Supporting Information for Capturing Molecular Interactions in Graph Neural Networks: A Case Study in Multi-Component Phase Equilibrium

Shiyi Qin,[a] Shengli Jiang,[a] Jianping Li[a]
Prasanna Balaprakash,[b] Reid C. Van Lehn,[a] and Victor M. Zavala,[a]*

All data and scripts needed to reproduce the results of the manuscript are available as open-source code in the GitHub repo https://github.com/zavalab/ML/tree/SolvGNN.

S1 Solvent Categorization

Each solvent was categorized as one of the 22 solvent categories based on its primary functional group. We first defined a list of functional groups and specified their order of priority. The list was adapted from the conventional functional group hierarchy for organic nomenclature[1] with a few modifications (*e.g.*, addition of arbitrary function groups such as "Sulfur" and "Silicon" that capture solvents with certain elements but were not identified as the precedent functional groups) to accommodate our solvent data set. SMARTS, or SMILES arbitrary target specification, was used to search for the functional groups within a molecule. For each functional group, the identifiable SMARTS strings were gathered from the source documentation from a collection of cheminformatics tools, including RDKit,[2] OpenBabel,[3] and Daylight Chemical Information Systems.[4] For each solvent, the searching algorithm goes through the functional group list from high priority to low priority and returns the category with which a SMART pattern is first matched. For example, an ester may also be identified as a ketone due to the carbonyl group, but since ester is defined to have higher order of priority than ketone, the ester-like molecule is eventually categorized as an ester. Although the included SMARTS helped categorize most of the solvents, there were a few exceptions; for these, we manually categorized them based on their nomenclatures. Despite its limitation in categorizing complex chemical structures that require functional group identification by human expertise, the algorithm provides a simple method for fast solvent categorization and can be easily modified for other studies.

[a]Department of Chemical and Biological Engineering, University of Wisconsin - Madison, 1415 Engineering Dr, Madison, WI 53706, USA
[b]Mathematics and Computer Science Division & Leadership Computing Facility, Argonne National Laboratory, 9700 S Cass Ave, Lemont, IL 60439, USA
*Corresponding Author: victor.zavala@wisc.edu.

S2 Additional Details on Model Training and Validation

Table S1: List of atom features.[5]

Node Feature Index	Meaning
1-43	One-hot-encoded atom type
44-54	One-hot-encoded atom degree (0-10)
55-61	One-hot-encoded number of implicit H's (0-6)
62	Formal Charge
63	Number of radical electrons
64-68	One-hot-encoded atom hybridization (SP, SP2, SP3, SP3D, SP3D2)
69	Whether the atom is aromatic
70-74	One-hot-encoded number of total H's (0-4)

Table S2: List of hyperparameters.

Hyperparameter	SolvGNN	SolvGCN	SolvCAT
# local graph convolution layers	2	2	2
# local hidden layer size	256	256	256
# global graph convolution layers	1	1	-
# global hidden layer size	256	256	-
# readout layers	2	2	2
# readout hidden layer size	256	256	256
# trainable weights	2.8M	283K	283K(binary)/349K(ternary)
Optimizer	Adam	Adam	Adam
Batch size	100	100	100
Learning rate	0.001	0.001	0.001
Loss function	MSE	MSE	MSE

S3 Model Comparison Between Data with or without $\ln \gamma^{\infty}$

Here, we tested SolvCAT, SolvGCN, and SolvGNN on binary mixtures and compared two cases, one trained on a data set with infinite dilution activity coefficients and one without. In both cases, SolvGNN performs the best, followed by SolvCAT and SolvGCN. Although the cross-validation RMSE increased slightly when we train the models with infinite dilution activity coefficients, the R^2 and MAE are comparable. As discussed in the main text, training with infinite dilution activity coefficients improved predictions at extreme concentrations and therefore was kept as the benchmark model for analysis in the manuscript.

