Journal of Cheminformatics

RESEARCH Open Access

Multi-fidelity graph neural networks for predicting toluene/water partition coefficients

Thomas Nevolianis[1], Jan G. Rittig[2], Alexander Mitsos[2,3,4] and Kai Leonhard[1*]

Abstract

Accurate prediction of toluene/water partition coefficients of neutral species is crucial in drug discovery and separation processes; however, data-driven modeling of these coefficients remains challenging due to limited available experimental data. To address the limitation of available data, we apply multi-fidelity learning approaches leveraging a quantum chemical dataset (low fidelity) of approximately 9000 entries generated by COSMO-RS and an experimental dataset (high fidelity) of about 250 entries collected from the literature. We explore the *transfer learning*, *feature-augmented learning*, and *multi-target learning* approaches in combination with graph neural networks, validating them on two external datasets: one with molecules similar to training data (EXT-Zamora) and one with more challenging molecules (EXT-SAMPL9). Our results show that *multi-target learning* significantly improves predictive accuracy, achieving a root-mean-square error of 0.44 log *P* units for the EXT-Zamora, compared to a root-mean-square error of 0.63 log *P* units for single-task models. For the EXT-SAMPL9 dataset, *multi-target learning* achieves a root-mean-square error of 1.02 log *P* units, indicating reasonable performance even for more complex molecular structures. These findings highlight the potential of multi-fidelity learning approaches that leverage quantum chemical data to improve toluene/water partition coefficient predictions and address challenges posed by limited experimental data. We expect the applicability of the methods used beyond just toluene/water partition coefficients.

Scientific contribution

We investigate the benefits of transfer learning, feature-augmented learning, and multi-target learning approaches in combination with graph neural networks for the prediction of toluene–water partition coefficients. We show how a combination of a large number of cheap data from the semi-empirical COSMO-RS model with a few high-fidelity experimental data and multi-target learning efficiently leads to machine learning models with broad applicability and low uncertainties of 0.44 to 1.02 log units in the partition coefficient, depending on the test set.

Keywords Graph neural network, Multi-fidelity learning, Partition coefficients

*Correspondence:
Kai Leonhard
kai.leonhard@ltt.rwth-aachen.de
Full list of author information is available at the end of the article

Introduction

The partition coefficient log *P* of neutral species between water and an organic species is an important physical property, playing a significant role in various fields such as drug discovery [1–4] and separation processes [5, 6]. This property captures the ratio of concentrations of a chemical species in two immiscible solvents. For pharmaceutical applications, the partition coefficient of an Active Pharmaceutical Ingredient (API) indicates its hydrophobicity/hydrophilicity and is thus a critical indicator for its pharmacokinetics and physical properties of potential drug candidates [7, 8]. In separation processes, the partition coefficient between water and an organic solvent is key for determining the most effective methods for separating species impacting both the yield and purity [9–11]. While water/octanol partition coefficients of neutral species are widely available, data for water and other organic solvents, such as toluene/water are limited. Toluene/water partition coefficients offer better physiological relevance compared to water/octanol [12, 13]. Consequently, models that predict toluene/water partition coefficients for a wide spectrum of neutral species are highly desired.

Existing computational methods, such as the COnductor like Screening MOdel for Real Solvents (COSMO-RS) [14, 15], Group Contribution (GC), and Molecular Dynamics (MD) have been employed to predict toluene/water log *P* of neutral species [16–23]. Recently, the SAMPL9 blind challenge [24] allowed different groups to compare such predictive methods against a set of 16 drug-like molecules for predicting toluene/water log *P*. We also participated in the challenge using the COSMO-RS method. Among 18 contributions, we ranked second with an Root-Mean-Square Error (RMSE) of 1.24 log *P* units [25]. The best-performing method in the SAMPL9 blind challenge [24] achieved an RMSE of 1.12 log *P* units [24]. COSMO-RS is a semi-empirical model, partially physics-based, allowing application to any system, though its performance varies depending on the specific system and property being studied [26–28]. Nevertheless, COSMO-RS shows good agreement with experimental log *P* values in our dataset that we performed in this study, supporting our choice to use it for generating low-fidelity training data. Machine Learning (ML) offers new possibilities for predicting toluene/water log *P* by utilizing experimental data. Recent advances in ML such as Graph Neural Networks (GNN) models and transformers enable end-to-end learning of molecular properties directly from the structure and have demonstrated success across various applications [29–37]. The general idea is to find a representation of molecules, e.g., in the form of descriptors, strings, or graphs, which can be mapped to properties of interest by applying regressions methods. For instance, in predicting the toluene/water partition coefficient of APIs as a post-SAMPL9 study, Zamora et al. [38] used a variety of molecular descriptors—related to the topological structure and properties such as the Ghose–Crippen water/octanol partition coefficient—on which they fitted a multiple linear regression model for the 251 experimental log *P* values from their collected dataset. These 251 experimental toluene/water log *P* of neutral species [38] are currently the largest available dataset in the literature. This multiple linear regression model achieved an RMSE of 1.05 log *P* units on the test dataset and an RMSE of 0.86 log *P* units on the SAMPL9 dataset [38]. These promising results are constrained by the limited amount of training data, which may restrict the model's broader applicability and potentially its effectiveness across diverse solutes and log *P* ranges. The direct prediction of toluene/water log *P* of neutral species using ML therefore remains limited due to data scarcity, necessitating the exploration of alternative approaches.

To address scarcity of molecular property data, previous literature studies [32, 39–41] have employed various multi-fidelity learning approaches. A recent review by Qian et al. [42] summarizes the different multi-fidelity methods, suggesting that pretraining models on low fidelity data such as a large dataset derived from Quantum Chemical (QC) calculations and semi-empirical models, followed by fine-tuning with high fidelity data such as experimental data, can significantly enhance their applicability and reliability in predicting molecular properties. In particular, three multi-fidelity approaches are promising in molecular ML: *transfer learning*, *feature-augmented learning*, and *multi-target learning* [42]. *Transfer learning* leverages pretrained models to improve predictions, *feature-augmented learning* integrates predictions as additional features, and *multi-target learning* simultaneously predicts multiple related properties. To this end, to overcome the challenges posed by limited experimental data in predicting toluene/water log *P* of neutral species, we investigate these three multi-fidelity learning approaches that leverage QC and experimental data to increase the effectiveness of GNN models.

We apply various ML models and multi-fidelity learning approaches to predict the toluene/water log *P* of neutral species. Initially, we generate a low fidelity QC dataset consisting of approximately 9000 toluene/water log *P* values of neutral species using the COSMO-RS approach, which we chose due to its balance of accuracy and computational efficiency. We use this dataset to pretrain GNN models, so they encompass a wide range of chemical classes and atom types. We then fine-tune and test the pretrained GNN models with different multi-fidelity learning approaches using the high fidelity datasets of Zamora et al. [38] and the SAMPL9 blind

challenge. Specifically, a part of the Zamora dataset, comprising 212 out of 250 experimental log *P* values, is used for fine-tuning while the remaining part 38 out of 250 is reserved for testing, similar to the approach taken with the SAMPL9 dataset, which includes 16 experimental log *P* values. The Zamora dataset originally contained 251 values, but we removed one duplicate molecule that appeared as both entry 79 (Aflukin) and entry 266 (Quinine) in External-SAMPL9 (EXT-SAMPL9) [38]. Next, we compare the GNN models with a GNN trained only on the experimental data and additional semi-empirical and data-driven approaches for the prediction of toluene/water log *P*. Finally, we discuss the strengths and limitations of the different approaches. Thereby, we address how multi-fidelity strategies leveraging both QC and experimental values can play a crucial role in ML for accurately predicting molecular properties, especially when only a limited amount of experimental data is available.

Dataset

We first present the low and high fidelity datasets of toluene/water partition coefficients and describe the data splitting process for training and testing of the computational methods. An overview of the datasets is shown in Table 1. The SMILES from all molecules used in this study are provided in the supporting information as a CSV file.

Low fidelity-quantum chemical dataset

To generate the Low fidelity-quantum chemical (LF-QC) dataset of log *P* values, we initially collect molecules represented by SMILES strings from the iBonD database [43], covering a diverse range of chemical classes and atom types. The iBonD database is chosen because it contains many drug-like molecules similar to those in the experimental datasets investigated in this work while also covering a broad spectrum of chemical diversity. The molecules are selected on the basis of standard ranges of acid dissociation constants. The final selection consists of molecules, predominantly featuring substituted benzoic and phenolic acids, alkyl carboxylic acids, alkylamines, and derivatives of pyridine and aniline. We then use these SMILES strings as input to obtain the 3D geometric structures using the software RDKit [44, 45]. Next, we optimize the molecular structures obtained from RDKit at the GFN2-xTB level of theory [46]. We further refine the geometries of each molecule in the COSMO state using COSMOconf 23 [47], with the BP86/TZVPD parametrization and FINE COSMO cavity [48–50]. Finally, we calculate the $\log P_{\mathrm{tol/w}}$ values for each molecule at 25 °C and at low finite dilution (0.0002 mol%) using COSMOtherm 23 [51], based on the difference in chemical potential between the water and toluene phases. We utilize small finite fractions of the molecules in both the aqueous phase and toluene to match the solute concentration range used in the experimental studies, which is 2.0–0.5 mM [52]. The error of log *P* in the LF-QC dataset is determined by propagating the uncertainties of the solvation free energies in water and toluene using Eq. 2. Given the uncertainty of 0.45 kcal mol^{-1} [53] for the solvation free energy, the resulting error in log *P* is 0.47 log *P* units. The LF-QC dataset consists of 8891 molecules (see Table 1). The LF-QC set is not publicly available due to licensing restrictions. Consultation with the commercial software provider confirmed that data sharing is not permitted under our academic license terms. However, the $\log P_{\mathrm{tol/w}}$ values for each molecule in the LF-QC can be generated by applying the described approach to the provided SMILES strings, which are available in the supporting information as a CSV file.

