https://doi.org/10.5194/egusphere-2025-1191
Preprint. Discussion started: 20 March 2025

Improved vapor pressure predictions using group contribution-assisted graph convolutional neural networks (GC^{2}NN)

Matteo Krüger[1,*], Tommaso Galeazzo[2,*], Ivan Eremets[1], Bertil Schmidt[3], Ulrich Pöschl[1], Manabu Shiraiwa[2], and Thomas Berkemeier[1]

[1]Multiphase Chemistry Department, Max Planck Institute for Chemistry, Hahn-Meitner-Weg 1, 55128 Mainz, Germany
[2]Department of Chemistry, University of California Irvine, Irvine, California, USA
[3]Department of Computer Science, Johannes Gutenberg University Mainz, Staudingerweg 9, 55128 Mainz, Germany
[*]These authors contributed equally to this work.

Correspondence: Manabu Shiraiwa (m.shiraiwa@uci.edu) and Thomas Berkemeier (t.berkemeier@mpic.de)

Abstract.

The vapor pressures (p_{vap}) of organic molecules play a crucial role in the partitioning of secondary organic aerosol (SOA). Given the vast diversity of atmospheric organic compounds, experimentally determining p_{vap} of each compound is unfeasible. Machine Learning (ML) algorithms allow the prediction of physicochemical properties based on complex representations of molecular structure, but their performance crucially depends on the availability of sufficient training data. We propose a novel approach to predict p_{vap} using group contribution-assisted graph convolutional neural networks (GC^{2}NN). The models use molecular descriptors like molar mass alongside molecular graphs containing atom and bond features as representations of molecular structure. Molecular graphs allow the ML model to better infer molecular connectivity compared to methods using other, non-structural embeddings. We achieve best results with an adaptive-depth GC^{2}NN, where the number of evaluated graph layers depends on molecular size. We present two vapor pressure estimation models that achieve strong agreement between predicted and experimentally-determined p_{vap}. The first is a general model with broad scope that is suitable for both organic and inorganic molecules and achieves a mean absolute error (MAE) of 0.67 log-units (R^2 = 0.86). The second model is specialized on organic compounds with functional groups often encountered in atmospheric SOA, achieving an even stronger correlation with the test data (MAE = 0.36 log-units, R^2 = 0.97). The adaptive-depth GC^{2}NN models clearly outperform existing methods, including parameterizations and group-contribution methods, demonstrating that graph-based ML techniques are powerful tools for the estimation of physicochemical properties, even when experimental data are scarce.

1 Introduction

Secondary organic aerosol (SOA) account for a substantial mass fraction (20-90%) of tropospheric aerosols (Jimenez et al., 2009). They affect the atmosphere's radiative budget and serve as nuclei in cloud droplet and ice crystal formation (Kanakidou et al., 2005; Shrivastava et al., 2017). Furthermore, SOA play a major role in the context of air quality and have been linked to adverse health effects (Pöschl and Shiraiwa, 2015). Understanding SOA formation and evolution is complicated by the

https://doi.org/10.5194/egusphere-2025-1191
Preprint. Discussion started: 20 March 2025

large number and variety of involved organic species and associated reactions and properties, making SOA a source of large uncertainties in climate and air quality modelling (Intergovernmental Panel on Climate Change, 2023).

The saturation vapor pressure (p_{vap}) of a compound determines its partitioning equilibrium between the condensed and gas phase. In the following, we will classify compounds into volatility ranges based on their saturation mass concentrations over the pure liquid (C_0) as proposed by Donahue et al. (2009). The classes are extremely low-volatility organic compounds (ELVOC, $C_0 < 3\times10^{-6}$ μg m^{-3}), low-volatility organic compounds (LVOC, $3\times10^{-6} < C_0 < 3\times10^{-4}$ μg m^{-3}), semi-volatile organic compounds (SVOC, $3\times10^{-4} < C_0 < 300$ μg m^{-3}), intermediate-volatility organic compounds (IVOC, $300 < C_0 < 3\times10^{6}$ μg m^{-3}) and volatile organic compounds (IVOC, $C_0 > 3\times10^{6}$ μg m^{-3}). In the atmosphere, saturation vapor pressure governs new particle formation and gas-particle partitioning, such that SOA mass yield is largely determined by p_{vap} (Pankow, 1987; Kulmala and Kerminen, 2008). However, due to the large number of atmospherically-relevant compounds, exhaustive experimental determination of p_{vap} is not feasible (Goldstein and Galbally, 2007; Bilde et al., 2015).

Various quantitative structure-activity relationship (QSAR) methods for the approximation of thermodynamic properties like p_{vap} or reactivity have been developed to address this limitation: empirical structure-property relationship models often map a sum formula to a thermodynamic property of interest, using algebraic equations with parameters that are fitted to experimental data (Donahue et al., 2011; Li et al., 2016). Group contribution models such as SIMPOL (Pankow and Asher, 2008) and EVAPORATION (Compernolle et al., 2011) can be classified as semi-empirical (Gani, 2019) as they incorporate existing theoretical knowledge about the relationships of structural features and chemical behavior into mathematical equations. This often includes the consideration the occurrences, positions, or interactions of functional groups, while also determining fit parameters using experimental data (Nannoolal et al., 2004; Moller et al., 2008). The consideration of specific functional groups limits group contribution models to certain compound classes, possibly leading to significant errors when applied to molecules outside their applicable range (Tahami et al., 2019). Quantum-mechanical calculation (QM) models based on density functional theory are a common non-empirical approach to property determination (Geerlings et al., 2003), and can be combined with empirical approaches (Ratcliff et al., 2017). Such quantum-mechanical calculations have been used for the generation of large data sets (Wang et al., 2017; Tabor et al., 2019; Besel et al., 2023), facilitating the development of machine learning (ML)-based QSAR models (Lumiaro et al., 2021; Krüger et al., 2022). When categorising ML-based QSAR models, we can distinguish the actual algorithm and the molecular representation that encodes molecular structures into suitable model input, which together majorly determine a ML model's performance in deriving properties from molecular structures (Lumiaro et al., 2021). Combinations successfully applied in previous studies include one-hot encoded Simplified Molecular Input Line Entry System (SMILES) strings with convolutional neural networks (OHE-CNN; Krüger et al., 2022), specific molecular descriptors with decision trees (Armeli et al., 2023) or topological fingerprints with Gaussian process regression (Besel et al., 2024). Galeazzo and Shiraiwa (2022) developed a method to predict glass transition temperature and melting points of small molecules using Extreme Gradient Boosting (XGBoost) and a neural network, respectively, in combination with derived molecular embeddings as molecular fingerprints. The transformation of molecular structures into such machine-readable molecular representations requires the ML models to learn the representation principles along with the physicochemical principles that determine the target property, to the detriment of limiting their application to the prediction of properties with extensive amounts of data (von

https://doi.org/10.5194/egusphere-2025-1191
Preprint. Discussion started: 20 March 2025

Lilienfeld and Burke, 2020). Data curation techniques can improve model accuracy, e.g., through identification and deletion of data points associated with large experimental uncertainty (Gadaleta et al., 2018; Ulrich et al., 2021). Within atmospheric chemistry, only few ML-based QSAR models have been trained exclusively on experimental measurements, as they generally require a large quantity of training data for sufficient model generalization, and a careful and computationally expensive error estimation when only limited amounts of data are available (Galeazzo and Shiraiwa, 2022; Armeli et al., 2023). The overall moderate to poor accuracy of existing QSAR models for p_{vap} prediction exemplifies the need for more accurate, publicly available models (Longnecker et al., 2025).