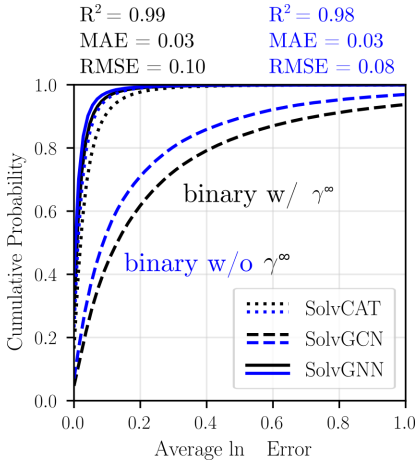

Figure S1: Cumulative frequency plots for SolvCAT, SolvGCN, and SolvGNN trained and validated on data with (black) and without (blue) infinite dilution activity coefficients.

S4 Baseline Model using XGBoost

We used Extreme Gradient Boosting (XGBoost)[6] to develop a baseline model for comparison to the SolvGNN. XGBoost is a decision tree-based model that incorporates the idea of ensemble learning and gradient boosting. It is accurate and scalable due to the parallelization of tree building and has been used extensively for benchmarking and comparing with deep learning models in molecular property predictions.[7] In our study, the XGBoost model was implemented using the Python package XGBoost (version 1.5.0). The major hyperparameters we tuned includes the number of estimators (100, **300**), learning rate (0.1, **0.2**), and max depth of the trees (4, **8**). For the input data representation, we used Morgan fingerprints[8] concatenated with mole fractions. Input data from the binary mixtures (with infinite dilution activity coefficients) were used to train the model. The same data splitting method used when training the graph neural nets was incorporated with random order switching of the components during training; the cross-validation MAE of this model is 0.21, which is substantially higher than the MAE of SolvGNN (0.03), and leads to less accurate predictions as illustrated by the parity plot below.

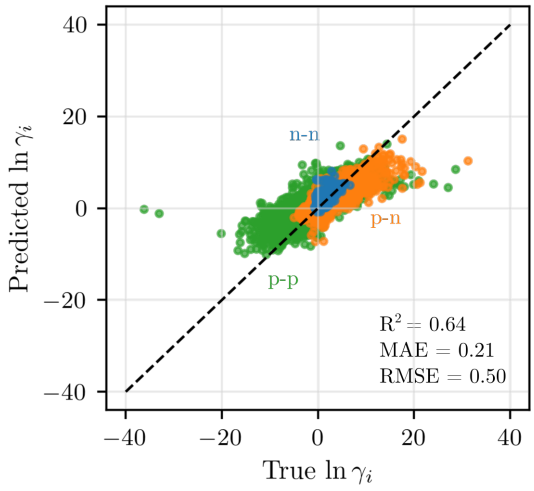

Figure S2: Parity plots for activity coefficient predictions (including infinite dilution activity coefficients) of binary mixtures using XGBoost. The same data splitting was used during training and validation. Individual $\ln \gamma_i$'s with the true (COSMO-RS) and predicted (XGBoost) values from CV are displayed. The points are colored by the type of mixtures following the naming convention of the main text.

