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

Thomas Nevolianis[a] & Jan G. Rittig[b] & Alexander Mitsos[b,c,d] & Kai Leonhard[a,*]

[a] RWTH Aachen University, Institute of Technical Thermodynamics, Aachen, Germany

[b] RWTH Aachen University, Process Systems Engineering (AVT.SVT), Aachen, Germany

[c] Forschungszentrum Jülich GmbH, Institute of Climate and Energy Systems ICE-1: Energy Systems Engineering, Jülich, Germany

[d] JARA-ENERGY, Aachen, Germany

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 (RMSE) of 0.44 $\log P$ units for the EXT-Zamora, compared to an RMSE of 0.63 $\log P$ units for single-task models. For the EXT-SAMPL9 dataset, *multi-target learning* achieves an RMSE 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 applicability of the methods used beyond just toluene/water partition coefficients.

1 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 (Testa et al., 2000; Klopman and Zhu, 2005; Mannhold et al., 2009; Andrés et al., 2015) and separation processes (Hostrup et al., 1999; Paes et al., 2022). 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 (Arnott and Planey, 2012; Johnson et al., 2018). 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 (Dunn et al., 1986; Otsuka, 2005; Polte et al., 2022). 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 (Caron and Ermondi, 2005; David et al., 2021). 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) (Klamt, 1995; Klamt et al., 1998), Group Contribution (GC), and Molecular Dynamics (MD) have been employed to predict toluene/water $\log P$ of neutral species (Platts et al., 1999; Lin and Sandler, 2000; Buggert et al., 2009; Loschen and Klamt, 2014; Ince et al., 2015; Bannan et al., 2016; Müller et al., 2024). Recently, the SAMPL9 blind challenge (Amezcua et al., 2023) 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, a semi-empirical model that is partially physics-based and allows application to a variety of systems. Among 18 contributions, we ranked second with an Root-Mean-Square Error (RMSE) of 1.24 $\log P$ units (Nevolianis et al., 2023).

[*] Corresponding author, E-mail: kai.leonhard@ltt.rwth-aachen.de

The best-performing method in the SAMPL9 blind challenge (Amezcua et al., 2023) achieved an RMSE of 1.12 $\log P$ units (Amezcua et al., 2023). Thus, for predicting toluene/water partition coefficients, COSMO-RS has been shown to be more accurate than the GC and MD approaches (Klamt, 2018). 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 (Gilmer et al., 2017; Schweidtmann et al., 2020; Rong et al., 2020; Vermeire and Green, 2021; Winter et al., 2022; Sanchez Medina et al., 2022; Felton et al., 2021; Rittig et al., 2023c; Sun et al., 2023). 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. (2023) 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 250 experimental $\log P$ values from their collected dataset. These 250 experimental toluene/water $\log P$ of neutral species (Zamora et al., 2023) 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 (Zamora et al., 2023). 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 (Vermeire and Green, 2021; Greenman et al., 2022; Fare et al., 2022; Buterez et al., 2024) have employed various multi-fidelity learning approaches. A recent review by Qian et al. (2024) 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* (Qian et al., 2024). *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 9,000 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. (2023) and the SAMPL9 blind challenge. Specifically, a part of the Zamora dataset, comprising 250 experimental $\log P$ values, is used for fine-tuning while the remaining part is reserved for testing, similar to the approach taken with the SAMPL9 dataset, which includes 16 experimental $\log P$ values. 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.

2 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.

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

Name	#data points	origin
LF-QC†	8,891	QC
HF-Exp (Zamora et al., 2023)	213	Exp.
EXT-Zamora (Zamora et al., 2023)	38	Exp.
EXT-SAMPL9 (Amezcua et al., 2023)	16	Exp.

2.1 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 (Cheng et al., 2023), 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 (Ebejer et al., 2012; Landrum et al., 2020). Next, we optimize the molecular structures obtained from RDKit at the GFN2-xTB level of theory (Bannwarth et al., 2019). We further refine the geometries of each molecule in the COSMO state using COSMOconf 23 (Dassault Systèmes, 2023a), with the BP86/TZVPD parametrization and FINE COSMO cavity (Becke, 1988; Perdew, 1986; Rappoport and Furche, 2010). 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 (Dassault Systèmes, 2023b), 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 (Ruiz et al., 2022). 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 Equation 2. Given the uncertainty of 0.45 kcal/mol (Letcher, 2007) for the solvation free energy, the resulting error in $\log P$ is 0.47 $\log P$ units. The LF-QC dataset consists of 8,891 molecules (see Table 1). The LF-QC set is not publicly available due to licensing restrictions. 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.