Graph neural networks (GNNs) are a class of algorithms within the domain of geometric deep learning which have emerged as a powerful addition to machine learning methods in computational chemistry and material sciences in the last decade (von Lilienfeld and Burke, 2020; Reiser et al., 2022). GNNs can be interpreted as an extension of convolutional neural networks beyond fixed dimension grids of data to include irregularly shaped structures (Kipf and Welling, 2017; Bronstein et al., 2017), such as graph-based representations of molecules (Duvenaud et al., 2015; Atz et al., 2021). Molecular graph representations and algorithms that operate on such graphs omit an additional representation learning step and can directly infer intramolecular spatial relations along with properties assigned to graph elements. Furthermore, in contrast to sum formula-based methods, structure-based methods can distinguish structural isomers, which may differ significantly in their properties (Isaacman-VanWertz and Aumont, 2021). Lumiaro et al. (2021) compared a variety of molecular fingerprints in combination with Kernel Ridge Regression, finding graph-based representations to be advantageous compared to canonical descriptive chemical features based methods. For the prediction of absorption, distribution, metabolism, excretion and toxicity (ADMET) properties, Xiong et al. (2021) employed a multi-task graph attention framework addressing classification and regression tasks.

In this work, we propose group contribution-assisted graph convolutional neural network (GC^2NN) models that are simultaneously trained on lists of molecular descriptors as well as graph representations of molecules, in which atom features are mapped to nodes, and bond features mapped to edges of a graph structure. We test model performance on data sets from experimental measurements and QM calculations (Besel et al., 2023), and compare our models with established methods for the determination of p_{vap}: one ML approach, where convolutional neural networks are trained on one-hot encoded SMILES representations (Krüger et al., 2022), two parameterizations, where p_{vap} are derived only from the compounds' elemental composition (Donahue et al., 2011; Li et al., 2016), and SIMPOL (Pankow and Asher, 2008), EVAPORATION (Compernolle et al., 2011), and EPI-Suite (EPI), which are commonly used semi-empirical group-contribution methods.

2 Methods

2.1 Vapor pressure data

We assembled a data set of SMILES representations of 6128 compounds with experimental saturation vapor pressure (p_{vap}) measurements at 298 K by crawling data from pubchem (Kim et al., 2016). In addition, we retrieved the data set published in Naef and Acree (2021), comprised of 2070 compounds. After removal of species present in both data sets, and species that contain elements that occur in fewer than 30 compounds, a total of 6256 unique compounds with experimental p_{vap}

https://doi.org/10.5194/egusphere-2025-1191
Preprint. Discussion started: 20 March 2025

measurements are obtained and referred to as broad data. An overview of molecular substructures in the broad data set is displayed in Fig. 1A. It encompasses various compound types, such as aromatics, alcohols, carboxylic acids, esters, amines, amides, carbonyls, sulfides and nitriles. As the broad data set also contains $\sim 5\%$ inorganic compounds, we refer to compounds in this data set more generally as extremely low-volatility compounds (ELVOC), low-volatility compounds (LVOC), semi-volatile compounds (SVOC), intermediate-volatility compounds (IVOC) and volatile compounds (VOC), thus keeping the same acronyms and vapor pressures bins as Donahue et al. (2009) established for organic compounds. Experimental p_{vap} measurements range from 10^{-10} to 10^{7} Pa. The distribution of saturation concentrations and the number of ELVOC, LVOC, SVOC, IVOC and VOC are summarized in Fig. 1E. For a comparison with established methods for p_{vap} prediction, and to test the method on a data set of compounds that are relevant for the atmosphere, we extract all compounds that lie within the scope of these methods (Pankow and Asher, 2008; Compernolle et al., 2011; Donahue et al., 2011; Li et al., 2016), confining the data set to molecules only consisting of C, H, and O atoms and belonging to the following compound classes: alkanes, (non-aromatic) alkenes, aldehydes, ketones, ethers, esters, peroxides, nitrates, peroxy acyl nitrates, alcohols, acids, hydroperoxides and peracids. This subset of the broad data, referred to as confined data, contains a total of 1371 compounds with much smaller variety of compound classes, including carboxyl, hydroxyl, ester and carbonyl functional groups (Fig. 1B). While the overall p_{vap} range is very similar, the confined data set exhibits a smaller fraction of ELVOC, LVOC and SVOC than the broad data set (Fig. 1C,D,E,F).

In addition to the experimental data, we train and evaluate GC^{2}NN models based on the quantum-mechanical (QM) data set GeckoQ (Besel et al., 2023). This data set contains a total of 31,637 compounds with calculated p_{vap}. Compounds in this data are carbon backbones derived from decane, toluene and α-pinene with various functional groups (including C, O, H). These structures were generated by the GECKO-A mechanism generator that simulates the atmospheric oxidation of hydrocarbons to ensure atmospheric relevance (Aumont et al., 2005). Besel et al. conducted a conformer search using the COSMO*conf* program, calculated individual conformer p_{vap} values with COSMO*therm*, and determined a single p_{vap} accounting for the population of conformers according to the Boltzmann distribution (Wang et al., 2017; Kurtén et al., 2018; Hyttinen et al., 2022).

From each data set, we sample test sets (10% of compounds) that are fully withheld from model training and used to evaluate the trained GC^{2}NN models. The remaining compounds in each data set (90%) are used for training of the GC^{2}NN models, applying 5-fold cross-validation with 80% of data in the training and 20% in the validation set. The resulting data set sizes are the following: broad training: 4505, broad validation: 1126, broad test: 625, confined training: 987, confined validation: 247, confined test: 137, GeckoQ training: 22,778, GeckoQ validation: 5695, and GeckoQ test: 3164. p_{vap} measurements in Pa are logarithmized and scaled to a [0, 1]-interval using min-max scaling.

Of the 1371 molecules in the confined data set, 474 are also contained in the EVAPORATION training data (Compernolle et al., 2011). We ensure that no EVAPORATION training data are present in the test set that is used for comparison between the methods. Note that this only applies to EVAPORATION due to data availability and practicability; any other pre-trained or fitted method is likely to contain some fraction of the test set used in this study in their training data, including Donahue et al. (2011), Li et al. (2016), SIMPOL, and EPI-Suite.

https://doi.org/10.5194/egusphere-2025-1191
Preprint. Discussion started: 20 March 2025