S5 Convergence Profiles for Robustness Study

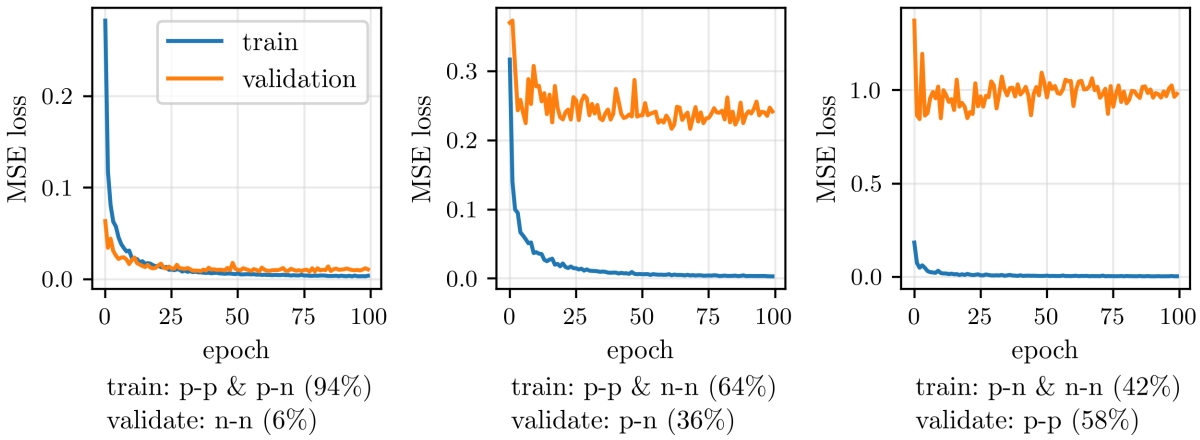

Figure S3: Convergence plots. MSE loss is plotted against epoch for each of the three CV folds. In each CV fold, SolvGNN was trained on two of the mixture types and validated on the rest. The mixture types and percentage of the data included for training/validation are listed underneath each figure. p-p stands for polar-polar; p-n stands for polar-nonpolar, and n-n stands for nonpolar-nonpolar.

S6 Significance of Physics-informed Edge Feature

To illustrate that the physics-informed edge features (H-bonds) are non-trivial in the proposed architecture, we conducted an experiment where we set all the intermolecular interactions to 1 and applied the same network. The results are shown below in Table S3. The performance decreased by 15% and 9% for CV MSE and CV MAE compared to the proposed SolvGNN where H-bond information is incorporated, suggesting the implementation of the global interaction network with H-bond information as edge features plays a role in improving the performance.

Table S3: Comparison between physics-informed edge features and non physics-informed edge features of SolvGNN.

	CV MSE	CV MAE
Proposed SolvGNN	0.0088	0.033
SolvGNN with all interaction edges set to 1	0.0102	0.036
% difference	+15%	+9%

S7 Performance Metrics

The performance metrics in Table 2 are the same metrics used by Medina et al.,[9] where the unscaled γ^{∞} values were used for evaluation. Each metric is defined in the subsections that follow.

S7.1 Mean Absolute Error (MAE)

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\gamma_i^{\infty} - \hat{\gamma}_i^{\infty}|$$

S7.2 Standard Deviation of the Errors of Prediction

$$SDEP = \sqrt{\frac{\sum_{i=1}^{N}(r_i - \mu_r)^2}{N}}$$

where $r_i = |\gamma_i^{\infty} - \hat{\gamma}_i^{\infty}|$ and $\mu_r = \frac{1}{N}\sum_{i=1}^{N} r_i$

S7.3 Mean Squared Error (MSE)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\gamma_i^{\infty} - \hat{\gamma}_i^{\infty})^2$$

S7.4 Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\gamma_i^{\infty} - \hat{\gamma}_i^{\infty})^2}$$

S7.5 Mean Absolute Percentage Error (MAPE)

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|\gamma_i^{\infty} - \hat{\gamma}_i^{\infty}|}{\gamma_i^{\infty}} \times 100\%$$

S7.6 Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\gamma_i^{\infty} - \hat{\gamma}_i^{\infty})^2}{\sum_{i=1}^{N}(\gamma_i^{\infty} - \mu_{\gamma})^2}$$

where $\mu_{\gamma} = \frac{1}{N}\sum_{i=1}^{N}\gamma_i^{\infty}$

S8 Phase Diagram Comparison with Experimental Data

Figure 5 in the main text shows binary phase diagrams to highlight that SolvGNN-predicted activity coefficients lead to good agreement with phase diagrams generated by COSMO-RS and Aspen (using the UNIFAC approach). To highlight that these approaches are also suitable for reproducing experimentally determined phase diagrams, Figure S4 compares binary phase diagrams between two experimental data sets[10] and the computational predictions for a cyclohexane-ethanol mixture. The two experimental data sets are at $T = 293.15$ and $T = 303.15$ K, respectively, while the computational results are at $T = 298$. The COSMO-RS, Aspen, and SolvGNN predicted data lie in between the two experimental data sets and exhibit similar patterns and azeotrope compositions, highlighting their applicability.

