Journal of Cheminformatics

RESEARCH Open Access

# Geometric deep learning for molecular property predictions with chemical accuracy across chemical space

Maarten R. Dobbelaere[1], István Lengyel[1,2], Christian V. Stevens[3] and Kevin M. Van Geem[1*]

## Abstract

Chemical engineers heavily rely on precise knowledge of physicochemical properties to model chemical processes. Despite the growing popularity of deep learning, it is only rarely applied for property prediction due to data scarcity and limited accuracy for compounds in industrially-relevant areas of the chemical space. Herein, we present a geometric deep learning framework for predicting gas- and liquid-phase properties based on novel quantum chemical datasets comprising 124,000 molecules. Our findings reveal that the necessity for quantum-chemical information in deep learning models varies significantly depending on the modeled physicochemical property. Specifically, our top-performing geometric model meets the most stringent criteria for "chemically accurate" thermochemistry predictions. We also show that by carefully selecting the appropriate model featurization and evaluating prediction uncertainties, the reliability of the predictions can be strongly enhanced. These insights represent a crucial step towards establishing deep learning as the standard property prediction workflow in both industry and academia.

### Scientific contribution

We propose a flexible property prediction tool that can handle two-dimensional and three-dimensional molecular information. A thermochemistry prediction methodology that achieves high-level quantum chemistry accuracy for a broad application range is presented. Trained deep learning models and large novel molecular databases of real-world molecules are provided to offer a directly usable and fast property prediction solution to practitioners.

**Keywords** Artificial intelligence, Deep learning, Thermochemistry, Liquid-phase thermodynamics, Representation learning

## Introduction

Chemical engineering hinges on accurate understanding of physicochemical properties to effectively model processes, design products, and assess environmental impacts [1–3]. Since experimental determination of all properties of every chemical compound is practically infeasible, they are typically estimated computationally [4]. The classical property prediction toolkit comprises, next to quantum chemical calculations, empirical methods such as group contributions. Although various machine learning (ML) approaches have shown higher accuracies and wider application ranges than empirical

*Correspondence:
Kevin M. Van Geem
Kevin.VanGeem@UGent.be
[1] Laboratory for Chemical Technology, Department of Materials, Textiles and Chemical Engineering, Faculty of Engineering and Architecture, Ghent University, Technologiepark 125, 9052 Ghent, Belgium
[2] ChemInsights LLC, Dover, DE 19901, USA
[3] SynBioC Research Group, Department of Green Chemistry and Technology, Faculty of Bioscience Engineering, Ghent University, Coupure Links 653, 9000 Ghent, Belgium

methods [5, 6], they are not yet a standard tool for many researchers.

Industrial experts emphasize the pressing need for faster and more accurate property prediction methods. These methods should consistently predict properties with high accuracy, particularly of chemical structures that are found in various industrial processes, thereby enabling faster decision-making [7, 8]. The target accuracy for a computational method is only well-defined for thermochemistry, where "chemical accuracy" (approximately 1 kcal $mol^{-1}$) is demanded to construct thermodynamically consistent kinetic models [9, 10]. By extending this definition, it is reasoned that "chemically accurate" octanol–water partitioning coefficients ($logK_{OW}$) correspond to errors below 0.7 log units [11]. On the other hand, relative accuracies are suggested for other thermodynamic properties in analogy to process modeling assessment [12]. The main limitation of molecular machine learning models is the lack of high-quality data, which hampers their reliability and application range. To overcome challenges related to the low data regime, methods such as transfer learning and Δ -ML have increasingly been adopted. In transfer learning, a model is trained on a large database with low-accuracy data and a vast application domain to learn a molecular representation [13–15]. The knowledge from that model is then "transferred"; that is, the model is trained for a few epochs on a small dataset with highly accurate data to finally obtain a model that can predict with the accuracy of the small dataset for the application range of the large dataset. Δ-ML consists of training a model on the residual between high-quality and low-quality data [16, 17]. This method is especially effective for quantum chemical data, where a consistent difference exists between high level-of-theory and low level-of-theory data.

Graph neural networks, particularly message-passing neural networks (MPNN), have emerged as the primary model type for property prediction. The neural message-passing framework was introduced in 2017 by Gilmer et al. [18]. Initially, these models only considered two-dimensional (2D) information with a string-based identifier as the sole input. A molecule is mathematically represented as a graph, with the nodes representing the atoms and the edges representing the bonds. The molecular graph is then converted with a graph-traversing algorithm into a numerical representation, which is the input for a nonlinear regression model, typically a neural network. The algorithm's core comprises the message-passing phase, where atom representations are iteratively updated using "messages" from neighboring atoms. Yang et al. [19] adapted this framework to directed MPNNs (D-MPNN) in which messages are related to directed edges rather than nodes. The inclusion of directed edges was motivated to prevent noise in the model training by avoiding unnecessary loops during the message-passing stage. The inclusion of 3D molecular information in a D-MPNN necessitates the handling of DFT-optimized 3D molecular coordinates. Such 3D models fall under the umbrella term *geometric deep learning* and are reviewed in detail by Atz et al. [20] and Duval et al. [21]. There are various approaches to combine 3D information and MPNNs. Biswas et al. [22] incorporated quantum chemically calculated descriptors in the featurization of nodes and edges of a 2D D-MPNN. Axelrod et al. [23], on the other hand, utilized a 3D graph with node and edge featurization. Powerful graph neural network interatomic potentials use invariant geometric information, such as radial distances or angles, to learn representations [24–29]. Individual studies have reported that 3D MPNNs outperform their 2D counterparts on quantum chemical data and in virtual screening tasks [23, 30]. However, it remains unclear whether using geometric information in D-MPNNs is a prerequisite to achieving the desired accuracy for compounds and physicochemical properties that are relevant in an industrial setting.

This study introduces a novel tool designed for rapid prediction of physicochemical properties crucial to a wide array of industrial applications. By constructing four new quantum chemical databases comprising over 124,000 molecules relevant to the chemical and pharmaceutical sectors, we ensure applicability across diverse chemical systems. Additionally, we compile 26,000 experimental data points from public databases, covering six key physicochemical properties. Our model, built on the D-MPNN architecture, is capable of handling both 2D and 3D graph representations. We investigate whether incorporating 3D chemical information enhances prediction accuracy significantly. To achieve chemical accuracy across all properties, we employ Δ-ML for thermochemical properties and transfer learning for liquid-phase thermodynamic properties. Extrapolative tests using various data splits and analysis of learning curves are conducted to assess model robustness. Open-source access to the source code, datasets, and optimized models is provided on https://github.com/mrodobbe/chemperium/ for transparency and reproducibility.

## Results and discussion

### Chemical datasets

Existing large training and pretraining datasets utilized for physicochemical property prediction serve as benchmarks for algorithm evaluation but lack specific alignment with industrial demands. The molecules targeted for reliable prediction tools vary significantly

depending on industry sectors (e.g., base chemicals, pharmaceuticals) and applications (e.g., kinetic modeling, solvent selection). To address this diversity of needs, we have taken into account several criteria in the creation of quantum chemical databases, including molecule size, presence of heteroatoms, and the constituent elements of the molecule. A descriptive evaluation of the composition of the four databases ThermoG3, ThermoCBS, ReagLib20, and DrugLib36 is provided in Fig. 1.

ThermoG3 is a database with quantum chemical properties of 53,550 structures, including radicals, calculated at the B3LYP/6-31G* and the G3MP2B3 levels. ThermoCBS is similar to ThermoG3 but contains 52,837 compounds with properties calculated at the CBS-QB3 level. Compared to the principal quantum chemical benchmark, QM9 [31], ThermoG3 and ThermoCBS have a greater diversity of chemical species, including radical species, different conformers for several compounds, and molecules with up to 23 heavy atoms (Fig. 1a and e). These species are representative of detailed kinetic modeling tasks involving renewable feedstocks. In contrast, QM9 comprises molecules up to nine heavy atoms, and 98% of its molecules belong to four classes (HCON, HCO, HCN, and HC), while only 63% of ThermoG3's and 52% of ThermoCBS's molecules belong to these classes (Fig. 1f). As shown in Fig. 1e, only 3,898 compounds from ThermoG3 and ThermoCBS are found in QM9, making it unique benchmarks for thermochemical property prediction.

ReagLib20 and DrugLib36 are two quantum chemical solvation datasets containing 48 physicochemical properties, constructed using COSMO-RS [32, 33] as pretraining sets in the transfer learning tasks. ReagLib20 contains 45,478 organic molecules of biological and industrial relevance, selected from internal databases, and DrugLib36 counts 40,080 organic molecules selected from Enamine's DDS-50 [34]. It is illustrated in Fig. 1a and b that the databases are complementary in terms of molecular size. ReagLib20 is focused on smaller molecules than DrugLib36, with a much greater diversity in heteroatoms (Fig. 1d and f) in terms of heteroatoms. Hence, ReagLib20 is considered to represent the chemical space of reagent-like molecules, while DrugLib36 covers drug-like molecules.