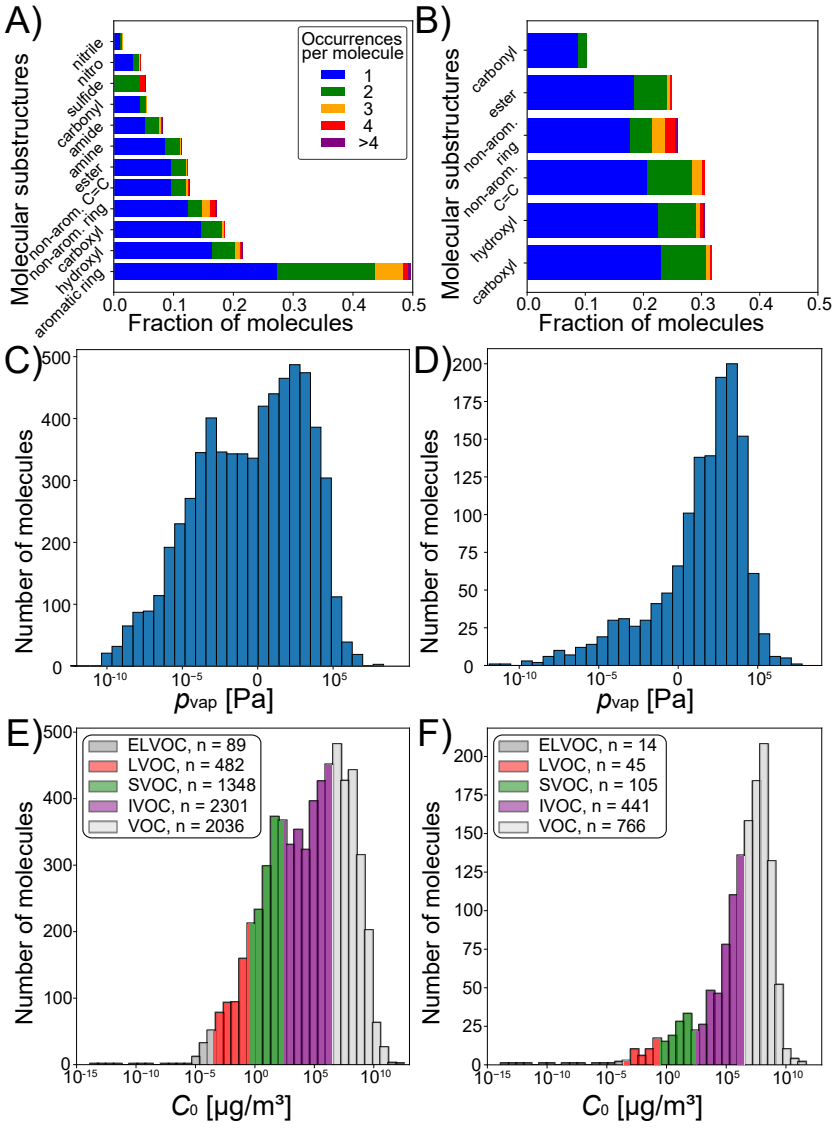

Figure 1. Overview of the two experimental data sets used in this study: broad data set (n = 6256; A, C, E) and confined subset (n = 1371; B, D, F). Panels A and B show all substructures which are present in more than 1% of molecules in the respective data set (not shown: A: nitrate, sulfo, peroxide, organosulfate, peroxy acyl nitrate; B: peroxide). Panels C and D display histograms of experimental vapor pressure measurements in each data set, whereas Panels E and F show the same data as saturation mass concentrations (C_0). The volatility classes are adopted from Donahue et al. (2009).

https://doi.org/10.5194/egusphere-2025-1191
Preprint. Discussion started: 20 March 2025

2.2 Molecular representation

For the graph convolution component of the GC^{2}NN, we transform SMILES representations of molecular structures into graph-representations where atom features are mapped to node features, and bond features to edge features (Tables S1 and S2). The final graph structure is comprised of three tensors. Each node and bond in the graph is associated with a vector of atom features and bond features, respectively. An adjacency matrix indicates the connectivity of atoms in the molecule.

For the model's group contribution component, a list of molecular descriptors (including molar mass, number of atoms for each element, and the number of common functional groups) are derived directly from the SMILES representation of the molecule. The descriptors are specific to each data set and are summarized in Tab. S3. All descriptors and features are one-hot encoded or normalized to a [0, 1] interval.

2.3 Model architecture

We test and compare two group contribution-assisted graph convolutional neural networks (GC^{2}NN) models in this work: a fixed-depth GC^{2}NN (fdGC^{2}NN) model with a fixed number of graph layers, and an adaptive-depth GC^{2}NN (adGC^{2}NN) model where the number of graph layers is dynamically adapted based on a compound's size. Schematic overviews of the adGC^{2}NN and fdGC^{2}NN models are shown in Fig. 2 and Fig. S1, respectively. All GC^{2}NN models encompass two components with separate inputs that are derived from the SMILES-encoded molecular structure. The graph convolution component is comprised of multiple graph convolution layers and graph attention layers. Graph convolution layers apply convolution operations on each node, deriving information from the current node's properties, as well as its neighbors (Kipf and Welling, 2017; Zhang et al., 2019). Graph attention layers enable the model to also derive information from edge attributes (Veličković et al., 2017; Withnall et al., 2020; Tang et al., 2020). Each graph attention or convolution layer increases the nodes' receptive fields, i.e. the distance between two nodes (and hence atoms) that still affect each other. To account for variable molecule sizes, we use the maximum distance between two atoms of a compound ($maxdist$) to determine the number of processing graph layers in the adGC^{2}NN, with a maximum of five layers for molecules with $maxdist > 4$. In the fdGC^{2}NN, all compounds are indiscriminately passed through five graph layers. The models' group contribution component is comprised of fully connected hidden layers that process additional molecular descriptors in parallel. Graph layer-specific merging layers map the information obtained from both model components to the output layer and a vapor pressure prediction. We use the Python packages RDKit and PyTorch (and PyTorch_Geometric) to generate the graph representations of molecular species from SMILES and train GC^{2}NN models (Landrum, 2013; Paszke et al., 2019).

The Python package Optuna (Akiba et al., 2019) is used to efficiently optimize hyper-parameters of each GC^{2}NN model, using 5-fold cross-validation to mitigate variability due to the small data sets. We select MAE as loss function and optimize hyperparameters by minimizing average validation loss across all cross-validation folds, but reject models if the MAE standard deviation is larger than 0.08, to ensure robust model architectures. All models are trained to a maximum of 400 training epochs, unless validation loss does not decrease for 20 consecutive epochs. If so, model parameters are reset to the state of the epoch where the last validation loss decrease occurred, and training is terminated to avoid over-fitting. After the selection of suitable

https://doi.org/10.5194/egusphere-2025-1191
Preprint. Discussion started: 20 March 2025

