002
Actor learning rate	0.001
Activation function for actor/critic networks	*ReLU*
Activation function for the last (outputs) dense layer of actor network	*Tanh*
Standard deviation for OU action noise (for exploration)	0.2
Target network updating factor (τ)	0.005

During the current study, we emphasized on the applicability of incremental learning combined with causality analysis to efficiently accelerate optimization process without the need for GPU utilization and parallelization. This is to take into consideration performing optimization with the least possible computational requirements to be applicable and handy for process experts.

Regarding hyperparameters optimization, trial and error with human expert involvement were considered firstly to give a good picture about the range of hyperparameters. We followed the values that are published in the literature and confirmed among the ML community in addition to human supervision. Besides, one of the proposed methods to overcome this dimensionality problem is to firstly optimize the RL agent while being connected to a surrogate model representing the studied system to have optimized values in a short time. Then, these values will be applied to the agent connected to the physics-based simulator, e.g. Aspen HYSYS, directly.

6.6.2 Results of MEA-based carbon capture process design

The ranges of the actions in this case are $T_{solvent}$ (24.5-45 ^{o}C), $T_{flue\ gas}$ (24.5-45 ^{o}C), $P_{Absorber}$ (1.5-9 bar) and BUR (0.085-0.225 molar). The total simulation runs are around 11000 runs with an average simulation run time of 30 seconds. Figure 6.7 summarizes the results of first optimization attempts done using the energy prices, emissions factor and costing data related to *Quebec* region. Based on the proposed incremental multi-objective optimization procedure, the first baseline for capture cost was put as 0.3 USD/kg CO_2 in the first trial/attempt. This is a starting point in our incremental learning-based methodology to give the agent the opportunity to start searching the space and follow a certain policy for cost minimization, which will be updated in the next trials. As shown in Figure 6.7 **(a)**, the RL agent is following a policy to decrease the capture cost over simulation runs (steps) while focusing on its exploration capabilities to search the specified design space. The baseline in this trial was reached after around 5500 runs. With respect to the relative process emissions, Figure 6.7 **(b),** the first baseline was put as 1 which means that for each kg of CO_2 to be captured a kg of CO_2 will be emitted. By the end of this trial, the agent could suggest several actions (design variables) to reduce the relative emissions over consecutive simulation runs but the minimum value was higher than the specified baseline by 27%. This highlights the challenging nature of optimizing these two objectives simultaneously and means that more exploration is still required. Thus, a second trial (Figure 6.7 **(c)** and **(d)**) was performed while decreasing the capture cost and relative emissions baselines to the values of 0.065 USD/kg CO_2 and 0.5 kg CO_2 emitted/kg CO_2 captured. The ranges of actions were maintained at the same initial values. Fortunately, the replay buffer is relatively large (50000) giving the DDPG agent the opportunity to learn from all the past simulations (in the first trial) and further follow the path of cost and emissions minimization. As shown in Figure 6.7 **(c)** and **(d),** the agent could reach lower costs and emissions after smaller numbers of simulation runs compared to the first trial, where it

also used its exploration abilities to try to search other parts of the search space to not be trapped in local optima. In these figures, there are some points that seem a bit lower than the final values of capture costs or total emissions at the end of the second trial. However, at these points the purity and capture percentage/rate were lower than the predefined thresholds in this trial. In addition, there is a competition between the two main objectives/KPIs, i.e. capture cost and total emissions, where the cost is at values higher than the threshold while the total emissions are at lower levels and vice versa. Thus, the agent continued working to overcome this competition to obtain acceptable values of both capture costs and total emissions below the defined thresholds in this optimization trial.

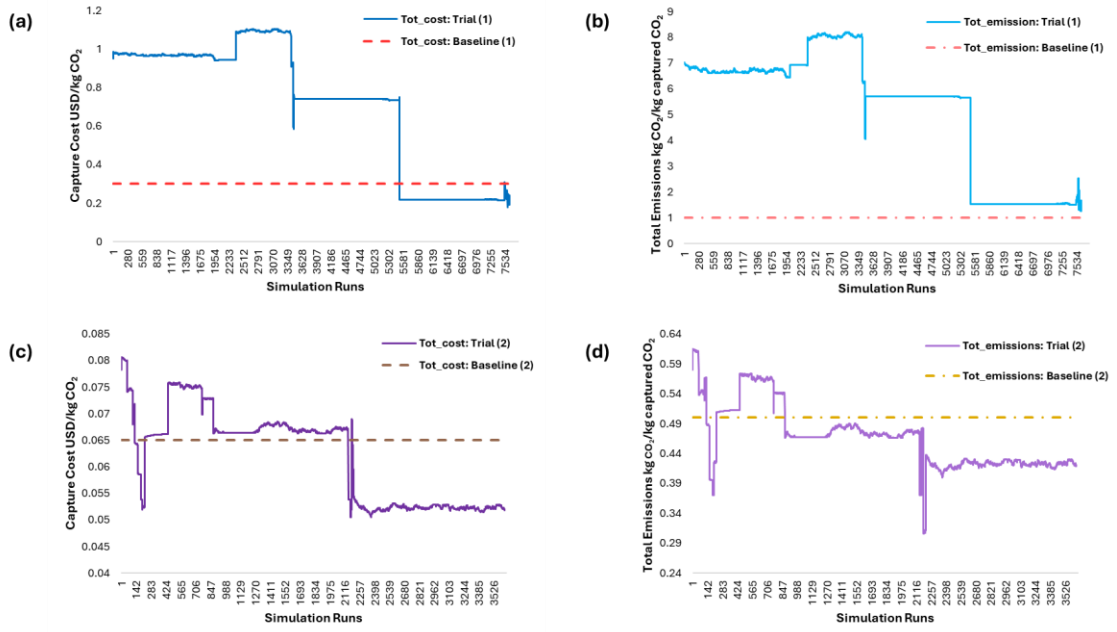


Figure 6.7 Results of first RL-based process design optimization trials (MEA-based capture process)

The results of these two trials were collected, and causal relationships were extracted through following the causality analysis methodology presented in **Section 3.3** to further enhance the optimization results. As a result, XAI-based methodology (regression-based) identified the flue gas stream (entering the absorption column) temperature and inlet solvent temperature as the most impactful variables on the capture cost. The absorption pressure was ranked third in the importance. From the total emissions perspective, absorption pressure was found as the most influential variable followed by flue gas stream temperature and solvent temperature. Fortunately, the direction of the

impact of these variables on both capture cost and total emissions are the same. Absorption pressure and solvent temperature affect the total cost and emissions inversely, which recommends the operation at relatively higher pressures (higher than 2.5 bar) and higher temperatures (> 30 ºC) but not too high to avoid equilibrium limitations. XAI-based analysis suggests operating at lower flue gas temperatures to decrease both capture cost and total emissions. Whereas the IML-based methodology identified the absorption pressure as the most influential variable on capture cost followed by boil-up ratio (BUR), flue gas temperature and the inlet solvent temperature was ranked fourth. The threshold put to classify the process as high cost or low cost is 0.065 USD/kg CO_2 captured (the second baseline). This result is obtained due to the correlation between absorption pressure and flue gas temperature. It is worth noting that similar results were obtained when XAI was applied with a classification ML-based model, where BUR was found significant, unlike the case of regression-based XAI analysis. That is the reason we did not rely solely on IML+XAI and other methodologies were considered to extract the most possible meaningful information/insights. In this regard, GC was employed and gave more insights. For instance, GC highlighted the significance of absorption pressure, solvent temperature, flue gas temperature in agreement with XAI results (regression-based and classification-based). The optimal lag was found to be 9 for both raw dataset and the modified one through conversion to stationary by differencing. The ***p-value*** for absorption pressure, solvent temperature, flue gas temperature and boil-up ratio (BUR) are <0.001, <0.001, <0.001 and 0.9, respectively. This means that BUR statistically has no significant effect among the studied variables/actions upon using the collected data from as the ***p-value*** is above 0.05. In the case of total emissions, the ***p-value*** for absorption pressure, solvent temperature, flue gas temperature and boil-up ratio (BUR) are <0.001, <0.001, <0.001 and 0.62, respectively. This means that BUR still statistically has no significant effect on the total emissions based on GC analysis of the data collected from the first two trials. However, as a side important note, GC does not give a complete vision about the causal relationships besides the need for transformation of the data to stationary type. That is why we only use it as a guide along with IML+XAI analysis.

FCA as a descriptive method was further adopted to extract more hidden patterns/rules. Accordingly, based on the contexts and formal concepts extracted, BUR influence on capture cost was found significant where the capture cost is directly proportional to BUR. This recommends operating at lower BUR values to have capture processes with lower costs. This observation remained true in the case of relative total emissions. Absorption pressure was confirmed as a

significant variable and increasing it leads to increasing the CO_2 recovery, which lowers the total capture cost. However, to ensure acceptable relative total emissions, moderate levels of absorption pressures are preferable. As this is a multi-objective optimization process, then it is recommended to operate at moderate pressures to optimize both KPIs, i.e. capture cost and total emissions. The results of IML+XAI, GC and FCA were used to guide building the DAG (structural causal model) essential to perform robust causal inference through causal BNs. Essential edges are kept in the graphs that represent the effect of significant variables on the KPIs of interest, i.e. capture cost and total emissions. The effect of CO_2 recovery on both total cost and total emissions was included in the causal graphs besides the effect of design variables on CO_2 recovery itself. Whereas any illogical edges were removed such as the effect of KPIs on design variables. The structural causal graphs/models constructed with its weights representing how different variables affecting/causing each other to recommend designing the process at moderate absorption pressures. The absorption pressure directly increases the total cost and emissions but has an inverse indirect effect through increasing the CO_2 recovery. The increase in CO_2 recovery leads to lower total capture cost. Lower flue gas temperatures are recommended as they increase capture cost directly and decrease the CO_2 recovery. BUR has a strong direct effect on capture cost and total emissions, where its increase leads to an increase in their values. Accordingly, lower BUR values are recommended. With respect to inlet solvent temperature, this variable has a direct effect on capture cost (with relatively low significance), where the cost increases due to increasing inlet solvent temperature. At the same time, it has an inverse indirect effect on the KPIs. This result suggests the operation at moderate inlet solvent temperatures to achieve an optimized capture process with maximum recovery and minimum cost/emissions. This range of moderate solvent temperature is presumably preferred in this current case due to decreasing the viscosity of the solvent. This can decrease the mass transfer resistance allowing more gas molecules to dissolve in the liquid solvent phase (higher CO_2 absorption recovery). Another reason could be the increase in reaction rate constant according to *Arrhenius* equation as the MEA-based capture process in a chemisorption one. However, higher temperatures can affect the absorption equilibrium negatively; hence, further optimization of solvent temperature entering the absorption column is crucial. It is worth noting that these results are related to the current study, which may be generalized to other chemisorption systems with similar characteristics. Yet, reasonable experimentation and simulations should be performed under process expert supervision to ensure the validity this hypothesis.

New baselines and action ranges were considered based on the results of the causality analysis for further improvement. In particular, the new baseline was put as 0.045 USD/kg captured CO_2 and the absorption pressure ranged from 3 to 6 bars. Higher pressures (above 6 bars) were not considered based on process expert recommendations and common practices. Operating at higher pressures can lead to excessive emissions (confirmed by FCA and XAI "classification-based" analysis). Solvent temperature ranged from 28 ºC to 38 ºC while the range of flue gas temperature started from around 25 ºC and ended at 30 ºC. BUR new range is specified from 0.085 to 0.125 (molar). Figure 6.8 summarizes the results of this causality-informed optimization trial. Based on our developed causal incremental reinforcement learning (CIRL) methodology, the capture cost was further decreased till reaching around 41.3 USD/ton CO_2 (0.0413 USD/kg CO_2) with a reduction percentage of 8%. In addition, the total emissions were improved, where its value dropped to around 0.283 kg CO_2 emitted/kg CO_2 captured. These values were obtained at absorption pressure of 3 bars (300.5 kPa), flue gas temperature of 25 ºC, solvent temperature of 31 ºC and BUR of 0.087 (molar). The number of theoretical stages for both absorption and desorption columns are 4 and 6 stages, respectively. This relatively low number of stages of absorption column is due to the recommended pressure, which is higher than the common near atmospheric pressure usually used in industrial scale MEA-based absorbers. The average column efficiency for both absorber and stripper is 50% and 65% which is equal to an average height equivalent to theoretical plate (*HETP*) of 1 m. The average diameter of both columns is 6 m. These values were used to determine the actual dimensions essential for calculating actual costs of both columns.

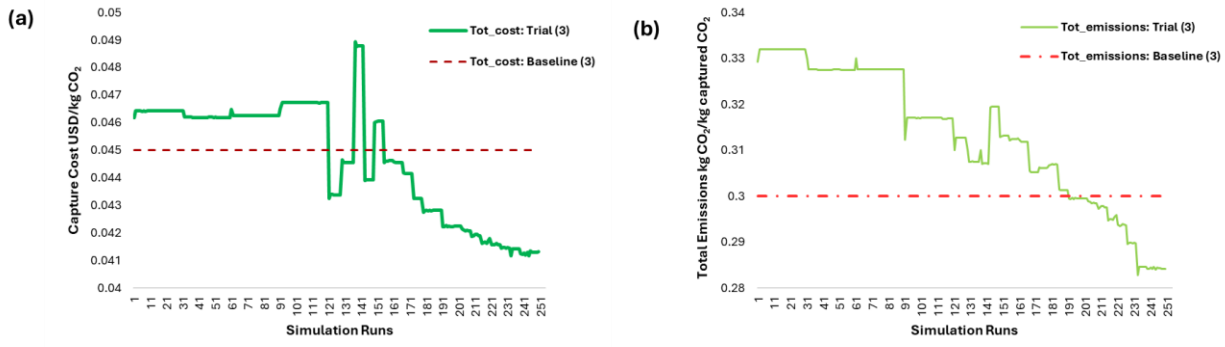


Figure 6.8 Result of RL-based optimization after causality inclusion (MEA-based capture process)

All the data were then collected to build the final causal graphs that can be transferred to new processes and optimization procedures. Figure 6.9 presents the final causal graphs for both cases of capture cost and total emissions of MEA-based capture process. The representations resulted from applying the combined IML and XAI methodology to evaluate the features importance are presented in the supplementary materials. The patterns extracted through employing IML to determine the rules to be followed to have lower capture costs and total emissions are also summarized in supplementary materials (Appendix J). One of the strongest prime patterns obtained using IML in the case of relative total emissions is that regardless the values of other variables if the temperature of the flue gas is higher than 27.6 ^{o}C then the emissions will be higher than 0.5 kg CO_2 emitted/kg CO_2 captured. Another strong pattern suggests that at flue gas temperature lower than 27.6 ^{o}C and absorption pressures lower than 2.96 bar then the emissions will be also higher than 0.5 kg CO_2 emitted/kg CO_2 captured. In the case of capture cost, IML analysis suggests that at BUR lower than 0.145 (molar) and absorption pressures lower than 2.96 bar then the capture cost will be higher than 0.065 USD/kg CO_2. However, if BUR is lower than 0.145 (molar), absorption pressure is higher than 2.96 bar and the flue gas temperature is lower than 30 ^{o}C then the capture cost will be lower than 0.065 USD/kg CO_2. It is worth noting that, through using GC analysis, boil-up ratio (BUR) was evaluated as a significant variable as well and *Granger*-cause the change in capture cost and total emissions after analyzing the whole simulation history combining the three trials. This effect was evaluated as a significant negative one from the cost and emissions minimization point of views after performing the "*Impulse Response Functions"* test. This means that increasing boil-up ratio can lead to higher costs and emissions and vice versa. The optimal lag was found to be 9 as in the case of the data extracted from the first two trials. In addition, the final ***p-value*** for absorption pressure, solvent temperature, flue gas temperature and boil-up ratio are <0.001, <0.001, <0.001 and 0.003, respectively. This means, from GC point of view, that BUR remains the least significant variable among the studied variables/actions but still significant as the ***p-value*** is below 0.05. The full patterns extracted from conducting FCA are tabulated in Table S2 in supplementary materials. Based on the results of FCA, relatively lower capture costs (< 0.065 USD/kg CO_2) most probably can be achieved at BUR values below 0.15 (molar), solvent temperatures above 26 ^{o}C, flue gas temperatures below 27 ^{o}C and absorption pressures between 3 and 4 bars. These ranges of design variables achieve lower relative total emissions (< 0.5 kg CO_2 emitted/kg CO_2 captured) and higher CO_2 recoveries (> 95%). These results are in good agreement

with and supported by the causal relationships presented in Figure 6.9. This figure also contains other important information about the effect of some input variables on others. In particular, the choice of boil-up ratio is significantly affected by the selected values of absorption pressure (direct proportionality), solvent temperature (inverse proportionality) and flue gas temperature (direct proportionality). These results will guide the process designer for the proper selection of different variables in order of having feasible operation range while achieving the optimal values of process KPIs. It is worth noting that the coefficients on the arrows are dimensionless to show the relative cause-effect relationships between different variables. The higher the magnitude of the coefficient, the stronger the relation and the higher the conditional probability that the variable or the KPI will change.

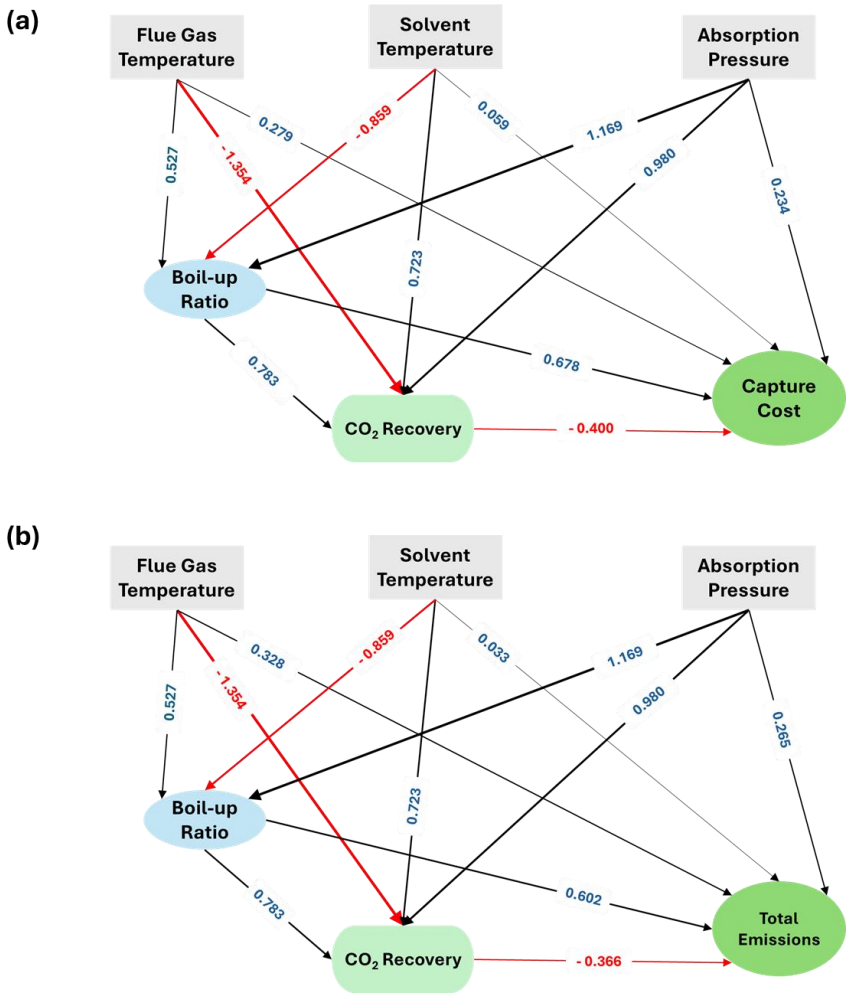


Figure 6.9 Causality analysis graphical representations of **(a)** capture cost and **(b)** total emissions of MEA-based CO_2 capture process

From all the knowledge gained through, we were able to construct empirical equations to represent the relation between design variables and process KPIs in a causality-guided way. They will be used afterwards by process experts and other stakeholders for quick decision making. A compartmental modeling-like approach was adopted in this case, where different ranges of KPIs were defined and a separate equation was developed for each range/region. In the case of capture cost, three main regions were considered, i.e. (0-0.15, 0.15-0.5 and 0.5-1.5 USD/kg CO_2). Similar approach was followed in the case of relative total emissions where the three regions are (0.25-1.25, 1.25-2.6 and 2.6-8.5 kg CO_2 emitted/kg CO_2 captured). The final mathematical representations of these equations for the (0-0.15 USD/kg CO_2) cost range and (0.25-1.25 kg CO_2 emitted/kg CO_2 captured) total emissions range where the CO_2 recovery is above 96% are presented below. The equations for other ranges are reported in the supplementary materials. The first four terms in each equation represent the direct effect of each variable whereas the following four terms represent the indirect effect of each variable on the KPIs of interest. The last terms represent the possible interactions between different variables.

$$\begin{aligned}\boldsymbol{Capture\ Cost} = {} & -2.77e-02 \times P^{0.234} + T_S^{0.059} - 0.46 \times T_{FG}^{0.279} - 0.2 \times BUR^{0.678} \\ & - \frac{1.23}{P^{0.392}} + \frac{0.67}{T_S^{0.289}} - 0.11 \times T_{FG}^{0.542} - \frac{0.33}{BUR^{0.313}} \\ & + \left(BUR^{0.678} + \frac{1}{BUR^{0.313}}\right) \times \left(-1.95e-06 \times P^{1.169} - \frac{4.26e-02}{T_S^{0.859}} + 5.92e-02 \times T_{FG}^{0.527}\right)\end{aligned} \tag{6.11}$$

$$\begin{aligned}\boldsymbol{Total\ Emissions} = {} & -1.03 \times P^{0.256} + 13.44 \times T_S^{0.033} - 4.19 \times T_{FG}^{0.328} - BUR^{0.602} \\ & - \frac{28.79}{P^{0.359}} + \frac{5.06}{T_S^{0.265}} + 0.62 \times T_{FG}^{0.496} - \frac{2.24}{BUR^{0.287}} \\ & + \left(BUR^{0.602} + \frac{1}{BUR^{0.287}}\right) \times \left(1.2e-04 \times P^{1.169} - \frac{0.46}{T_S^{0.859}} + 0.41 \times T_{FG}^{0.527}\right)\end{aligned} \tag{6.12}$$

CatBoost ML model trained on modified input variables and mapping them to both costs and emissions showed a good performance, where the determination coefficient (R^2) took values of around 99% with relatively low MSEs as well. The input variables were modified through multiplying their actual values by combined coefficients representing both direct and indirect effects. Figures presenting the actual versus predicted values of costs and emissions obtained by the developed ML model in both training and testing phases are in the supplementary materials (Appendix J) along with the optimized hyperparameters of this ML model. These models and

equations can be used for fast decision making by different stakeholders while considering the causal relationships implicitly.

6.6.3 Results of DEPG-based carbon capture process design

In the case of DEPG-based capture this number is around 8000 runs with an average simulation run time of 10 seconds. The ranges of actions are $P_{Absorber}$ (1.5-30 bar), S/F (0.75-5 molar) and $NTS_{Absorber}$ (3-12 stages). In a similar way as followed in the case of MEA-based carbon capture, two trials were firstly conducted, where the two baselines for inverse capture effectiveness and inverse capture eco-friendliness scores (in first trial) were put as 1.11 and 0.0148. These two values are corresponding to capture costs of 0.15 USD/kg CO_2 and 0.002 kg CO_2 emitted/kg CO_2 captured. In addition, the CO_2 recovery and purity thresholds included in these two baselines are 15% and 90%. In the case of second trial, the inverse capture effectiveness score baseline was put as 0.1579 corresponding to 0.075 USD/kg CO_2 capture cost, 50% CO_2 recovery and 95% CO_2 purity thresholds. The inverse capture eco-friendliness score baseline was put as 0.0021 corresponding to total emissions threshold of 0.001 kg CO_2 emitted/kg CO_2 captured and the same thresholds for both CO_2 recovery and purity used for inverse capture effectiveness score's second baseline. Figure 6.10 illustrates the results of these two trials. As illustrated, the RL agent in the first trial achieved considerably lower scores than the specified thresholds during the first 400 simulation runs. Then it continued and used its exploration abilities to discover other possible optimal solutions. With the aid of agent's replay buffer and updating the baselines, other lower scores were obtained during the first 150 simulation runs in the second trial.

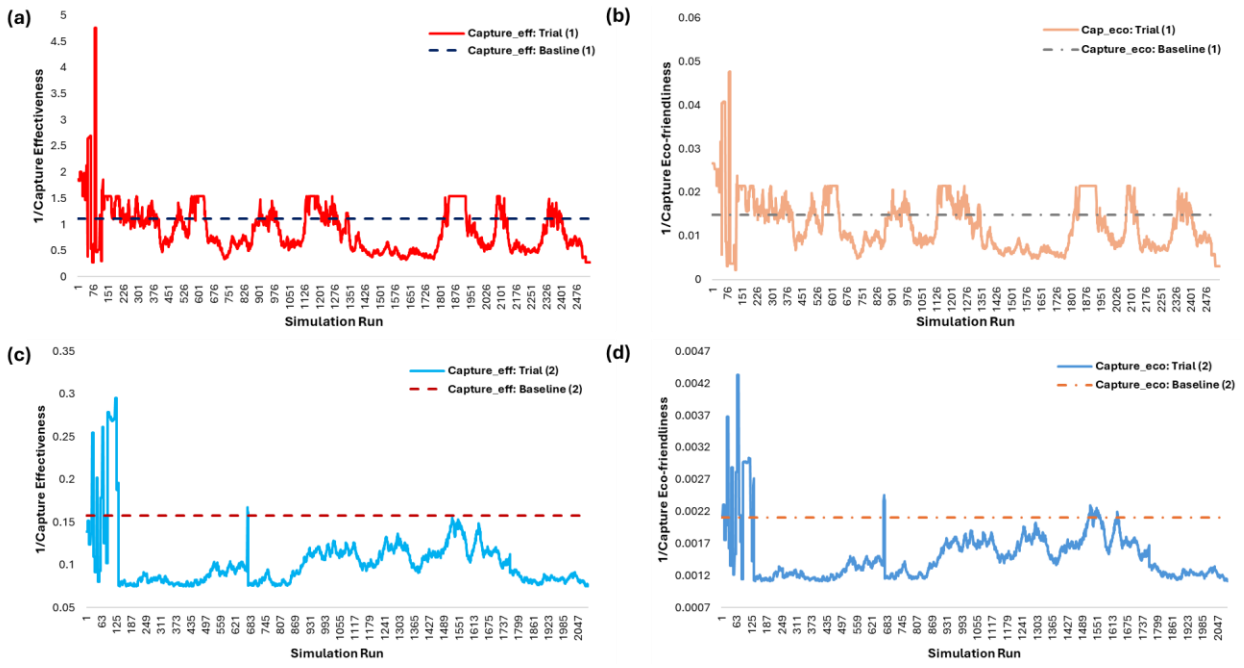


Figure 6.10 Results of first RL-based process design optimization trials (DEPG-based carbon capture)

IML+XAI-based analysis of the simulation history collected from these two trials were performed. Solvent-to-feed molar ratio was indicated as the most impactful variable on the inverse capture effectiveness score of DEPG-based process followed by absorption pressure and number of stages of absorption column. Inverse capture effectiveness score is inversely proportional to solvent-to-feed molar ratio and directly proportional to both absorber number of stages and absorption pressure. The same results and effects remain accurate in the case of inverse capture eco-friendliness score. Fortunately, IML-based analyses supported these results and have perfect agreement with them in both cases of inverse capture effectiveness and eco-friendliness scores. XAI combined with classification-based analysis suggests that higher pressures can lead to inverse effectiveness scores lower than 0.15 which corresponds to capture costs lower than 0.075 USD/kg CO_2, CO_2 purities higher than 90% and recoveries higher than 50%. GC analysis confirmed the significance of solvent-to-feed molar ratio, absorption pressure and number of stages of absorption column effects on both inverse capture effectiveness and inverse capture eco-friendliness, where all of them had ***p*-values** < 0.001. The optimal maximum lag in this case was found to be 20. Moreover, GC analysis showed that the number of stages of absorption column is affected significantly by both solvent-to-feed molar ratio and absorption pressure. FCA confirmed the

significance of the three main input variables. FCA-based patterns suggest making it clear that higher inverse capture effectiveness scores (corresponding to higher costs) can be obtained at lower solvent-to-feed molar ratios (< 3) and higher number of stages of absorption column (> 9 stages). Pressures lower than 7.5 bars lead to lower recoveries and consequently higher inverse effectiveness scores "higher capture costs". Same pattern was observed in the case of inverse capture eco-friendliness scores. For instance, solvent-to-feed molar ratios < 3 lead to higher inverse eco-friendliness scores (> 0.002) corresponding to higher relative total emissions. Based on the abovementioned results and human expertise, preliminary DAGs and CBNs were constructed for both cases of inverse capture effectiveness and inverse capture eco-friendliness scores. The structural models and causal graphs confirmed the significant effect of solvent-to-feed molar ratio, where increasing its value decreases the inverse capture effectiveness score. this means lower costs can be obtained at higher solvent-to-feed molar ratios. Both direct and indirect effects of this important variable lead to decreasing the capture cost when increased. In the case of absorption pressure, it has a positive direct effect on the inverse effectiveness score and capture cost, where its increase leads to significant direct increase of both. However, it has another indirect effect, where its increase leads to a decrease of capture cost and inverse effectiveness score due to increasing CO_2 recovery. A similar observation was obtained in the case of number of stages of the absorption column. This suggests designing the absorption process at moderate pressures while considering relatively higher solvent-to-feed molar ratios and moderate number of stages as well. Similar results were obtained in the case of both inverse capture eco-friendliness score and total relative emissions.

Combining all the previous results suggests performing the absorption process at moderate pressures between 5 bars to 15 bars instead of the wide range suggested in the first two trials (between 1.5 bars to 30 bars). In addition, as higher solvent-to-feed molar ratio are recommended to have lower inverse capture eco-friendliness and effectiveness scores, its range will be narrowed to be (3 to 5) instead of the wide range of (0.75 to 5) as in the first two trials. The maximum number of stages considered in the third trial was decreased from 15 (in the first trials) to only 12 stages. Accordingly, the optimal design variables (actions) for DEPG-based capture process were found to be solvent-to-feed molar ratio of 4.5, absorption pressure of 7.5 bars and absorption column number of stages of 9. At these values of design variables, the capture cost and total emissions took the values of 0.0445 USD/kg CO_2 and 0.00071 kg CO_2 emitted/kg CO_2 captured. These values

were obtained based on the data related to *Quebec* province where electricity is used directly to drive the compressors.

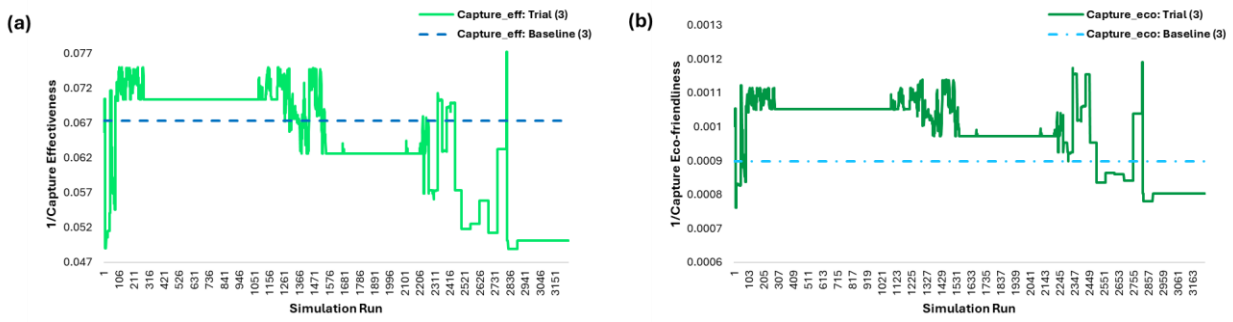


Figure 6.11 Result of RL-based optimization after causality inclusion (DEPG-based carbon capture)

Similar to what have been done in the case of MEA-based process, all the simulation-based optimization history was collected, and final causal graphs were constructed. Figure 6.12 summarizes these causal graphs and relationships. As shown, solvent-to-feed has significant direct and indirect negative effects on all the KPIs of interest, which highly recommends in general using higher ratios (> 3.7 supported by strong patterns obtained via IML analysis) to achieve cost-effective and eco-friendly physical absorption process. Among the three studied design variables, solvent-to-feed ratio has the most significant effect on CO_2 recovery, which aligns with the physics-based interpretation of the system. Increasing the solvent amount leads to increased absorption factor and enhanced mass transfer inside the absorption column. As a result, lower number of stages are required in this case, where the fixed cost will be decreased. However, care should be taken as raising the solvent-to-feed ratio means an increase in the circulation/recycle rates, which elevates pumping and other operating costs. This the reason why 4.5 molar ratio was found and selected as the optimum one. Causality analysis also confirmed the interactions between design variables which are in good agreement with the physics-based interpretation of physical absorption. For instance, it was found that changing the absorption pressure strongly affects the selection of solvent-to-feed molar ratios, where raising the pressure tends to select lower values of molar ratios. Increasing the pressure affects the solvent-gas equilibrium so that more gas molecules will be dissolved in the liquid solvent phase, hence, lower solvent amounts will be required. However, decreasing the molar ratios at higher pressures can lead to an increase in absorption column number of stages as a compensation to get the required recovery. This illustrates the importance of the

developed causal incremental RL methodology in the current study to optimize this multi-objective process with competing design variables. Fortunately, it succeeded to achieve optimal values of both capture cost and relative total emissions at reasonable values of design variables (4.5 molar ratio, 9 stages and 7.5 bars absorption pressure) while considering their interactions.

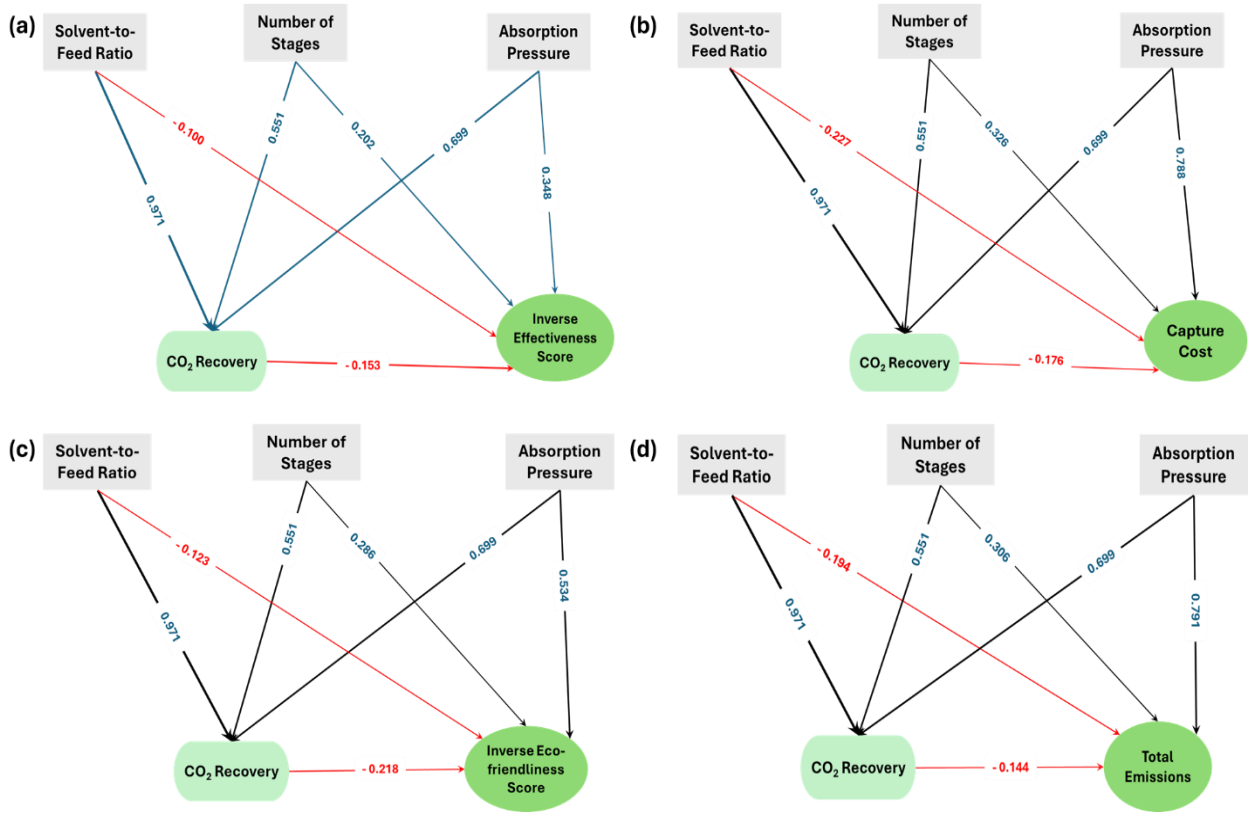


Figure 6.12 Causal graphs for **(a)** inverse capture effectiveness, **(b)** capture cost, **(c)** inverse capture eco-friendliness and **(d)** relative total emissions of DEPG-based carbon capture process

The empirical equations relating capture cost and total emissions to the input variables in this case of DEPG-based capture process are presented by **Equations 6.13 & 6.14**. Similar to the case of MEA-based process, the compartmental modeling-like approach was adopted. **Equations 6.13** determines the capture cost in the cost range of (0-0.11 USD/kg CO_2) and CO_2 recovery (> 90%) corresponding to inverse capture score range of (0-0.125). **Equations 6.14** can be used to calculate the relative total emissions based on the design variables in the emissions range of (0-0.0019 kg CO_2 emitted/kg CO_2 captured) and CO_2 recovery (> 90%) corresponding to inverse capture score range of (0-0.00215).

$$\boldsymbol{Capture\ Cost} = (4.91e-03 \times P^{0.348} + 0.83 \times N_{ABS}^{0.202} - \frac{15.13}{SFR^{0.1}} - \frac{0.15}{P^{0.107}} + \frac{3.47}{N_{ABS}^{0.084}} + \frac{11.16}{SFR^{0.149}}) \times (CO_2\ Recovery\ \times CO_2\ Purity) \quad (6.13)$$

$$\boldsymbol{Total\ Emissions} = (3.58e-05 \times P^{0.534} + 3.19e-03 \times N_{ABS}^{0.286} - \frac{6.44e-02}{SFR^{0.123}} - \frac{5.4e-03}{P^{0.152}} + \frac{1.71e-02}{N_{ABS}^{0.12}} + \frac{4.4e-02}{SFR^{0.212}}) \times (CO_2\ Recovery\ \times CO_2\ Purity) \quad (6.14)$$

The *Catboost* model is developed with the aid of the results of the causality analysis where each design variable is multiplied by a combined coefficient representing both direct and indirect effects. The determination coefficients (R^2) obtained took the value of around 99.99 with relatively low MSEs of (< 1e-5) in both cases of capture cost and relative total emissions.

6.6.4 Preliminary sensitivity analysis

This section performs a comparison between different scenarios for carbon capture based on the source of energy and the location of the plant. Several attempts were made to evaluate the possible capture costs and total emissions produced from the different capture processes. Table 6.4 summarizes the different scenarios/processes considered in this study. It should be noted that each scenario was applied to both *Quebec* and *Alberta*, where they represent two extremes on the scale with respect to emissions factor, energy sources and prices. It is worth noting that the CO_2 recovery is higher than 90% and the purity of the captured CO_2 is higher than 95% in all scenarios defined below.

Table 6.4 Different scenarios

Scenario/process ID	**Brief Description**
Process_1	MEA-based capture, steam produced by natural gas to be used for heating and electricity is used directly for driving the compressor(s)
Process_2	MEA-based capture, steam produced by electricity to be used for heating and electricity is used directly for driving the compressor(s)
Process_3	MEA-based capture, steam produced by natural gas to be used for heating and steam is used for driving the compressor(s)
Process_4	MEA-based capture, steam produced by electricity to be used for heating and steam is used for driving the compressor(s)
Process_5	MEA-based capture, steam produced by natural gas to be used for heating and natural gas is used for driving the compressor(s)
Process_6	MEA-based capture, steam produced by electricity to be used for heating and natural gas is used for driving the compressor(s)
Process_7	DEPG-based capture and electricity is used directly for driving the compressor(s)
Process_8	DEPG-based capture and natural gas is used directly for driving the compressor(s)
Process_9	DEPG-based capture and steam is used for driving the compressor(s) where the steam is produced by natural gas
Process_10	DEPG-based capture and steam is used for driving the compressor(s) where the steam is produced by electricity

Results of the optimal processes are summarized in Figure 6.13. Processes depending on the utilization of electricity are eco-friendlier specially in the case of *Quebec* as the source of energy for electricity generation is hydropower, which has low emissions factor. DEPG-based capture process is generally cleaner and eco-friendlier as it does not involve regeneration based on supplying any sources/forms of thermal energy. On the other hand, MEA-based capture process has significantly higher total relative emissions due to the dependence on steam as a thermal energy source for solvent regeneration. However, MEA-based process in general is more cost effective compared to DEPG-based process except for rare cases, where stripping steam is produced in *Alberta* via electric power. These results suggest that it is preferable to adopt MEA-based capture process in *Alberta* where steam is produced by natural gas-powered boilers and compressors are driven by steam or natural gas combustion. Whereas, in *Quebec* DEPG-based capture process would be a promising choice due to being cleaner and having comparable costs compared to MEA-based process.

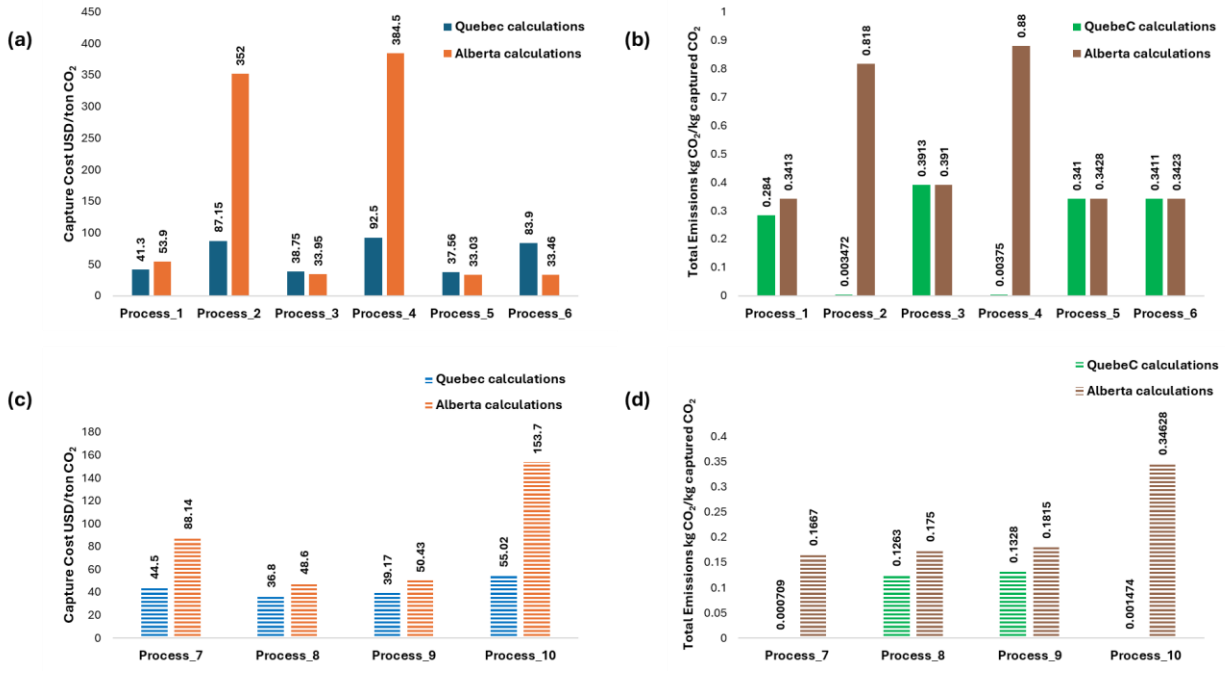


Figure 6.13 Preliminary sensitivity analysis results

6.7 Remarks, recommendations and future work

The recommendations and future work perspectives of this study can be divided into two main branches, i.e. data science and technological viewpoints. Regarding the data science viewpoint, it is recommended to adopt, in the future, new algorithms such as multi-arm bandit and other learnable optimization algorithms and comparing them with DDPG in terms of optimization time and final results within the same range of operation/design. New problem formulation can be then suggested in these future studies to help accelerating the optimization. Additionally, new reward formulations can be designed to help the agents achieve optimal values of specified KPIs in shorter times. It is highly recommended to design new more efficient metrics to fairly compare the carbon capture processes/technologies, which can be also incorporated in the reward definition such as those suggested in this study, i.e. capture effectiveness and eco-friendliness scores. In addition, other noise forms/algorithms such as Gaussian noise can be tested and compared to the performance of OU noise for exploration in future works.

On the technological side, for a cleaner carbon capture, it is recommended to adopt the DEPG-based carbon capture in *Quebec.* Besides, any steam required for the capture process should be produced by electricity which is generated through clean hydropower. On the other hand, MEA-based capture process is more efficient, from the capture cost point of view, to be adopted in *Alberta* where the natural gas prices are relatively low compared to *Quebec*. In this case, natural gas will be used to produce steam and drive directly the compressor required to raise the pressure of inlet flue gas stream. A detailed study considering different provinces in *Canada* and other regions around the world should be conducted in the near future to determine the best capture process for each region. In this study, variability and uncertainties in energy prices and other aspects should be considered for robust and informative analysis. Our developed methodology for causality analysis can be adopted in this future study. It will play a key role in extracting insights about the most significant variables that affects the change in capture costs and total emissions of various capture process in different regions in uncertain situations.

6.8 Conclusion

This study presents a simulation-based optimization framework for process design, leveraging causal incremental reinforcement learning (RL) to systematically explore and optimize design variables. The RL agent functions as an intelligent optimizer, searching a predefined design space

for optimal configurations. To enhance decision-making, we integrated formal concept analysis, Granger causality, and causal Bayesian networks, enabling a well-informed causality analysis. The insights gained from this analysis were incorporated into the RL agent's learning process, allowing it to make more rational design choices based on causal relationships. The proposed methodology was applied to two industrial carbon capture processes, MEA-based and DEPG-based, optimizing two key performance indicators: capture cost and total relative emissions. For the MEA-based process, an optimal capture cost of 41.3 USD/ton CO_2 was achieved, representing an 8.14% reduction from the expert-provided baseline of 44.96 USD/ton CO_2. The optimal total relative emissions were 0.28 kg CO_2 emitted/kg CO_2 captured, obtained at design conditions of 3 bar absorption pressure, 25.5 °C flue gas temperature, 31 °C lean MEA temperature, and a boil-up ratio of 0.087 (molar). Similarly, for the DEPG-based process, the optimal design variables were identified as 7.5 bar absorption pressure, a solvent-to-feed ratio of 4.5, and nine absorption stages, resulting in a capture cost of 44.5 USD/ton CO_2 and total relative emissions of 0.00071 kg CO_2 emitted/kg CO_2 captured. Beyond carbon capture, this methodology can be extended to other industrial absorption and chemical processes. Furthermore, all extracted causal graphs, dependencies, and design patterns are systematically stored to facilitate transfer learning, enhancing the generalizability of the approach.[4]

[4] The supplementary materials related to this paper can be found in (Appendix J)

CHAPTER 7 GENERAL DISCUSSION, CONCLUSION AND RECOMMENDATIONS

This doctoral research developed a human-centric, physics- and causality AI-assisted framework for the accelerated design and optimization of industrial processes and materials in support of industrial decarbonization and sustainable value chains. The proposed framework bridges advanced AI with industrial engineering principles to overcome key limitations of conventional process and material design/optimization approaches, namely rigidity, high computational cost, and limited adaptability to complex integrated systems.

The research contributions are structured around three interconnected methodological pillars:

1) Interpretable and explainable machine learning (IML+XAI) for transparent understanding and selection of material and processes understanding;

2) Physics- and chemistry-aware generative AI for accelerated discovery of sustainable materials and process configurations and

3) Causal incremental reinforcement learning (CIRL) for simulation-based process optimization.

Together, these pillars establish a coherent foundation for trustworthy, physically consistent, and data-efficient AI-enabled decision support framework in complex industrial systems.

7.1 Advancing AI Methodologies for Industrial Systems Design Optimization

A central methodological contribution of this thesis is the systematic integration of interpretability, explainability, physical consistency, and causality into AI-driven industrial design workflows. The proposed IML/XAI ensemble framework combines intrinsically interpretable models with post-hoc explainability techniques, enabling both global and local understanding of model behavior while preserving high predictive accuracy. Unlike purely black-box approaches or strictly interpretable but limited models, this hybrid strategy achieves a balanced trade-off between predictive performance, transparency, and knowledge extraction.

The resulting framework supports the development of data-driven design rules linking material properties and process variables to key performance indicators (KPIs) such as capture capacity,

solubility, and efficiency. These insights transform AI outputs into process and engineering knowledge, strengthening trust and adoption in industrial settings.

To further enhance robustness and scalability, the research introduced NNMF-based bi-clustering combined with nonlinear clustering techniques for feature selection and dimensionality reduction. This enabled simultaneous clustering of features and samples, improving model interpretability and robustness while reducing dimensionality from thousands of variables to a few dozen key descriptors. Such data reduction not only enhanced model performance but also strengthened computational scalability, facilitating real-time applications in industrial analytics.

Building on these predictive and interpretative capabilities, the research developed a physics- and chemistry-aware generative AI framework for the sustainable material and process discovery. The approach integrates GANs, VAEs, and transformer-based models within an ensemble architecture constrained by chemical feasibility rules. These constraints were incorporated through informed auto-labeling, physics-guided validation, and hybrid modeling that ensures generated molecular structures are chemically feasible and aligned with sustainability targets. Unlike conventional data-driven generative approaches that rely solely on pattern learning, this hybrid design enforces physical consistency, thereby advancing AI for science beyond empirical correlations. A human-in-the-loop mechanism was incorporated during generation and evaluation to enhance robustness and trustworthiness. The framework successfully generated novel, valid, and environmentally preferable solvents for CO_2 capture and lignin dissolution with high consistency, reliability and interpretability relative to targeted KPIs.

The third contribution, causal incremental reinforcement learning (CIRL), addresses long-standing challenges in simulation-based process optimization. By incorporating causal reasoning into reinforcement learning, the proposed methodology enhances sample efficiency, interpretability, and transferability within a unified framework. . The approach couples an RL agent directly with physics-based process simulators, ensuring physically realistic learning, while incremental optimization decomposes complex design problems into tractable sub-tasks, substantially reducing computational costs. Causal reasoning, based on the combination of formal concept analysis (FCA), Granger causality, and causal Bayesian networks (BNs), guides exploration by identifying true directional dependencies between design variables and KPIs, thereby reducing search space and accelerating convergence toward physically meaningful optima.

7.2 Industrial Relevance and Decision-Support Value

From an industrial engineering perspective, the developed AI-based methodologies were validated across several decarbonization case studies. For material selection, the IML+XAI framework enabled accurate classification of carbon capture adsorbents and identification of design rules relating pore structure and surface properties to capture performance. For process optimization, ensemble regression models achieved near-perfect prediction of solubility parameters for bio-based solvents, reducing model error while providing interpretable explanations of solvent behavior. The generative AI framework expanded these capabilities by designing novel solvent molecules that satisfied multi-objective KPIs, including environmental safety, flammability, and cost. These results demonstrate the framework's capacity to transition from model development to actionable decision-support tools for engineers, accelerating material discovery and process design under sustainability constraints.

In simulation-based optimization, the CIRL methodology delivered tangible performance improvements. Applied to real flue gas conditions from cement and power generation plants, the optimized monoethanolamine (MEA) absorption-based CO_2 capture systems achieved an 8.14% cost reduction, while dimethyl ethers of polyethylene glycol (DEPG) system reached a capture cost of 44.5 USD per ton of CO_2 with minimal emissions. These results were validated by industrial experts, confirming the robustness and practical relevance of the developed framework. The modular nature of CIRL allows its extension to other process systems, supporting adaptive, data-informed decision-making throughout the industrial lifecycle.

7.3 Expected Social Impacts and Economic Benefits

This doctoral research contributes to sustainable industrial transformation by improving knowledge accessibility, workforce empowerment, and responsible AI integration within decarbonization pathways. Its impact extends beyond optimization performance to fostering transparent, data-driven decision practices that strengthen long-term economic resilience under Net-Zero transitions.

- Open Knowledge Infrastructure and Data Transparency: the consolidation and standardization of carbon capture materials data into an accessible database improves research efficiency, reduces duplication of effort, and fosters collaboration between

academia and industry. This shared knowledge infrastructure supports transparency and evidence-based material screening and accelerates innovation cycles.

- Knowledge Extraction and Human Expertise Augmentation: the extraction of interpretable design rules and causal knowledge structures enhances expert understanding rather than replacing it, facilitating skill transfer, informed decision-making, and trust in AI-assisted systems within engineering teams.
- Sustainable Material and Technology Adoption: By enabling structured sustainability evaluation, e.g. environmental and safety-oriented ranking criteria, the framework supports the selection of safer and more resource-efficient materials, encouraging responsible industrial adoption of green alternatives.
- Economic Competitiveness and Risk Reduction

AI-assisted adaptive optimization reduces uncertainty, lowers experimentation and development costs, and improves investment confidence in decarbonization technologies, thereby supporting economically viable and scalable climate solutions.

Collectively, these contributions advance responsible and economically sustainable industrial decarbonization aligned with *Canada's* strategic climate objectives, strengthening national innovation capacity and clean technology competitiveness. By enabling evidence-based decision making and scalable deployment of AI-enhanced low-carbon technologies, they support long-term industrial resilience and effective climate action.

7.4 Cross-Perspective Integration and Human-Centric Design

The contributions of this thesis lie primarily in the integration, adaptation, and application of existing AI methodologies within a unified, physics-aware framework for industrial design, rather than in the development of entirely new standalone algorithms. The novelty resides in the way interpretable ML, generative modeling, and reinforcement learning are synergistically combined with domain knowledge, human expertise, and process constraints to address complex engineering problems. In addition, the framework explicitly incorporates a human-centric perspective by ensuring that expert judgment, usability, transparency, and decision support remain central to the proposed solutions. This systems-level integration represents a meaningful contribution in applied industrial AI, where practical deployment requires not only algorithmic performance but also interpretability, consistency, user trust, and domain alignment.

The integration of these methodologies reflects a coherent human-centric AI philosophy, where human expertise is essential across all stages of model design and application. Experts define meaningful reward functions, validate generated materials, and interpret causal graphs, ensuring that AI operates as an intelligent collaborator rather than a fully autonomous system. This human-in-the-loop architecture enhances reliability, accountability, and adoption potential in industrial contexts, particularly where safety, feasibility, and interpretability are paramount. By unifying interpretable, generative, and causal learning within a single framework, the research proposes a new paradigm for AI-augmented industrial decision-making, in which human reasoning and machine intelligence are mutually reinforcing.

7.5 Key Clarifications and Physical Interpretations of Published Results

While the preceding chapters present the published works and their results, several aspects merit further clarification and deeper physical interpretation. The following discussion addresses key limitations, assumptions, and conceptual nuances highlighted during the thesis examination, with the aim of strengthening the scientific consistency and practical relevance of the proposed framework.

The analysis in Chapter 3 captures key structural and operating variables governing adsorption; however, two important limitations must be acknowledged. First, the absence of moisture (H_2O partial pressure) reflects inconsistencies in available data in literature, and thus the models should be interpreted as representative of dry or weakly humid conditions. Given the known effects of moisture, competitive adsorption, pore blocking, and material degradation, its inclusion is essential for realistic CCUS applications [634]. Second, the datasets exhibit non-uniform distributions, particularly for pressure and temperature, with strong clustering near atmospheric conditions and sparse coverage of intermediate regimes. As a result, model predictions in underrepresented regions rely on interpolation with limited physical constraints and should be treated with caution. These limitations highlight the need for targeted data generation and/or physics-informed constraints to improve robustness and generalizability. Additionally, several interpretations derived from the models also require careful contextualization. Correlations between structural properties (e.g., pore size, porosity) and regenerability should not be viewed as causal, as regenerability is a system-level outcome influenced by process conditions such as heat transfer, reactor configuration, and

regeneration strategy. Similarly, adsorption selectivity cannot be explained solely by kinetic diameter differences between CO_2 and N_2; rather, it is largely governed by electrostatic interactions arising from the higher quadrupole moment of CO_2, particularly in microporous or functionalized materials. In addition, the identified importance of CO_2 partial pressure must be interpreted in terms of influence magnitude rather than benefit, as its correlation with selectivity is negative. The strong correlation between total pressure and CO_2 partial pressure is physically expected and does not provide independent insight. Moreover, the observed preference for Ca- and Li-based materials at elevated temperatures is consistent with a transition from physisorption to chemisorption mechanisms, where CO_2 reacts with metal oxides to form stable carbonates under thermodynamically favorable high-temperature conditions. While Chapter 3 demonstrates that interpretable ensemble learning can achieve strong predictive/descriptive performance and extract useful design rules, these results should be understood within the context of data limitations and underlying physical mechanisms. This combined perspective forms the basis for the more advanced, physics-guided generative and optimization frameworks developed in subsequent chapters. On the methodological side, training-testing splits as well as validation set in Chapter 3 were primarily constructed using random partitioning to ensure statistical representativeness, as described in the chapter. While this approach is standard, it does not explicitly test distributional generalization (e.g., across material families). This limitation is acknowledged, and stratified partitioning strategies were also explored and successfully leveraged in subsequent studies. Hence, future works will incorporate this partitioning mechanism. The same was applied in Chapter 4, random split (primarily) and modified stratification to fit the regression problem were employed along with k-fold cross validation to avoid overfitting and distributional shift problems.