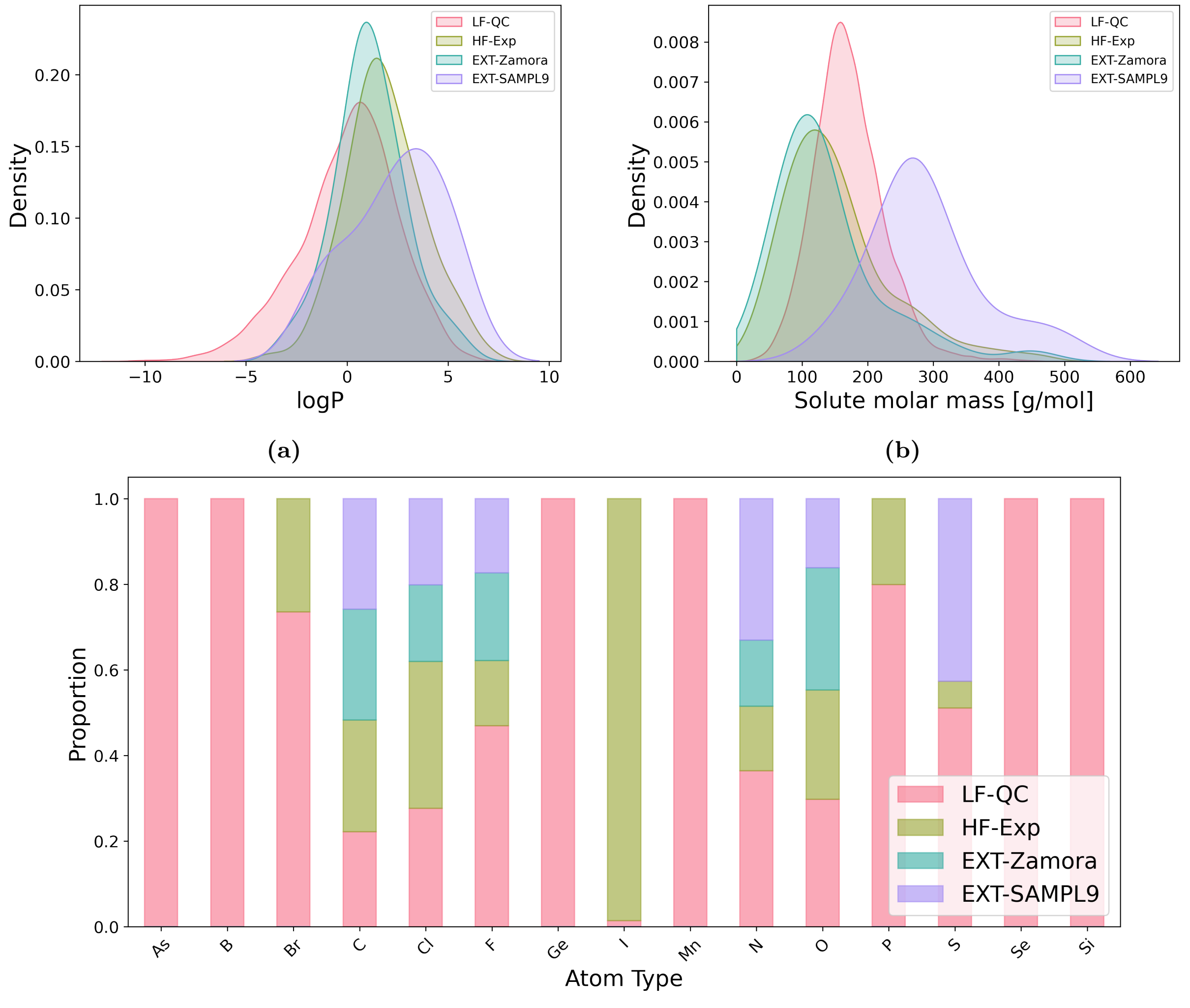
2.2 High fidelity - Experimental dataset

The High fidelity - Experimental (HF-Exp) dataset is obtained from Zamora et al. (2023) 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 (Ruiz et al., 2022; Zamora et al., 2023). 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 (Işık et al., 2019). 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 (Nevolianis et al., 2023). The HF-Exp dataset consists of 213 molecules (see Table 1).

2.3 External Zamora and SAMPL9 datasets

The External - Zamora (EXT-Zamora) and External - SAMPL9 (EXT-SAMPL9) datasets are taken from previous studies (Amezcua et al., 2023; Zamora et al., 2023). 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).

2.4 Dataset comparison and analysis



(a)

(b)

(c)

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.

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 EXT-SAMPL9 datasets. The LF-QC dataset (red) shows a peak around 170 g/mol, 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, suggesting a range of smaller molecule sizes. The EXT-SAMPL9 dataset (purple) displays a peak at higher molar masses, around 300 g/mol, 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.

3 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.

3.1 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 (Klamt, 1995; Klamt et al., 1998). 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. (Eckert and Klamt, 2002; Klamt and Eckert, 2000; Klamt et al., 2002; Loschen et al., 2020; Warnau et al., 2021).

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 (Lipinski et al., 2001; Warnau et al., 2021)