Table 1 Overview of $\log P_{\mathrm{tol/w}}$ datasets used for model (pre-) training and testing

Name	Number of data points	Origin
LF-QC[a]	8891	QC
HF-Exp [38]	212	Exp.
EXT-Zamora [38]	38	Exp.
EXT-SAMPL9 [24]	16	Exp.

[a] The LF-QC set is generated in this work and is not publicly available due to licensing restrictions. We describe how to generate the LF-QC set in the text

High fidelity-experimental dataset

The High fidelity-experimental (HF-Exp) dataset is obtained from Zamora et al. [38] who determined the partition coefficients $\log P_{\mathrm{tol/w}}$ through sample titrations, following a procedure similar to that used for aqueous acid dissociation constants determination but in the presence of varying amounts of the partitioning solvent. All measurements were conducted at 25 °C under an inert gas atmosphere, with at least three titrations performed for each compound to ensure accuracy. The solute concentration range estimations are based on the details provided in the experimental study [38, 52]. While these studies do not report the uncertainty of the toluene/water partition coefficient measurements, similar methods used for octanol/water partition coefficients typically report uncertainties around 0.04 log *P* units [54]. Therefore, it is reasonable to expect a similar level of uncertainty for the toluene/water measurements. An additional uncertainty arises from the fact that experimental concentrations are not provided for individual molecules, resulting in the calculations potentially being

at slightly different concentrations. For most molecules, this difference will be negligible, but for molecules forming dimers in the toluene phase, the discrepancy can be in the order of 2 $\log P$ units [25]. The HF-Exp dataset consists of 212 molecules (see Table 1).

External Zamora and SAMPL9 datasets

The External-Zamora (EXT-Zamora) and EXT-SAMPL9 datasets are taken from previous studies [24, 38]. The experiments conducted to measure the $\log P_{\mathrm{tol/w}}$ values in these datasets follow similar protocols to those used for obtaining the HF-Exp dataset. The EXT-Zamora and EXT-SAMPL9 datasets consist of 38 and 16 molecules, respectively (see Table 1).

Dataset comparison and analysis

Figure 1a shows the density distributions of $\log P$ values for the LF-QC, HF-Exp, EXT-Zamora, and EXT-SAMPL9 datasets. The LF-QC dataset (red) shows a wide distribution range from −10 to 7, reflecting the extensive chemical diversity captured by the Quantum Mechanics (QM) dataset. The HF-Exp dataset (green) and EXT-Zamora dataset (cyan) have a more narrow and peaked distribution centered around −1 to 3 $\log P$ values, indicating that the experimental measurements are focused on a more homogenous set of species. The EXT-SAMPL9 dataset (purple) peaks around −1 to 3 $\log P$ values and 3 to 6 $\log P$ values, indicating differences in the molecules compared to the other datasets. The broad range of the LF-QC dataset shows the variability in computational predictions, while the narrower distributions of the experimental datasets (HF-Exp, EXT-Zamora, EXT-SAMPL9) reflect controlled conditions and specific chemical spaces. This variation is crucial for evaluating the performance and generalizability of predictive models across different types of data.

Figure 1b depicts the density distributions of solute molar masses for the LF-QC, HF-Exp, EXT-Zamora, and

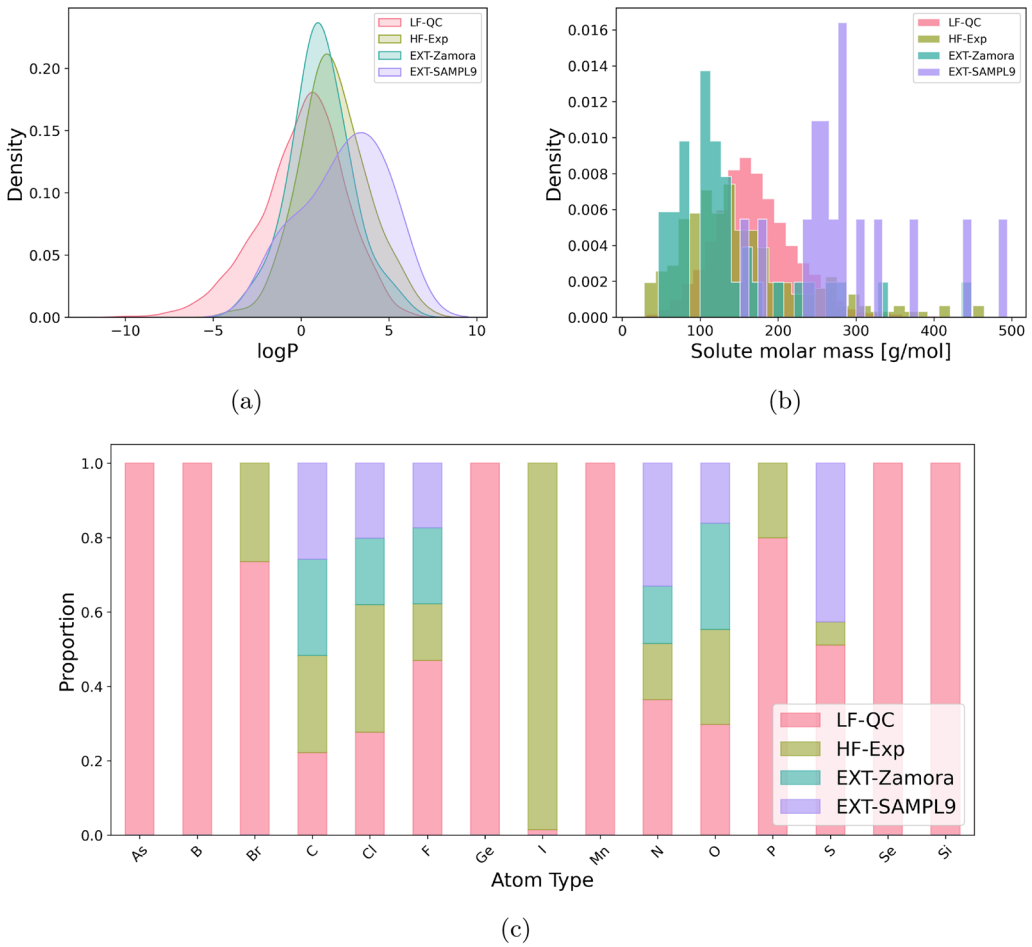

Fig. 1 Comparison of chemical properties across four datasets: LF-QC, HF-Exp, EXT-Zamora, and EXT-SAMPL9. The subfigures show **a** density plots of log *P* values, **b** density distribution of molar masses, and **c** analysis of atom type distributions

EXT-SAMPL9 datasets. The LF-QC dataset (red) shows a peak around 170 g mol^{-1}, indicating a relatively uniform distribution of molecules. The HF-Exp dataset (green) and the EXT-Zamora dataset (cyan) have a peak around 110 g mol^{-1}, suggesting a range of smaller molecule sizes. The EXT-SAMPL9 dataset (purple) displays a peak at higher molar masses, around 300 g mol^{-1}, indicating a tendency towards larger molecules.

Figure 1c shows the normalized distribution of different atom types across the LF-QC, HF-Exp, EXT-Zamora, and EXT-SAMPL9 datasets. This distribution is defined as the frequency of each atom type appearing in the datasets, adjusted so that the total frequency adds up to one. The LF-QC dataset (red) exhibits a broad distribution with significant representation across various atom types, highlighting its diverse chemical composition. The HF-Exp dataset (green) shows a more constrained distribution, indicating a focus on a narrower range of chemical species. The EXT-Zamora (cyan) and EXT-SAMPL9 (purple) datasets display even more distinct distributions, with the EXT-SAMPL9 dataset showing significant representation of specific atom types. This comparison highlights the diverse chemical compositions and focuses of the datasets, with LF-QC covering a wide array of atom types, while the experimental datasets (HF-Exp, EXT-Zamora, EXT-SAMPL9) are more specialized.

Methodology

Next, we present the different computational methods, both semi-empirical and data-driven that we explore for predicting toluene/water partition coefficients. We choose COSMO-RS, a physics-based model, to generate low fidelity data because it performs better than the other available methods like GC and MD. Based on this low fidelity data, we develop several multi-fidelity ML approaches to address the issue of limited high fidelity experimental data.

COSMO-RS

COSMO-RS is a computational model utilized for predicting thermodynamic properties and solvation behavior of molecules in solution. It combines quantum chemistry and statistical thermodynamics to estimate the chemical potentials of components in a system [14, 15]. Molecules are represented by surface segments, with segment interactions approximated as independent entities. The model relies on the σ-profile calculated from quantum chemical calculations, to predict the properties of interest. For a detailed description of COSMO-RS, we refer the interested reader to Refs. [55–59].