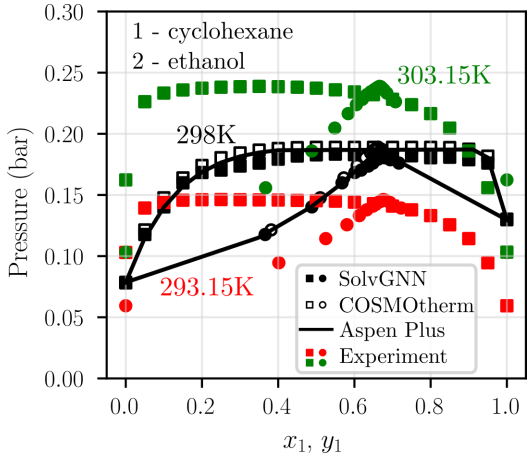

Figure S4: P-x-y comparison between SolvGNN and experimental data.

S9 Additional Binary Phase Diagrams

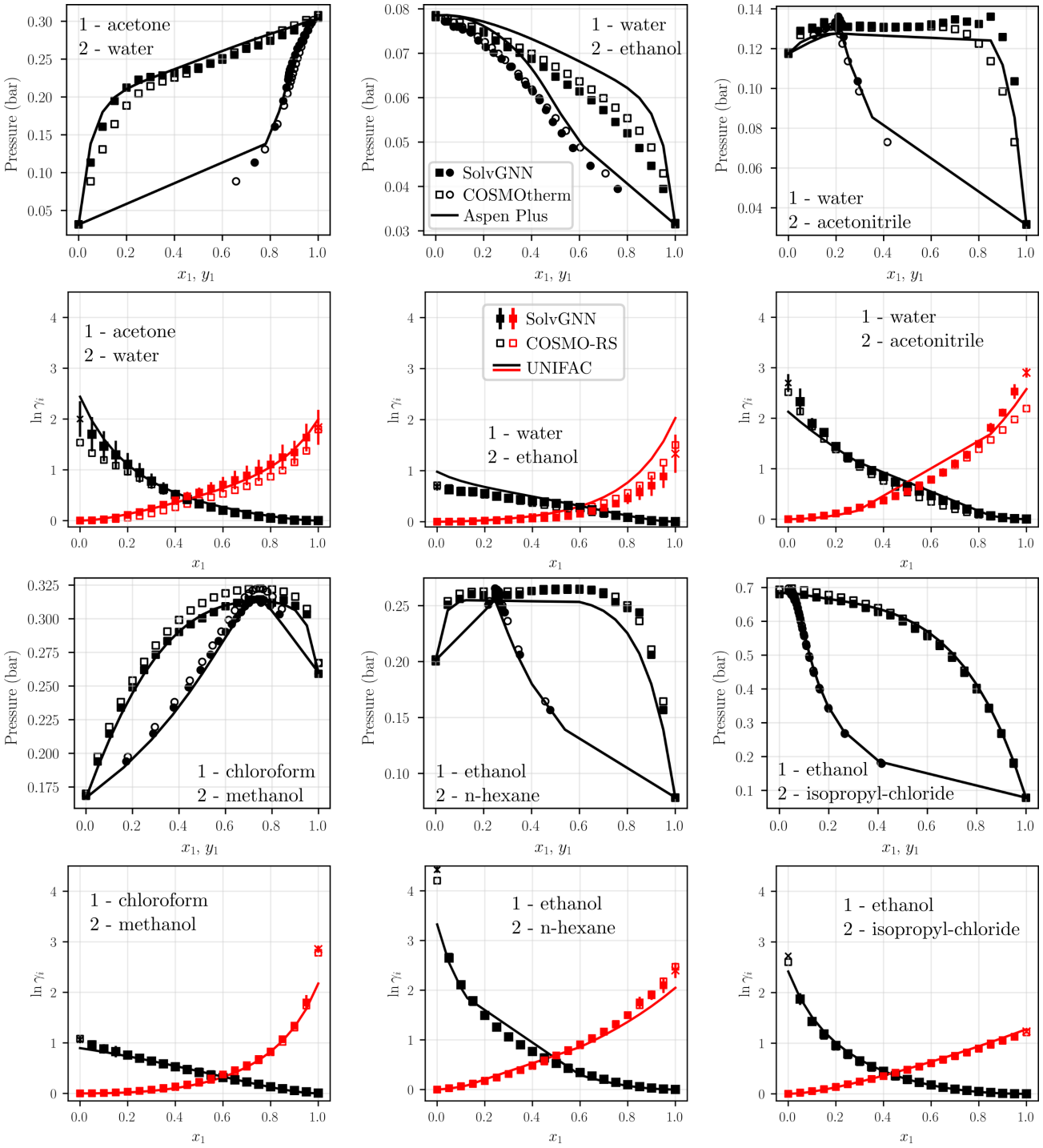

Figure S5: Six example P-x-y phase diagrams generated from SolvGNN displayed along with their activity coefficient predictions (shown below each P-x-y diagram). The phase diagrams are compared with those generated from two other state-of-the-art tools, including COSMOtherm that implements COSMO-RS[11] and Aspen Plus that implements UNIFAC.[12] The vapor compositions are represented as circles and liquid compositions are represented as squares. "x" denotes activity coefficients at infinite dilution. In all calculations, the $\ln \gamma_i$'s are obtained by averaging the predictions of SolvGNN trained from each CV fold, and standard deviations are visualized as error bars.

S10 Numerical Comparison of VLE Data for Water-Acetone-MIBK

To test whether vapor-liquid equilibrium data could also be reliably obtained for a ternary system, we computed isobars using modified Raoult's Law for the ternary water(1)-acetone(2)-MIBK(3) mixture using SolvGNN and compared to results from COSMO-RS. To do so, we sampled liquid-phase compositions (x_i) within the ternary space, calculated activity coefficients for all three components, and used modified Raoult's Law to determine the equilibrium bubble pressure (denoted as P) in Table S4. Pressures (either from SolvGNN or COSMO-RS) within 2% of each target isobar pressure and corresponding vapor-phase compositions (y_i) are included in the table below alongside corresponding COSMO-RS predictions (from COSMOtherm). The comparison shows that in general SolvGNN and COSMO-RS are in excellent agreement, with the MAE shown at bottom in the table.

Table S4: Numerical comparison between SolvGNN and COSMO-RS of VLE data for a ternary mixture water(1)-acetone(2)-MIBK(3).