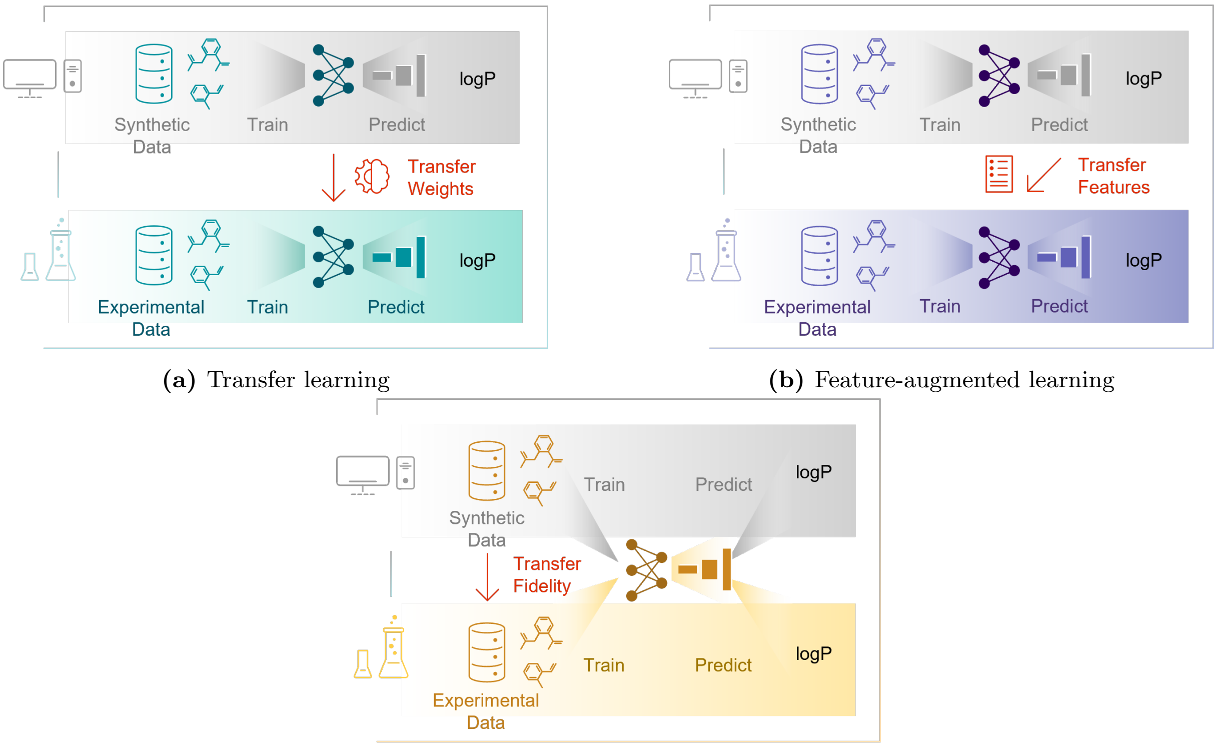
$$\log P_{\mathrm{tol/w}} = \frac{\Delta \mathrm{G}_{\mathrm{Transfer}}}{RT\ln 10} = \frac{\Delta \mathrm{G}_{\mathrm{w}}^{\mathrm{solv}} - \Delta \mathrm{G}_{\mathrm{tol}}^{\mathrm{solv}}}{RT\ln 10}, \tag{2}$$

where $\Delta G_{\mathrm{Transfer}}$ 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. $\Delta G^{\mathrm{solv}}_{\mathrm{w}}$ and $\Delta G^{\mathrm{solv}}_{\mathrm{tol}}$ 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 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:

$$\log P^{\infty}_{\mathrm{tol/w}} = \log \frac{\gamma^{\infty,s}_{\mathrm{w}}}{\gamma^{\infty,s}_{\mathrm{tol}}} \frac{\nu_{\mathrm{w}}}{\nu_{\mathrm{tol}}} \tag{3}$$

3.2 Graph neural networks



(a) Transfer learning

(b) Feature-augmented learning

(c) Multi-target learning

Fig. 2. Overview of different multi-fidelity approaches for training the graph neural network models.

GNN models learn properties directly from the molecular structure and have shown high prediction accuracies for a variety of both pure component (Coley et al., 2017; Schweidtmann et al., 2020; Brozos et al., 2024) and mixture properties (Vermeire and Green, 2021; Sanchez Medina et al., 2022; Rittig et al., 2023a; Qin et al., 2023). 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. (Gilmer et al., 2017; Rittig et al., 2023c; Reiser et al., 2022; Schweidtmann et al., 2023).

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 (Heid et al., 2023). We use the default molecular features (Heid et al., 2023) 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). 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 (Qian et al., 2024) to enhance the prediction of molecular properties: *transfer learning*, *feature-augmented learning*, and *multi-target learning* (see Figure 2). *Transfer learning* (cf. Refs. (Pan and Yang, 2009; Torrey and Shavlik, 2010)) 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. (Buterez et al., 2024)) 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. (Ruder, 2017; Zhang and Yang, 2017)) 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.

4 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.

4.1 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*.

Tab. 2. D-MPNN models performance comparison for EXT-Zamora (Zamora et al., 2023) and EXT-SAMPL9 (Amezcua et al., 2023) datasets.

Model	Mode	dataset	split	**EXT-Zamora** (Zamora et al., 2023)		**EXT-SAMPL9** (Amezcua et al., 2023)	
				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

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).

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.

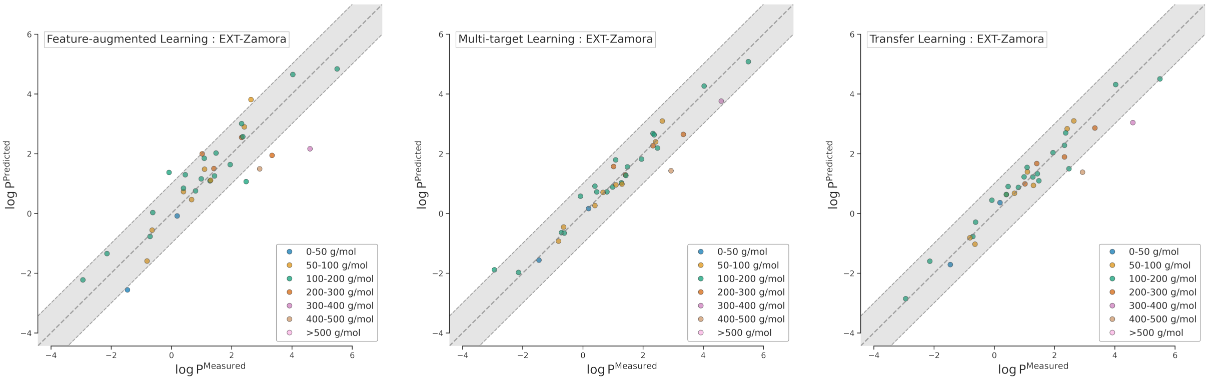
4.2 Impact of molar mass

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 to 500 g/mol is out of the range of ± 1 $\log P$ units (see Figure 3b). The *transfer learning* approach also performs well, though slightly less effectively for larger molecules >300 g/mol. The *feature-augmented learning* approach, however, shows higher variability, particularly for the middle-weight range (100 g/mol to 200 g/mol and 200 g/mol to 300 g/mol).

