Ensemble machine learning for empowered industrial decarbonization: High-Fidelity predictions of Hansen solubility parameters for streamlined solvent selection

Eslam G. Al-Sakkari[1,2], Ahmed Ragab[1,2], Mostafa Amer[3], Olumoye Ajao[4], Marzouk Benali[2,*], Daria C. Boffito[5], Hanane Dagdougui[1] and Mouloud Amazouz[2]

[1] Department of Mathematics and Industrial Engineering, Polytechnique Montréal, 2500 Chemin de Polytechnique, Montréal, Québec, H3T 1J4, Canada

[2] Natural Resources Canada, CanmetENERGY, 1615 Lionel-Boulet Blvd, P.O. Box 4800, Varennes, Québec, J3X 1P7, Canada

[3] Department of Electrical Engineering, Polytechnique Montréal, 2500 Chemin de Polytechnique, Montréal, Québec, H3T 1J4, Canada

[4] Natural Resources Canada, Clean Fuels Branch, Fuel Diversification Division, 580 Booth Street, Ottawa, K1A 0E4, Canada

[5] Department of Chemical Engineering, Polytechnique Montréal, 2500 Chemin de Polytechnique, Montréal, Québec, H3T 1J4, Canada

*Corresponding author: marzouk.benali@nrcan-rncan.gc.ca

Supplementary Materials

To complement the findings presented in our manuscript, this supplementary material offers comprehensive insights into the machine learning algorithms, data preprocessing steps, and validation processes used to enhance the prediction of Hansen solubility parameters for efficient chemical solvent selection.

Descriptive captions

TABLES		
Number	**Title**	**Description**
S1	Key descriptors based on NNMF	This table summarizes the most significant descriptors extracted from RDKit with their corresponding scores after using NNMF.
S2	Decision fusion results (Solubility_2 "δ_P")	This table introduces a summary of the results of each fusion step in the case of predicting polarization solubility parameter
S3	Decision fusion results (Solubility_3 "δ_H")	This table introduces a summary of the results of each fusion step in the case of predicting hydrogen-bonding solubility parameter
S4	Comparison with selected previous studies considering ML models to predict solubility parameters	In this table, we compare the results of our methodology with those previously published in recent studies considering ML for the prediction of various solubility parameters. This comparison highlights the novelty of our work.

FIGURES		
Number	**Title**	**Description**
S1	NNMF concept simple schematics	This figure depicts the common schematics of NNMF technique reported in literature
S2	NNMF results for ($n_{descriptors}$ = 208 & $n_{components}$ = 3)	This figure depicts the data biclustering at the optimal number of components with the scores of the whole molecular descriptors extracted from RDKit. The key descriptors were then determined based on these results.
S3	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	This figure illustrates the effect of data size during the training of different ML models to predict the polarization solubility parameters.
S4	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	This figure illustrates the effect of data size during the training of different ML models to predict the

		hydrogen-bonding solubility parameters.
S5	Results of different individual techniques (Polarization solubility parameter).	This figure visualizes the results of all the optimized models that build the developed fused model to predict polarization solubility parameter. The presented results represent 100 randomly selected data points.
S6	Results of final decision fusion vs. selected different individual techniques (Polarization solubility parameter).	This figure compares the performance of individual models with that of the fused model to illustrate its superiority to predict polarization solubility parameter based on SMILES codes.
S7	Predicted Vs. experimental values of polarization solubility parameter (final decision fusion results).	This figure compares the predicted values of the polarization solubility parameter using the developed methodology to the experimental values to show the prediction accuracy of this new method.
S8	Results of different individual techniques (Hydrogen bonding solubility parameter).	This figure visualizes the results of all the optimized models that build the developed fused model to predicted hydrogen-bonding solubility parameter. The presented results represent 100 randomly selected data points.
S9	Results of final decision fusion vs. selected different individual techniques (Hydrogen bonding solubility parameter).	This figure compares the performance of individual models with that of the fused model to illustrate its superiority to predict hydrogen-bonding solubility parameter based on SMILES codes.
S10	Predicted Vs. experimental values of hydrogen bonding solubility parameter (final decision fusion results).	This figure compares the predicted values of the hydrogen-bonding using the developed methodology to the experimental values to show the prediction accuracy of this new method.
S11	SHAP values of sugar cane bagasse-based lignin solvents classification based on RED (**Descriptors**)	This figure shows the relative importance of different key descriptors that represent their ability to predict the solvent's goodness towards sugar cane bagasse-based lignin solvation.
S12	SHAP values of sugar cane bagasse-based lignin solvents classification based on RED (**Hansen solubility parameters**)	This figure focuses on the relative importance of HSPs and how they affect the ability of solvents to dissolve sugar cane bagasse-based lignin. This will help in selecting the

		appropriate solvent based on their HSPs and will help during the design of new solvents.
S13	SHAP values of CO_2 solvents classification based on RED **(Descriptors)**	This figure shows the relative importance of different key descriptors that represent their ability to predict the solvent's goodness towards CO_2 solvation.
S14	SHAP values of CO_2 solvents classification based on RED **(Hansen solubility parameters)**	This figure focuses on the relative importance of HSPs and how they affect the ability of solvents to dissolve CO_2. This will help in selecting the appropriate solvent based on their HSPs and will help during the design of new solvents for carbon capture.

Table S1: 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
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
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