Chapter 4 is primarily intended to introduce a generalized ensemble-based machine learning framework for predicting Hansen solubility parameters (HSP) and demonstrating its applicability across different domains, rather than to exhaustively optimize solvent systems for specific industrial processes. Accordingly, the current implementation focuses on physical solubility descriptors and available datasets, which explains the limited representation of chemically reactive systems such as amine-based CO_2 capture, ionic liquids [635], or potential lignin-solvent reactions. Extending the framework to explicitly account for reactive absorption (e.g., carbamate formation), electronic interactions, or solvolysis would require augmenting the feature space with quantum-chemical descriptors and enriching the training database with targeted data from experiments or

first-principles simulations. Similarly, the inclusion of ionic liquids or industrially validated solvent families is a data availability and scope consideration rather than a methodological limitation. Importantly, the proposed framework is modular and data-driven, and can be readily adapted or expanded, through database curation, feature engineering, or hybrid physics-informed integration, to address chemically complex systems and industrially relevant solvent design problems in future work. On the methodological side, all reported values of accuracy, R^2 and MSE/RMSE correspond to testing performance rather than training performance. During model development, high training accuracy and R^2 values (> 0.99) with minimal error were targeted to ensure effective learning; however, careful safeguards, including cross validation and the use of separate testing and validation datasets, were implemented to avoid overfitting. The same principle also applies to Chapter 3. The ensemble fusion strategy in Chapter 4 combines predictions from multiple base learners using a structured multi-stage aggregation approach designed to improve robustness and accuracy. Specifically, individual model outputs are first combined within homogeneous groups, followed by cross-model aggregation, and finally refined through a higher level fusion step to balance model biases and leverage complementary strengths. This hierarchical fusion allows progressive error reduction and improved generalization compared to single-step aggregation. The multi-stage fusion is designed to progressively reduce model-specific errors and improve robustness by first aggregating similar learners to stabilize predictions, then combining diverse model outputs to capture complementary patterns, and finally applying a higher-level weighting to optimize overall performance. In practice, this is implemented through sequential averaging, weighted aggregation and learnable fusion of predictions across these stages, as outlined in Section 4.3 and Figure 4.1. With respect to features selection, specifically the adoption of clustering techniques to select the optimal number of components for the NNMF graphically, the silhouette test, originally introduced by Peter J. Rousseeuw [636], is used to evaluate clustering quality by measuring how well each data point fits within its assigned cluster relative to other clusters. Specifically, it quantifies the balance between intra-cluster cohesion and inter-cluster separation, with higher silhouette values indicating more well-defined clustering structures. This metric was employed in Section 4.5.2 to support the selection and validation of the clustering results. In addition, regarding the UMAP visualization, the apparent overlap between clusters in Figure 4.10 arises from the projection of high-dimensional data onto a lower-dimensional (2D) space, which can obscure true separations. As discussed in Chapter 4, clustering was performed in a higher-

dimensional feature space, and the inclusion of 3D visualizations was intended to better reveal this separation. Consequently, points that appear overlapping in 2D may still be correctly assigned based on their position in the full feature space.

Chapter 5 is intended to demonstrate a generative AI framework for exploring solvent design spaces guided by Hansen solubility parameters (HSP), rather than to deliver a complete inverse design-to-synthesis pipeline. Accordingly, the work focuses on data-driven generation, screening, and validation at the computational level, and does not extend to experimental synthesis or laboratory realization of candidate molecules. The representation of solvents using SMILES strings reflects standard practice for molecular systems; however, for room-temperature ionic liquids (RTILs), this representation should be enhanced to fully capture their paired cation-anion nature. These aspects highlight practical limitations of current datasets rather than conceptual shortcomings of the methodology. Importantly, the proposed framework remains modular and extensible: it can be enhanced by incorporating explicit cation-anion representations for RTILs, expanding databases with chemically valid solvent families, and integrating first-principles or experimental validation steps to enable fully closed-loop inverse design workflows in future research. On the methodological side, synthesis feasibility is indeed a critical factor in evaluating candidate molecules, and in this work it is addressed through a post-generation retrosynthesis analysis step. This modular separation reflects practical considerations, as generative design and synthesis planning rely on different data sources and modeling approaches. While full integration of synthesis prediction within the generative loop is conceptually appealing, it would significantly increase model complexity and computational cost. The current framework therefore adopts a sequential approach, where candidates are first generated and screened, then evaluated for synthetic feasibility. This strategy, as implemented in Chapter 5, provides a balance between exploration efficiency and practical validation, while remaining extensible toward tighter integration in future work. It is worth noting that Molecular Validity in this chapter refers to the extent to which AI-generated molecular structures correspond to chemically feasible compounds that satisfy fundamental rules of chemical bonding and valency. In practical terms, a valid molecule is one that can exist according to established chemical principles, without structural inconsistencies or impossible configurations. In Chapter 5, validity is quantified using standard cheminformatics checks applied to generated SMILES strings, as detailed in the chapter and the associated

publication. This metric ensures that the generative model produces not only novel but also chemically meaningful candidates.

Chapter 6 employs process simulations based on conventional rate-based or equilibrium models in Aspen HYSYS, which rely on simplified assumptions such as plug flow and idealized packing behavior. While these models are widely used and computationally efficient for process design and optimization, they do not explicitly capture detailed hydrodynamics governed by packing geometry, maldistribution, or transient effects. The integration of higher-fidelity approaches, such as 2D/3D or dynamic computational fluid dynamics (CFD) coupled with user-defined reaction kinetics, represents a promising direction toward more realistic digital twins of absorption systems. However, such integration can introduce significant computational cost and complexity, which can limit its direct use within iterative optimization frameworks such as reinforcement learning. In this work, the adopted modeling strategy reflects a practical trade-off between fidelity and scalability. Nevertheless, the proposed AI-based optimization framework is inherently modular and can, in principle, be coupled with higher fidelity CFD-based models or reduced order surrogates derived from them, enabling future extensions toward more physically detailed and reliable process representations. On the methodological side, the agent in this chapter is defined as the decision-making component that interacts with the simulator by proposing actions and receiving rewards following standard reinforcement learning formulation. The reward structure defined in Chapter 6 incorporates multiple objectives to reflect process performance, economic, and environmental criteria. While such composite rewards can be challenging to optimize, the formulation was iteratively refined to ensure stable learning behavior and meaningful policy convergence. As discussed, normalization and weighting strategies were applied to balance competing objectives and avoid dominance of individual terms. The results indicate that the agent was able to learn effective policies under this structure; yet, simplification or adaptive reward shaping can be investigated in future implementations based on the problem in hand. The causal graphs presented in Figures (6.9) and (6.12) were obtained by first constructing the DAG, followed by Bayesian network analysis. The resulting probabilistic relationships were subsequently transformed into the numerical values shown on the arrows, providing an interpretable indication of the relative influence of each design variable on other variables and on the final KPIs, namely costs and emissions.

7.6 Limitations and Comparative Context

This thesis spans multiple domains, including interpretable ML, generative modeling, and reinforcement learning. Such breadth is intentional and directly aligned with the objective of developing an integrated AI-assisted design framework. Rather than optimizing a single algorithmic component in isolation, the work emphasizes the orchestration of complementary methodologies across the design pipeline, from prediction to discovery and optimization. As a result, certain components are developed to a level sufficient for integration and validation within the overall framework, while opportunities for deeper methodological refinement are identified as avenues for future work.

Additionally, it is worth noting that the high predictive and classification performances reported across several chapters should be interpreted within the context of the datasets and modeling assumptions employed. The use of heterogeneous data sources, extensive preprocessing, and multi-stage modeling pipelines introduces potential sources of bias and limits the direct generalizability of the results. While cross-validation and external validation were performed where possible, further assessment under unseen conditions and with independent datasets would be required to fully establish robustness. Accordingly, the results are best viewed as demonstrating the effectiveness and promise of the proposed methodologies, rather than definitive predictive capability across all operating regimes.

While the developed methodologies demonstrated significant results, direct benchmarking against alternative approaches, such as standalone XAI models, physics-informed neural networks, and other conventional RL optimizers, was limited. In several cases, comparisons were conducted against standard or representative models; however, a more systematic evaluation across diverse baselines and datasets would strengthen the generality of the conclusions. This limitation is partly due to differences in data availability (where the data acquired was a mix between literature data, simulated data and confidential datasets provided by industrial partners), preprocessing requirements, and problem formulations across studies, which complicate direct comparisons. Future studies should include systematic comparisons with representative baselines, conduct structured comparative analyses to quantify gains in accuracy, interpretability, and computational efficiency.

It should be noticed that although extensive benchmarking is not always presented explicitly within the main chapters (3 to 6), a systematic comparison with existing studies was conducted during the development of this work. In particular, relevant literature on the prediction of adsorption capacity, selectivity, and Hansen solubility parameters as well as generating new molecules was reviewed and summarized in the supplementary materials associated with the published articles (available online and the links are mentioned in footnotes after each chapter). These compilations include model types, datasets, and reported performance metrics, providing a contextual baseline for evaluating the proposed approaches. For instance, in the case of capture materials evaluation based on their capacities and selectivities, the majority of the research studies presented in the literature focused on predicting the values of these two KPIs through black-box models (with prediction accuracies around 90%) without taking into consideration extracting the interpretable patterns to reach high performances to reduce the vast search space and enrich human expert's knowledge to accelerate carbon capture material design. Additionally, in the case of HSPs prediction, the accuracy ranged from 0.35 to 0.93 in most cases with RMSE around 0.8 while models used are black-box as well. From this literature survey and based on discussions with industrial partners as well as the available data, the baseline accuracy for capture materials evaluation was put as 90% and for predicting HSPs was put as 95% with minimal possible error (RMSE< 0.15) while maintaining transparent modeling. However, direct one-to-one benchmarking remains challenging due to differences in datasets, feature definitions, preprocessing strategies, and problem formulations across studies. Consequently, the comparisons should be interpreted as indicative rather than strictly equivalent. Future work would benefit from standardized datasets and evaluation protocols to enable more rigorous and reproducible benchmarking. Fortunately, the developed methodologies in this doctoral research were adopted in other applications (biomass valorization and zeolitic materials experimental synthesis) and fine-tuned using domain-specific data, where the obtained accuracies and results were high and comparable with those reported in the literature, supporting the generalizability of the proposed framework across applications. However, care should be taken during data collection (representative data with reasonable size) and preprocessing (human guidance and features extraction/engineering) as the data quality and quantity are among the most important aspects of having good accuracies. In addition, k-fold cross validation and reasonable selection of data for testing and validation are essential steps to avoid any overfitting. Foe instance, during this doctoral research, 5-10 fold cross validation was

employed along with robust hyperparameters optimization (Bayesian optimization) to ensure avoidance of overfitting. Furthermore, the testing size was around 15-20% of original data and always separate validation datasets (around 10% of original dataset's size) were employed to avoid overfitting in the range of the study and under the assumptions/baselines put in the beginning of the study. While validation strategies such as cross-validation and external testing were employed where possible, the generalization of these models beyond the studied domains, particularly in sparsely populated regions of the input space, remains uncertain. Therefore, the results are best viewed as demonstrating strong performance within defined conditions, rather than universally transferable predictive capability.

Integration between AI tools and legacy physics-based simulators also posed challenges, as these tools were not originally designed for unified AI coupling. Although effective solutions were implemented, re-engineering simulation interfaces will be necessary to further streamline optimization workflows. Scalability from pilot-scale studies to full industrial deployment remains another challenge, particularly under data sparsity and real-time constraints.

Nevertheless, when our CIRL-based methodology applied for process design it accelerated the whole design process. For the selected carbon capture case study, it reduced the time needed to perform the optimization from several weeks (using the static optimizer embedded in the simulator and/or following the traditional procedure of building surrogates and couple them with non-learnable optimizers) to only four days, while simultaneously extracting causal insights transferable to other processes. These four days include optimizing two processes not only one and most importantly extracting the causal relationships between different variables enriching the designer's knowledge for future considerations. More information about the models and optimization strategies developed in the literature for process design and control optimization are summarized in the related work section in Chapter 6 and in Table (E.1) in Appendix E. In this related work section, the merits of the developed methodology (CIRL) in this doctoral research over the existing ones in literature (used as practical baseline for comparison) are also clearly highlighted. These comparisons are intended to position the proposed framework within the broader state of the art rather than to provide an exhaustive or fully standardized benchmarking across all existing methods.

In addition, to summarize the high-level merits of the developed approaches in this thesis, Table 7.1 provides a comparative assessment of three dominant paradigms for material and process design and optimization. These paradigms are (i) traditional physics-based modeling coupled with classical optimization techniques, (ii) purely data-driven modeling approaches integrated with optimization, and (iii) the hybrid physics-guided AI methodologies developed in this thesis. This high-level comparison focuses on criteria critical for industrial decision-making, including predictive reliability, interpretability, computational scalability, suitability for early-stage design, and adaptability to complex, multi-objective decarbonization problems. Instead of positioning any paradigm as entirely superior, the table highlights their respective strengths and limitations, thereby clarifying the design contexts in which hybrid modeling approaches can complement and extend established methods.

As summarized in the table, traditional physics-based approaches remain indispensable for high-fidelity analysis and final design validation, while purely data-driven methods offer computational efficiency and flexibility when sufficient data are available. However, both paradigms exhibit limitations when applied independently to early-stage, large-scale, and highly uncertain design problems typical of industrial decarbonization. The hybrid physics-guided AI methodologies developed in this thesis are not intended to replace established methodologies, but rather to bridge their gaps by combining physical consistency, data-efficiency, interpretability, and scalable optimization. This integration enables informed decision-making across the full design lifecycle, from conceptual screening to simulation-based optimization, thereby supporting more reliable and accelerated development of sustainable materials and industrial processes.

Table 7.1 Comparison of materials and processes design optimization paradigms

Criterion	**Physics-Based Modeling + Classical Optimization**	**Purely Data-Driven Modeling + Optimization**	**Hybrid Physics-Guided AI Methodologies (This Thesis)**
Primary modeling principle	First-principles equations (thermodynamics, kinetics, transport phenomena)	Statistical learning from historical or simulated data	Integrated use of physical constraints, causal knowledge, and data-driven learning
Data dependency	Low (requires physical parameters and assumptions)	High (requires large, representative datasets)	Moderate (leverages limited data augmented by physics and simulation)
Physical consistency	High by construction	Not guaranteed; must be enforced post hoc	Explicitly embedded through physics-guided learning and simulation coupling
Interpretability and transparency	High (equation-based)	Often limited (black-box behavior)	High to moderate, enabled by IML-XAI and causal analysis
Predictive accuracy in complex, nonlinear systems	High when models are well specified; degrades with uncertainty	High within data domain; limited extrapolation	High and more robust under uncertainty due to hybridization
Scalability to large design spaces	Limited by model complexity and computational cost	High, but constrained by data quality and availability	High, supported by surrogate modeling, ensembles, and incremental learning
Suitability for early-stage design and screening	Limited (requires detailed specifications)	Moderate (data availability dependent)	High, enabling rapid screening under uncertainty

Table 7.1 Comparison of materials and processes design optimization paradigms (Cont'd and end)

Criterion	**Physics-Based Modeling + Classical Optimization**	**Purely Data-Driven Modeling + Optimization**	**Hybrid Physics-Guided AI Methodologies (This Thesis)**
Suitability for early-stage design and screening	Limited (requires detailed specifications)	Moderate (data availability dependent)	High, enabling rapid screening under uncertainty
Optimization capability	Mature, but sensitive to model fidelity	Efficient, but may converge to non-physical solutions	Physics-guided, causality-aware optimization ensuring feasible solutions
Adaptability and transferability	Low to moderate; requires re-derivation	Limited across domains	High, supported by causal learning and incremental updates
Handling of uncertainty and variability	Explicit but computationally expensive	Implicit and data-dependent	Explicit and efficient via ensemble learning and CIRL
Role in industrial decarbonization	Validation and detailed design	Pattern discovery and prediction	End-to-end decision support: understanding, discovery, and optimization
Typical limitations	Computationally intensive; limited flexibility	Data hunger; lack of trust and extrapolation	Increased methodological complexity; requires careful integration

7.7 Future Research Directions

Future work will focus on enhancing scalability, adaptability, and computational efficiency. One promising direction is the integration of quantum-inspired optimization for hyperparameter tuning, replacing conventional grid search and Bayesian optimization. Preliminary proof-of-concept

results demonstrated up to 50% reduction in optimization time while maintaining or improving model performance, highlighting strong potential for further acceleration. For example, the time required for hyperparameters optimization of *CatBoost* regression model was reduced from 27 minutes to only 14 minutes when quantum-inspired optimization was applied instead of Bayesian optimization. The case study involved the prediction of hydrothermal carbonization process's performance based on biomass type and operating conditions as a representative pathway of carbon management and industrial decarbonization through biomass valorization [637]. In this case, the determination coefficient (R^2) was 0.9624 compared to 0.94 obtained by Bayesian optimization. The new approach developed and followed to achieve these results is summarized in Figure 7.1.

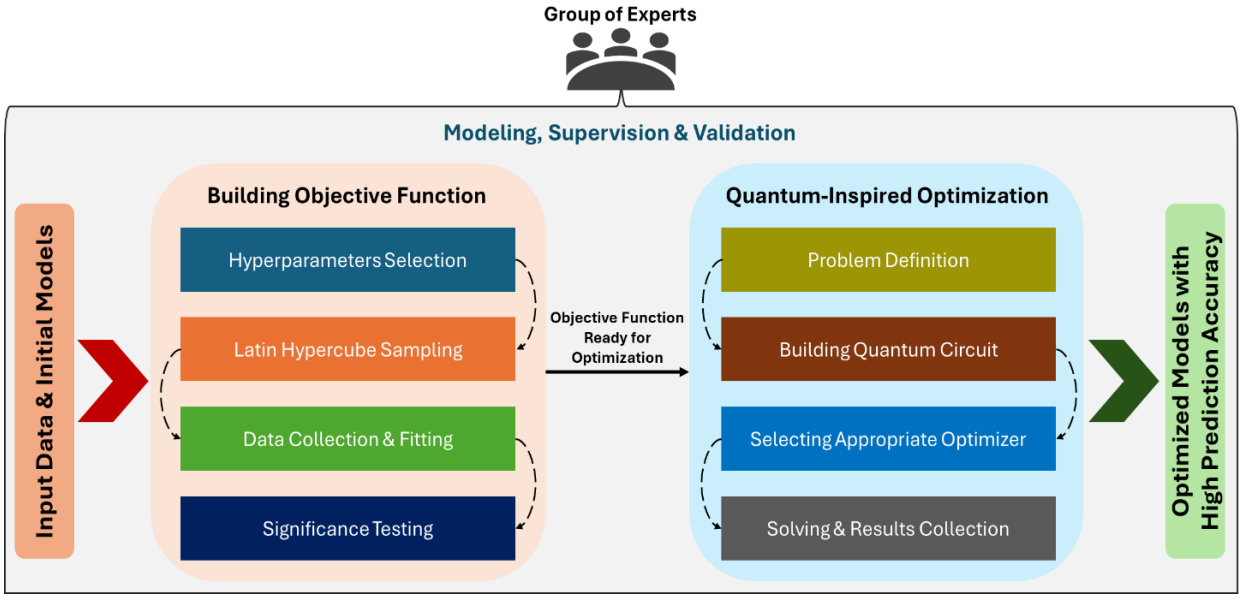


Figure 7.1 Quantum-inspired optimization for accelerated hyperparameters optimization

Continuous model updating through new data and expert feedback will also be explored, potentially supported by quantum-inspired techniques. The integration of quantum machine learning models into ensemble frameworks represents another avenue for future investigation.

On the generative side, expanding the ensemble to include models tailored to different complexity levels and coupling physics-informed GenAI with reinforcement learning in an end-to-end optimization framework will be pursued as illustrated in Figure 7.2. Experimental validation at bench and pilot scales will be essential to support industrial translation.

Additionally, the developed methodologies along this doctoral research could benefit from extensive ablation analysis in future studies for further improvement on the deployment level. Yet, during the work on developing an ensemble model for HSPs prediction, the descriptors/features selection, based primarily on the NNMF-based biclustering approach identifying coherent feature-sample patterns and helps prioritize the most informative descriptors, was guided by an empirical ablation analysis. In this procedure, subsets of features were iteratively evaluated to quantify their impact on model performance (e.g. XGBoost and Random Forest) for predicting Hansen solubility parameters. The selection of the top-ranked 25 descriptors per component of the NNMF reflects a performance plateau observed during this combined analysis, balancing predictive accuracy and model complexity. This hybrid strategy provided a practical, data-driven, and interpretable approach, as described in Chapter 4. As a result, the performance of these models increased from only (R^2 = 0.69-0.72) to (R^2 around 0.96) with the proper NNMF-based selection guided by the ablation analysis. Furthermore, this preliminary ablation study was done on two levels, the first one is the abovementioned features level and the second is the models level. In particular, at the model level, additional models were originally incorporated into the ensemble, namely CNN and GPR. However, based on the ablation study they were removed to keep high performance. In particular, the performance of CNN was (R^2 = 0.7) and GPR was (R^2 = 0.5) which deteriorated the performance of the whole ensemble and made it reach (R^2 below 0.85), significantly lower than the defined baseline of 0.95, while also increasing the model complexity.

In the case of simulation-based optimization, new learnable optimization techniques such as multi-arm bandit will be developed and compared to DDPG algorithm. This will be associated with possible new formulations of the design problem as well as new definitions of the reward function. Applying these methods on other applications and processes while considering processes dynamics for simultaneous industrial processes design and control optimization is another future research direction. In this case, where temporal effects are present, dynamic Bayesian networks and process mining will be added to the developed causality analysis framework to extract causal relationships and guide the considered learnable optimizers.

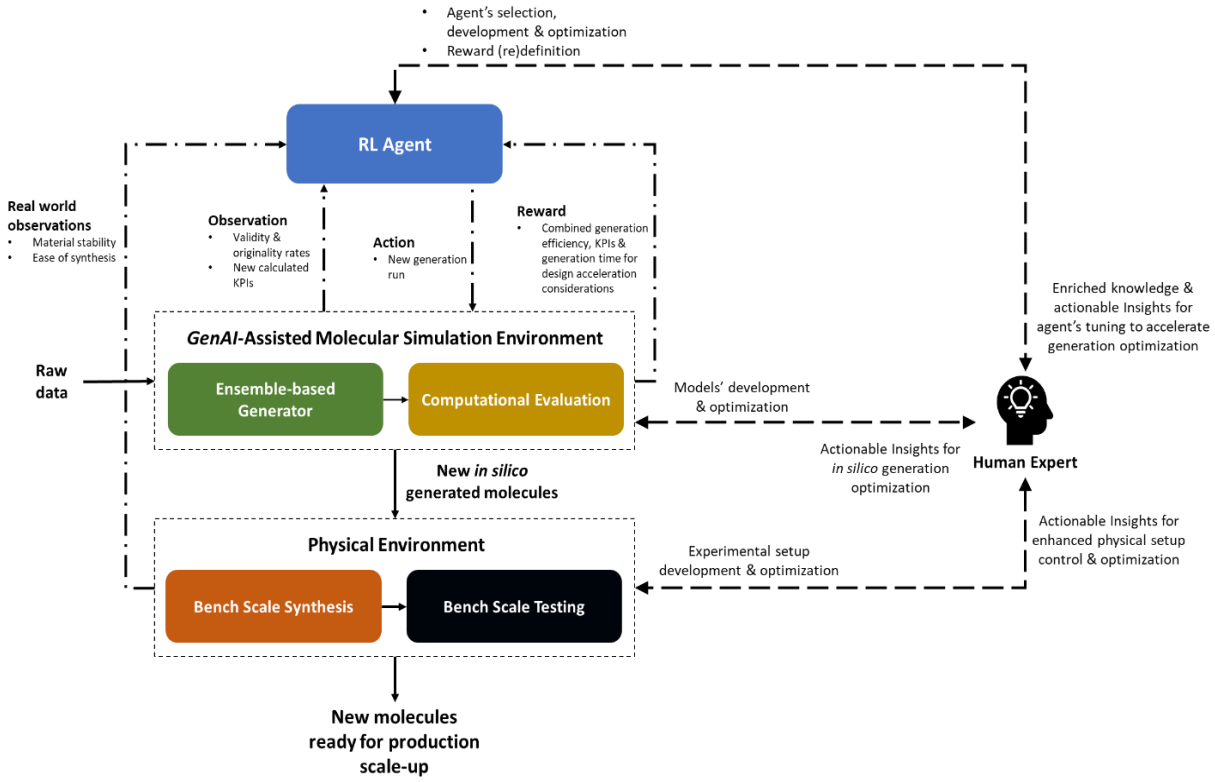


Figure 7.2 New research directions based on combined physics-informed GenAI and RL

In conclusion, this doctoral research demonstrates that integrating interpretability, physical realism, and causal reasoning into AI systems fundamentally enhances their value for industrial engineering applications. The proposed human-centric framework not only advances AI methodology but also delivers tangible industrial impact, supporting transparent, data-driven, and sustainable decision-making. By bridging AI innovation with industrial decarbonization challenges, this research work provides a robust foundation for future digital transformation efforts aimed at achieving *Net-Zero* emissions, with Canada positioned as a key contributor to global climate mitigation strategies.

REFERENCES

[1] The World Economic Forum, “How the industrial sector is turning net zero goals into practice,” 2024. https://www.weforum.org/stories/2024/06/industrial-sector-turning-net-zero-goals-into-practice/.

[2] A. de Pee, D. Pinner, O. Roelofsen, and K. Somers, “Decarbonization of industrial sectors: the next frontier,” 2018.

[3] De Nora, “The role of green hydrogen in decarbonizing steel production,” 2025. https://energytransition.denora.com/en/content/news-role-of-green-hydrogen-in-decarbonizing-steel-production.

[4] O. Roelofsen, K. Somers, E. Speelman, and M. Witteveen, “Plugging in: What electrification can do for industry,” *McKinsey \& Co.*, vol. 28, 2020.

[5] R. Hoggett, “Hydrogen, heat, and the decarbonisation of energy systems.” Issue September, 2020.

[6] P. Spiller, “Making supply-chain decarbonization happen,” *McKinsey \& Co.*, vol. 17, 2021.

[7] M. & Company, “Tackling heat electrification to decarbonize industry,” 2024. https://www.mckinsey.com/industries/industrials-and-electronics/our-insights/tackling-heat-electrification-to-decarbonize-industry.

[8] M. & Company, “Identifying opportunities and starting to build a new green business in the industrial sector,” 2022. https://www.mckinsey.com/industries/industrials-and-electronics/our-insights/identifying-opportunities-and-starting-to-build-a-new-green-business-in-the-industrial-sector.

[9] IEAGHG, “Techno-Economic Assessment of Small-Scale Carbon Capture for Industrial and Power Systems,” 2024. [Online]. Available: https://ieaghg.org/publications/techno-economic-assessment-of-small-scale-carbon-capture-for-industrial-and-power-systems/.

[10] IEA, “Net Zero by 2050 A Roadmap for the Global Energy Sector,” 2021. [Online]. Available: https://www.iea.org/reports/net-zero-by-2050.

[11] IEAGHG, “Cost of CO_2 capture in the industrial sector cement and iron and steel industries,” 2018. [Online]. Available: https://ieaghg.org/publications/cost-of-co2-capture-in-the-

industrial-sector-cement-and-iron-and-steel-industries/.

[12] Energy Transitions Commission, “Making Mission Possible: Delivering a Net-Zero Economy,” 2020. [Online]. Available: https://www.energy-transitions.org/publications/making-mission-possible/.

[13] and B. Z. Ajitesh Anand, Toralf Hagenbruch, Anoop Muppalla, “Tackling the challenge of decarbonizing steelmaking,” 2021. [Online]. Available: https://www.mckinsey.com/industries/metals-and-mining/our-insights/tackling-the-challenge-of-decarbonizing-steelmaking.

[14] J. Hack, N. Maeda, and D. M. Meier, “Review on CO2 capture using amine-functionalized materials,” *ACS omega*, vol. 7, no. 44, pp. 39520–39530, 2022.

[15] S. H. A. M. Leenders *et al.*, “Amine Adsorbents Stability for Post-Combustion CO2 Capture: Determination and Validation of Laboratory Degradation Rates in a Multi-staged Fluidized Bed Pilot Plant,” *ChemSusChem*, vol. 16, no. 24, p. e202300930, 2023.

[16] V. Pawar, S. Appari, D. S. Monder, and V. M. Janardhanan, “Study of the combined deactivation due to sulfur poisoning and carbon deposition during biogas dry reforming on supported Ni catalyst,” *Ind. \& Eng. Chem. Res.*, vol. 56, no. 30, pp. 8448–8455, 2017.

[17] D. Rolnick *et al.*, “Tackling climate change with machine learning,” *ACM Comput. Surv.*, vol. 55, no. 2, pp. 1–96, 2022.

[18] E. Pahija, S. Golshan, B. Blais, and D. C. Boffito, “Perspectives on the Process Intensification of CO2 Capture and Utilization,” *Chem. Eng. Process. Intensif.*, p. 108958, 2022.

[19] “CCUS projects at CanmetENERGY.” https://www.nrcan.gc.ca/energy/offices-labs/canmet/ottawa-research-centre/co2-capture-utilization-and-storage-ccus/23320 (accessed May 14, 2022).

[20] P. Crispeels, D. Inia, H. Legge, T. Nauclér, and P. Radtke, “Decarbonize and create value: how incumbents can tackle the steep challenge,” *McKinsey \& Co.*, 2023.

[21] J. Finkelstein, D. Frankel, and J. Noffsinger, “How to decarbonize global power systems,” *McKinsey Co.*, 2020.

[22] C. AKERMANP and S. CHRISTIANSENE, “Reachingzero with renewables: eliminating CO2 emissions from industry and transport in line with the1? 5℃ climate goal,” 2020.

[23] F. Superchi, A. Mati, C. Carcasci, and A. Bianchini, “Techno-economic analysis of wind-powered green hydrogen production to facilitate the decarbonization of hard-to-abate sectors: A case study on steelmaking,” *Appl. Energy*, vol. 342, p. 121198, 2023.

[24] The World Economic Forum, “Net-Zero Industry Tracker 2023,” 2023. [Online]. Available: https://www.weforum.org/publications/net-zero-industry-tracker-2023/.

[25] P. R. Shukla *et al.*, “Climate change 2022: Mitigation of climate change,” *Contrib. Work. Gr. III to sixth Assess. Rep. Intergov. Panel Clim. Chang.*, vol. 10, p. 9781009157926, 2022.

[26] A. W. Dowling, “Artificial Intelligence and Machine Learning for Sustainable Molecular-to-Systems Engineering,” *Syst. Control Trans*, vol. 3, pp. 22–31, 2024.

[27] C. M. Aras, L. Ntaimo, and M. M. F. Hasan, “Optimal design of Carbon Capture, Utilization and Storage (CCUS) networks under uncertain geological storage volumes and CO2 price,” *Comput. \& Chem. Eng.*, p. 109203, 2025.

[28] M. M. F. Hasan, M. S. Zantye, and M.-K. Kazi, “Challenges and opportunities in carbon capture, utilization and storage: A process systems engineering perspective,” *Comput. \& Chem. Eng.*, vol. 166, p. 107925, 2022.

[29] K. Nadim, M.-S. Ouali, H. Ghezzaz, and A. Ragab, “Learn-to-supervise: Causal reinforcement learning for high-level control in industrial processes,” *Eng. Appl. Artif. Intell.*, vol. 126, p. 106853, 2023.

[30] R. Feldmann, “Artificial intelligence helps cut emissions and costs in cement plants.”

[31] Y. KBC, “Value Chain Optimization Manifesto: Digitalize with purpose.”

[32] J. A. Marmolejo-Saucedo, “Design and Development of Digital Twins: a Case Study in Supply Chains,” *Mob. Networks Appl.*, p. 1, 2020.

[33] W. Wang, H. Yang, Y. Zhang, and J. Xu, “IoT-enabled real-time energy efficiency optimisation method for energy-intensive manufacturing enterprises,” *Int. J. Comput. Integr. Manuf.*, vol. 31, no. 4–5, pp. 362–379, 2018.

[34] J. A. Marmolejo-Saucedo, M. Hurtado-Hernandez, and R. Suarez-Valdes, “Digital twins in

supply chain management: a brief literature review," in *International Conference on Intelligent Computing & Optimization*, 2019, pp. 653–661.

[35] Z. Aung, I. S. Mikhaylov, and Y. T. Aung, "Artificial Intelligence Methods Application in Oil Industry," in *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, 2020, pp. 563–567.

[36] S. Avalos, W. Kracht, and J. M. Ortiz, "Machine learning and deep learning methods in mining operations: A data-driven SAG mill energy consumption prediction application," *Mining, Metall. Explor.*, vol. 37, no. 4, pp. 1197–1212, 2020.

[37] T. Walsh, A. Evatt, and C. S. de Witt, "Artificial Intelligence & Climate Change: Supplementary Impact Report," 2020.

[38] A. Kramer and F. Morgado-Dias, "Artificial intelligence in process control applications and energy saving: a review and outlook," *Greenh. Gases Sci. Technol.*

[39] Y. KBC, "Digitaliztion Manifesto: Now is the time."

[40] J. J. Marchais, "Digitalization, Artificial Intelligence, and Related Technologies for Energy Efficency and GHG Emissions Reduction in Industry," *First webinar inputs, IETS Annex XVIII*, 2018.

[41] D. Weichert, P. Link, A. Stoll, S. Rüping, S. Ihlenfeldt, and S. Wrobel, "A review of machine learning for the optimization of production processes," *Int. J. Adv. Manuf. Technol.*, vol. 104, no. 5–8, pp. 1889–1902, 2019.

[42] S. R. Saufi, Z. A. Bin Ahmad, M. S. Leong, and M. H. Lim, "Challenges and opportunities of deep learning models for machinery fault detection and diagnosis: A review," *IEEE Access*, vol. 7, pp. 122644–122662, 2019.

[43] Y. Lei, N. Li, L. Guo, N. Li, T. Yan, and J. Lin, "Machinery health prognostics: A systematic review from data acquisition to RUL prediction," *Mech. Syst. Signal Process.*, vol. 104, pp. 799–834, 2018.

[44] E. A. Rogers, "The energy savings potential of smart manufacturing," 2014.

[45] Y. Han, H. Wu, Z. Geng, Q. Zhu, X. Gu, and B. Yu, "Energy efficiency evaluation of complex petrochemical industries," *Energy*, p. 117893, 2020.

[46] A. Rasheed, O. San, and T. Kvamsdal, “Digital twin: Values, challenges and enablers from a modeling perspective,” *IEEE Access*, vol. 8, pp. 21980–22012, 2020.

[47] D. Rolnick *et al.*, “Tackling climate change with machine learning,” *arXiv Prepr. arXiv1906.05433*, 2019.

[48] “AI Use Cases For Energy Management.” .

[49] “Digital in industry: From buzzword to value creation | McKinsey.” .

[50] M. G. C. Company, “Stabilization of Process and Improvement of Operation through Process Data Analysis.” Yokogawa.

[51] “MOdel based coNtrol framework for Site-wide OptimizatiON of data-intensive processes | SPIRE.” .

[52] Q. Min, Y. Lu, Z. Liu, C. Su, and B. Wang, “Machine learning based digital twin framework for production optimization in petrochemical industry,” *Int. J. Inf. Manage.*, vol. 49, pp. 502–519, 2019.

[53] H. Lasi, P. Fettke, H.-G. Kemper, T. Feld, and M. Hoffmann, “Industry 4.0,” *Bus. Inf. Syst. Eng.*, vol. 6, no. 4, pp. 239–242, 2014.

[54] Y. Xu, Y. Sun, J. Wan, X. Liu, and Z. Song, “Industrial big data for fault diagnosis: Taxonomy, review, and applications,” *IEEE Access*, vol. 5, pp. 17368–17380, 2017.

[55] “2019 Yokogawa Sustainability Report Published | Yokogawa Electric Corporation.” .

[56] M. Liao and Y. Yao, “Applications of artificial intelligence-based modeling for bioenergy systems: A review,” *GCB Bioenergy*, vol. 13, no. 5, pp. 774–802, 2021.

[57] S. K. Jha, J. Bilalovic, A. Jha, N. Patel, and H. Zhang, “Renewable energy: Present research and future scope of Artificial Intelligence,” *Renew. Sustain. Energy Rev.*, vol. 77, pp. 297–317, 2017.

[58] B. Pomeroy, M. Grilc, and B. Likozar, “Artificial neural networks for bio-based chemical production or biorefining: A review,” *Renew. Sustain. Energy Rev.*, vol. 153, p. 111748, 2022.

[59] L. Zhou, Y. Song, W. Ji, and H. Wei, “Machine learning for combustion,” *Energy AI*, vol.

7, p. 100128, 2022.

[60] M. Mowbray *et al.*, “Machine learning for biochemical engineering: A review,” *Biochem. Eng. J.*, vol. 172, p. 108054, 2021.

[61] B. Sundui, O. A. Ramirez Calderon, O. M. Abdeldayem, J. Lázaro-Gil, E. R. Rene, and U. Sambuu, “Applications of machine learning algorithms for biological wastewater treatment: updates and perspectives,” *Clean Technol. Environ. Policy*, vol. 23, no. 1, pp. 127–143, 2021.

[62] K. Mazloomi and C. Gomes, “Hydrogen as an energy carrier: Prospects and challenges,” *Renew. Sustain. Energy Rev.*, vol. 16, no. 5, pp. 3024–3033, 2012.

[63] M. Yu, K. Wang, and H. Vredenburg, “Insights into low-carbon hydrogen production methods: Green, blue and aqua hydrogen,” *Int. J. Hydrogen Energy*, vol. 46, no. 41, pp. 21261–21273, 2021.

[64] A. Boretti, “There are hydrogen production pathways with better than green hydrogen economic and environmental costs,” *Int. J. Hydrogen Energy*, vol. 46, no. 46, pp. 23988–23995, 2021.

[65] S. Atilhan, S. Park, M. M. El-Halwagi, M. Atilhan, M. Moore, and R. B. Nielsen, “Green hydrogen as an alternative fuel for the shipping industry,” *Curr. Opin. Chem. Eng.*, vol. 31, p. 100668, 2021.

[66] M. R. Shaner, H. A. Atwater, N. S. Lewis, and E. W. McFarland, “A comparative technoeconomic analysis of renewable hydrogen production using solar energy,” *Energy \& Environ. Sci.*, vol. 9, no. 7, pp. 2354–2371, 2016.

[67] A. Mostafaeipour *et al.*, “Evaluating the wind energy potential for hydrogen production: a case study,” *Int. J. Hydrogen Energy*, vol. 41, no. 15, pp. 6200–6210, 2016.

[68] K. Christopher and R. Dimitrios, “A review on exergy comparison of hydrogen production methods from renewable energy sources,” *Energy \& Environ. Sci.*, vol. 5, no. 5, pp. 6640–6651, 2012.

[69] E. E. Ozbas, D. Aksu, A. Ongen, M. A. Aydin, and H. K. Ozcan, “Hydrogen production via biomass gasification, and modeling by supervised machine learning algorithms,” *Int. J.*

Hydrogen Energy, vol. 44, no. 32, pp. 17260–17268, 2019.

[70] A. Ahmed and D. J. Siegel, "Predicting hydrogen storage in MOFs via machine learning," *Patterns*, vol. 2, no. 7, p. 100291, 2021.

[71] A. J. Abbas, H. Hassani, M. Burby, and I. J. John, "An investigation into the volumetric flow rate requirement of hydrogen transportation in existing natural gas pipelines and its safety implications," *Gases*, vol. 1, no. 4, pp. 156–179, 2021.

[72] W. Wang, Y. Ma, A. Maroufmashat, N. Zhang, J. Li, and X. Xiao, "Optimal design of large-scale solar-aided hydrogen production process via machine learning based optimisation framework," *Appl. Energy*, vol. 305, p. 117751, 2022.

[73] M. Byun, H. Lee, C. Choe, S. Cheon, and H. Lim, "Machine learning based predictive model for methanol steam reforming with technical, environmental, and economic perspectives," *Chem. Eng. J.*, vol. 426, p. 131639, 2021.

[74] J. Li, L. Pan, M. Suvarna, and X. Wang, "Machine learning aided supercritical water gasification for H2-rich syngas production with process optimization and catalyst screening," *Chem. Eng. J.*, vol. 426, p. 131285, 2021.

[75] S. Ascher, I. Watson, and S. You, "Machine learning methods for modelling the gasification and pyrolysis of biomass and waste," *Renew. Sustain. Energy Rev.*, p. 111902, 2021.

[76] B. V. Ayodele, S. I. Mustapa, R. Kanthasamy, M. Zwawi, and C. K. Cheng, "Modeling the prediction of hydrogen production by co-gasification of plastic and rubber wastes using machine learning algorithms," *Int. J. Energy Res.*, vol. 45, no. 6, pp. 9580–9594, 2021.

[77] S. Shenbagaraj, P. K. Sharma, A. K. Sharma, G. Raghav, K. B. Kota, and V. Ashokkumar, "Gasification of food waste in supercritical water: An innovative synthesis gas composition prediction model based on Artificial Neural Networks," *Int. J. Hydrogen Energy*, vol. 46, no. 24, pp. 12739–12757, 2021.

[78] Z. Lian *et al.*, "Hydrogen Production by Fluidized Bed Reactors: A Quantitative Perspective Using the Supervised Machine Learning Approach," *J*, vol. 4, no. 3, pp. 266–287, 2021.

[79] Q. Tao, T. Lu, Y. Sheng, L. Li, W. Lu, and M. Li, "Machine learning aided design of perovskite oxide materials for photocatalytic water splitting," *J. Energy Chem.*, vol. 60, pp.

351–359, 2021.

[80] A. Govind Rajan, J. M. P. Martirez, and E. A. Carter, “Why do we use the materials and operating conditions we use for heterogeneous (photo) electrochemical water splitting?,” *ACS Catal.*, vol. 10, no. 19, pp. 11177–11234, 2020.

[81] H. Jin *et al.*, “Data-driven systematic search of promising photocatalysts for water splitting under visible light,” *J. Phys. Chem. Lett.*, vol. 10, no. 17, pp. 5211–5218, 2019.

[82] X. Sun *et al.*, “Machine-learning-accelerated screening of hydrogen evolution catalysts in MBenes materials,” *Appl. Surf. Sci.*, vol. 526, p. 146522, 2020.

[83] M. Amirinejad, N. Tavajohi-Hasankiadeh, S. S. Madaeni, M. A. Navarra, E. Rafiee, and B. Scrosati, “Adaptive neuro-fuzzy inference system and artificial neural network modeling of proton exchange membrane fuel cells based on nanocomposite and recast Nafion membranes,” *Int. J. Energy Res.*, vol. 37, no. 4, pp. 347–357, 2013.

[84] L. Zheng, Y. Hou, T. Zhang, and X. Pan, “Performance prediction of fuel cells using long short-term memory recurrent neural network,” *Int. J. Energy Res.*, vol. 45, no. 6, pp. 9141–9161, 2021.

[85] J. Milewski and others, “Artificial Neural Network Based Model for Calculating the Flow Composition Influence of Solid Oxide Fuel Cell,” 2014.

[86] J. Milewski, A. Szczęśniak, Ł. Szabłowski, O. Dybiński, and A. Miller, “Artificial neural network model of molten carbonate fuel cells: Validation on experimental data,” *Int. J. Energy Res.*, vol. 43, no. 13, pp. 6740–6761, 2019.

[87] S. Bozorgmehri and M. Hamedi, “Modeling and optimization of anode-supported solid oxide fuel cells on cell parameters via artificial neural network and genetic algorithm,” *Fuel Cells*, vol. 12, no. 1, pp. 11–23, 2012.

[88] J. Arriagada, P. Olausson, and A. Selimovic, “Artificial neural network simulator for SOFC performance prediction,” *J. Power Sources*, vol. 112, no. 1, pp. 54–60, 2002.

[89] F. Jurado, “Power supply quality improvement with a SOFC plant by neural-network-based control,” *J. Power Sources*, vol. 117, no. 1–2, pp. 75–83, 2003.

[90] M. Rahimi, M. H. Abbaspour-Fard, and A. Rohani, “Machine learning approaches to

rediscovery and optimization of hydrogen storage on porous bio-derived carbon," *J. Clean. Prod.*, vol. 329, p. 129714, 2021.

[91] R. M. Giappa, E. Tylianakis, M. Di Gennaro, K. Gkagkas, and G. E. Froudakis, "A combination of multi-scale calculations with machine learning for investigating hydrogen storage in metal organic frameworks," *Int. J. Hydrogen Energy*, vol. 46, no. 54, pp. 27612–27621, 2021.

[92] G. Anderson, B. Schweitzer, R. Anderson, and D. A. Gómez-Gualdrón, "Attainable volumetric targets for adsorption-based hydrogen storage in porous crystals: molecular simulation and machine learning," *J. Phys. Chem. C*, vol. 123, no. 1, pp. 120–130, 2018.

[93] D. Grondin, J. Deseure, P. Ozil, J.-P. Chabriat, B. Grondin-Perez, and A. Brisse, "Solid oxide electrolysis cell 3d simulation using artificial neural network for cathodic process description," *Chem. Eng. Res. Des.*, vol. 91, no. 1, pp. 134–140, 2013.

[94] E. Can and R. Yildirim, "Data mining in photocatalytic water splitting over perovskites literature for higher hydrogen production," *Appl. Catal. B Environ.*, vol. 242, pp. 267–283, 2019.

[95] J. George and G. Hautier, "Chemist versus machine: Traditional knowledge versus machine learning techniques," *Trends Chem.*, vol. 3, no. 2, pp. 86–95, 2021.

[96] Z. Ding, Z. Chen, T. Ma, C.-T. Lu, W. Ma, and L. Shaw, "Predicting the hydrogen release ability of LiBH4-based mixtures by ensemble machine learning," *Energy Storage Mater.*, vol. 27, pp. 466–477, 2020.

[97] N. S. Bobbitt and R. Q. Snurr, "Molecular modelling and machine learning for high-throughput screening of metal-organic frameworks for hydrogen storage," *Mol. Simul.*, vol. 45, no. 14–15, pp. 1069–1081, 2019.

[98] M. O. J. Jäger, E. V Morooka, F. Federici Canova, L. Himanen, and A. S. Foster, "Machine learning hydrogen adsorption on nanoclusters through structural descriptors," *npj Comput. Mater.*, vol. 4, no. 1, pp. 1–8, 2018.

[99] A. Hosseinzadeh, J. L. Zhou, A. Altaee, and D. Li, "Machine learning modeling and analysis of biohydrogen production from wastewater by dark fermentation process," *Bioresour. Technol.*, vol. 343, p. 126111, 2022.

[100] Y. Wang *et al.*, "Modeling biohydrogen production using different data driven approaches," *Int. J. Hydrogen Energy*, vol. 46, no. 58, pp. 29822–29833, 2021.

[101] I. Monroy and G. Buitrón, "Diagnosis of undesired scenarios in hydrogen production by photo-fermentation," *Water Sci. Technol.*, vol. 78, no. 8, pp. 1652–1657, 2018.

[102] I. Monroy, E. Guevara-López, and G. Buitrón, "Biohydrogen production by batch indoor and outdoor photo-fermentation with an immobilized consortium: a process model with Neural Networks," *Biochem. Eng. J.*, vol. 135, pp. 1–10, 2018.

[103] A. Hosseinzadeh, J. L. Zhou, A. Altaee, M. Baziar, and D. Li, "Effective modelling of hydrogen and energy recovery in microbial electrolysis cell by artificial neural network and adaptive network-based fuzzy inference system," *Bioresour. Technol.*, vol. 316, p. 123967, 2020.

[104] L. Wang, F. Long, D. Liang, X. Xiao, and H. Liu, "Hydrogen production from lignocellulosic hydrolysate in an up-scaled microbial electrolysis cell with stacked bio-electrodes," *Bioresour. Technol.*, vol. 320, p. 124314, 2021.

[105] S. Faizollahzadeh Ardabili, B. Najafi, S. Shamshirband, B. Minaei Bidgoli, R. C. Deo, and K. Chau, "Computational intelligence approach for modeling hydrogen production: A review," *Eng. Appl. Comput. Fluid Mech.*, vol. 12, no. 1, pp. 438–458, 2018.

[106] A. Zamaniyan, F. Joda, A. Behroozsarand, and H. Ebrahimi, "Application of artificial neural networks (ANN) for modeling of industrial hydrogen plant," *Int. J. Hydrogen Energy*, vol. 38, no. 15, pp. 6289–6297, 2013.

[107] L. L. Teck Chan and J. Chen, "Economic model predictive control of an absorber-stripper CO2 capture process for improving energy cost," *IFAC-PapersOnLine*, vol. 51, no. 18, pp. 109–114, 2018, doi: https://doi.org/10.1016/j.ifacol.2018.09.284.

[108] F. Li, J. Zhang, E. Oko, and M. Wang, "Modelling of a post-combustion CO2 capture process using neural networks," *Fuel*, vol. 151, pp. 156–163, 2015.

[109] N. Sipöcz, F. A. Tobiesen, and M. Assadi, "The use of artificial neural network models for CO2 capture plants," *Appl. Energy*, vol. 88, no. 7, pp. 2368–2376, 2011.

[110] J. Zhan *et al.*, "Simultaneous absorption of H2S and CO2 into the MDEA+ PZ aqueous

solution in a rotating packed bed," *Ind. \& Eng. Chem. Res.*, vol. 59, no. 17, pp. 8295–8303, 2020.

[111] X. Wu, J. Shen, M. Wang, and K. Y. Lee, "Intelligent predictive control of large-scale solvent-based CO2 capture plant using artificial neural network and particle swarm optimization," *Energy*, vol. 196, p. 117070, 2020.

[112] A. Nuchitprasittichai and S. Cremaschi, "Optimization of CO2 Capture Process with Aqueous Amines A Comparison of Two Simulation--Optimization Approaches," *Ind. \& Eng. Chem. Res.*, vol. 52, no. 30, pp. 10236–10243, 2013.

[113] S. Perathoner, K. M. Van Geem, G. B. Marin, and G. Centi, "Reuse of CO 2 in energy intensive process industries," *Chem. Commun.*, vol. 57, no. 84, pp. 10967–10982, 2021.

[114] D. M. Abouelella, S.-E. K. Fateen, and M. M. K. Fouad, "Multiscale modeling study of the adsorption of CO2 using different capture materials," *Evergr. Jt. J. Nov. Carbon Resour. Sci. \& Green Asia Strateg.*, vol. 5, no. 01, pp. 43–51, 2018.

[115] S. G. Subraveti, Z. Li, V. Prasad, and A. Rajendran, "Machine learning-based multiobjective optimization of pressure swing adsorption," *Ind. \& Eng. Chem. Res.*, vol. 58, no. 44, pp. 20412–20422, 2019.

[116] T. D. Burns *et al.*, "Prediction of MOF performance in vacuum swing adsorption systems for postcombustion CO2 capture based on integrated molecular simulations, process optimizations, and machine learning models," *Environ. Sci. \& Technol.*, vol. 54, no. 7, pp. 4536–4544, 2020.

[117] M. Khurana and S. Farooq, "Integrated Adsorbent Process Optimization for Minimum Cost of Electricity Including Carbon Capture bya VSA Process," *AIChE J.*, vol. 65, no. 1, pp. 184–195, 2019.

[118] J. Xiao *et al.*, "Machine learning--based optimization for hydrogen purification performance of layered bed pressure swing adsorption," *Int. J. Energy Res.*, vol. 44, no. 6, pp. 4475–4492, 2020.

[119] K. N. Pai, V. Prasad, and A. Rajendran, "Experimentally validated machine learning frameworks for accelerated prediction of cyclic steady state and optimization of pressure swing adsorption processes," *Sep. Purif. Technol.*, vol. 241, p. 116651, 2020.

[120] S. M. Moosavi, K. M. Jablonka, and B. Smit, “The role of machine learning in the understanding and design of materials,” *J. Am. Chem. Soc.*, vol. 142, no. 48, pp. 20273–20287, 2020.

[121] M. Rahimi, S. M. Moosavi, B. Smit, and T. A. Hatton, “Toward smart carbon capture with machine learning,” *Cell Reports Phys. Sci.*, vol. 2, no. 4, p. 100396, 2021.

[122] W. Ampomah *et al.*, “Optimum design of CO2 storage and oil recovery under geological uncertainty,” *Appl. Energy*, vol. 195, pp. 80–92, 2017.

[123] B. Chen, D. R. Harp, Y. Lin, E. H. Keating, and R. J. Pawar, “Geologic CO2 sequestration monitoring design: A machine learning and uncertainty quantification based approach,” *Appl. Energy*, vol. 225, pp. 332–345, 2018.

[124] S. Sinha *et al.*, “Normal or abnormal? Machine learning for the leakage detection in carbon sequestration projects using pressure field data,” *Int. J. Greenh. Gas Control*, vol. 103, p. 103189, 2020.

[125] D. Moriarty, L. Dobeck, and S. Benson, “Rapid surface detection of CO2 leaks from geologic sequestration sites,” *Energy Procedia*, vol. 63, pp. 3975–3983, 2014.

[126] M. N. Amar, “Towards improved genetic programming based-correlations for predicting the interfacial tension of the systems pure/impure CO2-brine,” *J. Taiwan Inst. Chem. Eng.*, vol. 127, pp. 186–196, 2021.

[127] H. Ni and S. M. Benson, “Using unsupervised machine learning to characterize capillary flow and residual trapping,” *Water Resour. Res.*, vol. 56, no. 8, p. e2020WR027473, 2020.

[128] A. Daryasafar, A. Keykhosravi, and K. Shahbazi, “Modeling CO2 wettability behavior at the interface of brine/CO2/mineral: Application to CO2 geo-sequestration,” *J. Clean. Prod.*, vol. 239, p. 118101, 2019.

[129] M. N. Amar and A. J. Ghahfarokhi, “Prediction of CO2 diffusivity in brine using white-box machine learning,” *J. Pet. Sci. Eng.*, vol. 190, p. 107037, 2020.

[130] Y. Yan, L. Wang, T. Wang, X. Wang, Y. Hu, and Q. Duan, “Application of soft computing techniques to multiphase flow measurement: A review,” *Flow Meas. Instrum.*, vol. 60, pp. 30–43, 2018.

[131] L. Wang, Y. Yan, X. Wang, T. Wang, Q. Duan, and W. Zhang, “Mass flow measurement of gas-liquid two-phase CO2 in CCS transportation pipelines using Coriolis flowmeters,” *Int. J. Greenh. Gas Control*, vol. 68, pp. 269–275, 2018.

[132] L. Wang, Y. Yan, X. Wang, and T. Wang, “Input variable selection for data-driven models of Coriolis flowmeters for two-phase flow measurement,” *Meas. Sci. Technol.*, vol. 28, no. 3, p. 35305, 2017.

[133] D. Shao, Y. Yan, W. Zhang, S. Sun, C. Sun, and L. Xu, “Dynamic measurement of gas volume fraction in a CO2 pipeline through capacitive sensing and data driven modelling,” *Int. J. Greenh. Gas Control*, vol. 94, p. 102950, 2020.

[134] M. Henry *et al.*, “Two-phase flow metering of heavy oil using a Coriolis mass flow meter: A case study,” *Flow Meas. Instrum.*, vol. 17, no. 6, pp. 399–413, 2006.

[135] M. J. Mølnvik, “NCCS-Annual Report 2019,” *SINTEF Rapp.*, 2019.

[136] V. E. Onyebuchi, A. Kolios, D. P. Hanak, C. Biliyok, and V. Manovic, “A systematic review of key challenges of CO2 transport via pipelines,” *Renew. Sustain. Energy Rev.*, vol. 81, pp. 2563–2583, 2018.

[137] F. Krasnov, N. Glavnov, and A. Sitnikov, “A machine learning approach to enhanced oil recovery prediction,” in *International Conference on Analysis of Images, Social Networks and Texts*, 2017, pp. 164–171.