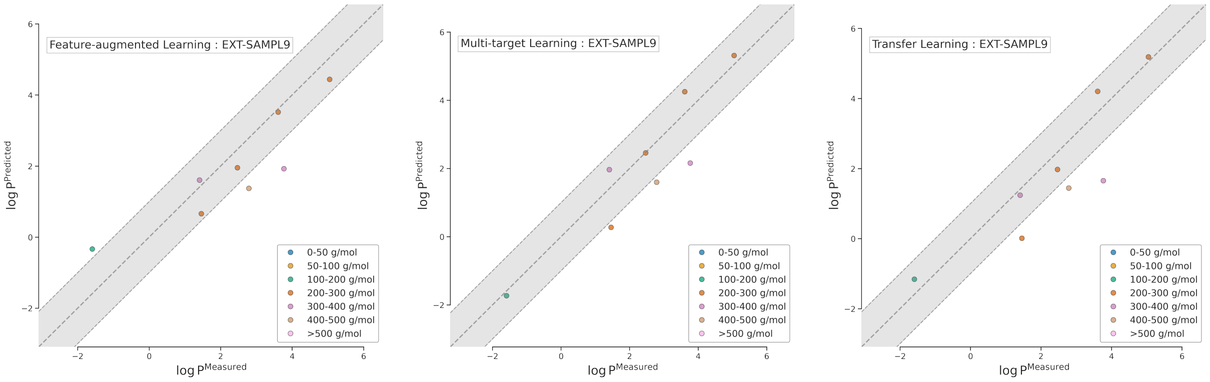
Similarly, for the EXT-SAMPL9 dataset, the *multi-target learning* approach maintains the best performance across most weight categories (see Figure 4). It shows particularly strong results for light molecules and less strong results for heavier molecules. *Transfer learning* remains competitive but 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 to 300 g/mol and >500 g/mol.

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



(a) **(b)** **(c)**

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.



(a) **(b)** **(c)**

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.

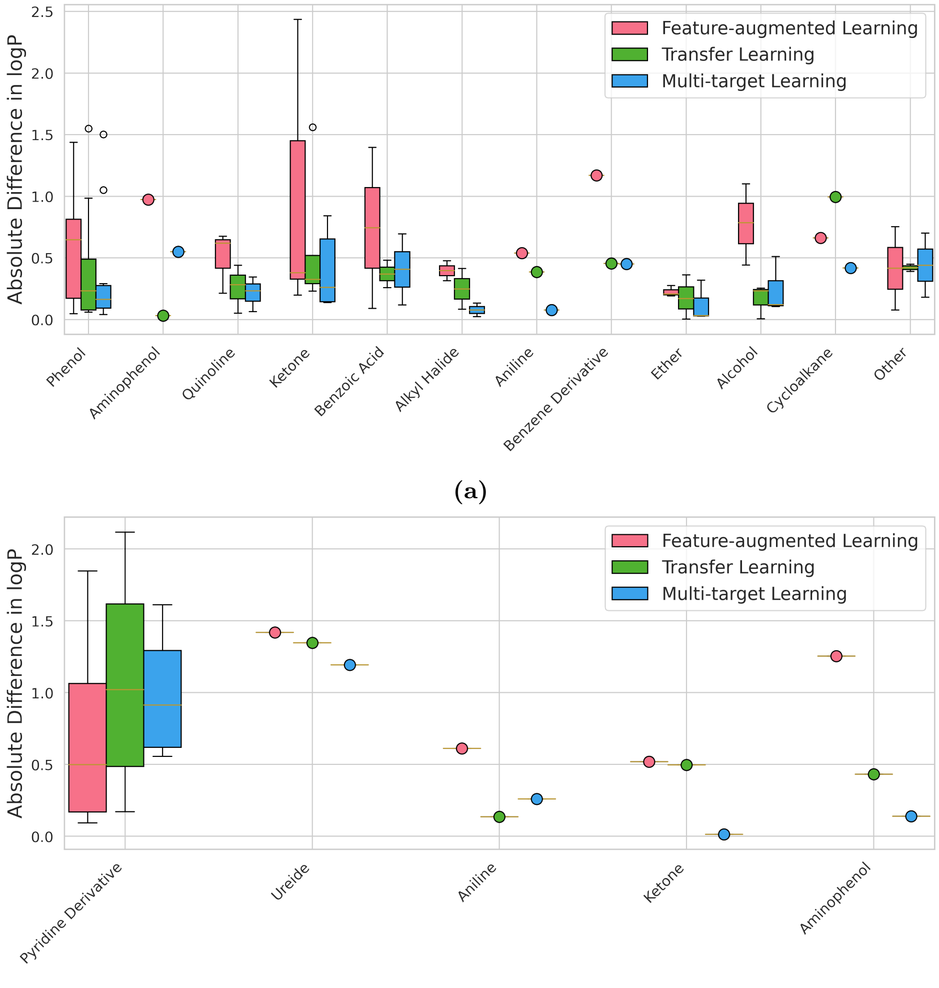
4.3 Impact of chemical classes

We also investigate the model performance across different chemical classes, and illustrate the results in Figure 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 Figure 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 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 (2021) 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 sequen-



(a)

(b)

Fig. 5. Predictive performance of the different multi-fidelity learning approaches accross various chemical classes in the (a) EXT-Zamora and (b) EXT-SAMPL9 datasets.

tial 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.