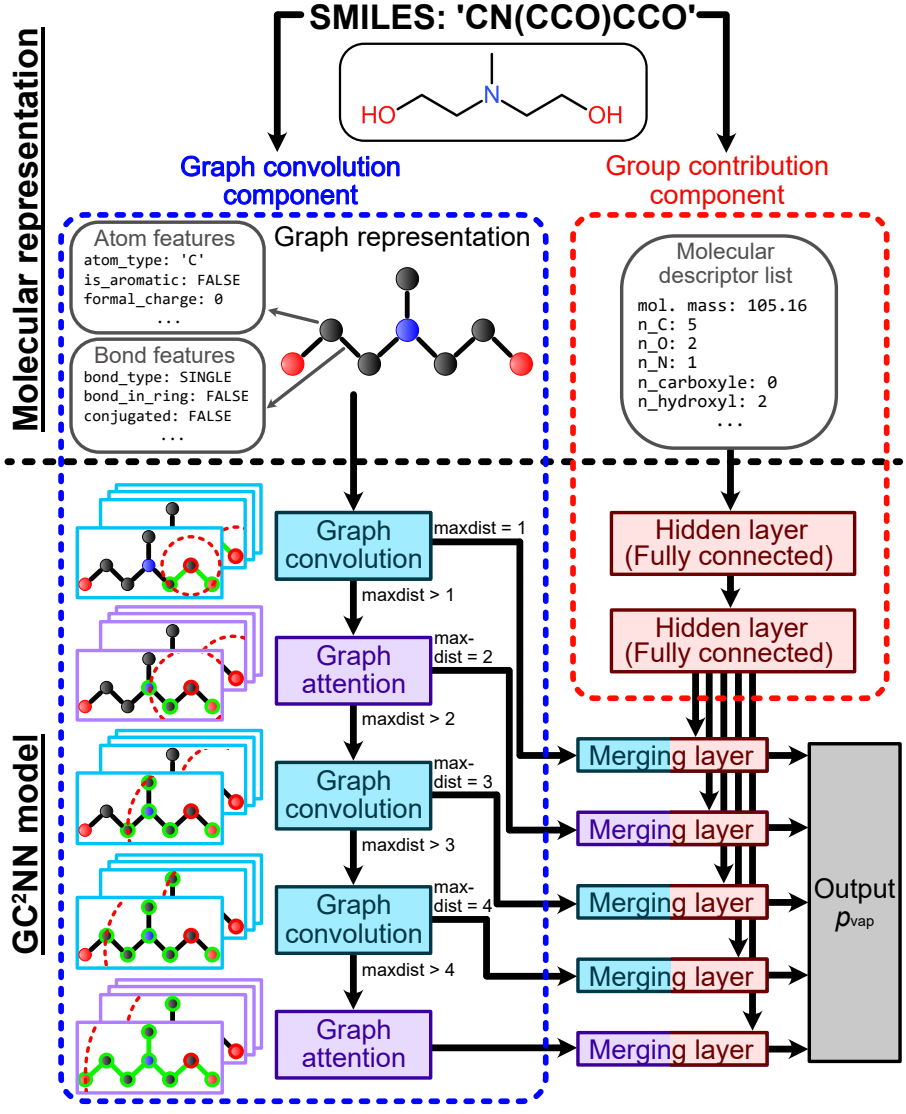

Figure 2. Schematic overview of molecular representation and model functionality in the adaptive-depth GC^{2}NN models. Right: for the group contribution component, Simplified Molecular Input Line Entry System (SMILES) strings are used to derive holistic information on the molecule, such as its molar mass and the presence of atoms and functional groups (Tab. S3). Left: for the model's graph convolution component, SMILES strings are transformed into graph representations, encoded as adjacency matrices, node features, and edge features. This molecular representation is transformed using graph attention and graph convolution layers. The maximum distance ($maxdist$) between two nodes in the input graph determines the number of utilized graph layers, matching the nodes' receptive fields with the respective compound's size. Fully-connected merging layers process information from both model components and map them to the single-node output layer, the p_{vap} prediction.

https://doi.org/10.5194/egusphere-2025-1191
Preprint. Discussion started: 20 March 2025

hyper-parameters, a single model is trained by merging training and validation data to a single training data set, referred to as T+V model. To account for the additional training data, we locally optimize the number of training epochs around the number determined during hyper-parameter tuning. A summary of the relevant hyper-parameters including descriptions and tested ranges is displayed in Tab. S4.

3 Results and Discussion

We train and evaluate group contribution-assisted graph convolutional neural network (GC^{2}NN) models on two sets of experimental vapor pressure (p_{vap}) data and the GeckoQ data set where p_{vap} was derived from quantum-mechanical calculations (Besel et al., 2023). We distinguish between models trained on experimental data sets with different scopes: the *GC^{2}NN-confined* are trained on a confined data set that only contains compounds relevant in the atmosphere within the scope of the methods used for benchmarking, i.e. only containing C, H, and O, and excluding aromatics and some additional functional groups (Fig. 1B,D,F). *GC^{2}NN-broad* are trained on the full experimental data set (Fig. 1A,C,E).

3.1 GC^{2}NN-confined

Figure 3A shows that the adGC^{2}NN model exhibits excellent agreement with the experimental measurements in the independent test set, except from a small number of outliers (MAE = 0.36 log-units). Average training time of the five adGC^{2}NN cross-validation models is 57 minutes on a Nvidia A100, and the average test set mean absolute error (MAE) is 0.39 log-units with a standard deviation of 1.37×10^{-2}. The T+V fdGC^{2}NN performs worse with an MAE of 0.49 log-units. Average training time of the five fdGC^{2}NN cross-validation models is 22 minutes on a Nvidia A100, and the average test set mean absolute error (MAE) is 0.51 log-units with a standard deviation of 1.8×10^{-2}. The selected hyper-parameters for all fdGC^{2}NN models are summarized in Tab. S5. The adGC^{2}NN model is more robust regarding the choice of hyper-parameters, which permits the use of a single model architecture for all data sets (Tab. S6). The adGC^{2}NN significantly outperforms the Krüger et al. (2022) one hot-encoding convolutional neural network approach (OHE-CNN; MAE = 0.66 log-units; average MAE = 0.71 log-units for five cross-validation folds), the Donahue et al. (2011) (MAE = 1.58 log-units) and Li et al. (2016) (MAE = 1.06 log-units) parameterizations, as well as EPI-Suite (MAE = 0.51 log-units), SIMPOL (MAE = 0.55 log-units) and EVAPORATION (MAE = 0.51 log-units) group contribution methods (Fig. 3). Note that the exclusion of a large fraction of molecules (>30 %) from the test set biases the populations of chemical species in the training and test set for the GC^{2}NN and OHE-CNN models (Fig. S2). This may be disadvantageous for the GC^{2}NN models, however, separate calculations with unbiased test set sampling show that the choice of the test set does not have a strong effect on the test set error of the GC^{2}NN models.

Figure 4 shows the distributions of the individual errors for chemical species in the test set for all methods. The fdGC^{2}NN-confined, SIMPOL and EVAPORATION methods exhibit near-identical error distributions where the majority of predictions are very accurate (MAE < 0.5 log-units), and few predictions fall within the range of 0.5 to 1.5 log-units. Only the adGC^{2}NN model has a larger density of very accurate predictions with only few compounds exceeding an MAE of 1.0. Significant outliers (MAE > 1.5 log-units) only occur in the low-volatility range and are predominantly the same compounds across all

https://doi.org/10.5194/egusphere-2025-1191
Preprint. Discussion started: 20 March 2025

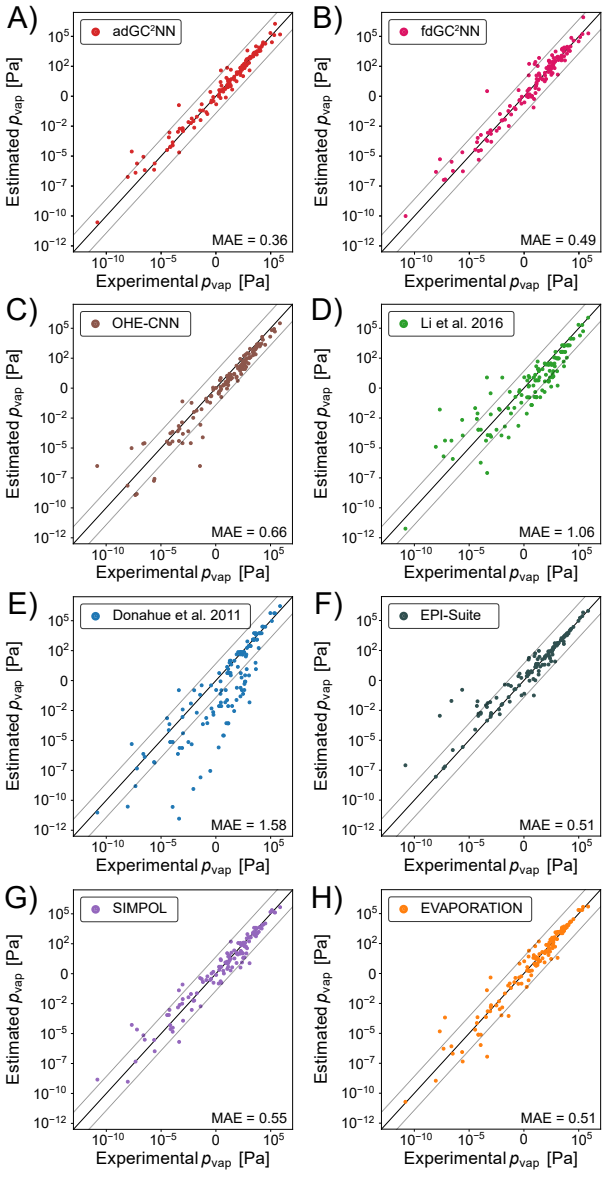

Figure 3. Correlation scatter plots of model-predicted and experimentally-measured vapor pressures for the confined data set. Displayed are data from the independent test set only. The adGC^{2}NN-confined (panel A) and fdGC^{2}NN-confined (panel B) models are compared with established methods: panel C shows the results using a convolutional neural network approach on one-hot encoded SMILES strings following Krüger et al. (2022). (D) Li et al. (2016) and (E) Donahue et al. (2011) are empirical parameterizations, whereas (F) EPI-Suite (EPI), (G) Pankow and Asher (2008) and (H) Compernolle et al. (2011) are group contribution methods. All molecules present in the EVAPORATION training data have been excluded from the test data set. Mean absolute error (MAE) values are in $\log_{10}(p_{vap}$ / [Pa]). The dashed lines ($\pm$1.5 log-units from the 1:1 line) are used to indicate significant outliers.