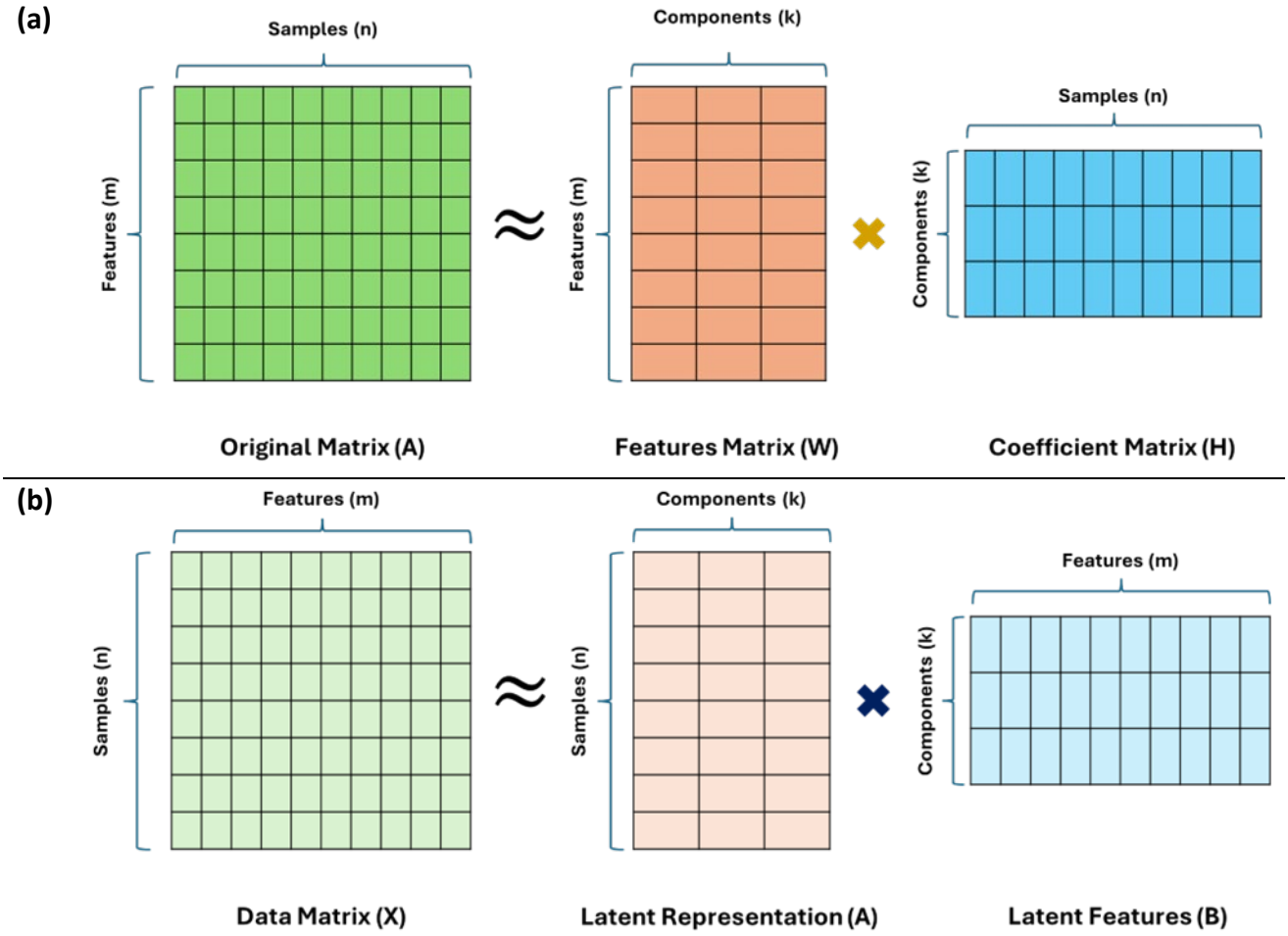


Figure S1: NNMF concept simple schematics

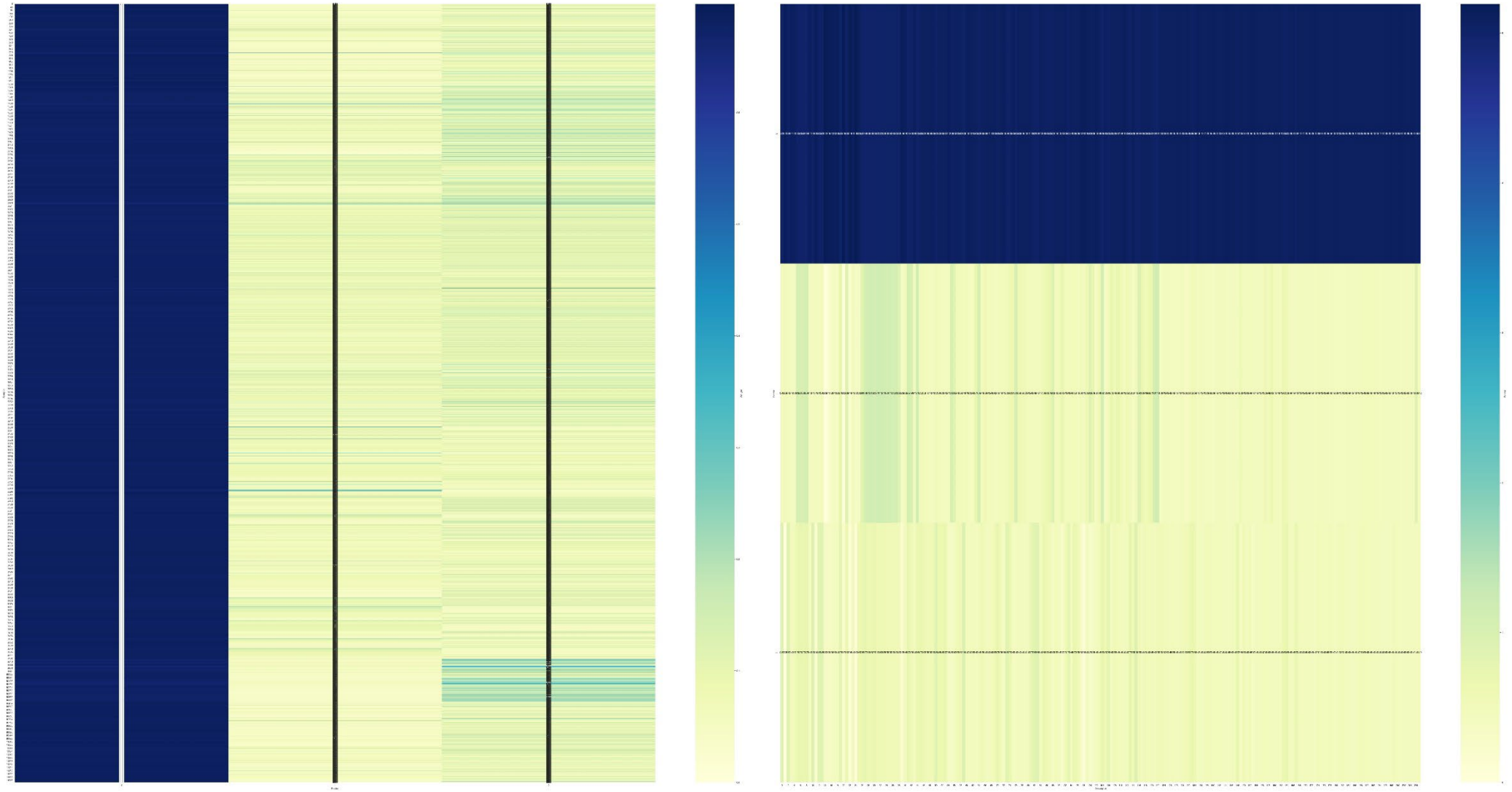


Figure S2: NNMF results for ($n_{descriptors}$ = 208 & $n_{components}$ = 3)