The logarithm of the toluene/water partition coefficient $\log P$ can be calculated according to

$$\log P_{\mathrm{tol/w}} = \log\left(\frac{[\mathrm{S}]_{\mathrm{tol}}}{[\mathrm{S}]_{\mathrm{w}}}\right), \tag{1}$$

where $[\mathrm{S}]_{\mathrm{tol}}$ and $[\mathrm{S}]_{\mathrm{wat}}$ are the concentrations of a solute [S] in toluene and water, respectively. In the COSMO-RS framework, the toluene/water $\log P$ is calculated according to [59, 60]

$$\log P_{\mathrm{tol/w}} = \frac{\Delta \mathrm{G}_{\mathrm{Transfer}}}{RT\ln 10} = \frac{\Delta \mathrm{G}_{\mathrm{w}}^{\mathrm{solv}} - \Delta \mathrm{G}_{\mathrm{tol}}^{\mathrm{solv}}}{RT\ln 10}, \tag{2}$$

where $\Delta \mathrm{G}_{\mathrm{Transfer}}$ is the transfer free energy of a solute from the pure aqueous phase to toluene. R is the gas constant and T is the temperature. $\Delta \mathrm{G}_{\mathrm{w}}^{\mathrm{solv}}$ and $\Delta \mathrm{G}_{\mathrm{tol}}^{\mathrm{solv}}$ are the solvation free energies of a solute in water and toluene, respectively. For all calculations, the temperature of 25 °C and the reference state of 1 mol L^{-1} in the liquid and the gas is used.

Alternatively, the partition coefficient at infinite dilution can be calculated from infinite dilution activity coefficients γ^{∞} and liquid molar volumes ν of toluene and water:

$$\log P_{\mathrm{tol/w}}^{\infty} = \log \frac{\gamma_{\mathrm{w}}^{\infty,s}}{\gamma_{\mathrm{tol}}^{\infty,s}} \frac{\nu_{\mathrm{w}}}{\nu_{\mathrm{tol}}} \tag{3}$$

We also evaluated openCOSMO-RS [22, 23] as an open-source alternative but found it produced significant errors (see Table 3 and supporting information for more detail). This performance gap likely stems from openCOSMO-RS currently being limited to single conformers, which is particularly problematic for polar molecules that require multiple conformer consideration for accurate predictions. Development of conformer ensemble capabilities for openCOSMO-RS is currently underway to address this single-conformer limitation.

Graph neural networks

GNN models learn properties directly from the molecular structure and have shown high prediction accuracies for a variety of both pure component [30, 61, 62] and mixture properties [32, 34, 63, 64]. Each molecule is represented as a graph with atoms as nodes and bonds as edges with corresponding feature vectors that contain atom and bond information, respectively. GNN models learn to extract local structural information about the molecular graph in graph convolutions that are then encoded into a vector representation. This molecular vector is then mapped to the property of interest by using a feedforward neural network. For a detailed description of GNN models, we refer the interested reader to overviews in Refs. [29, 36, 65, 66].

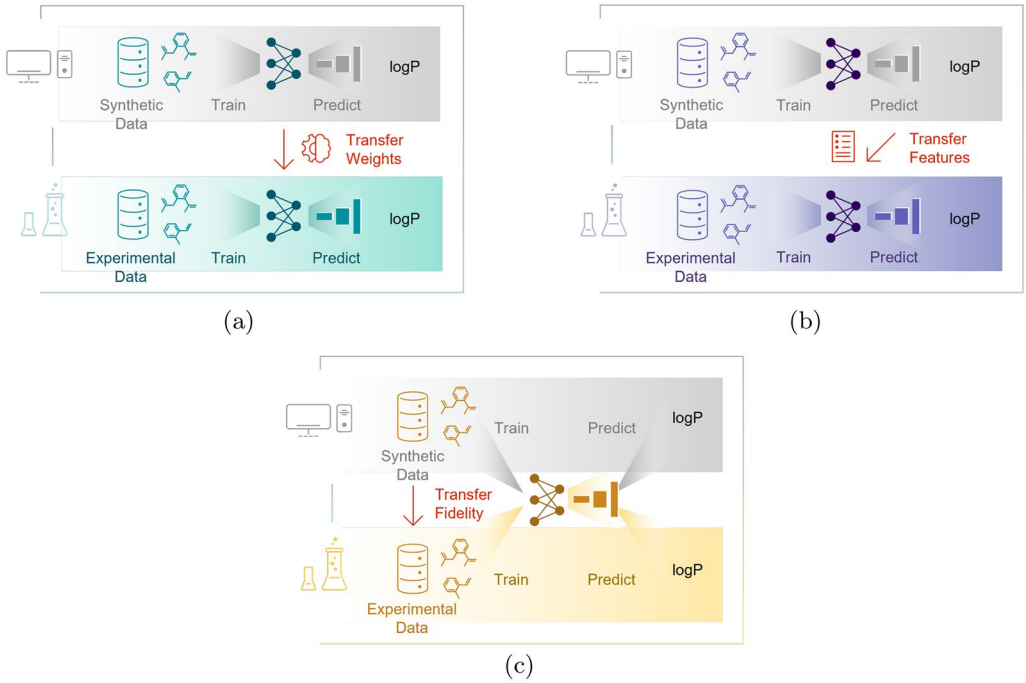

Fig. 2 Overview of different multi-fidelity approaches for training the graph neural network models. Panels **a**–**c** depict the transfer learning, the feature-augmented learning, and the multi-target learning approach, respectively

We use the Directed-Message Passing Neural Network (D-MPNN) model implemented in python library chemprop v1.7, which has achieved high accuracies in a variety of molecular property prediction tasks [67]. All datasets were split into 80-10-10 training-validation-test proportions using chemprop's default random splitting. We use the default molecular features [67] and we tune the model hyperparameters of the chemprop library using 100 iterations of Bayesian optimization for hyperparameter search (see supporting information for more detail), with the test set remaining completely unseen during model development. The best set of parameters is chosen based on the validation error to train the final model, which is provided in the supporting information. We then explore different training approaches.

We utilize three multi-fidelity approaches [42] to enhance the prediction of molecular properties: *transfer learning*, *feature-augmented learning*, and *multi-target learning* (see Fig. 2). *Transfer learning* (cf. Refs. [68, 69]) leverages pretrained models on LF-QC dataset to fine-tune predictions on the HF-Exp dataset. The idea is to use the low fidelity QC data (LF-QC) to develop a broadly applicable model and then employ the high fidelity experimental data (HF-Exp) to increase model's accuracy, thus enhancing the model's predictive capability with limited high fidelity data. *Feature-augmented learning* (cf. Ref. [41]) combines the HF-Exp dataset and LF-QC dataset: first a model is trained on the LF-QC dataset and then the predictions are used as an additional feature to existing ones for training a new model on the HF-Exp dataset. The purpose of *feature-augmented learning* is to integrate data of varying fidelities with high correlation to improve the predictive accuracy. *Multi-target learning* or multi-task learning (cf. Refs. [70, 71]) simultaneously predicts both experimental (HF-Exp dataset) and synthetic (LF-QC dataset) properties using a single model, aiming to exploit the interdependencies between different properties. This approach therefore aims to utilize information from multiple related tasks (predicted and experimental data) to improve the overall learning process and model robustness.

Results & discussion

We now present a comparison of the D-MPNN prediction performance, focusing on the different multi-fidelity learning approaches, to conclude if one is more suitable than the others. We then compare these models with other existing models from the literature that can be used for toluene/water partition coefficient prediction to evaluate the multi-fidelity learning approaches overall.

Table 2 D-MPNN and Random Forrest (baseline) models performance comparison for EXT-Zamora [38] and EXT-SAMPL9 [24] datasets

Model	Mode	Dataset	Split	EXT-Zamora [38]		EXT-SAMPL9 [24]	
				RMSE	R^2	RMSE	R^2
D-MPNN (this work)	Single	HF-Exp	Random	0.63	0.86	1.32	0.65
D-MPNN (this work)	Single	LF-QC	Random	0.71	0.83	1.34	0.64
D-MPNN Transfer Learning (this work)	sequential	LF-QC + HF-Exp	Random	0.51	0.91	1.14	0.74
D-MPNN Multi-target (this work)	simultaneous	LF-QC + HF-Exp	Random	0.44	0.93	1.02	0.79
D-MPNN Feature-augmented (this work)	sequential	LF-QC + HF-Exp	Random	0.81	0.78	1.16	0.73
Random Forrest (baseline)	Single	LF-QC + HF-Exp	Random	0.88	0.74	2.51	−0.26

Comparison of multi-fidelity learning approaches

Table 2 first shows the performance of the D-MPNN models on the EXT-Zamora and EXT-SAMPL9 datasets. As described in Section "Dataset", the EXT-Zamora contains molecules that are similar to the training sets (LF-QC and HF-Exp) in terms of molecular weight and log *P* range, thereby providing insight into the predictive capability within a similar molecular space. In contrast, the EXT-SAMPL9 dataset consists of relatively larger molecules, allowing us to evaluate the models' generalization capabilities. We report the performance of various D-MPNN models, including single-task, *transfer learning*, *multi-target learning*, and *feature-augmented learning*.