[138] J. You *et al.*, “Assessment of enhanced oil recovery and CO2 storage capacity using machine learning and optimization framework,” 2019.

[139] B. Chen and R. J. Pawar, “Characterization of CO2 storage and enhanced oil recovery in residual oil zones,” *Energy*, vol. 183, pp. 291–304, 2019.

[140] J. You *et al.*, “Machine learning based co-optimization of carbon dioxide sequestration and oil recovery in CO2-EOR project,” *J. Clean. Prod.*, vol. 260, p. 120866, 2020.

[141] J. You, W. Ampomah, and Q. Sun, “Development and application of a machine learning based multi-objective optimization workflow for CO2-EOR projects,” *Fuel*, vol. 264, p. 116758, 2020.

[142] J. You, W. Ampomah, A. Morgan, Q. Sun, and X. Huang, “A comprehensive techno-eco-

assessment of CO2 enhanced oil recovery projects using a machine-learning assisted workflow," *Int. J. Greenh. Gas Control*, vol. 111, p. 103480, 2021.

[143] R. Das *et al.*, "Noble-Metal-Free Heterojunction Photocatalyst for Selective CO2 Reduction to Methane upon Induced Strain Relaxation," *ACS Catal.*, vol. 12, no. 1, pp. 687–697, 2021.

[144] Y. Chen, Y. Huang, T. Cheng, and W. A. Goddard III, "Identifying active sites for CO2 reduction on dealloyed gold surfaces by combining machine learning with multiscale simulations," *J. Am. Chem. Soc.*, vol. 141, no. 29, pp. 11651–11657, 2019.

[145] N. Zhang *et al.*, "Machine Learning in Screening High Performance Electrocatalysts for CO2 Reduction," *Small Methods*, vol. 5, no. 11, p. 2100987, 2021.

[146] A. Chen, X. Zhang, L. Chen, S. Yao, and Z. Zhou, "A machine learning model on simple features for CO2 reduction electrocatalysts," *J. Phys. Chem. C*, vol. 124, no. 41, pp. 22471–22478, 2020.

[147] X. Ma, Z. Li, L. E. K. Achenie, and H. Xin, "Machine-learning-augmented chemisorption model for CO2 electroreduction catalyst screening," *J. Phys. Chem. Lett.*, vol. 6, no. 18, pp. 3528–3533, 2015.

[148] H. Li, D. Yan, Z. Zhang, and E. Lichtfouse, "Prediction of CO2 absorption by physical solvents using a chemoinformatics-based machine learning model," *Environ. Chem. Lett.*, vol. 17, no. 3, pp. 1397–1404, 2019.

[149] A. Naghizadeh, A. Larestani, M. N. Amar, and A. Hemmati-Sarapardeh, "Predicting viscosity of CO2--N2 gaseous mixtures using advanced intelligent schemes," *J. Pet. Sci. Eng.*, vol. 208, p. 109359, 2022.

[150] S. Babamohammadi, A. Shamiri, T. N. G. Borhani, M. S. Shafeeyan, M. K. Aroua, and R. Yusoff, "Solubility of CO2 in aqueous solutions of glycerol and monoethanolamine," *J. Mol. Liq.*, vol. 249, pp. 40–52, 2018.

[151] G. Truc, N. Rahmanian, and M. Pishnamazi, "Assessment of Cubic Equations of State: Machine Learning for Rich Carbon-Dioxide Systems," *Sustainability*, vol. 13, no. 5, p. 2527, 2021.

[152] B. Chehaibou, M. Badawi, T. Bucko, T. Bazhirov, and D. Rocca, "Computing RPA

adsorption enthalpies by machine learning thermodynamic perturbation theory," *J. Chem. Theory Comput.*, vol. 15, no. 11, pp. 6333–6342, 2019.

[153] P. V Acharya and V. Bahadur, "Thermodynamic features-driven machine learning-based predictions of clathrate hydrate equilibria in the presence of electrolytes," *Fluid Phase Equilib.*, vol. 530, p. 112894, 2021.

[154] Z. Zhang *et al.*, "Prediction of carbon dioxide adsorption via deep learning," *Angew. Chemie*, vol. 131, no. 1, pp. 265–269, 2019.

[155] Z. Zhang, H. Li, H. Chang, Z. Pan, and X. Luo, "Machine learning predictive framework for CO2 thermodynamic properties in solution," *J. CO2 Util.*, vol. 26, pp. 152–159, 2018.

[156] E. G. Al-Sakkari, A. Ragab, H. Dagdougui, D. C. Boffito, and M. Amazouz, "Carbon capture, utilization and sequestration systems design and operation optimization: Assessment and perspectives of artificial intelligence opportunities," *Sci. Total Environ.*, p. 170085, 2024.

[157] H. S. Doma and S. I. Abou-Elela, "Treatment of black liquor derived from non-woody feedstocks," *WIT Trans. Ecol. Environ.*, vol. 63, 2003.

[158] A. Demirbas, "Waste management, waste resource facilities and waste conversion processes," *Energy Convers. Manag.*, vol. 52, no. 2, pp. 1280–1287, 2011.

[159] D. Núñez, P. Oulego, S. Collado, F. A. Riera, and M. D\'\iaz, "Recovery of organic acids from pre-treated Kraft black liquor using ultrafiltration and liquid-liquid extraction," *Sep. Purif. Technol.*, vol. 284, p. 120274, 2022.

[160] M. Naqvi, J. Yan, and E. Dahlquist, "Black liquor gasification integrated in pulp and paper mills: A critical review," *Bioresour. Technol.*, vol. 101, no. 21, pp. 8001–8015, 2010.

[161] M. El Koujok, H. Ghezzaz, and M. Amazouz, "Energy inefficiency diagnosis in industrial process through one-class machine learning techniques," *J. Intell. Manuf.*, vol. 32, no. 7, pp. 2043–2060, 2021.

[162] A. Ragab, M. El-Koujok, B. Poulin, M. Amazouz, and S. Yacout, "Fault diagnosis in industrial chemical processes using interpretable patterns based on Logical Analysis of Data," *Expert Syst. Appl.*, vol. 95, pp. 368–383, 2018.

[163] M. Elhefnawy, A. Ragab, and M.-S. Ouali, “Fault classification in the process industry using polygon generation and deep learning,” *J. Intell. Manuf.*, pp. 1–14, 2021.

[164] A. Ragab, M. El Koujok, H. Ghezzaz, M. Amazouz, M.-S. Ouali, and S. Yacout, “Deep understanding in industrial processes by complementing human expertise with interpretable patterns of machine learning,” *Expert Syst. Appl.*, vol. 122, pp. 388–405, 2019.

[165] Q. Jiang, X. Yan, and W. Zhao, “Fault detection and diagnosis in chemical processes using sensitive principal component analysis,” *Ind. \& Eng. Chem. Res.*, vol. 52, no. 4, pp. 1635–1644, 2013.

[166] H. Li and D. Xiao, “Fault diagnosis of Tennessee Eastman process using signal geometry matching technique,” *EURASIP J. Adv. Signal Process.*, vol. 2011, no. 1, pp. 1–19, 2011.

[167] I. Lomov, M. Lyubimov, I. Makarov, and L. E. Zhukov, “Fault detection in Tennessee eastman process with temporal deep learning models,” *J. Ind. Inf. Integr.*, vol. 23, p. 100216, 2021.

[168] S. Heo and J. H. Lee, “Statistical process monitoring of the Tennessee Eastman process using parallel autoassociative neural networks and a large dataset,” *Processes*, vol. 7, no. 7, p. 411, 2019.

[169] G. Wang, J. Liu, and Y. Li, “Fault diagnosis using kNN reconstruction on MRI variables,” *J. Chemom.*, vol. 29, no. 7, pp. 399–410, 2015.

[170] R. Pelalak, A. T. Nakhjiri, A. Marjani, M. Rezakazemi, and S. Shirazian, “Influence of machine learning membership functions and degree of membership function on each input parameter for simulation of reactors,” *Sci. Rep.*, vol. 11, no. 1, pp. 1–11, 2021.

[171] M. Babanezhad, I. Behroyan, A. T. Nakhjiri, A. Marjani, M. Rezakazemi, and S. Shirazian, “High-performance hybrid modeling chemical reactors using differential evolution based fuzzy inference system,” *Sci. Rep.*, vol. 10, no. 1, pp. 1–11, 2020.

[172] M. Pourtousi, M. Zeinali, P. Ganesan, and J. N. Sahu, “Prediction of multiphase flow pattern inside a 3D bubble column reactor using a combination of CFD and ANFIS,” *Rsc Adv.*, vol. 5, no. 104, pp. 85652–85672, 2015.

[173] L. Jofre, Z. R. del Rosario, and G. Iaccarino, “Data-driven dimensional analysis of heat

transfer in irradiated particle-laden turbulent flow," *Int. J. Multiph. Flow*, vol. 125, p. 103198, 2020.

[174] Y. Liu, N. Dinh, Y. Sato, and B. Niceno, "Data-driven modeling for boiling heat transfer: using deep neural networks and high-fidelity simulation results," *Appl. Therm. Eng.*, vol. 144, pp. 305–320, 2018.

[175] N. Nabipour, M. Babanezhad, A. Taghvaie Nakhjiri, and S. Shirazian, "Prediction of nanofluid temperature inside the cavity by integration of grid partition clustering categorization of a learning structure with the fuzzy system," *ACS omega*, vol. 5, no. 7, pp. 3571–3578, 2020.

[176] Y. ElShazly, "The use of neural network to estimate mass transfer coefficient from the bottom of agitated vessel," *Heat Mass Transf.*, vol. 51, no. 4, pp. 465–475, 2015.

[177] E. Alvarez, J. M. Correa, C. Riverol, and J. M. Navaza, "Model based in neural networks for the prediction of the mass transfer coefficients in bubble columns. Study in Newtonian and non-Newtonianian fluids," *Int. Commun. heat mass Transf.*, vol. 27, no. 1, pp. 93–98, 2000.

[178] J. Cho, H. Kim, A. L. Gebreselassie, and D. Shin, "Deep neural network and random forest classifier for source tracking of chemical leaks using fence monitoring data," *J. Loss Prev. Process Ind.*, vol. 56, pp. 548–558, 2018.

[179] M. Babanezhad, S. Zabihi, I. Behroyan, A. T. Nakhjiri, A. Marjani, and S. Shirazian, "Prediction of gas velocity in two-phase flow using developed fuzzy logic system with differential evolution algorithm," *Sci. Rep.*, vol. 11, no. 1, pp. 1–14, 2021.

[180] S. Shu, D. Vidal, F. Bertrand, and J. Chaouki, "Multiscale multiphase phenomena in bubble column reactors: A review," *Renew. Energy*, vol. 141, pp. 613–631, 2019.

[181] M. Babanezhad, I. Behroyan, A. T. Nakhjiri, A. Marjani, and S. Shirazian, "Simulation of liquid flow with a combination artificial intelligence flow field and Adams--Bashforth method," *Sci. Rep.*, vol. 10, no. 1, pp. 1–10, 2020.

[182] M. Babanezhad, M. Pishnamazi, A. Marjani, and S. Shirazian, "Bubbly flow prediction with randomized neural cells artificial learning and fuzzy systems based on k--ε turbulence and Eulerian model data set," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, 2020.

[183] M. Pishnamazi, M. Babanezhad, A. T. Nakhjiri, M. Rezakazemi, A. Marjani, and S. Shirazian, “ANFIS grid partition framework with difference between two sigmoidal membership functions structure for validation of nanofluid flow,” *Sci. Rep.*, vol. 10, no. 1, pp. 1–11, 2020.

[184] M. Babanezhad, A. Taghvaie Nakhjiri, M. Rezakazemi, and S. Shirazian, “Developing intelligent algorithm as a machine learning overview over the big data generated by Euler--Euler method to simulate bubble column reactor hydrodynamics,” *ACS omega*, vol. 5, no. 32, pp. 20558–20566, 2020.

[185] M. Babanezhad, I. Behroyan, A. Marjani, and S. Shirazian, “Artificial intelligence simulation of suspended sediment load with different membership functions of ANFIS,” *Neural Comput. Appl.*, vol. 33, no. 12, pp. 6819–6833, 2021.

[186] M. Rezakazemi and S. Shirazian, “Gas-liquid phase recirculation in bubble column reactors: development of a hybrid model based on local CFD--adaptive neuro-fuzzy inference system (ANFIS),” *J. Non-Equilibrium Thermodyn.*, vol. 44, no. 1, pp. 29–42, 2019.

[187] S. L. Brunton, B. R. Noack, and P. Koumoutsakos, “Machine learning for fluid mechanics,” *Annu. Rev. Fluid Mech.*, vol. 52, pp. 477–508, 2020.

[188] R. Wang, K. Kashinath, M. Mustafa, A. Albert, and R. Yu, “Towards physics-informed deep learning for turbulent flow prediction,” in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery \& Data Mining*, 2020, pp. 1457–1466.

[189] A. Mannarino and P. Mantegazza, “Nonlinear aeroelastic reduced order modeling by recurrent neural networks,” *J. Fluids Struct.*, vol. 48, pp. 103–121, 2014.

[190] K. Li, J. Kou, and W. Zhang, “Deep neural network for unsteady aerodynamic and aeroelastic modeling across multiple Mach numbers,” *Nonlinear Dyn.*, vol. 96, no. 3, pp. 2157–2177, 2019.

[191] K. Duraisamy, G. Iaccarino, and H. Xiao, “Turbulence modeling in the age of data,” *Annu. Rev. Fluid Mech.*, vol. 51, pp. 357–377, 2019.

[192] Y. Cao, M. Babanezhad, M. Rezakazemi, and S. Shirazian, “Prediction of fluid pattern in a shear flow on intelligent neural nodes using ANFIS and LBM,” *Neural Comput. Appl.*, vol. 32, no. 17, pp. 13313–13321, 2020.

[193] M. Pourtousi, J. N. Sahu, P. Ganesan, S. Shamshirband, and G. Redzwan, “A combination of computational fluid dynamics (CFD) and adaptive neuro-fuzzy system (ANFIS) for prediction of the bubble column hydrodynamics,” *Powder Technol.*, vol. 274, pp. 466–481, 2015.

[194] B. Wang and J. Wang, “Application of artificial intelligence in computational fluid dynamics,” *Ind. \& Eng. Chem. Res.*, vol. 60, no. 7, pp. 2772–2790, 2021.

[195] B. Huang and O. A. von Lilienfeld, “Ab initio machine learning in chemical compound space,” *Chem. Rev.*, vol. 121, no. 16, pp. 10001–10036, 2021.

[196] J. Roberts, J. R. S. Bursten, and C. Risko, “Genetic Algorithms and Machine Learning for Predicting Surface Composition, Structure, and Chemistry: A Historical Perspective and Assessment,” *Chem. Mater.*, vol. 33, no. 17, pp. 6589–6615, 2021.

[197] R. Gómez-Bombarelli *et al.*, “Automatic chemical design using a data-driven continuous representation of molecules,” *ACS Cent. Sci.*, vol. 4, no. 2, pp. 268–276, 2018.

[198] E. Heid and W. H. Green, “Machine learning of reaction properties via learned representations of the condensed graph of reaction,” *J. Chem. Inf. Model.*, 2021.

[199] J. N. Wei, D. Duvenaud, and A. Aspuru-Guzik, “Neural networks for the prediction of organic chemistry reactions,” *ACS Cent. Sci.*, vol. 2, no. 10, pp. 725–732, 2016.

[200] P. P. Plehiers, I. Lengyel, D. H. West, G. B. Marin, C. V Stevens, and K. M. Van Geem, “Fast estimation of standard enthalpy of formation with chemical accuracy by artificial neural network correction of low-level-of-theory ab initio calculations,” *Chem. Eng. J.*, vol. 426, p. 131304, 2021.

[201] K. Ahuja, W. H. Green, and Y.-P. Li, “Learning to optimize molecular geometries using reinforcement learning,” *J. Chem. Theory Comput.*, vol. 17, no. 2, pp. 818–825, 2021.

[202] R. Zubatyuk, J. S. Smith, B. T. Nebgen, S. Tretiak, and O. Isayev, “Teaching a neural network to attach and detach electrons from molecules,” *Nat. Commun.*, vol. 12, no. 1, pp. 1–11, 2021.

[203] M. G. Taylor, A. Nandy, C. C. Lu, and H. J. Kulik, “Deciphering Cryptic Behavior in Bimetallic Transition-Metal Complexes with Machine Learning,” *J. Phys. Chem. Lett.*, vol.

12, no. 40, pp. 9812–9820, 2021.

[204] J. A. Keith *et al.*, “Combining machine learning and computational chemistry for predictive insights into chemical systems,” *Chem. Rev.*, vol. 121, no. 16, pp. 9816–9872, 2021.

[205] M. Zhong *et al.*, “Accelerated discovery of CO2 electrocatalysts using active machine learning,” *Nature*, vol. 581, no. 7807, pp. 178–183, 2020.

[206] C. W. Park and C. Wolverton, “Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery,” *Phys. Rev. Mater.*, vol. 4, no. 6, p. 63801, 2020.

[207] J. Nam and J. Kim, “Linking the neural machine translation and the prediction of organic chemistry reactions,” *arXiv Prepr. arXiv1612.09529*, 2016.

[208] P. Schwaller *et al.*, “Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction,” *ACS Cent. Sci.*, vol. 5, no. 9, pp. 1572–1583, 2019.

[209] K. McCullough, T. Williams, K. Mingle, P. Jamshidi, and J. Lauterbach, “High-throughput experimentation meets artificial intelligence: a new pathway to catalyst discovery,” *Phys. Chem. Chem. Phys.*, vol. 22, no. 20, pp. 11174–11196, 2020.

[210] S.-H. Wang, H. S. Pillai, S. Wang, L. E. K. Achenie, and H. Xin, “Infusing theory into deep learning for interpretable reactivity prediction,” *Nat. Commun.*, vol. 12, no. 1, pp. 1–9, 2021.

[211] W. M. Abdelmoula *et al.*, “Peak learning of mass spectrometry imaging data using artificial neural networks,” *Nat. Commun.*, vol. 12, no. 1, pp. 1–13, 2021.

[212] A. H. Lipkus *et al.*, “Structural diversity of organic chemistry. A scaffold analysis of the CAS Registry,” *J. Org. Chem.*, vol. 73, no. 12, pp. 4443–4451, 2008.

[213] J. J. Naveja and J. L. Medina-Franco, “Finding constellations in chemical space through core analysis,” *Front. Chem.*, p. 510, 2019.

[214] J.-L. Reymond and M. Awale, “Exploring chemical space for drug discovery using the chemical universe database,” *ACS Chem. Neurosci.*, vol. 3, no. 9, pp. 649–657, 2012.

[215] P. S. Gromski, A. B. Henson, J. M. Granda, and L. Cronin, “How to explore chemical space using algorithms and automation,” *Nat. Rev. Chem.*, vol. 3, no. 2, pp. 119–128, 2019.

[216] X. Li, Y. Xu, H. Yao, and K. Lin, “Chemical space exploration based on recurrent neural networks: applications in discovering kinase inhibitors,” *J. Cheminform.*, vol. 12, no. 1, pp. 1–13, 2020.

[217] K. Kaitoh and Y. Yamanishi, “TRIOMPHE: Transcriptome-Based Inference and Generation of Molecules with Desired Phenotypes by Machine Learning,” *J. Chem. Inf. Model.*, vol. 61, no. 9, pp. 4303–4320, 2021.

[218] R. Zubatyuk, J. S. Smith, J. Leszczynski, and O. Isayev, “Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network,” *Sci. Adv.*, vol. 5, no. 8, p. eaav6490, 2019.

[219] F. H. Vermeire and W. H. Green, “Transfer learning for solvation free energies: From quantum chemistry to experiments,” *Chem. Eng. J.*, vol. 418, p. 129307, 2021.

[220] O. T. Unke and M. Meuwly, “PhysNet: A neural network for predicting energies, forces, dipole moments, and partial charges,” *J. Chem. Theory Comput.*, vol. 15, no. 6, pp. 3678–3693, 2019.

[221] T. W. Ko, J. A. Finkler, S. Goedecker, and J. Behler, “A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer,” *Nat. Commun.*, vol. 12, no. 1, pp. 1–11, 2021.

[222] F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks, and F. Glorius, “A structure-based platform for predicting chemical reactivity,” *Chem*, vol. 6, no. 6, pp. 1379–1390, 2020.

[223] A. Hajibabaei, M. Ha, S. Pourasad, J. Kim, and K. S. Kim, “Machine Learning of First-Principles Force-Fields for Alkane and Polyene Hydrocarbons,” *J. Phys. Chem. A*, vol. 125, no. 42, pp. 9414–9420, 2021.

[224] G. P. Peters *et al.*, “The challenge to keep global warming below 2 C,” *Nat. Clim. Chang.*, vol. 3, no. 1, pp. 4–6, 2013.

[225] S. H. Schneider, “What is’ dangerous’ climate change?,” *Nature*, vol. 411, no. 6833, pp. 17–19, 2001.

[226] M. Rückamp, U. Falk, K. Frieler, S. Lange, and A. Humbert, “The effect of overshooting

1.5° C global warming on the mass loss of the Greenland ice sheet," *Earth Syst. Dyn.*, vol. 9, no. 4, pp. 1169–1189, 2018.

[227] M. V. K. Sivakumar, "Interactions between climate and desertification," *Agric. For. Meteorol.*, vol. 142, no. 2–4, pp. 143–155, 2007.

[228] M. M. Verstraete, A. B. Brink, R. J. Scholes, M. Beniston, and M. S. Smith, "Climate change and desertification: Where do we stand, where should we go?," *Global and Planetary Change*, vol. 64, no. 3–4. Elsevier, pp. 105–110, 2008.

[229] P. A. Owusu and S. Asumadu-Sarkodie, "A review of renewable energy sources, sustainability issues and climate change mitigation," *Cogent Eng.*, vol. 3, no. 1, p. 1167990, 2016.

[230] T. Rajabloo, W. De Ceuninck, L. Van Wortswinkel, M. Rezakazemi, and T. Aminabhavi, "Environmental management of industrial decarbonization with focus on chemical sectors: A review," *J. Environ. Manage.*, vol. 302, p. 114055, 2022.

[231] G. C. Iyer *et al.*, "The contribution of Paris to limit global warming to 2 C," *Environ. Res. Lett.*, vol. 10, no. 12, p. 125002, 2015.

[232] L. Warszawski *et al.*, "All options, not silver bullets, needed to limit global warming to 1.5 C: A scenario appraisal," *Environ. Res. Lett.*, vol. 16, no. 6, p. 64037, 2021.

[233] H. Singh, "CCUS: status and priorities for research and development." https://decarbonisationtechnology.com/article/54/ccus-status-and-priorities-for-research-and-development#.Yn1syYfMJPZ (accessed May 12, 2022).

[234] NBC News, "Global carbon emissions bounce back to nearly 2019 levels, study finds." https://www.nbcnews.com/news/world/global-carbon-emissions-bounce-back-nearly-2019-levels-n1283167 (accessed Jan. 29, 2023).

[235] S. Bierbaumer *et al.*, "Enzymatic Conversion of CO2: From Natural to Artificial Utilization," *Chem. Rev.*, vol. 123, no. 9, pp. 5702–5754, 2023.

[236] T. Beyazay *et al.*, "Ambient temperature CO2 fixation to pyruvate and subsequently to citramalate over iron and nickel nanoparticles," *Nat. Commun.*, vol. 14, no. 1, p. 570, 2023.

[237] A. Demessence, D. M. D'Alessandro, M. L. Foo, and J. R. Long, "Strong CO2 binding in a

water-stable, triazolate-bridged metal-organic framework functionalized with ethylenediamine," *J. Am. Chem. Soc.*, vol. 131, no. 25, pp. 8784–8786, 2009, doi: 10.1021/ja903411w.

[238] R. Pardemann and B. Meyer, "Pre-Combustion Carbon Capture," *Handb. Clean Energy Syst.*, pp. 1–28, 2015.

[239] S. Yadav and S. S. Mondal, "A review on the progress and prospects of oxy-fuel carbon capture and sequestration (CCS) technology," *Fuel*, vol. 308, p. 122057, 2022.

[240] O. Shekhah *et al.*, "Made-to-order metal-organic frameworks for trace carbon dioxide removal and air capture," *Nat. Commun.*, vol. 5, no. May, pp. 1–7, 2014, doi: 10.1038/ncomms5228.

[241] O. Aschenbrenner and P. Styring, "Comparative study of solvent properties for carbon dioxide absorption," *Energy \& Environ. Sci.*, vol. 3, no. 8, pp. 1106–1113, 2010.

[242] Y. Cao *et al.*, "Evaluation of the rapid phase change absorbents based on potassium glycinate for CO2 capture," *Chem. Eng. Sci.*, vol. 273, p. 118627, 2023.

[243] L. Liu, M. Fang, S. Xu, J. Wang, and D. Guo, "Development and testing of a new post-combustion CO2 capture solvent in pilot and demonstration plant," *Int. J. Greenh. Gas Control*, vol. 113, p. 103513, 2022.

[244] C. Font-Palma, D. Cann, and C. Udemu, "Review of cryogenic carbon capture innovations and their potential applications," *C*, vol. 7, no. 3, p. 58, 2021.

[245] S. Abuelgasim, W. Wang, and A. Abdalazeez, "A brief review for chemical looping combustion as a promising CO2 capture technology: Fundamentals and progress," *Sci. Total Environ.*, vol. 764, p. 142892, 2021.

[246] A. M. Arias, M. C. Mussati, P. L. Mores, N. J. Scenna, J. A. Caballero, and S. F. Mussati, "Optimization of multi-stage membrane systems for CO2 capture from flue gas," *Int. J. Greenh. Gas Control*, vol. 53, pp. 371–390, 2016.

[247] Y. Zhao, X. Liu, K. X. Yao, L. Zhao, and Y. Han, "Superior Capture of CO 2 Achieved by Introducing Extra-framework Cations into N - doped Microporous Carbon," 2012.

[248] Q. Liu, N. C. O. Cheung, A. E. Garcia-Bennett, and N. Hedin, "Aluminophosphates for CO2

separation," *ChemSusChem*, vol. 4, no. 1, pp. 91–97, 2011, doi: 10.1002/cssc.201000256.

[249] R. V. Siriwardane, M. S. Shen, E. P. Fisher, and J. Losch, "Adsorption of CO2 on zeolites at moderate temperatures," *Energy and Fuels*, vol. 19, no. 3, pp. 1153–1159, 2005, doi: 10.1021/ef040059h.

[250] Y. Shen, Z. Niu, R. Zhang, and D. Zhang, "Vacuum pressure swing adsorption process with carbon molecular sieve for CO2 separation from biogas," *J. CO2 Util.*, vol. 54, p. 101764, 2021.

[251] J. Wilcox, "An electro-swing approach," *Nat. Energy*, vol. 5, no. 2, pp. 121–122, 2020.

[252] X. Shi *et al.*, "Sorbents for the direct capture of CO2 from ambient air," *Angew. Chemie Int. Ed.*, vol. 59, no. 18, pp. 6984–7006, 2020.

[253] V. S. Derevschikov, J. V Veselovskaya, T. Y. Kardash, D. A. Trubitsyn, and A. G. Okunev, "Direct CO2 capture from ambient air using K2CO3/Y2O3 composite sorbent," *Fuel*, vol. 127, pp. 212–218, 2014.

[254] J. V Veselovskaya, V. S. Derevschikov, T. Y. Kardash, O. A. Stonkus, T. A. Trubitsina, and A. G. Okunev, "Direct CO2 capture from ambient air using K2CO3/Al2O3 composite sorbent," *Int. J. Greenh. Gas Control*, vol. 17, pp. 332–340, 2013.

[255] J. Wang, M. Wang, W. Li, W. Qiao, D. Long, and L. Ling, "Application of polyethylenimine-impregnated solid adsorbents for direct capture of low-concentration CO2," *AIChE J.*, vol. 61, no. 3, pp. 972–980, 2015.

[256] C. Chen, J. Kim, and W. S. Ahn, "CO2capture by amine-functionalized nanoporous materials: A review," *Korean J. Chem. Eng.*, vol. 31, no. 11, pp. 1919–1934, 2014, doi: 10.1007/s11814-014-0257-2.

[257] C. A. Trickett, A. Helal, B. A. Al-Maythalony, Z. H. Yamani, K. E. Cordova, and O. M. Yaghi, "The chemistry of metal--organic frameworks for CO2 capture, regeneration and conversion," *Nat. Rev. Mater.*, vol. 2, no. 8, pp. 1–16, 2017.

[258] R. Krishna and J. M. van Baten, "A comparison of the CO2 capture characteristics of zeolites and metal--organic frameworks," *Sep. Purif. Technol.*, vol. 87, pp. 120–126, 2012.

[259] A. E. Creamer and B. Gao, "Carbon-based adsorbents for postcombustion CO2 capture: a

critical review," *Environ. Sci. \& Technol.*, vol. 50, no. 14, pp. 7276–7289, 2016.

[260] M. Bermeo, L. F. Vega, M. R. M. Abu-Zahra, and M. Khaleel, "Critical assessment of the performance of next-generation carbon-based adsorbents for CO2 capture focused on their structural properties," *Sci. Total Environ.*, vol. 810, p. 151720, 2022.

[261] L. Zou *et al.*, "Porous organic polymers for post-combustion carbon capture," *Adv. Mater.*, vol. 29, no. 37, p. 1700229, 2017.

[262] J. Wang *et al.*, "Recent progress in porous organic polymers and their application for CO2 capture," *Chinese J. Chem. Eng.*, vol. 42, pp. 91–103, 2022.

[263] S. K. Tiwari, B. S. Giri, S. Tantuvoy, S. M. S. Nagendra, and V. Katiyar, "CO2 removal using alkaline waste as a solid adsorbent: Challenges and forthcoming directions," in *Novel Materials for Environmental Remediation Applications*, Elsevier, 2023, pp. 399–411.

[264] R. Wu and A. Bao, "Preparation of cellulose carbon material from cow dung and its CO2 adsorption performance," *J. CO2 Util.*, vol. 68, p. 102377, 2023.

[265] H. A. Patel, J. Byun, and C. T. Yavuz, "Carbon dioxide capture adsorbents: chemistry and methods," *ChemSusChem*, vol. 10, no. 7, pp. 1303–1317, 2017.

[266] W. Gao, T. Zhou, Y. Gao, B. Louis, D. O'Hare, and Q. Wang, "Molten salts-modified MgO-based adsorbents for intermediate-temperature CO2 capture: A review," *J. energy Chem.*, vol. 26, no. 5, pp. 830–838, 2017.

[267] X. Gao, S. Yang, L. Hu, S. Cai, L. Wu, and S. Kawi, "Carbonaceous materials as adsorbents for CO2 capture: synthesis and modification," *Carbon Capture Sci. \& Technol.*, p. 100039, 2022.

[268] H. Madzaki, W. A. W. A. K. Ghani, T. C. S. Yaw, U. Rashid, and N. Muda, "Carbon dioxide adsorption on activated carbon hydrothermally treated and impregnated with metal oxides," *J. Kejuruter.*, vol. 30, no. 1, pp. 31–38, 2018.

[269] T. C. Drage, A. Arenillas, K. M. Smith, C. Pevida, S. Piippo, and C. E. Snape, "Preparation of carbon dioxide adsorbents from the chemical activation of urea--formaldehyde and melamine--formaldehyde resins," *Fuel*, vol. 86, no. 1–2, pp. 22–31, 2007.

[270] B. Kaur, R. K. Gupta, and H. Bhunia, "Chemically activated nanoporous carbon adsorbents

from waste plastic for CO2 capture: Breakthrough adsorption study," *Microporous Mesoporous Mater.*, vol. 282, pp. 146–158, 2019.

[271] Y. Jia, X. Hou, Z. Wang, and X. Hu, "Machine learning boosts the design and discovery of nanomaterials," *ACS Sustain. Chem. \& Eng.*, vol. 9, no. 18, pp. 6130–6147, 2021.

[272] Y. Juan, Y. Dai, Y. Yang, and J. Zhang, "Accelerating materials discovery using machine learning," *J. Mater. Sci. \& Technol.*, vol. 79, pp. 178–190, 2021.

[273] H. Mai, T. C. Le, D. Chen, D. A. Winkler, and R. A. Caruso, "Machine Learning in the Development of Adsorbents for Clean Energy Application and Greenhouse Gas Capture," *Adv. Sci.*, p. 2203899, 2022.

[274] A. Grimm and M. Gazzani, "A Machine Learning-Aided Equilibrium Model of VTSA Processes for Sorbents Screening Applied to CO2 Capture from Diluted Sources," *Ind. \& Eng. Chem. Res.*, vol. 61, no. 37, pp. 14004–14019, 2022.

[275] X. Zhu *et al.*, "Machine learning exploration of the direct and indirect roles of Fe impregnation on Cr (VI) removal by engineered biochar," *Chem. Eng. J.*, vol. 428, p. 131967, 2022.

[276] W. Jiang, X. Xing, S. Li, X. Zhang, and W. Wang, "Synthesis, characterization and machine learning based performance prediction of straw activated carbon," *J. Clean. Prod.*, vol. 212, pp. 1210–1223, 2019.

[277] E. O. Pyzer-Knapp *et al.*, "Accelerating materials discovery using artificial intelligence, high performance computing and robotics," *npj Comput. Mater.*, vol. 8, no. 1, pp. 1–9, 2022.

[278] A. M. Mroz, V. Posligua, A. Tarzia, E. H. Wolpert, and K. E. Jelfs, "Into the Unknown: How Computation Can Help Explore Uncharted Material Space," *J. Am. Chem. Soc.*, vol. 144, no. 41, pp. 18730–18743, 2022.

[279] G. Wang *et al.*, "ALKEMIE: An intelligent computational platform for accelerating materials discovery and design," *Comput. Mater. Sci.*, vol. 186, p. 110064, 2021.

[280] S. G. Subraveti, "Machine learning-based design and techno-economic assessments of adsorption processes for CO2 capture," 2021.

[281] M. J. Regufe *et al.*, "Adsorption material composition and process optimization, a

systematical approach based on Deep Learning," *IFAC-PapersOnLine*, vol. 54, no. 3, pp. 43–48, 2021.

[282] M. Rahimi, M. H. Abbaspour-Fard, A. Rohani, O. Yuksel Orhan, and X. Li, "Modeling and Optimizing N/O-Enriched Bio-Derived Adsorbents for CO2 Capture: Machine Learning and DFT Calculation Approaches," *Ind. \& Eng. Chem. Res.*, vol. 61, no. 30, pp. 10670–10688, 2022.

[283] T. E. Akinola, P. L. B. Prado, and M. Wang, "Experimental studies, molecular simulation and process modelling$\backslash$simulation of adsorption-based post-combustion carbon capture for power plants: A state-of-the-art review," *Appl. Energy*, vol. 317, p. 119156, 2022.

[284] X. Zhu *et al.*, "Machine learning exploration of the critical factors for CO2 adsorption capacity on porous carbon materials at different pressures," *J. Clean. Prod.*, vol. 273, p. 122915, 2020.

[285] H. A. H. Mahmoud, N. A. Hakami, and A. M. Hafez, "An Intelligent Deep Learning Model for Adsorption Prediction," *Adsorpt. Sci. \& Technol.*, vol. 2022, 2022.

[286] C. Zhang, Y. Xie, C. Xie, H. Dong, L. Zhang, and J. Lin, "Accelerated discovery of porous materials for carbon capture by machine learning: A review," *MRS Bull.*, pp. 1–8, 2022.

[287] X. Yuan *et al.*, "Applied machine learning for prediction of CO2 adsorption on biomass waste-derived porous carbons," *Environ. Sci. \& Technol.*, vol. 55, no. 17, pp. 11925–11936, 2021.

[288] F. Fathalian, S. Aarabi, A. Ghaemi, and A. Hemmati, "Intelligent prediction models based on machine learning for CO2 capture performance by graphene oxide-based adsorbents," *Sci. Rep.*, vol. 12, no. 1, pp. 1–20, 2022.

[289] H. Dureckova, M. Krykunov, M. Z. Aghaji, and T. K. Woo, "Robust machine learning models for predicting high CO2 working capacity and CO2/H2 selectivity of gas adsorption in metal organic frameworks for precombustion carbon capture," *J. Phys. Chem. C*, vol. 123, no. 7, pp. 4133–4139, 2019.

[290] R. Anderson, J. Rodgers, E. Argueta, A. Biong, and D. A. Gómez-Gualdrón, "Role of pore chemistry and topology in the CO2 capture capabilities of MOFs: from molecular simulation

to machine learning," *Chem. Mater.*, vol. 30, no. 18, pp. 6325–6337, 2018.

[291] M. Fernandez, P. G. Boyd, T. D. Daff, M. Z. Aghaji, and T. K. Woo, "Rapid and accurate machine learning recognition of high performing metal organic frameworks for CO2 capture," *J. Phys. Chem. Lett.*, vol. 5, no. 17, pp. 3056–3060, 2014.

[292] K. Low, M. L. Coote, and E. I. Izgorodina, "Explainable solvation free energy prediction combining graph neural networks with chemical intuition," *J. Chem. Inf. Model.*, vol. 62, no. 22, pp. 5457–5470, 2022.

[293] X. Zhong, B. Gallagher, S. Liu, B. Kailkhura, A. Hiszpanski, and T. Han, "Explainable machine learning in materials science," *npj Comput. Mater.*, vol. 8, no. 1, pp. 1–19, 2022.

[294] F. Oviedo, J. L. Ferres, T. Buonassisi, and K. T. Butler, "Interpretable and explainable machine learning for materials science and chemistry," *Accounts Mater. Res.*, vol. 3, no. 6, pp. 597–607, 2022.

[295] J. Dieber and S. Kirrane, "Why model why? Assessing the strengths and limitations of LIME," *arXiv Prepr. arXiv2012.00093*, 2020.

[296] D. Bowen and L. Ungar, "Generalized SHAP: Generating multiple types of explanations in machine learning," *arXiv Prepr. arXiv2006.07155*, 2020.

[297] A. S. Anker *et al.*, "Extracting structural motifs from pair distribution function data of nanostructures using explainable machine learning," *npj Comput. Mater.*, vol. 8, no. 1, pp. 1–11, 2022.

[298] C. Xie, Y. Xie, C. Zhang, H. Dong, and L. Zhang, "Explainable machine learning for carbon dioxide adsorption on porous carbon," *J. Environ. Chem. Eng.*, vol. 11, no. 1, p. 109053, 2023.

[299] C. Altintas *et al.*, "Database for CO2 separation performances of MOFs based on computational materials screening," *ACS Appl. Mater. \& interfaces*, vol. 10, no. 20, pp. 17257–17268, 2018.

[300] "COSMOS (Computational Simulations of MOFs for Gas Separations)." https://cosmoserc.ku.edu.tr/.

[301] M. Irani, A. T. Jacobson, K. A. M. Gasem, and M. Fan, "Modified carbon

nanotubes/tetraethylenepentamine for CO2 capture," *Fuel*, vol. 206, pp. 10–18, 2017.

[302] I. A. Iugai *et al.*, "MgO/carbon nanofibers composite coatings on porous ceramic surface for CO2 capture," *Surf. Coatings Technol.*, vol. 400, p. 126208, 2020.

[303] N. Iqbal, X. Wang, J. Yu, and B. Ding, "Robust and flexible carbon nanofibers doped with amine functionalized carbon nanotubes for efficient CO2 capture," *Adv. Sustain. Syst.*, vol. 1, no. 3–4, p. 1600028, 2017.

[304] C. Lu, H. Bai, B. Wu, F. Su, and J. F. Hwang, "Comparative study of CO2 capture by carbon nanotubes, activated carbons, and zeolites," *Energy and Fuels*, vol. 22, no. 5, pp. 3050–3056, 2008, doi: 10.1021/ef8000086.

[305] S. Shan, Q. Jia, L. Jiang, Q. Li, Y. Wang, and J. Peng, "Novel Li4SiO4-based sorbents from diatomite for high temperature CO2 capture," *Ceram. Int.*, vol. 39, no. 5, pp. 5437–5441, 2013.

[306] Y. Zhang *et al.*, "Recent advances in lithium containing ceramic based sorbents for high-temperature CO 2 capture," *J. Mater. Chem. A*, vol. 7, no. 14, pp. 7962–8005, 2019.

[307] M. Minelli, V. Medri, E. Papa, F. Miccio, E. Landi, and F. Doghieri, "Geopolymers as solid adsorbent for CO2 capture," *Chem. Eng. Sci.*, vol. 148, pp. 267–274, 2016.

[308] R. Rodr\'\iguez-Mosqueda and H. Pfeiffer, "High CO2 capture in sodium metasilicate (Na2SiO3) at low temperatures (30--60 C) through the CO2--H2O chemisorption process," *J. Phys. Chem. C*, vol. 117, no. 26, pp. 13452–13461, 2013.

[309] M. L. T. Triviño, V. Hiremath, and J. G. Seo, "Stabilization of NaNO3-promoted magnesium oxide for high-temperature CO2 capture," *Environ. Sci. \& Technol.*, vol. 52, no. 20, pp. 11952–11959, 2018.

[310] K. Ho, S. Jin, M. Zhong, A.-T. Vu, and C.-H. Lee, "Sorption capacity and stability of mesoporous magnesium oxide in post-combustion CO2 capture," *Mater. Chem. Phys.*, vol. 198, pp. 154–161, 2017.

[311] W. Liu *et al.*, "Performance enhancement of calcium oxide sorbents for cyclic CO2 capture□ A review," *Energy \& Fuels*, vol. 26, no. 5, pp. 2751–2767, 2012.

[312] L. Li, D. L. King, Z. Nie, and C. Howard, "Magnesia-stabilized calcium oxide absorbents

with improved durability for high temperature CO2 capture," *Ind. \& Eng. Chem. Res.*, vol. 48, no. 23, pp. 10604–10613, 2009.

[313] H. Gao, Q. Li, and S. Ren, "Progress on CO2 capture by porous organic polymers," *Curr. Opin. Green Sustain. Chem.*, vol. 16, pp. 33–38, 2019.

[314] C. Bläker, J. Muthmann, C. Pasel, and D. Bathen, "Characterization of activated carbon adsorbents--state of the art and novel approaches," *ChemBioEng Rev.*, vol. 6, no. 4, pp. 119–138, 2019.

[315] R. Bardestani, G. S. Patience, and S. Kaliaguine, "Experimental methods in chemical engineering: specific surface area and pore size distribution measurements—BET, BJH, and DFT," *Can. J. Chem. Eng.*, vol. 97, no. 11, pp. 2781–2791, 2019.

[316] J. R. Li *et al.*, "Porous materials with pre-designed single-molecule traps for CO 2 selective adsorption," *Nat. Commun.*, vol. 4, pp. 1–8, 2013, doi: 10.1038/ncomms2552.

[317] J. An, S. J. Geib, and N. L. Rosi, "High and selective CO2 uptake in a cobalt adeninate metal- organic framework exhibiting pyrimidine-and amino-decorated pores," *J. Am. Chem. Soc.*, vol. 132, no. 1, pp. 38–39, 2010.

[318] M. Ismail, M. A. Bustam, N. E. F. Kari, and Y. F. Yeong, "Ideal Adsorbed Solution Theory (IAST) of Carbon Dioxide and Methane Adsorption Using Magnesium Gallate Metal-Organic Framework (Mg-gallate)," *Molecules*, vol. 28, no. 7, p. 3016, 2023.

[319] T. Ben *et al.*, "Selective adsorption of carbon dioxide by carbonized porous aromatic framework (PAF)," *Energy \& Environ. Sci.*, vol. 5, no. 8, pp. 8370–8376, 2012.

[320] S. Himeno, T. Tomita, K. Suzuki, and S. Yoshida, "Characterization and selectivity for methane and carbon dioxide adsorption on the all-silica DD3R zeolite," *Microporous Mesoporous Mater.*, vol. 98, no. 1–3, pp. 62–69, 2007.

[321] Natural Resources Canada NRCan, "EXPLORE." https://www.nrcan.gc.ca/maps-tools-and-publications/tools/modelling-tools/explore/24824 (accessed Jan. 01, 2023).

[322] M. S. Gold and P. M. Bentler, "Treatments of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization," *Struct. Equ. Model.*, vol. 7, no. 3, pp. 319–355, 2000.

[323] F. V Nelwamondo, S. Mohamed, and T. Marwala, “Missing data: A comparison of neural network and expectation maximization techniques,” *Curr. Sci.*, pp. 1514–1521, 2007.

[324] A. Pantanowitz and T. Marwala, “Missing data imputation through the use of the random forest algorithm,” in *Advances in computational intelligence*, 2009, pp. 53–62.

[325] J. Yoon, J. Jordon, and M. Schaar, “Gain: Missing data imputation using generative adversarial nets,” in *International conference on machine learning*, 2018, pp. 5689–5698.

[326] T. K. Moon, “The expectation-maximization algorithm,” *IEEE Signal Process. Mag.*, vol. 13, no. 6, pp. 47–60, 1996.

[327] B. K. Beaulieu-Jones, J. H. Moore, and P. R. O.-A. A. L. S. C. T. CONSORTIUM, “Missing data imputation in the electronic health record using deeply learned autoencoders,” in *Pacific symposium on biocomputing 2017*, 2017, pp. 207–218.

[328] P. Khosravi, A. Vergari, Y. Choi, Y. Liang, and G. Van den Broeck, “Handling missing data in decision trees: A probabilistic approach,” *arXiv Prepr. arXiv2006.16341*, 2020.

[329] X. Zhang, C. Yan, C. Gao, B. A. Malin, and Y. Chen, “Predicting missing values in medical data via XGBoost regression,” *J. Healthc. informatics Res.*, vol. 4, pp. 383–394, 2020.

[330] R. Polikar, “Ensemble learning,” *Ensemble Mach. Learn. Methods Appl.*, pp. 1–34, 2012.

[331] R. Polikar, “Ensemble based systems in decision making,” *IEEE Circuits Syst. Mag.*, vol. 6, no. 3, pp. 21–45, 2006.

[332] K. Nadim, A. Ragab, and M.-S. Ouali, “Data-driven dynamic causality analysis of industrial systems using interpretable machine learning and process mining,” *J. Intell. Manuf.*, pp. 1–27, 2022.

[333] S. E. Kentish, C. A. Scholes, and G. W. Stevens, “Carbon dioxide separation through polymeric membrane systems for flue gas applications,” *Recent Patents Chem. Eng.*, vol. 1, no. 1, pp. 52–66, 2008.

[334] H. N. Nguyen and D. A. Khuong, “New Trends in Pyrolysis Methods: Opportunities, Limitations, and Advantages,” in *Engineered Biochar: Fundamentals, Preparation, Characterization and Applications*, Springer, 2022, pp. 105–126.

[335] M. Patil *et al.*, “Synthesis and Characterization of Microwave-Assisted Copolymer

Membranes of Poly (vinyl alcohol)-g-starch-methacrylate and Their Evaluation for Gas Transport Properties," *Polymers (Basel).*, vol. 14, no. 2, p. 350, 2022.

[336] R. S. Maier and M. R. Schure, "Transport properties and size exclusion effects in wide-pore superficially porous particles," *Chem. Eng. Sci.*, vol. 185, pp. 243–255, 2018.

[337] X. Hu *et al.*, "Insights on size-exclusion effect of ordered mesoporous carbon for selective antibiotics adsorption under the interference of natural organic matter," *Chem. Eng. J.*, vol. 458, p. 141440, 2023.

[338] N. Drossis, M. A. Gauthier, and H. W. de Haan, "Elucidating the mechanisms of the molecular sieving phenomenon created by comb-shaped polymers grafted to a protein--a simulation study," *Mater. Today Chem.*, vol. 24, p. 100861, 2022.

[339] S. Kunze, R. Groll, B. Besser, and J. Thöming, "Molecular diameters of rarefied gases," *Sci. Rep.*, vol. 12, no. 1, p. 2057, 2022.

[340] P. Z. Sun *et al.*, "Exponentially selective molecular sieving through angstrom pores," *Nat. Commun.*, vol. 12, no. 1, p. 7170, 2021.

[341] S. H. Hashemi, Z. Besharati, and S. A. Hashemi, "Salicylic acid solubility prediction in different solvents based on machine learning algorithms," *Digit. Chem. Eng.*, p. 100157, 2024.

[342] A. A. Adeleke *et al.*, "Comparative studies of machine learning models for predicting higher heating values of biomass," *Digit. Chem. Eng.*, vol. 12, p. 100159, 2024.

[343] S. A. A. Taqvi, H. Zabiri, L. D. Tufa, F. Uddin, S. A. Fatima, and A. S. Maulud, "A review on data-driven learning approaches for fault detection and diagnosis in chemical processes," *ChemBioEng Rev.*, vol. 8, no. 3, pp. 239–259, 2021.

[344] D. A. Akinpelu, O. A. Adekoya, P. O. Oladoye, C. C. Ogbaga, and J. A. Okolie, "Machine learning applications in biomass pyrolysis: from biorefinery to end-of-life product management," *Digit. Chem. Eng.*, vol. 8, p. 100103, 2023.

[345] E. Y. Emori, M. A. S. S. Ravagnani, and C. B. B. Costa, "Application of a predictive Q-learning algorithm on the multiple-effect evaporator in a sugarcane ethanol biorefinery," *Digit. Chem. Eng.*, vol. 5, p. 100049, 2022.

[346] K. E. S. Pilario, P. M. L. Ching, A. M. A. Calapatia, and A. B. Culaba, “Predicting drying curves in algal biorefineries using Gaussian process autoregressive models,” *Digit. Chem. Eng.*, vol. 4, p. 100036, 2022.

[347] S. Chmiela *et al.*, “Accurate global machine learning force fields for molecules with hundreds of atoms,” *Sci. Adv.*, vol. 9, no. 2, p. eadf0873, 2023.

[348] J. Götz *et al.*, “High-throughput synthesis provides data for predicting molecular properties and reaction success,” *Sci. Adv.*, vol. 9, no. 43, p. eadj2314, 2023.

[349] D. York, I. Vidal-Daza, C. Segura, J. Norambuena-Contreras, and F. J. Martin-Martinez, “Data-driven representative models to accelerate scaled-up atomistic simulations of bitumen and biobased complex fluids,” *Digit. Discov.*, vol. 3, no. 6, pp. 1108–1122, 2024.

[350] A. Arias, G. Feijoo, and M. T. Moreira, “How could Artificial Intelligence be used to increase the potential of biorefineries in the near future? A review,” *Environ. Technol. \& Innov.*, vol. 32, p. 103277, 2023.

[351] V. Zuin Zeidler, “Digitalization paving the ways for sustainable chemistry: switching on more green lights,” *Science*, vol. 384, no. 6701. American Association for the Advancement of Science, p. eadq3537, 2024.

[352] H. Wen *et al.*, “A systematic modeling methodology of deep neural network-based structure-property relationship for rapid and reliable prediction on flashpoints,” *AIChE J.*, vol. 68, no. 1, p. e17402, 2022.

[353] Y. Su, Z. Wang, S. Jin, W. Shen, J. Ren, and M. R. Eden, “An architecture of deep learning in QSPR modeling for the prediction of critical properties using molecular signatures,” *AIChE J.*, vol. 65, no. 9, p. e16678, 2019.

[354] J. Zhang, Q. Wang, M. Eden, and W. Shen, “A deep learning-based framework towards inverse green solvent design for extractive distillation with multi-index constraints,” *Comput. \& Chem. Eng.*, vol. 177, p. 108335, 2023.

[355] Z. Wang *et al.*, “Insights into ensemble learning-based data-driven model for safety-related property of chemical substances,” *Chem. Eng. Sci.*, vol. 248, p. 117219, 2022.

[356] Y. Su, S. Jin, X. Zhang, W. Shen, M. R. Eden, and J. Ren, “Stakeholder-oriented multi-

objective process optimization based on an improved genetic algorithm," *Comput. \& Chem. Eng.*, vol. 132, p. 106618, 2020.

[357] H. Wen *et al.*, "A Systematic Review on Intensifications of Artificial Intelligence Assisted Green Solvent Development," *Ind. \& Eng. Chem. Res.*, vol. 62, no. 48, pp. 20473–20491, 2023.

[358] C. L. Ritt, M. Liu, T. A. Pham, R. Epsztein, H. J. Kulik, and M. Elimelech, "Machine learning reveals key ion selectivity mechanisms in polymeric membranes with subnanometer pores," *Sci. Adv.*, vol. 8, no. 2, p. eabl5771, 2022.

[359] O. T. Unke *et al.*, "Biomolecular dynamics with machine-learned quantum-mechanical force fields trained on diverse chemical fragments," *Sci. Adv.*, vol. 10, no. 14, p. eadn4397, 2024.

[360] B. Sanchez-Lengeling and A. Aspuru-Guzik, "Inverse molecular design using machine learning: Generative models for matter engineering," *Science (80-.).*, vol. 361, no. 6400, pp. 360–365, 2018.

[361] J. Zhang, Q. Wang, and W. Shen, "Message-passing neural network based multi-task deep-learning framework for COSMO-SAC based $σ$-profile and VCOSMO prediction," *Chem. Eng. Sci.*, vol. 254, p. 117624, 2022.

[362] N. Khan and S. A. Ammar Taqvi, "Machine learning an intelligent approach in process industries: A perspective and overview," *ChemBioEng Rev.*, vol. 10, no. 2, pp. 195–221, 2023.

[363] J. Zhang *et al.*, "An accurate and interpretable deep learning model for environmental properties prediction using hybrid molecular representations," *AIChE J.*, vol. 68, no. 6, p. e17634, 2022.

[364] Z. Wang *et al.*, "A novel unambiguous strategy of molecular feature extraction in machine learning assisted predictive models for environmental properties," *Green Chem.*, vol. 22, no. 12, pp. 3867–3876, 2020.

[365] Z. Wang *et al.*, "Predictive deep learning models for environmental properties: the direct calculation of octanol--water partition coefficients from molecular graphs," *Green Chem.*, vol. 21, no. 16, pp. 4555–4565, 2019.

[366] K. M. Jablonka *et al.*, “Machine learning for industrial processes: Forecasting amine emissions from a carbon capture plant,” *Sci. Adv.*, vol. 9, no. 1, p. eadc9576, 2023.

[367] A. F. Zahrt, J. J. Henle, B. T. Rose, Y. Wang, W. T. Darrow, and S. E. Denmark, “Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning,” *Science (80-.).*, vol. 363, no. 6424, p. eaau5631, 2019.

[368] M. Meuwly, “Machine learning for chemical reactions,” *Chem. Rev.*, vol. 121, no. 16, pp. 10218–10239, 2021.

[369] L. He *et al.*, “Reaction condition-and functional group-specific knowledge discovery: Data- and computation-based analysis on transition-metal-free transformation of organoborons,” *Artif. Intell. Chem.*, vol. 2, no. 1, p. 100034, 2024.

[370] O. Ajao, M. Benali, and N. El Mehdi, “Experimental and computer aided solubility quantification of diverse lignins and performance prediction,” *Chem. Commun.*, vol. 57, no. 14, pp. 1782–1785, 2021.

[371] R. M. O’Dea *et al.*, “Ambient-pressure lignin valorization to high-performance polymers by intensified reductive catalytic deconstruction,” *Sci. Adv.*, vol. 8, no. 3, p. eabj7523, 2022.

[372] G. Ginni *et al.*, “Valorization of agricultural residues: Different biorefinery routes,” *J. Environ. Chem. Eng.*, vol. 9, no. 4, p. 105435, 2021.

[373] M. Li, X. Liu, C. Sun, L. Stevens, and H. Liu, “Synthesis and characterization of advanced bio-carbon materials from Kraft lignin with enhanced CO2 capture properties,” *J. Environ. Chem. Eng.*, vol. 10, no. 3, p. 107471, 2022.

[374] R. Li *et al.*, “Selective value-added conversion of lignin derivatives over heterogeneous catalysts of TEMPO-functionalized metal-organic frameworks,” *J. Environ. Chem. Eng.*, vol. 11, no. 3, p. 109700, 2023.

[375] A. Arias, S. Estévez-Rivadulla, R. Rebolledo-Leiva, G. Feijoo, S. González-Garc\’\ia, and M. T. Moreira, “Boosting the transition to biorefineries in compliance with sustainability and circularity criteria,” *J. Environ. Chem. Eng.*, vol. 12, no. 5, p. 113361, 2024.

[376] H. K. Balsora *et al.*, “Machine learning approach for the prediction of biomass pyrolysis kinetics from preliminary analysis,” *J. Environ. Chem. Eng.*, vol. 10, no. 3, p. 108025, 2022.

[377] J. Lofgren, D. Tarasov, T. Koitto, P. Rinke, M. Balakshin, and M. Todorovic, “Machine learning optimization of lignin properties in green biorefineries,” *ACS Sustain. Chem. \& Eng.*, vol. 10, no. 29, pp. 9469–9479, 2022.

[378] N. H. Khashaba, R. S. Ettouney, M. M. Abdelaal, F. H. Ashour, and M. A. El-Rifai, “Artificial neural network modeling of biochar enhanced anaerobic sewage sludge digestion,” *J. Environ. Chem. Eng.*, vol. 10, no. 4, p. 107988, 2022.