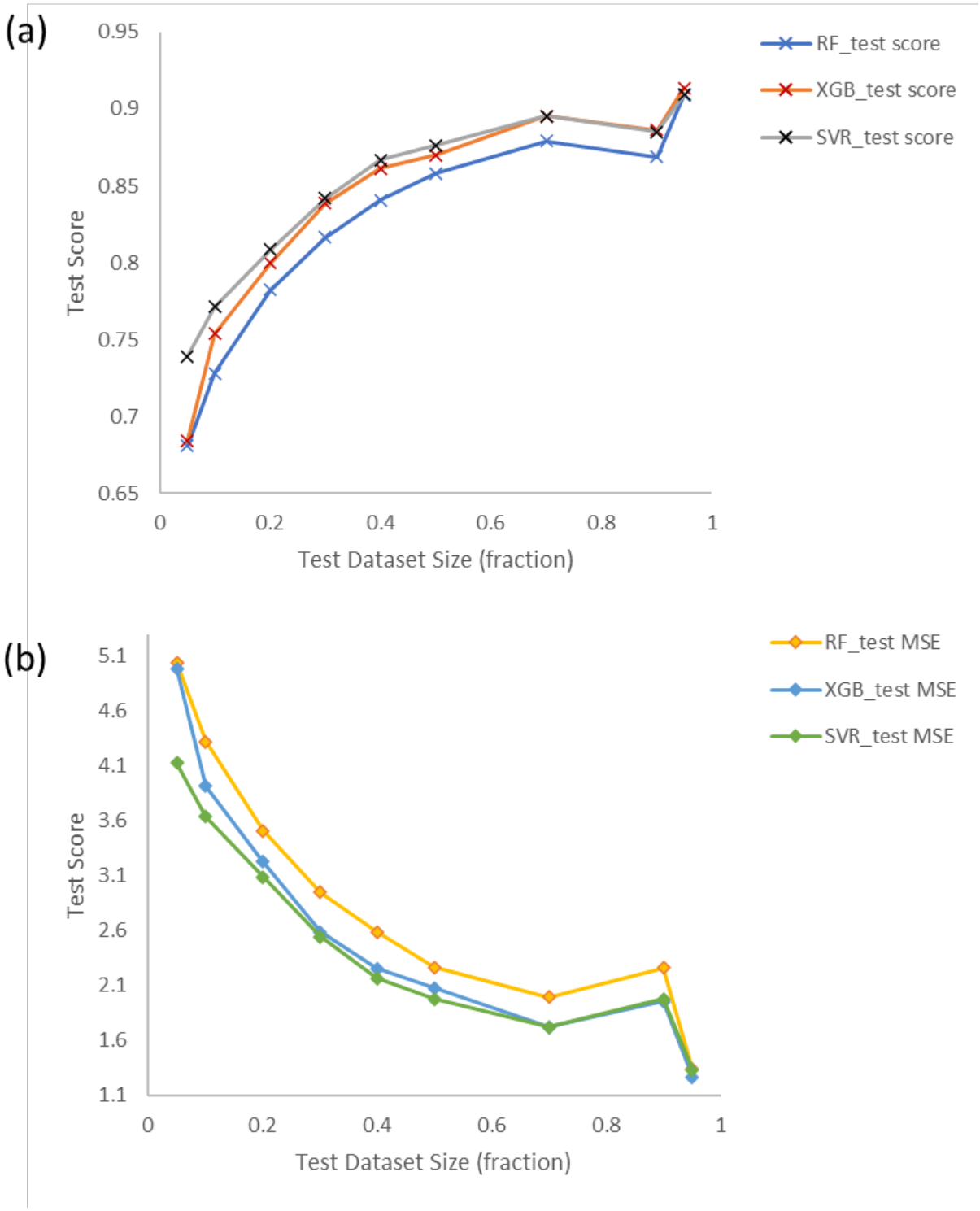


Figure S3: Effect of dataset size on (a) the accuracy/score and (b) mean squared error of ML modeling using RF , XGB and SVR for Polarization solubility parameter

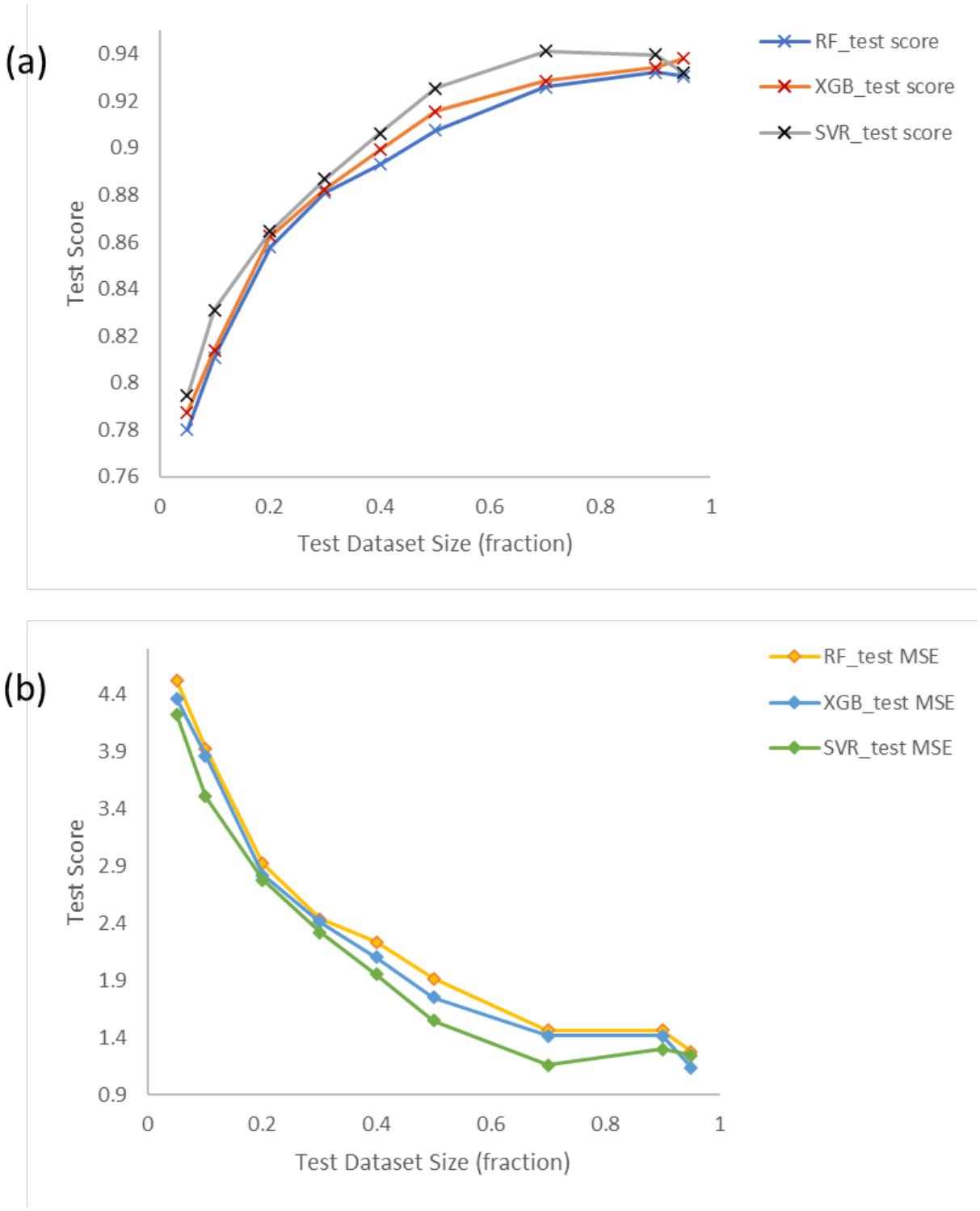


Figure S4: Effect of dataset size on (a) the accuracy/score and (b) mean squared error of ML modeling using RF , XGB and SVR for Hydrogen bonding solubility parameter

Table S2: Decision fusion results (Solubility_2 "δ_P")

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

** All the values are rounded to the second decimal*

Table S3: Decision fusion results (Solubility_3 "δ_H")