https://doi.org/10.5194/egusphere-2025-1191
Preprint. Discussion started: 20 March 2025

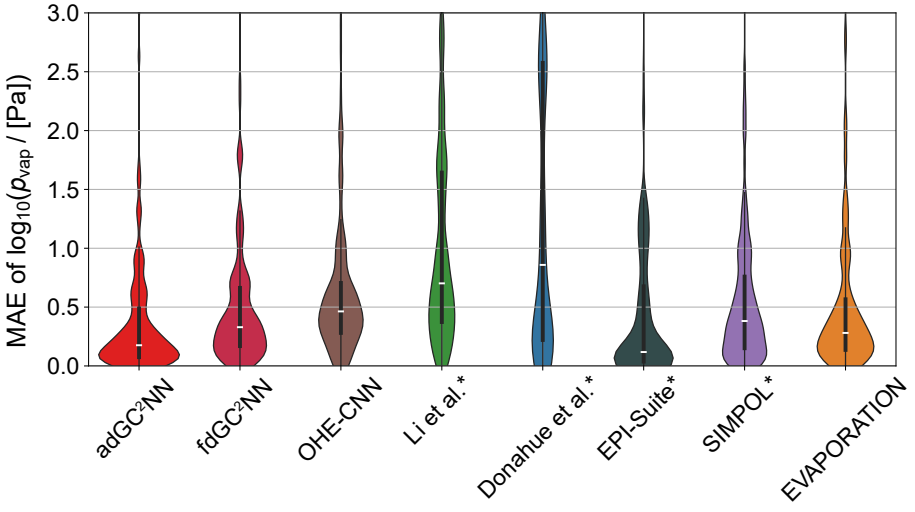

Figure 4. Violin plots representing confined test set error distribution of models shown in Fig. 3. Medians are displayed as white markers, interquartile ranges as vertical wide black lines and 1.5 × interquartile ranges as vertical narrow black lines. Outliers with an MAE > 3 log-units are not shown. Methods marked with an asterisk likely used a fraction of our test data in their training.

four methods, which may be an indicator for experimental measurement errors. EPI-Suite shows an hour-glass shaped profile with a large fraction of very accurate predictions, as well as a large fraction of outliers. Similarly to EPI-Suite, the violin plot for the SIMPOL method has its maximum density at very low error values, even though SIMPOL does not exhibit the overall lowest MAE in our comparison. This is likely due to the presence of EPI-Suite and SIMPOL training data in our test set. Methods for which this is likely the case are marked with an asterisk in Fig. 4. All methods generally perform better at higher p_{vap} (Fig. S3). This behavior correlates with a similar, but weaker bias with regards to molar mass (Fig. S4). The parameterization methods (Li et al., 2016; Donahue et al., 2011) exhibit the highest percentage of significant outliers.

To investigate the effect of experimental error in the low volatility range, we train fdGC^{2}NN models on a subset of the confined data with $\log_{10}(p_{vap}$ / [Pa]) > 0, encompassing only VOC and IVOC, resulting in 1057 compounds. The average test set MAE of the cross-validation folds of this high-volatility fdGC^{2}NN model is 0.32 log-units. This suggests that not only does experimental uncertainty of ELVOC and LVOC lead to model uncertainty in this low-volatility range, but it impedes the accuracy of fdGC^{2}NN models in general.

We use the trained adGC^{2}NN-confined model to review the concept of molecular corridors, following Shiraiwa et al. (2014), where the chemical evolution of molecules constituting SOA is contextualized through their vapor pressure, molar mass, and oxygen-to-carbon (O:C) ratio. The tight inverse correlation between volatility and molar mass mostly holds for the confined test set (Fig. 5A) as well as a data set of atmospherically-relevant compounds from Shiraiwa et al. (2014) (Fig. 5B). For the confined test set, the adGC^{2}NN predictions even tend to fall more strictly into these molecular corridors than the experimental measurements, a potential indicator for experimental uncertainties. When applied to the data from Shiraiwa et al. (2014), we

https://doi.org/10.5194/egusphere-2025-1191
Preprint. Discussion started: 20 March 2025

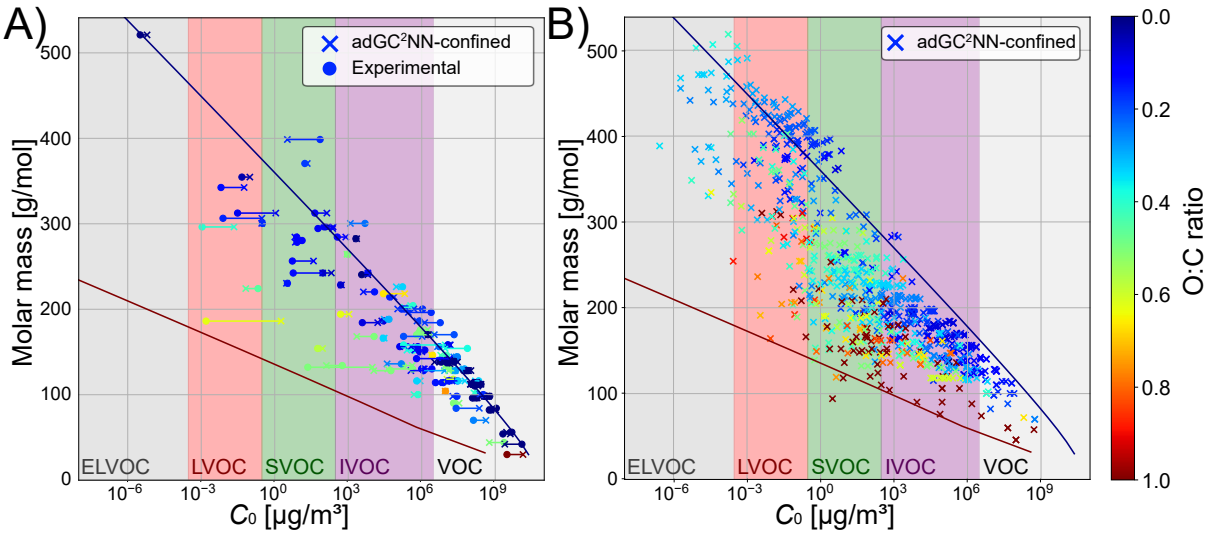

Figure 5. Molecular corridor plots following Shiraiwa et al. (2014). Left: comparison between adGC^{2}NN-confined predictions and experimental measurements in the confined test set. Right: application of the adGC^{2}NN-confined to a data set of atmospherically relevant compounds (Shiraiwa et al., 2014). Blue and red boundary lines correspond to the volatility of n-alkanes and sugar alcohols (as determined by EVAPORATION), respectively.

observe a large number of compounds classified as LVOC by the adGC^{2}NN, that appear to deviate from the molecular corridors, even exceeding upper boundary line corresponding to n-alkanes (O:C = 0). This deviation is either due to a mismatch between the adGC^{2}NN and the EVAPORATION model that was used to determine the boundary lines established in Shiraiwa et al. (2014), or could be due to a systematic error of the adGC^{2}NN as a result of the sparsity of ELVOC data in the training set (Fig. S2B). Furthermore, the difficulties of accurately determining vapor pressures of ELVOC experimentally (Huisman et al., 2013; Bilde et al., 2015) may contribute to this error. In atmospheric context, the accurate determination of ELVOC vapor pressure is not critical with regards to SOA formation, as such compounds condense anyway. Note however, that the accurate determination of ELVOC may be relevant in the context of nucleation, as recent experimental studies found ultra-low-volatility organic compounds (ULVOC) to nucleate, but not LVOC or ELVOC (Kirkby et al., 2023). Attempts have thus been undertaken previously to increase the representation of ELVOC molecules in training data sets for vapor pressure estimation models (Besel et al., 2024).