[379] A. C. Garcia, C. Shuo, and J. S. Cross, “Machine learning based analysis of reaction phenomena in catalytic lignin depolymerization,” *Bioresour. Technol.*, vol. 345, p. 126503, 2022.

[380] H. Ge *et al.*, “Machine learning prediction of delignification and lignin structure regulation of deep eutectic solvents pretreatment processes,” *Ind. Crops Prod.*, vol. 203, p. 117138, 2023.

[381] S. Varshney, C. V. Lakshmi, and C. Patvardhan, “Madhubani Art Classification using transfer learning with deep feature fusion and decision fusion based techniques,” *Eng. Appl. Artif. Intell.*, vol. 119, p. 105734, 2023.

[382] Y. He, R. Lou, Y. Wang, J. Wang, and X. Fang, “A dual attribute weighted decision fusion system for fault classification based on an extended analytic hierarchy process,” *Eng. Appl. Artif. Intell.*, vol. 114, p. 105066, 2022.

[383] M. Morimoto, K. Fukami, K. Zhang, and K. Fukagata, “Generalization techniques of neural networks for fluid flow estimation,” *Neural Comput. Appl.*, pp. 1–23, 2022.

[384] M. Sester, Y. Feng, and F. Thiemann, “Building generalization using deep learning,” *ISPRS-International Arch. Photogramm. Remote Sens. Spat. Inf. Sci. XLII-4*, vol. 42, pp. 565–572, 2018.

[385] L. Hui, M. Belkin, and P. Nakkiran, “Limitations of neural collapse for understanding generalization in deep learning,” *arXiv Prepr. arXiv2202.08384*, 2022.

[386] K. Tidriri, T. Tiplica, N. Chatti, and S. Verron, “A generic framework for decision fusion in fault detection and diagnosis,” *Eng. Appl. Artif. Intell.*, vol. 71, pp. 73–86, 2018.

[387] W. Wang, M. Zhang, G. Chen, H. V Jagadish, B. C. Ooi, and K.-L. Tan, “Database meets

deep learning: Challenges and opportunities," *ACM Sigmod Rec.*, vol. 45, no. 2, pp. 17–22, 2016.

[388] L. Brigato and L. Iocchi, "A close look at deep learning with small data," in *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021, pp. 2490–2497.

[389] L. Peng, M. Peng, B. Liao, G. Huang, W. Li, and D. Xie, "The advances and challenges of deep learning application in biological big data processing," *Curr. Bioinform.*, vol. 13, no. 4, pp. 352–359, 2018.

[390] R. Ghorbani and R. Ghousi, "Comparing different resampling methods in predicting students' performance using machine learning techniques," *IEEE Access*, vol. 8, pp. 67899–67911, 2020.

[391] S. P. Adam, S.-A. N. Alexandropoulos, P. M. Pardalos, and M. N. Vrahatis, "No free lunch theorem: A review," *Approx. Optim. Algorithms, Complex. Appl.*, pp. 57–82, 2019.

[392] D. H. Wolpert, "The supervised learning no-free-lunch theorems," *Soft Comput. Ind. Recent Appl.*, pp. 25–42, 2002.

[393] A. Ragab, H. Ghezzaz, and M. Amazouz, "Decision fusion for reliable fault classification in energy-intensive process industries," *Comput. Ind.*, vol. 138, p. 103640, 2022.

[394] C. D. Sutton, "Classification and regression trees, bagging, and boosting," *Handb. Stat.*, vol. 24, pp. 303–329, 2005.

[395] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, pp. 123–140, 1996.

[396] R. E. Schapire, "The boosting approach to machine learning: An overview," *Nonlinear Estim. Classif.*, pp. 149–171, 2003.

[397] R. E. Schapire and Y. Freund, "Boosting: Foundations and algorithms," *Kybernetes*, vol. 42, no. 1, pp. 164–166, 2013.

[398] R. E. Schapire, "Explaining adaboost," in *Empirical Inference: Festschrift in Honor of Vladimir N. Vapnik*, Springer, 2013, pp. 37–52.

[399] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.

[400] A. I. Naimi and L. B. Balzer, "Stacked generalization: an introduction to super learning,"

Eur. J. Epidemiol., vol. 33, pp. 459–464, 2018.

[401] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*, 2000, pp. 1–15.

[402] R. Atallah and A. Al-Mousa, "Heart disease detection using machine learning majority voting ensemble method," in *2019 2nd international conference on new trends in computing sciences (ictcs)*, 2019, pp. 1–6.

[403] M. A. I. Neloy, N. Nahar, M. S. Hossain, and K. Andersson, "A weighted average ensemble technique to predict heart disease," in *Proceedings of the Third International Conference on Trends in Computational and Cognitive Engineering: TCCE 2021*, 2022, pp. 17–29.

[404] A. Dogan and D. Birant, "A weighted majority voting ensemble approach for classification," in *2019 4th International Conference on Computer Science and Engineering (UBMK)*, 2019, pp. 1–6.

[405] S. I. Abba *et al.*, "Hybrid machine learning ensemble techniques for modeling dissolved oxygen concentration," *IEEE Access*, vol. 8, pp. 157218–157237, 2020.

[406] S. Mehta, P. Rana, S. Singh, A. Sharma, and P. Agarwal, "Ensemble learning approach for enhanced stock prediction," in *2019 twelfth international conference on contemporary computing (IC3)*, 2019, pp. 1–5.

[407] K. Sentz and S. Ferson, "Combination of evidence in Dempster-Shafer theory," 2002.

[408] Z. Mian *et al.*, "A literature review of fault diagnosis based on ensemble learning," *Eng. Appl. Artif. Intell.*, vol. 127, p. 107357, 2024.

[409] A. K. Asri *et al.*, "A machine learning-based ensemble model for estimating diurnal variations of nitrogen oxide concentrations in Taiwan," *Sci. Total Environ.*, vol. 916, p. 170209, 2024.

[410] A. K. Asri *et al.*, "What is the spatiotemporal pattern of benzene concentration spread over susceptible area surrounding the Hartman Park community, Houston, Texas?," *J. Hazard. Mater.*, p. 134666, 2024.

[411] C. M. Hansen, "The three dimensional solubility parameter," *Danish Tech. Copenhagen*, vol. 14, 1967.

[412] C. M. Hansen, *Hansen solubility parameters: a user's handbook*. CRC press, 2007.

[413] S. Bapat, S. O. Kilian, H. Wiggers, and D. Segets, "Towards a framework for evaluating and reporting Hansen solubility parameters: Applications to particle dispersions," *Nanoscale Adv.*, vol. 3, no. 15, pp. 4400–4410, 2021.

[414] R. Han *et al.*, "Predicting physical stability of solid dispersions by machine learning techniques," *J. Control. Release*, vol. 311, pp. 16–25, 2019.

[415] T. V. M. Sreekanth, S. Ramanaiah, K. D. Lee, and K. S. Reddy, "Hansen solubility parameters in the analysis of solvent--solvent interactions by inverse gas chromatography," *J. Macromol. Sci. Part B*, vol. 51, no. 6, pp. 1256–1266, 2012.

[416] S. Venkatram, C. Kim, A. Chandrasekaran, and R. Ramprasad, "Critical assessment of the Hildebrand and Hansen solubility parameters for polymers," *J. Chem. Inf. Model.*, vol. 59, no. 10, pp. 4188–4194, 2019.

[417] W. C. O. Ribeiro, P. F. M. Martinez, and V. Lobosco, "Solubility parameters analysis of Eucalyptus urograndis kraft lignin," *BioResources*, vol. 15, no. 4, p. 8577, 2020.

[418] E. Stefanis and C. Panayiotou, "Prediction of Hansen solubility parameters with a new group-contribution method," *Int. J. Thermophys.*, vol. 29, pp. 568–585, 2008.

[419] Y. S. Sistla, L. Jain, and A. Khanna, "Validation and prediction of solubility parameters of ionic liquids for CO2 capture," *Sep. Purif. Technol.*, vol. 97, pp. 51–64, 2012.

[420] F. Gharagheizi and M. Torabi Angaji, "A new improved method for estimating Hansen Solubility Parameters of polymers," *J. Macromol. Sci. Part B Phys.*, vol. 45, no. 2, pp. 285–290, 2006.

[421] T. Tamura and H. Yamamoto, "Calculation of Hansen Solubility Parameters Based on Solvatochromic Dye," 2019.

[422] G. Zhao, H. Ni, L. Jia, S. Ren, and G. Fang, "Quantitative analysis of relationship between Hansen solubility parameters and properties of alkali lignin/acrylonitrile--butadiene--styrene blends," *ACS omega*, vol. 3, no. 8, pp. 9722–9728, 2018.

[423] J. Ruwoldt, M. Tanase-Opedal, and K. Syverud, "Ultraviolet Spectrophotometry of Lignin Revisited: Exploring Solvents with Low Harmfulness, Lignin Purity, Hansen Solubility

Parameter, and Determination of Phenolic Hydroxyl Groups," *ACS omega*, vol. 7, no. 50, pp. 46371–46383, 2022.

[424] M. Mohan *et al.*, "Prediction of solubility parameters of lignin and ionic liquids using multi-resolution simulation approaches," *Green Chem.*, vol. 24, no. 3, pp. 1165–1176, 2022.

[425] M. Przybyłek, T. Jeliński, P. Cysewski, and others, "Application of multivariate adaptive regression splines (MARSplines) for predicting hansen solubility parameters based on 1D and 2D molecular descriptors computed from SMILES string," *J. Chem.*, vol. 2019, 2019.

[426] N. E. Jackson, M. A. Webb, and J. J. de Pablo, "Recent advances in machine learning towards multiscale soft materials design," *Curr. Opin. Chem. Eng.*, vol. 23, pp. 106–114, 2019.

[427] K. C. Leonard, F. Hasan, H. F. Sneddon, and F. You, "Can artificial intelligence and machine learning be used to accelerate sustainable chemistry and engineering?," *ACS Sustainable Chemistry \& Engineering*, vol. 9, no. 18. ACS Publications, pp. 6126–6129, 2021.

[428] G. Jarvas, C. Quellet, and A. Dallos, "Estimation of Hansen solubility parameters using multivariate nonlinear QSPR modeling with COSMO screening charge density moments," *Fluid Phase Equilib.*, vol. 309, no. 1, pp. 8–14, 2011.

[429] A. Chandrasekaran, C. Kim, S. Venkatram, and R. Ramprasad, "A deep learning solvent-selection paradigm powered by a massive solvent/nonsolvent database for polymers," *Macromolecules*, vol. 53, no. 12, pp. 4764–4769, 2020.

[430] J. D. Perea *et al.*, "Combined computational approach based on density functional theory and artificial neural networks for predicting the solubility parameters of fullerenes," *J. Phys. Chem. B*, vol. 120, no. 19, pp. 4431–4438, 2016.

[431] B. Sanchez-Lengeling, L. M. Roch, J. D. Perea, S. Langner, C. J. Brabec, and A. Aspuru-Guzik, "A Bayesian approach to predict solubility parameters," *Adv. Theory Simulations*, vol. 2, no. 1, p. 1800069, 2019.

[432] D. Obradović, F. Andrić, M. Zlatović, and D. Agbaba, "Modeling of Hansen's solubility parameters of aripiprazole, ziprasidone, and their impurities: A nonparametric comparison of models for prediction of drug absorption sites," *J. Chemom.*, vol. 32, no. 4, p. e2996,

2018.

[433] F. Delbecq, G. Adenier, Y. Ogue, and T. Kawai, “Gelation properties of various long chain amidoamines: Prediction of solvent gelation via machine learning using Hansen solubility parameters,” *J. Mol. Liq.*, vol. 303, p. 112587, 2020.

[434] F. Rexhepi, M. Woolever, J. Nabity, and S. Banerjee, “Metal oxide solvation with ionic liquids: A solubility parameter analysis,” *J. Mol. Liq.*, p. 122314, 2023.

[435] A. S. Alshehri, R. Gani, and F. You, “Deep learning and knowledge-based methods for computer-aided molecular design—toward a unified approach: State-of-the-art and future directions,” *Comput. \& Chem. Eng.*, vol. 141, p. 107005, 2020.

[436] US-Environmental Protection Agency, “SMILES Tutorial.” https://archive.epa.gov/med/med_archive_03/web/html/smiles.html (accessed Aug. 11, 2023).

[437] D. Fan *et al.*, “Deep learning model based on Bayesian optimization for predicting the infinite dilution activity coefficients of ionic liquid-solute systems,” *Eng. Appl. Artif. Intell.*, vol. 126, p. 107127, 2023.

[438] X. Chen and Q. Qian, “subGE: Enhancing the subgraph representation of molecular compounds structure--activity relationship discovery,” *Eng. Appl. Artif. Intell.*, vol. 119, p. 105727, 2023.

[439] K. Kobayashi and S. B. Alam, “Explainable, interpretable, and trustworthy AI for an intelligent digital twin: A case study on remaining useful life,” *Eng. Appl. Artif. Intell.*, vol. 129, p. 107620, 2024.

[440] O. M. Abdeldayem *et al.*, “Viral outbreaks detection and surveillance using wastewater-based epidemiology, viral air sampling, and machine learning techniques: A comprehensive review and outlook,” *Sci. Total Environ.*, vol. 803, p. 149834, 2022.

[441] S. C. Madeira and A. L. Oliveira, “Biclustering algorithms for biological data analysis: a survey,” *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 1, no. 1, pp. 24–45, 2004.

[442] G. Kalna, J. Keith Vass, and D. J. Higham, “Multidimensional partitioning and bi-partitioning: Analysis and application to gene expression data sets,” *Int. J. Comput. Math.*,

vol. 85, no. 3–4, pp. 475–485, 2008.

[443] A. A. Farhan, J. Lu, J. Bi, A. Russell, B. Wang, and A. Bamis, "Multi-view bi-clustering to identify smartphone sensing features indicative of depression," in *2016 IEEE first international conference on connected health: applications, systems and engineering technologies (CHASE)*, 2016, pp. 264–273.

[444] H. Liu and H. Motoda, "Computational methods of feature selection," *Chapman \&*, 1998.

[445] J. G. Dy, "Unsupervised feature selection," in *Computational methods of feature selection*, Chapman and Hall/CRC, 2007, pp. 35–56.

[446] A. Prelić *et al.*, "A systematic comparison and evaluation of biclustering methods for gene expression data," *Bioinformatics*, vol. 22, no. 9, pp. 1122–1129, 2006.

[447] Y. Kluger, R. Basri, J. T. Chang, and M. Gerstein, "Spectral biclustering of microarray data: coclustering genes and conditions," *Genome Res.*, vol. 13, no. 4, pp. 703–716, 2003.

[448] R. Henriques and S. C. Madeira, "Biclustering with flexible plaid models to unravel interactions between biological processes," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 12, no. 4, pp. 738–752, 2015.

[449] J. Gu and J. S. Liu, "Bayesian biclustering of gene expression data," *BMC Genomics*, vol. 9, no. 1, pp. 1–10, 2008.

[450] E. Meeds and S. Roweis, "Nonparametric bayesian biclustering," 2007.

[451] M. Lee, H. Shen, J. Z. Huang, and J. S. Marron, "Biclustering via sparse singular value decomposition," *Biometrics*, vol. 66, no. 4, pp. 1087–1095, 2010.

[452] Y. Li and A. Ngom, "The non-negative matrix factorization toolbox for biological data mining," *Source Code Biol. Med.*, vol. 8, no. 1, pp. 1–15, 2013.

[453] P. Carmona-Saez, R. D. Pascual-Marqui, F. Tirado, J. M. Carazo, and A. Pascual-Montano, "Biclustering of gene expression data by non-smooth non-negative matrix factorization," *BMC Bioinformatics*, vol. 7, pp. 1–18, 2006.

[454] "Non-Negative Matrix Factorization." https://www.geeksforgeeks.org/non-negative-matrix-factorization/ (accessed May 23, 2023).

[455] D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” *Adv. Neural Inf. Process. Syst.*, vol. 13, 2000.

[456] Y.-X. Wang and Y.-J. Zhang, “Nonnegative matrix factorization: A comprehensive review,” *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 6, pp. 1336–1353, 2012.

[457] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[458] L. McInnes, J. Healy, and J. Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” *arXiv Prepr. arXiv1802.03426*, 2018.

[459] B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley, “Uniform manifold approximation and projection (UMAP) and its variants: Tutorial and survey,” *arXiv Prepr. arXiv2109.02508*, 2021.

[460] Y. Wang, X. Gao, X. Ru, P. Sun, and J. Wang, “Using feature selection and Bayesian network identify cancer subtypes based on proteomic data,” *J. Proteomics*, vol. 280, p. 104895, 2023.

[461] P. G. Poličar, M. Stražar, and B. Zupan, “openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding,” *BioRxiv*, p. 731877, 2019.

[462] K. P. Sinaga and M.-S. Yang, “Unsupervised K-means clustering algorithm,” *IEEE access*, vol. 8, pp. 80716–80727, 2020.

[463] G. T. Reddy *et al.*, “Analysis of dimensionality reduction techniques on big data,” *Ieee Access*, vol. 8, pp. 54776–54788, 2020.

[464] B. M. S. Hasan and A. M. Abdulazeez, “A review of principal component analysis algorithm for dimensionality reduction,” *J. Soft Comput. Data Min.*, vol. 2, no. 1, pp. 20–30, 2021.

[465] Y. Wang, H. Huang, C. Rudin, and Y. Shaposhnik, “Understanding how dimension reduction tools work: an empirical approach to deciphering t-SNE, UMAP, TriMAP, and PaCMAP for data visualization,” *J. Mach. Learn. Res.*, vol. 22, no. 1, pp. 9129–9201, 2021.

[466] A. Gisbrecht, A. Schulz, and B. Hammer, “Parametric nonlinear dimensionality reduction using kernel t-SNE,” *Neurocomputing*, vol. 147, pp. 71–82, 2015.

[467] F. Anowar, S. Sadaoui, and B. Selim, “Conceptual and empirical comparison of

dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne),” *Comput. Sci. Rev.*, vol. 40, p. 100378, 2021.

[468] I. Choi, W. Koh, B. Koo, and W. C. Kim, “Network-based exploratory data analysis and explainable three-stage deep clustering for financial customer profiling,” *Eng. Appl. Artif. Intell.*, vol. 128, p. 107378, 2024.

[469] L. der Maaten and G. Hinton, “Visualizing data using t-SNE.,” *J. Mach. Learn. Res.*, vol. 9, no. 11, 2008.

[470] P. Hu, Z. Jiao, Z. Zhang, and Q. Wang, “Development of solubility prediction models with ensemble learning,” *Ind. \& Eng. Chem. Res.*, vol. 60, no. 30, pp. 11627–11635, 2021.

[471] E. G. Al-Sakkari *et al.*, “Machine learning-assisted selection of adsorption-based carbon dioxide capture materials,” *J. Environ. Chem. Eng.*, p. 110732, 2023.

[472] S. Abbott and C. M. Hansen, *Hansen solubility parameters in practice*. Hansen-Solubility, 2008.

[473] “Hansen Solubility Parameters in Practice (Official Web Page).” https://www.hansen-solubility.com/HSPiP/ (accessed Aug. 07, 2023).

[474] M. de los R\’\ios and E. Hernández Ramos, “Determination of the Hansen solubility parameters and the Hansen sphere radius with the aid of the solver add-in of Microsoft Excel,” *SN Appl. Sci.*, vol. 2, pp. 1–7, 2020.

[475] A. De La Peña-Gil, J. F. Toro-Vazquez, and M. A. Rogers, “Simplifying Hansen solubility parameters for complex edible fats and oils,” *Food Biophys.*, vol. 11, pp. 283–291, 2016.

[476] D. Schieppati *et al.*, “Chemical and biological delignification of biomass: a review,” *Ind. \& Eng. Chem. Res.*, vol. 62, no. 33, pp. 12757–12794, 2023.

[477] A. P. Bento *et al.*, “An open source chemical structure curation pipeline using RDKit,” *J. Cheminform.*, vol. 12, pp. 1–16, 2020.

[478] A. Duval, F. Vilaplana, C. Crestini, and M. Lawoko, “Solvent screening for the fractionation of industrial kraft lignin,” *Holzforschung*, vol. 70, no. 1, pp. 11–20, 2016.

[479] L. P. Novo and A. A. S. Curvelo, “Hansen solubility parameters: a tool for solvent selection for organosolv delignification,” *Ind. \& Eng. Chem. Res.*, vol. 58, no. 31, pp. 14520–14527,

2019.

[480] L. L. Williams, “10 Determination of Hansen Solubility Parameter Values for Carbon Dioxide,” 2007.

[481] L. L. Williams, J. B. Rubin, and H. W. Edwards, “Calculation of Hansen solubility parameter values for a range of pressure and temperature conditions, including the supercritical fluid region,” *Ind. \& Eng. Chem. Res.*, vol. 43, no. 16, pp. 4967–4972, 2004.

[482] B. Zhao *et al.*, “Lignin-based porous supraparticles for carbon capture,” *ACS Nano*, vol. 15, no. 4, pp. 6774–6786, 2021.

[483] D. Barker-Rothschild *et al.*, “Lignin-based porous carbon adsorbents for CO 2 capture,” *Chem. Soc. Rev.*, 2024.

[484] E. G. Al-Sakkari *et al.*, “New alginate-based interpenetrating polymer networks for water treatment: A response surface methodology based optimization study,” *Int. J. Biol. Macromol.*, Mar. 2020, doi: 10.1016/j.ijbiomac.2020.03.220.

[485] N. Supanchaiyamat, K. Jetsrisuparb, J. T. N. Knijnenburg, D. C. W. Tsang, and A. J. Hunt, “Lignin materials for adsorption: Current trend, perspectives and opportunities,” *Bioresour. Technol.*, vol. 272, pp. 570–581, 2019.

[486] P. J. M. Carrott, M. M. L. R. Carrott, and others, “Lignin--from natural adsorbent to activated carbon: a review,” *Bioresour. Technol.*, vol. 98, no. 12, pp. 2301–2312, 2007.

[487] K. Fu, Q. Yue, B. Gao, Y. Sun, and L. Zhu, “Preparation, characterization and application of lignin-based activated carbon from black liquor lignin by steam activation,” *Chem. Eng. J.*, vol. 228, pp. 1074–1082, 2013.

[488] M. M. Naeem, E. G. Al-Sakkari, D. C. Boffito, M. A. Gadalla, and F. H. Ashour, “One-pot conversion of highly acidic waste cooking oil into biodiesel over a novel bio-based bi-functional catalyst,” *Fuel*, vol. 283, Jan. 2021, doi: 10.1016/j.fuel.2020.118914.

[489] M. M. Naeem, E. G. Al-Sakkari, D. C. Boffito, E. R. Rene, M. A. Gadalla, and F. H. Ashour, “Single-stage waste oil conversion into biodiesel via sonication over bio-based bifunctional catalyst: optimization, preliminary techno-economic and environmental analysis,” *Fuel*, vol. 341, p. 127587, 2023.

[490] S. H. Dhawane, E. G. Al-Sakkari, and G. Halder, “Kinetic Modelling of Heterogeneous Methanolysis Catalysed by Iron Induced on Microporous Carbon Supported Catalyst,” *Catal. Letters*, Jul. 2019, doi: 10.1007/s10562-019-02905-5.

[491] H. Jeong, S. Kim, M. Gil, S. Song, T.-H. Kim, and K. J. Lee, “Preparation of poly-1-butene nanofiber mat and its application as shutdown layer of next generation lithium ion battery,” *Polymers (Basel).*, vol. 12, no. 10, p. 2267, 2020.

[492] R. Subrahmanyam, P. Gurikov, P. Dieringer, M. Sun, and I. Smirnova, “On the road to biopolymer aerogels—Dealing with the solvent,” *Gels*, vol. 1, no. 2, pp. 291–313, 2015.

[493] M. Li *et al.*, “New parameter derived from the hansen solubility parameter used to evaluate the solubility of asphaltene in solvent,” *ACS omega*, vol. 7, no. 16, pp. 13801–13807, 2022.

[494] V. D. Hähnke, S. Kim, and E. E. Bolton, “PubChem chemical structure standardization,” *J. Cheminform.*, vol. 10, pp. 1–40, 2018.

[495] J. G. Ethier *et al.*, “Predicting phase behavior of linear polymers in solution using machine learning,” *Macromolecules*, vol. 55, no. 7, pp. 2691–2702, 2022.

[496] V. M. Nagulapati, H. M. R. U. Rehman, J. Haider, M. A. Qyyum, G. S. Choi, and H. Lim, “Hybrid machine learning-based model for solubilities prediction of various gases in deep eutectic solvent for rigorous process design of hydrogen purification,” *Sep. Purif. Technol.*, vol. 298, p. 121651, 2022.

[497] H. Liu *et al.*, “A generic machine learning model for CO2 equilibrium solubility into blended amine solutions,” *Sep. Purif. Technol.*, vol. 334, p. 126100, 2024.

[498] Y.-D. Hsiao and C.-T. Chang, “Joint incremental learning network for flexible modeling of carbon dioxide solubility in aqueous mixtures of amines,” *Sep. Purif. Technol.*, vol. 330, p. 125299, 2024.

[499] M.-H. Lee, “Interpretable machine-learning for predicting power conversion efficiency of non-halogenated green solvent-processed organic solar cells based on Hansen solubility parameters and molecular weights of polymers,” *Sol. Energy*, vol. 261, pp. 7–13, 2023.

[500] S. E. Jerng, Y. J. Park, and J. Li, “Machine learning for CO2 capture and conversion: A review,” *Energy AI*, vol. 16, p. 100361, 2024, doi:

https://doi.org/10.1016/j.egyai.2024.100361.

[501] T. Elliott *et al.*, “Synergetic effects on the capture and release of CO2 using guanidine and amidine superbases,” *RSC Sustain.*, 2024.

[502] F. P. Byrne *et al.*, “Tools and techniques for solvent selection: green solvent selection guides,” *Sustain. Chem. Process.*, vol. 4, pp. 1–24, 2016.

[503] J. Hu, Q. Zhang, and D.-J. Lee, “Kraft lignin biorefinery: A perspective,” *Bioresour. Technol.*, vol. 247, pp. 1181–1183, 2018.

[504] Y. Cao *et al.*, “Advances in lignin valorization towards bio-based chemicals and fuels: Lignin biorefinery,” *Bioresour. Technol.*, vol. 291, p. 121878, 2019.

[505] J. G. Linger *et al.*, “Lignin valorization through integrated biological funneling and chemical catalysis,” *Proc. Natl. Acad. Sci.*, vol. 111, no. 33, pp. 12013–12018, 2014.

[506] D. Kun and B. Pukánszky, “Polymer/lignin blends: Interactions, properties, applications,” *Eur. Polym. J.*, vol. 93, pp. 618–641, 2017.

[507] J. V Vermaas, M. F. Crowley, and G. T. Beckham, “Molecular lignin solubility and structure in organic solvents,” *ACS Sustain. Chem. \& Eng.*, vol. 8, no. 48, pp. 17839–17850, 2020.

[508] A. Dastpak, T. V Lourençon, M. Balakshin, S. F. Hashmi, M. Lundström, and B. P. Wilson, “Solubility study of lignin in industrial organic solvents and investigation of electrochemical properties of spray-coated solutions,” *Ind. Crops Prod.*, vol. 148, p. 112310, 2020.

[509] R. Saadan, C. Hachimi Alaoui, A. Ihammi, M. Chigr, and A. Fatimi, “A Brief Overview of Lignin Extraction and Isolation Processes: From Lignocellulosic Biomass to Added-Value Biomaterials,” *Environ. Earth Sci. Proc.*, vol. 31, no. 1, p. 3, 2024.

[510] R. Rinaldi *et al.*, “Paving the way for lignin valorisation: recent advances in bioengineering, biorefining and catalysis,” *Angew. Chemie Int. Ed.*, vol. 55, no. 29, pp. 8164–8215, 2016.

[511] K. H. Kim and C. G. Yoo, “Challenges and perspective of recent biomass pretreatment solvents,” *Front. Chem. Eng.*, vol. 3, p. 785709, 2021.

[512] L. Soh and M. J. Eckelman, “Green solvents in biomass processing,” *ACS Sustain. Chem. \& Eng.*, vol. 4, no. 11, pp. 5821–5837, 2016.

[513] C. Libretti, L. S. Correa, and M. A. R. Meier, “From waste to resource: advancements in sustainable lignin modification,” *Green Chem.*, 2024.

[514] R. Hardian, Z. Liang, X. Zhang, and G. Szekely, “Artificial intelligence: The silver bullet for sustainable materials development,” *Green Chem.*, vol. 22, no. 21, pp. 7521–7528, 2020.

[515] Y. Huang, G. Cui, H. Wang, Z. Li, and J. Wang, “Tuning ionic liquids with imide-based anions for highly efficient CO2 capture through enhanced cooperations,” *J. CO2 Util.*, vol. 28, pp. 299–305, 2018.

[516] Q. Liu *et al.*, “Novel deep eutectic solvents with different functional groups towards highly efficient dissolution of lignin,” *Green Chem.*, vol. 21, no. 19, pp. 5291–5297, 2019.

[517] H. Strübing, P. G. Karamertzanis, E. N. Pistikopoulos, A. Galindo, and C. S. Adjiman, “Solvent design for a Menschutkin reaction by using CAMD and DFT calculations,” in *Computer Aided Chemical Engineering*, vol. 28, Elsevier, 2010, pp. 1291–1296.

[518] J. L. McDonagh *et al.*, “Machine guided discovery of novel carbon capture solvents,” *arXiv Prepr. arXiv2303.14223*, 2023.

[519] V. Ponnuchamy, O. Gordobil, R. H. Diaz, A. Sandak, and J. Sandak, “Fractionation of lignin using organic solvents: A combined experimental and theoretical study,” *Int. J. Biol. Macromol.*, vol. 168, pp. 792–805, 2021.

[520] M. Islam, I. M. Khan, S. Shakya, and N. Alam, “Design, synthesis, characterizing and DFT calculations of a binary CT complex co-crystal of bioactive moieties in different polar solvents to investigate its pharmacological activity,” *J. Biomol. Struct. Dyn.*, vol. 41, no. 20, pp. 10813–10829, 2023.

[521] P. Hou *et al.*, “Rational design of hydrogen-donor solvents for direct coal liquefaction,” *Energy \& fuels*, vol. 32, no. 4, pp. 4715–4723, 2018.

[522] L. König-Mattern *et al.*, “High-throughput computational solvent screening for lignocellulosic biomass processing,” *Chem. Eng. J.*, vol. 452, p. 139476, 2023.

[523] M. Zavrel, D. Bross, M. Funke, J. Büchs, and A. C. Spiess, “High-throughput screening for ionic liquids dissolving (ligno-) cellulose,” *Bioresour. Technol.*, vol. 100, no. 9, pp. 2580–2587, 2009.

[524] F. Porcheron *et al.*, “High Throughput Screening of amine thermodynamic properties applied to post-combustion CO2 capture process evaluation,” *Energy Procedia*, vol. 4, pp. 15–22, 2011.

[525] F. Porcheron, A. Gibert, P. Mougin, and A. Wender, “High throughput screening of CO2 solubility in aqueous monoamine solutions,” *Environ. Sci. \& Technol.*, vol. 45, no. 6, pp. 2486–2492, 2011.

[526] D. Ongari, L. Talirz, and B. Smit, “Too many materials and too many applications: an experimental problem waiting for a computational solution,” *ACS Cent. Sci.*, vol. 6, no. 11, pp. 1890–1900, 2020.

[527] L. M. Mayr and D. Bojanic, “Novel trends in high-throughput screening,” *Curr. Opin. Pharmacol.*, vol. 9, no. 5, pp. 580–588, 2009.

[528] L. M. Mayr and P. Fuerst, “The future of high-throughput screening,” *SLAS Discov.*, vol. 13, no. 6, pp. 443–448, 2008.

[529] D. C. Elton, Z. Boukouvalas, M. D. Fuge, and P. W. Chung, “Deep learning for molecular design—a review of the state of the art,” *Mol. Syst. Des. \& Eng.*, vol. 4, no. 4, pp. 828–849, 2019.

[530] D. Rothchild, A. Tamkin, J. Yu, U. Misra, and J. Gonzalez, “C5t5: Controllable generation of organic molecules with transformers,” *arXiv Prepr. arXiv2108.10307*, 2021.

[531] W. P. Walters and R. Barzilay, “Applications of deep learning in molecule generation and molecular property prediction,” *Acc. Chem. Res.*, vol. 54, no. 2, pp. 263–270, 2020.

[532] O. Zhang *et al.*, “Deep lead optimization: leveraging generative AI for structural modification,” *J. Am. Chem. Soc.*, vol. 146, no. 46, pp. 31357–31370, 2024.

[533] S. Wang, Y. Guo, Y. Wang, H. Sun, and J. Huang, “Smiles-bert: large scale unsupervised pre-training for molecular property prediction,” in *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, 2019, pp. 429–436.

[534] J. Pang, A. W. R. Pine, and A. Sulemana, “Using natural language processing (NLP)-inspired molecular embedding approach to predict Hansen solubility parameters,” *Digit.*

Discov., vol. 3, no. 1, pp. 145–154, 2024.

[535] A. Sultan, J. Sieg, M. Mathea, and A. Volkamer, "Transformers for molecular property prediction: Lessons learned from the past five years," *arXiv Prepr. arXiv2404.03969*, 2024.

[536] K. M. Jablonka, P. Schwaller, A. Ortega-Guerrero, and B. Smit, "Leveraging large language models for predictive chemistry," *Nat. Mach. Intell.*, pp. 1–9, 2024.

[537] A. R. Flanagan, D. Dalal, and F. G. Glavin, "Exploring Generative Artificial Intelligence and Data Augmentation Techniques for Spectroscopy Analysis," *Chem. Rev.*, 2025.

[538] P. Ertl, R. Lewis, E. Martin, and V. Polyakov, "In silico generation of novel, drug-like chemical matter using the LSTM neural network," *arXiv Prepr. arXiv1712.07449*, 2017.

[539] H. Yu, R. Wang, J. Qiao, and L. Wei, "Multi-CGAN: Deep Generative Model-Based Multiproperty Antimicrobial Peptide Design," *J. Chem. Inf. Model.*, vol. 64, no. 1, pp. 316–326, 2023.

[540] V. Bagal, R. Aggarwal, P. K. Vinod, and U. D. Priyakumar, "MolGPT: molecular generation using a transformer-decoder model," *J. Chem. Inf. Model.*, vol. 62, no. 9, pp. 2064–2076, 2021.

[541] A. D. White *et al.*, "Assessment of chemistry knowledge in large language models that generate code," *Digit. Discov.*, vol. 2, no. 2, pp. 368–376, 2023.

[542] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," *Adv. Neural Inf. Process. Syst.*, vol. 36, 2024.

[543] E. G. Al-Sakkari *et al.*, "Ensemble machine learning to accelerate industrial decarbonization: Prediction of Hansen solubility parameters for streamlined chemical solvent selection," *Digit. Chem. Eng.*, vol. 14, p. 100207, 2025.

[544] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[545] Y. Vasiliev, *Natural language processing with Python and spaCy: A practical introduction*. No Starch Press, 2020.

[546] M. Grohe, “word2vec, node2vec, graph2vec, x2vec: Towards a theory of vector embeddings of structured data,” in *Proceedings of the 39th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, 2020, pp. 1–16.

[547] P. Chen, L. He, L. Fu, L. Fan, E. F. Chang, and Y. Li, “Do self-supervised speech and language models extract similar representations as human brain?,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 2225–2229.

[548] X. Liu *et al.*, “Self-supervised learning: Generative or contrastive,” *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857–876, 2021.

[549] L. Ericsson, H. Gouk, C. C. Loy, and T. M. Hospedales, “Self-supervised representation learning: Introduction, advances, and challenges,” *IEEE Signal Process. Mag.*, vol. 39, no. 3, pp. 42–62, 2022.

[550] Hugging Face, “OpenAI GPT2.” https://huggingface.co/docs/transformers/model_doc/gpt2#openai-gpt2.

[551] A. Radford *et al.*, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[552] S. Adilov, “Generative pre-training from molecules,” 2021.

[553] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein generative adversarial networks,” in *International conference on machine learning*, 2017, pp. 214–223.

[554] Z. Li, P. Xia, R. Tao, H. Niu, and B. Li, “A new perspective on stabilizing GANs training: Direct adversarial training,” *IEEE Trans. Emerg. Top. Comput. Intell.*, vol. 7, no. 1, pp. 178–189, 2022.

[555] Y. Kossale, M. Airaj, and A. Darouichi, “Mode collapse in generative adversarial networks: An overview,” in *2022 8th International Conference on Optimization and Applications (ICOA)*, 2022, pp. 1–6.

[556] A. Sajeeda and B. M. M. Hossain, “Exploring generative adversarial networks and adversarial training,” *Int. J. Cogn. Comput. Eng.*, vol. 3, pp. 78–89, 2022.

[557] J. Adler and S. Lunz, “Banach wasserstein gan,” *Adv. Neural Inf. Process. Syst.*, vol. 31,

2018.

[558] A. Kensert, “WGAN-GP with R-GCN for the generation of small molecular graphs.” https://keras.io/examples/generative/wgan-graphs/.

[559] V. Basu, “Drug Molecule Generation with VAE,” 2022. https://keras.io/examples/generative/molecule_generation/#drug-molecule-generation-with-vae.

[560] C. Doersch, “Tutorial on variational autoencoders,” *arXiv Prepr. arXiv1606.05908*, 2016.

[561] J. Rocca, “Understanding Variational Autoencoders (VAEs),” 2019. https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73.

[562] TensorFlow, “Convolutional Variational Autoencoder.” https://www.tensorflow.org/tutorials/generative/cvae.

[563] M. Leon, Y. Perezhohin, F. Peres, A. Popovič, and M. Castelli, “Comparing SMILES and SELFIES tokenization for enhanced chemical language modeling,” *Sci. Rep.*, vol. 14, no. 1, p. 25016, 2024.

[564] X. Li and D. Fourches, “SMILES pair encoding: a data-driven substructure tokenization algorithm for deep learning,” *J. Chem. Inf. Model.*, vol. 61, no. 4, pp. 1560–1569, 2021.

[565] L. Zhao, Q. Wang, and K. Ma, “Solubility parameter of ionic liquids: a comparative study of inverse gas chromatography and Hansen solubility sphere,” *ACS Sustain. Chem. \& Eng.*, vol. 7, no. 12, pp. 10544–10551, 2019.

[566] Y. Agata and H. Yamamoto, “Determination of Hansen solubility parameters of ionic liquids using double-sphere type of Hansen solubility sphere method,” *Chem. Phys.*, vol. 513, pp. 165–173, 2018.

[567] G. Landrum, “Rdkit documentation,” *Release*, vol. 1, no. 1–79, p. 4, 2013.

[568] H. E. Pence and A. Williams, “ChemSpider: an online chemical information resource.” ACS Publications, 2010.

[569] U. V Ucak, I. Ashyrmamatov, J. Ko, and J. Lee, “Retrosynthetic reaction pathway prediction through neural machine translation of atomic environments,” *Nat. Commun.*, vol. 13, no. 1,

p. 1186, 2022.

[570] K. F. Jensen, “Automated System for Knowledge-Based Continuous Organic Synthesis (ASKCOS),” 2021.

[571] K. Sankaranarayanan and K. F. Jensen, “Computer-assisted multistep chemoenzymatic retrosynthesis using a chemical synthesis planner,” *Chem. Sci.*, vol. 14, no. 23, pp. 6467–6475, 2023.

[572] S. Genheden, A. Thakkar, V. Chadimová, J.-L. Reymond, O. Engkvist, and E. Bjerrum, “AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning,” *J. Cheminform.*, vol. 12, no. 1, p. 70, 2020.

[573] K. Satyanarayana and P. G. Rao, “Improved equation to estimate flash points of organic compounds,” *J. Hazard. Mater.*, vol. 32, no. 1, pp. 81–85, 1992.

[574] S. M. Santos, D. C. Nascimento, M. C. Costa, A. M. B. Neto, and L. V Fregolente, “Flash point prediction: Reviewing empirical models for hydrocarbons, petroleum fraction, biodiesel, and blends,” *Fuel*, vol. 263, p. 116375, 2020.

[575] M. Hristova and S. Tchaoushev, “Calculation of flash points and flammability limits of substances and mixtures,” *J. Univ. Chem. Technol. Metall.*, vol. 41, no. 3, pp. 291–296, 2006.

[576] A. Alibakhshi, H. Mirshahvalad, and S. Alibakhshi, “Prediction of flash points of pure organic compounds: Evaluation of the DIPPR database,” *Process Saf. Environ. Prot.*, vol. 105, pp. 127–133, 2017.

[577] C. Wu *et al.*, “A comprehensive review of carbon capture science and technologies,” *Carbon Capture Sci. \& Technol.*, vol. 11, p. 100178, 2024.

[578] C. Song, Y. Kitamura, and S. Li, “Optimization of a novel cryogenic CO2 capture process by response surface methodology (RSM),” *J. Taiwan Inst. Chem. Eng.*, vol. 45, no. 4, pp. 1666–1676, 2014.

[579] P. Daoutidis *et al.*, “Machine learning in process systems engineering: Challenges and opportunities,” *Comput. Chem. Eng.*, vol. 181, p. 108523, 2024, doi: https://doi.org/10.1016/j.compchemeng.2023.108523.

[580] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, “Deep reinforcement learning: A brief survey,” *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26–38, 2017.

[581] C. Yu *et al.*, “The surprising effectiveness of ppo in cooperative multi-agent games,” *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 24611–24624, 2022.

[582] T. P. Lillicrap *et al.*, “Continuous control with deep reinforcement learning,” *arXiv Prepr. arXiv1509.02971*, 2015.

[583] L. Stops, R. Leenhouts, Q. Gao, and A. M. Schweidtmann, “Flowsheet generation through hierarchical reinforcement learning and graph neural networks,” *AIChE J.*, vol. 69, no. 1, p. e17938, 2023.

[584] Q. Göttl, J. Pirnay, J. Burger, and D. G. Grimm, “Deep reinforcement learning uncovers processes for separating azeotropic mixtures without prior knowledge,” *arXiv Prepr. arXiv2310.06415*, 2023.

[585] Q. Gao, H. Yang, S. M. Shanbhag, and A. M. Schweidtmann, “Transfer learning for process design with reinforcement learning,” *arXiv Prepr. arXiv2302.03375*, 2023.

[586] L. I. Midgley, “Deep reinforcement learning for process synthesis,” *arXiv Prepr. arXiv2009.13265*, 2020.

[587] Q. Göttl, Y. Tönges, D. G. Grimm, and J. Burger, “Automated flowsheet synthesis using hierarchical reinforcement learning: proof of concept,” *Chemie Ing. Tech.*, vol. 93, no. 12, pp. 2010–2018, 2021.

[588] S. J. Plathottam, B. Richey, G. Curry, J. Cresko, and C. O. Iloeje, “Solvent extraction process design using deep reinforcement learning,” *J. Adv. Manuf. Process.*, vol. 3, no. 2, p. e10079, 2021.

[589] A. A. Khan and A. A. Lapkin, “Designing the process designer: Hierarchical reinforcement learning for optimisation-based process design,” *Chem. Eng. Process. Intensif.*, vol. 180, p. 108885, 2022.

[590] Q. Göttl, D. G. Grimm, and J. Burger, “Automated synthesis of steady-state continuous processes using reinforcement learning,” *Front. Chem. Sci. Eng.*, pp. 1–15, 2021.

[591] Q. Gao and A. M. Schweidtmann, “Deep reinforcement learning for process design: Review

and perspective," *arXiv Prepr. arXiv2308.07822*, 2023.

[592] E. G. Al-Sakkari, A. Ragab, M. Ali, H. Dagdougui, D. C. Boffito, and M. Amazouz, "Learn-to-design: reinforcement learning-assisted chemical process optimization," *Syst. \& Control Trans.*, vol. 3, pp. 245–252, 2024.

[593] M.-R. Mohammadi, A. Larestani, M. Schaffie, A. Hemmati-Sarapardeh, and M. Ranjbar, "Predictive modeling of CO2 solubility in piperazine aqueous solutions using boosting algorithms for carbon capture goals," *Sci. Rep.*, vol. 14, no. 1, p. 22112, 2024.

[594] P.-C. Chen, M.-W. Yang, C.-H. Wei, and S. Z. Lin, "Selection of blended amine for CO2 capture in a packed bed scrubber using the Taguchi method," *Int. J. Greenh. Gas Control*, vol. 45, pp. 245–252, 2016.

[595] F. Mufarrij, P. Navarri, and Y. Khojasteh, "Development and comparative eco-techno-economic analysis of two carbon capture and utilization pathways for polyethylene production," *J. Environ. Chem. Eng.*, p. 115920, 2025.

[596] H. Pashaei, A. Ghaemi, M. Nasiri, and B. Karami, "Experimental modeling and optimization of CO2 absorption into piperazine solutions using RSM-CCD methodology," *ACS omega*, vol. 5, no. 15, pp. 8432–8448, 2020.

[597] Y. H. Yan *et al.*, "Harnessing the power of machine learning for carbon capture, utilisation, and storage (CCUS)--a state-of-the-art review," *Energy \& Environ. Sci.*, 2021.

[598] W. G. Davies, S. Babamohammadi, Y. Yang, and S. M. Soltani, "The rise of the machines: A state-of-the-art technical review on process modelling and machine learning within hydrogen production with carbon capture," *Gas Sci. Eng.*, vol. 118, p. 205104, 2023.

[599] F. Hussin, S. A. N. M. Rahim, N. S. M. Hatta, M. K. Aroua, and S. A. Mazari, "A systematic review of machine learning approaches in carbon capture applications," *J. CO2 Util.*, vol. 71, p. 102474, 2023.

[600] M. Hosseinpour, M. J. Shojaei, M. Salimi, and M. Amidpour, "Machine learning in absorption-based post-combustion carbon capture systems: A state-of-the-art review," *Fuel*, vol. 353, p. 129265, 2023.

[601] D.-H. Oh, D. Adams, N. D. Vo, D. Q. Gbadago, C.-H. Lee, and M. Oh, "Actor-critic

reinforcement learning to estimate the optimal operating conditions of the hydrocracking process," *Comput. \& Chem. Eng.*, vol. 149, p. 107280, 2021.

[602] Z. Zhang, D.-N. Vo, T. B. H. Nguyen, J. Sun, and C.-H. Lee, "Advanced process integration and machine learning-based optimization to enhance techno-economic-environmental performance of CO2 capture and conversion to methanol," *Energy*, vol. 293, p. 130758, 2024.

[603] D.-N. Vo, Z. Zhang, and X. Yin, "Superstructure design, data-driven mixed-integer optimization, and guidelines for techno-economic-environmental enhancement of blended amine-based CO2 capture in natural gas combined cycle power plants," *J. Clean. Prod.*, vol. 491, p. 144757, 2025.

[604] N. D. Vo, D. H. Oh, S.-H. Hong, M. Oh, and C.-H. Lee, "Combined approach using mathematical modelling and artificial neural network for chemical industries: Steam methane reformer," *Appl. Energy*, vol. 255, p. 113809, 2019.

[605] N. D. Vo, D. H. Oh, J.-H. Kang, M. Oh, and C.-H. Lee, "Dynamic-model-based artificial neural network for H2 recovery and CO2 capture from hydrogen tail gas," *Appl. Energy*, vol. 273, p. 115263, 2020.

[606] H. Ali *et al.*, "Process Monitoring and Dynamic Fusion of Complex Industrial Systems: A Reconstruction-based Bayesian Framework," *Comput. \& Chem. Eng.*, p. 109352, 2025.

[607] H.-T. Oh, J. Kum, J. Park, N. D. Vo, J.-H. Kang, and C.-H. Lee, "Pre-combustion CO2 capture using amine-based absorption process for blue H2 production from steam methane reformer," *Energy Convers. Manag.*, vol. 262, p. 115632, 2022.

[608] M. Qi *et al.*, "Proposal and surrogate-based cost-optimal design of an innovative green ammonia and electricity co-production system via liquid air energy storage," *Appl. Energy*, vol. 314, p. 118965, 2022.

[609] M. Srilekha and A. Dutta, "Simulation-based optimization framework for design of a self-sustainable LNG regasification process recovering waste cold energy," *Therm. Sci. Eng. Prog.*, p. 103821, 2025.

[610] G. O. Cassol, G. V. K. Campos, D. M. Thomaz, B. D. O. Capron, and A. R. Secchi, "Reinforcement learning applied to process control: a van der Vusse reactor case study," in

Computer Aided Chemical Engineering, vol. 44, Elsevier, 2018, pp. 553–558.

[611] A. Khan and A. Lapkin, “Searching for optimal process routes: A reinforcement learning approach,” *Comput. \& Chem. Eng.*, vol. 141, p. 107027, 2020.

[612] P. Petsagkourakis, I. O. Sandoval, E. Bradford, D. Zhang, and E. A. del Rio-Chanona, “Reinforcement learning for batch-to-batch bioprocess optimisation,” in *Computer Aided Chemical Engineering*, vol. 46, Elsevier, 2019, pp. 919–924.

[613] J. Shin, T. A. Badgwell, K.-H. Liu, and J. H. Lee, “Reinforcement learning--overview of recent progress and implications for process control,” *Comput. \& Chem. Eng.*, vol. 127, pp. 282–294, 2019.

[614] A. Abuelnasr, A. Ragab, M. Amer, B. Gosselin, and Y. Savaria, “Incremental reinforcement learning for multi-objective analog circuit design acceleration,” *Eng. Appl. Artif. Intell.*, vol. 129, p. 107426, 2024.

[615] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, “Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor,” in *International conference on machine learning*, 2018, pp. 1861–1870.

[616] T. Haarnoja *et al.*, “Soft actor-critic algorithms and applications,” *arXiv Prepr. arXiv1812.05905*, 2018.

[617] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, “Deterministic policy gradient algorithms,” in *International conference on machine learning*, 2014, pp. 387–395.

[618] E. H. Sumiea *et al.*, “Deep deterministic policy gradient algorithm: A systematic review,” *Heliyon*, 2024.

[619] A. Shojaie and E. B. Fox, “Granger causality: A review and recent advances,” *Annu. Rev. Stat. Its Appl.*, vol. 9, no. 1, pp. 289–319, 2022.

[620] B. Ganter and R. Wille, *Formal concept analysis: mathematical foundations*. Springer Nature, 2024.

[621] R. Belohlavek, “Introduction to formal concept analysis,” *Palacky Univ. Dep. Comput. Sci. Olomouc*, vol. 47, 2008.

[622] K. Sumangali and C. A. Kumar, “A comprehensive overview on the foundations of formal concept analysis,” *Knowl. Manag. \& E-Learning*, vol. 9, no. 4, p. 512, 2017.

[623] J. F. Hair Jr *et al.*, “An introduction to structural equation modeling,” *Partial least squares Struct. Equ. Model. using R a Workb.*, pp. 1–29, 2021.

[624] G. Cooper, “An overview of the representation and discovery of causal relationships using Bayesian networks,” *Comput. causation, Discov.*, pp. 4–62, 1999.

[625] D. Heckerman, “A Bayesian approach to learning causal networks,” *arXiv Prepr. arXiv1302.4958*, 2013.

[626] J. Pearl, “From Bayesian networks to causal networks,” in *Mathematical models for handling partial knowledge in artificial intelligence*, Springer, 1995, pp. 157–182.

[627] B. Ellis and W. H. Wong, “Learning causal Bayesian network structures from experimental data,” *J. Am. Stat. Assoc.*, vol. 103, no. 482, pp. 778–789, 2008.

[628] S. Mueller, A. Li, and J. Pearl, “Learning Causes of Effects and Individual Responses from Population Data and Bayesian Networks,” 2024.

[629] T. C. Williams, C. C. Bach, N. B. Matthiesen, T. B. Henriksen, and L. Gagliardi, “Directed acyclic graphs: a tool for causal studies in paediatrics,” *Pediatr. Res.*, vol. 84, no. 4, pp. 487–493, 2018.

[630] J. C. Digitale, J. N. Martin, and M. M. Glymour, “Tutorial on directed acyclic graphs,” *J. Clin. Epidemiol.*, vol. 142, pp. 264–267, 2022.

[631] S. L. Hoe, “Issues and procedures in adopting structural equation modelling technique,” *J. Quant. Methods*, vol. 3, no. 1, p. 76, 2008.

[632] M. S. Peters, K. D. Timmerhaus, R. E. West, K. Timmerhaus, and R. West, *Plant design and economics for chemical engineers*, vol. 4. McGraw-hill New York, 1968.

[633] E. G. Al-Sakkari *et al.*, “Comparative Technoeconomic Analysis of Using Waste and Virgin Cooking Oils for Biodiesel Production,” *Front. Energy Res.*, p. 278, 2020.

[634] I. Iliuta, F. Larachi, J. Shabanian, and K. Zanganeh, “Dynamic modeling and energy optimization of CO2 and H2O desorption in zeolite 13$\times$ adsorbent DAC units: Enhancing regeneration efficiency in internally joule-heated vacuum beds,” *Chem. Eng. J.*,

p. 172064, 2025.

[635] A. Sarvaramini, O. Gravel, and F. Larachi, “Torrefaction of ionic-liquid impregnated lignocellulosic biomass and its comparison to dry torrefaction,” *Fuel*, vol. 103, pp. 814–826, 2013.

[636] P. J. Rousseeuw, “Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,” *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.

[637] O. M. Abdeldayem *et al.*, “Human-centric ensemble AI for hydrothermal carbonization modeling and hydrochar properties prediction,” *J. Environ. Chem. Eng.*, p. 117826, 2025.

[638] Z. Zhang, “Introduction to machine learning: k-nearest neighbors,” *Ann. Transl. Med.*, vol. 4, no. 11, 2016.

[639] S. Ray, “A quick review of machine learning algorithms,” in *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*, 2019, pp. 35–39.

[640] Y. Zhao, Y. Qian, and C. Li, “Improved KNN text classification algorithm with MapReduce implementation,” in *2017 4th International Conference on Systems and Informatics (ICSAI)*, 2017, pp. 1417–1422.

[641] Y. Liu, Y. Wang, and J. Zhang, “New machine learning algorithm: Random forest,” in *Information Computing and Applications: Third International Conference, ICICA 2012, Chengde, China, September 14-16, 2012. Proceedings 3*, 2012, pp. 246–252.

[642] M. Schonlau and R. Y. Zou, “The random forest algorithm for statistical learning,” *Stata J.*, vol. 20, no. 1, pp. 3–29, 2020.

[643] G. Biau and E. Scornet, “A random forest guided tour,” *Test*, vol. 25, pp. 197–227, 2016.

[644] T. Chen *et al.*, “Xgboost: extreme gradient boosting,” *R Packag. version 0.4-2*, vol. 1, no. 4, pp. 1–4, 2015.

[645] X. Yu, Y. Wang, L. Wu, G. Chen, L. Wang, and H. Qin, “Comparison of support vector regression and extreme gradient boosting for decomposition-based data-driven 10-day streamflow forecasting,” *J. Hydrol.*, vol. 582, p. 124293, 2020.

[646] F. Zhang and L. J. O’Donnell, “Support vector regression,” in *Machine learning*, Elsevier,

2020, pp. 123–140.

[647] A. J. Smola and B. Schölkopf, “A tutorial on support vector regression,” *Stat. Comput.*, vol. 14, pp. 199–222, 2004.

[648] M. Awad, R. Khanna, M. Awad, and R. Khanna, “Support vector regression,” *Effic. Learn. Mach. Theor. concepts, Appl. Eng. Syst. Des.*, pp. 67–80, 2015.

[649] D. F. Specht and others, “A general regression neural network,” *IEEE Trans. neural networks*, vol. 2, no. 6, pp. 568–576, 1991.

[650] S. Sen, K. P. Singh, and P. Chakraborty, “Dealing with imbalanced regression problem for large dataset using scalable Artificial Neural Network,” *New Astron.*, vol. 99, p. 101959, 2023.

[651] A. Heiat, “Comparison of artificial neural network and regression models for estimating software development effort,” *Inf. Softw. Technol.*, vol. 44, no. 15, pp. 911–922, 2002.

[652] A. Sherstinsky, “Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network,” *Phys. D Nonlinear Phenom.*, vol. 404, p. 132306, 2020.

[653] R. C. Staudemeyer and E. R. Morris, “Understanding LSTM--a tutorial into long short-term memory recurrent neural networks,” *arXiv Prepr. arXiv1909.09586*, 2019.

[654] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, “A survey of convolutional neural networks: analysis, applications, and prospects,” *IEEE Trans. neural networks Learn. Syst.*, 2021.

[655] J. Gu *et al.*, “Recent advances in convolutional neural networks,” *Pattern Recognit.*, vol. 77, pp. 354–377, 2018.

[656] K. O’Shea and R. Nash, “An introduction to convolutional neural networks,” *arXiv Prepr. arXiv1511.08458*, 2015.