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

** All the values are rounded to the second decimal*

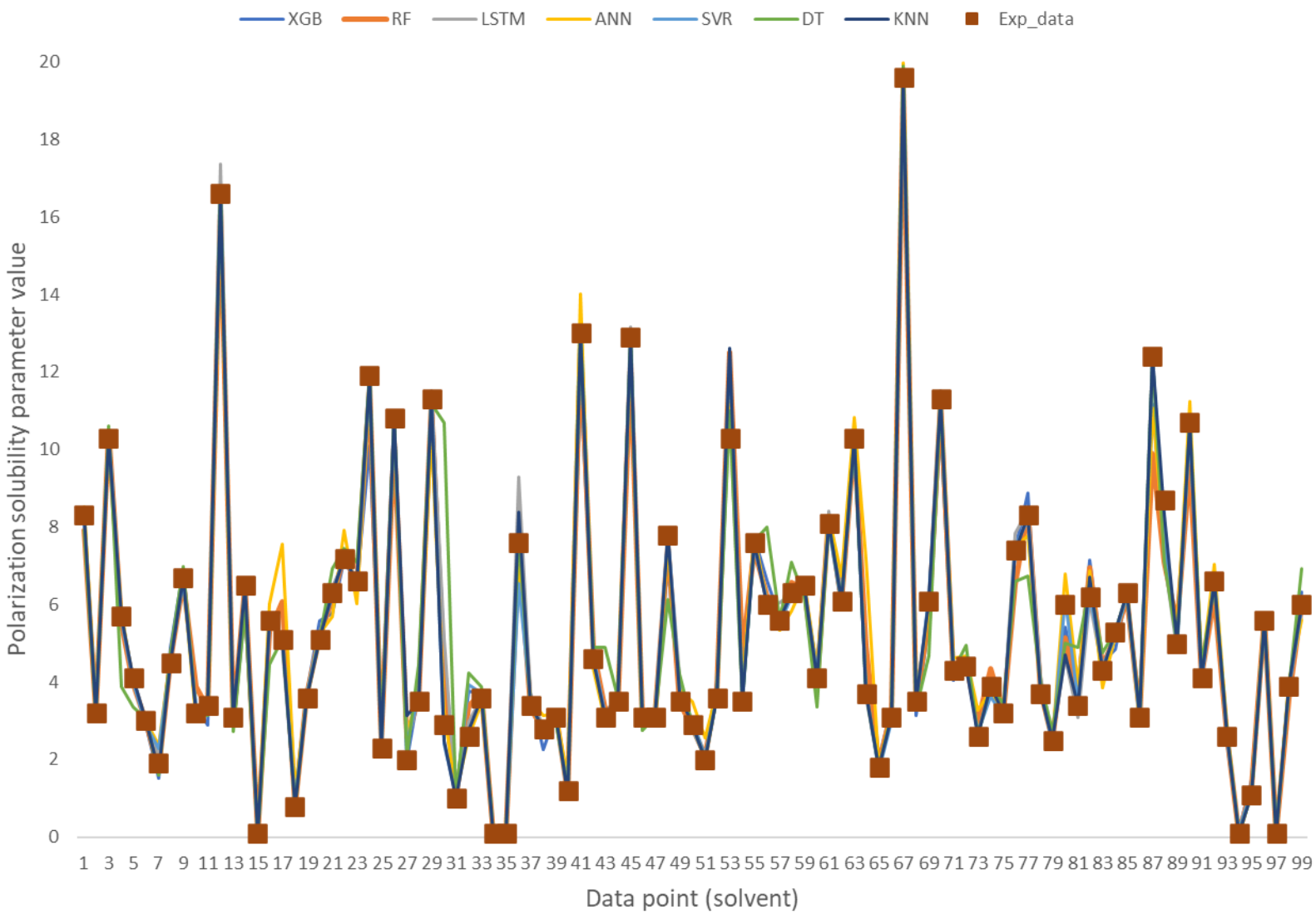


Figure S5: Results of different individual techniques (Polarization solubility parameter).

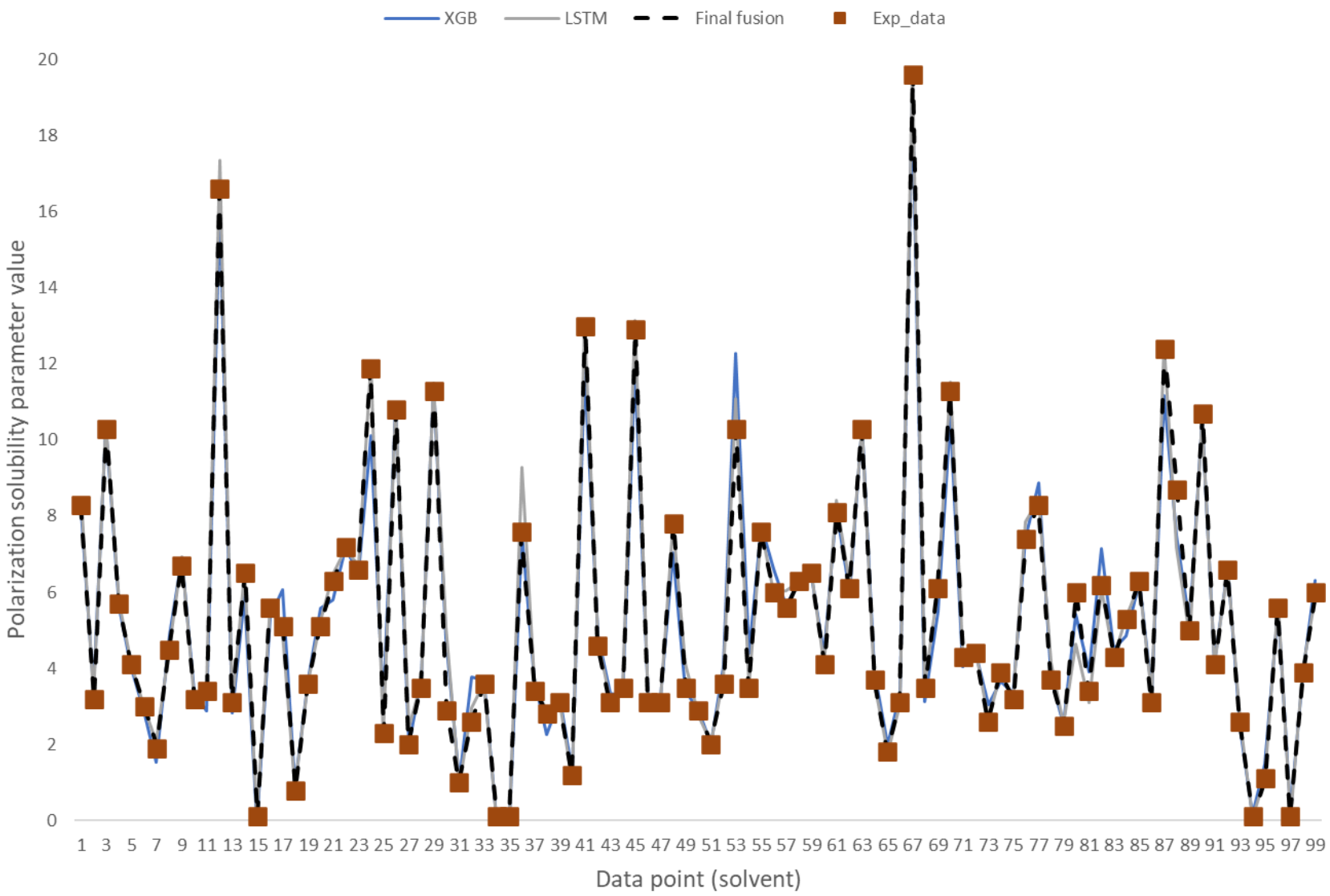


Figure S6: Results of final decision fusion vs. selected different individual techniques (Polarization solubility parameter).

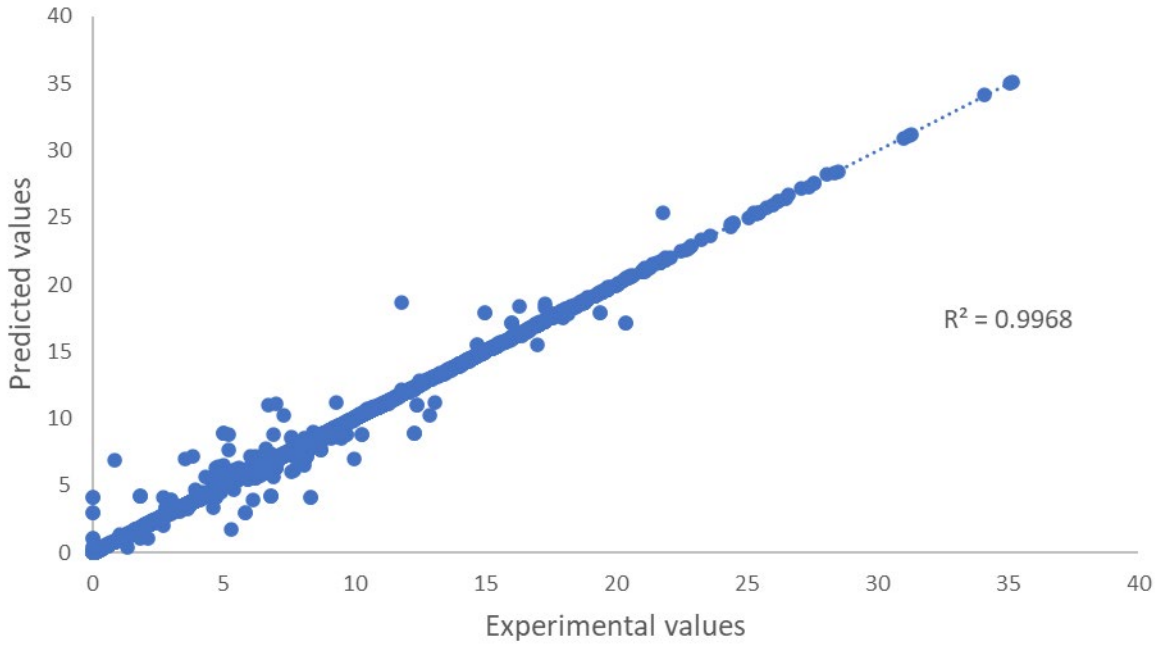


Figure S7: Predicted Vs. experimental values of polarization solubility parameter (final decision fusion results).