4.4 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 (Rittig et al., 2023b) and the Gibbs-Helmholtz (GH)-GNN (Sanchez Medina et al., 2023) trained on experimental infinite dilution activity coefficient (IDAC) data from the DECHEMA Chemistry Data Series (Gmehling et al., 2008). To predict the partition coefficients, we employ the already trained models from Ref. (Rittig et al., 2023b) and Ref. (Sanchez Medina et al., 2023), using Equation 3. We calculate the molar volumes with densities

Tab. 3. Model performance comparison for EXT-Zamora (Zamora et al., 2023) and EXT-SAMPL9 (Amezcua et al., 2023) datasets.

Model	Mode	dataset	split	EXT-Zamora (Zamora et al., 2023)		EXT-SAMPL9 (Amezcua et al., 2023)	
				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† by Rittig et al. (2023b)	ensemble	COSMO-AC	-	0.77	0.80	1.56	0.51
GH-GNN† by Sanchez Medina et al. (2023)	ensemble	DECHEMA IDAC	-	1.23	0.48	1.69	0.43
Solvation GNN† by Vermeire and Green (2021)	ensemble	COSMO & exp. G	-	0.27	0.97	1.07	0.77
DirectML† by Chung et al. (2022)	ensemble	COSMO & exp. G	-	0.37	0.95	1.04	0.78
MLR by Zamora et al. (2023)	single	exp	-	1.05	-	0.86	0.85
RFR by Zamora et al. (2023)	single	exp	-	1.13	-	0.84	0.86
COSMO-RS† by Nevolianis et al. (2023)	-	COSMO	-	0.60	0.88	1.23	0.70
MM/PBSA† by Amezcua et al. (2023)	-	-	-	-	-	1.12	0.75

†Models are not trained on partition coefficients.

Some molecules of the test set are included in the training set.

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

and molecular weights for toluene and water from the National Institute of Standards and Technology (NIST) Chemistry webbook (Linstrom and Mallard, 2001). 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 (Vermeire and Green, 2021) and DirectML (Chung et al., 2022). Here, the partition coefficients are calculated using the already trained models from Ref. (Chung et al., 2022) and Ref. (Vermeire and Green, 2021) along with Equation 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 Multi-Linear Regression (MLR) and Random Forest Regression (RFR) from Zamora et al. (2023) that were fitted on the HF-Exp set. The partition coefficient values are taken directly from the original publication (Zamora et al., 2023). These two regression models use 11 input descriptors, including AlogP (octanol/water partition coefficient using Ghose–Crippen atomic contributions (Ghose et al., 1998)), which shows a 58% correlation to the toluene-water partition coefficient, cf. (Zamora et al., 2023). Lastly, we compare to two semi-empirical models: COSMO-RS and Molecular Mechanics Poisson–Boltzmann Surface Area (MM/PBSA) (Amezcua et al., 2023). In the COSMO-RS approach (Nevolianis et al., 2023), the geometry of each molecule is optimized at GFN2-xTB (Bannwarth et al., 2019) level and further in the COSMO state using COSMOconf (Dassault Systèmes, 2022a). Next, the solvation free energies of the molecules are calculated in water and toluene at infinite dilution using COSMOtherm (Dassault Systèmes, 2022b). 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 (Amezcua et al., 2023).

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 (Amezcua et al., 2023). 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. (2023) 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 (Li et al., 2024). 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) (Zamora et al., 2023). This duplication might explain the better performance observed on the EXT-SAMPL9 dataset compared to the EXT-Zamora dataset. 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.

Last, 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.

5 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 9,000 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 (Rittig and Mitsos, 2024; Specht et al., 2024), aiming at more general predictive models.

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.

Acknowledgments

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. 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. The authors thank Prof. William H. Green, Jonathan W. Zheng, Dr. Kevin Greenman, and Dr. Florence Vermeire for stimulating discussion on this work.

Appendix

5.1 Chemical class distribution in the LF-QC and HF-Exp datasets

Tab. 4. Overview of the chemical classes in the LF-QC dataset.

Chemical Class	Count
Ketone	2165
Other	1308
Phenol	1006
Pyridine Derivative	788
Aniline	744
Benzene Derivative	474
Benzoic Acid	399
Alcohol	345
Quinoline	285
Alkyl Halide	272
Sulfonamide	246
Pyrimidine Derivative	201
Aminophenol	175
Phenylbutylamine	102
Ether	93
Thiophene Derivative	92
Ureide	84
Phenylethanolamine	43
Cycloalkane	42
Indole	15
Piperazine Derivative	12

Tab. 5. Overview of the chemical classes in the HF-Exp dataset.

Chemical Class	**Count**
Phenol	47
Other	37
Ketone	29
Alkyl Halide	14
Benzene Derivative	13
Quinoline	11
Alcohol	10
Ether	10
Aniline	9
Phenylbutylamine	8
Pyridine Derivative	7
Benzoic Acid	7
Ureide	2
Aminophenol	2
Pyrimidine Derivative	2
Phenylethanolamine	2
Cycloalkane	2
Sulfonamide	1

5.2 Hyperparameters

Tab. 6. Searchable hyperparameters using `chemprop_hyperopt` taken from (Heid et al., 2023).

Keyword	**Description**
activation	The activation function used after each linear layer, when necessary.
aggregation	The aggregation function used when constructing a molecule-level representation from node-level representations.
aggregation_norm	The normalization factor if using norm aggregation.
batch_size	The minibatch size.
depth	The number of message-passing iterations.
dropout	The dropout probability after each layer in both the D-MPNN encoder and FFN.
ffn_hidden_size	The size of each hidden layer in the FFN.
ffn_num_layers	The number of layers in the FFN.
hidden_size	The message size in the D-MPNN encoder.
max_lr	The maximum learning rate used in the learning rate scheduler.
init_lr	The initial learning rate expressed as the ratio of init_lr to max_lr.
final_lr	The final learning rate expressed as the ratio of final_lr to max_lr.
warmup_epochs	The number of epochs over which to ramp up the learning rate from init_lr to max_lr, expressed as a fraction of the total training epochs.