[657] S. Albawi, T. A. Mohammed, and S. Al-Zawi, “Understanding of a convolutional neural network,” in *2017 international conference on engineering and technology (ICET)*, 2017, pp. 1–6.

[658] E. Schulz, M. Speekenbrink, and A. Krause, “A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions,” *J. Math. Psychol.*, vol. 85, pp. 1–16, 2018.

[659] A. Shalaby, A. Elkamel, P. L. Douglas, Q. Zhu, and Q. P. Zheng, “A machine learning approach for modeling and optimization of a CO2 post-combustion capture unit,” *Energy*, vol. 215, p. 119113, 2021.

[660] W. M. Ashraf and V. Dua, “Machine learning based modelling and optimization of post-combustion carbon capture process using MEA supporting carbon neutrality,” *Digit. Chem. Eng.*, vol. 8, p. 100115, 2023.

[661] J. Chen and F. Wang, “Cost reduction of CO2 capture processes using reinforcement learning based iterative design: A pilot-scale absorption--stripping system,” *Sep. Purif. Technol.*, vol. 122, pp. 149–158, 2014.

[662] S. Ghavamipour, L. Vafajoo, G. Pourhossein, P. Parthasarathy, and G. McKay, “Post-combustion CO2 capturing by KOH solution: An experimental and statistical optimization modeling study,” *Energy \& Environ.*, p. 0958305X241230944, 2024.

[663] A. Ardeshiri and H. Rashidi, “Performance of Water-Lean Solvent for Postcombustion Carbon Dioxide Capture in a Process-Intensified Absorber: Experimental, Modeling, and Optimization Using RSM and ML,” *Ind. \& Eng. Chem. Res.*, vol. 62, no. 48, pp. 20821–20832, 2023.

[664] Z. Li, M. Sharma, R. Khalilpour, and A. Abbas, “Optimal operation of solvent-based post-combustion carbon capture processes with reduced models,” *Energy Procedia*, vol. 37, pp. 1500–1508, 2013.

[665] A. Hemmati and H. Rashidi, “Optimization of industrial intercooled post-combustion CO2 absorber by applying rate-base model and response surface methodology (RSM),” *Process Saf. Environ. Prot.*, vol. 121, pp. 77–86, 2019.

[666] B. Alexanda Petrovic and S. Masoudi Soltani, “Optimization of post combustion CO2 capture from a combined-cycle gas turbine power plant via taguchi design of experiment,” *Processes*, vol. 7, no. 6, p. 364, 2019.

[667] S. Sahraie, H. Rashidi, and P. Valeh-e-Sheyda, “An optimization framework to investigate the CO2 capture performance by MEA: Experimental and statistical studies using Box-Behnken design,” *Process Saf. Environ. Prot.*, vol. 122, pp. 161–168, 2019.

[668] R. Nazerifard, M. Mohammadpourfard, and S. Z. Heris, “Optimization of the integrated

ORC and carbon capture units coupled to the refinery furnace with the RSM-BBD method," *J. CO2 Util.*, vol. 66, p. 102289, 2022.

[669] F.-M. Ilea, A.-M. Cormos, V.-M. Cristea, and C.-C. Cormos, "Enhancing the post-combustion carbon dioxide carbon capture plant performance by setpoints optimization of the decentralized multi-loop and cascade control system," *Energy*, vol. 275, p. 127490, 2023.

[670] E. Marcoulaki, P. Linke, and A. Kokossis, "Design of separation trains and reaction-separation networks using stochastic optimization methods," *Chem. Eng. Res. Des.*, vol. 79, no. 1, pp. 25–32, 2001.

[671] R. Chebbi, "Optimizing reactors selection and sequencing: minimum cost versus minimum volume," *Chinese J. Chem. Eng.*, vol. 22, no. 6, pp. 651–656, 2014.

[672] L. E. Øi, "Aspen HYSYS simulation of CO2 removal by amine absorption from a gas based power plant," in *The 48th Scandinavian Conference on Simulation and Modeling (SIMS 2007), 30-31 October, 2007, Göteborg (Särö)*, 2007, pp. 73–81.

[673] N. Wang, D. Wang, A. Krook-Riekkola, and X. Ji, "MEA-based CO2 capture: a study focuses on MEA concentrations and process parameters," *Front. Energy Res.*, vol. 11, 2023.

[674] D. Mehta, "State-of-the-art reinforcement learning algorithms," *Int. J. Eng. Res. Technol.*, vol. 8, pp. 717–722, 2020.

[675] Y. Hou, L. Liu, Q. Wei, X. Xu, and C. Chen, "A novel DDPG method with prioritized experience replay," in *2017 IEEE international conference on systems, man, and cybernetics (SMC)*, 2017, pp. 316–321.

[676] Y. Zhang and X. Xu, "Solubility predictions through LSBoost for supercritical carbon dioxide in ionic liquids," *New J. Chem.*, vol. 44, no. 47, pp. 20544–20567, 2020.

[677] T.-L. Liu, L.-Y. Liu, F. Ding, and Y.-Q. Li, "A machine learning study of polymer-solvent interactions," *Chinese J. Polym. Sci.*, vol. 40, no. 7, pp. 834–842, 2022.

[678] H. Feng, P. Zhang, W. Qin, W. Wang, and H. Wang, "Estimation of solubility of acid gases in ionic liquids using different machine learning methods," *J. Mol. Liq.*, vol. 349, p. 118413, 2022.

[679] K. Ge and Y. Ji, “Novel computational approach by combining machine learning with molecular thermodynamics for predicting drug solubility in solvents,” *Ind. \& Eng. Chem. Res.*, vol. 60, no. 25, pp. 9259–9268, 2021.

[680] Z. Ye and D. Ouyang, “Prediction of small-molecule compound solubility in organic solvents by machine learning algorithms,” *J. Cheminform.*, vol. 13, no. 1, p. 98, 2021.

[681] N. AlQasas and D. Johnson, “The use of neural network modeling for the estimation of the Hansen solubility parameters of polymer films from contact angle measurements,” *Surfaces and Interfaces*, vol. 44, p. 103721, 2024.

[682] Y. Cui *et al.*, “Performance and Mechanism Study on Functionalized Phosphonium-Based Deep Eutectic Solvents for CO2 Absorption,” *Int. J. Thermophys.*, vol. 44, no. 7, p. 98, 2023.

[683] R. H. Myers, D. C. Montgomery, and C. M. Anderson-Cook, *Response surface methodology: process and product optimization using designed experiments*. John Wiley \& Sons, 2016.

[684] E. G. Al-Sakkari *et al.*, “A bi-functional alginate-based composite for catalyzing one-pot methyl esters synthesis from waste cooking oil having high acidity,” *Fuel*, vol. 306, p. 121637, 2021.

[685] L. M. S. Pereira, T. M. Milan, and D. R. Tapia-Blácido, “Using Response Surface Methodology (RSM) to optimize 2G bioethanol production: A review,” *Biomass and Bioenergy*, vol. 151, p. 106166, 2021.

[686] I. Veza, M. Spraggon, I. M. R. Fattah, and M. Idris, “Response surface methodology (RSM) for optimizing engine performance and emissions fueled with biofuel: Review of RSM for sustainability energy transition,” *Results Eng.*, vol. 18, p. 101213, 2023.

[687] H. Ghaedi, S. Akbari, H. Zhou, W. Wang, and M. Zhao, “Excess properties of and simultaneous effects of important parameters on CO2 solubility in binary mixture of water-phosphonium based-deep eutectic solvents: response surface methodology (RSM) and Taguchi method,” *Energy \& Fuels*, vol. 36, no. 4, pp. 1960–1972, 2022.

[688] X. Rodr\’\iguez-Mart\’\inez, E. Pascual-San-José, and M. Campoy-Quiles, “Accelerating organic solar cell material’s discovery: high-throughput screening and big data,” *Energy \& Environ. Sci.*, vol. 14, no. 6, pp. 3301–3322, 2021.

[689] Z. Chen *et al.*, “Development of high-throughput wet-chemical synthesis techniques for material research,” *Mater. Genome Eng. Adv.*, vol. 1, no. 1, p. e5, 2023.

[690] H. Lim, H. Cho, and J. Kim, “SAGE-Amine: Generative Amine Design with Multi-Property Optimization for Efficient CO2 Capture,” *arXiv Prepr. arXiv2503.02534*, 2025.

[691] E. Walker, J. Kammeraad, J. Goetz, M. T. Robo, A. Tewari, and P. M. Zimmerman, “Learning to predict reaction conditions: relationships between solvent, molecular structure, and catalyst,” *J. Chem. Inf. Model.*, vol. 59, no. 9, pp. 3645–3654, 2019.

[692] B. Selvaratnam, R. T. Koodali, and P. Miró, “Application of symmetry functions to large chemical spaces using a convolutional neural network,” *J. Chem. Inf. Model.*, vol. 60, no. 4, pp. 1928–1935, 2020.

[693] L. Zhang, S. Cignitti, and R. Gani, “Generic mathematical programming formulation and solution for computer-aided molecular design,” *Comput. \& Chem. Eng.*, vol. 78, pp. 79–84, 2015.

[694] A. Pratap, “AI will be everywhere in chemistry,” *C\&EN Glob. Enterp.*, vol. 103, no. 1, p. 27, 2025, doi: 10.1021/cen-10301-feature12.

[695] E. C. Achinivu *et al.*, “A predictive toolset for the identification of effective lignocellulosic pretreatment solvents: a case study of solvents tailored for lignin extraction,” *Green Chem.*, vol. 23, no. 18, pp. 7269–7289, 2021.

[696] F. Eckert and A. Klamt, “Fast solvent screening via quantum chemistry: COSMO-RS approach,” *AIChE J.*, vol. 48, no. 2, pp. 369–385, 2002.

[697] A. Klamt, F. Eckert, and W. Arlt, “COSMO-RS: an alternative to simulation for calculating thermodynamic properties of liquid mixtures,” *Annu. Rev. Chem. Biomol. Eng.*, vol. 1, pp. 101–122, 2010.

[698] A. Klamt, “The COSMO and COSMO-RS solvation models,” *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, vol. 8, no. 1, p. e1338, 2018.

[699] D. O. Abranches *et al.*, “Using COSMO-RS to design choline chloride pharmaceutical eutectic solvents,” *Fluid Phase Equilib.*, vol. 497, pp. 71–78, 2019.

[700] J. Scheffczyk *et al.*, “COSMO-CAMPD: a framework for integrated design of molecules

and processes based on COSMO-RS," *Mol. Syst. Des. \& Eng.*, vol. 3, no. 4, pp. 645–657, 2018.

[701] J. Palomar, V. R. Ferro, J. S. Torrecilla, and F. Rodr\'\iguez, "Density and molar volume predictions using COSMO-RS for ionic liquids. An approach to solvent design," *Ind. \& Eng. Chem. Res.*, vol. 46, no. 18, pp. 6041–6048, 2007.

[702] R. Potyrailo, K. Rajan, K. Stoewe, I. Takeuchi, B. Chisholm, and H. Lam, "Combinatorial and high-throughput screening of materials libraries: review of state of the art," *ACS Comb. Sci.*, vol. 13, no. 6, pp. 579–633, 2011.

[703] M. Zeng *et al.*, "High-throughput printing of combinatorial materials from aerosols," *Nature*, vol. 617, no. 7960, pp. 292–298, 2023.

[704] S. S. Mao and P. E. Burrows, "Combinatorial screening of thin film materials: An overview," *J. Mater.*, vol. 1, no. 2, pp. 85–91, 2015.

[705] P. Yoo *et al.*, "Deep learning workflow for the inverse design of molecules with specific optoelectronic properties," *Sci. Rep.*, vol. 13, no. 1, p. 20031, 2023.

[706] A. P. Samudra and N. V Sahinidis, "Optimization-based framework for computer-aided molecular design," *AIChE J.*, vol. 59, no. 10, pp. 3686–3701, 2013.

[707] T. Schiex, "Coupling CP with Deep Learning for Molecular Design and SARS-CoV2 Variants Exploration (Invited Talk)," 2023.

[708] M. MacNeil and M. Bodur, "Constraint programming approaches for the discretizable molecular distance geometry problem," *Networks*, vol. 79, no. 4, pp. 515–536, 2022.

[709] N. D. Austin, N. V Sahinidis, and D. W. Trahan, "Computer-aided molecular design: An introduction and review of tools, applications, and solution techniques," *Chem. Eng. Res. Des.*, vol. 116, pp. 2–26, 2016.

[710] S. Cignitti, L. Zhang, and R. Gani, "Computer-aided framework for design of pure, mixed and blended products," in *Computer Aided Chemical Engineering*, vol. 37, Elsevier, 2015, pp. 2093–2098.

[711] T. Zhou, Y. Zhou, and K. Sundmacher, "A hybrid stochastic--deterministic optimization approach for integrated solvent and process design," *Chem. Eng. Sci.*, vol. 159, pp. 207–

216, 2017.

[712] Y. Carissan, D. Hagebaum-Reignier, N. Prcovic, C. Terrioux, and A. Varet, “How constraint programming can help chemists to generate Benzenoid structures and assess the local Aromaticity of Benzenoids,” *Constraints*, vol. 27, no. 3, pp. 192–248, 2022.

[713] X. Zhang, J. Wang, Z. Song, and T. Zhou, “Data-driven ionic liquid design for CO2 capture: Molecular structure optimization and DFT verification,” *Ind. \& Eng. Chem. Res.*, vol. 60, no. 27, pp. 9992–10000, 2021.

[714] D. P. Kroese, T. Brereton, T. Taimre, and Z. I. Botev, “Why the Monte Carlo method is so important today,” *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 6, no. 6, pp. 386–392, 2014.

[715] T. M. Dieb, S. Ju, K. Yoshizoe, Z. Hou, J. Shiomi, and K. Tsuda, “MDTS: automatic complex materials design using Monte Carlo tree search,” *Sci. Technol. Adv. Mater.*, vol. 18, no. 1, pp. 498–503, 2017.

[716] K. Ohno, K. Esfarjani, and Y. Kawazoe, *Computational materials science: from ab initio to Monte Carlo methods*. Springer, 2018.

[717] T. Bereau, D. Andrienko, and K. Kremer, “Research Update: Computational materials discovery in soft matter,” *APL Mater.*, vol. 4, no. 5, 2016.

[718] K.-D. Luong and A. Singh, “Application of transformers in cheminformatics,” *J. Chem. Inf. Model.*, vol. 64, no. 11, pp. 4392–4409, 2024.

[719] N. Fu *et al.*, “Material transformers: deep learning language models for generative materials design,” *Mach. Learn. Sci. Technol.*, vol. 4, no. 1, p. 15001, 2023.

[720] C. P. Gomes, B. Selman, and J. M. Gregoire, “Artificial intelligence for materials discovery,” *MRS Bull.*, vol. 44, no. 7, pp. 538–544, 2019.

[721] J. Jumper *et al.*, “Highly accurate protein structure prediction with AlphaFold,” *Nature*, vol. 596, no. 7873, pp. 583–589, 2021.

[722] S. Ismail, H. Safari, and M. Bavarian, “Design of Supported Ionic Liquid Membranes for CO2 Capture Using a Generative AI-Based Approach,” *Ind. \& Eng. Chem. Res.*, vol. 64, no. 8, pp. 4439–4449, 2025.

[723] A. Hagg and K. N. Kirschner, “Open-source machine learning in computational chemistry,”

J. Chem. Inf. Model., vol. 63, no. 15, pp. 4505–4532, 2023.

[724] Z. Niu *et al.*, "Generative artificial intelligence for designing multi-scale hydrogen fuel cell catalyst layer nanostructures," *ACS Nano*, vol. 18, no. 31, pp. 20504–20517, 2024.

[725] L. Wossnig, N. Furtmann, A. Buchanan, S. Kumar, and V. Greiff, "Best practices for machine learning in antibody discovery and development," *Drug Discov. Today*, p. 104025, 2024.

[726] Y. Wang, H. Zhao, S. Sciabola, and W. Wang, "cMolGPT: A conditional generative pre-trained transformer for target-specific de novo molecular generation," *Molecules*, vol. 28, no. 11, p. 4430, 2023.

[727] W. Fan, Y. He, and F. Zhu, "RM-GPT: Enhance the comprehensive generative ability of molecular GPT model via LocalRNN and RealFormer," *Artif. Intell. Med.*, p. 102827, 2024.

[728] A. E. Cleves, S. R. Johnson, and A. N. Jain, "Synergy and complementarity between focused machine learning and physics-based simulation in affinity prediction," *J. Chem. Inf. Model.*, vol. 61, no. 12, pp. 5948–5966, 2021.

[729] X. Hu, G. Liu, Y. Zhao, and H. Zhang, "De novo Drug Design using Reinforcement Learning with Multiple GPT Agents," *Adv. Neural Inf. Process. Syst.*, vol. 36, 2024.

[730] S. Haroon, C. A. Hafsath, and A. S. Jereesh, "Generative pre-trained transformer (GPT) based model with relative attention for de novo drug design," *Comput. Biol. Chem.*, vol. 106, p. 107911, 2023.

[731] H. Lu, Z. Wei, X. Wang, K. Zhang, and H. Liu, "GraphGPT: A Graph Enhanced Generative Pretrained Transformer for Conditioned Molecular Generation," *Int. J. Mol. Sci.*, vol. 24, no. 23, p. 16761, 2023.

[732] E. Mazuz, G. Shtar, B. Shapira, and L. Rokach, "Molecule generation using transformers and policy gradient reinforcement learning," *Sci. Rep.*, vol. 13, no. 1, p. 8799, 2023.

[733] Y. Gu, Y. Wang, K. Zhu, W. Li, G. Liu, and Y. Tang, "DBPP-Predictor: a novel strategy for prediction of chemical drug-likeness based on property profiles," *J. Cheminform.*, vol. 16, no. 1, p. 4, 2024.

[734] W. Jin, R. Barzilay, and T. Jaakkola, "Junction tree variational autoencoder for molecular

graph generation," in *International conference on machine learning*, 2018, pp. 2323–2332.

[735] M. Simonovsky and N. Komodakis, "Graphvae: Towards generation of small graphs using variational autoencoders," in *Artificial Neural Networks and Machine Learning--ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part I 27*, 2018, pp. 412–422.

[736] Ł. Maziarka, A. Pocha, J. Kaczmarczyk, K. Rataj, T. Danel, and M. Warchoł, "Mol-CycleGAN: a generative model for molecular optimization," *J. Cheminform.*, vol. 12, no. 1, p. 2, 2020.

[737] B. Macedo, I. Ribeiro Vaz, and T. Taveira Gomes, "MedGAN: optimized generative adversarial network with graph convolutional networks for novel molecule design," *Sci. Rep.*, vol. 14, no. 1, p. 1212, 2024.

[738] Z. Zhou, S. Kearnes, L. Li, R. N. Zare, and P. Riley, "Optimization of molecules via deep reinforcement learning," *Sci. Rep.*, vol. 9, no. 1, p. 10752, 2019.

[739] P. Pogány, N. Arad, S. Genway, and S. D. Pickett, "De novo molecule design by translating from reduced graphs to SMILES," *J. Chem. Inf. Model.*, vol. 59, no. 3, pp. 1136–1146, 2018.

[740] L. Liu, X. Zhao, and X. Huang, "Generating Potential RET-Specific Inhibitors Using a Novel LSTM Encoder--Decoder Model," *Int. J. Mol. Sci.*, vol. 25, no. 4, p. 2357, 2024.

[741] M. Popova, O. Isayev, and A. Tropsha, "Deep reinforcement learning for de novo drug design," *Sci. Adv.*, vol. 4, no. 7, p. eaap7885, 2018.

[742] Y. Li, C. Gao, X. Song, X. Wang, Y. Xu, and S. Han, "DrugGPT: A GPT-based strategy for designing potential ligands targeting specific proteins," *bioRxiv*, pp. 2006–2023, 2023.

[743] Z. Xie, X. Evangelopoulos, Ö. H. Omar, A. Troisi, A. I. Cooper, and L. Chen, "Fine-tuning GPT-3 for machine learning electronic and functional properties of organic molecules," *Chem. Sci.*, vol. 15, no. 2, pp. 500–510, 2024.

[744] S. Kim, Y. Jung, and J. Schrier, "Large Language Models for Inorganic Synthesis Predictions," 2024.

[745] Z. Zheng, Z. Rong, N. Rampal, C. Borgs, J. T. Chayes, and O. M. Yaghi, "A GPT-4 Reticular Chemist for Guiding MOF Discovery," *Angew. Chemie Int. Ed.*, vol. 62, no. 46, p.

e202311983, 2023.

[746] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, and P. Schwaller, “ChemCrow: Augmenting large-language models with chemistry tools,” *arXiv Prepr. arXiv2304.05376*, 2023.

[747] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White, and P. Schwaller, “Augmenting large language models with chemistry tools,” *Nat. Mach. Intell.*, pp. 1–11, 2024.

[748] J. Abramson *et al.*, “Accurate structure prediction of biomolecular interactions with AlphaFold 3,” *Nature*, pp. 1–3, 2024.

[749] Y. Wang, Z. Li, L. Wang, M. Wang, and others, “A Scale Invariant Feature Transform Based Method.,” *J. Inf. Hiding Multim. Signal Process.*, vol. 4, no. 2, pp. 73–89, 2013.

[750] T. Ahonen, A. Hadid, and M. Pietikäinen, “Face recognition with local binary patterns,” in *Computer Vision-ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part I 8*, 2004, pp. 469–481.

[751] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)*, 2005, vol. 1, pp. 886–893.

[752] J. Alammar, “The Illustrated GPT-2 (Visualizing Transformer Language Models),” 2019. https://jalammar.github.io/illustrated-gpt2/.

[753] A. Vaswani *et al.*, “Attention is all you need,” *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[754] A. Arora, “The Annotated GPT-2.” https://amaarora.github.io/posts/2020-02-18-annotatedGPT2.html.

[755] A. Aggarwal, M. Mittal, and G. Battineni, “Generative adversarial network: An overview of theory and applications,” *Int. J. Inf. Manag. Data Insights*, vol. 1, no. 1, p. 100004, 2021.

APPENDIX A LIST OF PUBLICATIONS

The following is the list of publications and presentations related directly to the work done throughout this doctoral research.

Al-Sakkari, E. G., Ragab, A., Dagdougui, H., Boffito, D. C., & Amazouz, M. (2024). Carbon capture, utilization and sequestration systems design and operation optimization: Assessment and perspectives of artificial intelligence opportunities. Science of The Total Environment, 917, 170085. (Journal Article)

Al-Sakkari, E. G., Ragab, A., So, T. M., Shokrollahi, M., Dagdougui, H., Navarri, P., Elkamel A., & Amazouz, M. (2023). Machine learning-assisted selection of adsorption-based carbon dioxide capture materials. Journal of Environmental Chemical Engineering, 11(5), 110732. (Journal Article)

Al-Sakkari, E. G., Ragab, A., Amer, M., Ajao, O., Benali, M., Boffito, D. C., ... & Amazouz, M. (2025). Ensemble machine learning to accelerate industrial decarbonization: Prediction of Hansen solubility parameters for streamlined chemical solvent selection. Digital Chemical Engineering, 14, 100207. (Journal Article)

Al-Sakkari, E. G., Ragab, A., Benali, M., Ajao, O., Boffito, D. C., & Dagdougui, H. (2026). Accelerated green material and solvent discovery with chemistry-and physics-guided generative AI. Artificial Intelligence Chemistry, 4(1), 100106. (Journal Article)

Al-Sakkari, E. G., Ragab, A., Ali, M., Dagdougui, H., Boffito, D. C., & Amazouz, M. (2024). Learn-To-Design: Reinforcement Learning-Assisted Chemical Process Optimization. Systems and Control Transactions, 3. (Conference Paper)

Al-Sakkari, E. G., Ragab, A., Ali, M., Dagdougui, H., & Boffito, D. C. (2025). Simulate Intelligently: Causal Incremental Reinforcement Learning for Streamlined Industrial Chemical Process Design Optimization. Journal of Environmental Chemical Engineering, 13(6), 120167. (Journal Article)

Al-Sakkari, E. G., Ahmed Ragab, Mohamed Ali, Hanane Dagdougui, Daria C. Boffito and Mouloud Amazouz, (2025), Causal Incremental Reinforcement Learning for Sustainable Process Design, Presentation in the Canadian Chemical Engineering Conference (CSChE 2025), Montreal-QC, Canada (Conference Presentation)

APPENDIX B TABLES OF ARTICLE 1

Table B.1 Decision tree patterns for capacity (Dataset I)

Pattern ID	Pressure (bar)	Temperature (°C)	BET (m^2/g)	Ads_Lab	DP (nm)	VP (cc/g)	CPP (bar)	Capacity class
P1		<= 24	<= 854.5		<= 1.145	> 0.067		
P2		> 24	<= 854.5		1.145 >= DP > 0.705			
P3			<= 727	<= 130	> 1.145	<= 0.515	<= 1.02	
P4			<= 854.5		5.8 >= DP > 1.145	> 0.515	<= 1.02	
P5		<= 7.5	<= 854.5		7.94 >= DP > 5.8	> 0.515	<= 1.02	
P6		<= 7.5	<= 854.5	<= 171	> 7.94	> 0.515	<= 1.02	
P7		> 7.5	<= 271		> 5.8	> 0.515	> 0.3	High (above 2)
P8			854.5 >= BET > 264.515		> 1.145		> 1.02	
P9		<= 27.5	> 854.5			<= 1.415	> 0.75	
P10		> 27.5	> 854.5	<= 46		<= 1.415	> 0.75	
P11	> 5.5	> 27.5	> 854.5	> 46		<= 1.415	> 0.75	
P12		<= 20	> 854.5		> 5.35	> 1.415	28>=CPP> 0.75	
P13		<= 45	> 854.5			> 1.415	> 28	

Table B.1 Decision tree patterns for capacity (Dataset I) (Cont'd and end)

Pattern ID	Pressure (bar)	Temperature (°C)	BET (m^2/g)	Ads_Lab	DP (nm)	VP (cc/g)	CPP (bar)	Capacity class
P14		> 45	> 1655			> 1.415	> 0.75	High (above 2)
P15		<= 24	<= 854.5		<= 1.145	<= 0.067		Low (below 2)
P16		> 24	<= 854.5		<= 0.705			
P17			854.5 >= BET > 727	<= 130	> 1.145	<= 0.515	<= 1.02	
P18			<= 854.5	> 130	> 1.145	<= 0.515	<= 1.02	
P19		<= 7.5	<= 854.5	> 171	> 7.94	> 0.515	<= 1.02	
P20		> 7.5	<= 271		> 5.8	> 0.515	<= 0.3	
P21		> 7.5	> 271		> 5.8	> 0.515	> 0.3	
P22			<= 264.515		> 1.145		> 1.02	
P23			> 854.5				<= 0.75	
P24	<= 5.5	> 27.5	> 854.5	> 46		<= 1.415	> 0.75	
P25		<= 45	> 854.5		<= 5.35	> 1.415	28>=CPP> 0.75	
P26		> 20	> 854.5		> 5.35	> 1.415	28>=CPP> 0.75	
P27		> 45	1655 >= BET > 854.5			> 1.415	> 0.75	

Table B.2 Decision tree patterns for selectivity (Dataset I)

Pattern ID	Pressure (bar)	Temperature (°C)	BET (m^2/g)	Ads_Lab	CPP (bar)	Selectivity class
P1	<= 3.52	<= 62.5	> 22.25	<= 226		High (above 20)
P2	1.02 < Press. <= 3.52	<= 62.5	1031 >= BET > 22.25	<= 226		
P3	<= 1.02	<= 62.5	512 >= BET > 22.25	<= 226		
P4	<= 1.02	<= 12.5	525.5 >= BET > 512	<= 226		
P5	> 3.52	> 37.5		<= 70	<= 12.5	
P6	<= 3.52		<= 22.25			Low (below 20)
P7	<= 3.52		> 22.25	>226		
P8	<= 3.52	> 62.5	> 22.25	<= 226		
P9	1.02 < Press. <= 3.52	<= 62.5	> 1031	<= 226		
P10	> 3.52			> 70		
P11	> 3.52			<= 70	> 12.5	
P12	> 3.52	<= 37.5		<= 70	<= 12.5	

Table B.3 Decision tree patterns for combined capacity-selectivity classification (Dataset I)

Pattern ID	Pressure (bar)	Temperature (°C)	BET (m^2/g)	Ads_Lab	VP (cc/g)	CPP (bar)	Capacity-Selectivity class
P1	<= 12.5	<= 24	632 >= BET > 53.7	> 98	<= 0.295		High capacity-High selectivity
P2	<= 12.5	<= 12.5		<= 225	0.391 >= VP > 0.295	1.02>=CPP > 0.575	
P3	<= 12.5			<= 225	> 0.391	1.02>=CPP > 0.575	
P4	<= 12.5		<=1031	<= 225	> 0.295	> 1.02	
P5	<= 12.5		> 1031	<= 225	> 0.295	> 1.02	High capacity-Low selectivity
P6	<= 12.5			> 225	> 0.295	> 0.575	
P7	> 12.5				> 0.031		
P8	<= 5.5		<= 53.7	<= 72	<= 0.295		Low capacity-High selectivity
P9	12.5 >= P > 5.5	> 37.5	<= 53.7	<= 70	<= 0.295		
P10	<= 12.5		> 53.7	<= 98	<= 0.295		
P11	<= 12.5	> 24	632 >= BET > 53.7	> 98	<= 0.295		
P12	<= 12.5		> 632	> 98	<= 0.295		

Table B.3 Decision tree patterns for combined capacity-selectivity classification (Dataset I) (Cont'd and end)

Pattern ID	Pressure (bar)	Temperature (°C)	BET (m^2/g)	Ads_Lab	VP (cc/g)	CPP (bar)	Capacity-Selectivity class
P13	<= 12.5				1.02 >= VP > 0.405	<= 0.575	Low capacity-High selectivity
P14	<= 12.5	> 12.5		<= 225	0.391 >= VP > 0.36	1.02>=CPP > 0.575	
P15	12.5 >= P > 5.5	<= 37.5	<= 53.7	<= 70	<= 0.295		Low capacity-Low selectivity
P16	12.5 >= P > 5.5		<= 53.7	72 >= Ads_Lab > 70			
P17	<= 12.5		<= 53.7	> 72	<= 0.295		
P18	<= 12.5				0.405 >= VP > 0.295	<= 0.575	
P19	<= 12.5				> 1.02	<= 0.575	
P20	<= 12.5	> 12.5		<= 225	0.36 >= VP > 0.295	1.02>=CPP > 0.575	
P21	> 12.5				<= 0.031		

APPENDIX C TABLES OF ARTICLE 2

Table C.1 Overview of the different ML techniques and their key merits

Model	Description	Merits	Reference(s)
***k*-NN**	A non-parametric supervised ML technique that does not assume certain distribution. It is widely used for both regression and classification based on the number of nearest similar neighbours.	Simple and fast model that requires inexpensive training.	[638]–[640]
DT	Supervised interpretable ML technique that performs classification and regression through splitting the data based on specific parameters (cut points). It is a flow chart-like algorithm.	Easy handling and interpretation. Gives interpretable patterns.	[639]
RF	An ensemble-based ML technique consists of multiple weak decision trees and represents an advanced form of bagging.	Easy to handle and it can overcome the problem of overfitting.	[641]–[643]
XGB	An ensemble of decision trees that considers the boosting mechanism and level-wise tree growth.	Fast and high accuracy predictions besides overcoming the overfitting problem and handling the missing data	[644], [645]

Table C.1 Overview of the different ML techniques and their key merits (Cont'd)

Model	Description	Merits	Reference(s)
SVR	Supervised ML model having the objective of finding/creating a hyperplane that fits the training data while keeping a specific margin around this hyperplane. The regression model and its efficiency are determined by the closest points around this hyperplane which forms the "support vectors".	Easy tuning, memory efficiency, high performance even when the data has noise or outliers and the flexibility in handling both linear and non-linear relationships in the data.	[646]–[648]
ANN	A class of ML models inspired by the human brain cells (neurons) and the structure of the human brain that consists of three main types of layers, i.e. input layer, hidden layer(s) and output layer.	It can handle a large amount of structured and tabular data, possess high flexibility when dealing with complex and non-linear data, and can adapt to new data points.	[649]–[651]
LSTM	A type of recurrent neural network that is designed and used to handle/model sequential data.	It can handle large datasets, excellent performance when dealing with sequential data having long-term dependencies such as language structures, can avoid the vanishing gradient problem and can handle irregularly spaced data points (sequences with varying lengths).	[652], [653]

Table C.1 Overview of the different ML techniques and their key merits (Cont'd and end)

Model	Description	Merits	Reference(s)
CNN	A special type of artificial neural network is developed to process and identify patterns in grid-like data and images. Its main components are convolutional, pooling, and fully connected layers.	It can capture both high and low levels of information from the available data. Further, it can handle large datasets and offers the advantage of transfer learning	[654]–[657]
GPR	A non-parametric probabilistic machine learning model. It works according to the Gaussian processes concept representing distributions over functions. It models the entire space of the function and gives predictions in probability distributions form instead of estimating the parameters of a specific model.	It can handle noisy and limited data	[658]

Table C.2 Dataset excerpt - Solvents with their SMILES codes and Hansen Solubility Parameters

Solvent Name	**Raw Input Variable**	**Output Variables**		
	SMILES Codes	**D**	**P**	**H**
Acetone	CC(C)=O	15.5	10.4	7
Acetonitrile	CC#N	15.3	18	6.1
Benzyl Benzoate	O=C(OCC2=CC=CC=C2)C1=CC=CC=C1	20	5.1	5.2
tert-Butanol	O[C@](C)(C)C	15.2	5.1	14.7
n-Butyl benzoate	O=C(C1=CC=CC=C1)OCCCC	18.3	5.6	5.5
Butyldiglycol acetate	CCCCOCCOCCOC(=O)C	16	4.1	8.2
Dimethyl isosorbide (DMI)	O([C@@H]1[C@H]2OC[C@@H](OC)[C@H]2OC1)C	17.6	7.1	7.5
(±)-Limonene	CC1=CCC(CC1)C(=C)C	17.2	1.8	4.3
Methanol	OC([H])([H])[H]	14.7	12.3	22.3
Methyl Oleate	CCCCCCCC\C=C/CCCCCCCC(OC)=O	16.2	3.8	4.5
N-Methylpyrrolidone	CN1C(CCC1)=O	18	12.3	7.2
Dichloromethane	ClCCl	17	7.3	7.1
Vinyl trifluoroacetate	C=COC(=O)C(F)(F)F	13.9	4.3	7.6

Table C.2 Dataset excerpt - Solvents with their SMILES codes and Hansen Solubility Parameters (Cont'd and end)

Solvent Name	Raw Input Variable	Output Variables		
	SMILES Codes	D	P	H
o-Vinyltoluene	CC1=CC=CC=C1C=C	18.6	1	3.8
Vinylsilicon	C=C[Si]	15.5	2.6	4
Vinyl S-ethyl mercapto ethyl ether	C=CSCCOCC	16.4	7	6
4-Vinylpyridine	C=CC1=CC=NC=C1	18.1	7.2	6.8
Vanillin	OC1=CC=C(C([H])=O)C=C1OC	19.4	9.8	11.2
Valeronitrile	CCCCC#N	15.3	11	4.8
L-(−)-Tyrosine	N[C@@H](CC1=CC=C(O)C=C1)C(O)=O	17.5	6.9	17.2
1-[2-(2-Methoxy-1-methylethoxy)-1-methylethoxy]-2-propanol	CC(COC(COC(COC)C)C)O	15.3	5.5	10.4
Phosphoric acid, triphenyl ester	O=[P](OC2=CC=CC=C2)(OC3=CC=CC=C3)OC1=CC=CC=C1	20.1	6.4	6.8
Trioctyl phosphate	CCCCCCCCOP(=O)(OCCCCCCCC)OCCCCCCCC	16.2	5.9	4.2
2,4,6-Trinitrotoluene	CC1=C(C=C(C=C1[N+](=O)[O-])[N+](=O)[O-])[N+](=O)[O-]	19.5	10	4.5

Table C.3 ML modeling results

Variable	*k*-NN		DT		RF		XGB		SVR		ANN		LSTM	
	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE
Solubility1 (δ_D)	0.92	0.26	0.92	0.28	0.95	0.16	0.96	0.15	0.96	0.14	0.92	0.26	0.94	0.19
Solubility2 (δ_P)	0.89	1.72	0.81	2.96	0.91	1.37	0.93	1.06	0.93	1.11	0.90	1.49	0.91	1.44
Solubility3 (δ_H)	0.90	2.17	0.89	2.22	0.95	1.14	0.95	0.96	0.96	0.76	0.93	1.40	0.93	1.49

Table C.4 Decision fusion results (Solubility_1 "δ_D")

Fusion No.	Non-learnable Fusion								Learnable Fusion					
	Average-based		R^2-based		MSE-based		R^2/MSE-based		XGB		ANN		SVR	
	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE
Fusion 1	0.97	0.12	0.97	0.12	0.97	0.11	0.97	0.11	0.97	0.12	0.96	0.14	0.97	0.12
Fusion 2	0.97	0.11	0.97	0.11	0.97	0.11	0.97	0.11	0.97	0.10	--------	--------	0.98	0.06
Fusion 3	0.98	0.05	0.98	0.05	0.98	0.04	0.99	0.03	--------	--------	--------	--------	0.99	0.02

***** *All the values are rounded to the second decimal.*

APPENDIX D TABLES OF ARTICLE 3

Table D.1 Toxicity-flammability categories and their corresponding scores with average color ranking

LC50 (mg/L) [Vapors Inhalation]	**Value & Category**				
	<0.5 (Category 1) Highly toxic	<2 (Category 2) Moderately toxic	<10 (Category 3) Mildly toxic	<20 (Category 4) Minimally toxic	>20 (Category 5) Likely not toxic
	Score				
	1	2	3	4	5
LD50 (mg/kg) [Oral Exposure]	**Value & Category**				
	<5 (Category 1) Highly toxic	<50 (Category 2) Moderately toxic	<300 (Category 3) Mildly toxic	<2000 (Category 4) Minimally toxic	>2000 (Category 5) Likely not toxic
	Score				
	1	2	3	4	5
Developmental Toxicity	**Category**				
	Developmental Toxicant			Developmental Non-toxicant	
	Score				
	1			5	
Flash point (ºC)	**Value & Category**				
	<23 [with boiling point < 35] (Category 3) Extremely Flammable	<23 [with boiling point > 35] (Category 3) Highly Flammable	<60 (Category 3) Flammable	<93 (Category 3) Mildly Flammable	>93 (Category 5) Non-flammable
	Score				
	1	2	3	4	5

Table D.2 GPT-2 generation results

Case study	Validity (%)	Originality (%)	Generation samples
CO_2 solvents	56	87.72	FC(F)(F)C(F)(F)OC(OBr)C(F)(F)F CCCCCCCCC(=O)OCC\C=C\C=C\C=C
Softwood lignin solvents	40	90	O=S(=O)OC(C(O)C)c2ccc(N)cc2 FS(=O)CCC
Bagasse lignin solvents	37	89.19	O=C(O)CCNNC(=O)NO OC1=CC(O)C=CC=C1N

Table D.3 Predicted HSPs, REDs and labels of generated molecules sample

Case study	Model	Generated SMILES	Predicted KPIs				Label
			δ_D	δ_P	δ_H	RED	
CO_2 solvents	GPT-2	O=C(C)OCCCC(F)C	15.91	4.26	5.37	0.30	Good
		CCCCCCCCC(=O)OCC\C=C\C=C\C=C	18.18	11.20	5.39	1.98	Bad
	W-GAN	CCCCOC	15.04	4.11	4.60	0.49	Good
		CCC(C)SC	16.43	3.08	2.67	1.03	Bad
	VAE	CPF	15.33	8.31	3.93	0.92	Good
		COBr	16.79	8.42	7.75	1.11	Bad
Softwood lignin solvents	GPT-2	O=S(=O)OC(C(O)C)c2ccc(N)cc2	19.00	8.69	8.98	0.76	Good
		FS(=O)CCC	15.24	9.10	5.14	1.16	Bad
Bagasse lignin solvents	GPT-2	OC1=CC(O)C=CC=C1N	18.91	8.87	12.52	0.78	Good
		CCNCCNCCNCCCOCCN	16.65	5.84	7.38	1.29	Bad

Table D.4 Toxicity and flammability results (Color-based Ranking)

Case study	SMILES	Toxicity			Flammability	Individual greenness score
		LC50 (mg/L)	LD50 (mg/kg)	Developmental Toxicity	Flash point (°C)	
CO_2 solvents	CCOCCC(CC)OCC	85.42	2335.15	Developmental NON-toxicant	154.10	20
	CCCCCCCCOC(=O)CCCCCCCCCCCC	481.49	16599.04	Developmental toxicant	412.18	16
	CC1=COCCOCCC1	1228.22	734.64	Developmental toxicant	121.97	15
	O=C(C)OCCCC(F)C	1161.50	2279.07	Developmental toxicant	130.98	16
	CCCCCOC(=O)SCC	243.27	3087.88	Developmental toxicant	172.31	16
Softwood lignin solvents	O=C(O)CC(O)C(COC)O	2748.30	7868.07	Developmental NON-toxicant	268.43	20
	O=C(O)CCC(CC)N	423.08	2091.49	Developmental NON-toxicant	100.73	20
	C(Cl)C=CN=C1C=CC=C1	116.38	2859.68	Developmental NON-toxicant	240.90	20
	O=S(=O)OC(C(O)C)c2ccc(N)cc2	22.18	6903.37	Developmental toxicant	411.80	16
	CC(=O)CN(CCO)CO	2785.27	3109.85	Developmental NON-toxicant	214.78	20
Bagasse lignin solvents	CCOCC(O)C(=O)C(O)CO	1997.40	3247.28	Developmental NON-toxicant	335.21	20
	OCCNNCO	3416.11	2117.85	Developmental NON-toxicant	165.33	20
	OC1=CC(O)C=CC=C1N	83.43	1323.53	Developmental toxicant	130.42	15
	OC1=CC=C(C=C1)N	12.40	1509.57	Developmental toxicant	147.43	14
	OCCC(O)CO	2385.62	2906.84	Developmental toxicant	108.02	16

APPENDIX E TABLES OF ARTICLE 4

Table E.1 Selected studies on data-driven-assisted solvent-based carbon capture process design/control optimization

Model type	Model	Inputs (Design variables)	Outputs (KPIs)	References
AI/ML models	Bootstrap neural network	▪ Flue gas temperature ▪ Flue gas pressure ▪ Flue gas flowrate ▪ Flue gas CO_2 concentration ▪ MEA flowrate ▪ MEA temperature ▪ MEA concentration	▪ Capture rate	[108]
	Artificial neural network (ANN)	▪ Flue gas flowrate ▪ Flue gas inlet temperature ▪ Flue gas CO_2 concentration ▪ Reboiler duty ▪ MEA circulation rate	▪ Capture rate ▪ Reboiler specific duty	[109]
		▪ MDEA and PZ concentrations in inlet solvent ▪ Solvent liquid volumetric flow rate ▪ Solvent temperature ▪ Solvent high gravity factor	▪ Absorption efficiency	[110]
		▪ Lean solvent flowrate ▪ Steam flowrate supplied to reboiler.	▪ Capture level ▪ Reboiler temperature	[111]

Table E.1 Selected studies on data-driven-assisted solvent-based carbon capture process design/control optimization (Cont'd)

Model type	Model	Inputs (Design variables)	Outputs (KPIs)	References
AI/ML models	Artificial neural network (ANN)	▪ Amine-based solvent circulation rate ▪ Solvents inlet concentration ▪ Reboiler duty ▪ Stripper inlet temperature ▪ Number of stages of absorption columns ▪ Number of stages of stripping columns.	▪ Capture cost	[112]
		▪ Reboiler duty ▪ Reboiler pressure ▪ Condenser duty ▪ Flue gas pressure ▪ Flue gas temperature ▪ Flue gas flowrate	▪ Capture rate ▪ System energy requirement ▪ CO_2 purity	[659]
	ANN and support vector machine (SVM)	▪ inlet flue gas flowrate ▪ inlet CO_2 concentration ▪ inlet flue gas pressure ▪ inlet flue gas temperature ▪ lean solvent flowrate ▪ MEA concentration ▪ lean solvent temperature	▪ CO_2 capture level	[660]
	RL	▪ Lean MEA flowrate ▪ Reboiler heat duty	▪ Capture operating cost	[661]

Table E.1 Selected studies on data-driven-assisted solvent-based carbon capture process design/control optimization (Cont'd)

Model type	Model	Inputs (Design variables)	Outputs (KPIs)	References
Statistical models	Response surface methodology (RSM)	▪ Liquid volume in the absorption column ▪ The alkaline concentration ▪ Temperature ▪ Gas hold-up	▪ CO_2 removal efficiency	[662]
		▪ Inlet solvent flowrate ▪ MEA concentration ▪ Inlet solvent temperature	▪ CO_2 removal efficiency ▪ Mass transfer coefficient	[663]
		▪ Liquid-to-gas ratio ▪ Solvent concentration ▪ Rich loading ▪ Lean loading ▪ Stripper inlet temperature ▪ Condenser duty ▪ Recycle cooling duty ▪ Capture rate	▪ Reboiler duty	[664]
		▪ Height of cooler I ▪ duty of cooler I ▪ height of cooler II ▪ Duty of cooler II (D)	▪ CO_2 absorption efficiency ▪ Total cooling duty	[665]

Table E.1 Selected studies on data-driven-assisted solvent-based carbon capture process design/control optimization (Cont'd and end)

Model type	Model	Inputs (Design variables)	Outputs (KPIs)	References
Statistical models	Taguchi design	▪ Inlet flue gas temperature ▪ Absorber column operating pressure ▪ Amount of exhaust gas recycle ▪ MEA concentration	▪ Specific energy requirements	[666]
	RSM and Box-Behnken design (BBD)	▪ Inlet solvent temperature ▪ Solvent flow rate ▪ Gas flow rate ▪ Reboiler heat load ▪ MEA concentration ▪ CO_2 concentration	▪ Absorption percentage ▪ Overall gas phase mass transfer coefficient ▪ Reboiler specific heat duty	[667]
		▪ Lean MEA loading ▪ Flue gas inlet temperature ▪ Temperature of rich MEA entering the stripping column ▪ Reboiler pressure	▪ Capture cost	[668]
		▪ MEA concentration setpoint ▪ Reboiler temperature setpoint ▪ MEA Buffer tank temperature setpoint	▪ Carbon capture rate ▪ Absorption rate ▪ Energy performance index	[669]

APPENDIX F ARTICLE 5: LEARN-TO-DESIGN: REINFORCEMENT LEARNING-ASSISTED CHEMICAL PROCESS OPTIMIZATION

Eslam G. Al-Sakkari (Eslam Ibrahim), Ahmed Ragab, Mohamed Ali, Hanane Dagdougui, Daria C. Boffito and Mouloud Amazouz

Article Submitted on December 15, 2023, Accepted on May 22, 2024 to be presented in FOCAPD 2024, Published Online on July 10, 2024 in Systems & Control Transactions, PSE Press, Volume 3, 2024

Online Publication date: July 2024

DOI: https://doi.org/10.69997/sct.103483

Abstract

This paper proposes an AI-assisted approach aimed at accelerating chemical process design through causal incremental reinforcement learning (CIRL) where an intelligent agent is interacting iteratively with a process simulation environment (e.g., Aspen HYSYS, DWSIM, etc.). The proposed approach is based on an incremental learnable optimizer capable of guiding multi-objective optimization towards optimal design variable configurations, depending on several fac-tors including the problem complexity, selected RL algorithm and hyperparameters tuning. One advantage of this approach is that the agent-simulator interaction significantly reduces the vast search space of design variables, leading to an accelerated and optimized design process. This is a generic causal approach that enables exploring new process configurations and provides actionable insights to designers to improve not only the process design but also the design pro-cess across various applications. The approach was validated on industrial processes including an absorption-based carbon capture, considering the economic and technological uncertainties of different capture processes, such as energy price, production cost, and storage capacity. The design of the considered process was accelerated with a significant cost reduction of 5.5%. In particular, the optimal capture cost was achieved after 25 episodes in the second iteration. The proposed approach paves the way for accelerating the adoption of decarbonization technologies (CCUS value chains, clean fuel production, etc.) at a larger scale, thus catalyzing climate change mitigation.

Introduction

Climate change has become a crucial problem, where massive efforts are underway to mitigate its disastrous outcomes. In this regard, there is a major trend to develop efficient carbon capture technologies, recognized as one of the promising solutions to the challenges posed by climate change. Furthermore, the research is accelerating in this hot field due to the clear and prioritized need for developing cost-efficient and environmentally friendly carbon capture processes. Various methods based on mechanistic models and combined with traditional operation research methods, including linear and non-linear programming, are used to optimize existing carbon capture processes. Nevertheless, despite their robustness, most existing methods for optimizing process design suffer from certain limitations that hinder the development and optimization of carbon capture systems. These limitations include time consumption, high computational costs, and a lack of adaptability and transferability.

Therefore, there is a pressing need for the development of new methodologies to help overcome the abovementioned limitations. Artificial intelligence (AI) and its related machine learning (ML) technologies holds significant promise to address this critical gap. Specifically, reinforcement learning (RL), as a prescriptive machine learning technique, offers an effective approach due to being a learnable optimizer. In addition, its application has been successful in various fields, including process control [29], [579] which underscores its potential for accelerating chemical process design. RL encompasses various types, ranging from model-free to model-based algorithms, as well as policy-based and value-based agents. Examples of policy-based RL agents include deep Q-network (DQN) [580] and proximal policy optimization (PPO) [581]. Besides, there are agents that combine both policy-based and value-based calculations, such as deep deterministic policy gradient (DDPG) [582]. These diverse types offer versatile tools for enhancing chemical process design acceleration.

Recently, several research studies have employed RL within the realm of chemical engineering to optimize process sequences and to suggest new process flow diagrams for simple processes [583]–[589]. For instance, in [587], the authors used a hierarchical deep RL (DRL) agent to automate the synthesis of flowsheets for the ethyl tert-butyl ether (ETBE) production process. The study adopted the *SynGameZero* approach outlined in [590]. The Soft Actor Critic (SAC) algorithm was used in [586] to design a distillation train to separate a mixture of benzene, toluene, and *P*-xylene. The

SAC agent interacts with a Distillation Gym environment that interfaces with *COCO* open-source simulator. The paper [591] conducted a survey on the RL-assisted flow sheet synthesis, highlighting the interaction with other simulators, i.e. DWSIM, and the concept of transfer learning for process design [585]. Nonetheless, some of these attempts are still at early stages as the agents developed suggest technically infeasible flowsheets in some cases due to the lack of reasoning for the trained agent. Besides, one major limitation of these studies is the absence of causality. Analyzing the historical data gathered during the optimization process can offer new insights to the designers, enabling them to identify the causal variables that impact selected key performance indicators (KPIs), e.g. costs. Extracting knowledge from causality analysis part will lead to more generalizable results and provide insights into hidden information and relationships that can be transferred to enhance other processes.

The novelty of the current work lies in the introduction of causal DRL to accelerate the design of chemical processes. Two systems were considered for illustration: 1) ammonia-water system (hypothetical physical absorption-desorption process as a simplified example) and 2) mono-ethanolamine (MEA)-based carbon capture process. Introducing the concept of causal reinforcement learning to process design, specifically in applications related to solvent-based capture processes, can enhance the agent's reasoning capabilities, and enrich the knowledge of design experts with new insights. This approach can be generalized to other systems, under the full supervision of human experts, to streamline and accelerate the design process. The specific research objectives of this study are outlined as follows: 1) Develop an incremental DRL (IDRL) simulation-based optimization methodology to accelerate the design of steady state processes; 2) Integrate a newly developed causality analysis methodology, employing various algorithms with the IRDL to capture the underlying cause-effect relationships among different process variables and the specified KPI(s); 3) Deploy the developed IDRL coupled with causality analysis methodology to optimize different processes, including absorption-based carbon capture process utilizing amine solvents.

Methodology

General DRL simulation-based optimization approach

The general simulation-based optimization approach and its underlying methodology achieving its objectives is depicted in Figure F.1. A DRL agent is linked with a simulation environment in an

interactive and learnable manner that minimizes the search space. Within this approach, the term "rewards" is used to denote the key performance indicators, whereas the term "states" is used to represent the observations defining the current condition of the process. The agent provides ”actions” (representing design variables including equipment sizing and operating conditions) to the simulation environment based on the current state and the observed reward. The agent tries to find the optimal point without sticking in local minima where it may follow the shortest path, depending on factors such as problem complexity, reward formulation, the RL algorithm selected, and available computational resources. Unlike most of the traditional optimizers, DRL-based optimization is driven by the concept of reward and penalty. The agent receives a reward when following the right path of optimizing the objective function, i.e. minimizing the process costs, or minimizing its environmental impact. Conversely, a penalty is incurred if the agent followed the wrong direction, where the actions result in cost or environmental impacts exceeding specified values. To avoid settling in local minima, the agent posses the ability to explore new scenarios, aiming to identify paths leading to the global minimum point. Balancing between exploitation and exploration is crucial to speed up achieving the optimal solution. Additionally, achieving an accelerated optimization requires proper formulation of reward function and optimization of the agent’s hyperparameters.

DRL and problem formulation as Markov Decision Process.

From a mathematical standpoint, when employing RL, the problem is typically formulated as a Markov Decision Process (MDP). This formulation involves a set of observations that characterize the state of the environment, along with a distribution of initial states. Additionally, it entails a set of actions, whether continuous or discrete, that can modify the state of the environment based on rewards received by the agent. An MDP comprises transition dynamics or probabilities, which describe the relation between states and actions. Two other important components are the immediate rewards and the discount factor (which ranges from 0 to 1), a lower discount factor places greater emphasis on the immediate reward. For a more detailed explanation of the mathematical formulation of RL and MDP, refer to this study [580]. It is worth noting that one of the primary differences between dynamic and steady-state problems lies in the definition of discount factor during MDP formulation. The discount factor is an optional element in defining an MDP and can be disregarded (assigned a zero value) when considering the optimization of steady-state processes, where immediate rewards take precedence over long-term ones. Disregarding the

discount factor in case of steady-state processes facilitates the design optimization problem definition/formulation. Nevertheless, finding the optimal policy based on the cumulative reward for the RL agent remains relevant. This design problem adheres to the Markov Property, resulting in a memoryless agent, where the current state depends only on the previous state and influences the selection of the subsequent actions that modify the successive states. Consequently, the transition dynamics or probabilities over time and the whole states history will not be taken into consideration. The transition probabilities between two successive states based on the proposed or taken actions are primarily deterministic. In chemical process design, these transitions may involve thermodynamic/equilibrium relations, kinetic/rate equations, or heuristic rules.

RL Algorithm selection & development of the simulation environment.

Various RL algorithms can be developed, optimized, and then utilized for process design optimization. As previously mentioned, the PPO and SAC algorithms were employed process flowsheet synthesis. In this work, we consider the DDPG algorithm [582], which is one of the promising algorithms. It combines the value-based and policy-based optimization principles.

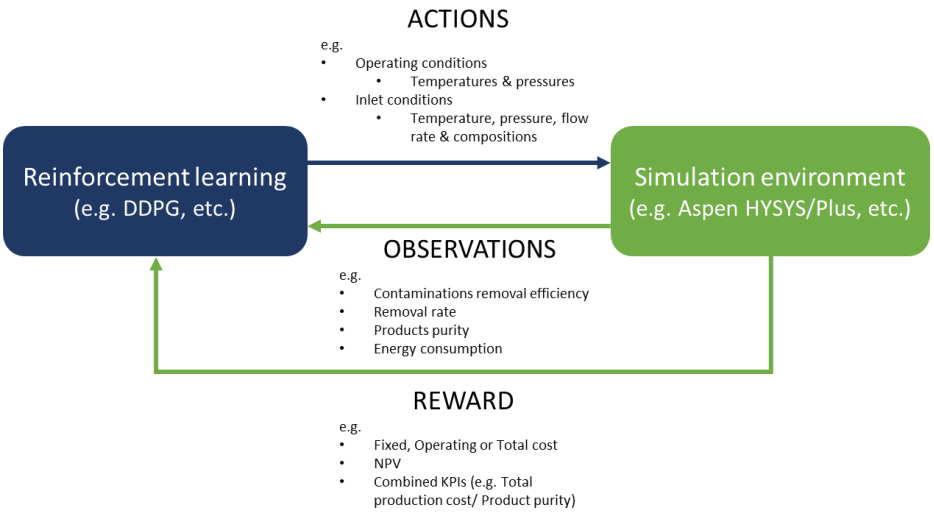


Figure F.1: General simulation-based optimization approach

Upon developing the simulation environment, the potential actions (design variables) under consideration include equipment sizing and process operating conditions. Additionally, parameters such as recycle rate, purge percentage and flowrates of inlet streams are deemed significant actions to be prescribed or optimized. Equipment sizing varies according to the equipment type. For instance, the number of stages is a crucial variable for separation columns, defining their heights

and costs. Continuous stirred reactors (CSTRs) are sized by volume, which is also related to the percentage conversion of limiting reactants. Heat exchange equipment sizing is determined by the heat exchange area, which is also related to other process operating conditions. Process operating conditions encompass different streams/equipment temperatures and pressures. Key observations describing the simulation environment's state include the purities of specified product streams/components, recovery rates of components of interest, and utilities consumption. These states and actions vary from a problem to another depending on the specific requirements of the problem and the specifications set by the process expert.