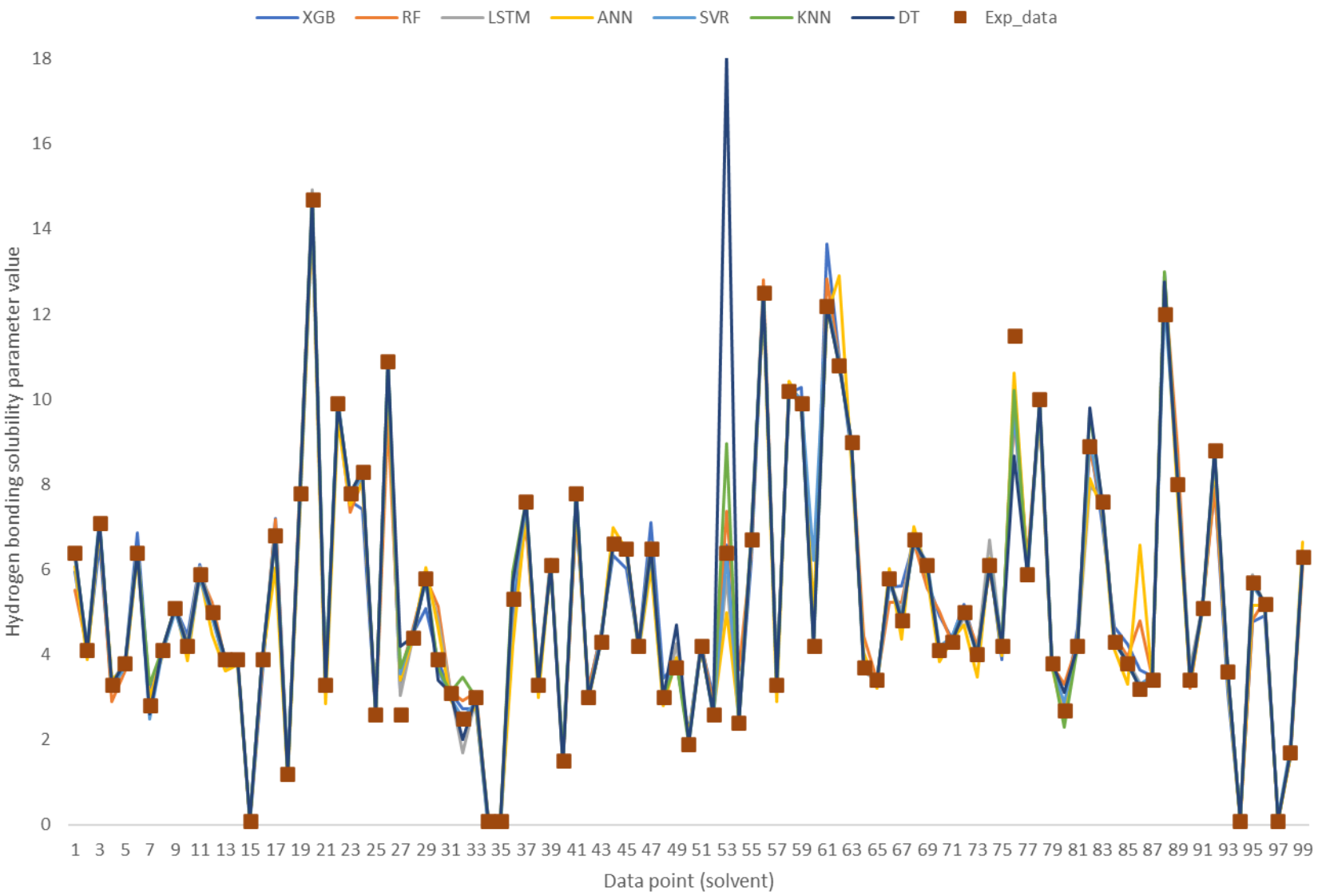


Figure S8: Results of different individual techniques (Hydrogen bonding solubility parameter).

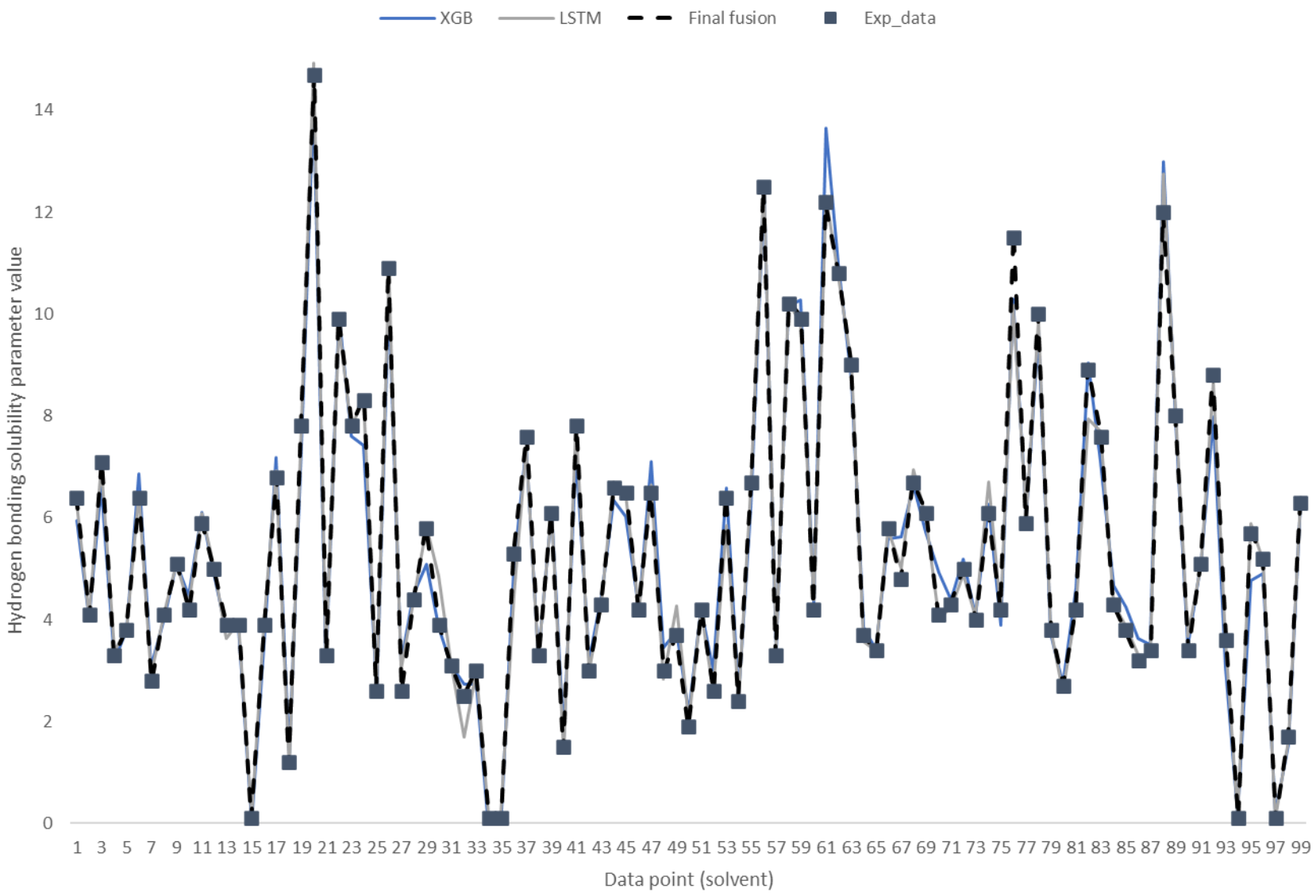


Figure S9: Results of final decision fusion vs. selected different individual techniques (Hydrogen bonding solubility parameter).