Experimental data points for six properties ($T_b$, $T_c$, $P_c$, $V_c$, $\log K_{OW}$, $\log S_{aq}$) of 17,156 chemical compounds are collected from various public sources [5, 22, 35]. All chemical compounds in the experimental database have at least one property listed with an experimental value. There is an imbalance in the distribution of compounds since $T_b$ data is mainly available for compounds with up to 12 heavy atoms, while most experimental data points for $\log K_{OW}$ and $\log S_{aq}$ are for compounds with 12 to 36 heavy atoms. An overview of the data statistics is given in Table S1.

Figure 2 shows the accuracy for six COSMO-RS-calculated properties. For each of the properties, experimental data is available for only a small subset of the molecules. Experimental data for the boiling point and the critical parameters does not overlap with the DrugLib36 dataset, as larger, drug-like compounds will likely decompose before reaching their critical state or even boiling point. The calculated data is especially, but not surprisingly, accurate for the critical volume, and the boiling point, octanol–water partition coefficient, and critical temperature are also in good agreement with experiments. The lower accuracy of $\log S_{aq}$ is related to the accuracy of Abraham's linear free energy relationship [36, 37], of which the descriptors are calculated from COSMO-RS $\sigma$-moments. $T_c$ has the lowest accuracy against experimental data, which might be explained by large experimental uncertainties [38].

#### Geometric directed message-passing neural networks

We used directed message-passing neural networks (D-MPNN) to learn the relationship between the molecular structure and a physicochemical property. We based the architecture of the 2D D-MPNN on the methodology described by Yang et al. [19]. To include the third dimension of molecular information, two different geometric D-MPNNs are created which differ from each other by the initial featurization of nodes and edges. We considered in this work geometric D-MPNNs using 3D graphs that differ from 2D graphs by the incorporation of the xyz-coordinates of the atoms. The first geometric D-MPNN uses the same initial atomic featurization as the 2D D-MPNN, which is a well-documented approach [23, 24, 30]. In the second model, we introduce the atomic radial distribution function (RDF) [39] as a novel atom featurization for geometric D-MPNNs. RDFs were chosen as an atom descriptor in accordance with the findings from Wojtuch et al. [40] that information about the atomic neighborhood boosts the predictive performance. In both 3D models, the edges correspond to all atom pairs that are separated from each other by a distance shorter than a cutoff radius $r_C$. An illustration of the RDF-featurized geometric D-MPNN is shown in Fig. 3.

We have trained the two geometric D-MPNNs with $r_C$ values ranging from 1.5 Å to 3.0 Å on the ThermoG3 dataset. The ground truth data is composed of the residual between the standard enthalpy of formation at 298 K ($\Delta H^\circ_{f,298\,K}$) values calculated at G3MP2B3 and B3LYP level-of-theory. $\Delta H^\circ_{f,298\,K}$ is, as a physicochemical property of a molecular conformation, an appropriate

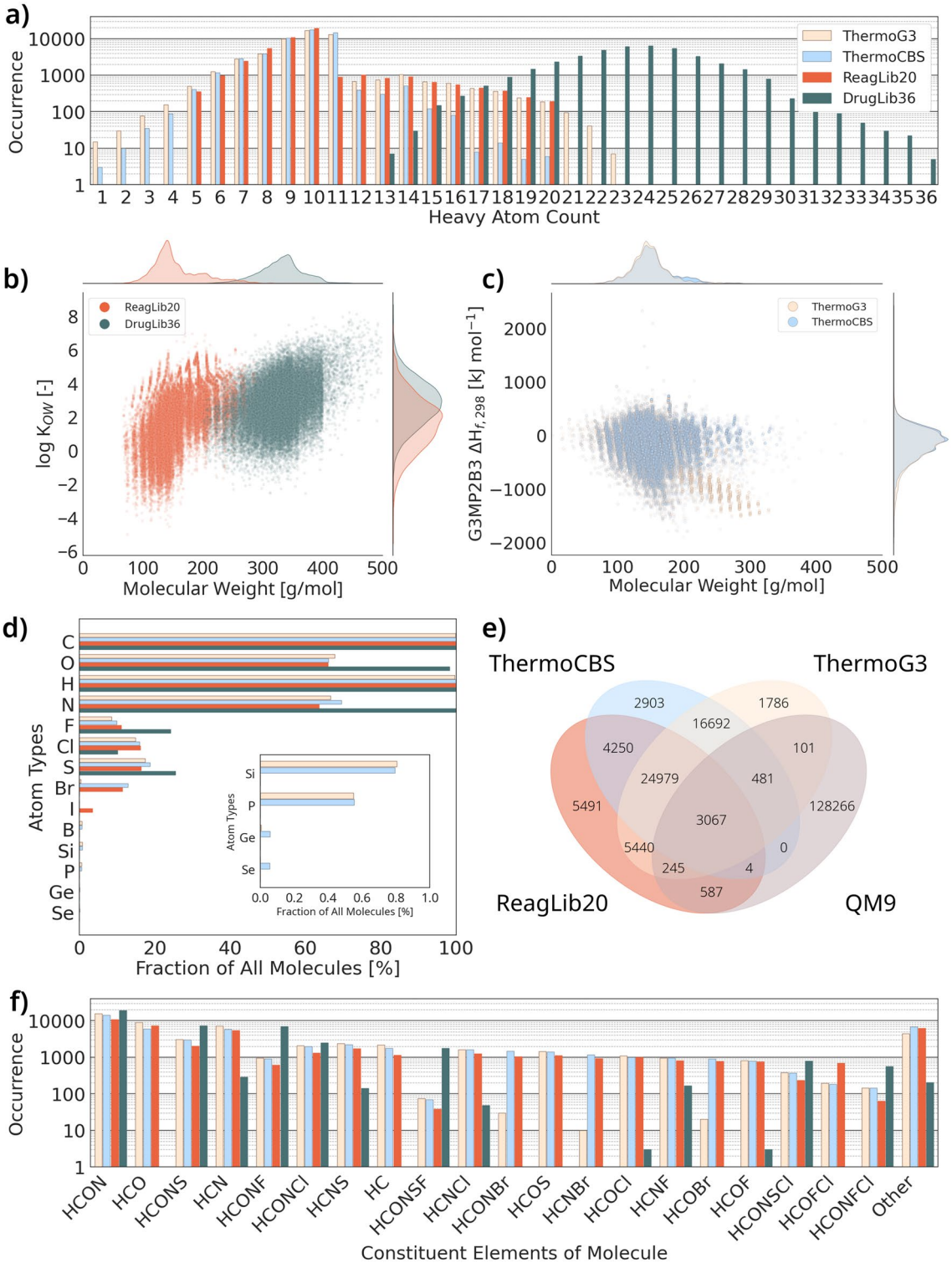


**Fig. 1** Overview of the new quantum chemical databases ThermoG3 (yellow), ThermoCBS (blue), ReagLib20 (orange), and DrugLib36 (green). **a** Heavy atom distribution. **b** Relationship between $\log K_{ow}$ as function of molecular weight for liquid-phase databases. **c** $\Delta H^{\circ}_{f,298K}$ as function of molecular weight for ThermoG3 and ThermoCBS. **d** Atom types distribution. **e** Overlap between the ReagLib20, ThermoG3, and QM9 database. DrugLib36 does not overlap with any of the datasets. **f** Number of molecules per type, classified by constituent elements

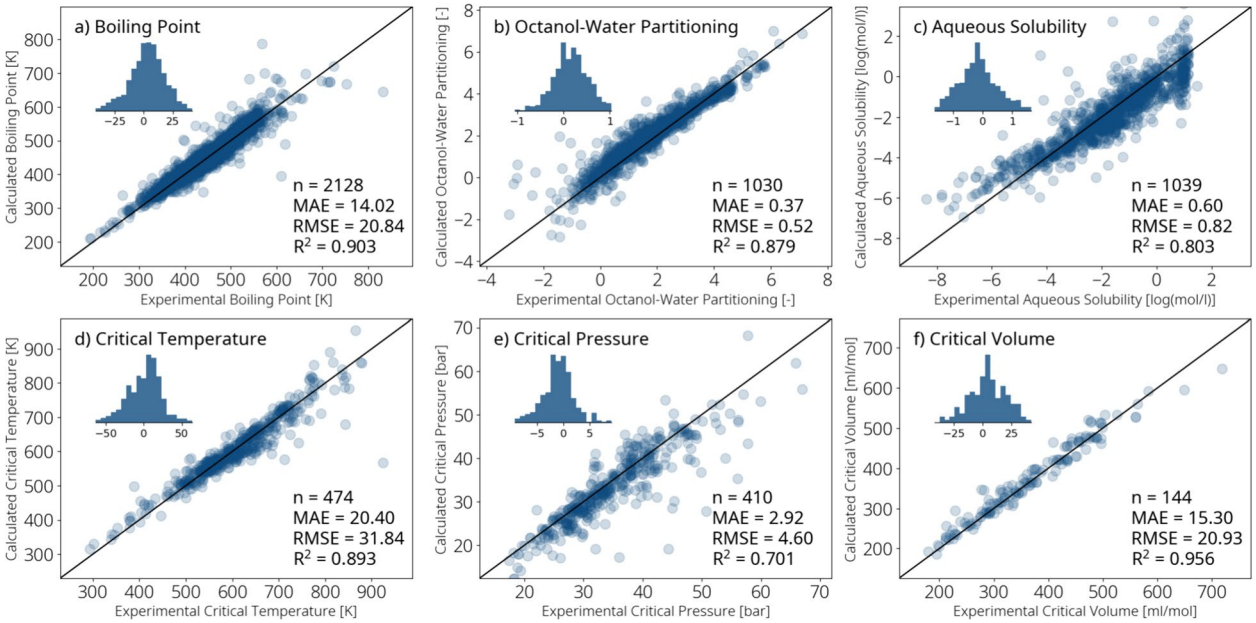


**Fig. 2** Parity plots showing the agreement between experimental and COSMO-RS calculated data for six properties. The bar plots show the difference between calculated and experimental data