5.3 Hyperparameter results

Tab. 7. Hyperparameter results for the *multi-target learning* method.

Hyperparameter	**Value**
activation	ReLU
aggregation	sum
aggregation_norm	76.0
batch_size	140
depth	4
dropout	0.05
ffn_hidden_size	300
ffn_num_layers	3
final_lr	7.88154413271554e-05
hidden_size	800
init_lr	6.193713383888547e-06
max_lr	0.004155920461904726
warmup_epochs	15

Tab. 8. Hyperparameter results for the *transfer learning* method.

Hyperparameter	**Value**
activation	ReLU
aggregation	norm
aggregation_norm	192.0
batch_size	50
depth	6
dropout	0.2
ffn_hidden_size	700
ffn_num_layers	3
final_lr	0.0026602019364376866
hidden_size	900
init_lr	5.3010992525841e-05
max_lr	0.004540534678184651
warmup_epochs	5

Tab. 9. Hyperparameter results for the *feature-augmented learning* method.

Hyperparameter	**Value**
activation	LeakyReLU
aggregation	mean
aggregation_norm	99.0
batch_size	30
depth	3
dropout	0.1
ffn_hidden_size	2400
ffn_num_layers	1
final_lr	0.00022766782311393923
hidden_size	2300
init_lr	1.145260591279654e-06
max_lr	0.00354533150456784
warmup_epochs	10

Bibliography

Amezcua, M., Mobley, D. L., and Bergazin, T. D. (2023). samplchallenges/sampl9: 0.8.

Andrés, A., Rosés, M., Ràfols, C., Bosch, E., Espinosa, S., Segarra, V., and Huerta, J. M. (2015). Setup and validation of shake-flask procedures for the determination of partition coefficients (logd) from low drug amounts. *European Journal of Pharmaceutical Sciences*, 76:181–191.

Arnott, J. A. and Planey, S. L. (2012). The influence of lipophilicity in drug discovery and design. *Expert Opinion on Drug Discovery*, 7(10):863–875.

Bannan, C. C., Calabró, G., Kyu, D. Y., and Mobley, D. L. (2016). Calculating partition coefficients of small molecules in octanol/water and cyclohexane/water. *Journal of Chemical Theory and Computation*, 12(8):4015–4024.

Bannwarth, C., Ehlert, S., and 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. *Journal of Chemical Theory and Computation*, 15(3):1652–1671. PMID: 30741547.

Becke, A. D. (1988). Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A*, 38:3098–3100.

Brozos, C., Rittig, J. G., Bhattacharya, S., Akanny, E., Kohlmann, C., and Mitsos, A. (2024). Graph neural networks for surfactant multi-property prediction. *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, 694:134133.

Buggert, M., Cadena, C., Mokrushina, L., Smirnova, I., Maginn, E. J., and Arlt, W. (2009). Cosmo-rs calculations of partition coefficients: Different tools for conformation search. *Chemical Engineering & Technology*, 32:977–986.

Buterez, D., Janet, J. P., Kiddle, S. J., Oglic, D., and Lió, P. (2024). Transfer learning with graph neural networks for improved molecular property prediction in the multi-fidelity setting. *Nature Communications*, 15(1).

Caron, G. and Ermondi, G. (2005). Calculating virtual log p in the alkane/water system ($\log \mathrm{p}^{n}_{alk}$) and its derived parameters $\delta \log p^{n}_{\text{oct-alk}}$ and $\log \mathrm{d}^{\mathrm{pH}}_{\mathrm{alk}}$. *Journal of Medicinal Chemistry*, 48(9):3269–3279.

Cheng, J.-P., Yang, J.-D., Xue, X.-S., Ji, P., Li, X., and Wang, Z. (2023). ibond website. http://ibond.nankai.edu.cn/.

Chung, Y., Vermeire, F. H., Wu, H., Walker, P. J., Abraham, M. H., and Green, W. H. (2022). Group contribution and machine learning approaches to predict abraham solute parameters, solvation free energy, and solvation enthalpy. *Journal of chemical information and modeling*.

Coley, C. W., Barzilay, R., Green, W. H., Jaakkola, T. S., and Jensen, K. F. (2017). Convolutional embedding of attributed molecular graphs for physical property prediction. *Journal of Chemical Information and Modeling*, 57(8):1757–1772.

Dassault Systèmes (2022a). BIOVIA COSMOconf 2022.

Dassault Systèmes (2022b). BIOVIA COSMOtherm 2022.

Dassault Systèmes (2023a). BIOVIA COSMOconf 2023.

Dassault Systèmes (2023b). BIOVIA COSMOtherm 2023.

David, L., Wenlock, M., Barton, P., and Ritzén, A. (2021). Prediction of chameleonic efficiency. *ChemMedChem*, 16(17):2669–2685.

Dunn, W. J., Block, J. H., and Pearlman, R. S. (1986). *Partition Coefficient: Determination and Estimation*. Pergamon Press, New York. Published in cooperation with the American Pharmaceutical Association, Academy of Pharmaceutical Sciences.

Ebejer, J.-P., Morris, G. M., and Deane, C. M. (2012). Freely Available Conformer Generation Methods: How Good Are They? *J. Chem. Inf. Model.*, 52(5):1146–1158.

Eckert, F. and Klamt, A. (2002). Fast solvent screening via quantum chemistry: COSMO-RS approach. *AIChE J.*, 48(2):369–385.