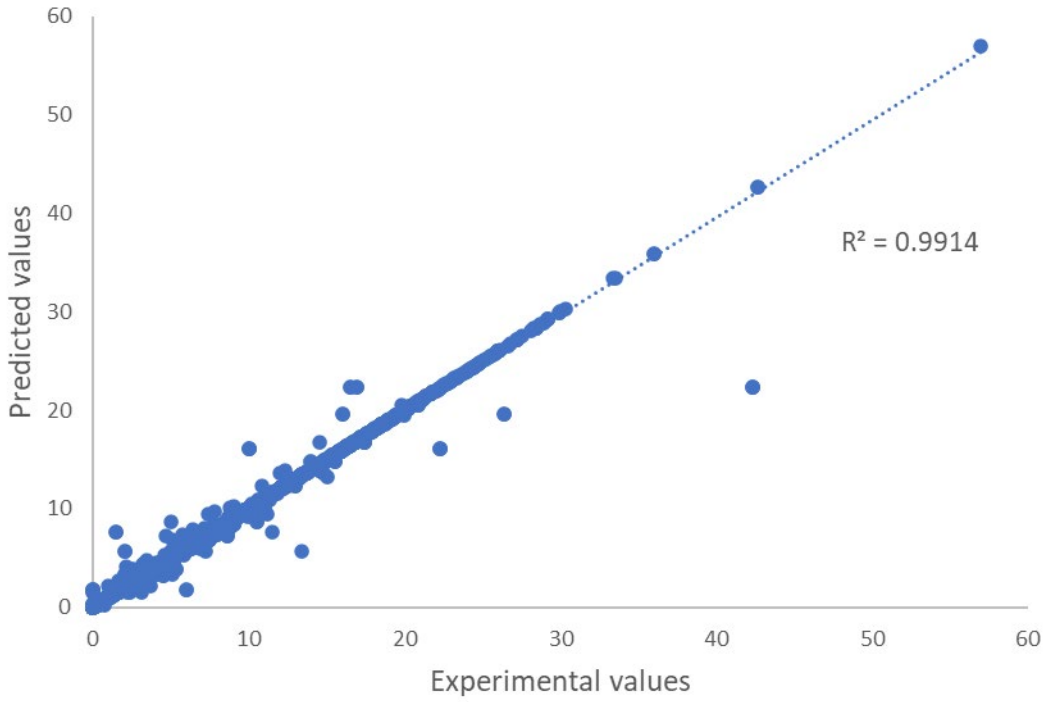
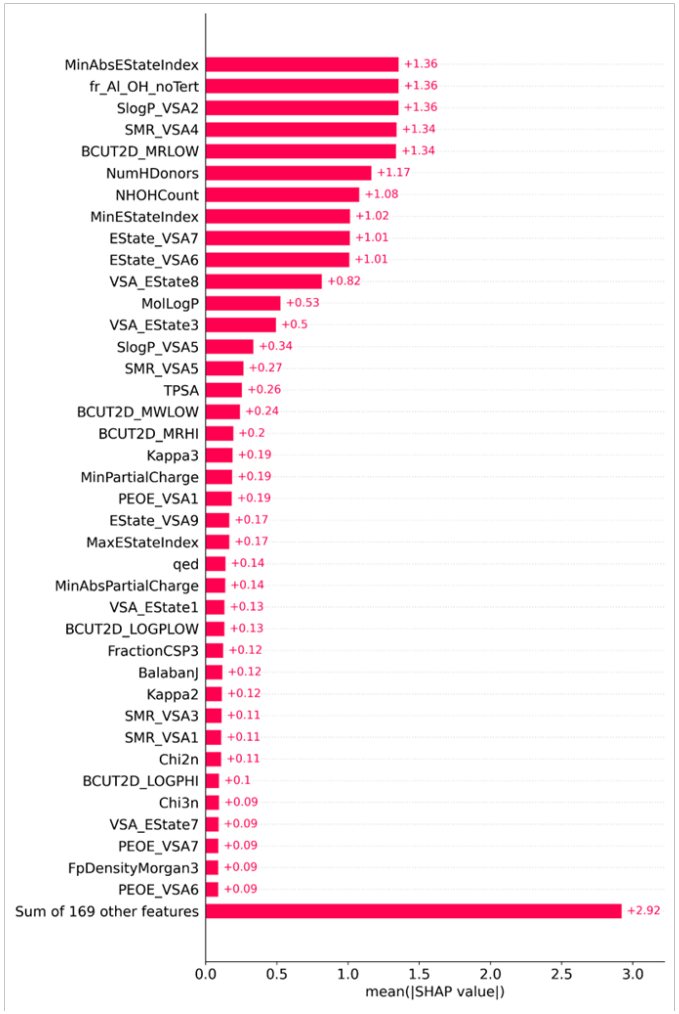


Figure S10: Predicted Vs. experimental values of hydrogen bonding solubility parameter (final decision fusion results).

(a)



(b)

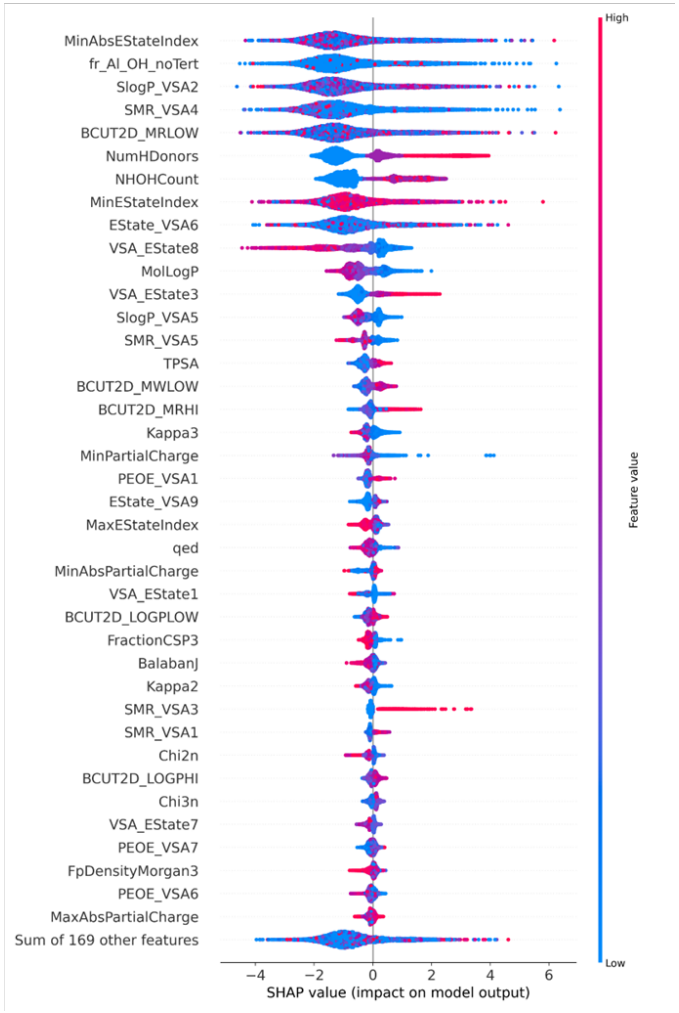


Figure S11: SHAP values of sugar cane bagasse-based lignin solvents classification based on RED (**Descriptors**)