The single-task D-MPNN model is trained on HF-Exp only and thus serves as a baseline to evaluate whether the inclusion of LF-QC data in the different multi-fidelity approaches can improve prediction accuracy. The single-task model achieves an RMSE of 0.63 log *P* units and R^2 of 0.86 on the EXT-Zamora and an RMSE of 1.32 log *P* units and R^2 of 0.65 on the EXT-SAMPL9 dataset. The lower accuracy observed on EXT-SAMPL9 dataset is expected, as this dataset tests the generalization to larger molecules. For completeness, we train also a single-task D-MPNN model on the LF-QC and the models shows comparable performance, with slight differences in RMSE and R^2 values (Table 2). For comparison with a traditional Quantitative StructureActivity Relationships (QSAR) method, we also trained a Random-Forest regressor on ECFP4 fingerprints generated from the combined LF-QC+HF-Exp set, retaining the experimental value whenever a molecule appeared in both sources. This fingerprint model attains an RMSE/R^2 of 0.88 log *P* units/0.74 on EXT-Zamora and 2.51 log *P* units/−0.26 on EXT-SAMPL9, markedly worse than any D-MPNN variant and thus underscoring the benefit of the graph-based approach.

Now considering the multi-fidelity approaches, we find that *transfer learning*, where the model is sequentially trained on the LF-QC dataset and HF-Exp dataset, shows an improvement over single-task training with an RMSE of 0.51 log *P* units and R^2 of 0.91 on the EXT-Zamora and an RMSE of 1.14 log *P* units and R^2 of 0.74 on the EXT-SAMPL9 dataset. The *multi-target learning* approach, which simultaneously trains on both LF-QC and HF-Exp datasets, performs even better, achieving an RMSE of 0.44 log *P* units and R^2 of 0.93 on the EXT-Zamora and an RMSE of 1.02 log *P* units and R^2 of 0.79 on the EXT-SAMPL9 dataset. The *feature-augmented learning* approach, which sequentially trains on LF-QC and HF-Exp datasets, does not perform as well as the *multi-target learning* approach, with an RMSE of 0.81 log *P* units and R^2 of 0.78 on the EXT-Zamora and an RMSE of 1.16 log *P* units and R^2 of 0.73 on the EXT-SAMPL9 dataset. It thus does not improve the predictive quality compared to the single-task model on the EXT-Zamora, but only on the EXT-SAMPL9 dataset. For the overall predictive quality in terms of RMSE and R^2, *multi-target learning* thus yields the highest improvement over single-task learning and is therefore most effective, see Table 2.

Impact of molar mass

Figures 3 and 4 further show the parity plots, i.e., predicted against the experimental data, of EXT-Zamora and EXT-SAMPL9 datasets for the different multi-fidelity approaches. The dashed lines indicate an error of ± 1 log *P* units. To analyze the impact of the molar mass on the performance of the models, we also indicate different weight ranges with colors.

For the EXT-Zamora dataset, the *multi-target learning* approach consistently shows the best performance across all molar masses. Only one molecule of 400 g mol^{-1} to 500 g mol^{-1} is out of the range of ± 1 log *P* units (see Fig. 3b). The *transfer learning* approach also performs well, though slightly less effectively for larger molecules > 300 g mol^{-1}. The *feature-augmented learning* approach, however, shows higher variability, particularly for the middle-weight range (100 g mol^{-1} to 200 g mol^{-1} and 200 g mol^{-1} to 300 g mol^{-1}).

Similarly, for the EXT-SAMPL9 dataset, the *multi-target learning* approach maintains the best performance

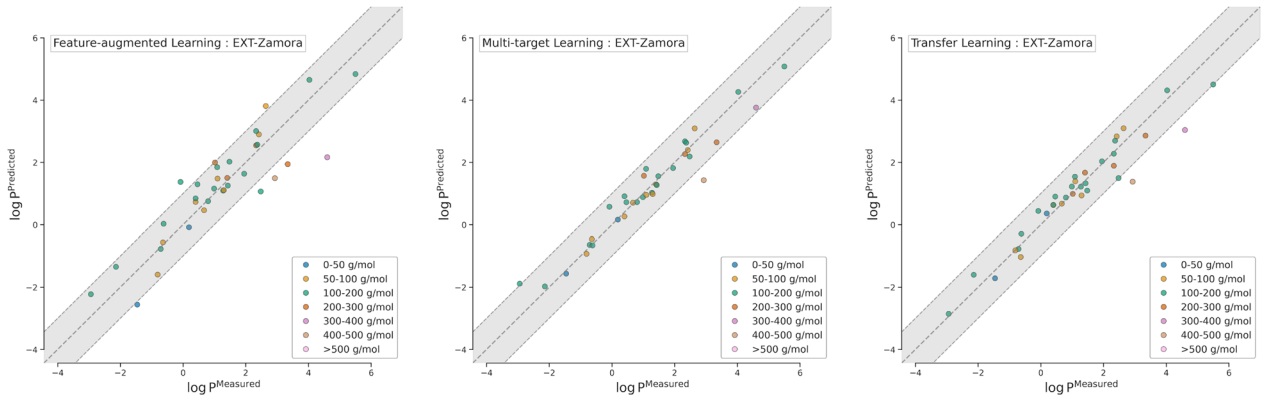

Fig. 3 Comparison of the multi-fidelity learning approaches on EXT-Zamora dataset colored by molar mass range for **a** *feature-augmented learning*, **b** *multi-target learning*, and **c** *transfer learning*. Dashed lines indicate an error margin of ± 1 log *P* units

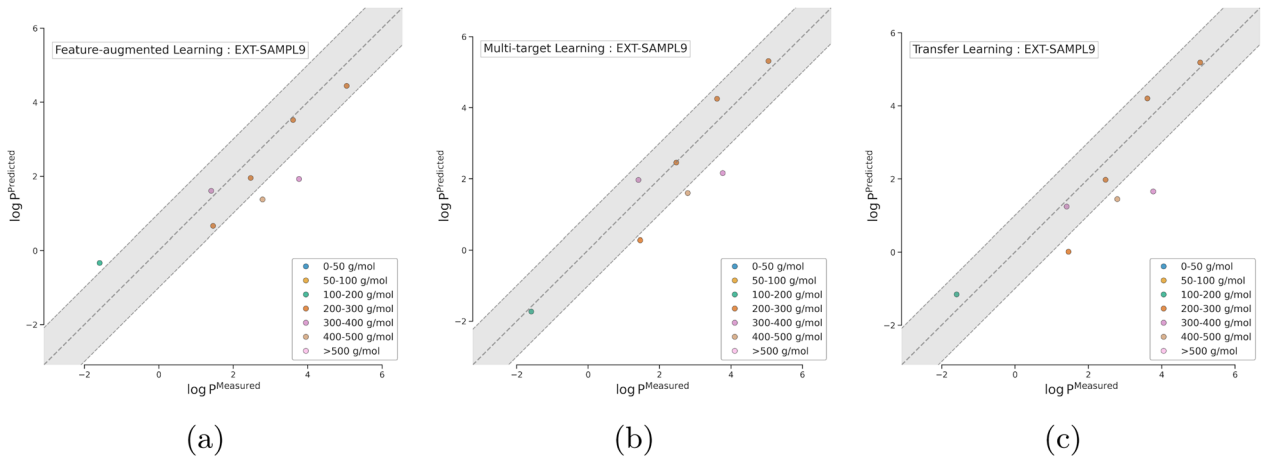

Fig. 4 Comparison of the multi-fidelity learning approaches on EXT-SAMPL9 dataset colored by molar mass range for **a** *feature-augmented learning*, **b** *multi-target learning*, and **c** *transfer learning*. Dashed lines indicate an error margin of ± 1 log *P* units

across most weight categories (see Fig. 4). It shows particularly strong results for light molecules and less strong results for heavier molecules. *Transfer learning* remains competitive but again shows slight performance degradation for heavier molecules. The *feature-augmented learning* approach continues to exhibit higher variability, especially for molecules in the 200 g mol^{-1} to 300 g mol^{-1} and > 500 g mol^{-1}.

Overall, the *multi-target learning* approach shows the highest predictive robustness across different molar masses.

Impact of chemical classes

We also investigate the model performance across different chemical classes, and illustrate the results in Fig. 5. To analyze the impact of chemical classes on model performance, we categorize molecules based on their chemical structures using SMARTS patterns and substructure matching. It is important to note that the overall number of molecules per class is very low (sometimes as few as one), indicating that additional data and further evaluations will be needed to confirm these findings. In Fig. 5, the boxes represent the interquartile range with lines indicating the median values and the whiskers extend to 1.5 times the interquartile range. The EXT-Zamora dataset features a diverse set of chemical classes, including 11 phenols, 5 ketones, 3 quinoline, 3 ethers, 3 alcohols, 2 benzoic acids, 2 alkyl halides, and one each of aminophenol, aniline, benzene derivative, and cycloalkane (5 molecules classified as other). The EXT-SAMPL9 dataset, in