property to evaluate the effect of atomic featurization of geometric D-MPNNs on. Since DFT-optimized molecular geometries are used as input, the $\Delta H^{\circ}_{f,298K}$ with DFT quality is calculated without extra cost and can therefore be used as input, validating the choice for Δ-learning. A summary of the results is given in Fig. 4a. The baseline model is a 3D D-MPNN with simple atomic features, in which messages are sent only through edges corresponding to covalent bonds. This baseline model is outperformed by models with the same atomic featurization that use spherical message-passing. Remarkably, the value of the cutoff radius only affects RDF-based models, where large errors are found for $r_C$ smaller than 1.9 Å. This is explained by the presence of bonds in the molecule with larger bond lengths, such as Si-Cl bonds with a length of 2.08 Å. Indeed, the combination of RDFs with a small cutoff radius and limited spherical message-passing lead to disconnected parts in the molecule. In such case, the complete graph lacks information about that bond and the test set error increases exponentially with the number of unmodeled bonds (see Figure S1). This scenario is not observed with simple atomic features, as these contain information about the atomic neighborhood regardless of the $r_C$ value. Even at the lowest evaluated $r_C$ value of 1.5 Å, which is shorter than most covalent bonds, the feature-based model is able to accurately learn a representation of the molecule. Since message-passing is then only performed for the shortest bonds (*e.g.,* C–H, C≡C, …), the molecular representation will then mainly consist of learned atomic contributions, which appears to be sufficient when using a large training set. It is assumed that the low variance in prediction errors over the $r_C$ values is due to using averaged predictions with ensemble learning.

An ideal $r_C$ is found around 2.1 Å and corresponds to a graph with mainly covalent bonds. More details about the uncertainties as function of the cutoff radius are provided in Figure S2 and S3. Increasing the cutoff radius does not lead to a better predictive performance but to a higher computational effort. Therefore, it is recommended to keep $r_C$ as small as reasonably possible. A similar finding is given in the work by Isert et al. [30]. Radii above 3.0 Å are not evaluated to ensure sufficient memory during training.

### Predictive performance for thermochemistry

To probe what the impact of geometric information inclusion is, we compared the performance of 2D and 3D D-MPNNs for predicting $\Delta H^{\circ}_{f,298\ K}$ and $\Delta H_{f,1000K}$ directly and via Δ-ML. The results are summarized in Table 1 by means of the mean absolute error (MAE) and the root-mean-squared error (RMSE) on a random split test set. There is a difference between the order of magnitude of the errors in the direct prediction models and in the Δ -ML models. This discrepancy can be attributed to the output range, which spans over 3,000 kJ/mol for the direct predictions and around 300 kJ/mol for the residual prediction. None of the tested D-MPNNs is capable of

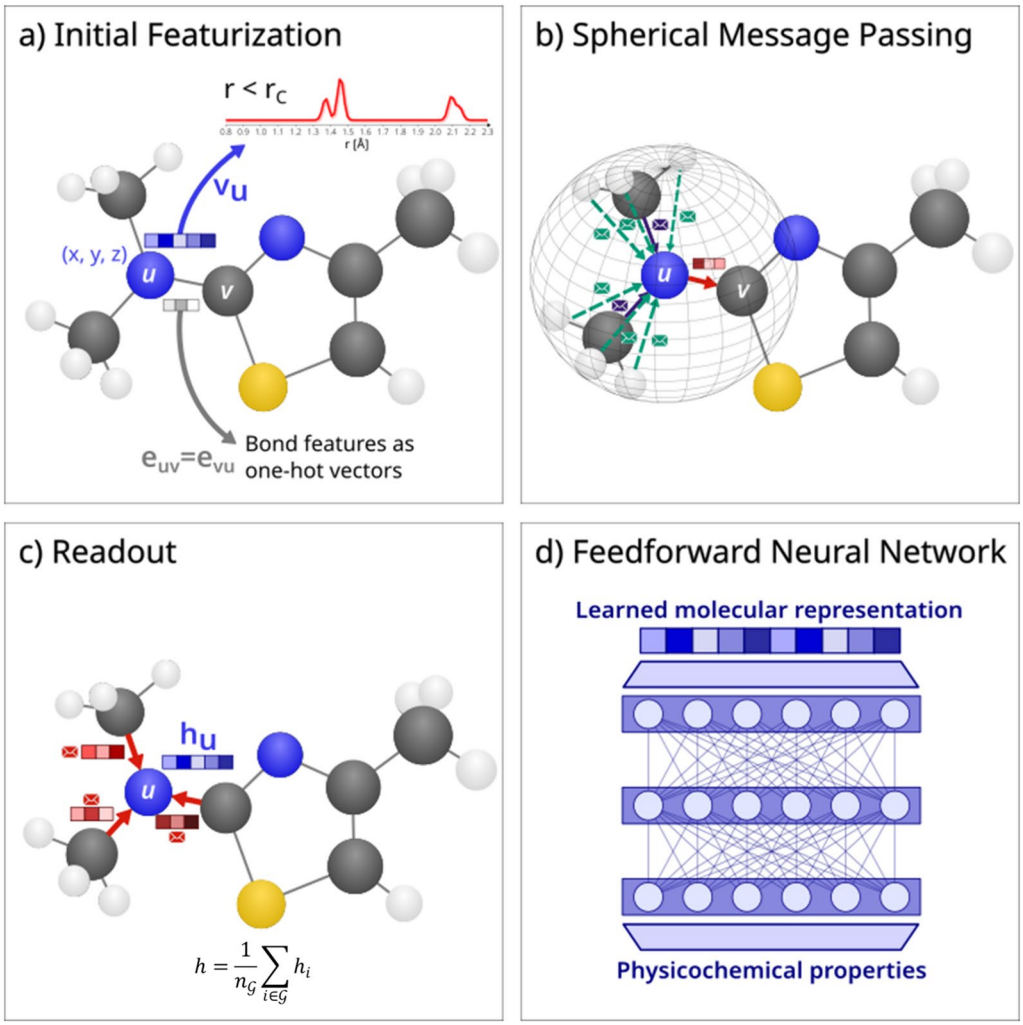


**Fig. 3** Working principle of the geometric D-MPNN. **a** Initial atomic featurization by atomic RDFs with maximal radius $r_C$. **b** Directional edges are updated for $T$ steps using messages from incoming edges within a sphere with radius $r_C$. **c** Atomic representations are created by averaging the incoming edges from covalently bonded atoms. The molecular representation is made by averaging the atomic representations. **d** A feedforward neural network is used for making a regression between the learned molecular representation and the physicochemical properties

reaching "chemical accuracy" in the direct prediction test. Hereby, it is tacitly assumed that the 2D model predicts the value of the lowest-energy conformer. Despite being incapable of distinguishing various conformations, which are present in ThermoG3 and ThermoCBS for thousands of molecules, the 2D models reach a comparable performance with the best 3D models. Additionally, where previous work failed to accurately account for radical species using 2D D-MPNNs [41], the 2D model in this work is able to do so since hydrogens are explicitly added to the graph.

Indeed, with a 2D model, an RMSE as low as 1.84 kJ/mol is reached on the ThermoCBS dataset. However, the use of Δ-ML requires a DFT calculation per se so that the optimized molecular geometry is given without a cost. Since a 2D model is not able to distinguish between various conformations of the same compound, it is not possible to use it for tasks such as conformer search when a workflow is created with a conformer ensemble generation software [42–44]. In that case, a 2D D-MPNN will predict the same output value for every conformation, while the 3D D-MPNN can differentiate. Moreover, conformer ordering is dependent on the accuracy of the computational method and lower-level-of-theory optimized conformers do not guarantee that the high-level-of-theory minimum energy conformer is found [17]. The necessity for predictions with high-level-of-theory accuracy motivates the use of geometric Δ-ML for thermochemistry tasks.