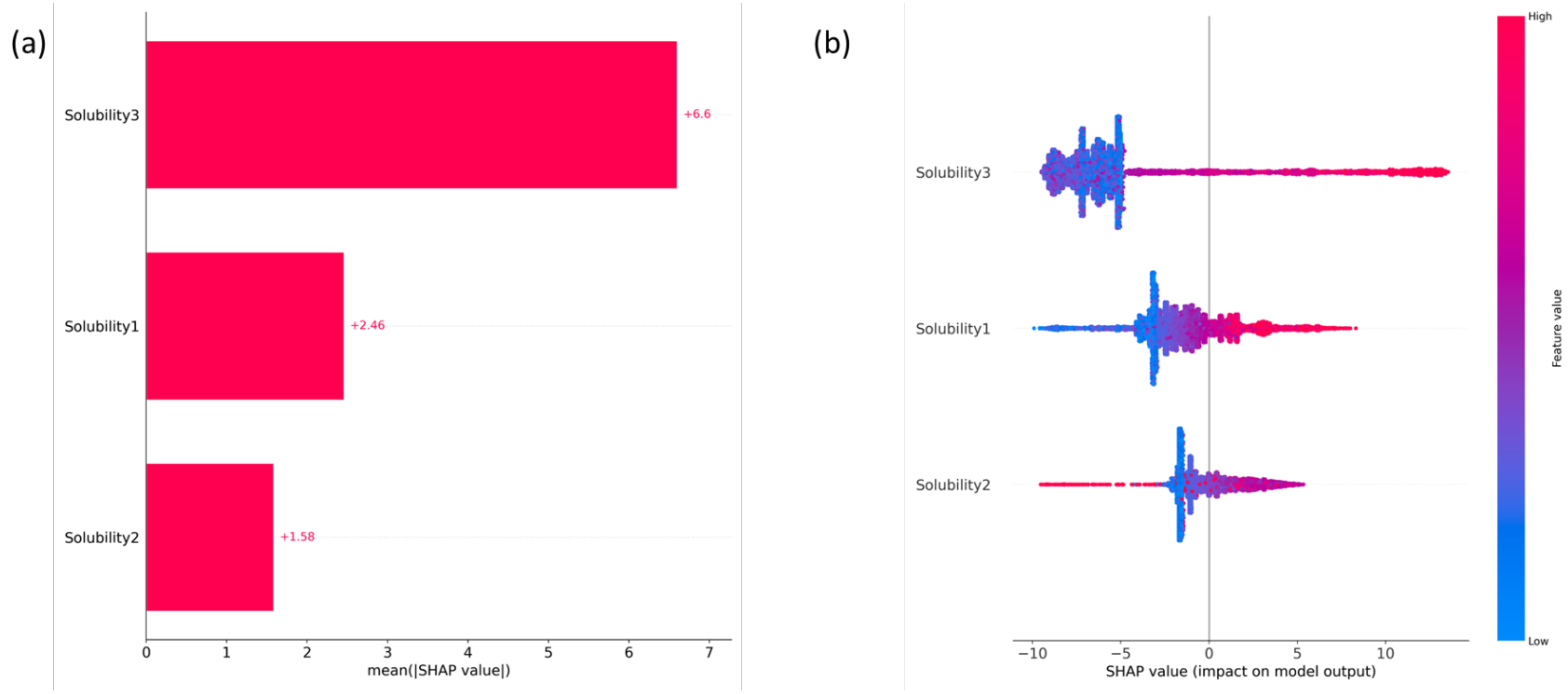


Figure S12: SHAP values of sugar cane bagasse-based lignin solvents classification based on RED (**Hansen solubility parameters**)

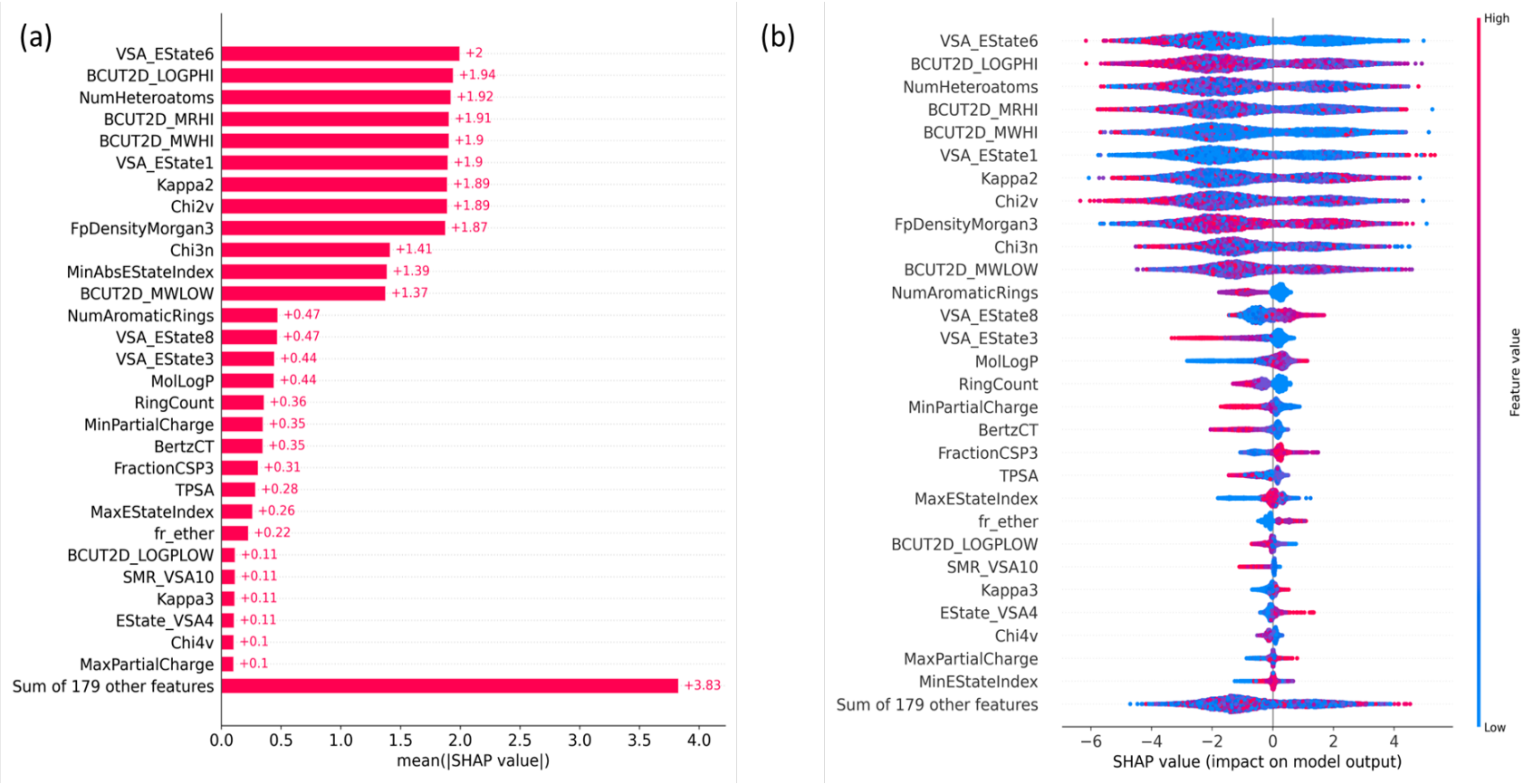


Figure S13: SHAP values of CO_2 solvents classification based on RED **(Descriptors)**