Fare, C., Fenner, P., Benatan, M., Varsi, A., and Pyzer-Knapp, E. O. (2022). A multi-fidelity machine learning approach to high throughput materials screening. *npj Computational Materials*, 8(1).

Felton, K. C., Ben-Safar, H., and Lapkin, A. A. (2021). Deepgamma : A deep learning model for activity coefficient prediction. *1st Annual AAAI Workshop on AI to Accelerate Science and Engineering*.

Ghose, A. K., Viswanadhan, V. N., and Wendoloski, J. J. (1998). Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: An analysis of alogp and clogp methods. *The Journal of Physical Chemistry A*, 102(21):3762–3772.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). Neural message passing for quantum chemistry. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1263–1272. PMLR.

Gmehling, J., Tiegs, D., Medina, A., Soares, M., Bastos, J., Alessi, P., Kikic, I., Schiller, M., and Menke, J. (2008). Dechema chemisry data series, volume ix activity coefficients at infinite dilution. *DECHEMA Chemistry Data Series*, 9.

Greenman, K. P., Green, W. H., and Gómez-Bombarelli, R. (2022). Multi-fidelity prediction of molecular optical peaks with deep learning. *Chemical Science*, 13(4):1152–1162.

Heid, E., Greenman, K. P., Chung, Y., Li, S.-C., Graff, D. E., Vermeire, F. H., Wu, H., Green, W. H., and McGill, C. J. (2023). Chemprop: A machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling*, 64(1):9–17.

Hostrup, M., Harper, P. M., and Gani, R. (1999). Design of environmentally benign processes: Integration of solvent design and separation process synthesis. *Comput. Chem. Eng.*, 23:1395–1414.

Ince, A., Carstensen, H.-H., Reyniers, M.-F., and Marin, G. B. (2015). First-principles based group additivity values for thermochemical properties of substituted aromatic compounds. *AIChE J.*, 61:3858–3870.

Işık, M., Levorse, D., Mobley, D. L., Rhodes, T., and Chodera, J. D. (2019). Octanol–water partition coefficient measurements for the sampl6 blind prediction challenge. *Journal of Computer-Aided Molecular Design*, 34(4):405–420.

Johnson, T. W., Gallego, R. A., and Edwards, M. P. (2018). Lipophilic efficiency as an important metric in drug design. *Journal of Medicinal Chemistry*, 61(15):6401–6420.

Klamt, A. (1995). Conductor-like Screening Model for Real Solvents: A New Approach to the Quantitative Calculation of Solvation Phenomena. *The Journal of Physical Chemistry*, 99(7):2224–2235.

Klamt, A. (2018). The cosmo and cosmo-rs solvation models. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 8(1):e1338.

Klamt, A. and 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. 00530.

Klamt, A., Jonas, V., Bürger, T., and Lohrenz, J. C. (1998). Refinement and parametrization of cosmo-rs. *The Journal of Physical Chemistry A*, 102(26):5074–5085.

Klamt, A., Krooshof, G. J. P., and Taylor, R. (2002). COSMOSPACE: Alternative to conventional activity-coefficient models. *AIChE J.*, 48(10):2332–2349. 00070.

Klopman, G. and 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-Reviews in Medicinal Chemistry*, 5(2):127–133.

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., and Saconne, M. (2020). rdkit/rdkit: 2020/03/1 (q1 2020) release.

Letcher, T. M. (2007). *Development and Applications in Solubility.* Royal Society of Chemistry.

Li, S.-C., Wu, H., Menon, A., Spiekermann, K. A., Li, Y.-P., and Green, W. H. (2024). When do quantum mechanical descriptors help graph neural networks to predict chemical properties? *Journal of the American Chemical Society.*

Lin, S.-T. and Sandler, S. I. (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.

Linstrom, P. J. and Mallard, W. G. (2001). The NIST chemistry webbook: A chemical data resource on the internet. *Journal of Chemical & Engineering Data*, 46(5):1059–1063.

Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. (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. Special issue dedicated to Dr. Eric Tomlinson, Advanced Drug Delivery Reviews, A Selection of the Most Highly Cited Articles, 1991-1998.

Loschen, C. and Klamt, A. (2014). Prediction of solubilities and partition coefficients in polymers using COSMO-RS. *Ind. Eng. Chem. Res.*, 53:11478–11487.

Loschen, C., Reinisch, J., and Klamt, A. (2020). Cosmo-rs based predictions for the sampl6 logp challenge. *Journal of Computer-Aided Molecular Design*, 34(4):385–392.

Mannhold, R., Poda, G. I., Ostermann, C., and Tetko, I. V. (2009). Calculation of molecular lipophilicity: State-of-the-art and comparison of logp methods on more than 96, 000 compounds. *Journal of Pharmaceutical Sciences*, 98(3):861–893.

Müller, S., Nevolianis, T., Garcia-Ratés, M., Riplinger, C., Leonhard, K., and Smirnova, I. (2024). Predicting solvation free energies for neutral molecules in any solvent with opencosmo-rs.

Nevolianis, T., Ahmed, R. A., Hellweg, A., Diedenhofen, M., and 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.

Otsuka, H. (2005). *Purification by Solvent Extraction Using Partition Coefficient*, pages 269–273. Humana Press, Totowa, NJ.

Paes, F. C., Privat, R., Jaubert, J.-N., and 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 Equilibria*, 561:113540.

Pan, S. J. and Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Perdew, J. P. (1986). Density-functional approximation for the correlation energy of the inhomogeneous electron gas. *Phys. Rev. B*, 33:8822–8824.