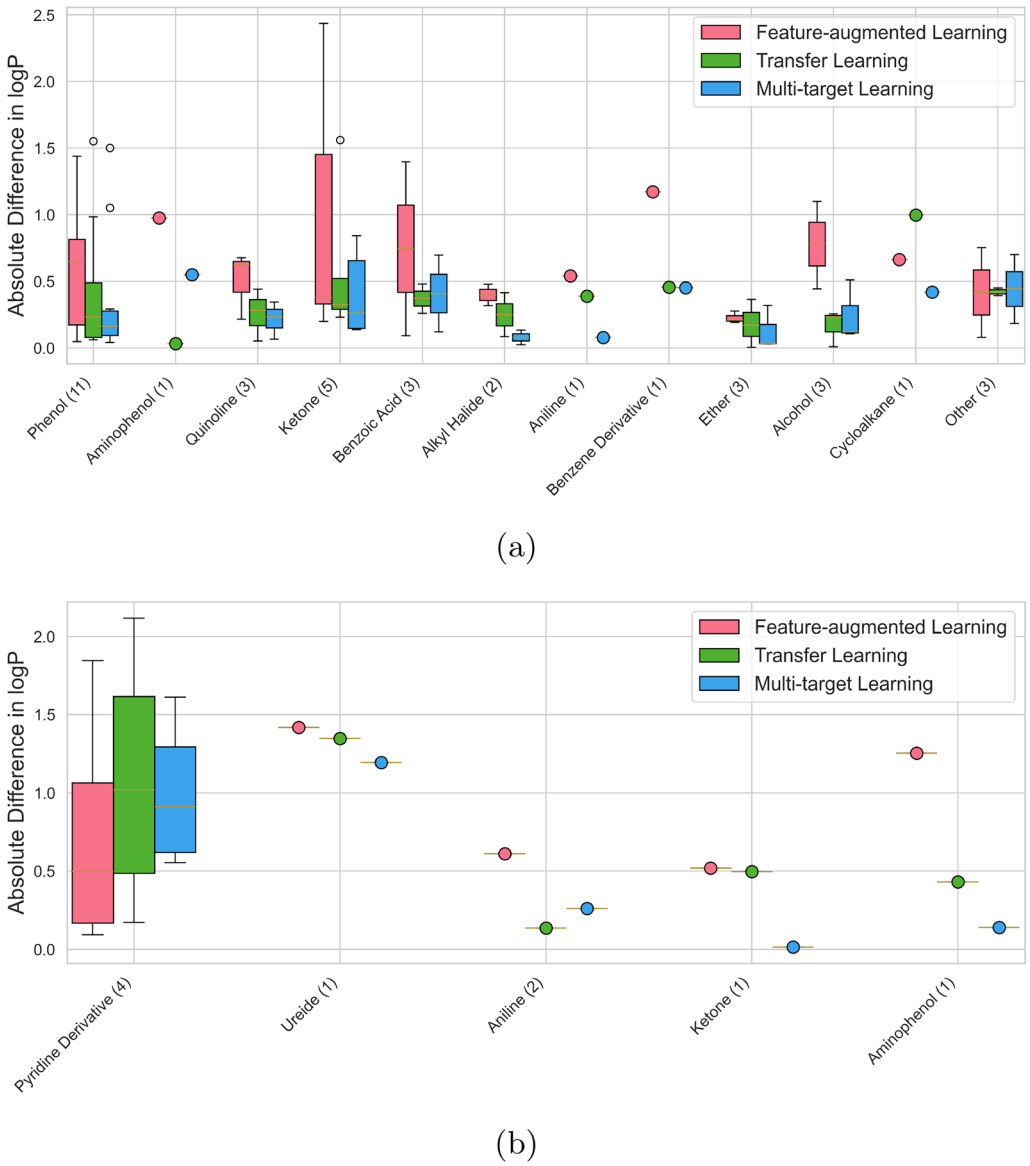

Fig. 5 Predictive performance of the different multi-fidelity learning approaches across various chemical classes in the **a** EXT-Zamora and **b** EXT-SAMPL9 datasets. Numbers in parentheses indicate the number of molecules in each chemical class

contrast, is less diverse compared to EXT-Zamora dataset. It is a smaller dataset comprising a limited range of chemical classes, containing 4 pyridine derivatives, 2 benzene derivatives, 2 anilines, and one each of phenol, ureide, ketone, aminophenol, and sulfonamide (3 molecules classified as other). An overview of the chemical class distributions in the LF-QC and HF-Exp datasets can be found in the supporting information.

The *multi-target learning* approach demonstrates the most consistent and lowest absolute differences in log *P* predictions across various chemical classes. For example, in the classes of alcohols, ethers, and alkyl halides, it shows significantly lower errors compared to *feature-augmented learning* and *transfer learning* approaches. Interestingly, *multi-target learning* shows a great agreement between predictions and experiments with a mean absolute lower lower than 0.5 log *P* units for the chemical classes aniline, ketone, and aminophenol for the EXT-SAMPL9 dataset and keeps the same consistency for the EXT-Zamora dataset except for the chemical class aminophenol. This indicates *multi-target learning* effectively captures the distinct characteristics of different chemical

structures by leveraging both LF-QC and HF-Exp datasets during training.

Transfer learning also performs well across various chemical classes but shows higher variability in classes such as benzene derivatives and amides. This variability suggests that while *transfer learning* can improve model accuracy by integrating different data types, it may still face challenges in fully capturing the intricate properties of more complex molecules. For instance, the errors are more pronounced in the benzene derivatives class in the EXT-Zamora dataset, indicating a potential limitation in handling aromatic systems. This might be due to the fact that not enough data are available for the fine-tuning step, as Vermeire and Green [32] have shown that *transfer learning* can achieve a great agreement between predictions and experiments if enough high fidelity data are available.

The *feature-augmented learning* approach shows the highest absolute differences in several chemical classes, including ketones and benzene derivatives. This performance suggests that the method's sequential training on LF-QC and HF-Exp datasets may not be as effective in capturing the detailed chemical properties required for accurate log P predictions. The higher errors in the ketone class, particularly in the SAMPL9 dataset, highlight the approach's difficulty in balancing data contributions from different fidelities, especially for complex chemical structures. This indicates that *feature-augmented learning* requires careful handling to avoid poor performance in chemically diverse datasets, especially when few data is available for fine-tuning.

Comparison to other models

We further compare the best performing D-MPNN model, *multi-target learning*, to other semi-empirical and data-driven models from the literature, as shown in Table 3. Specifically, we consider two GNN models that provide infinite dilution activity coefficient (AC) predictions, namely Gibbs-Duhem-informed (GDI)-GNNs trained on COSMO-RS activity coefficient data from our previous work [72] and the Gibbs–Helmholtz (GH)-GNN [73] trained on experimental infinite dilution activity coefficient (IDAC) data from the DECHEMA Chemistry Data Series [78]. To predict the partition coefficients, we employ the already trained models from Refs. [72, 73], using Eq. 3. We calculate the molar volumes with densities and molecular weights for toluene and water from the National Institute of Standards and Technology (NIST) Chemistry webbook [79]. We further include two GNN models based on the D-MPNN architecture trained on diverse datasets of COSMO-RS and experimental solvation Gibbs free energies, namely Solvation GNN [32] and DirectML [74]. Here, the partition coefficients are calculated using the already trained models from Refs. [32, 74] along with Eq. 2. All GNN models use an ensemble approach, i.e., the prediction of multiple models trained on different data splits are averaged to obtain a final prediction. In addition, we consider the MLR and RFR from Zamora et al. [38] that were fitted on the HF-Exp set. The partition coefficient values are taken directly from the original publication [38]. These two regression models use 11 input descriptors, including AlogP (octanol/water partition coefficient using Ghose–Crippen atomic contributions [80]), which shows a 58%

Table 3 Model performance comparison for EXT-Zamora [38] and EXT-SAMPL9 [24] datasets

Model	Mode	Dataset	Split	EXT-Zamora [38]		EXT-SAMPL9 [24]	
				RMSE	R^2	RMSE	R^2
D-MPNN Multi-target (this work)	Simultaneous	LF-QC + HF-Exp	Random	0.44	0.93	1.02	0.79
GDI-GNN[a] by Rittig et al. [72]	Ensemble	COSMO-AC	–	**0.77**	**0.80**	1.56	0.51
GH-GNN[a] by Sanchez Medina et al. [73]	Ensemble	DECHEMA IDAC	–	*1.23*	*0.48*	*1.69*	*0.43*
Solvation GNN[a] by Vermeire and Green [32]	Ensemble	COSMO & exp. G	–	**0.27**	**0.97**	**1.07**	**0.77**
DirectML[a] by Chung et al. [74]	Ensemble	COSMO & exp. G	–	**0.37**	**0.95**	**1.04**	**0.78**
MLR by Zamora et al.[38]	Single	exp	–	1.05	–	**0.86**	**0.85**
RFR by Zamora et al. [38]	Single	exp	–	1.13	–	**0.84**	**0.86**
COSMO-RS[a] results from Nevolianis et al. [25]	–	COSMO	–	0.60	0.88	1.23	0.70
openCOSMO-RS[a] by Müller et al. [23]	–	–	–	1.74	−0.73	2.37	−0.35
MM/PBSA[a] by Amezcua et al. [24]	–	–	–	–	–	1.12	0.75
HANNA[a] (Clapeyron.jl) by Walker et al. [75]	–	–	–	1.71	0.01	2.01	0.19
UNIFAC[a] (thermo) by Bell et al. [76, 77]	–	–	–	3.34	−2.80	3.89	−2.01

[a] Models are not trained on partition coefficients

Some molecules of the test set are included in the training set (in bold)

Some molecules of the test set might be included in the training set (the training set is not publicly available)

correlation to the toluene-water partition coefficient, cf. [38]. Lastly, we compare to two semi-empirical models: COSMO-RS and MM/PBSA [24]. In the COSMO-RS approach [25], the geometry of each molecule is optimized at GFN2-xTB [46] level and further in the COSMO state using COSMOconf [81]. Next, the solvation free energies of the molecules are calculated in water and toluene at infinite dilution using COSMOtherm [82]. In the MM/PBSA approach, each molecule is optimized using QM, followed by molecular dynamics geometry optimization, and solvation free energies in water and toluene are calculated. In this case, the partition coefficient values are obtained directly from the original publication [24].