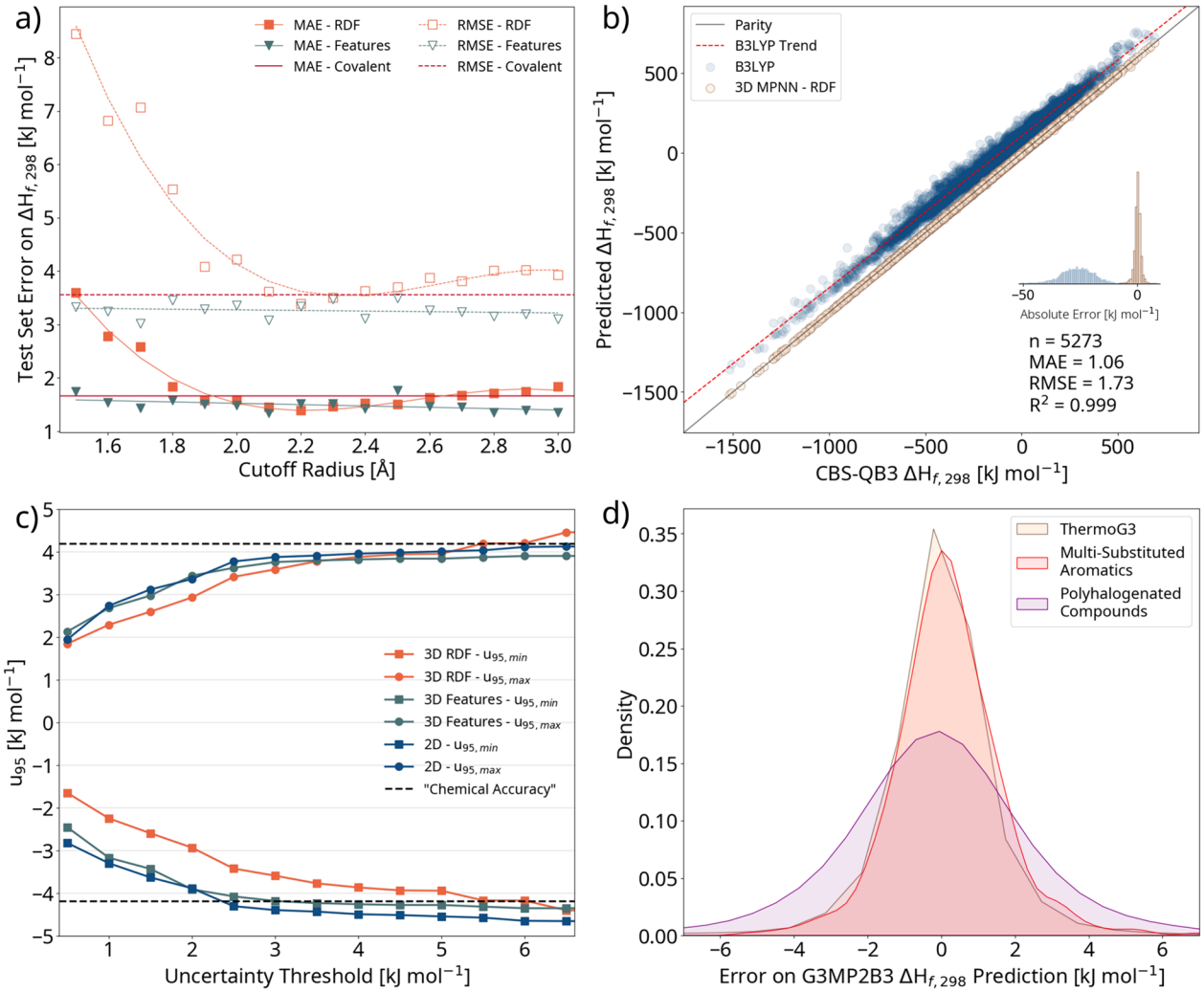


**Fig. 4** Predictive performance for thermochemistry predictions on a random split of test set molecules. **a** Effect of cutoff radius on the mean absolute error (MAE) and root-mean-squared error (RMSE) using tenfold ensemble models for ThermoG3 $\Delta H^{\circ}_{f,298\,K}$ predictions. 3D D-MPNNs are used of which the atoms are featurized with RDFs and simple atomic features. Error bars have been omitted for visual clarity and are given in Figure S2 and S3. **b** Parity plot and error distribution of B3LYP-calculated (blue) and ML-predicted (yellow) $\Delta H^{\circ}_{f,298\,K}$ values against CBS-QB3 calculated data. An optimized tenfold 3D MPNN with RDF featurization and molecular feature descriptor is used for ML predictions. **c** 95% confidence interval for ensemble when predictions with an uncertainty above a threshold are excluded. **d** Error distribution of all data (yellow), multi-substituted aromatics (orange), and polyhalogenated compounds (purple) for predictions with an RDF-featurized 3D D-MPNN

The $\Delta H_f$ at temperatures above 298.15 K are calculated using predicted heat capacity values. Only a slight increase in the prediction error is observed at 1000 K as the prediction error is mainly determined by the error on $\Delta H^{\circ}_{f,298\,K}$. The heat capacity values at 45 different temperatures between 298.15 and 1500 K are predicted in a multitask model that also includes $S_{298}$. NASA polynomials are fitted from the predicted values. These polynomial fits allow to calculate $\Delta H_f$, $S_f$, $c_p$, and $\Delta G_f$ at any temperature between 298.15 and 1500 K. Furthermore, the NASA coefficients allow thermochemistry prediction in CHEMKIN® input format, and direct integration into reaction network generation and reactor simulation packages [45]. A complete overview of the prediction accuracies for each model is given in Table S2 to S9.

Figure 4b depicts the tenfold ensemble performance of the best performing 3D D-MPNN with RDF featurization and molecular feature descriptor using Δ -learning, trained on ThermoCBS $\Delta H^{\circ}_{f,298\,K}$. A systematic deviation is noticeable for the low-level-of-theory data, which is larger at the lower end of the value range and smaller for positive $\Delta H^{\circ}_{f,298\,K}$ values. This deviation arises from approximations that are made in the

**Table 1** Performance of various D-MPNN models on a random test set. Two types of learning are tested: directly predicting $\Delta H^{\circ}_{f,298K}$ and predicting the $\Delta H^{\circ}_{f,298K}$ residual (Δ-ML)

| D-MPNN configuration | | | ThermoG3 | | | | ThermoCBS | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\Delta H^{\circ}_{f,298K}$ [kJ mol$^{-1}$] | | $\Delta H_{f,1000K}$ [kJ mol$^{-1}$] | | $\Delta H^{\circ}_{f,298K}$ [kJ mol$^{-1}$] | | $\Delta H_{f,1000K}$ [kJ mol$^{-1}$] | |
| **Model** | **Atom descriptor** | **Method** | **MAE** | **RMSE** | **MAE** | **RMSE** | **MAE** | **RMSE** | **MAE** | **RMSE** |
| 2D | Features | Direct | 8.66 | 14.24 | 8.91 | 14.51 | **6.94** | **13.63** | 7.20 | 13.90 |
| 3D | Features | Direct | 8.08 | **11.68** | 8.21 | 11.87 | 7.83 | 13.78 | 8.03 | 14.07 |
| 3D | RDF | Direct | **7.09** | 20.44 | 7.25 | 20.36 | 7.67 | 16.96 | 7.81 | 17.18 |
| 2D | Features | Δ-ML | 1.42 | 3.00 | 1.71 | 3.29 | 1.26 | 1.84 | 1.71 | 2.44 |
| 3D | Features | Δ-ML | 1.36 | **2.93** | 1.67 | 3.27 | 1.20 | 1.78 | 1.62 | 2.36 |
| 3D | RDF | Δ-ML | **1.23** | 3.44 | 1.62 | 3.65 | **1.06** | **1.73** | 1.46 | 2.34 |

The optimal values per property are given in bold

Assessment made using mean absolute error (MAE), root-mean-squared error (RMSE). All models are trained as a tenfold ensemble

lower-level-of-theory quantum chemistry calculations, in this case with B3LYP/6-31G*. Nevertheless, lower-level-of-theory methods are sufficiently accurate to locate the minima on the potential energy surface. That is why higher-level-of-theory quantum chemistry methods (*e.g.* G3MP2B3 or CBS-QB3), which take electron correlations into account in more detail, use DFT-optimized molecular geometries. Therefore, the difference in energies is related to local structural features around the atom. The ML models can learn this residual with high accuracy and the data points coincide with the parity line. As such, the Δ-ML approach reduces the computational effort to obtain the thermochemistry to the time required for the DFT optimization of the molecular geometry. The DFT optimization and subsequent vibrational frequency calculations of a molecule with more than 15 non-hydrogen atoms require a computational time in the order of $10^3$ s on a workstation with 8 central processing units (CPU). The most time-consuming part consists of the sequence of single-point calculations with a high-level-of-theory quantum chemistry method, taking approximately $10^6$ s, and is sped up to less than a second using the trained ML models.

Ensemble learning allows for determining uncertainty in the prediction by calculating the standard deviation over the individual model predictions. By selecting a threshold standard deviation, the 95% confidence intervals ($u_{95}$) can be tightened to meet even the most stringent chemical accuracy definition. A detailed overview for the ThermoG3 predictions is given in Fig. 4c. Selecting an uncertainty threshold of 2 kJ/mol, can lower the $u_{95}$ to 3 kJ/mol, so that only 1% of the remaining values have a test set error above 4.184 kJ/mol. The model confidence can be further increased by lowering the threshold value. The RDF-featurized model appears to be the most reliable one, as it can most effectively remove poor predictions and tighten $u_{95}$.