	Sampled Composition			**Prediction Comparison** ˆindicates SolvGNN results (COSMO-RS otherwise)							
	x_1	x_2	x_3	$\hat{y}_1$	y_1	$\hat{y}_2$	y_2	$\hat{y}_3$	y_3	$\hat{P}$ (bar)	P (bar)
P∼0.15 bar	0.00	0.40	0.60	0.00	0.00	0.89	0.89	0.11	0.10	0.15	0.16
	0.05	0.35	0.60	0.07	0.08	0.81	0.81	0.12	0.11	0.14	0.15
	0.10	0.35	0.55	0.12	0.12	0.77	0.78	0.11	0.11	0.15	0.15
	0.40	0.30	0.30	0.23	0.20	0.66	0.70	0.11	0.10	0.14	0.15
	0.45	0.30	0.25	0.22	0.20	0.67	0.71	0.11	0.09	0.14	0.15
	0.55	0.30	0.15	0.19	0.18	0.71	0.74	0.10	0.08	0.15	0.16
	0.60	0.25	0.15	0.21	0.20	0.67	0.70	0.13	0.10	0.14	0.15
	0.70	0.20	0.10	0.24	0.19	0.63	0.69	0.12	0.12	0.13	0.15
	0.70	0.25	0.05	0.21	0.15	0.73	0.78	0.06	0.06	0.15	0.18
	0.75	0.25	0.00	0.22	0.13	0.78	0.87	0.00	0.00	0.15	0.20
	0.80	0.15	0.05	0.31	0.19	0.61	0.68	0.09	0.13	0.12	0.15
P∼0.20 bar	0.00	0.55	0.45	0.00	0.00	0.93	0.93	0.07	0.06	0.19	0.20
	0.05	0.55	0.40	0.04	0.05	0.89	0.89	0.07	0.06	0.20	0.20
	0.10	0.55	0.35	0.07	0.07	0.87	0.87	0.06	0.06	0.20	0.20
	0.25	0.50	0.25	0.14	0.11	0.81	0.84	0.05	0.05	0.20	0.20
	0.30	0.50	0.20	0.14	0.11	0.81	0.84	0.04	0.05	0.20	0.20
	0.35	0.50	0.15	0.14	0.12	0.82	0.84	0.04	0.04	0.20	0.21
	0.40	0.50	0.10	0.14	0.11	0.83	0.85	0.03	0.03	0.20	0.21
	0.45	0.45	0.10	0.15	0.13	0.81	0.84	0.04	0.04	0.19	0.20
	0.55	0.45	0.00	0.14	0.11	0.86	0.89	0.00	0.00	0.20	0.23
	0.60	0.35	0.05	0.16	0.13	0.80	0.83	0.04	0.04	0.17	0.20
MAE	-	-	-	0.03		0.03		0.01		0.01	

References

[1] Libretexts. 18.2: Functional group order of precedence for organic nomenclature, Aug 2020. https://chem.libretexts.org/Bookshelves/Organic_Chemistry/Organic_Chemistry_I_(Cortes)/18%3A_Important_Concepts_in_Alkyne_Chemistry/18.02%3A_Functional_Group_Order_of_Precedence_For_Organic_Nomenclature [Online; accessed 2022-04-19].

[2] Greg Landrum. Rdkit documentation. Release, 1(1-79):4, 2013.

[3] Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. Open babel: An open chemical toolbox. Journal of cheminformatics, 3(1):1–14, 2011.

[4] Daylight Chemical Information Systems. Smarts examples. https://www.daylight.com/dayhtml_tutorials/languages/smarts/smarts_examples.html [Online; accessed 2022-04-19].

[5] Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, et al. Deep graph library: A graph-centric, highly-performant package for graph neural networks. arXiv preprint arXiv:1909.01315, 2019.

[6] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, pages 785–794, 2016.

[7] Hao Tian, Rajas Ketkar, and Peng Tao. Accurate admet prediction with xgboost. arXiv preprint arXiv:2204.07532, 2022.

[8] Harry L Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. Journal of chemical documentation, 5(2):107–113, 1965.

[9] Edgar Ivan Sanchez Medina, Steffen Linke, Martin Stoll, and Kai Sundmacher. Graph neural networks for the prediction of infinite dilution activity coefficients. Digital Discovery, 2022.

[10] Dortmund data bank, Aug 2022. https://www.ddbst.com.

[11] Andreas Klamt, Frank Eckert, and Wolfgang Arlt. Cosmo-rs: an alternative to simulation for calculating thermodynamic properties of liquid mixtures. Annual review of chemical and biomolecular engineering, 1:101–122, 2010.

[12] Aage Fredenslund, Russell L Jones, and John M Prausnitz. Group-contribution estimation of activity coefficients in nonideal liquid mixtures. AIChE Journal, 21(6):1086–1099, 1975.