The GDI-GNN model shows strong performance on EXT-Zamora dataset; however, its prediction accuracy is likely overestimated due to 16 of the 38 test set molecules being included in the training. In contrast, its performance on the EXT-SAMPL9 set, which has no overlap with the training data, is lower. The GH-GNN model generally shows lower performance, and since its training data is not publicly available, we could not identify potential overlaps of training and test data. Interestingly, activity coefficient GNN models are performing at level comparable to the top five models from the SAMPL9 challenge [24]. Yet, the activity coefficient GNN models show lower accuracy than the D-MPNNs directly trained on partition coefficients.

The Solvation GNN and DirectML models show high predictive quality; however, their accuracy is likely overestimated due to significant overlap between training and test molecules. For example, the experimental training data of Solvation GNN and DirectML contain, respectively, 29 (34 for pretraining) and 35 of the 38 molecules of EXT-Zamora, and, respectively, 4 (7 for pretraining) and 14 of the 16 molecules of EXT-SAMPL9. In fact, we observe a similar accuracy of the Solvation GNN and DirectML on EXT-SAMPL9 compared to the multi-target D-MPNN, although some molecules are already included in training, thus indicating at most comparable generalization capabilities.

The MLR and RFR models from Zamora et al. [38] show varying performance. Both models achieve higher accuracy on the EXT-SAMPL9 dataset compared to the EXT-Zamora dataset. The high predictive accuracy on the EXT-SAMPL9 indicates the effectiveness of using molecular descriptors when available training data is limited, which has also been reported in recent comparisons of ML/GNN models with and without using QC descriptors [83]. However, these models are typically limited in their generalizability to molecules dissimilar from the training data. The higher accuracy on the presumably more distinct EXT-SAMPL9 set compared to EXT-Zamora (cf. Section "Dataset") is thus unexpected. In fact, we find that the experimental data used for fitting contains a duplicate entry with EXT-SAMPL9, indexed as entries 79 (Aflukin) and 266 (Quinine) [38]. This duplication might explain the better performance observed on the EXT-SAMPL9 dataset compared to the EXT-Zamora dataset. However, after retraining the models without the duplicate entry, the RMSEs for EXT-Zamora are 1.12 (MLR) and 1.04 (RFR), while for EXT-SAMPL9, they are 0.94 (MLR) and 0.90 (RFR), indicating that the duplication had only a minor impact on the results. We thus find lower accuracy of the MLR and RFR compared to the ML models for EXT-Zamora and slightly reduced accuracy for EXT-SAMPL9.

The COSMO-RS and MM/PBSA models from the SAMPL9 challenge show moderate performance on the EXT-SAMPL9 dataset but perform better on the EXT-Zamora dataset. Despite their performance, they are outperformed by the D-MPNNs with multi-fidelity learning. It is important to note that the SAMPL9 challenge reports different r^2 values, which are not coefficients of determination R^2; therefore, the R^2 values here have been recalculated for consistency. Additionally, we evaluated openCOSMO-RS as an open-source alternative to COSMO-RS, which showed limited accuracy on both datasets, primarily due to its current single-conformer limitation as discussed in the methods section (see supporting information). Last, the thermodynamic models HANNA and UNIFAC show limited performance on both datasets, with particularly poor R^2 values, suggesting that these methods may struggle with the molecular diversity present in these datasets. The absence of conformational flexibility in these approaches could be a significant limitation for larger, more flexible molecules where conformational effects play a crucial role in partition behavior.

Conclusion

In this work, we investigated multi-fidelity learning approaches with GNN models for predicting toluene/water partition coefficients for which experimental data are only readily available in the order of a few hundred values. First, we used COSMO-RS to create a low fidelity dataset of partition coefficients for about 9000 molecules. The low fidelity data in combination with the available high fidelity experimental data was then utilized for training GNN models. Our results showed that *multi-target learning*, i.e., predicting low fidelity and high fidelity target properties with one GNN model, yields substantial accuracy increases to training a GNN model on the experimental data only and is superior to *transfer learning* and *feature-augmented learning*. We further found competitive accuracy of

the multi-target GNN model compared to other predictive models, e.g., based on activity coefficients and solvation free energies, and other methods such as COSMO-RS. Overall, the comparison of the different approaches for partition coefficient predictions shows that direct training on log *P* data is most effective. Here, multi-fidelity learning in the form of *multi-target learning* substantially increases the predictive accuracy. This is particularly interesting as the *multi-target learning* approach presumably requires the least training and model changes, i.e., just an additional model output, and is thus straightforward to implement. Generating additional molecular property data through QC calculations for training predictive ML models like GNN models is thus highly promising to enhance the predictive quality when available experimental data is limited, such as for toluene/water partition coefficients. However, it is important to acknowledge that the availability of high fidelity data remains a significant challenge and the extrapolation to new chemical classes cannot be fully resolved with multi-fidelity learning approaches leveraging large low-fidelity datasets.

Future work could consider *multi-target learning* with low and high fidelity datasets for multiple molecular properties, e.g., combining activity coefficients, solvation free energies, and partition coefficients. For this, also thermodynamics relationships between the properties could be integrated into the model training and architecture, as, e.g., in [84, 85], aiming at more general predictive models.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-025-01057-6.

Supplementary Material 1.

Acknowledgements
The authors thank Prof. William H. Green, Jonathan W. Zheng, Dr. Kevin Greenman, Dr. Florence Vermeire, and Dr. Simon Müller for stimulating discussion on this work. We thank William Zamora for removing the duplicate entry, retraining the models, and sharing the revised predictions after reading an earlier version of the preprint on ChemRxiv.

Author contributions
T.N. implemented the multi-fidelity graph neural networks, set up and conducted the computational experiments including the formal analysis and visualization, and wrote the original draft of the manuscript. J.G.R. conducted the formal analysis and wrote the original draft of the manuscript. A.M. acquired funding, provided supervision, and edited the manuscript. K.L. acquired funding, provided supervision, and edited the manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. T.N., A.M., and K.L. gratefully acknowledge financial support from the SFB 985, Functional Microgels and Microgel Systems of Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within project B4 (191948804). J.G.R. and A.M. gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—466417970—within the Priority Programme "SPP 2331: Machine Learning in Chemical Engineering. Simulations were performed with computing resources granted by RWTH Aachen University under project rwth1603.

Data availability
The datasets supporting the conclusions of this article are available in the Zenodo repository under DOI: 10.5281/zenodo.13236218. The SMILES for all molecules used in this study are provided in the supporting information as a CSV file, except for the LF-QC dataset, which is not publicly available due to licensing restrictions. The trained models are also not publicly available for the same reason. However, a Python notebook containing all the scripts and code to reproduce the results of this work is provided.

Declarations

Competing interests
The authors declare no competing interests.

Author details
[1]Institute of Technical Thermodynamics, RWTH Aachen University, 52062 Aachen, Germany. [2]Process Systems Engineering, RWTH Aachen University, 52074 Aachen, Germany. [3]Institute of Climate and Energy Systems ICE-1: Energy Systems Engineering, Forschungszentrum Jülich GmbH, 52425 Jülich, Germany. [4]JARA-SOFT, 52425 Jülich, Germany.

Received: 13 August 2024 Accepted: 12 July 2025
Published online: 08 August 2025

References

1. Testa B, Crivori P, Reist M, Carrupt P-A (2000) The influence of lipophilicity on the pharmacokinetic behavior of drugs: concepts and examples. Perspect Drug Discov Des 19(1):179–211. https://doi.org/10.1023/a:1008741731244
2. Klopman G, Zhu H (2005) Recent methodologies for the estimation of n-octanol/water partition coefficients and their use in the prediction of membrane transport properties of drugs. Mini-Rev Med Chem 5(2):127–133. https://doi.org/10.2174/1389557053402765
3. Mannhold R, Poda GI, Ostermann C, Tetko IV (2009) Calculation of molecular lipophilicity: state-of-the-art and comparison of logp methods on more than 96,000 compounds. J Pharm Sci 98(3):861–893. https://doi.org/10.1002/jps.21494
4. Andrés A, Rosés M, Ràfols C, Bosch E, Espinosa S, Segarra V, Huerta JM (2015) Setup and validation of shake-flask procedures for the determination of partition coefficients (logd) from low drug amounts. Eur J Pharm Sci 76:181–191. https://doi.org/10.1016/j.ejps.2015.05.008
5. Hostrup M, Harper PM, Gani R (1999) Design of environmentally benign processes: integration of solvent design and separation process synthesis. Comput Chem Eng 23:1395–1414. https://doi.org/10.1016/S0098-1354(99)00300-2
6. Paes FC, Privat R, Jaubert J-N, Sirjean B (2022) A comparative study of cosmo-based and equation-of-state approaches for the prediction of solvation energies based on the compsol databank. Fluid Phase Equilib 561:113540. https://doi.org/10.1016/j.fluid.2022.113540
7. Arnott JA, Planey SL (2012) The influence of lipophilicity in drug discovery and design. Expert Opin Drug Discov 7(10):863–875. https://doi.org/10.1517/17460441.2012.714363
8. Johnson TW, Gallego RA, Edwards MP (2018) Lipophilic efficiency as an important metric in drug design. J Med Chem 61(15):6401–6420. https://doi.org/10.1021/acs.jmedchem.8b00077
9. Dunn WJ, Block JH, Pearlman RS (1986) Partition coefficient: determination and estimation. Pergamon Press, New York . Published in cooperation with the American Pharmaceutical Association, Academy of Pharmaceutical Sciences
10. Otsuka H (2005) Purification by solvent extraction using partition coefficient. Humana Press, Totowa, pp 269–273. https://doi.org/10.1385/1-59259-955-9:269
11. Polte L, Raßpe-Lange L, Latz F, Jupke A, Leonhard K (2022) Cosmo-camped-solvent design for an extraction distillation considering