Analogies can be drawn between MPNNs and group contribution methods. Essentially, an MPNN implicitly learns to incorporate higher-order group neighborhoods in the message-passing phase and, as such, outperforms traditional second-order group contribution methods, which solely use the additive character in 2D graph information [46]. Some compound classes that are highly relevant in industrial projects, such as multi-substituted aromatics and polyhalogenated hydrocarbons, are deemed problematic to estimate accurately with group contributions [47, 48]. The limitations of the group additivity are then tried to overcome by introducing non-nearest neighbor interactions [49]. This is a futile job given the immense diversity of aromatic systems [50]. Figure 4d shows that 3D D-MPNNs exhibit a comparable performance for multi-substituted aromatics as for the complete test set. Polyhalogenated compounds, which are about 5% of the complete database, have a wider error distribution than the average molecule in the database.

#### *Prediction of solvation and phase transition properties*

Thermochemical property prediction benefits of using a gas-phase optimized molecular geometry because of the existing relationship between that structure and the property. For many other physical and chemical properties of a molecule, additional effects on the geometry must also be taken into account, which further increases the computation time. In this work, we have evaluated the predictive performance for six molecular properties using a gas-phase single-conformer geometry. The results of this evaluation are given in Fig. 5, in which 2D and

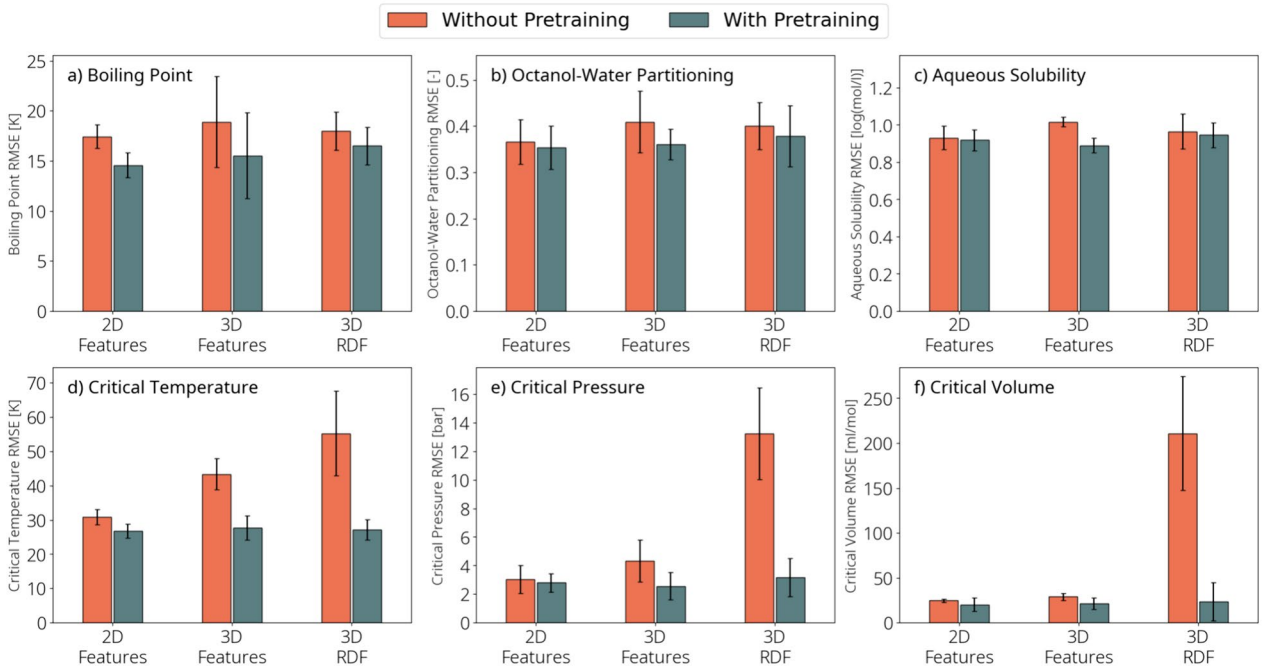


**Fig. 5** RMSE for various D-MPNNs with and without transfer learning tested on experimental $T_b$ (**a**), $logK_{OW}$ (**b**), $logS_{aq}$ (**c**), $T_c$ (**d**), $P_c$ (**e**), and $V_c$ data (**f**)

3D D-MPNN architectures are evaluated, as well as two learning strategies: direct property prediction and transfer learning by pretraining on the ReagLib20 and DrugLib36 databases. Pretraining is performed for $T_b$ and the critical properties by training models on the respective properties in the ReagLib20 database, since these properties are of interest for smaller compounds. Then, these models are fine-tuned on the experimental dataset. The transfer learning procedure for $logK_{OW}$ and $logS_{aq}$ is analogous with the difference that the models are initially trained on the combination of the ReagLib20 and DrugLib36 databases.

The 2D model is superior over 3D models in direct property prediction. This is especially the case for the critical properties of which the experimental data is scarce. However, the improvement with pretraining is statistically insignificant for 2D models as opposed to the 3D models. This might indicate that 3D models need larger data sets to effectively learn structure–property relationships.

#### Generalizability and extrapolative performance

To understand the model's reliability for unseen molecules, it is necessary to quantify the accuracy of extrapolative tests. This is done with scaffold-based test splits [51], where none of the molecules in the training set has the same Bemis-Murcko scaffold [52] as the molecules in the test set. Because the tested molecules are structurally different from the ones the model has seen, this method allows to assess the generalizability capacities. In Fig. 6, we show learning curves for the $\Delta H^{\circ}_{f,298K}$ residual and $logK_{OW}$ using the three D-MPNN architectures that have been evaluated throughout this article. In Fig. 6a and c, the learning curves are determined for random (interpolative) test sets and in Fig. 6b and d, a scaffold-based split is used. Common for all situations is that the RDF-based 3D D-MPNN has a much larger error than the models that use simple atomic features. This result is in line with the results in Fig. 5, where the RDF-based model has a lower accuracy for the properties with the smallest training set sizes, namely the directly predicted critical properties.

The errors from random splitting are lower than those from the scaffold-based splits, which is an expected result. However, the rate at which the error drops for RDF-based models is in all cases higher than for the models that use simple atomic features. Both the 2D and 3D model appear to learn at a similar rate, whereas the RDF-based model takes better advantage of increased training set sizes. In the $logK_{OW}$ results, the 2D D-MPNN outperforms the 3D D-MPNNs in the entire domain. This is in line with the results in Fig. 5, where the 2D model outperformed the 3D models in the direct prediction of liquid-phase properties. The $logK_{OW}$ data is still in the limited data range, since the learning curves follow a continuous linear trend. This indicates that adding more data will further improve the model performance. Both in random and scaffold splits, an error below 0.7 log units