3.2 GC^{2}NN-broad

Compared to the confined data set, the broad data set encompasses a much larger range of molecular complexity, going far beyond molecules relevant for atmospheric SOA. Thus, and despite a much larger training set size, the adGC^{2}NN-broad model achieves a lower test set accuracy than the adGC^{2}NN-confined model, with an MAE of 0.67 log-units for the T+V model (Fig. 6). Average training time of the cross-validation models is 4.4 hours on a Nvidia A100 GPU, and the average test set mean

https://doi.org/10.5194/egusphere-2025-1191
Preprint. Discussion started: 20 March 2025

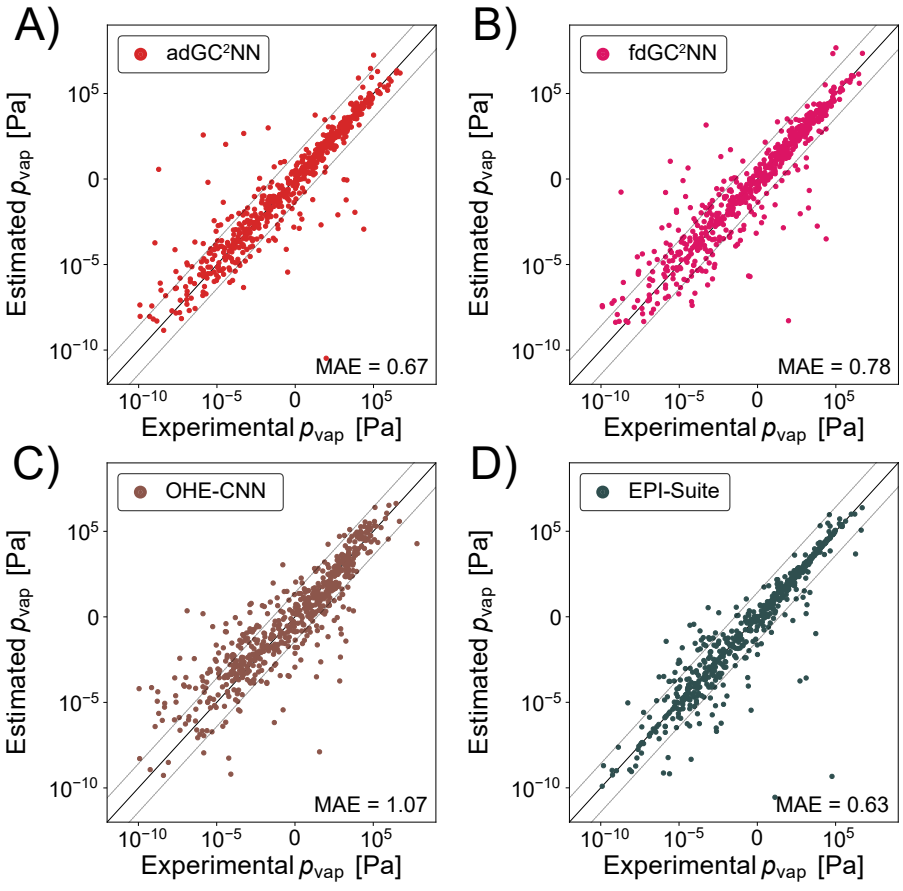

Figure 6. Correlation scatter plots of model-predicted and experimentally-measured vapor pressures for the broad data set. Displayed are data from the independent test set only. (A) adGC^{2}NN-broad model, (B) fdGC^{2}NN-broad model, (C) OHE-CNN method presented in Krüger et al. (2022), and (D) EPI-Suite (EPI). Mean absolute error (MAE) values are in $\log_{10}(p_{vap}$ / [Pa]). The dashed lines ($\pm$1.5 log-units from the 1:1 line) are used to indicate significant outliers.

absolute error (MAE) is 0.68 log-units with a standard deviation of 7.64×10^{-3}. The T+V fdGC^{2}NN model performs worse with an MAE of 0.78 log-units. Cross-validation fdGC^{2}NN models have an average test set MAE of 0.80 with a standard deviation of $1.49\times10^{-}2$ and an average training time of 2.4 hours. Both GC^{2}NN models outperform the OHE-CNN approach from Krüger et al. (2022) (MAE = 1.07 log-units; average MAE = 1.09 log-units for five cross-validation folds), but have a larger test set error than EPI-Suite (EPI) (MAE = 0.63 log-units). Error distributions for the broad test set are displayed in Fig. S5. Note that EPI-Suite was trained on larger data sets that are not publicly available. As discussed above, the MAE that EPI-Suite achieves in our test set is likely biased through overlap of training and test data and thus not fully representative for unknown molecules.

https://doi.org/10.5194/egusphere-2025-1191
Preprint. Discussion started: 20 March 2025

We also train a fdGC^{2}NN model on a subset of the broad data with $\log_{10}(p_{vap}$ / [Pa]) > 0 to investigate the effect of experimental uncertainty in the low-volatility range. Due to the large fraction of low-volatile compounds in the broad data, the high-volatility subset only contains roughly 50% of the original compounds (n = 3116). The cross-validation models achieve an average MAE of 0.37 log-units, greatly reducing the error by more than 50% and outperforming EPI-Suite (Fig. S6, S7). A molecular corridor plot following Shiraiwa et al. (2014) for the adGC^{2}NN-broad model is displayed in Fig. S8, exhibiting a similar bias than the confined model (Fig. 5).

3.3 GC^{2}NN-GeckoQ

In addition to the experimental data sets, we train GC^{2}NN models on the GeckoQ data from Besel et al. (2023), which were derived from quantum-mechanical calculations. For the T+V adGC^{2}NN model, the average test set mean absolute error (MAE) is 0.66 log-units (Fig. 7). The five adGC^{2}NN cross-validation models achieve an MAE of 0.67 log-units, average training time is 13.77 hours on a Nvidia A100. Again, the adGC^{2}NN model achieves a better result than the fdGC^{2}NN model (MSE = 0.71 log-units; average MAE = 0.74 log-units for five cross-validation folds with an average training time of 3.4 hours on a Nvidia A100), as well as the model adapted from Krüger et al. (2022) for p_{vap} prediction (MAE = 0.77 log-units; average MAE = 0.77 log-units for five cross-validation folds). It also outperforms the Gaussian Process Regression model presented in Besel et al. (2023) which achieved a test set MAE of 0.82 log-units.

https://doi.org/10.5194/egusphere-2025-1191
Preprint. Discussion started: 20 March 2025