molecular, process, equipment, and economic optimization. Chem Ing Tec 95(3):416–426. https://doi.org/10.1002/cite.202200144
12. Caron G, Ermondi G (2005) Iculating virtual log P in the alkane/water system (log P Nalk) and its derived parameters Δlog P Noct-alk and log D pHalk. J Med Chem 48(9):3269–3279. https://doi.org/10.1021/jm048980b
13. David L, Wenlock M, Barton P, Ritzén A (2021) Prediction of chameleonic efficiency. ChemMedChem 16(17):2669–2685. https://doi.org/10.1002/cmdc.202100306
14. Klamt A (1995) Conductor-like screening model for real solvents: a new approach to the quantitative calculation of solvation phenomena. J Phys Chem 99(7):2224–2235. https://doi.org/10.1021/j100007a062
15. Klamt A, Jonas V, Bürger T, Lohrenz JC (1998) Refinement and parametrization of COSMO-RS. J Phys Chem A 102(26):5074–5085. https://doi.org/10.1021/jp980017s
16. Platts JA, Abraham MH, Butina D, Hersey A (1999) Estimation of molecular linear free energy relationship descriptors by a group contribution approach. 2. Prediction of partition coefficients. J Chem Inf Comput Sci 40(1):71–80. https://doi.org/10.1021/ci990427t
17. Lin S-T, Sandler SI (2000) Multipole corrections to account for structure and proximity effects in group contribution methods: octanol-water partition coefficients. J Phys Chem A 104:7099–7105. https://doi.org/10.1021/jp000091m
18. Buggert M, Cadena C, Mokrushina L, Smirnova I, Maginn EJ, Arlt W (2009) COSMO-RS calculations of partition coefficients: different tools for conformation search. Chem Eng Technol 32:977–986. https://onlinelibrary.wiley.com/doi/abs/10.1002/ceat.200800654
19. Loschen C, Klamt A (2014) Prediction of solubilities and partition coefficients in polymers using COSMO-RS. Ind Eng Chem Res 53:11478–11487. https://doi.org/10.1021/ie501669z
20. Ince A, Carstensen H-H, Reyniers M-F, Marin GB (2015) First-principles based group additivity values for thermochemical properties of substituted aromatic compounds. AIChE J 61:3858–3870. https://aiche.onlinelibrary.wiley.com/doi/abs/10.1002/aic.15008
21. Bannan CC, Calabró G, Kyu DY, Mobley DL (2016) Calculating partition coefficients of small molecules in octanol/water and cyclohexane/water. J Chem Theory Comput 12(8):4015–4024. https://doi.org/10.1021/acs.jctc.6b00449
22. Müller S, Nevolianis T, Garcia-Ratés M, Riplinger C, Leonhard K, Smirnova I (2024) Predicting solvation free energies for neutral molecules in any solvent with openCOSMO-RS . https://arxiv.org/abs/2407.03434
23. Müller S, Nevolianis T, Garcia-Ratés M, Riplinger C, Leonhard K, Smirnova I (2025) Predicting solvation free energies for neutral molecules in any solvent with openCOSMO-RS. Fluid Phase Equilib 589:114250. https://doi.org/10.1016/j.fluid.2024.114250
24. Amezcua M, Mobley DL, Bergazin TD (2023) samplchallenges/SAMPL9: 0.8. Zenodo . https://doi.org/10.5281/ZENODO.7644720
25. Nevolianis T, Ahmed RA, Hellweg A, Diedenhofen M, Leonhard K (2023) Blind prediction of toluene/water partition coefficients using COSMO-RS: results from the sampl9 challenge. Phys Chem Chem Phys 25:31683–31691. https://doi.org/10.1039/D3CP04077A
26. Xue Z, Mu T, Gmehling J (2012) Comparison of the a priori COSMO-RS models and group contribution methods: Original unifac, modified unifac(do), and modified unifac(do) consortium. Ind Eng Chem Res 51(36):11809–11817. https://doi.org/10.1021/ie301611w
27. Fingerhut R, Chen W-L, Schedemann A, Cordes W, Rarey J, Hsieh C-M, Vrabec J, Lin S-T (2017) Comprehensive assessment of cosmo-sac models for predictions of fluid-phase equilibria. Ind Eng Chem Res 56(35):9868–9884. https://doi.org/10.1021/acs.iecr.7b01360
28. Klamt A (2018) The cosmo and COSMO-RS solvation models. Wiley Interdiscip Rev Comput Mol Sci 8(1):1338. https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1338
29. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE (2017) Neural message passing for quantum chemistry. In: Precup D, Teh YW (eds.) Proceedings of the 34th international conference on machine learning. proceedings of machine learning research, vol. 70, pp 1263–1272. PMLR. https://proceedings.mlr.press/v70/gilmer17a.html
30. Schweidtmann AM, Rittig JG, König A, Grohe M, Mitsos A, Dahmen M (2020) Graph neural networks for prediction of fuel ignition quality. Energy Fuels 34(9):11395–11407. https://doi.org/10.1021/acs.energyfuels.0c01533
31. Rong Y, Bian Y, Xu T, Xie W, Wei Y, Huang W, Huang J (2020) Self-supervised graph transformer on large-scale molecular data. In: Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H (eds.) Advances in Neural Information Processing Systems, vol. 33. Curran Associates, Inc., pp 12559–12571. https://proceedings.neurips.cc/paper_files/paper/2020/file/94aef38441efa3380a3bed3faf1f9d5d-Paper.pdf
32. Vermeire FH (2021) Transfer learning for solvation free energies: from quantum chemistry to experiments. Chem Eng J 418:129307. https://doi.org/10.1016/j.cej.2021.129307
33. Winter B, Winter C, Schilling J, Bardow A (2022) A smile is all you need: predicting limiting activity coefficients from smiles with natural language processing. Digital Discov 1(6):859–869. https://doi.org/10.1039/d2dd00058j
34. Sanchez Medina EI, Linke S, Stoll M, Sundmacher K (2022) Graph neural networks for the prediction of infinite dilution activity coefficients. Digital Discov 1(3):216–225. https://doi.org/10.1039/d1dd00037c
35. Felton KC, Ben-Safar H, Lapkin AA (2021) Deepgamma : A deep learning model for activity coefficient prediction. In: 1st annual AAAI workshop on AI to accelerate science and engineering
36. Rittig JG, Gao Q, Dahmen M, Mitsos A, Schweidtmann AM (2023) Graph neural networks for the prediction of molecular structure–property relationships. In: Zhang D, Del Río Chanona EA (eds) Machine learning and hybrid modelling for reaction engineering. Royal Society of Chemistry, pp 159–181. https://doi.org/10.1039/BK9781837670178-00159
37. Sun G, Zhao Z, Sun S, Ma Y, Li H, Gao X (2023) Vapor-liquid phase equilibria behavior prediction of binary mixtures using machine learning. Chem Eng Sci 282:119358. https://doi.org/10.1016/j.ces.2023.119358
38. Zamora WJ, Viayna A, Pinheiro S, Curutchet C, Bisbal L, Ruiz R, Ràfols C, Luque FJ (2023) Prediction of toluene/water partition coefficients in the sampl9 blind challenge: assessment of machine learning and ief-pcm/mst continuum solvation models. Phys Chem Chem Phys. https://doi.org/10.1039/D3CP01428B
39. Greenman KP, Green WH, Gómez-Bombarelli R (2022) Multi-fidelity prediction of molecular optical peaks with deep learning. Chem Sci 13(4):1152–1162. https://doi.org/10.1039/d1sc05677h
40. Fare C, Fenner P, Benatan M, Varsi A, Pyzer-Knapp EO (2022) A multi-fidelity machine learning approach to high throughput materials screening. NPJ Comput Mater. https://doi.org/10.1038/s41524-022-00947-9
41. Buterez D, Janet JP, Kiddle SJ, Oglic D, Lió P (2024) Transfer learning with graph neural networks for improved molecular property prediction in the multi-fidelity setting. Nat Commun. https://doi.org/10.1038/s41467-024-45566-8
42. Qian E, Chaudhuri A, Kang D, Sella V (2024) Multifidelity linear regression for scientific machine learning from scarce data. arXiv preprint arXiv:2403.08627
43. Cheng J-P, Yang J-D, Xue X-S, Ji P, Li X, Wang Z (2023) iBonD Website. http://ibond.nankai.edu.cn/
44. Ebejer J-P, Morris GM, Deane CM (2012) Freely available conformer generation methods: how good are they? J Chem Inf Model 52(5):1146–1158. https://doi.org/10.1021/ci2004658
45. Landrum G, Tosco P, Kelley B, Sriniker A, Dalke A, Vianello R, Cole B, Codrea V, Bain D, Halvorsen T, Wójcikowski M, Pahl A, Shadnia H, Jones M, Turk S, Vaucher A, Schwaller P, Johnson D, Fuller P, Saconne M (2020) rdkit/rdkit: 2020/03/1 (Q1 2020) release. Zenodo. https://doi.org/10.5281/zenodo.3732262
46. Bannwarth C, Ehlert S, Grimme S (2019) Gfn2-xtb-an accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. J Chem Theory Comput 15(3):1652–1671. https://doi.org/10.1021/acs.jctc.8b01176. (**PMID: 30741547**)
47. Dassault Systèmes: BIOVIA COSMOconf 2023 (2023). https://www.3ds.com
48. Becke AD (1988) Density-functional exchange-energy approximation with correct asymptotic behavior. Phys Rev A 38:3098–3100. https://doi.org/10.1103/PhysRevA.38.3098
49. Perdew JP (1986) Density-functional approximation for the correlation energy of the inhomogeneous electron gas. Phys Rev B 33:8822–8824. https://doi.org/10.1103/PhysRevB.33.8822
50. Rappoport D, Furche F (2010) Property-optimized gaussian basis sets for molecular response calculations. J Chem Phys 133(13):134105. https://doi.org/10.1063/1.3484283