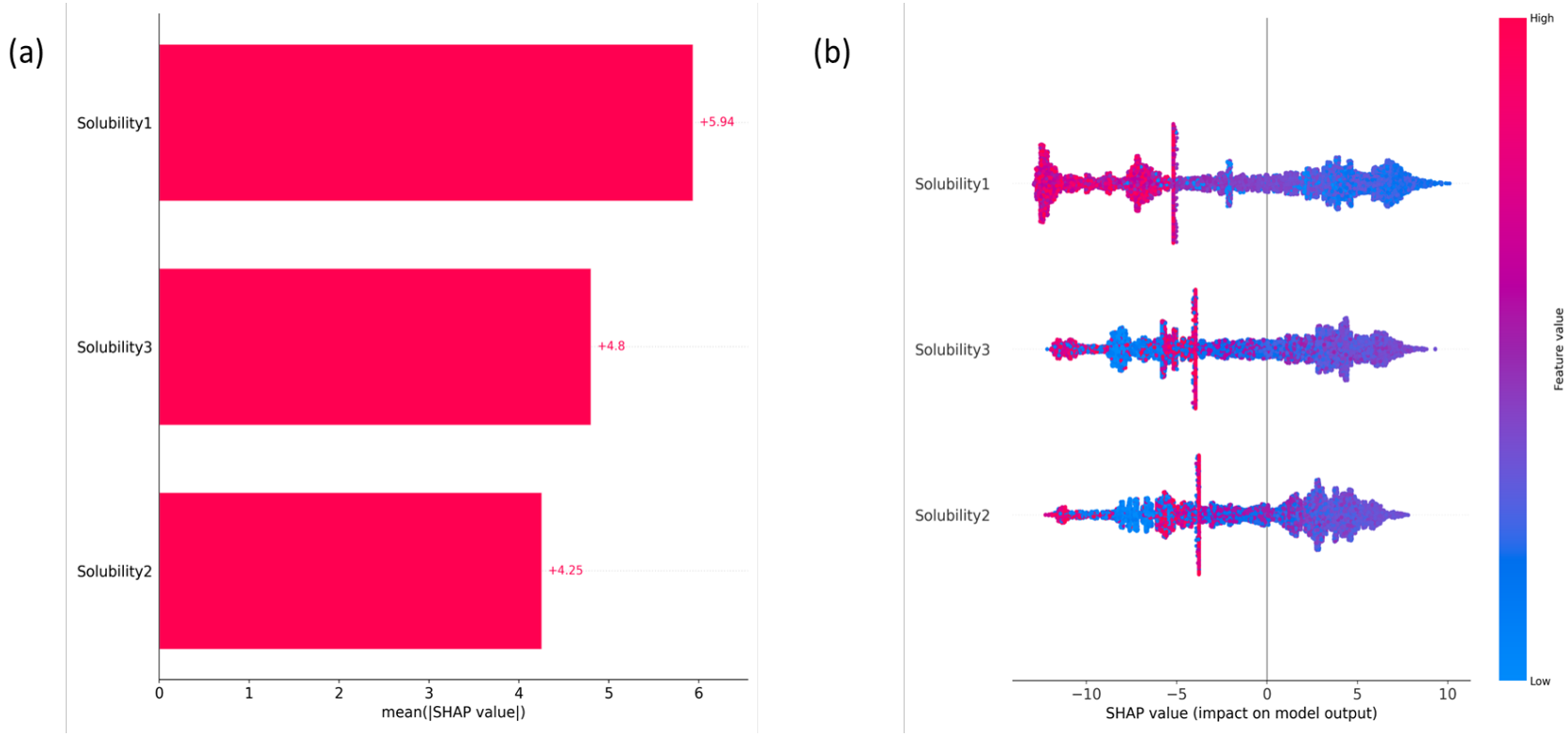


Figure S14: SHAP values of CO_2 solvents classification based on RED **(Hansen solubility parameters)**

Table S4: Comparison with selected previous studies considering ML models to predict solubility parameters

Model(s)	Solubility parameter(s)/representation(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	$R^2 = 0.69\text{-}0.83$ RMSE = 1.02-2.83	Limited	[1]
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	[2]
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	[3]
LightGBM	Hildebrand solubility and HSPs-based RED Range (0 to 15)	SMILES codes and extracted molecular descriptors	55272 samples representing 81 polymers	$R^2 = 0.86\text{-}0.94$	Limited	[4]

			and 1221 solvents interactions			
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	[5]
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	[6]
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	[7]
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	[8]
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	Classification accuracy = 93%	No	[9]

			representing 24 solvent and 4595 polymers			
Gradient Boosting and Bagging regressors (among 40 ML techniques)	HSPs Range (0 to 30)	Molecular descriptors & fingerprints	252 points representing various green solvents	R^2 = 0.35-0.78	No	[10]
NLP-based models (ChemBERTa)	HSPs Range (0 to 40)	SMILES codes	1200 points representing organic molecules	R^2 = 0.41-0.73 RMSE = 0.83-2.83	No	[11]
ANN, SVM, RF, ExtraTrees, Bagging regressor and GPR	LogS Range (-12 to 4)	14 descriptors representing the interactions between solutes and solvents (water & organic compounds) in addition to molecular properties of both solutes & solvents. They include solvation energy, solute melting point and solvent accessible surface area	Over 2500 points representing the interactions between different solutes and solvents (water, benzene, ethanol and acetone)	R^2 = 0.42-0.93 RMSE = 0.54-0.83	Limited	[12]
SVM	Hildebrand solubility Range (15 to 35) HSPs-related RED Range (0.3 to 2.5)	Heat capacity, TPSA, melting temperature, molar volume, density, molecular weight and system temperature	548 points representing the interaction between 100 ionic liquids and different metal oxides	R^2 = 0.98-0.99 RMSE = 0.03-0.05	No	[13]

			including zinc oxide (ZnO)			
Multilinear regression (MLR) and kernel ridge regression (KRR)	Hildebrand solubility represented by experimental heat of vaporization	15 molecular descriptors including number of heavy atoms, chemical hardness, atomization energy and electronegativity	Below 100 points representing the interaction between 61 small molecule solvents and 16 polymers	R^2 = 0.82 RMSE = 4.35	No	[14]
RF, Conditional Inference trees (CTREE) & Partial Least Squares Regression (PLSR)	Mole fraction Range (0 to 1)	Quantum chemical and molecular orbital-based descriptors of ionic liquids including polarizabilities and charge partial surface areas (CPSA)	10848 points representing solubility of CO_2 in 185 different ionic liquids at various temperatures & pressures	R^2 = 0.35-0.96 RMSE = 0.004-0.21	Limited	[15]
Gradient Boosting (GB) and RF	Material solvent extraction yield Range (0 “0%” - 1 “100%”)	Molecular descriptors/fingerprints of ionic liquids, herbaceous biomass composition (lignin, cellulose & hemicellulose percentages) and solvent extraction conditions (temperature, pressure & solvent concentration)	110 points representing the efficiency (yield) of lignin extraction from herbaceous biomasses by ionic liquids as sustainable solvents at different conditions	R^2 = 0.63-0.73 MSE = 0.02-0.03	Yes	[16]

Additional References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] A. Mahmood, Y. Sandali, and J.-L. Wang, "Easy and fast prediction of green solvents for small molecule donor-based organic solar cells through machine learning," *Phys. Chem. Chem. Phys.*, vol. 25, no. 15, pp. 10417–10426, 2023.

[11] 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.

[12] S. Boobier, D. R. J. Hose, A. J. Blacker, and B. N. Nguyen, "Machine learning with physicochemical relationships: solubility prediction in organic solvents and water," *Nat. Commun.*, vol. 11, no. 1, p. 5753, 2020.

[13] 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.

[14] M. Chi, R. Gargouri, T. Schrader, K. Damak, R. Maâlej, and M. Sierka, "Atomistic descriptors for machine learning models of solubility parameters for small molecules and polymers," *Polymers (Basel).*, vol. 14, no. 1, p. 26, 2021.

[15] V. Venkatraman and B. K. Alsberg, "Predicting CO2 capture of ionic liquids using machine learning," *J. CO2 Util.*, vol. 21, pp. 162–168, 2017.

[16] K. Baran, B. Barczak, and A. Kloskowski, "Modeling lignin extraction with ionic liquids using

machine learning approach," *Sci. Total Environ.*, p. 173234, 2024.