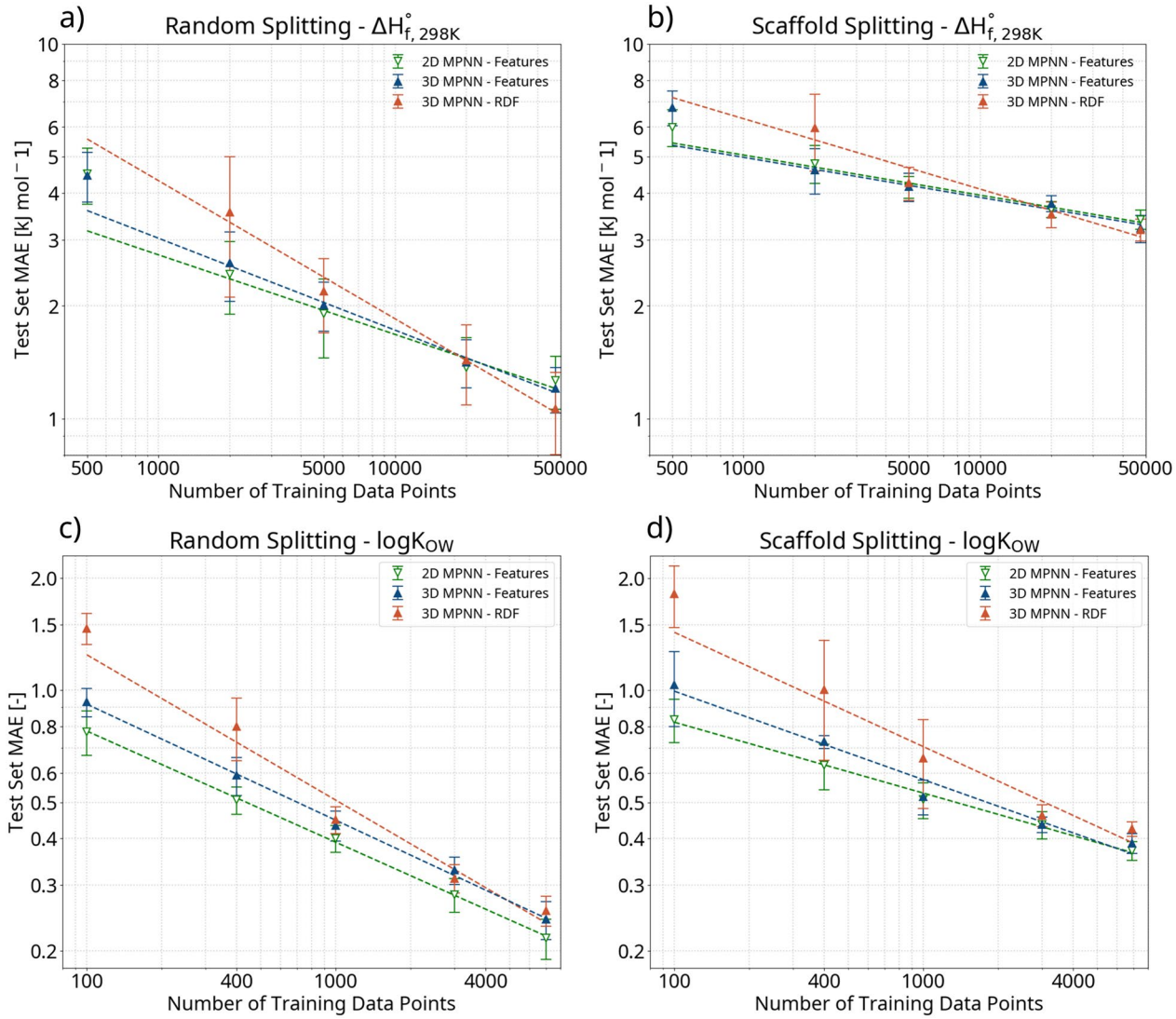


**Fig. 6** Learning curves for models trained on the CBS-QB3-B3LYP $\Delta H_{f,298\,K}$ correction (**a**, **b**) and experimental $logK_{OW}$ data (**c**, **d**). The plots show the test set mean absolute error (MAE) against the training set size for random and scaffold-based splits. In plots **a** and **b**, the value for RDF-based models trained on 500 data points is omitted for visual reasons

is obtained for $logK_{OW}$ models trained on only 1,000 data points, which corresponds to the "chemical accuracy" in theoretical chemistry [11].

The benefit of performing DFT optimization of molecular geometries prior to predicting liquid-phase and critical properties was found to be negligible with the tested D-MPNN models. The utilization of a single gas-phase conformation might even induce an undesired bias since these properties are less sensitive to conformational differences. On the other hand, gas-phase thermochemistry, and in particular energy, strongly depends on the 3D arrangement of atoms. Especially in cases where larger amounts of data are available, an RDF-featurized 3D D-MPNN with ensemble uncertainty increases accuracy and reliability of the predictions. Therefore, the main advantage of geometric models is found in building a sequence of molecular geometry optimization and Δ-ML to significantly accelerate gas-phase thermochemistry calculations while maintaining the accuracy of expensive single-point calculations.

## Conclusions

In this work, we have focused on the importance of geometric information in directed message-passing neural networks (D-MPNN), and their potential to reach "chemically accurate" property predictions

for molecules of industrial interest. To this extent, diverse quantum chemical datasets with more than 124,000 molecules, relevant to chemical and pharmaceutical processes, were developed for training or pretraining machine learning (ML) models. We have found that D-MPNNs are capable of meeting the strictest definition of "chemical accuracy" for $\Delta H^{\circ}_{f,298K}$ predictions by setting up threshold values for the prediction uncertainty. It was shown that only a slight drop in accuracy is witnessed in temperature-dependent thermochemistry predictions up to 1500 K. There are two main arguments for optimizing molecular geometries with DFT before performing ML predictions. Firstly, this enables the use of Δ-ML, in which a correction is learned for a low level-of-theory value, and is crucial for obtaining the desired accuracy. Secondly, 2D models cannot be used for conformational search because of their invariance. The use of a novel radial distribution function (RDF) based atomic featurization outperforms the other models on uncertainty quantification and learning rate tests, hence, increasing the predictive reliability in extrapolative tests. However, the benefits of using molecular geometries were not observed for the prediction of liquid-phase and critical properties. In fact, 2D models obtained similar or even better performance compared to 3D models on both inter- and extrapolative testing. One reason might be that a gas-phase geometry insufficiently relates to the desired properties, while the low number of available highly accurate data points might be another reason. In conclusion, we believe that D-MPNNs are ready for use in industrial chemical engineering applications if (1) the model architecture is carefully chosen depending on the application and available data, and (2) the reliability of the predictions is assessed by setting suitable uncertainty thresholds. The property prediction algorithm developed and used in this work is freely accessible at https://github.com/mrodobbe/chemperium/.

## Methods

### Quantum chemical calculations

The enthalpy of formation of molecules in the ThermoG3 database is computed with the G3MP2B3 method, which is a composite method based on G3 theory [53, 54]. The computation sequence starts with geometry optimizations at the B3LYP/6-31G* level. Then, vibrational frequency computations and a sequence of increasing accuracy single-point energy computations are performed. The enthalpy of formation is calculated from the primary data based on atomization energies.

The liquid-phase properties of molecules in the ReagLib20 and DrugLib36 databases are calculated using the commercial software COSMOtherm, which calculates data based on the COSMO-RS theory [33]. TurboMole [55] was used to perform geometry optimizations and single-point calculations at BP/TZVP level, followed by COSMO-RS/COSMOtherm calculations for solvent effects. Partition coefficients were calculated using the Abraham QSPR module in COSMOtherm [37].

### NASA polynomials

Thermochemical properties at higher temperatures are calculated using the empirical equations developed by Gordon and McBride [56]. The equations contain dimensionless coefficients ($a_1$ to $a_7$) which can be derived from fitting the heat capacity ($c_p$) at various temperatures. Equation (1) is the empirical NASA polynomial for $c_p$.

$$c_p(T) = R\left(a_1 + a_2 T + a_3 T^2 + a_4 T^3 + a_5 T^4\right) \quad (1)$$

The temperature-dependent enthalpy of formation ($\Delta H_f$) is obtained via Eq. (2), so that the NASA polynomial for $\Delta H_f$ is obtained in Eq. (3).

$$\Delta H_f(T) = \Delta H^{\circ}_{f,298.15K} + \int_{298.15K}^{T} c_p(T)dT \quad (2)$$

$$\Delta H_f(T) = R\left(a_1 T + \frac{a_2}{2}T^2 + \frac{a_3}{3}T^3 + \frac{a_4}{4}T^4 + \frac{a_5}{5}T^5 + a_6\right) \quad (3)$$

The temperature-dependent entropy of formation ($S_f$) is calculated via Eq. (4).

$$S_f(T) = R\left(a_1 \ln T + a_2 T + \frac{a_3}{2}T^2 + \frac{a_4}{3}T^3 + \frac{a_5}{4}T^4 + a_7\right) \quad (4)$$

### Geometric message-passing neural networks

This section describes the proposed geometric MPNN framework, which contains four parts: the initial featurization of a 3D molecular graph, the spherical message-passing phase, the readout phase, and a feedforward neural network. The architecture is depicted in Fig. 3.

#### *Initial featurization*

A molecule with $n$ atoms is treated as a 3D molecular graph$\mathcal{G} = (V, E, P)$. $V = \{\mathbf{v}_i\}_{i=1:n}$ is the set of node (atom) features with $\mathbf{v}_i \in \mathbb{R}^{d_v}$ the feature vector for atom$i$. $E = \left\{\mathbf{e}_{ij}\right\}_{j=1:n,k=1:n,k\in\mathcal{N}(j)}$ is the set of edge (bond) features with $\mathbf{e}_{ij} \in \mathbb{R}^{d_e}$ the feature vector for the bond between atom $i$ and atom$j$, where $\mathcal{N}(i)$ denotes the nearest neighboring atoms of atom$j$. It holds that$\mathbf{e}_{ij} = \mathbf{e}_{ji}$. $P = \{\mathbf{r}_i\}_{i=1:n}$ is the set of three-dimensional coordinates with $\mathbf{r}_i \in \mathbb{R}^3$ denoting the x, y, and z-coordinate of atom$i$. We compare two different initial atom embeddings$\mathbf{v}_i$: the atomic features as implemented in Chemprop [57] and an atomic radial distribution function. The atomic features consist of the atomic number, the aromaticity (0 or 1), and three one-hot vectors that denote the degree of the atom, the hybridization, and the chirality. The atomic radial distribution function $g_i(r)$ for atom $i$ is a convolution of the intramolecular distances around atom$i$, and is given in Eq. (5).

$$g_i(r) = \left[1 - \frac{1}{1+\exp(-c(r-b))}\right] \sum_k (m_i m_k)^{0.5} \exp\left[-B(r-d_{ik})^2\right] \tag{5}$$