Platts, J. A., Abraham, M. H., Butina, D., and Hersey, A. (1999). Estimation of molecular linear free energy relationship descriptors by a group contribution approach. 2. prediction of partition coefficients. *Journal of Chemical Information and Computer Sciences*, 40(1):71–80.

Polte, L., Raßpe-Lange, L., Latz, F., Jupke, A., and Leonhard, K. (2022). Cosmo-camped – solvent design for an extraction distillation considering molecular, process, equipment, and economic optimization. *Chemie Ingenieur Technik*, 95(3):416–426.

Qian, E., Chaudhuri, A., Kang, D., and Sella, V. (2024). Multifidelity linear regression for scientific machine learning from scarce data. *arXiv preprint arXiv:2403.08627*.

Qin, S., Jiang, S., Li, J., Balaprakash, P., Lehn, R. C. V., and Zavala, V. M. (2023). Capturing molecular interactions in graph neural networks: a case study in multi-component phase equilibrium. *Digital Discovery*, 2(1):138–151.

Rappoport, D. and Furche, F. (2010). Property-optimized gaussian basis sets for molecular response calculations. *The Journal of Chemical Physics*, 133(13):134105.

Reiser, P., Neubert, M., Eberhard, A., Torresi, L., Zhou, C., Shao, C., Metni, H., van Hoesel, C., Schopmans, H., Sommer, T., and Friederich, P. (2022). Graph neural networks for materials science and chemistry. *Communications Materials*, 3(1):93.

Rittig, J. G., Ben Hicham, K., Schweidtmann, A. M., Dahmen, M., and Mitsos, A. (2023a). Graph neural networks for temperature-dependent activity coefficient prediction of solutes in ionic liquids. *Computers and Chemical Engineering*, 171:108153.

Rittig, J. G., Felton, K. C., Lapkin, A. A., and Mitsos, A. (2023b). Gibbs–duhem-informed neural networks for binary activity coefficient prediction. *Digital Discovery*, 2:1752–1767.

Rittig, J. G., Gao, Q., Dahmen, M., Mitsos, A., and Schweidtmann, A. M. (2023c). Graph neural networks for the prediction of molecular structure–property relationships. In Zhang, D. and Del Río Chanona, E. A., editors, *Machine Learning and Hybrid Modelling for Reaction Engineering*, pages 159–181. Royal Society of Chemistry.

Rittig, J. G. and Mitsos, A. (2024). Thermodynamics-consistent graph neural networks. *arXiv preprint arXiv:2407.18372*. arXiv.

Rong, Y., Bian, Y., Xu, T., Xie, W., Wei, Y., Huang, W., and Huang, J. (2020). Self-supervised graph transformer on large-scale molecular data. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12559–12571. Curran Associates, Inc.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*. arXiv.

Ruiz, R., Zamora, W. J., Ràfols, C., and Bosch, E. (2022). Molecular characteristics of several drugs evaluated from solvent/water partition measurements: Solvation parameters and intramolecular hydrogen bond indicator. *European Journal of Pharmaceutical Sciences*, 168:106066.

Sanchez Medina, E. I., Linke, S., Stoll, M., and Sundmacher, K. (2022). Graph neural networks for the prediction of infinite dilution activity coefficients. *Digital Discovery*, 1(3):216–225.

Sanchez Medina, E. I., Linke, S., Stoll, M., and Sundmacher, K. (2023). Gibbs–helmholtz graph neural network: capturing the temperature dependency of activity coefficients at infinite dilution. *Digital Discovery*, 2:781–798.

Schweidtmann, A. M., Rittig, J. G., König, A., Grohe, M., Mitsos, A., and Dahmen, M. (2020). Graph neural networks for prediction of fuel ignition quality. *Energy Fuels*, 34(9):11395–11407.

Schweidtmann, A. M., Rittig, J. G., Weber, J. M., Grohe, M., Dahmen, M., Leonhard, K., and Mitsos, A. (2023). Physical pooling functions in graph neural networks for molecular property prediction. *Computers and Chemical Engineering*, 172:108202.

Specht, T., Nagda, M., Fellenz, S., Mandt, S., Hasse, H., and Jirasek, F. (2024). Hanna: Hard-constraint neural network for consistent activity coefficient prediction. *arXiv preprint arXiv:2407.18011*.

Sun, G., Zhao, Z., Sun, S., Ma, Y., Li, H., and Gao, X. (2023). Vapor-liquid phase equilibria behavior prediction of binary mixtures using machine learning. *Chem. Eng. Sci.*, 282:119358.

Testa, B., Crivori, P., Reist, M., and Carrupt, P.-A. (2000). The influence of lipophilicity on the pharmacokinetic behavior of drugs: Concepts and examples. *Perspectives in Drug Discovery and Design*, 19(1):179–211.

Torrey, L. and Shavlik, J. (2010). Transfer learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, pages 242–264. IGI Global.

Vermeire, F. H. and Green, W. H. (2021). Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chemical Engineering Journal*, 418:129307.

Warnau, J., Wichmann, K., and Reinisch, J. (2021). Cosmo-rs predictions of logp in the sampl7 blind challenge. *Journal of Computer-Aided Molecular Design*, 35(7):813–818.

Winter, B., Winter, C., Schilling, J., and Bardow, A. (2022). A smile is all you need: predicting limiting activity coefficients from smiles with natural language processing. *Digital Discovery*, 1(6):859–869.

Zamora, W. J., Viayna, A., Pinheiro, S., Curutchet, C., Bisbal, L., Ruiz, R., Ràfols, C., and Luque, F. J. (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.*, page 10.1039/D3CP01428B.

Zhang, Y. and Yang, Q. (2017). A survey on multi-task learning. *arXiv preprint arXiv:1707.08114*. arXiv.