51. Dassault Systèmes: BIOVIA COSMOtherm 2023 (2023). https://www.3ds.com
52. Ruiz R, Zamora WJ, Ràfols C, Bosch E (2022) Molecular characteristics of several drugs evaluated from solvent/water partition measurements: solvation parameters and intramolecular hydrogen bond indicator. Eur J Pharm Sci 168:106066. https://doi.org/10.1016/j.ejps.2021.106066
53. Letcher TM (2007) Development and applications in solubility. Royal Society of Chemistry. https://doi.org/10.1039/9781847557681
54. Işık M, Levorse D, Mobley DL, Rhodes T, Chodera JD (2019) Octanol-water partition coefficient measurements for the sampl6 blind prediction challenge. J Comput Aided Mol Des 34(4):405–420. https://doi.org/10.1007/s10822-019-00271-3
55. Eckert F, Klamt A (2002) Fast solvent screening via quantum chemistry: COSMO-RS approach. AIChE J 48(2):369–385. https://doi.org/10.1002/aic.690480220
56. Klamt A, Eckert F (2000) COSMO-RS: a novel and efficient method for the a priori prediction of thermophysical data of liquids. Fluid Phase Equilibr 172(1):43–72. https://doi.org/10.1016/S0378-3812(00)00357-5.00530
57. Klamt A, Krooshof GJP, Taylor R (2002) COSMOSPACE: alternative to conventional activity-coefficient models. AIChE J 48(10):2332–2349. https://doi.org/10.1002/aic.690481023.00070
58. Loschen C, Reinisch J, Klamt A (2020) COSMO-RS based predictions for the sampl6 logp challenge. J Comput Aided Mol Des 34(4):385–392. https://doi.org/10.1007/s10822-019-00259-z
59. Warnau J, Wichmann K, Reinisch J (2021) COSMO-RS predictions of logp in the sampl7 blind challenge. J Comput Aided Mol Des 35(7):813–818. https://doi.org/10.1007/s10822-021-00395-5
60. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings1pii of original article: S0169-409x(96)00423-1. the article was originally published in advanced drug delivery reviews 23 (1997) 3-25.1. Advanced Drug Delivery Reviews 46(1), 3–26 https://doi.org/10.1016/S0169-409X(00)00129-0 . Special issue dedicated to Dr. Eric Tomlinson, Advanced Drug Delivery Reviews, A Selection of the Most Highly Cited Articles, 1991-1998
61. Coley CW, Barzilay R, Green WH, Jaakkola TS, Jensen KF (2017) Convolutional embedding of attributed molecular graphs for physical property prediction. J Chem Inf Model 57(8):1757–1772. https://doi.org/10.1021/acs.jcim.6b00601
62. Brozos C, Rittig JG, Bhattacharya S, Akanny E, Kohlmann C, Mitsos A (2004) Graph neural networks for surfactant multi-property prediction. Colloids Surf A 694:134133. https://doi.org/10.1016/j.colsurfa.2024.134133
63. Rittig JG, Ben Hicham K, Schweidtmann AM, Dahmen M, Mitsos A (2023) Graph neural networks for temperature-dependent activity coefficient prediction of solutes in ionic liquids. Comput Chem Eng 171:108153. https://doi.org/10.1016/j.compchemeng.2023.108153
64. Qin S, Jiang S, Li J, Balaprakash P, Lehn RCV, Zavala VM (2023) Capturing molecular interactions in graph neural networks: a case study in multicomponent phase equilibrium. Digital Discov 2(1):138–151. https://doi.org/10.1039/d2dd00045h
65. Reiser P, Neubert M, Eberhard A, Torresi L, Zhou C, Shao C, Metni H, van Hoesel C, Schopmans H, Sommer T, Friederich P (2022) Graph neural networks for materials science and chemistry. Commun Mater 3(1):93. https://doi.org/10.1038/s43246-022-00315-6
66. Schweidtmann AM, Rittig JG, Weber JM, Grohe M, Dahmen M, Leonhard K, Mitsos A (2023) Physical pooling functions in graph neural networks for molecular property prediction. Comput Chem Eng 172:108202. https://doi.org/10.1016/j.compchemeng.2023.108202
67. Heid E, Greenman KP, Chung Y, Li S-C, Graff DE, Vermeire FH, Wu H, Green WH, McGill CJ (2023) Chemprop: a machine learning package for chemical property prediction. J Chem Inf Model 64(1):9–17. https://doi.org/10.1021/acs.jcim.3c01250
68. Pan SJ, Yang Q (2009) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359. https://doi.org/10.1109/TKDE.2009.191
69. Handbook of research on machine learning applications and trends: algorithms, methods, and techniques. IGI Global (2010). https://doi.org/10.4018/978-1-60566-766-9
70. Ruder S (2017) An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 arXiv
71. Zhang Y, Yang Q (2017) A survey on multi-task learning. arXiv preprint arXiv:1707.08114 . arXiv
72. Rittig JG, Felton KC, Lapkin AA, Mitsos A (2023) Gibbs-duhem-informed neural networks for binary activity coefficient prediction. Digital Discov 2:1752–1767. https://doi.org/10.1039/D3DD00103B
73. Sanchez Medina EI, Linke S, Stoll M, Sundmacher K (2023) Gibbs-helmholtz graph neural network: capturing the temperature dependency of activity coefficients at infinite dilution. Digital Discov 2:781–798. https://doi.org/10.1039/D2DD00142J
74. Chung Y, Vermeire FH, Wu H, Walker PJ, Abraham MH, Green WH (2022) Group contribution and machine learning approaches to predict Abraham solute parameters, solvation free energy, and solvation enthalpy. J Chem Inf Model. https://doi.org/10.1021/acs.jcim.1c01103
75. Walker PJ, Yew H-W, Riedemann A (2022) Clapeyron.jl: an extensible, open-source fluid thermodynamics toolkit. Ind Eng Chem Res 61(20):7130–7153. https://doi.org/10.1021/acs.iecr.2c00326
76. Bell C, Yoel seimen cintronej, Yu T, Alex Matthias A, Vasilyev A., Volpatto D, Felton K, RoryKurek Keskitalo T, Shimwell J CalebBell/thermo: 0.4.2. https://doi.org/10.5281/zenodo.15036476
77. Fredenslund A, Jones RL, Prausnitz JM (1975) Group-contribution estimation of activity coefficients in nonideal liquid mixtures. AIChE J 21(6):1086–1099. https://doi.org/10.1002/aic.690210607
78. Gmehling J, Tiegs D, Medina A, Soares M, Bastos J, Alessi P, Kikic I, Schiller M, Menke J (2008) Dechema chemistry data series, volume ix activity coefficients at infinite dilution. DECHEMA Chemistry Data Series 9. ISBN: 978-3-89746-107-9. https://dechema.de/Analysis+_+Consulting/Chemistry+Data+Series/Volume+IX.html
79. Linstrom PJ, Mallard WG (2001) The NIST chemistry webbook: a chemical data resource on the internet. J Chem Eng Data 46(5):1059–1063. https://doi.org/10.1021/je000236i
80. Ghose AK, Viswanadhan VN, Wendoloski JJ (1998) Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: an analysis of alogp and clogp methods. J Phys Chem A 102(21):3762–3772. https://doi.org/10.1021/jp980230o
81. Dassault Systèmes: BIOVIA COSMOconf 2022 (2022). https://www.3ds.com
82. Dassault Systèmes: BIOVIA COSMOtherm 2022 (2022). https://www.3ds.com
83. Li S-C, Wu H, Menon A, Spiekermann KA, Li Y-P, Green WH (2024) When do quantum mechanical descriptors help graph neural networks to predict chemical properties? J Am Chem Soc. https://doi.org/10.1021/jacs.4c04670
84. Rittig JG, Mitsos A (2024) Thermodynamics-consistent graph neural networks. arXiv preprint arXiv:2407.18372 . arXiv
85. Specht T, Nagda M, Fellenz S, Mandt S, Hasse H, Jirasek F (2024) Hanna: Hard-constraint neural network for consistent activity coefficient prediction. arXiv preprint arXiv:2407.18011

Publisher's Note