The radial distribution function is defined by the following parameters: $b$ and $c$ are respectively the decay position and width, $m_i$ is the atomic mass of atom$i$, $B$ is the smoothing parameter, $d_{ik}$ is the interatomic distance between atoms $i$ and$k$. The values of the parameters are taken from the work of Plehiers et al. [17]. The distance $r$ runs from 0.8 Å to$r_C$, which is a cutoff distance. The length of the atomic radial distribution function is taken as 100. In this directed MPNN, the directed edge $\mathbf{e}'_{ij}$ represents an interatomic distance between two atoms $i$ and$j$, which are not necessarily chemically bonded. A directed edge $\mathbf{e}'_{ij}$ is constructed if$d_{ij} < r_C$, and is defined in Eq. (6) as the concatenation of the atomic feature vector $\mathbf{v}_i$ and the edge feature vector$\mathbf{e}_{ij}$. In case atoms $i$ and $j$ are not chemically bonded, then $\mathbf{e}_{ij}$ is a vector of dimension $d_e$ consisting of all zeros. This approach can be considered to be a variant on spherical MPNNs, since edges are constructed for all$j \in \mathcal{U}(i)$, with $\mathcal{U}(i)$ the spherical environment (dt: *Umgebung*) of atom $i$ with radius$r_C$.

$$\mathbf{e}'_{ij} = \text{cat}([\mathbf{v}_i, \mathbf{e}_{ij}]) \tag{6}$$

Before starting the message-passing step, the directed edge hidden states are initialized as given by Eq. (7).

$$h_{ij}^0 = \tau(\mathbf{W}_0 \cdot \mathbf{e}'_{ij}) \tag{7}$$

Here, $\tau$ is the rectified linear unit (ReLU) activation function and $\mathbf{W}_0 \in \mathbb{R}^{d_v+d_e \times d_h}$ is a learned weight matrix with $d_h$ the size of the edge hidden state.

#### *Directional message-passing*

The message-passing phase is the first part of the MPNN and operates for $T$ iterations on the directed 3D molecular graph. In the message-passing phase, information is transmitted through the molecule using message functions. The MPNN updates in iteration $t$ the edge's hidden states $h_{ij}^t$ and messages $m_{ij}^t$ using message function $M_t$ and update function $U_t$. The updated hidden state $h_{ij}^{t+1}$ and message $m_{ij}^{t+1}$ are defined in Eqs. (8) and (9). $n_i$ is the number of atoms in the spherical atomic environment of atom $i$. $\mathbf{W}_m \in \mathbb{R}^{d_h \times d_h}$ is a learned weight matrix.

$$m_{ij}^{t+1} = \frac{1}{n_i - 1} \sum_{k\in\mathcal{U}(i)\setminus j} M_t\left(h_{ki}^t, d_{ik}\right) = \frac{1}{n_i - 1} \sum_{k\in\mathcal{U}(i)\setminus j} \frac{h_{ki}^t}{d_{ik}} \tag{8}$$

$$h_{ij}^{t+1} = U_t\left(h_{ij}^t, m_{ij}^{t+1}\right) = \tau\left(h_{ij}^t + \mathbf{W}_m m_{ij}^{t+1}\right) \tag{9}$$

#### *Readout phase*

In the readout phase, a molecular representation is created from the edge hidden states. First, an atomic message $m_i$ is created by averaging the incoming hidden edges at iteration $T$ [Eq. (10)]. The atomic representation $h_i$ is calculated by concatenating the atomic feature vector $\mathbf{v}_i$ and the atomic message $m_i$, multiplying this new vector with a weight matrix $\mathbf{W}_h \in \mathbb{R}^{d_h \times d_o}$ and sending it through a ReLU activation function $\tau$ [Eq. (11)].

$$m_i = \frac{1}{n_i} \sum_{k\in\mathcal{U}(i)} h_{ki}^T \tag{10}$$

$$h_i = \tau(\mathbf{W}_h \text{cat}([\mathbf{v}_i, m_i])) \tag{11}$$

A molecular representation $h$ is obtained by averaging the atomic representations, as shown in Eq. (12). Generic MPNNs aggregate by summing edge hidden states and atomic representations, but in agreement to the findings of Isert et al. [30] an averaging operation is used to prevent exploding gradients. The learned molecular representation $h$ is used as input for a feedforward neural network.

$$h = \frac{1}{n_\mathcal{G}} \sum_{i\in\mathcal{G}} h_i \tag{12}$$

### Hyperparameter optimization and training details

The model is written using the Python deep learning library Keras (version 2.15) [58], as implemented in TensorFlow (version 2.15) [59]. The training is performed on NVIDIA V100 GPUs. Hyperparameters were optimized using the Hyperband optimizer [60] in Keras-Tuner and a fixed set of hyperparameters is chosen that performs well for the various model configurations and datasets. The size of the edge hidden states $d_h$ is 512 and the size of the molecular representation $d_o$ is 256. The message-passing iteration depth $T$ equals 6. A feedforward neural network with 5 layers and hidden layers size 500 was used. The layers have a bias and are connected with Leaky ReLU activation functions. The weights and biases are initialized using the Glorot initialization scheme [61]. To avoid memory problems, a batch size of 16 was used. The neural network learning is performed with an Adam optimizer using an exponentially decaying learning rate schedule [62].

All model comparisons are made on a single trained model. The optimized performances are given based on the performance of a tenfold model ensemble. Ensemble learning is a common technique in literature to improve model performance by training independent models and averaging their predictions. The averaged predictions of the ten models is used as the final prediction value and the standard deviation on the predictions is used as an uncertainty estimate.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-024-00895-0.

Supplementary Material 1.

#### Acknowledgements

Maarten Dobbelaere acknowledges financial support from the Research Foundation—Flanders (FWO) through doctoral fellowship Grant 1S45522N. The authors acknowledge funding from the European Research Council under the European Union's Horizon 2020 Research and Innovation Programme/ERC Grant agreement No 818607. This project has received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant agreement No 101057816. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation—Flanders (FWO) and the Flemish Government—department EWI.

#### Author contributions

M.R.D. conceived the study, developed the software, trained the models, analyzed the results, and wrote the initial manuscript. I.L. compiled the databases and provided support in analyzing the results. C.V.S. and K.M.V.G. supervised the study. All authors contributed to writing and editing the manuscript.

#### Funding

Maarten Dobbelaere acknowledges financial support from the Research Foundation—Flanders (FWO) through doctoral fellowship Grant 1S45522N. The authors acknowledge funding from the European Research Council under the European Union's Horizon 2020 Research and Innovation Programme/ERC Grant agreement No 818607. This project has received funding from the European Union's Horizon Europe Research and Innovation Programme under Grant agreement No 101057816. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation—Flanders (FWO) and the Flemish Government—department EWI.

#### Availability of data and materials

The four quantum chemical datasets generated in this work (ThermoG3, ThermoCBS, DrugLib36, ReagLib20) and the experimental dataset can be downloaded from Zenodo (https://www.doi.org/https://doi.org/10.5281/zenodo.11409710). The entire source code is provided as open-source software under MIT license in the following repository: https://www.github.com/mrodobbe/chemperium. All conclusions from the paper can be reproduced using the provided scripts. A demo notebook is available in the folder notebooks/demo.ipynb.

## Declarations

#### Competing interests

The authors declare no competing interests.

Received: 8 June 2024 Accepted: 6 August 2024
Published online: 13 August 2024

#### References

1. Poling BE, Prausnitz JM, O'connell JP (2001) Properties of gases and liquids. McGraw-Hill Education, New York
2. Seider WD, Lewin DR, Seader JD, Widagdo S, Gani R, Ng KM (2017) Product and process design principles: synthesis, analysis, and evaluation. John Wiley & Sons, Hoboken
3. Alshehri AS, Gani R, You F (2020) Deep learning and knowledge-based methods for computer-aided molecular design—toward a unified approach: state-of-the-art and future directions. Comput Chem Eng 141:107005
4. Dobbelaere MR, Plehiers PP, Van de Vijver R, Stevens CV, Van Geem KM (2021) Machine learning in chemical engineering: strengths, weaknesses, opportunities, and threats. Engineering 7:1201–1211
5. 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 62:433–446
6. Dobbelaere MR, Ureel Y, Vermeire FH, Tomme L, Stevens CV, Van Geem KM (2022) Machine learning for physicochemical property prediction of complex hydrocarbon mixtures. Ind Eng Chem Res 61:8581–8594
7. Bollini P, Diwan M, Gautam P, Hartman RL, Hickman DA, Johnson M, Kawase M, Neurock M, Patience GS, Stottlemyer A et al (2023) Vision 2050: reaction engineering roadmap. ACS Eng Au. https://doi.org/10.1021/acsengineeringau.3c00023
8. Kontogeorgis GM, Dohrn R, Economou IG, de Hemptinne J-C, ten Kate A, Kuitunen S, Mooijer M, Žilnik LF, Vesovic V (2021) Industrial requirements for thermodynamic and transport properties: 2020. Ind Eng Chem Res 60:4987–5013
9. Pople JA (1999) Nobel lecture: quantum chemical models. Rev Mod Phys 71:1267–1274
10. Ruscic B (2014) Uncertainty quantification in thermochemistry, benchmarking electronic structure computations, and active thermochemical tables. Int J Quantum Chem 114:1097–1101
11. Salthammer T, Grimme S, Stahn M, Hohm U, Palm W-U (2022) Quantum chemical calculation and evaluation of partition coefficients for classical and emerging environmentally relevant organic compounds. Environ Sci Technol 56:379–391
12. van Speybroeck V, Gani R, Meier RJ (2010) The calculation of thermodynamic properties of molecules. Chem Soc Rev 39:1764–1779