Reward definition (formulation)

The reward function can take various forms, depending on the specified objective defined by the process expert. For instance, it can be formulated as the relative production cost, i.e. production cost per kg of product or material. Alternatively, it could represent the total capital investment (TCI), net present value (NPV) or any other form of economic KPIs. Furthermore, it might be a combined value of both techno-economic KPIs such as relative production cost per material/product recovery or purity. One objective function that can serve as a highly relevant reward function is the total environmental impact or the relative emissions per kg of the product. This function holds significant importance in current and future research studies for process design optimization, enabling the assessment of processes' environmental footprints and promoting the development of eco-friendly processes.

Incremental learning

It is worth noting that our approach for the DRL-based optimization is built upon the concept of incremental learning. Specified costs, environmental impacts and actions ranges change from one case to another until the maximum allowable performance is achieved, i.e. lowest possible relative production cost and minimum acceptable environmental impact. More specifically, the actions ranges are narrowed or redefined after analyzing the historical data collected from previous RL-assisted optimization cases. In addition, reward function and its value are redefined based on this analysis. For example, the specified cost value may be reduced to explore the agent's ability to surpass it and reach the minimum allowable cost.

Towards causality analysis for causal and rational DRL

Following the incremental learning-based optimization, historical data representing the simulation runs resulting from the agent-simulation interactions are stored in spreadsheets for the causal analysis, where each run represents a data point (observation). This is to gain more insights into the root causes and the importance of design variables impacting the KPIs. Three methods were employed to perform a preliminary causality analysis based on observations. These methods include combined interpretable ML (IML)-Explainable AI (XAI), Granger causality (GC), and Bayesian causal discovery & inference.

Combined IML-XAI

Initially, the decision tree classification algorithm (i.e. IML) is combined with XGBoost and SHAP explanation algorithm (i.e. XAI) to extract the patterns, assess variables importance and analyze the relationships governing the dependence of the specified KPI (reward) and observations (state) on the design variables (actions). The IML extracts the patterns (ranges of design variables) that lead to processes with reasonable feasibility. Additionally, it assesses the variables importance, determining the most significant ones. The XAI gives relative magnitudes and directions of the effect of each input variable on the final KPI (process overall production cost). This will help afterwards in the incremental learning procedure, helping to select/redefine the relevant input variables ranges for guided optimization.

Granger causality (GC)

This method is used preliminary to validate the relations between the actions (especially the most significant ones identified by the combined IML-XAI method) and the KPI. Since actions are provided continuously by agent and they depend on the previous set of actions and states, consequently actions, states, and rewards/KPIs are treated as a form of time series data, leading to the application of GC. As mentioned earlier, in case of optimizing steady-state processes, the current state (representing the current process configuration) is solely influenced by the preceding one, rendering longer-term relationships negligible. The dependence between each successive pair of states (St & St+1) and their associated set of actions based on the reward provided to the agent, justifies the utilization of approximated time series assumption to apply GC.

Bayesian causal discovery & inference

This method is used to generate causal graphs to discover the root causes that significantly affect changes of the KPI. It also reveals the interdependence among actions, discovering the root cause

design variable that affects both other design variables and KPI. Figure F.2 summarizes the proposed causality analysis methodology. Further details regarding the combined IML-XAI method can be found in our recent study [471]. The insights gained from this causality analysis will upgrade/enrich the experts' knowledge, where they can redefine the baselines/rewards and the range of actions to retrain the agent, thereby fostering further improvement in design optimization.

This methodology can be applied in flowsheet synthesis to enhance the DRL agent's reasoning capability by providing it with the extracted knowledge and causal graphs. This empowers the agent to operate as a rational causal DRL agent that avoids the synthesis of technically infeasible flowsheets/processes, thus accelerating the synthesis and design process. In addition to the other design variables, equipment and their sequencing will be added to the list of actions. Consequently, the causal graphs will account for these two crucial categorical variables and quantify their impact on the defined KPI/reward. Furthermore, integrating domain knowledge from process designers and operators as constraints into the causal DRL agent's structure will enhance its rationality. These domain knowledge-informed constraints ensure the avoidance of invalid streams/equipment selection, placement, and sequencing, e.g. separation train sequencing [670] and chemical reactors sequencing [671]. This will further shorten the time needed for the agents to converge on the optimal new process, thus ensuring the desired ML-assisted design acceleration.

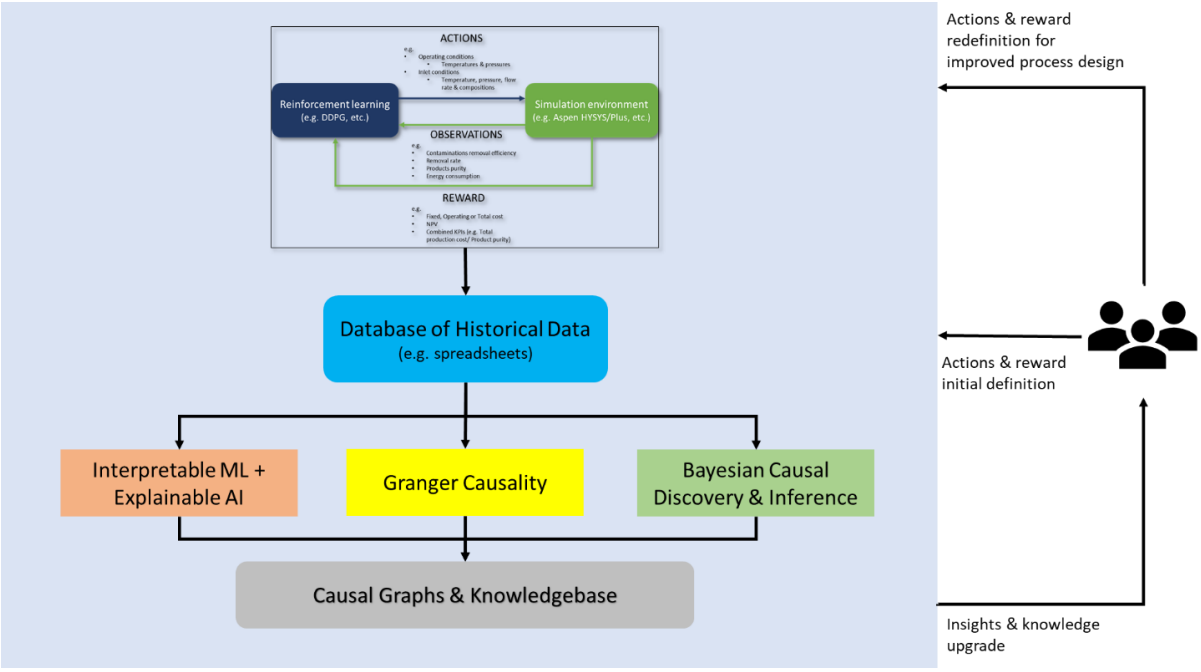


Figure F.2: Summary of Causality Analysis Methodology

Simulation case studies

Aspen HYSYS was selected as the simulation software for constructing the process flow diagrams and conducting the process design. Two simulation case studies were developed to validate the proposed CIRL approach. The simplified process flow diagrams of the two case studies are presented in the supplementary materials and their details are summarized in the following subsections.

Case study 1: Ammonia water system

An ammonia-water absorption-desorption system was developed with a stream of air contaminated with ammonia is introduced to the absorption column, where pure water is utilized as a physical solvent to separate ammonia from air. The captured ammonia is then removed from the water solvent using steam in a re-boiled stripping column. Heat exchange between the stream entering the stripping column and its bottoms product stream enables heat integration, minimizing the reboiler heat load. A condensation and a flash separation are applied for the top product stream of stripping column to remove any water associated with the separated ammonia, thus enhancing its purity and recovering any escaped solvent. Recycle was not considered in this simplified case study. The NRTL simulation fluid package was used. The actions considered include number of stages in both absorption and stripping columns, temperatures of fresh solvent and inlet contaminated gaseous stream, and the pressure of the inlet contaminated gaseous stream.

Case study 2: Mono-ethanolamine (MEA)-based carbon capture system

This case considers chemical absorption (CO_2-MEA system), where flue gas enters the absorption column after solids removal, pressurization, and cooling to the desired operating temperature. The absorption-desorption setup is like the case of ammonia (case study 1). However, in this case, lean MEA is recirculated, after carbon dioxide stripping, to the absorption column and mixed with a makeup stream to enhance the process feasibility. A purge stream is used to avoid excessive accumulation of carbon dioxide in the system and the recycled lean MEA stream. Effect of temperature and pressure of inlet flue gas stream and temperature of makeup solvent were considered as actions. Effect of recycle percentage and temperature of stream entering the stripping column were also considered for optimization. The *Acid Gas - Chemical Solvents* Aspen built-in fluid package was used to perform the simulation calculations. The stream of flue gases comes from a power plant, where all conditions defined according to that study [672]. A baseline/reference

for the carbon dioxide recovery cost of 1.12 USD/kg was adopted from a published study [661] that ultimately achieved the lowest cost after several iterations.

Reward definition

Important observations that describe the state of simulation are purity of purified gas stream, recovery of contaminants (for ammonia and carbon dioxide), purity of the product stream, and utilities consumption. The reward function is defined as the relative removal cost, i.e. removal cost per kg of captured material. The procedure for calculating the removal cost and other related costs follows the guidelines outlined in this textbook [632] and these two studies [489], [633]. Utilities and MEA prices were sourced from a study published in [673]. The reward was determined as follows: if the agent progresses in the specified direction (new cost <= specified cost), it receives a value equivalent to the total production cost (unity factor); otherwise, it receives a zero value. Furthermore, if the new cost in current step is lower than the cost of previous step, the agent receives the full total production cost value; otherwise, it receives zero.

Results and discussion

Optimized DRL agent

The agent developed and adopted in this study is based on DDPG, a policy-based model-free reinforcement learning algorithm. It comprises two neural networks, one for the actor and one for the critic. For a deeper understanding of the mathematical underpinnings of the DDPG, interested readers can refer to [29], [674], [675]. The structure and crucial hyperparameters of the optimized DDPG agent are summarized in Table F.1.

The *Tanh* activation function was selected to standardize the range of actions between -1 to 1, which are then converted back to their real values before being used in the simulation.

Table F.1 DDPG optimized hyperparameters

Hyperparameter	**Value**
Action noise standard deviation	0.2
Actor learning rate	0.001
Critic learning rate	0.002
Discount factor (gamma)	0
Tau parameter used to update target networks	0.005
Total number of episodes	100-1000
Activation function for the critic neural network dense layers	ReLU
Activation function for the actor neural network inputs dense layers	ReLU
Activation function for the actor neural network outputs dense layer	Tanh

Process design optimization

Ammonia-water system

For the ammonia-water absorption system, the initial range for both inlet solvent and flue gas streams temperatures was set between 20 to 100 ^{o}C, with pressure ranging from 600 to 950 kPa. It is important to mention that the absorption column pressure was set equal to the inlet pressure of flue gas. This simplification facilitates the manipulation of the absorber pressure in the simulation environment by the RL agent. It is important to note that the inlet flue gas pressure must be higher than that of the absorber to ensure propre flow inside the column. Hence, a pressure drop value will be added to specify the pressure of inlet flue gas to the designer to incorporate it in real life operations. Additionally, the range of number of stages for both absorption and desorption columns was from 3 to 15. The initial desired cost (baseline) had the value of 2 USD/kg of ammonia based on a simulation of base case. As a result, after 220 episodes, a cost of almost 0.880 USD/kg of

ammonia was obtained initially, which is significantly lower than the specified baseline. Accordingly, following the proposed incremental learning method, this value was given as a new baseline to the agent for further improvement of the process economic optimization, where the ranges of variables were the same. Surprisingly, the agent could reach another lower production/recovery cost of 0.775 USD/kg of ammonia. The preliminary causality analysis methodology was applied to extract some other knowledge about the most effective variables and the patterns to achieve lower process costs. Thus, the historical data from these two first optimization iterations was collected, preprocessed, and analyzed. The preprocessing was done through dividing the observations to high cost (above 0.9 USD/kg of ammonia) observations and labeled as Class (0). The other portion was labeled as Class (1), which represented the low cost (below 0.9 USD/kg of ammonia) observations. Figure F.3 represents the results of employing IML+XAI to get the dependency graphs.

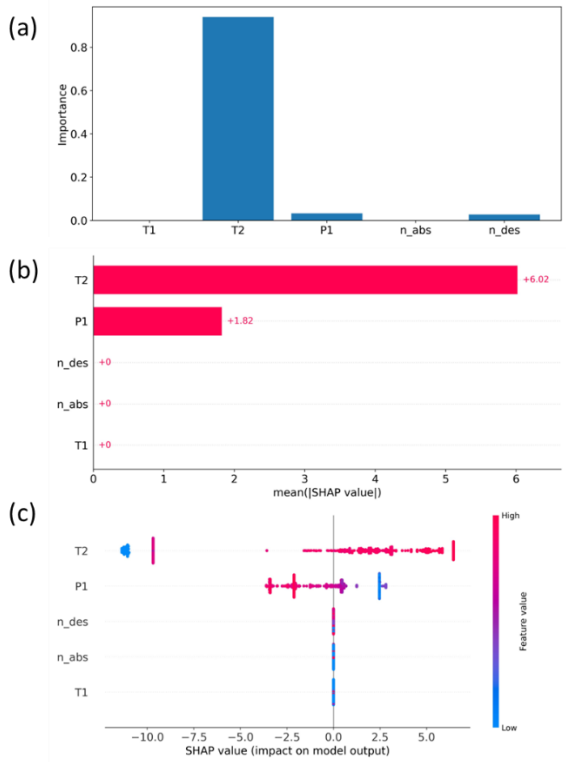


Figure F.3 Results of IML+XAI (Dependency graphs) for ammonia-water system

As shown, the most effective two variables are T2 (temperature of flue gases entering the absorption column) and the pressure of the solvent stream. These results were validated through performing the GC analysis to test the dependency of the continuous cost change corresponding to

the continuous actions given by the agent, in which the new actions depend on the previous actions, state and reward. *p*-value of the dependence between T2 and cost was < 0.001 which indicates high GC. Besides, according the XAI dependency graphs, higher values of T2 lead to lower costs (Class (1)). The variable P1 possesses clearly lower significance, where, after data preprocessing/analysis, the lower values in the range of 830-900 kPa (around 850 kPa) lead to lower costs (Class (1)). These results of the current case mean that the operating costs dominate the removal cost. In addition, upon analyzing the available data, to achieve higher purities and recoveries of ammonia, it is preferred to utilize absorption columns with low number of stages and stripping column with a higher number of stages. The causal incremental learning methodology aims at minimizing the search space to accelerate the process design optimization. Hence, after obtaining the dependency graphs, the ranges of the variables were redefined and narrowed to see if further improvement can be obtained. The new ranges are (85-99 ºC) for T2, (25-41 ºC) for T1, (800-900 kPa) for P1, (3-5) for absorption number of stages and (10-15) for stripping column number of stages. In addition, a new baseline was set to 0.75 USD/kg of ammonia. As a result, a new lower production cost of 0.719 USD/kg of ammonia was obtained at the new conditions of T1 = 40.5 ºC, T2 = 99 ºC, P1 = 900 kPa, absorption number of stages = 4.8 (almost 5) and stripping column number of stages = 15. At these conditions, the purity of outlet air and ammonia recovery were 97.2% and 96.3%, respectively. Figure F.4 summarises the results of the proposed causal incremental learning methodology for process design optimization. The Bayesian causal graph showed that the temperature of the flue gas stream is one of the root causes of changing the KPI (removal cost) and can even affect other design variables (absorption column stages (n_abs) & stripping column stages (n_des)) and state. This graph supports the results of the first two methods and is provided in the supplementary materials (Appendix K).

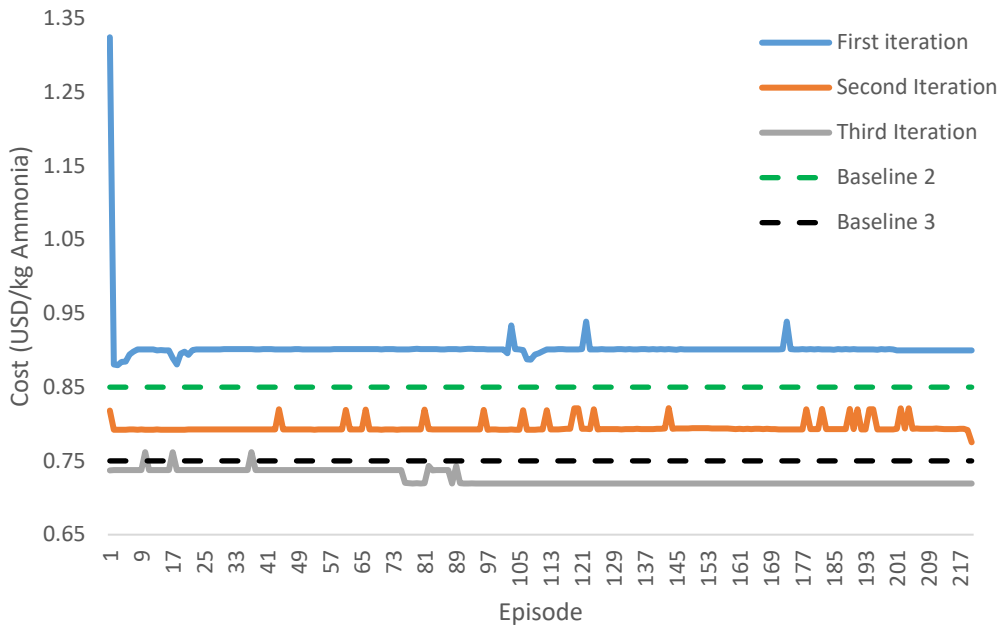


Figure F.4 Optimization results (Cost reduction along iterations and episodes) for ammonia-water system

MEA-based Carbon Capture system

In the case of MEA-based carbon capture system, the first ranges, according to literature [672], were inlet solvent and flue gas temperatures (T1 and T2) of 30-90 ^{o}C and inlet flue gas pressure (P1) of 150-500 kPa. Besides, recycle percentage ranged initially between 90% and 98%, while the range of temperatures of rich MEA stream entering the stripping column was 77-85 ^{o}C. As in the case of ammonia-water system, the same assumption of considering the absorption column pressure as the inlet flue gas pressure is still applicable. After using the first baseline of 1.12 USD/kg CO_2, this initial optimization led to a lower production cost of almost 0.190 USD/kg CO_2. The collected data was then preprocessed and labeled to the IML-XAI based low level causality analysis of the historical simulation data collected from this iteration show that the most significant variables are flue gas inlet temperature and its pressure. The GC analysis confirmed this where the *p*-value in the case of cost dependence on T2 did not exceed 0.0001. It is worth mentioning that when the test was done in the reverse direction the *p*-value exceeded 0.9; this means that T2 causes the change in the cost but not the vice versa. In addition, the *p*-values in the case of the other variables T1, Rec%, T before stripping column and the fresh solvent flowrate exceeded 0.1, which indicates low to no significance of these variables compared to T2 in the current study within the

specified range. Figure F.5 shows the results of the combined IML+XAI approach in the current case of MEA-based carbon capture system. According to these results the ranges of the variables were narrowed/redefined for the sake of exploration for further optimization through retraining the agent considering the extracted causal knowledge, which gives it more reasoning. This reasoning/causality accelerates the retraining step and the subsequent design optimization. In this regard, the ranges of the actions were narrowed/redefined to become inlet solvent and flue gas temperatures (T1 & T2) of 23-35 ºC and inlet gas pressure (P1) of 250-400 kPa in the second iteration. In addition, the new range of the recycle percentage was 85%-95% and accordingly the range of temperatures of rich amine stream entering the stripping column changed to 71-81 ºC. Furthermore, the baseline was decreased to 0.185 USD/kg CO_2. A third baseline, based on the expertise of carbon capture and process engineering experts, was put at 0.155 USD/kg CO_2 for comparison and more improvement in future research.

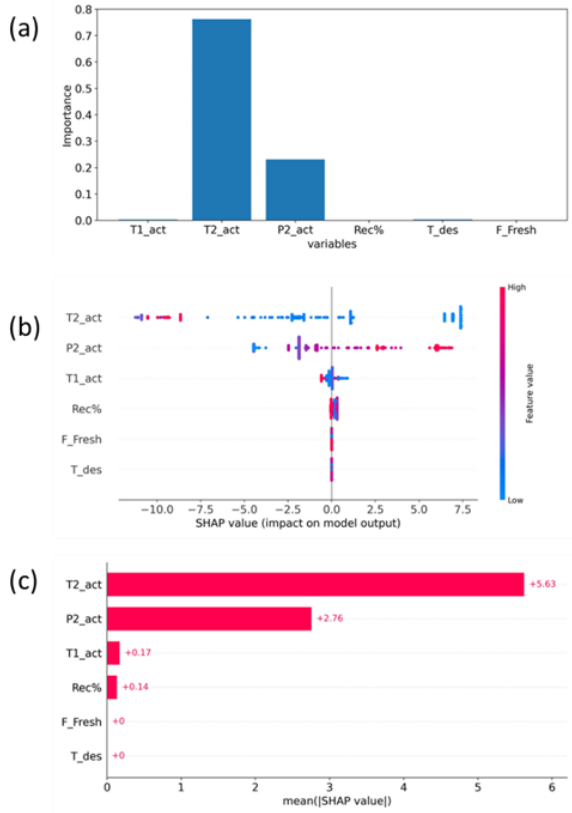


Figure F.5 Results of IML+XAI (Dependency graphs) for CO_2-MEA system

Figure F.6 summarizes the results of the two training iterations with their corresponding baselines. as shown, the minimum relative removal cost obtained was about 0.175 USD/kg CO_2. This cost

was obtained after 25 episodes during the second iteration (agent retraining step). The conditions that achieved this value were T1 of 24 ºC, T2 of 25 ºC, P2 of 270 kPa, recycle percentage of 85 % and temperature of the rich amine entering the desorption column of 71 ºC, respectively. Additionally, the carbon capture percentage at these conditions was 92.5% with a carbon dioxide purity of 98%. The causal graph based on Bayesian framework showed that one of clear the root causes of changing the KPI and can even affect other design variables and state is the temperature of the flue gas stream in agreement with what was obtained by the first two methods. This graph is put as supplementary material (Appendix K).

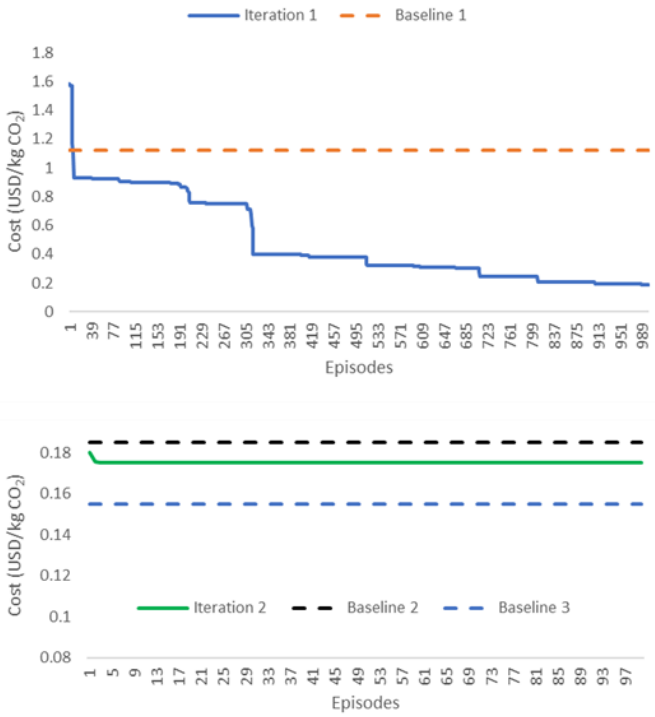


Figure F.6 Optimization results for CO_2-MEA system

Recommendations for future work

In forthcoming research, we aim to incorporate emissions and other environmental impacts as crucial optimization objectives. In particular, the net emissions should be considered as a reward, where the emissions of the capture process itself should be minimized as possible. We advocate for exploring the application of transformers-based RL for flowsheet synthesis and process design acceleration. This will allow adding more reasoning to the causal incremental learning-based RL agent to help in generating/suggesting new technically feasible process flow diagram. This will be done also through defining constraints based on process design experts' domain knowledge and

rules as a practical application of the interesting concept of RL with human preferences. This will guarantee the acceleration of process design under the full supervision of human experts to always insure simultaneous technical and economic feasibility of the generated flowsheets. Uncertainty of inlet streams composition and energy/fuel prices will be considered in our future research as well which is one of the advantages of RL to deal with uncertainties.

Conclusion

This study introduced and employed a causal Incremental Reinforcement Learning (CIRL) agent as a learnable optimizer with reasoning capabilities to accelerate efficient process design. Two case studies were considered to validate the approach, physical absorption (ammonia-water system) and chemical absorption (CO_2-MEA system). In the case of chemical absorption, the causal DRL agent reached an optimal cost after only 25 episodes. The causality analysis revealed that key variables, such as inlet flue gas temperature and absorption pressure, significantly influence the CO_2-MEA system's production cost. For instance, it showed that the relative capture cost of the process is inversely proportional to the flue gas temperature fed to the absorber. Feeding the agent with the extracted knowledge and causal information obtained from the causality analysis helped it to achieve more improved process design. These findings hold relevance beyond the studied systems and offer insights for future absorption system designs.[5]

[5] All supplementary materials related to this article are available in (Appendix K)

APPENDIX G SUPPLEMENTARY MATERIALS OF ARTICLE 1

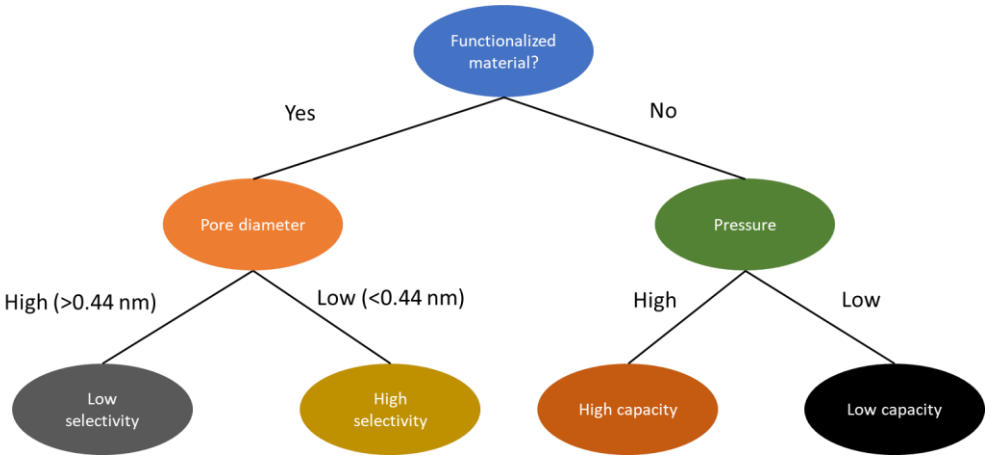


Figure G.1 Decision tree illustration

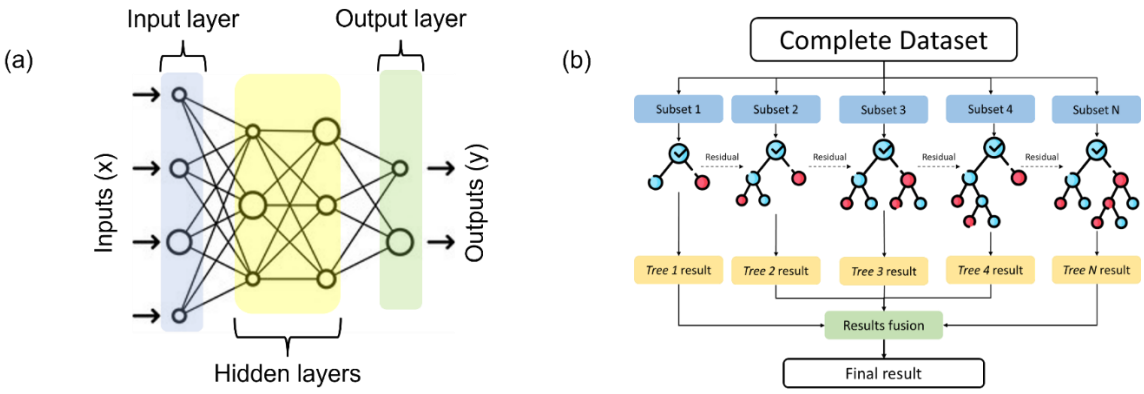


Figure G.2 Graphical representations of (a) ANN and (b) XGBoost

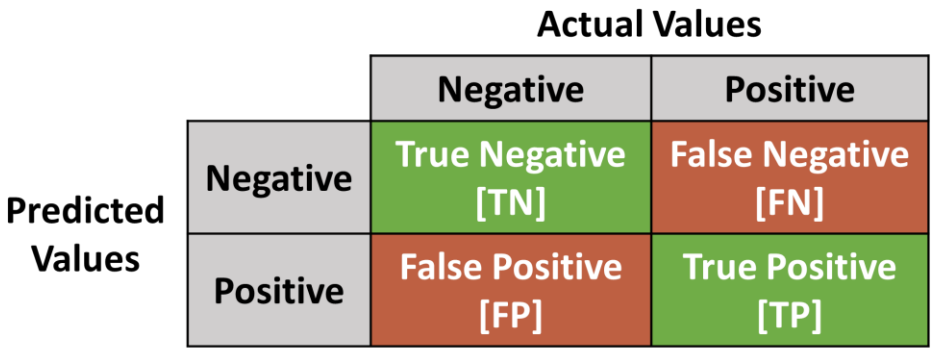


Figure G.3 General confusion matrix

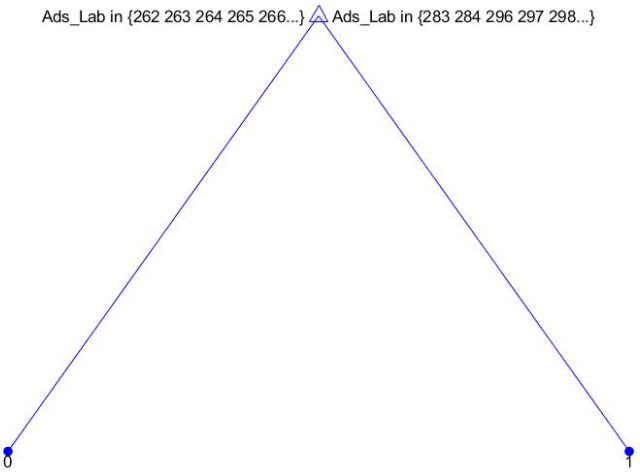


Figure G.4 Capacity classification DT for COSMOS dataset (Two patterns based on adsorbent label)

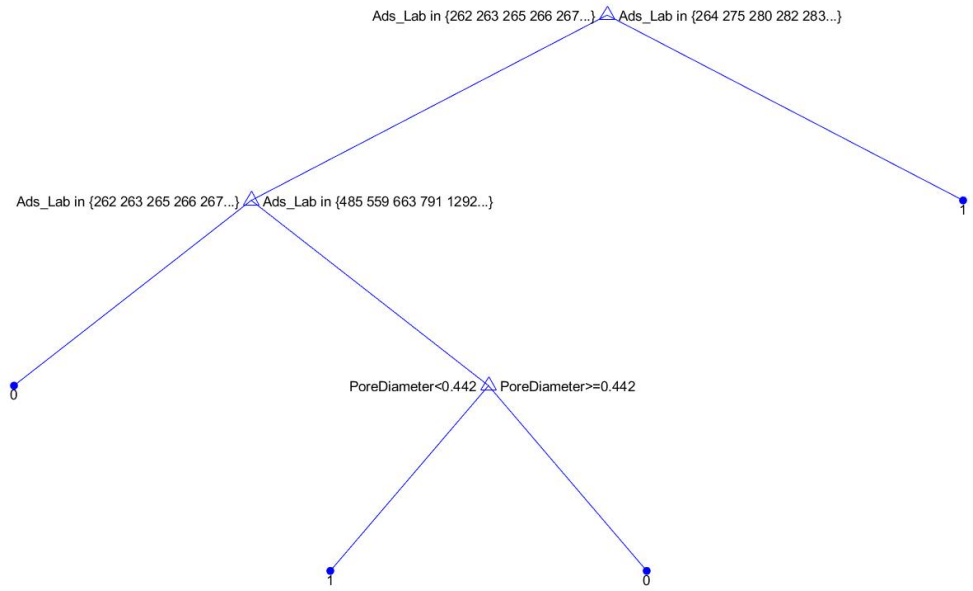


Figure G.5 Selectivity classification DT for COSMOS dataset (Four patterns based on adsorbent label and pore diameter)

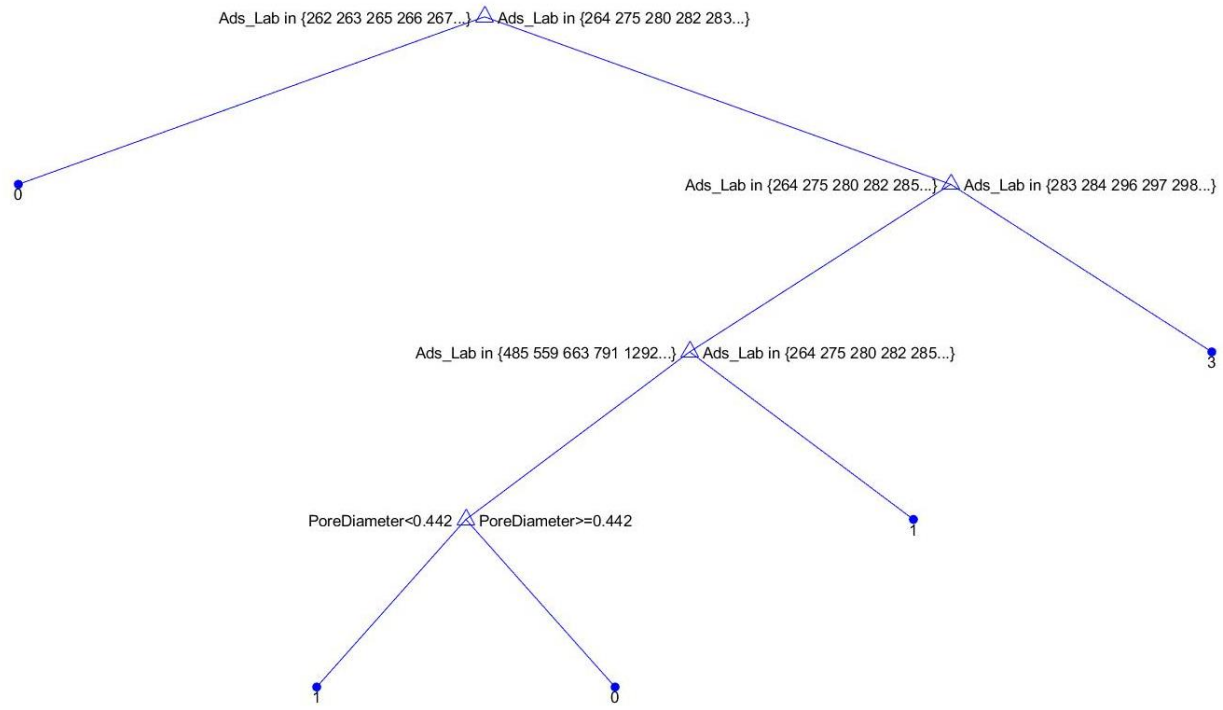


Figure G.6 Combined Capacity-Selectivity classification DT for COSMOS dataset (Five patterns based on adsorbent label and pore diameter)

APPENDIX H SUPPLEMENTARY MATERIALS OF ARTICLE 2

Table H.1 Key descriptors based on NNMF

Components of NNMF	Key Descriptors	Descriptor Score on NNMF
Component 1	BCUT2D_LOGPLOW	5.19314901
	BCUT2D_CHGLO	5.19090594
	BCUT2D_MRLOW	5.17520485
	FpDensityMorgan1	5.17439906
	MinEStateIndex	5.17077739
	VSA_EState5	5.16504564
	FpDensityMorgan2	5.16018481
	HallKierAlpha	5.15543013
	MinPartialCharge	5.15176211
	FpDensityMorgan3	5.13418228
	BCUT2D_MWLOW	5.13283145
	VSA_EState9	5.12312419
	PEOE_VSA5	5.12187283
	BCUT2D_MWHI	5.1208136
	VSA_EState4	5.12039519
	SMR_VSA2	5.11992984
	fr_nitrile	5.1195979
	fr_SH	5.11770179
	SlogP_VSA12	5.1172063
	fr_sulfide	5.11650573
	fr_term_acetylene	5.11623714
	fr_aryl_methyl	5.11556309
	fr_C_S	5.11542518
	MinAbsEStateIndex	5.114958
	fr_thiophene	5.11478889

Table H.1 Key descriptors based on NNMF (Cont'd)

Components of NNMF	**Key Descriptors**	**Descriptor Score on NNMF**
Component 2	Chi1n	1.14136999
	Chi2n	1.1383067
	Chi0v	1.1307961
	Chi0n	1.13000702
	MolMR	1.12519916
	Chi1v	1.12492772
	Chi3n	1.1122268
	LabuteASA	1.10107878
	NumValenceElectrons	1.08500496
	Chi4n	1.08206496
	Chi1	1.08001417
	MolLogP	1.07305269
	HeavyAtomCount	1.0723785
	Kappa1	1.07037957
	Chi2v	1.06481601
	Chi0	1.05950481
	Chi3v	1.05163482
	SMR_VSA5	1.03640199
	SlogP_VSA5	1.02852568
	ExactMolWt	1.01921727
	MolWt	1.01827609
	PEOE_VSA6	1.00765154
	Kappa2	0.98202774
	HeavyAtomMolWt	0.97524133
	Chi4v	0.95305644

Table H.1 Key descriptors based on NNMF (Cont'd and end)

Components of NNMF	**Key Descriptors**	**Descriptor Score on NNMF**
Component 3	SMR_VSA1	0.94045757
	NumHeteroatoms	0.89993835
	MaxPartialCharge	0.89794206
	EState_VSA10	0.87700721
	MaxEStateIndex	0.87054194
	MaxAbsEStateIndex	0.87054194
	MinAbsPartialCharge	0.86030133
	PEOE_VSA14	0.8595814
	SlogP_VSA2	0.84980554
	EState_VSA1	0.84581918
	MaxAbsPartialCharge	0.83199405
	NOCount	0.79721006
	VSA_EState1	0.7934961
	BCUT2D_CHGHI	0.78801383
	NumHAcceptors	0.77314412
	TPSA	0.77000754
	Chi0	0.76914984
	SlogP_VSA10	0.76354062
	NumValenceElectrons	0.7478514
	HeavyAtomMolWt	0.7423416
	HeavyAtomCount	0.74058984
	Kappa1	0.73293582
	ExactMolWt	0.72916723
	MolWt	0.72819776
	VSA_EState2	0.72702933

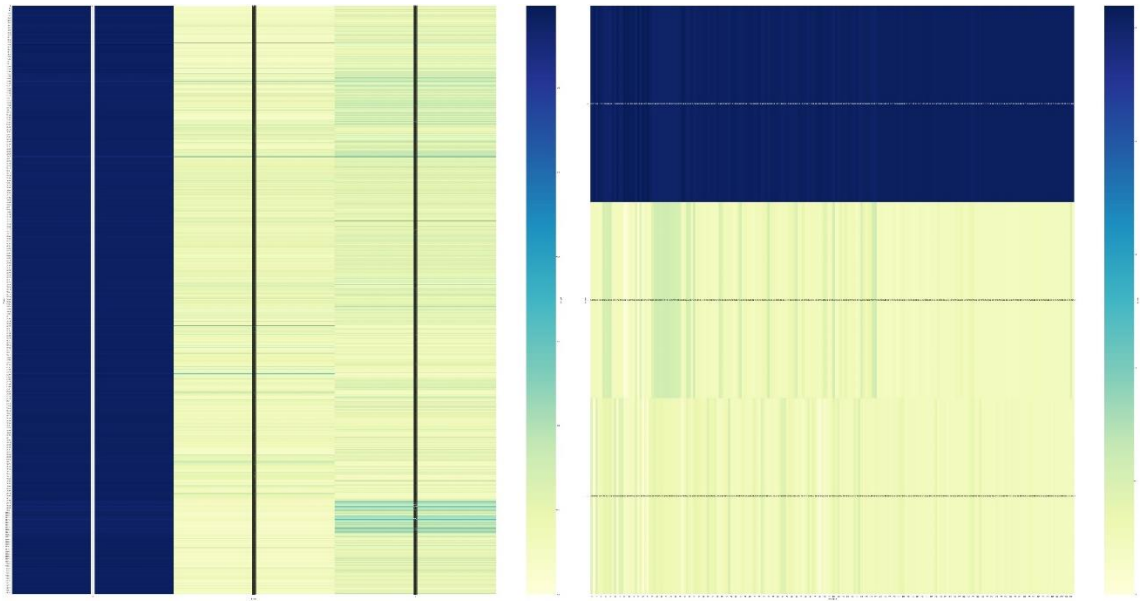


Figure H.1 NNMF results for (ndescriptors = 208 & ncomponents = 3)

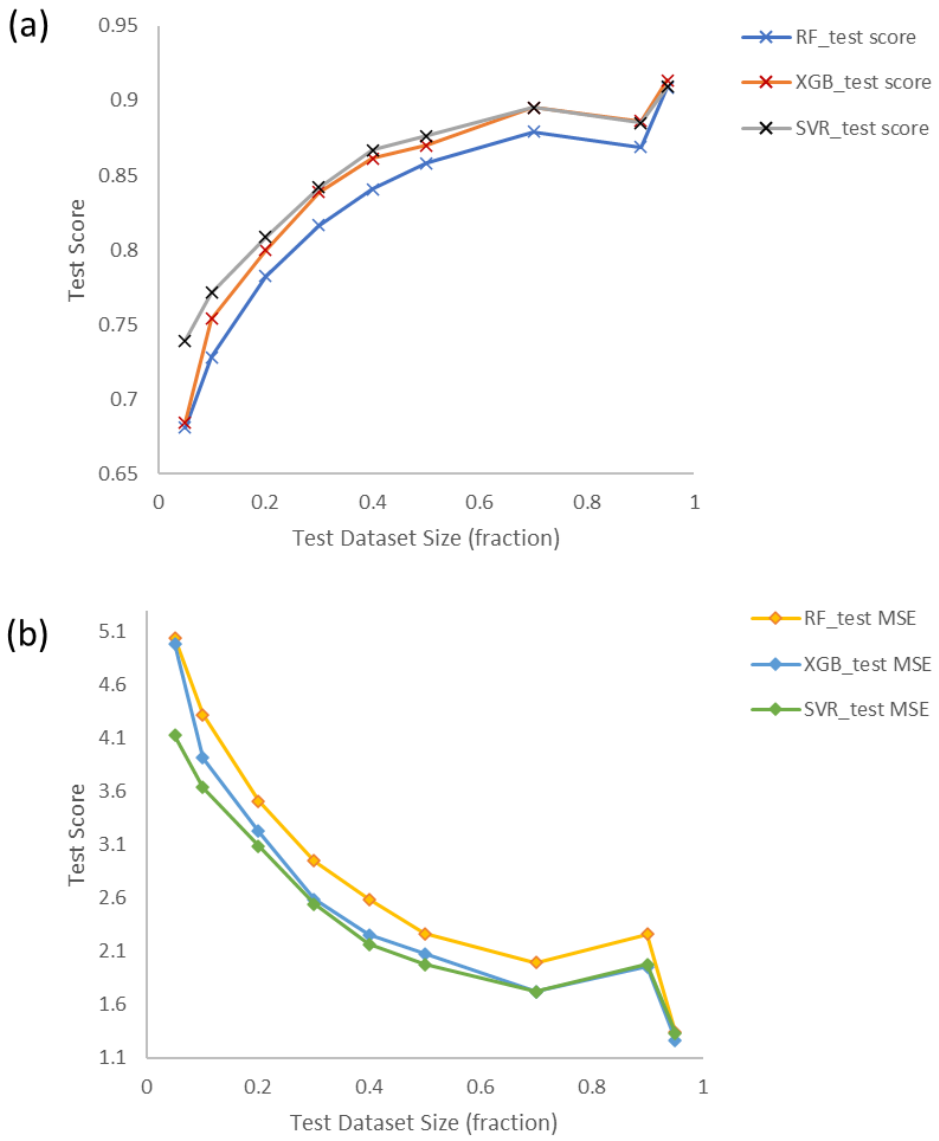


Figure H.2 Effect of dataset size on (a) the accuracy/score and (b) mean squared error of ML modeling using RF , XGB and SVR for Polarization solubility parameter

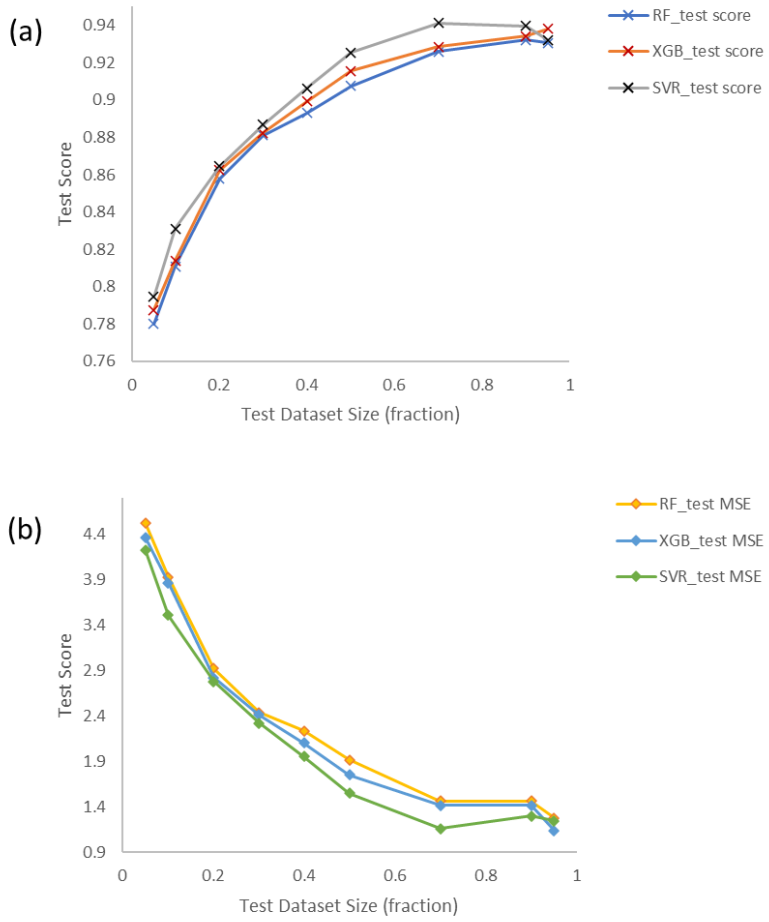


Figure H.3 Effect of dataset size on (a) the accuracy/score and (b) mean squared error of ML modeling using RF , XGB and SVR for Hydrogen bonding solubility parameter

Table H.2 Decision fusion results (Solubility_2 "δP")

Fusion No.	Non-learnable Fusion								Learnable Fusion					
	Average-based		R^2-based		MSE-based		R^2/MSE-based		XGB		ANN		SVR	
	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE
Fusion 1	0.94	0.93	0.94	0.92	0.94	0.88	0.94	0.87	0.95	0.75	0.94	0.76	0.96	0.59
Fusion 2	0.96	0.56	0.96	0.56	0.96	0.54	0.96	0.54	0.98	0.25	--------	--------	0.98	0.20
Fusion 3	0.99	0.15	0.99	0.12	0.99	0.10	0.99	0.08	--------	--------	--------	--------	0.99	0.05

Table H.3 Decision fusion results (Solubility_3)

Fusion No.	Non-learnable Fusion								Learnable Fusion					
	Average-based		R^2-based		MSE-based		R^2/MSE-based		XGB		ANN		SVR	
	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE	R^2	MSE
Fusion 1	0.96	0.75	0.96	0.74	0.97	0.68	0.97	0.68	0.97	0.45	0.97	0.55	0.98	0.44
Fusion 2	0.98	0.25	0.98	0.25	0.98	0.23	0.98	0.22	0.99	0.15	--------	--------	0.99	0.12
Fusion 3	0.99	0.10	0.99	0.10	0.99	0.09	0.99	0.08	--------	--------	--------	--------	0.99	0.03

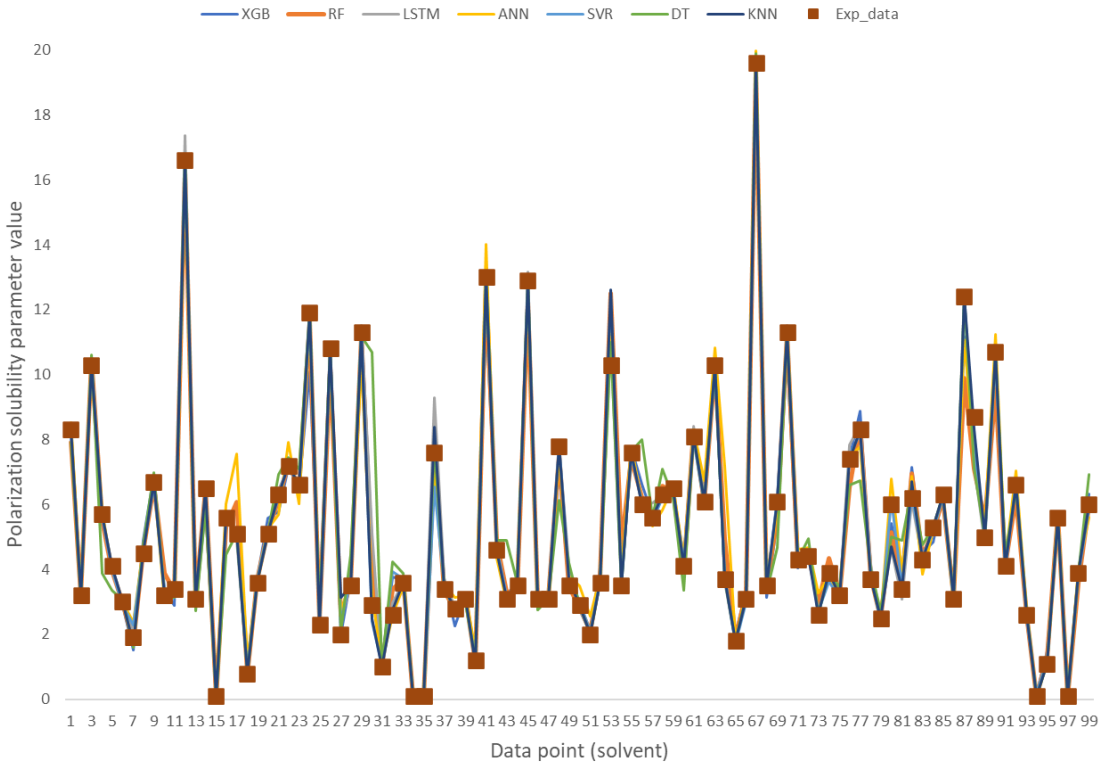


Figure H.4 Results of different individual techniques (Polarization solubility parameter)

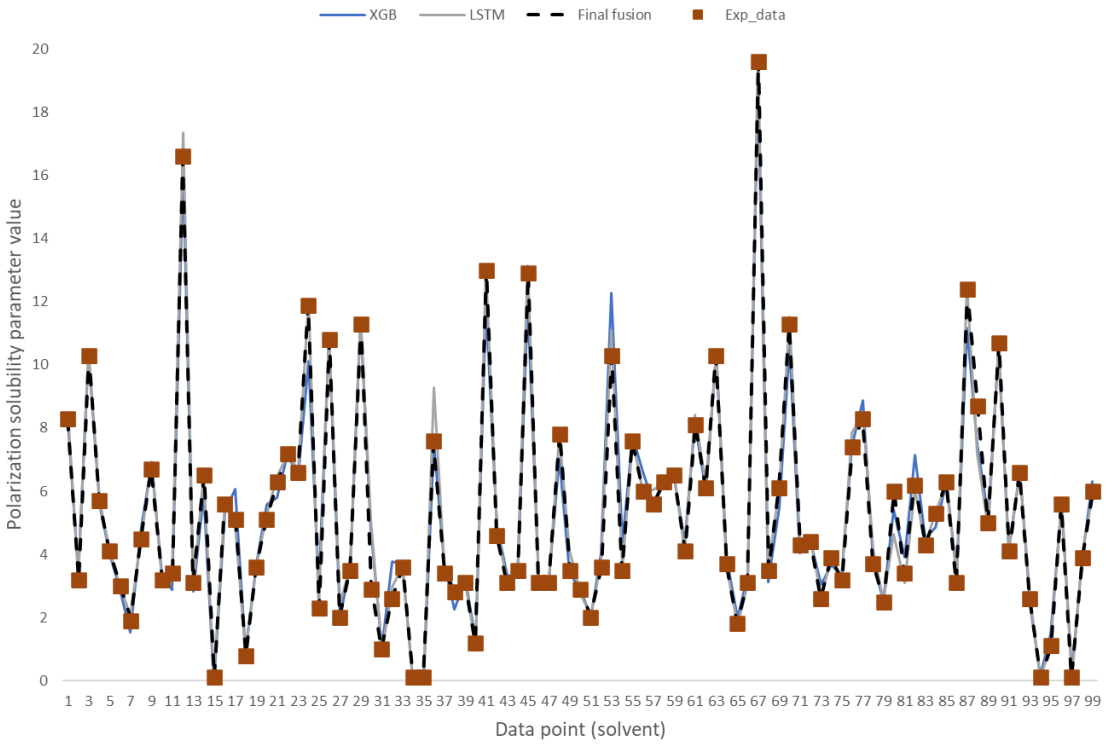


Figure H.5 Results of final decision fusion vs. selected different individual techniques (Polarization solubility parameter)

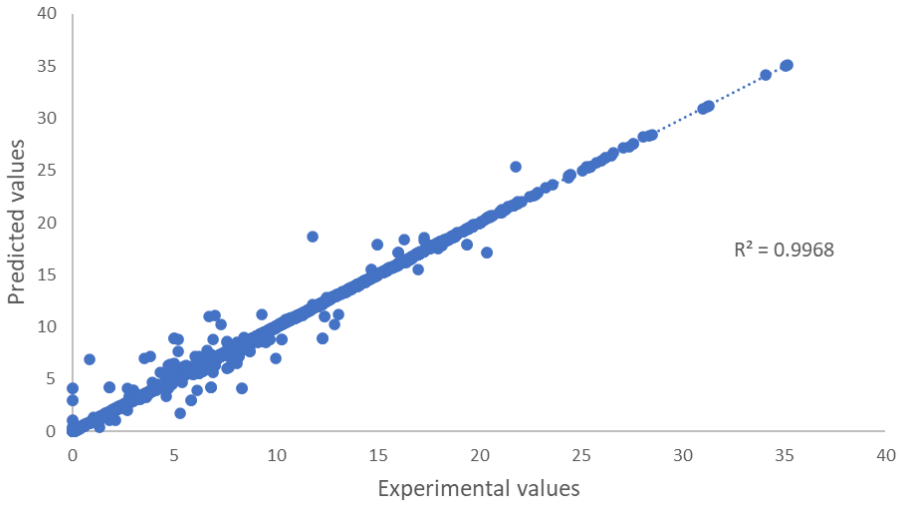


Figure H. 6 Predicted Vs. experimental values of polarization solubility parameter (final decision fusion results)