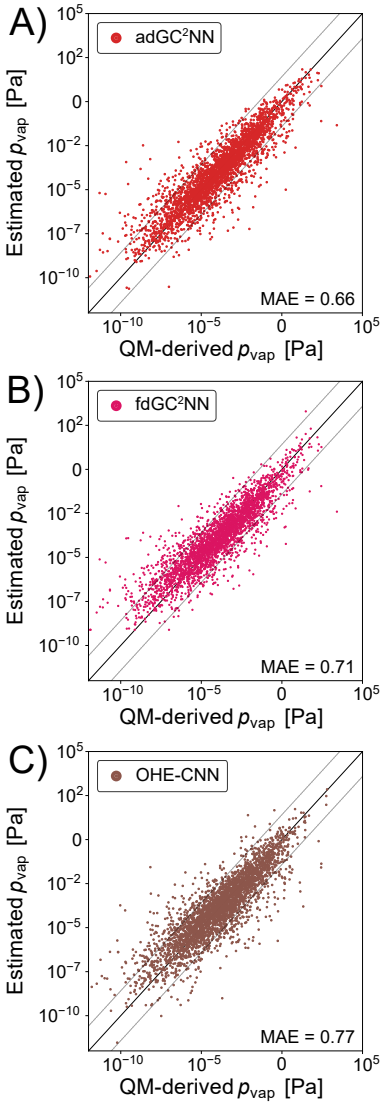

Figure 7. Correlation scatter plots of model-predicted and experimentally-measured vapor pressures for the GeckoQ data set. Displayed are data from the independent test set only. (A) adGC^{2}NN-GeckoQ model (B) fdGC^{2}NN-GeckoQ model, and (C) OHE-CNN method presented in Krüger et al. (2022). Mean absolute error (MAE) values are in $\log_{10}(p_{vap}$ / [Pa]). The dashed lines ($\pm$1.5 log-units from the 1:1 line) are used to indicate significant outliers.

https://doi.org/10.5194/egusphere-2025-1191
Preprint. Discussion started: 20 March 2025

3.4 Learning curves

For each of the three data sets, we obtain fdGC^{2}NN learning curves by training models on subsets of specific sizes. By sampling subsets from the confined, broad and GeckoQ data sets, we obtain three different learning curves that represent models based on different molecular variability and hence different scopes. In general, we observe that significantly more data are needed to achieve the same accuracy, if the data contain a larger variety of compound classes as found in the broad data set (Fig. 8). Gradients and convergence rates of the learning curves significantly differ between the data sets. The fdGC^{2}NN-confined models exhibit the steepest learning curve, possibly due to the limited variability of elements and molecular substructures. While the lack of distinct molecular features may impede models trained on few data, the same simplicity seems to permit good model performance with the full training data. In the broad and GeckoQ data, the high variability of molecular features and, potentially, their complex interactions require much more data for accurate predictions. None of the learning curves appear to fully level-off for large data set sizes, which means that the models can be expected to improve significantly with additional training data.

Note that we chose to display learning curves for fdGC^{2}NN models only because for adGC^{2}NN models, model performance in a test data set depends more strongly on the distribution of molecule sizes in the training data set. While all compounds are passed through the first graph layer, later layers are frequented less, which leads to a large variability of model errors for the smallest training data set sizes.

In addition to the fdGC^{2}NN models, we tested graph-only models without the additional input layer to obtain holistic molecular information (group-contribution component). These pure GCNN models are associated with significantly larger errors for all data sets and sizes (Fig. S9). This can be attributed to graph convolutions which, in principle, are merely a succession of local operations on subgraphs. In a GC^{2}NN, each additional convolution layer increases the distance allowed for two nodes (and hence atoms) to influence each other. Setting the number of graph convolution layers to the largest distance between two nodes in the data set would enable the model to derive information from each molecule as a whole. However, this is detrimental for most model training because it would result in very deep neural networks which would likely over-fit on most data sets. Therefore, since the graph neural network training might not effectively capture whole-molecule properties, the lack of information on general molecular properties, like molar mass, inhibits the graph-only models to generalize between molecules of different size. We observe that the addition of molar mass as an input is crucial for the performance of GC^{2}NN, while additional descriptors like element and functional group counts lead to further improvements.

4 Conclusions

Our findings suggest that group contribution-assisted graph convolutional neural networks (GC^{2}NN) and graph representations of molecules are a promising approach for quantitative structure-activity relationship (QSAR) models. Despite the challenging scarcity of experimental data available for atmospherically relevant compounds, the GC^{2}NN models surpass established methods, including parameterizations, group contribution methods, and machine learning (ML) approaches. Graph representations are a natural and unambiguous representation of molecular structures, encoding additional information related to individual

https://doi.org/10.5194/egusphere-2025-1191
Preprint. Discussion started: 20 March 2025

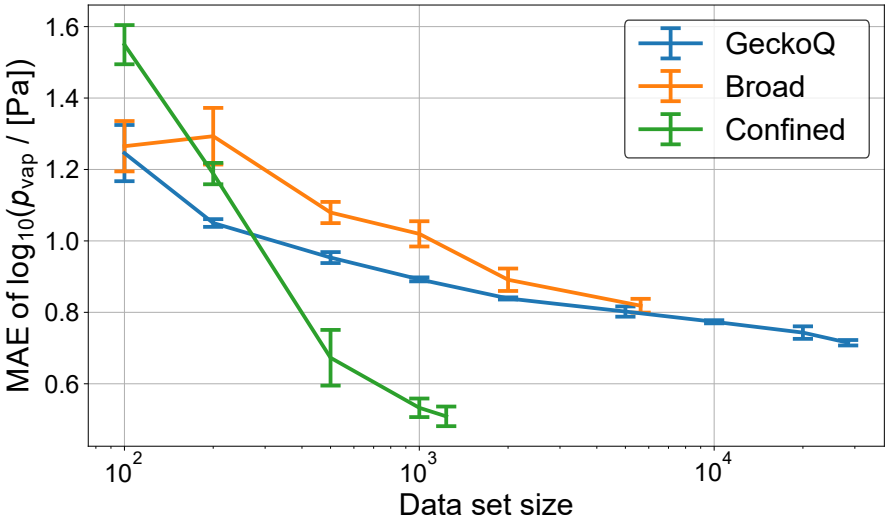

Figure 8. Mean absolute error (MAE) for independent test sets (confined: n = 137; broad: n = 625; GeckoQ: n = 3,163), as a function of training data set size of fdGC^{2}NN models trained on subsets of the three data sets. The experiment is performed by sampling subsets of various size from each of the respective data sets and training fdGC^{2}NN models on these. Hyper-parameter tuning is performed for each subset. Shown are the average test set log unit MAE of five cross-validation models in each subset. Error bars represent standard deviations across the cross-validation folds.

atoms (graph nodes) or bonds (graph edges), and making spatial relations between molecular substructures directly interpretable. With that, graph representations are advantageous over molecular representations in which spatial information are lost or not easily retrievable, such as one-hot encoded (OHE) SMILES strings, which we used previously in conjunction with convolutional neural networks (CNN) for the determination of quinone redox potentials Krüger et al. (2022). In this study, OHE-CNN models performed worse than GC^{2}NN models for every tested data set. Note, however, that we only performed a very basic tuning of the hyperparameters from the original study and correlation of the OHE-CNN model may improve with more extensive optimization.

We find that models that combine graph convolution with the direct interpretation of molecular properties like molar mass, element, and functional group occurrences outperform models that only process one of the two. The accuracy of graph-only GCNN models, without the additional input layer, falls behind pure group contribution models that process information on functional groups under consideration of known principles governing their effect on molecular properties. The provision of holistic information on the molecular structure, especially molar mass, is crucial for the performance of GC^{2}NN models, as graph convolutions only process structural information locally. The difficulty in the application of graph convolutional neural networks is their dependence on the size of the input graphs. Therefore, specialized fdGC^{2}NN models for narrow vapor pressure ranges achieved excellent results, given sufficient training data, in this study. Our adaptive-depth approach, however, enables

https://doi.org/10.5194/egusphere-2025-1191
Preprint. Discussion started: 20 March 2025

the GC^{2}NN to make use of the full training data, while matching the individual nodes' receptive fields with the compound size dynamically.