13. Grambow CA, Li Y-P, Green WH (2019) Accurate thermochemistry with small data sets: a bond additivity correction and transfer learning approach. J Phys Chem A 123:5826–5835
14. Smith JS, Nebgen BT, Zubatyuk R, Lubbers N, Devereux C, Barros K, Tretiak S, Isayev O, Roitberg AE (2019) Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. Nat Commun 10:2903
15. Vermeire FH, Green WH (2021) Transfer learning for solvation free energies: from quantum chemistry to experiments. Chem Eng J 418:129307
16. Ramakrishnan R, Dral PO, Rupp M, von Lilienfeld OA (2015) Big data meets quantum chemistry approximations: the Δ-machine learning approach. J Chem Theory Comput 11:2087–2096
17. Plehiers PP, Lengyel I, West DH, Marin GB, Stevens CV, Van Geem KM (2021) Fast estimation of standard enthalpy of formation with chemical accuracy by artificial neural network correction of low-level-of-theory ab initio calculations. Chem Eng J 426:131304
18. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE (2017) Neural message passing for quantum chemistry. PMLR
19. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, Guzman-Perez A, Hopper T, Kelley B, Mathea M et al (2019) Analyzing learned molecular representations for property prediction. J Chem Inf Model 59:3370–3388
20. Atz K, Grisoni F, Schneider G (2021) Geometric deep learning on molecular representations. Nat Mach Intell 3:1023–1032
21. Duval A, Mathis SV, Joshi CK, Schmidt V, Miret S, Malliaros FD, Cohen T, Lio P, Bengio Y, Bronstein M (2023) A Hitchhiker's guide to geometric GNNs for 3D atomic systems. Preprint at arXiv arXiv:2312.07511
22. Biswas S, Chung Y, Ramirez J, Wu H, Green WH (2023) Predicting critical properties and acentric factors of fluids using multitask machine learning. J Chem Inf Model 63:4574–4588
23. Axelrod S, Gómez-Bombarelli R (2023) Molecular machine learning with conformer ensembles. Mach Learn Sci Technol 4:035025
24. Gasteiger J, Groß J, Günnemann S (2020) Directional message passing for molecular graphs. Preprint at arXiv:2003.03123
25. Schütt KT, Sauceda HE, Kindermans PJ, Tkatchenko A, Müller KR (2018) SchNet—a deep learning architecture for molecules and materials. J Chem Phys 148:241722
26. Schütt K, Unke O, Gastegger M (2021) Equivariant message passing for the prediction of tensorial properties and molecular spectra
27. Unke OT, Meuwly M (2019) PhysNet: a neural network for predicting energies, forces, dipole moments, and partial charges. J Chem Theory Comput 15:3678–3693
28. Gasteiger J, Giri S, Margraf JT, Günnemann S (2020) Fast and uncertainty-aware directional message passing for non-equilibrium molecules. Preprint at arXiv arXiv:2011.14115
29. Batzner S, Musaelian A, Sun L, Geiger M, Mailoa JP, Kornbluth M, Molinari N, Smidt TE, Kozinsky B (2022) E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. Nat Commun 13:2453
30. Isert C, Kromann JC, Stiefl N, Schneider G, Lewis RA (2023) Machine learning for fast, quantum mechanics-based approximation of drug lipophilicity. ACS Omega 8:2046–2056
31. Ramakrishnan R, Dral PO, Rupp M, von Lilienfeld OA (2014) Quantum chemistry structures and properties of 134 kilo molecules. Sci Data 1:140022
32. Klamt A, Eckert F (2000) COSMO-RS: a novel and efficient method for the a priori prediction of thermophysical data of liquids. Fluid Phase Equilib 172:43–72
33. Klamt A (2018) The COSMO and COSMO-RS solvation models. WIREs Comput Mol Sci 8:e1338
34. Grygorenko OO (2021) Enamine Ltd.: the science and business of organic chemistry and beyond. Eur J Org Chem 2021:6474–6477
35. Mansouri K, Grulke CM, Richard AM, Judson RS, Williams AJ (2016) An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling. SAR QSAR Environ Res 27:911–937
36. Abraham MH, Le J (1999) The correlation and prediction of the solubility of compounds in water using an amended solvation energy relationship. J Pharm Sci 88:868–880
37. Zissimos AM, Abraham MH, Klamt A, Eckert F, Wood J (2002) A comparison between the two general sets of linear free energy descriptors of Abraham and Klamt. J Chem Inf Comput Sci 42:1320–1331
38. Gil L, Otín SF, Embid JM, Gallardo MA, Blanco S, Artal M, Velasco I (2008) Experimental setup to measure critical properties of pure and binary mixtures and their densities at different pressures and temperatures: determination of the precision and uncertainty in the results. J Supercrit Fluids 44:123–138
39. Hemmer MC (2007) Radial distribution functions in computational chemistry—theory and applications, Friedrich-Alexander-Universität Erlangen-Nürnberg
40. Wojtuch A, Danel T, Podlewska S, Maziarka Ł (2023) Extended study on atomic featurization in graph neural networks for molecular property prediction. J Cheminf 15:81
41. Dobbelaere MR, Plehiers PP, Van de Vijver R, Stevens CV, Van Geem KM (2021) Learning molecular representations for thermochemistry prediction of cyclic hydrocarbons and oxygenates. J Phys Chem A 125:5166–5179
42. Raush E, Abagyan R, Totrov M (2024) Efficient generation of conformer ensembles using internal coordinates and a generative directional graph convolution neural network. J Chem Theory Comput 20:4054–4063
43. Seidel T, Permann C, Wieder O, Kohlbacher SM, Langer T (2023) High-quality conformer generation with CONFORGE: algorithm and performance assessment. J Chem Inf Model 63:5549–5570
44. McNutt AT, Bisiriyu F, Song S, Vyas A, Hutchison GR, Koes DR (2023) Conformer generation for structure-based drug design: how many and how good? J Chem Inf Model 63:6598–6607
45. Vandewiele NM, Van Geem KM, Reyniers M-F, Marin GB (2012) Genesys: kinetic model construction using chemo-informatics. Chem Eng J 207–208:526–538
46. Benson SW (1976) Thermochemical kinetics: methods for the estimation of thermochemical data and rate parameters, 2d edn. Wiley, New York
47. Holmes JL, Aubry C (2011) Group additivity values for estimating the enthalpy of formation of organic compounds: an update and reappraisal. 1. C, H, and O. J Phys Chem A 115:10576–10586
48. Holmes JL, Aubry C (2012) Group additivity values for estimating the enthalpy of formation of organic compounds: an update and reappraisal. 2. C, H, N, O, S, and halogens. J Phys Chem A 116:7196–7209
49. 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
50. Dobbelaere MR, Lengyel I, Stevens CV, Van Geem KM (2024) Rxn-INSIGHT: fast chemical reaction analysis using bond-electron matrices. J Cheminf 16:37
51. Spiekermann KA, Stuyver T, Pattanaik L, Green WH (2023) Comment on 'physics-based representations for machine learning properties of chemical reactions.' Mach Learn Sci Technol 4:048001
52. Bemis GW, Murcko MA (1996) The properties of known drugs. 1. Molecular frameworks. J Med Chem 39:2887–2893
53. Baboul AG, Curtiss LA, Redfern PC, Raghavachari K (1999) Gaussian-3 theory using density functional geometries and zero-point energies. J Chem Phys 110:7650–7657
54. Redfern PC, Zapol P, Curtiss LA, Raghavachari K (2000) Assessment of Gaussian-3 and density functional theories for enthalpies of formation of C1–C16 alkanes. J Phys Chem A 104:5850–5854
55. Balasubramani SG, Chen GP, Coriani S, Diedenhofen M, Frank MS, Franzke YJ, Furche F, Grotjahn R, Harding ME, Hättig C et al (2020) TURBOMOLE: modular program suite for ab initio quantum-chemical and condensed-matter simulations. J Chem Phys 152:184107
56. Gordon S (1976) Computer program for calculation of complex chemical equilibrium compositions, rocket performance, incident and reflected shocks, and Chapman-Jouguet detonations. Scientific and Technical Information Office, National Aeronautics and Space Administration
57. 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. https://doi.org/10.1021/acs.jcim.3c01250
58. Chollet F (2015) keras. https://keras.io Accessed 15 May 2024.

59. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M (2016) Tensorflow: large-scale machine learning on heterogeneous distributed systems. Preprint at arXiv arXiv:1603.04467
60. Li L, Jamieson K, DeSalvo G, Rostamizadeh A, Talwalkar A (2018) Hyperband: a novel bandit-based approach to hyperparameter optimization. J Mach Learn Res 18:1–52
61. Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. In: JMLR Workshop and Conference Proceedings
62. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. Preprint at arXiv arXiv:1412.6980

## Publisher's Note