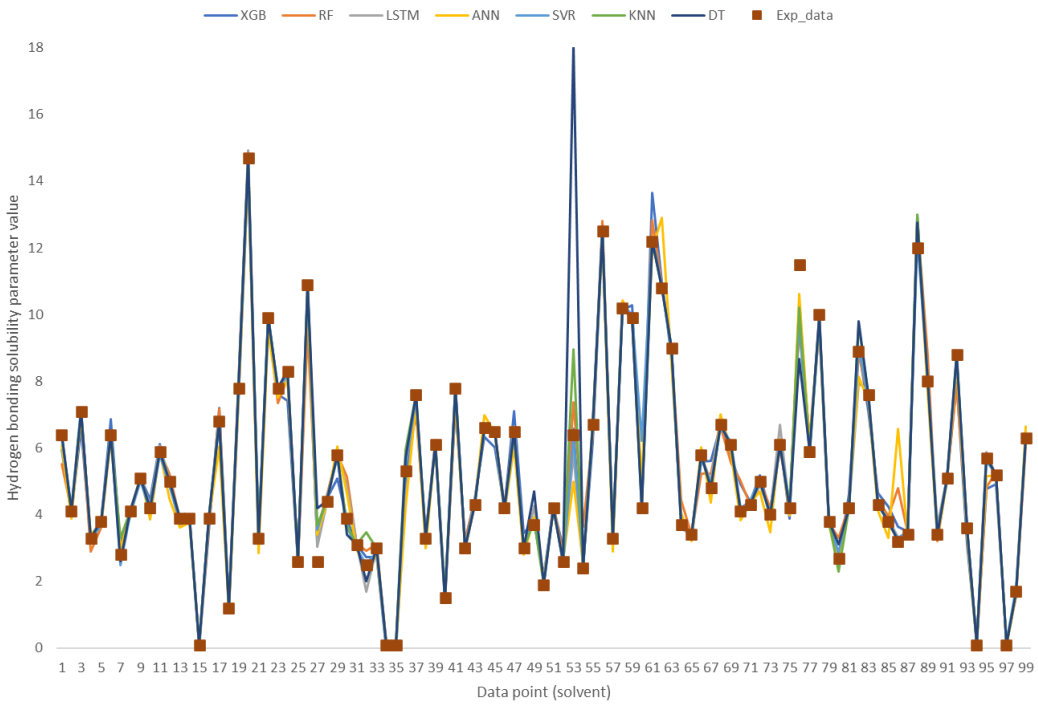


Figure H.7 Results of different individual techniques (Hydrogen bonding solubility parameter)

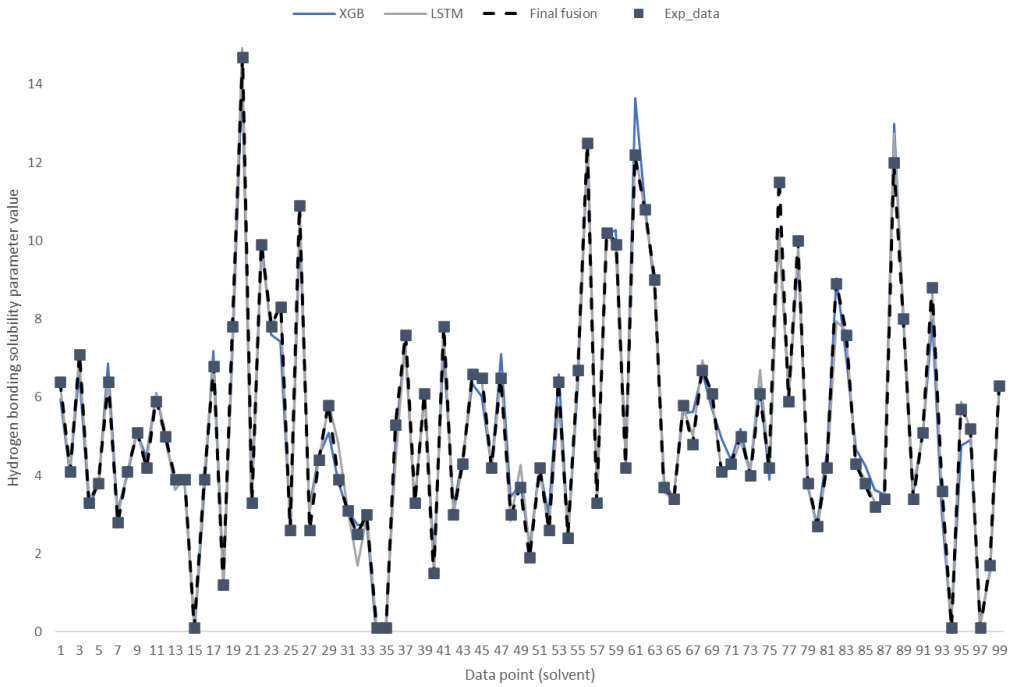


Figure H.8 Results of final decision fusion vs. selected different individual techniques (Hydrogen bonding solubility parameter)

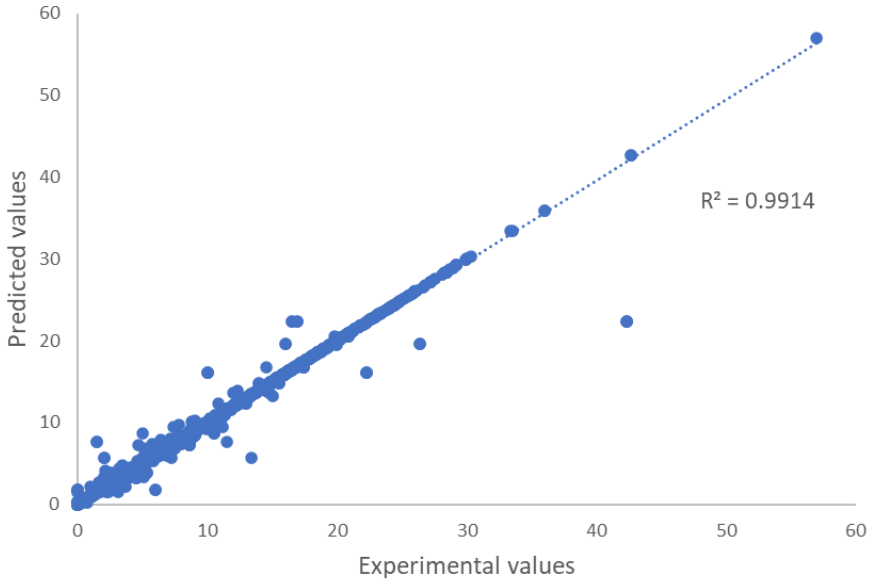


Figure H.9 Predicted Vs. experimental values of hydrogen bonding solubility parameter (final decision fusion results)

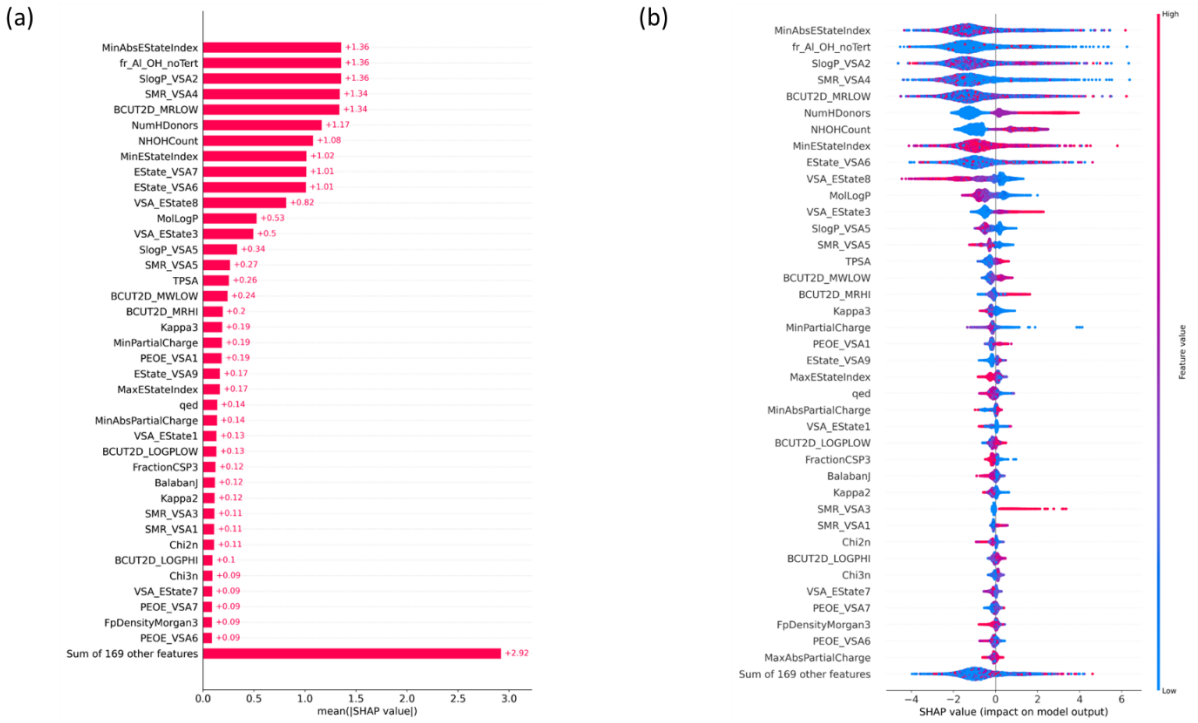


Figure H.10 SHAP values of sugar cane bagasse-based lignin solvents classification based on RED (**Descriptors**)

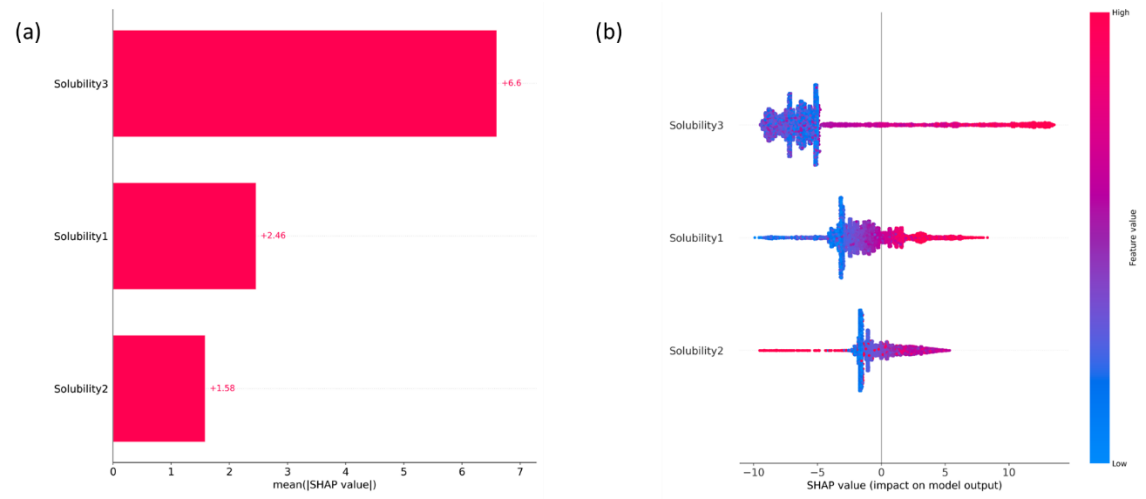


Figure H.11 SHAP values of sugar cane bagasse-based lignin solvents classification based on RED (**Hansen solubility parameters**)

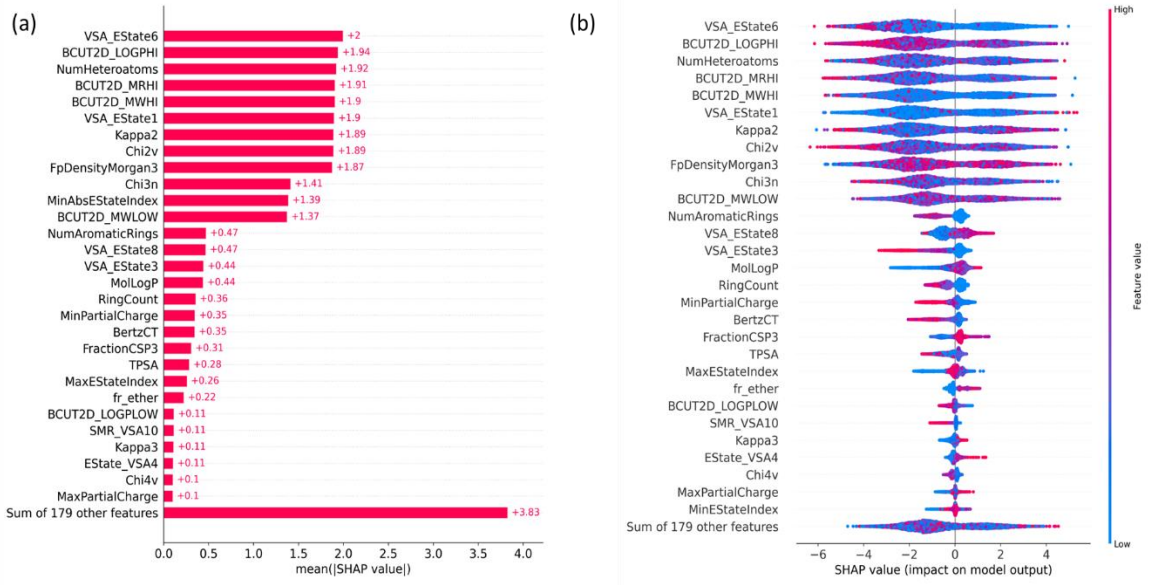


Figure H.12 SHAP values of CO_2 solvents classification based on RED (**Descriptors**)

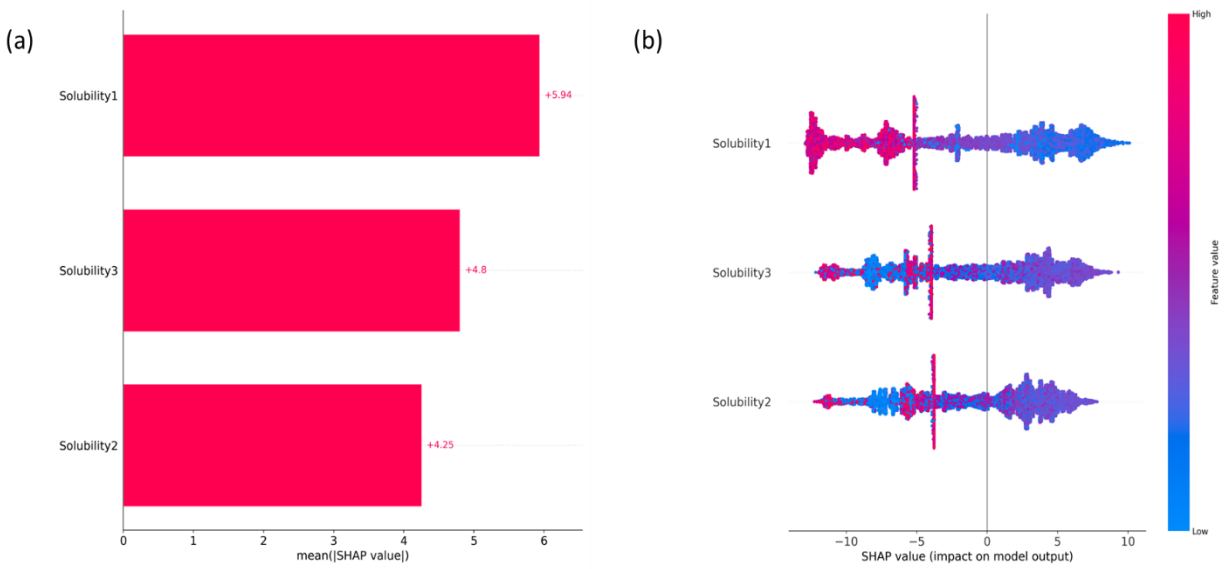


Figure H.13 SHAP values of CO_2 solvents classification based on RED (**Hansen solubility parameters**)

Table H.4 Comparison with selected previous studies predicting solubility parameters

Model(s)	**Solubility parameter(s)**	**Inputs**	**No. of dataset points**	**Metrics**	**Explainability**	**Reference**
Ensemble of several ML techniques and architectures	HSPs Range (0 to 60)	SMILES codes and extracted molecular descriptors & fingerprints	Almost 12000 points representing different solvents	$R^2 > 0.99$ MSE = 0.02	Yes	This work
GPR	HSPs Range (0 to 35)	Molecular shape & size, electrostatic forces, σ-profile, and molecular structure	193	R^2 = 0.69-0.83 RMSE = 1.02-2.83	Limited	[431]
Ensemble of tree-based models	HSPs Range (0 to 50)	Molecular weight, refractive index, boiling point, melting point, radius of gyration, van der Waals reduced volume, van der Waals area, parachor, dielectric constant, dipole moment, liquid molar volume	1889	$R^2 > 0.97$ RMSE < 0.8	No	[470]
LSBoost	Molar solubility Range (0.003 to 0.88)	Critical pressure, critical temperature and acentric factor of ionic liquids (specific type of solvents). System temperature and pressure	1140 samples representing 24 ionic liquids	$R^2 > 0.99$ MSE < 0.01	Limited	[676]

Table H.4 Comparison with selected previous studies predicting solubility parameters (Cont'd)

Model(s)	**Solubility parameter(s)**	**Inputs**	**No. of dataset points**	**Metrics**	**Explainability**	**Reference**
LightGBM	Hildebrand solubility and HSPs-based RED Range (0 to 15)	SMILES codes and extracted molecular descriptors	55272 samples representing 81 polymers and 1221 solvents interactions	R^2 = 0.86-0.94	Limited	[677]
Staking-based ensemble	Mole fraction Range (0 to 1)	Critical pressure, critical temperature and molecular weight of ionic liquids (specific type of solvents). System temperature and pressure	4107 points of CO2 solubility in 17 ionic liquids and 549 points of H_2S solubility in 10 ionic liquids	R^2 = 0.97	No	[678]
ANN	LogS Range (-6 to 4)	16 molecular descriptors including drugs melting point and molecular weight in addition to system temperature	4567 samples representing 103 drugs and 49 solvents interactions	R^2 = 0.98	Limited	[679]

Table H.4 Comparison with selected previous studies predicting solubility parameters (Cont'd and end)

Model(s)	**Solubility parameter(s)**	**Inputs**	**No. of dataset points**	**Metrics**	**Explainability**	**Reference**
LightGBM	LogS Range (-6 to 4)	Fingerprints extracted from canonical SMILES codes and system temperatures	5081 samples representing 266 compounds and 123 organic solvents interactions at different temperatures	R^2 = 0.91 MSE < 0.16	No	[680]
ANN	HSPs Range (0 to 20)	Polymer films-solvent contact angle, solvent surface tension and solvent viscosity	70 samples representing 5 polymer films and 14 solvents interactions	R^2 = 0.85-0.93 RMSE = 1.24	No	[681]
ANN (Classification)	Hildebrand solubility and HSPs (to determine the good and bad solvents for proper classification)	Polymers molecular descriptors & fingerprints besides solvents one-hot coding representations	11958 polymer-solvent combinations and a total of 8469 polymer-nonsolvent pairs representing 24 solvent and 4595 polymers	Classification accuracy = 93%	No	[429]

APPENDIX I SUPPLEMENTARY MATERIALS OF ARTICLE 3

Literature review on molecular discovery

There are various approaches for designing and discovering these new materials, such as solvents for CO_2 and lignins, including the experimental methods [682]. This approach primarily relies on wet-lab work, employing various synthesis procedures and experimental designs [501], [515], [516]. Common experimental design methods include one-factor-at-a-time designs, completely randomized designs, factorial designs, response surface methodology (RSM) [683]–[686], *Taguchi* methods [687] and *Bayesian* optimization techniques that account for interactions between factors and their effect on the properties of interest. However, the existing experimental approach is mainly limited by: (1) material consumption, (2) time requirements, and (3) limited search space and exploration capabilities [526]. Knowledge gaps and limited technical expertise continue to hinder progress in discovery. These experiments are also resource-intensive in terms of materials, energy, and labor [688]. Hence, there is a need for more advanced and accelerated approaches to fill the wet lab-based material design and discovery gaps [689], [690]. The emerging dry-lab approach leverages computational tools to design, synthesize, and screen materials prior to experimental validation [691], [692]. **Figure I.1** illustrates different computational methods for materials design and screening.

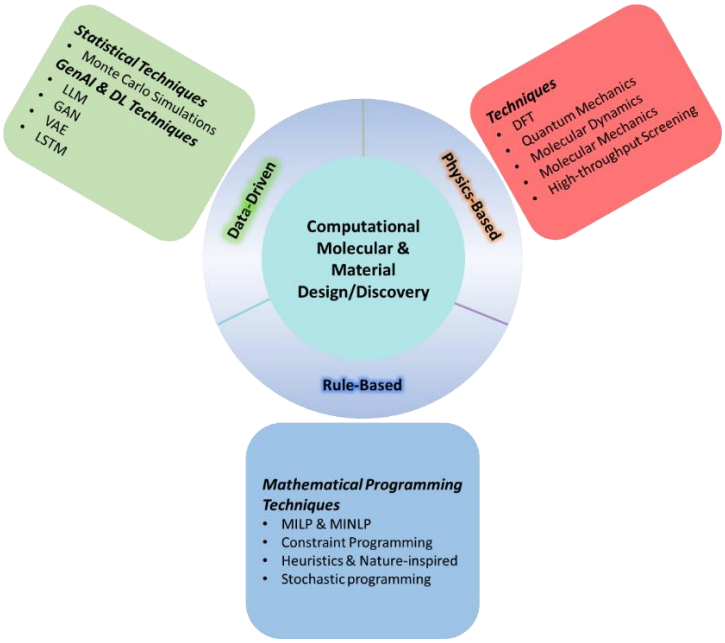


Figure I.1 Common computational methods for molecular and material design

Computational methods for materials design and discovery can be grouped into three main categories: physics-based, and data-driven, and rule-based approaches relying on mathematical programming [693], [694]. Among these, computational chemistry and its related physics-based methods have proven to be powerful tools for material design and discovery [517], [518]. Common examples of computational methods include density functional theory (DFT) [519]–[521], molecular dynamics, quantum mechanics [695], and high-throughput screening techniques [522]–[525] are among these existing. These methods have demonstrated excellent performance across various *in silico* material discovery and screening tasks [156]. For instance, *COSMO-RS*, a quantum chemistry-based method, has been widely used to predict solubility parameters of different materials and screen solvents [695]–[698]. It combines *COSMO* quantum chemistry with statistical thermodynamics for accurate property prediction of real solvents. Several solvents have been designed using this method with adequate performance for various applications [699]–[701]. High-throughput screening works in a combinatorial way to create a large library of materials with vast compositions, properties and structures, enabling their identification of optimum materials for specific application. [702]–[704]. Despite their high fidelity, physics-based methods are often computationally expensive and time-intensive [527], [528]. Therefore, rule-based and data-driven techniques are progressively used to overcome these limitations [705].

In rule-based methods, human experts define the constraints of the optimization problem and the objective function to be optimized for generating new molecules or materials [706], [707]. They typically use mathematical programming techniques such as constraint programming [708], mixed-integer linear programming (MILP) [709], mixed-integer nonlinear programming (MINLP) [710], and stochastic programming [711]. Rule-based methods are at the interface between physics-based and data-driven approaches. They can be used separately or in combination with others, depending on human-defined heuristics and objectives [357], [712], [713].

For data-driven design and discovery methods, many researchers started to utilize statistical methods, e.g. *Monte Carlo* simulations [714]–[717], and generative AI (GenAI) methods employing natural language processing (NLP) and image processing to create new molecule structures (or materials) [529]–[532]. **Table I.1** summarizes recent GenAI-based studies.

Recently, NLP-related techniques frequently employed are language models (LMs) represented by generative pretrained transformers (GPTs) [541], [718]. LMs and GPTs include a collection of

models having various architectures and sizes [719] such as *BERT, RoBERTa, GPT-2, GPT-3, GPT-3.5, GPT-4, Mistral and Llama*. Some studies also use these transformers to predict molecular and materials properties [533]–[537]. Generative deep learning (GDL) also plays a significant role in molecular/material design where various deep neural networks (NN) with different architectures are employed [720]–[722]. The common NN architectures are gated recurrent unit (GRU), long short-term memory (LSTM), variational autoencoders (VAEs) and generative adversarial networks (GANs) [723], [724]. VAEs and GANs depend mainly on images processing concepts. Whereas recurrent NNs including GRU and LSTM employ NLP concepts thanks to their attention mechanism facilitating character and text generation. Most recent AI an ML-based material discovery and design studies have focused on drug discovery [725]–[728], with notable success in generating new drug-like molecules with optimal desired properties, mainly *Quantitative Estimate of Drug-likeness* "QED" [729]–[733].

Although promising results, rule-based methods still face some limitations such as high reliance on human expertise and long computational time. They can also occasionally produce invalid molecular designs. Likewise, existing AI-driven generative techniques often proceed with black-box models, lacking interpretability and sometimes generating invalid designs or even undesired designs. These limitations are usually due to insufficient considerations of physics and chemistry, absence of incremental learning, and lack of targeting ensemble-based generative modeling. A human-centric hybrid modeling approach can address these limitations by combining expert guidance with machine learning to generate molecules and materials that meet multiple KPIs. This approach emphasizes human preference during dataset preparation and model training.

Table I.1 Selected recent studies on deep learning and GenAI for molecular and material design/discovery

Model	**Materials of interest and Application**	**Remarks**	**References**
Junction Tree VAE (JT-VAE)	Small molecules (Drug discovery)	Molecular graphs are generated in two main two steps: 1. A tree-structured scaffold over chemical substructures is constructed. 2. Graph message-passing network assembles these substructures into a complete molecule. The model is trained on a dataset of 250,000 molecules. SMILES codes are converted into graph operations.	[734]
VAE	Small drug-like molecules (Drug discovery)	SMILES codes are converted into continuous molecular representations in latent space. An encoder-decoder neural network (NN) system is used, coupled with a property predictor that estimates chemical properties based on these latent vectors The QM9 (108,000 molecules) and ZINC (250,000 molecules) datasets are used for training.	[197]
Graph VAE	Small drug-like molecules (Drug discovery)	While the model is well-suited for small molecules, it faces challenges in handling large and complex molecules.	[735]

Table I.1 Selected recent studies on deep learning and GenAI for molecular and material design/discovery (Cont'd)

Model	**Materials of interest and Application**	**Remarks**	**References**
CycleGAN	Small drug molecules (Drug design)	The Mol-CycleGAN developed model is trained on the ZINC dataset to generate optimized small drug molecules. SMILES codes are encoded into latent vectors using JT-VAE.	[736]
Multi-CGAN	Multi-property antimicrobial peptides (Drug design)	In addition to drug discovery, some generative models are applied for data augmentation, where newly generated molecules enrich the training datasets of other generative models. Different data sources including literature-based datasets, simulations, and publicly available datasets (e.g. *UniProt*) are used to construct the training dataset.	[539]
WGAN-GCN (MedGAN)	Small drug molecules (Drug discovery)	Both PubChem and ZINC datasets are used for model training. SMILES codes are embedded in latent space vectors and fed to WGAN to generate new molecular graphs. A Graph Convolutional Network (GCN) refines the generated molecules.	[737]
RL (MolDQN)	Small drug molecules (Drug discovery)	Authors stated that they performed no pretraining based on any dataset to avoid dataset-related biases. Molecular optimization was done through the combination of chemistry domain knowledge and RL concepts.	[738]

Table I.1 Selected recent studies on deep learning and GenAI for molecular and material design/discovery (Cont'd)

Model	**Materials of interest and Application**	**Remarks**	**References**
RNN	Small drug molecules (Drug discovery)	ChEMBL is also used in another model where Reduced Graphs (RGs) of molecules are used for the training, and the generated new RGs were then translated to valid SMILES codes. Approximately 800,000 unique RGs are gathered to form the training dataset.	[739]
LSTM	Large molecules (Inhibitor design)	Rearranged during transfection (RET)-specific inhibitors are generated using two LSTM networks acting as encoder and decoder trained on ChEMBL dataset.	[740]
LSTM	Drug-like molecules (Drug design)	ChEMBL datasets are used for extracting SMILES codes of 509,000 bioactive molecules for model training. LSTM learns the grammar of SMILES codes to generate valid molecules.	[538]
LSTM-RL	Various drug molecules (Drug design)	LSTM is used as a generative model and RL is responsible for molecular design optimization. Various databases were used for training and molecular properties extraction including ChEMBL, PubChem and PHYSPROP	[741]

Table I.1 Selected recent studies on deep learning and GenAI for molecular and material design/discovery (Cont'd)

Model	**Materials of interest and Application**	**Remarks**	**References**
Reduced GPT-1 (MolGPT)	Small drug molecules (Drug discovery)	MOSES and GuacaMo datasets are used for model fine-tuning/training. Conditional generation facilitates the generation of molecules having specific scaffolds and properties.	[540]
GPT-2	Large molecules (Protein ligands/drug design)	The Bite Pair Encoding (BPE) strategy was used for SMILES tokenization. Fine-tuning is performed using the jglaser/binding affinity and ZINC20. BPE reduces the number of required tokens to only 5,373 to represent the chemical compounds, despite ZINC dataset containing about 2 billion compounds.	[742]
GPT-3	Organic molecules (Organic semiconductors discovery)	An organic semiconductors dataset containing 48,82 molecules from the Cambridge Structural Database is used to predict electronic and functional properties of organic molecules using a prompt-based learning approach.	[743]
GPT-3.5	Inorganic materials (Various applications)	Datasets such as Materials Project (containing 393,053 inorganic compositions) are used to address synthesizability and precursor selection problems. The input consists solely of the target chemical formula provided through prompts. However, dataset bias limits generalization to rare chemistry cases.	[744]

Table I.1 Selected recent studies on deep learning and GenAI for molecular and material design/discovery (Cont'd and end)

Model	**Materials of interest and Application**	**Remarks**	**References**
GPT-4	Metal organic frameworks "MOFs" (Energy and environmental applications)	An iterative recommendation for reticular chemistry (e.g. MOFs) uses GPT-4 to interact with expert chemists. The interactive work of the design framework guides the discovery..	[745]
GPT-4	Different organic materials (various applications including synthesis of organo-catalysts, insect repellent and discovering new chromophore	The *ChemCrow* framework integrates GPT-4 (as an LLM) with experts-designed tools to support tasks in drug discovery, material design, and organic synthesis. It uses interactive prompting to suggest laboratory synthesis instruction for novel molecules.	[746], [747]
Diffusion Transformer (*AlphaFold*-3)	Large biomolecules (structure prediction of biomolecular interactions)	An updated version of the *AlphaFold* project, considering diffusion-based architecture, enables the joint structure prediction of complex molecules such as proteins and nucleic acids.	[748]

Text-based raw data tokenization and positional encoding procedure

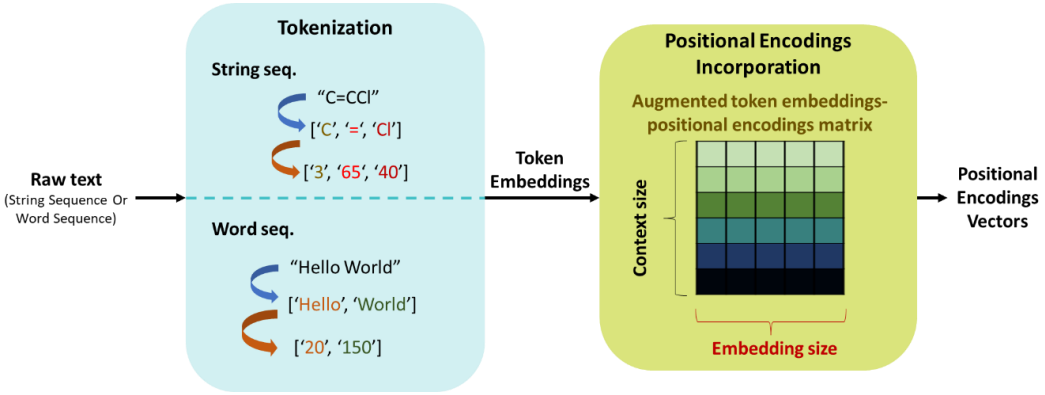


Figure I.2 Tokenization and positional encoding step related to GPTs

Image-based raw data vectorization procedure

Figure I.3 depicts an illustrative example of the general image vectorization procedure. All images in the dataset are uniformly resized to a standard resolution (e.g., 256×256 pixels) to ensure

consistency and enable efficient machine learning modeling. A data augmentation step is recommended at the beginning to enhance the generalization capability of ML models. This step includes transformations such as brightness adjustment, cropping, image rotation and other customized modifications. After these preprocessing steps, feature extraction is performed using several methods, including Scale-Invariant Feature Transform (SIFT) [749], Local Binary Patterns (LBP) [750] and Histogram of Oriented Gradient (HOG) [751]. These methods compute descriptors from the image's raw pixels, capturing various characteristics of the original image content comprising key points, shapes and textures. The extracted features are then concatenated into a flattened one-dimensional (1D) vector. A normalization step is essential to stabilize the training process. The vectorized features are scaled to values between 0 and 1. These normalized values are then fed to the deep learning model. It is recommended to apply normalization to improve training convergence and stability. The features extraction and vectors concatenation steps can be eliminated and only three essential steps of augmentation, resizing and normalization are performed in new deep learning (DL) practices. In this case the input data to the DL model will be in the shape of normalized image tensor keeping the spatial information.

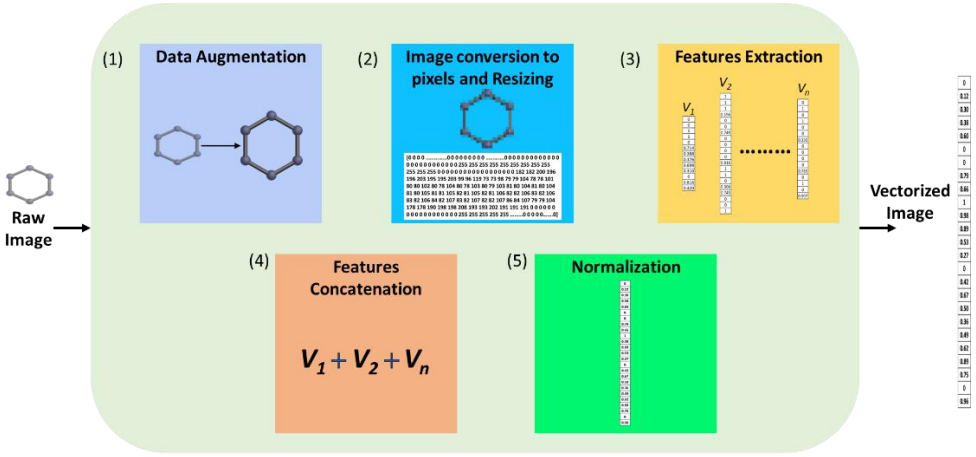


Figure I.3 General image vectorization procedure

Generative AI models

GPT-2

GPT-2 is a large language model with about 1.5 billion parameters that is an agnostic application model, which is fine-tuned for a wide range of tasks with acceptable accuracy. It is particularly designed to perform well in sequence-to-sequence learning applications such as text generation.

As a transformer-based model, GPT-2 efficiently captures complex patterns and learns dependencies from a large amount of text-based data. Once optimized, it generates diverse coherent and unique text sequences. GPT-2 is referred to as a decoder-only model, as it lacks an explicit encoder [752]. The general architecture of GPT-2 model (**Figure I.4**) contains several layers including tokenization and positional encoding generation, decoder transformer and stacked decoder layers. The decoder transformer layer contains important sub-layers such as self-attention, feed-forward neural network (NN) and normalization layers. During tokenization, processes such as masking and padding are applied to the input sequences to ensure they are uniform and fixed size for efficient neural network (NN) training.

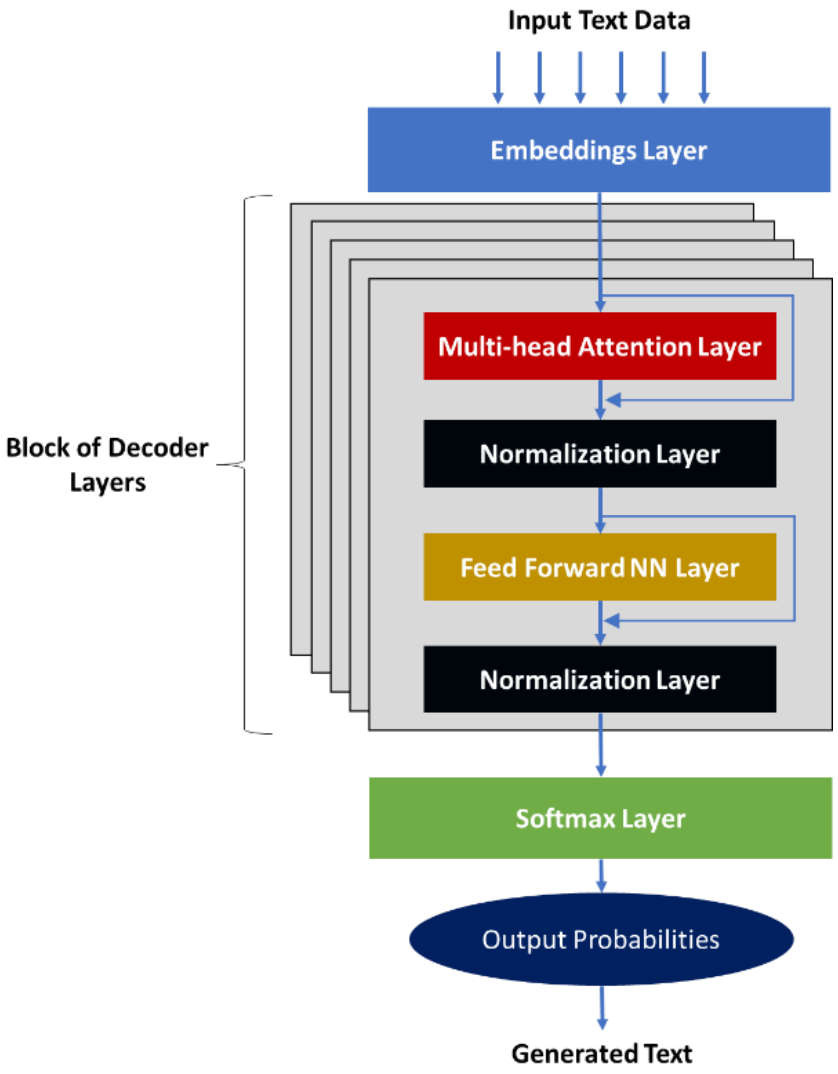


Figure I.4 Simplified illustration of GPT-2 general architecture [753], [754]

GAN

GANs are a special class of NNs widely used for generating high quality images [755]. They represent a paradigm shift from the discriminative focus of deep learning to generative capabilities. GANs are trained on a specific dataset to learn how to produce new information and new data samples that resemble those in the original dataset. This consists of two essential components: the generator, which produces images, and the discriminator, which evaluates the generated images. If the discriminator network cannot detect the difference between actual and generated images representation, then it is converged.

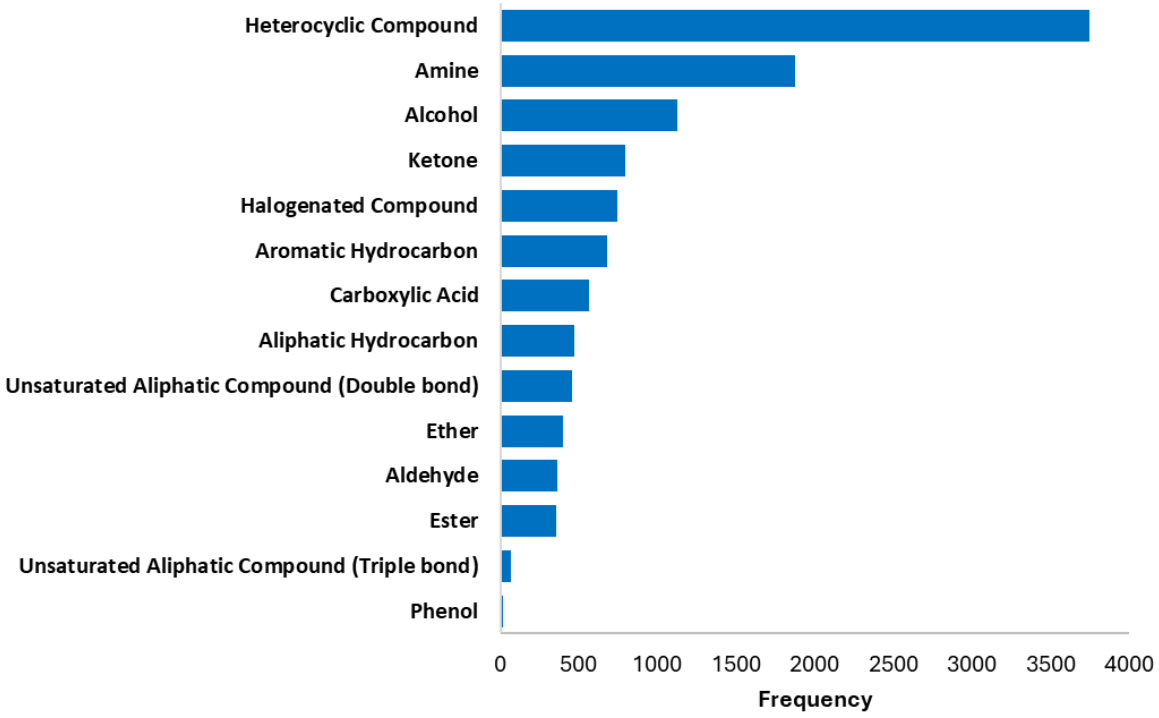


Figure I.5 Solvent type distribution of the parent dataset

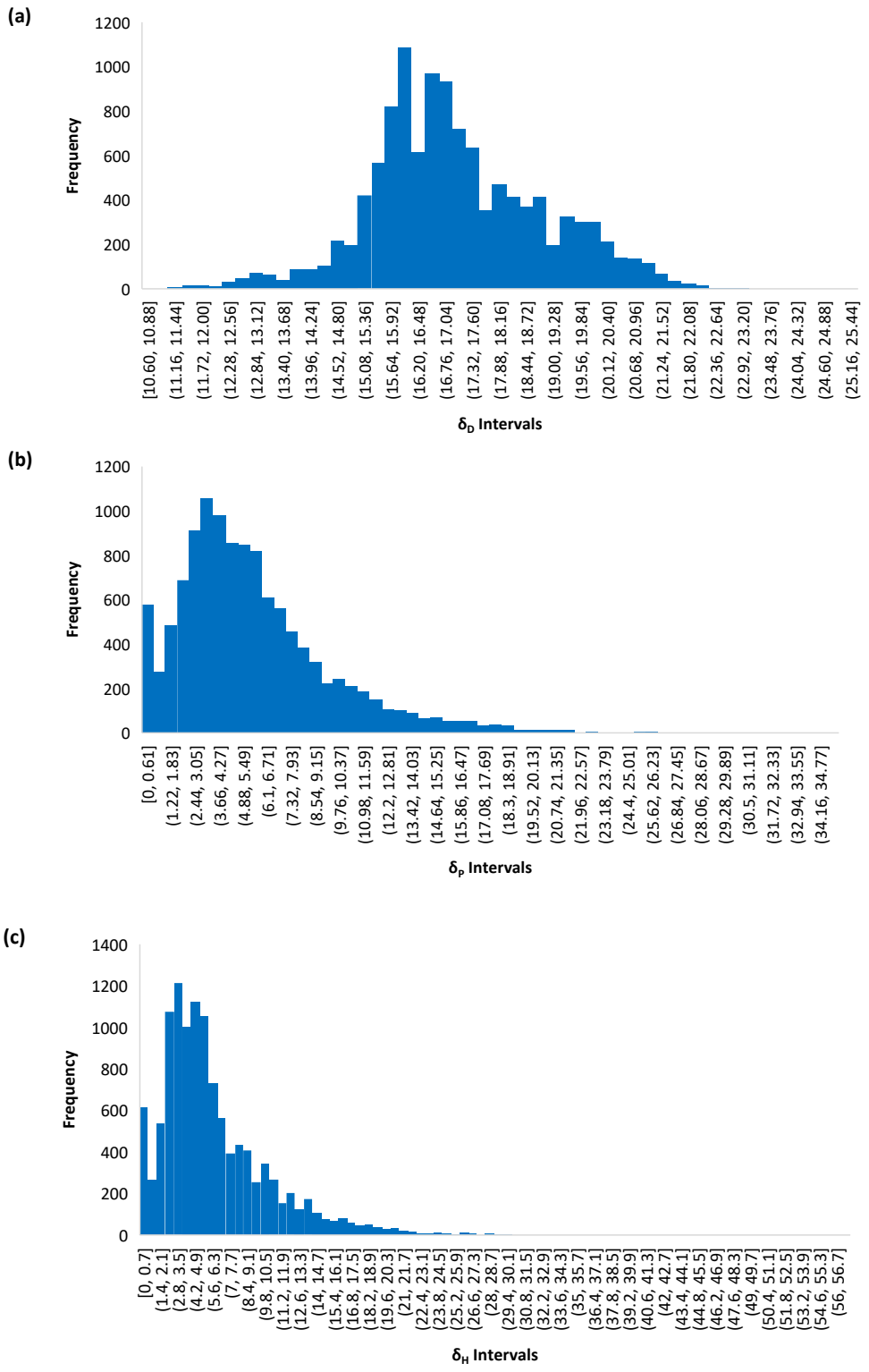


Figure I.6 Hansen solubility parameters (a) dispersion, (b) polarization and (c) hydrogen bonding distributions for the parent dataset

Table I.2 Simple statistical description of *Hansen* solubility parameters of the solvents in the raw parent dataset

Statistical parameter	Variables		
	δ_D [$MPa^{0.5}$]	δ_P [$MPa^{0.5}$]	δ_H [$MPa^{0.5}$]
Maximum	25.3	35.2	57
Minimum	10.6	0	0
Average	17.14	5.69	6.05
Median	16.9	4.8	4.8
Standard Deviation	1.80	3.98	4.54
Mode	16	0.1	0.1
Skewness	0.26	1.49	1.83
Kurtosis	0.20	3.68	6.18

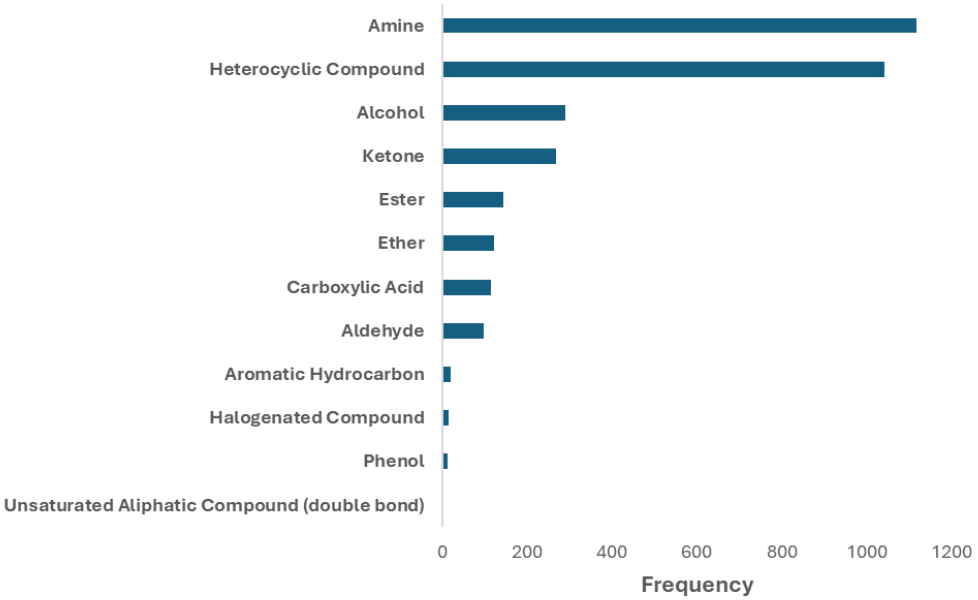


Figure I.7 Softwood Kraft Lignin Solvents distribution

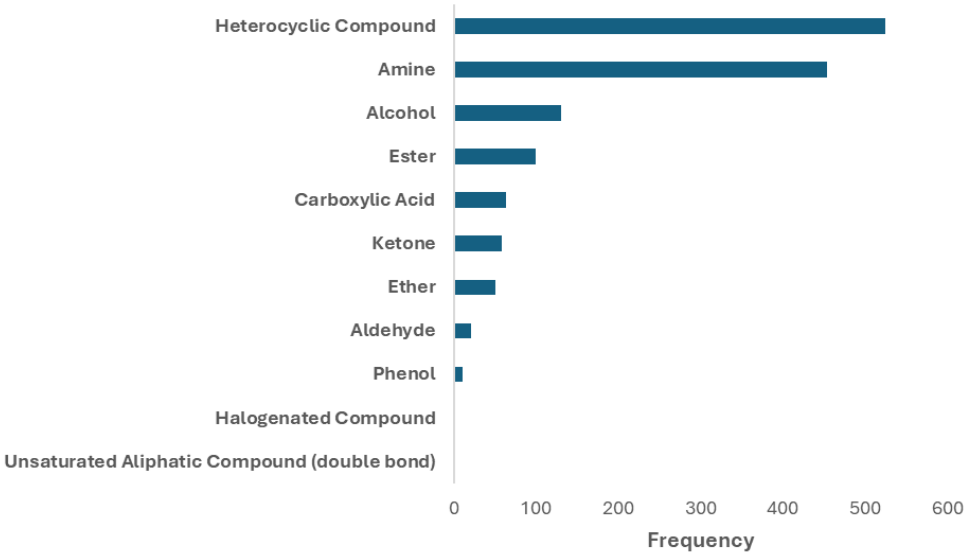


Figure I.8 Bagasse Lignin Solvents distribution

Table I.3 Toxicity and flammability results of each reactant and intermediate molecules in the suggested synthesis routes

SMILES	LC50	LD50	D.T.	FPt
CCOS(=O)(=O)C(F)(F)F	24.18	-----	NON	58.65
CCOCCC(O)CC	496.82	5632.83	D. Tox	60.92
CCl	325.58	1800	-----	-18.82
CCBr	245.73	1349.92	D. Tox	6.63
NNCO	897.05	401.75	D. Tox	82.84
NCO	2220	2054.38	D. Tox	49.28
O=CCC(O)C(O)CO	3438.19	---------	NON	168.43
CI	656.91	409.42	--------	15.81
OBr	---------	---------	---------	---------
OCCCl	802.10	32.86	D. Tox	59.61
O=C(NCO)C(F)(F)F	1333.79	--------	D. Tox	54.05
O=C(O)CC(O)C(O)CO	5452.25	10648.97	NON	209.86

Pseudocode for the proposed general approach

1: Convert all molecules in D_raw to canonical SMILES format

2: Remove invalid and duplicate molecules using RDKit + InChIKey

3: Construct cleaned dataset D_clean

4: Construct a cleaned dataset of good solvents for each case D_clean_good_i

5: Train auto-labeling models using:

- Valid SMILES from ChEMBL
- Synthetically generated invalid SMILES

6: Further validate auto-labeling models using D_val

7: Fine-tune generative models (GPT-2, W-GAN, VAE) on D_clean_good_i

8: Generate candidate molecules:

S_GPT2 ← Generate(GPT-2)

S_WGAN ← Generate(W-GAN)

S_VAE ← Generate(VAE)

S_all ← S_GPT2 ∪ S_WGAN ∪ S_VAE

9: For each SMILES s ∈ S_all do

10: If RDKit_is_valid(s) = False then

11: Remove s from S_all

12: Else

13: Convert s to molecular structure

14: End if

15: End for

16: Remove duplicates from S_all using InChIKey

17: For each molecule m ∈ S_all do

18: Predict Hansen solubility parameters (δD, δP, δH)

19: Compute RED_CO2(m) and RED_Lignin(m)

20: If RED_CO2(m) ≤ RED_max OR RED_Lignin(m) ≤ RED_max then

21: Predict:

- Greenness score

- Toxicity

- Flash point

- Boiling point

22: Else

23: Remove m from S_all

24: End if

25: End for

26: For each molecule m ∈ S_all do

27: If (Toxicity(m) > threshold) OR (Greenness(m) < threshold) then

28: Remove m from S_all

29: End if

30: End for

31: For each molecule m ∈ S_all do

32: Check exact-match novelty using:

- InChIKey

- ChemSpider

33: Compute structural similarity to training set using:

- Morgan fingerprints

- Tanimoto similarity

34: End for

35: For each molecule m ∈ S_all do

36: Perform retrosynthesis analysis to identify possible synthetic routes

37: Compute synthetic accessibility / route feasibility score

38: If (Synthesis score > synthesis_threshold) then

39: Remove m from S_all

40: End if

41: End for

42: Rank remaining molecules according to:

- RED value

- Greenness score

- Safety indicators

- Synthetic accessibility

43: Perform interpretability analysis:

- Extract key molecular descriptors
- Identify dominant chemical features

44: Return top-ranked solvent candidates for:

- CO_2 capture
- Lignin solubilization

45: Apply incremental learning procedure via

- Changing the weights of the ensemble-based procedure based on the generation results
- Enhancing D_clean_good_i with the generated good green solvents for each case

APPENDIX J SUPPLEMENTARY MATERIALS OF ARTICLE 4

Table J.1 Important parameters for capture cost and emissions calculations

Parameter [unit]	**Average Values**
Electrical boiler efficiency [%]	95-99
Natural gas boiler efficiency [%]	85
Steam turbine efficiency [%]	40
Gas turbine efficiency [%]	40
Electricity emissions factor (Quebec) [g CO_2 eq/ kWh]	2
Electricity emissions factor (Alberta) [g CO_2 eq/ kWh]	470
Natural gas emissions factor [kg CO_2 / GJ of energy produced]	55.82
Natural gas price (Quebec) [USD/GJ]	1.805
Natural gas price (Alberta) [USD/ GJ]	1.38
Electricity price (Quebec) [USD/kWh]	0.047
Electricity price (Alberta) [USD/kWh]	0.17
Quebec Steam price (electricity-based production) [USD/ton]	26.35
Quebec Steam price (natural gas-based production) [USD/ton]	7.64
Alberta Steam price (electricity-based production) [USD/ton]	127.78
Alberta Steam price (natural gas-based production) [USD/ton]	5.84
Cooling water price [USD/ton]	0.11
MEA Price [USD/kg]	1.4
DEPG Price [USD/kg]	1.8
Lean-Rich MEA heat exchanger heat transfer coefficient [$W/m^2.K$]	1500
Lean MEA cooler heat transfer coefficient [$W/m^2.K$]	1500
Reboiler heat transfer coefficient [$W/m^2.K$]	800
Stripper condenser transfer coefficient [$W/m^2.K$]	1500
Working hours [h/ year]	8000

IML Patterns

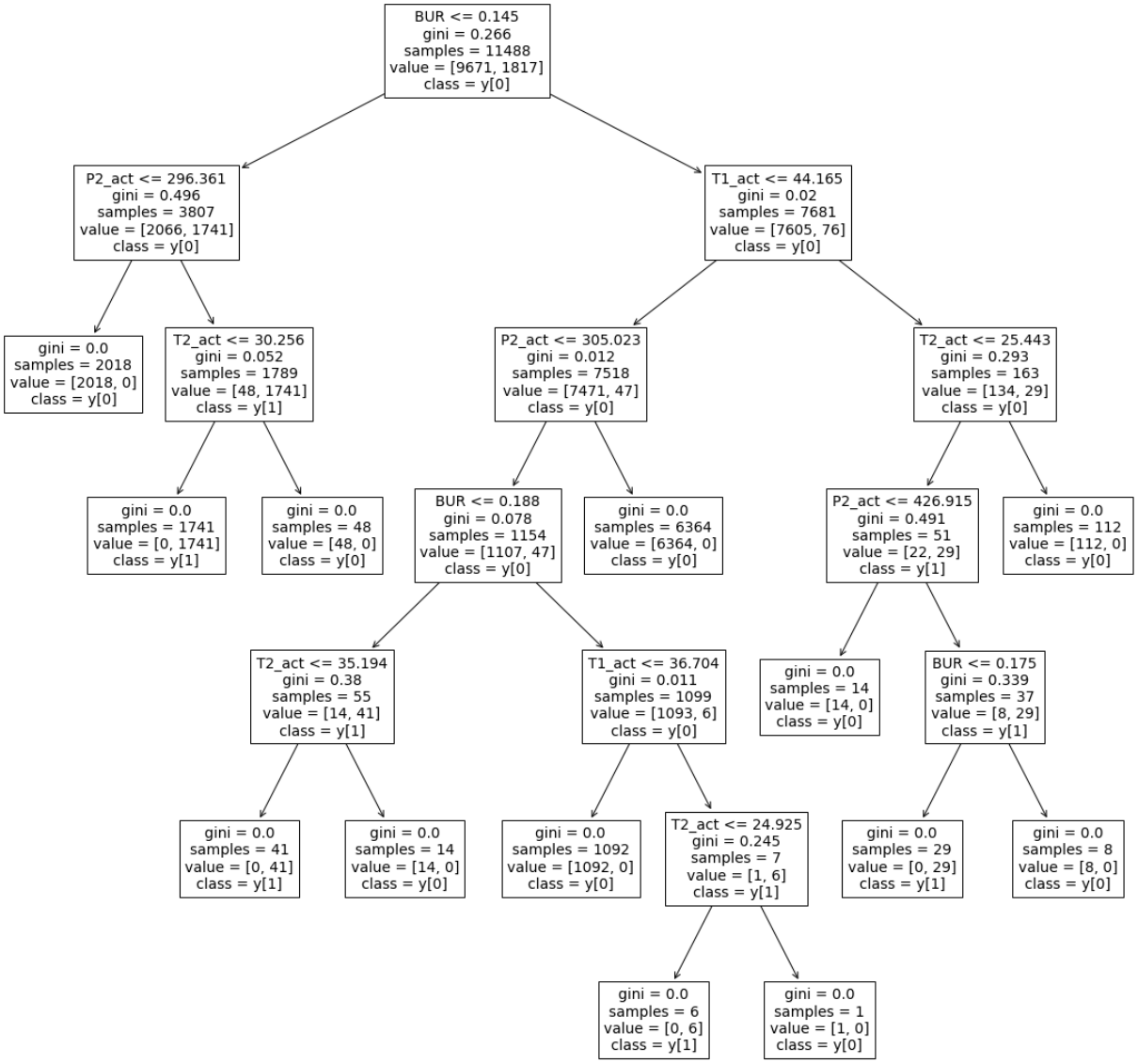


Figure J.1 Patterns for capture cost of MEA process

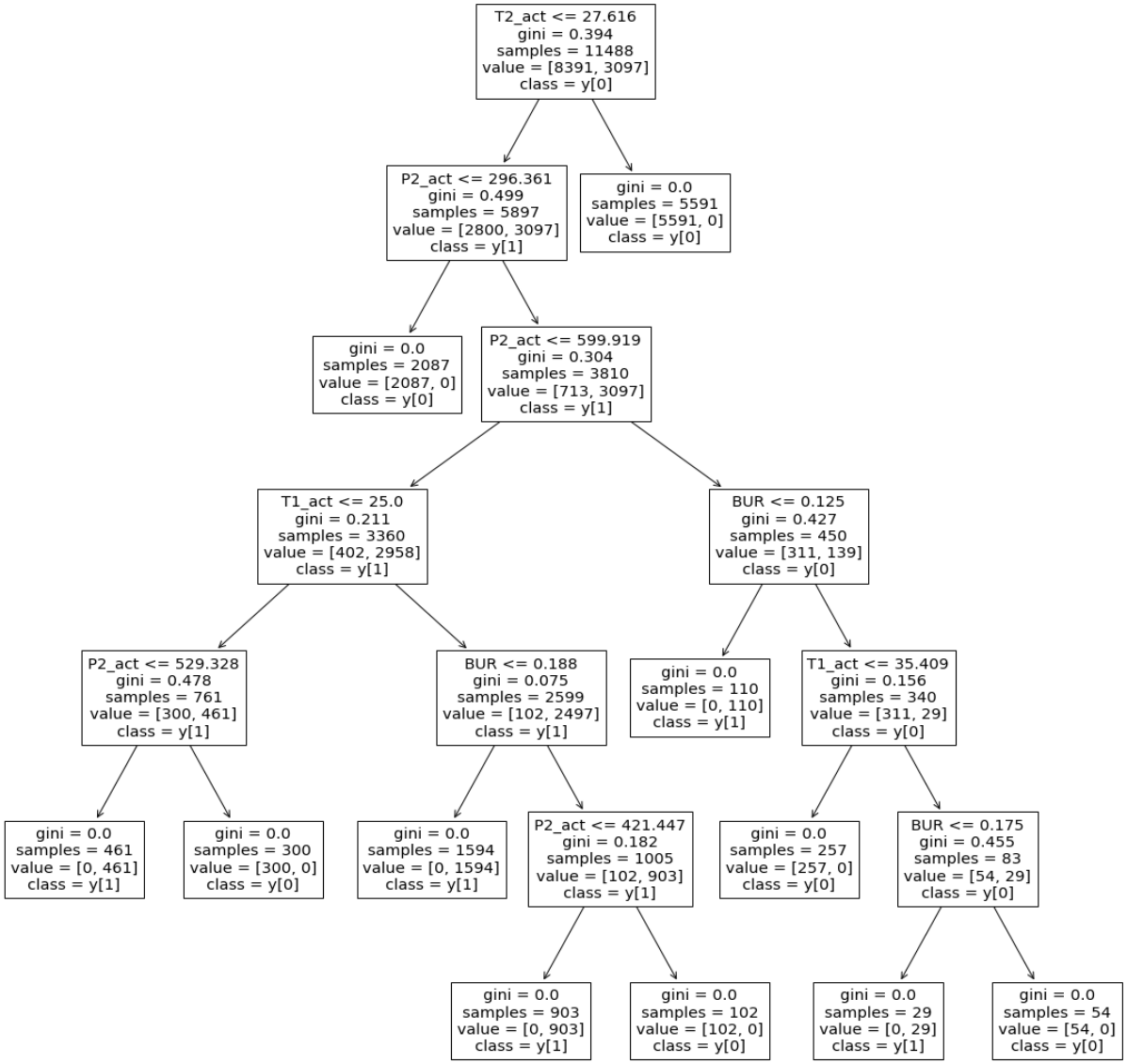


Figure J.2 Patterns for capture emissions of MEA process

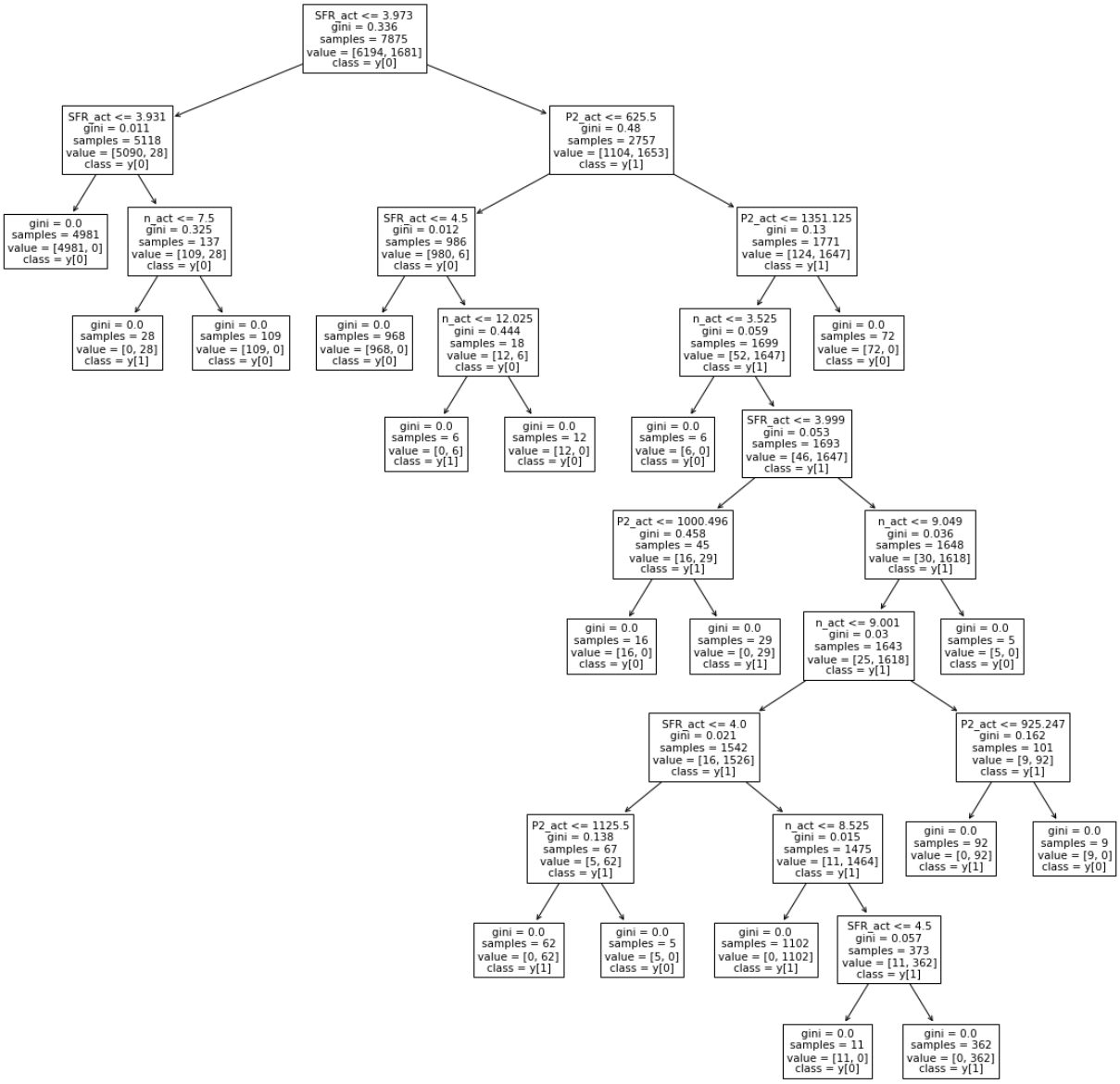


Figure J.3 Patterns for capture effectiveness of DEPG process

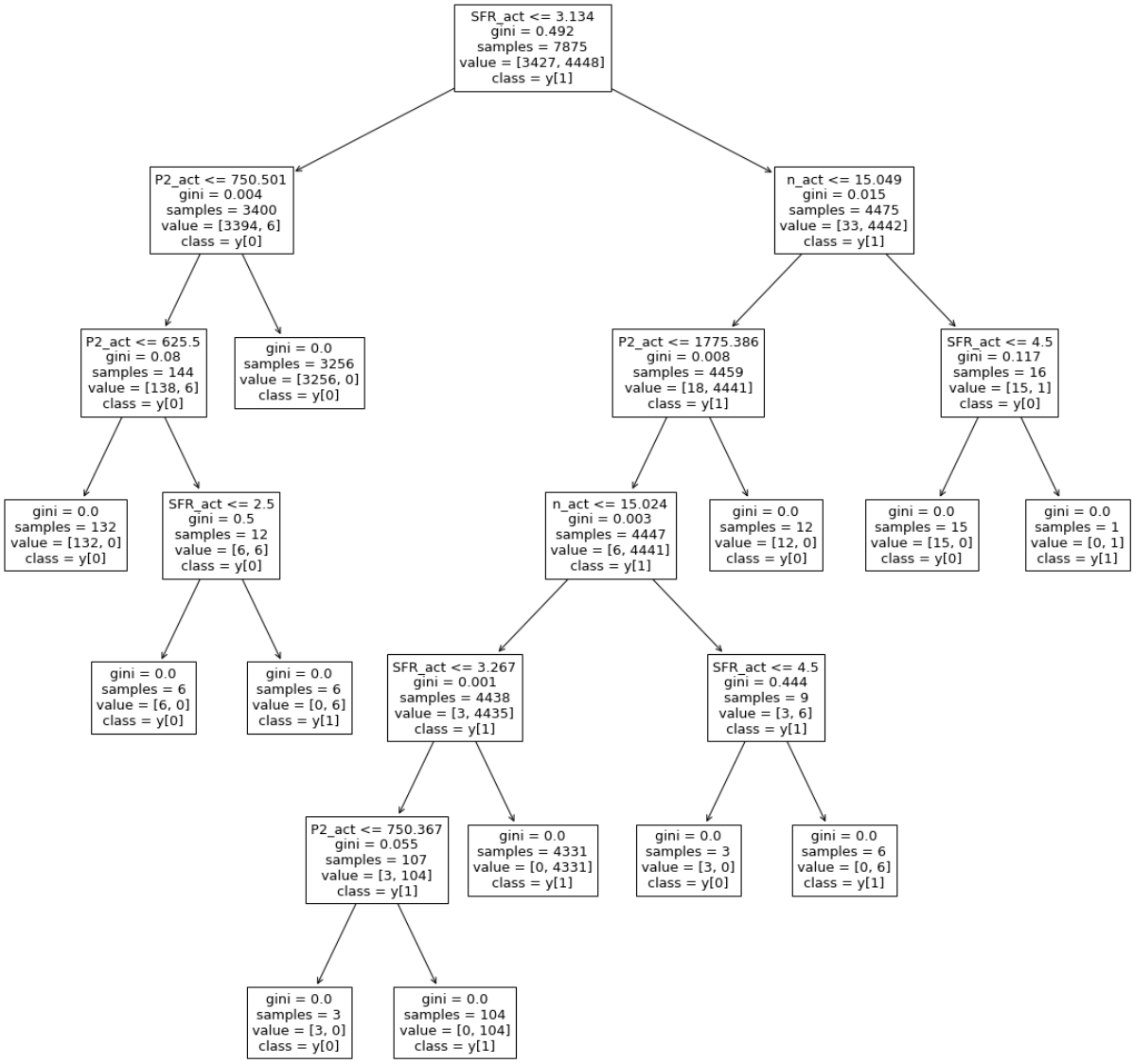


Figure J.4 Patterns for capture eco-friendliness of DEPG process

Importance and XAI figures

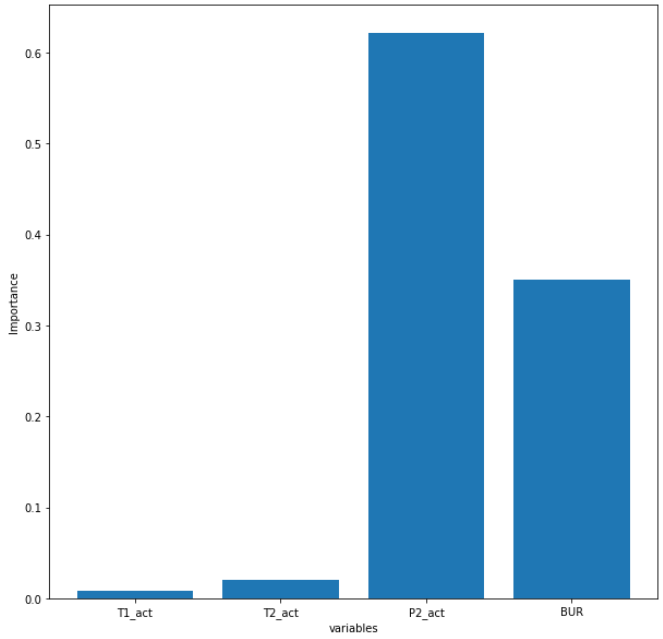


Figure J.5 Results of IML on MEA-based capture cost (classification-based)

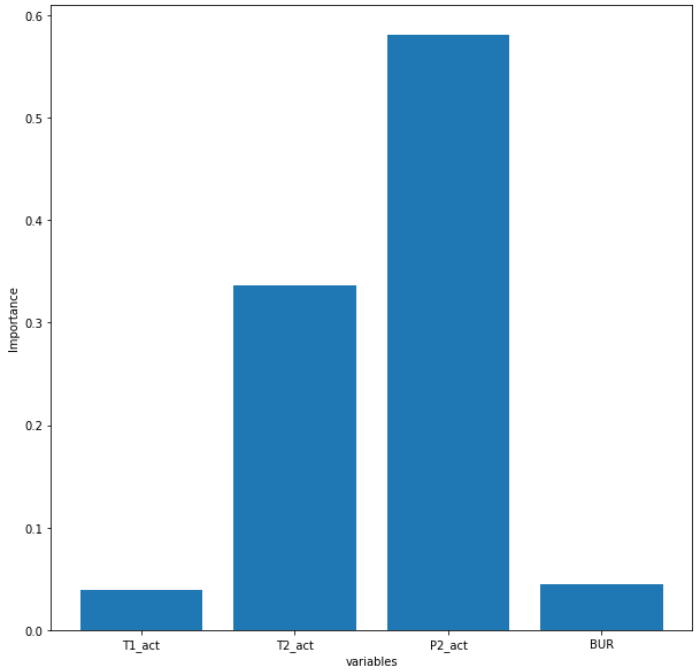


Figure J.6 Results of IML on MEA-based capture emissions (classification-based)

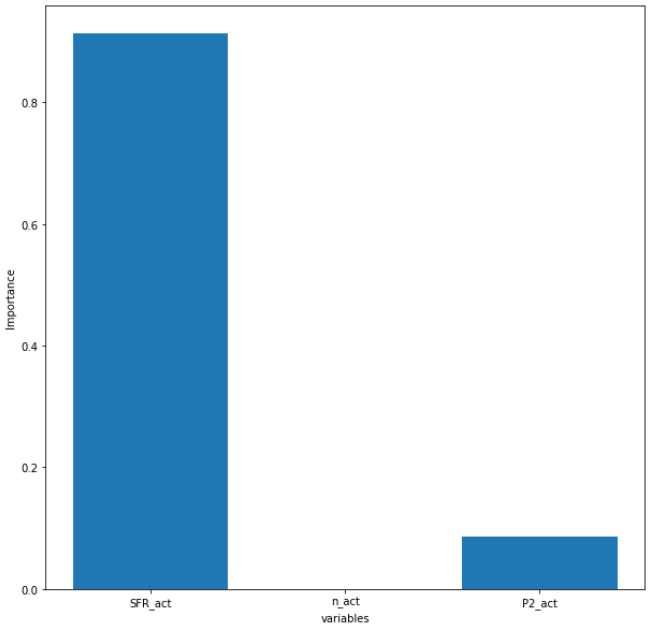


Figure J.7 Results of IML on DEPG-based capture effectiveness (classification-based)

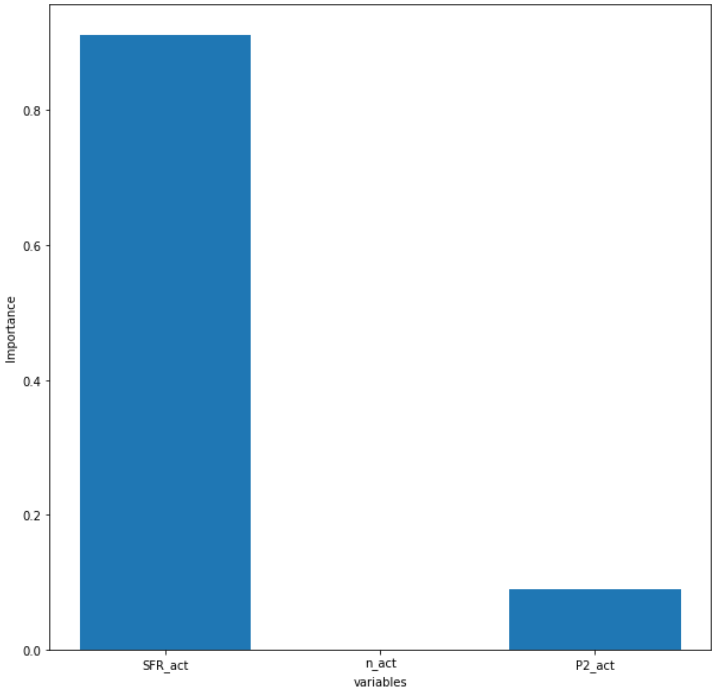


Figure J.8 Results of IML on DEPG-based capture eco-friendliness (classification-based)

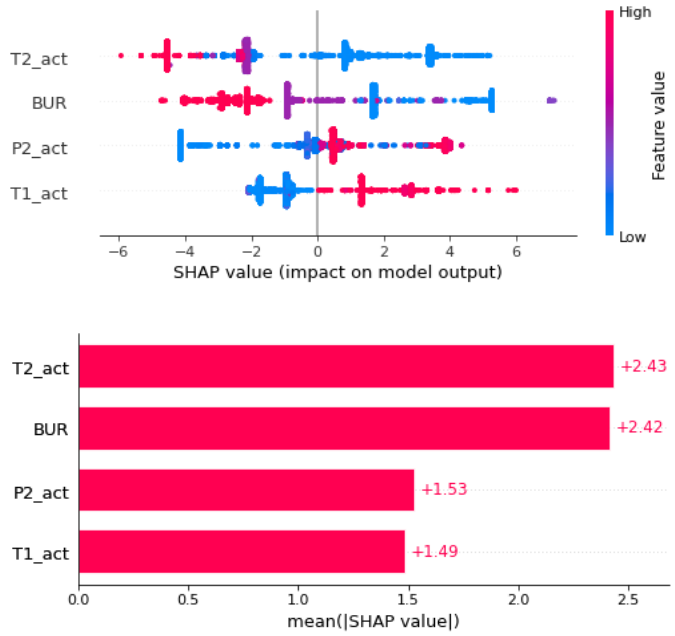


Figure J.9 Results of XAI on MEA-based capture cost (classification-based)

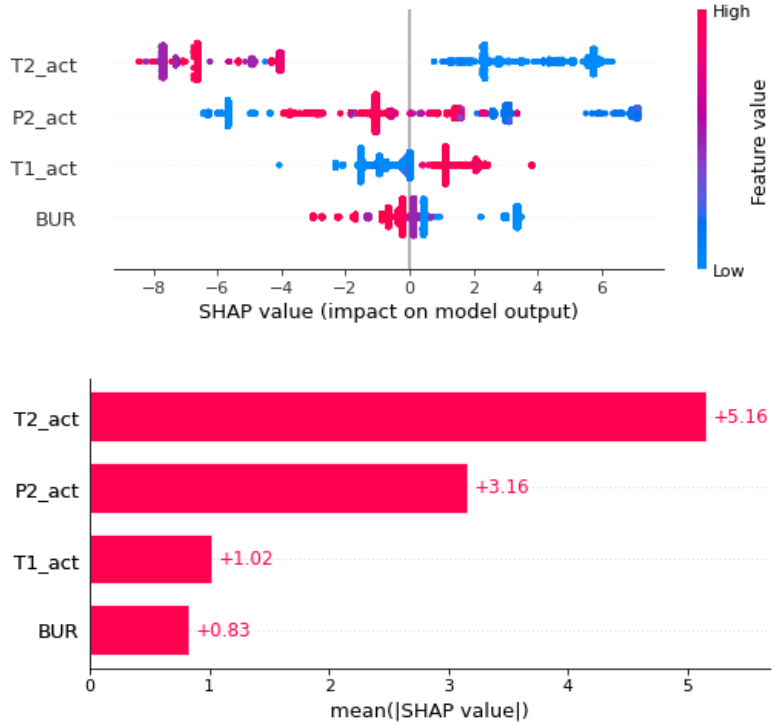


Figure J.10 Results of XAI on MEA-based capture emissions (classification-based)

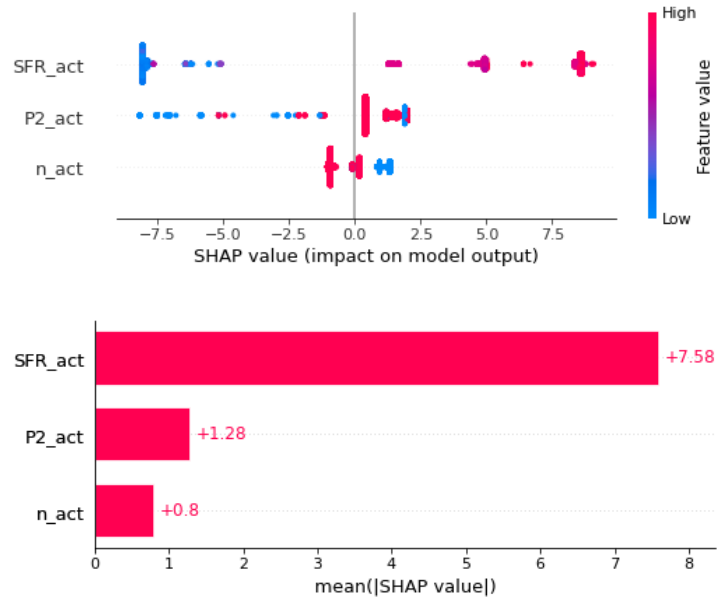


Figure J.11 Results of XAI on DEPG-based capture effectiveness (classification-based)

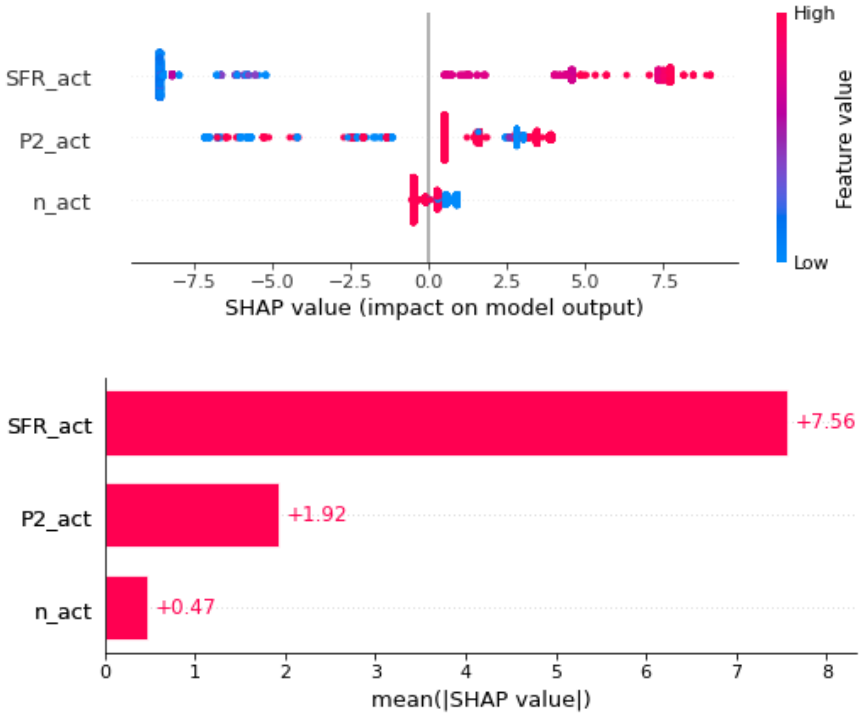


Figure J.12 Results of XAI on DEPG-based capture eco-friendliness (classification-based)

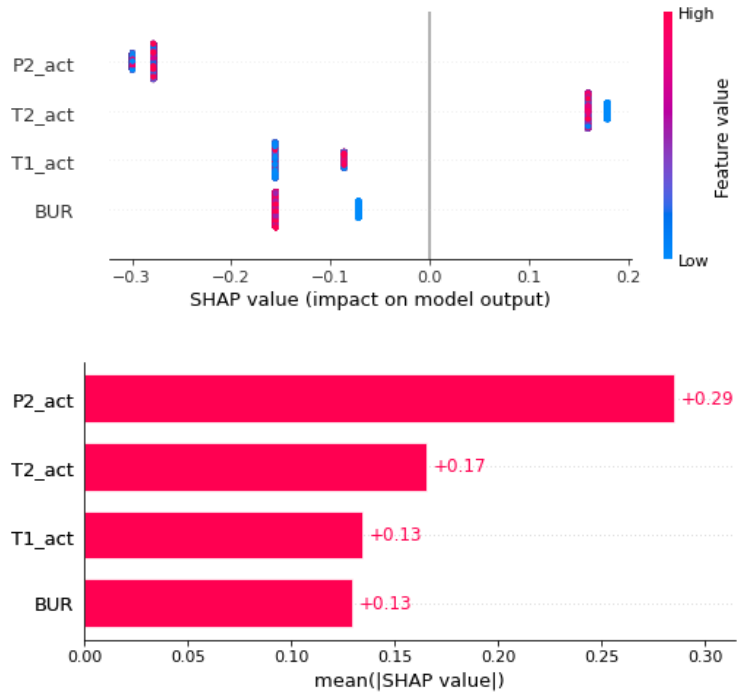


Figure J.13 Results of XAI on MEA-based capture cost (regression-based)

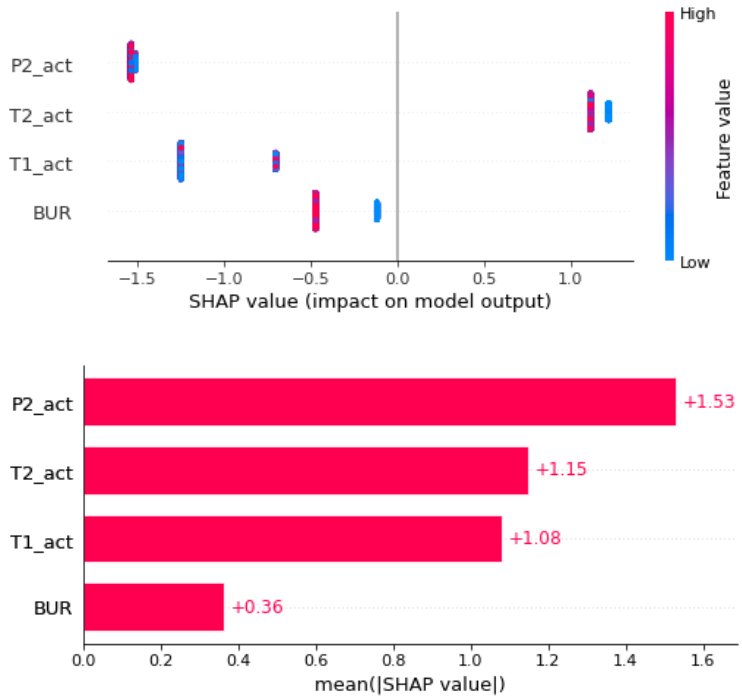


Figure J.14 Results of XAI on MEA-based capture emissions (regression-based)

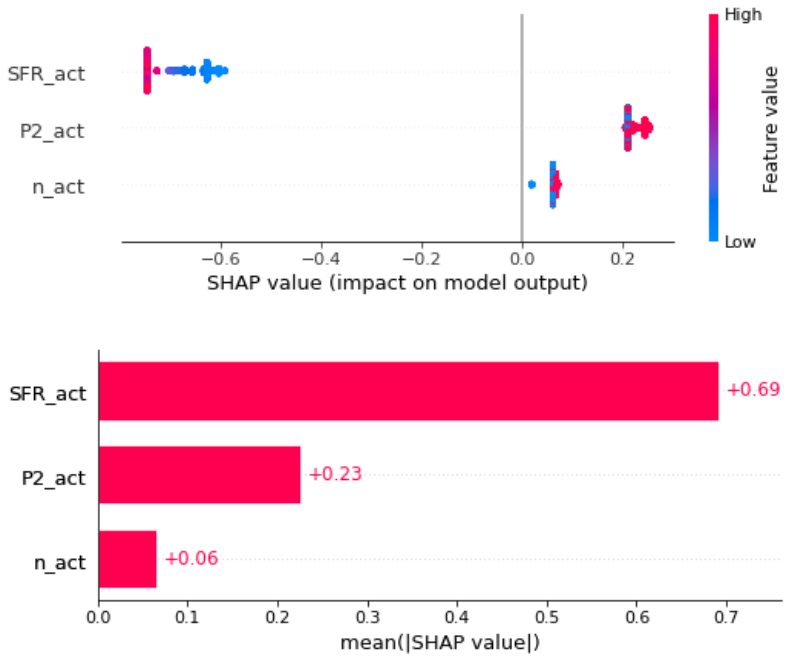


Figure J.15 Results of XAI on DEPG-based capture effectiveness (regression-based)

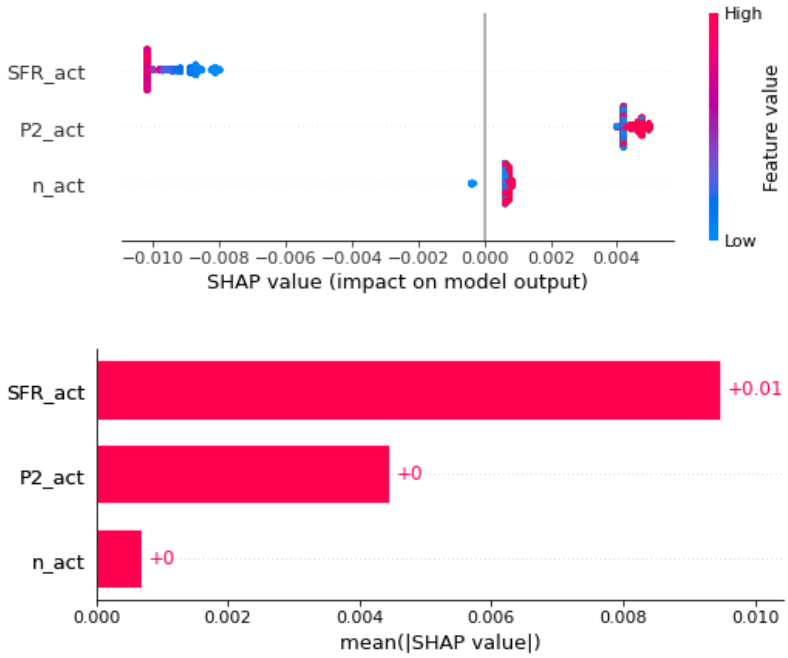


Figure J.16 Results of XAI on DEPG-based capture eco-friendliness (regression-based)

Table J.2 FCA results of capture cost classification (MEA Process)

Observation Number	Temp1_Low	Temp1_High	Temp2_Low	Temp2_High	Pressure_Low	Pressure_High	BUR_Low	BUR_High	CO2 Recovery_Low	CO2 Recovery_High	Cost_Low	Cost_High
1		X		X		X		X		X		X
2		X	X			X		X		X		X
3	X		X			X		X		X		X
4		X	X			X		X		X	X	
5		X		X	X			X		X		X
6		X	X		X			X		X		X
7	X		X		X			X		X	X	
8	X		X		X		X			X	X	
9		X	X		X		X			X	X	
10		X		X		X	X			X		X
11		X		X	X		X			X		X
12		X	X			X	X			X	X	
13	X			X		X	X			X		X
14	X			X	X		X			X		X
15	X		X			X	X			X	X	
16		X	X		X			X		X	X	
17	X		X		X		X		X			X
18	X		X		X		X			X		X
19		X	X		X		X			X		X
20		X		X		X		X	X			X
21		X		X	X			X	X			X
22	X		X		X			X	X			X
23		X	X		X			X	X			X
24	X			X		X		X	X			X
25	X			X	X			X	X			X
26	X		X		X			X		X		X

Table J.3 FCA results of capture emissions classification (MEA Process)

Observation Number	Temp1_Low	Temp1_High	Temp2_Low	Temp2_High	Pressure_Low	Pressure_High	BUR_Low	BUR_High	CO2 Recovery_Low	CO2 Recovery_High	em_Low	em_High
1		X		X		X		X		X		X
2		X	X			X		X		X		X
3	X		X			X		X		X		X
4		X	X			X		X		X	X	
5		X		X	X			X		X		X
6		X	X		X			X		X		X
7	X		X		X			X		X	X	
8	X		X		X		X			X	X	
9		X	X		X		X			X	X	
10		X		X		X	X			X		X
11		X		X	X		X			X		X
12		X	X			X	X			X	X	
13	X			X		X	X			X		X
14	X			X	X		X			X		X
15	X		X			X	X			X	X	
16		X	X		X			X		X	X	
17	X		X		X		X		X			X
18	X		X		X		X			X		X
19		X	X		X		X			X		X
20		X		X		X		X	X			X
21		X		X	X			X	X			X
22	X		X		X			X	X			X
23		X	X		X			X	X			X
24	X			X		X		X	X			X
25	X			X	X			X	X			X
26		X		X	X			X		X	X	

Table J.4 FCA results of capture effectiveness classification (DEPG Process)

Observation Number	SFR_act_Low	SFR_act_High	n_act_Low	n_act_High	Pressure_Low	Pressure_High	CO2_Recovery_Low	CO2_Recovery_High	cap_eff_Low	cap_eff_High
1		X	X			X		X	X	
2	X			X		X	X			X
3	X		X			X	X			X
4		X	X			X		X		X
5		X		X	X		X		X	
6		X	X		X		X		X	
7	X			X	X		X			X
8	X		X		X		X			X
9		X		X		X		X	X	
10		X		X	X			X	X	
11		X	X		X			X	X	
12		X		X	X		X			X
13		X	X		X		X			X
14	X		X			X		X	X	
15	X		X			X	X		X	

Table J.5 FCA results of capture eco-friendliness classification (DEPG Process)

Observation Number	SFR_act_Low	SFR_act_High	n_act_Low	n_act_High	Pressure_Low	Pressure_High	CO2_Recovery_Low	CO2_Recovery_High	cap_eco_Low	cap_eco_High
1		X	X			X		X		X
2	X			X		X	X			X
3	X		X			X	X			X
4		X		X	X		X		X	
5		X	X		X		X		X	
6	X			X	X		X			X
7	X		X		X		X			X
8		X		X		X		X	X	
9		X		X	X			X	X	
10		X	X			X		X	X	
11		X	X		X			X	X	
12		X		X	X		X			X
13		X	X		X		X			X
14	X		X			X		X	X	
15	X		X			X	X		X	

Causality-informed ML Modeling

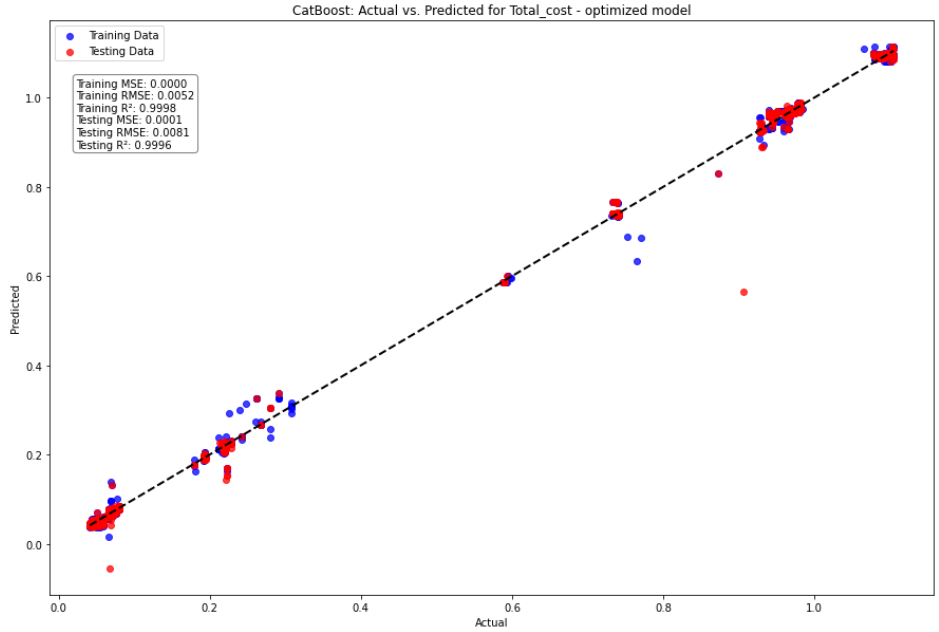


Figure J.17 Results of causality-informed CatBoost modeling of total capture cost for MEA process

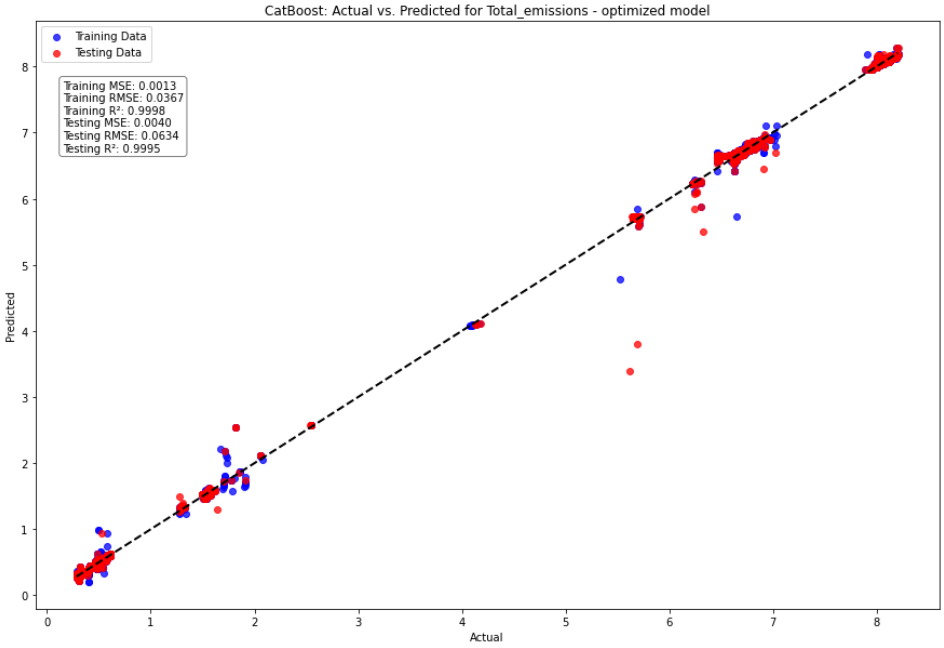


Figure J.18 Results of causality-informed CatBoost modeling of total capture emissions for MEA process

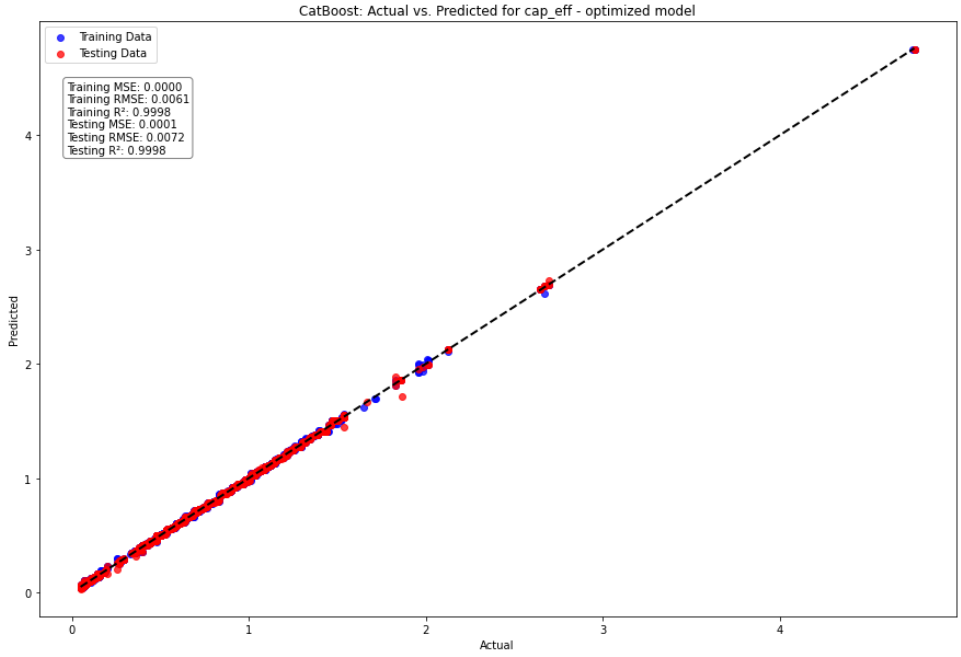


Figure J.19: Results of causality-informed CatBoost modeling of capture effectiveness for DEPG process

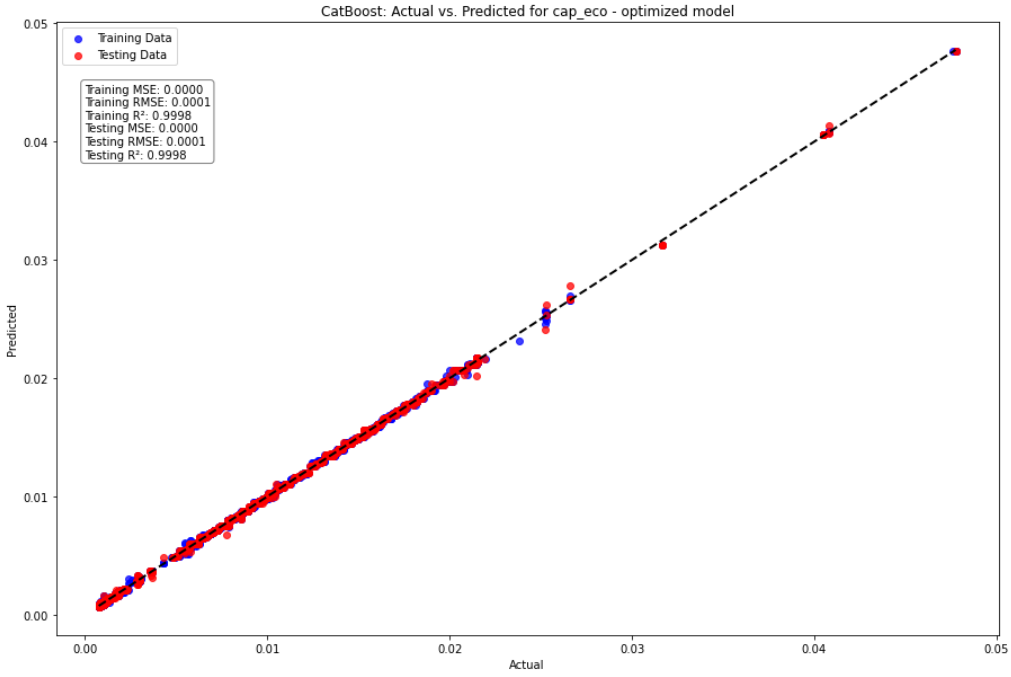


Figure J.20 Results of causality-informed CatBoost modeling of capture eco-friendliness for DEPG process

Table J.6 Optimal hyperparameters of the developed *CatBoost* model

Hyperparameter	**Value**
Iterations	250
Depth	4
Learning rate	0.167
L2_leaf_reg	0.087
Subsample	0.826
Colsample_bylevel	0.638

Pseudo-code of our CIRL method

Initialize:

Define action space A and state space S

Set initial design variable ranges based on literature/prior knowledge

Define initial KPI thresholds: GoalC (capture cost), GoalEM (emissions)

Initialize DDPG agent with hyperparameters (actor, critic, buffer, γ=0, OU noise)

Set simulation environment (e.g., Aspen HYSYS digital twin)

Initialize replay buffer B ← ∅

Initialize iteration counter ← 0

WHILE not converged DO:

FOR episode = 1 to N_episodes DO:

Initialize state s_0 (based on system reset or last optimal design)

FOR step t = 1 to max_steps_per_episode DO:

 # Agent selects action using actor policy + exploration noise

$a_t \leftarrow \text{Actor}(s_t) + \text{OU\_Noise}()$

 # Simulate process using Aspen HYSYS with action a_t

$s_{t+1}, \text{KPIs} \leftarrow \text{SimulateProcess}(a_t)$

 # Compute reward based on KPIs and thresholds

$r_t \leftarrow \text{ComputeReward}(\text{KPIs}, \text{GoalC}, \text{GoalEM})$

 # Store experience tuple in replay buffer

$B \leftarrow B \cup \{(s_t, a_t, r_t, s_{t+1})\}$

 # Train agent using mini-batch sampled from buffer

TrainDDPG(B)

 # Update current state

$s_t \leftarrow s_{t+1}$

END FOR

Store all simulation data from current episode

END FOR

 # Perform causal analysis on accumulated data

Causal_Info ← PerformCausalityAnalysis(Data):

- Phase 1: Apply IML, XAI, FCA, Granger Causality

- Phase 2: Construct DAGs → Fit Causal Bayesian Networks (CBNs)

- Phase 3: Derive empirical models for expert interpretability

 # Update action ranges and KPI thresholds using causal insights

UpdateDesignSpace(Causal_Info)

UpdateKPIThresholds(Causal_Info)

 # Increment iteration count

iteration ← iteration + 1

 # Check for convergence criteria (stabilized reward and KPIs)

IF Converged(Causal_Info, Rewards, KPIs) THEN:

BREAK

END WHILE

 # Final Output:

Return optimized design variables, causal graphs, empirical equations, and trained RL policy.

Equations of capture cost and emissions for MEA process

Equations for the range of (0.15-0.5 USD/kg) and (1.25-2.6 kg CO_2 emitted/kg CO_2 captured)

$$\boldsymbol{Capture\ Cost} = -12.66 \times P^{0.234} + 19.69 \times T_S^{0.059} + 13.97 \times T_{FG}^{0.279} + 3.73 \times BUR^{0.678} - \frac{1.44e02}{P^{0.392}} + \frac{31.54}{T_S^{0.289}} - 1.12 \times T_{FG}^{0.542} + \frac{3.46}{BUR^{0.313}} + \left(BUR^{0.678} + \frac{1}{BUR^{0.313}}\right) \times \left(2.3e-03 \times P^{1.169} - \frac{17.96}{T_S^{0.859}} - 8.06e-01 \times T_{FG}^{0.527}\right) \quad (J.1)$$

$$\boldsymbol{Total\ Emissions} = -12.84 \times P^{0.256} + 35.26 \times T_S^{0.033} + 31.56 \times T_{FG}^{0.328} - 6.72 \times BUR^{0.602} - \frac{1.51e02}{P^{0.359}} + \frac{15.15}{T_S^{0.265}} - 10.49 \times T_{FG}^{0.496} - \frac{3.81}{BUR^{0.287}} + \left(BUR^{0.602} + \frac{1}{BUR^{0.287}}\right) \times \left(4.25e-03 \times P^{1.169} + \frac{6.04}{T_S^{0.859}} + 0.37 \times T_{FG}^{0.527}\right) \quad (J.2)$$

APPENDIX K SUPPLEMENTARY MATERIALS OF ARTICLE 5

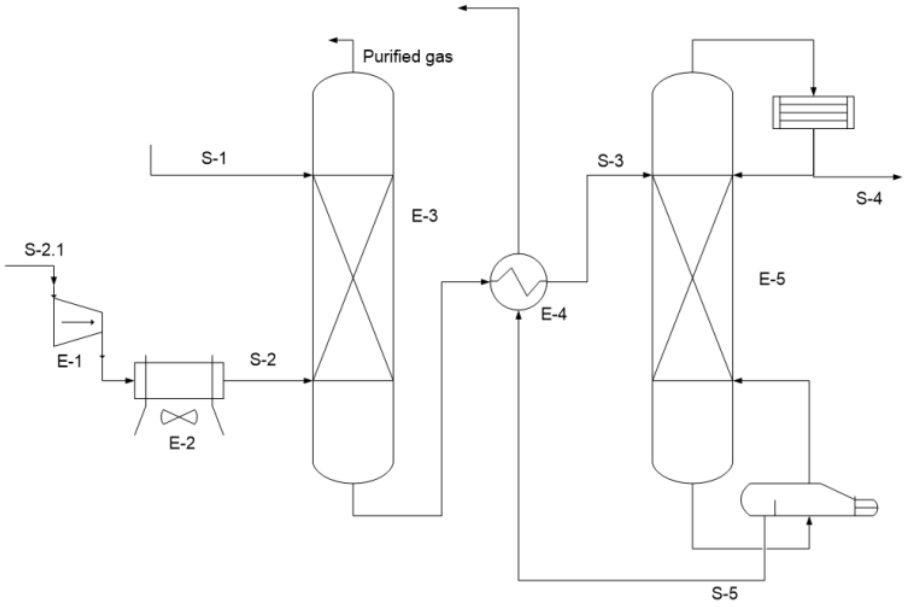


Figure K.1 Process flow diagram (ammonia-water system)

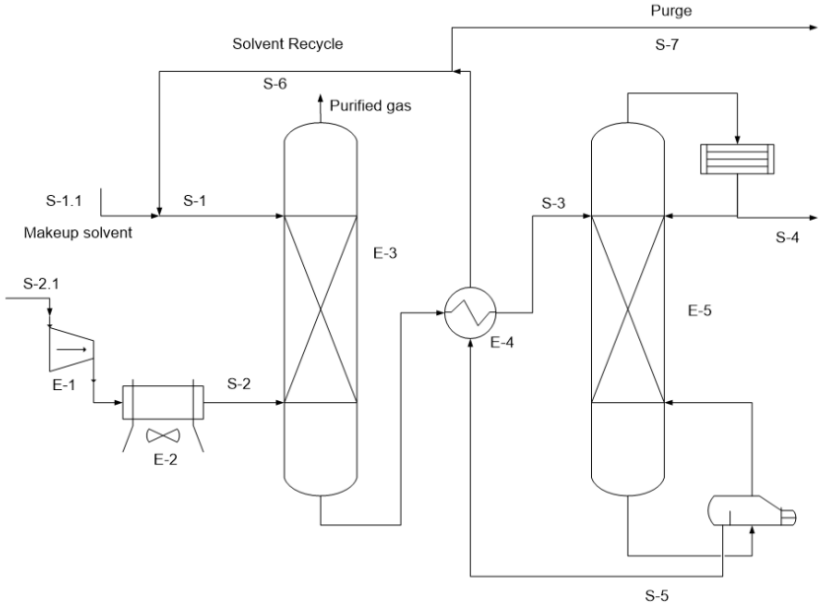


Figure K.2 Process flow diagram (MEA-based CO2 Capture system)

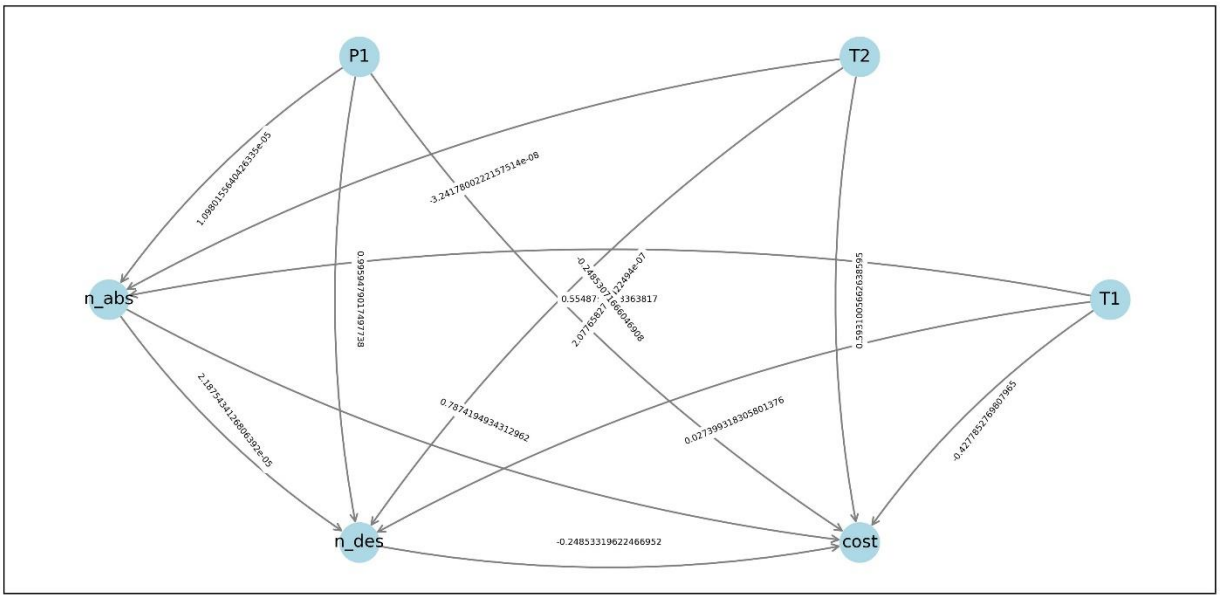


Figure K.3: Causal graph based on Bayesian causal discovery and inference (ammonia-water system)

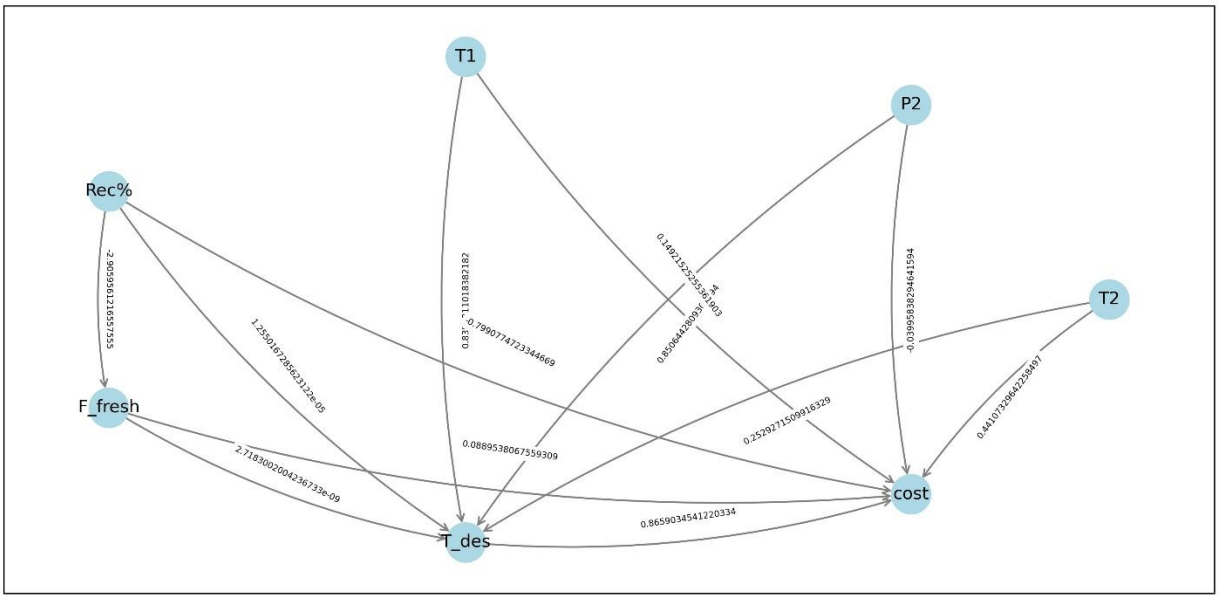


Figure K.4 Causal graph based on Bayesian causal discovery and inference (CO_2-MEA system)