In general, the application of machine learning with few data is challenging, and learning curves suggest that additional data would significantly improve model accuracy for all compound ranges. We hypothesize that ML QSAR models may furthermore improve through prediction of multiple related molecular properties at a time. For instance, vapor pressure-predicting models may benefit from the simultaneous prediction of melting points or glass transition temperature, as the addition of such properties in the training data possibly makes physical principles more accessible by the model. Additional molecular parameters that are known to affect vapor pressure, such as polarity and representations of secondary intermolecular bonding, might also increase prediction performances with a similar architecture in the future. However, this may pose further restrictions on the training data available while highlighting how the application of machine learning methods in atmospheric chemistry is currently limited by the scarcity of comprehensive experimental data sets involving atmospheric compounds. Furthermore, the multiple component approach to QSAR modelling permits the utilization of far more advanced group contribution components alongside the graph convolution component. While the shallow neural networks in our study can indiscriminately be applied to various molecular descriptors and data sets, the utilization of advanced group contribution methods like SIMPOL or EVAPORATION alongside the graph convolution component, or the utilization of additional molecular descriptors may significantly increase model accuracy.

By using data sets of differing molecular complexity, a broad data set using most web-crawled data and a data set confined for atmospherically-relevant compounds, we find that the more specialized model can achieve a higher test set accuracy. In turn, while the models training on the broad data set have the largest error of all GC^{2}NN models in this study, they are applicable to a large population of compounds with a diverse elemental composition and variety of functional groups, encompassing both organic and inorganic species. It is therefore recommended to train QSAR models that are specific to certain molecule scopes and applications. We also find that model accuracy significantly differs between models that are trained on subsets of the p_{vap} range, and that models that are trained on smaller ranges can outperform more general models despite training data scarcity. In practice, an ensemble approach with multiple models, e.g., specifically for the low and high volatility range may be a viable approach for ML methods, similarly to the ensemble utilization of the Modified Grain, Antoine and Mackay methods (EPI; Li et al., 2016). Further improvements may be achievable through data curation techniques, as common outliers between various methods indicate data points with large experimental uncertainty.

Code and data availability. The data and source code, as well as a model executable are openly available at:
https://doi.org/10.17617/3.GIKHJL

Author contributions. MS and TG conceived the study. All authors designed research. MK and TG wrote the code and performed model simulations. All authors discussed and interpreted calculation results. MK and TB wrote the manuscript with contributions from all authors.

https://doi.org/10.5194/egusphere-2025-1191
Preprint. Discussion started: 20 March 2025

Competing interests. The authors declare that they have no competing interests.

Acknowledgements. We thank Steven Compernolle for providing the list of molecules contained in EVAPORATION training data in machine-readable format. We thank Nadin Ulrich for helpful discussions. MS thanks the U.S. Department of Energy (DE-SC0022139) and the U. S. National Science Foundation (AGS-2246502) for funding. This work was funded by the Max Planck Society (MPG). MK is supported by the Max Planck Graduate Center with the Johannes Gutenberg University Mainz (MPGC).

Supplementary information The online version contains supplementary material:

Additional file 1: Table S1. Summary of atom features in the graph representation. **Table S2.** Summary of bond features in the graph representation. **Table S3.** Summary of molecular descriptors passed to the group contribution component of GC^{2}NN models. **Table S4.** GC^{2}NN hyper-parameter description. **Table S5.** Selected hyper-parameters of fdGC^{2}NN models. **Table S6.** Selected hyper-parameters of adGC^{2}NN models. **Figure S1.** Schematic overview of molecular representation and model functionality of fdGC^{2}NN models. **Figure S2.** Information on confined training and test data. **Figure S3.** Confined test set errors for volatility bins. **Figure S4.** Confined test set errors for molecular mass bins. **Figure S5.** Violinplots for broad test set errors. **Figure S6.** Broad test set errors for volatility bins. **Figure S7.** Broad test set errors for molecular mass bins. **Figure S8.** Molecular corridor plot for adGC^{2}NN-broad. **Figure S9.** Learning curve for graph-only GCNN models.

https://doi.org/10.5194/egusphere-2025-1191
Preprint. Discussion started: 20 March 2025

References

EPI Suite™-Estimation Program Interface, https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface, accessed: 2024-10-13.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M.: Optuna: A Next-generation Hyperparameter Optimization Framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2623–2631, ACM, Anchorage AK USA, ISBN 978-1-4503-6201-6, https://doi.org/10.1145/3292500.3330701, 2019.

Armeli, G., Peters, J.-H., and Koop, T.: Machine-Learning-Based Prediction of the Glass Transition Temperature of Organic Compounds Using Experimental Data, ACS Omega, 8, 12 298–12 309, https://doi.org/10.1021/acsomega.2c08146, 2023.

Atz, K., Grisoni, F., and Schneider, G.: Geometric deep learning on molecular representations, Nature Machine Intelligence, 3, 1023–1032, https://doi.org/10.1038/s42256-021-00418-8, 2021.

Aumont, B., Szopa, S., and Madronich, S.: Modelling the evolution of organic carbon during its gas-phase tropospheric oxidation: development of an explicit model based on a self generating approach, Atmos. Chem. Phys., 5, 2497–2517, https://doi.org/10.5194/acp-5-2497-2005, 2005.

Besel, V., Todorović, M., Kurtén, T., Rinke, P., and Vehkamäki, H.: Atomic structures, conformers and thermodynamic properties of 32k atmospheric molecules, Sci Data, 10, 450, https://doi.org/10.1038/s41597-023-02366-x, 2023.

Besel, V., Todorović, M., Kurtén, T., Vehkamäki, H., and Rinke, P.: The search for sparse data in molecular datasets: Application of active learning to identify extremely low volatile organic compounds, Journal of Aerosol Science, 179, 106 375, https://doi.org/10.1016/j.jaerosci.2024.106375, 2024.

Bilde, M., Barsanti, K., Booth, M., Cappa, C. D., Donahue, N. M., Emanuelsson, E. U., McFiggans, G., Krieger, U. K., Marcolli, C., Topping, D., Ziemann, P., Barley, M., Clegg, S., Dennis-Smither, B., Hallquist, M., Hallquist, A. M., Khlystov, A., Kulmala, M., Mogensen, D., Percival, C. J., Pope, F., Reid, J. P., Ribeiro da Silva, M. A. V., Rosenoern, T., Salo, K., Soonsin, V. P., Yli-Juuti, T., Prisle, N. L., Pagels, J., Rarey, J., Zardini, A. A., and Riipinen, I.: Saturation Vapor Pressures and Transition Enthalpies of Low-Volatility Organic Molecules of Atmospheric Relevance: From Dicarboxylic Acids to Complex Mixtures, Chem. Rev., 115, 4115–4156, https://doi.org/10.1021/cr5005502, 2015.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P.: Geometric Deep Learning: Going beyond Euclidean data, IEEE Signal Processing Magazine, 34, 18–42, https://doi.org/10.1109/MSP.2017.2693418, 2017.

Compernolle, S., Ceulemans, K., and Müller, J.-F.: EVAPORATION: a new vapour pressure estimation methodfor organic molecules including non-additivity and intramolecular interactions, Atmos. Chem. Phys., 11, 9431–9450, https://doi.org/10.5194/acp-11-9431-2011, 2011.

Donahue, N. M., Robinson, A. L., and Pandis, S. N.: Atmospheric organic particulate matter: From smoke to secondary organic aerosol, Atmospheric Environment, 43, 94–106, https://doi.org/10.1016/j.atmosenv.2008.09.055, 2009.

Donahue, N. M., Epstein, S. A., Pandis, S. N., and Robinson, A. L.: A two-dimensional volatility basis set: 1. organic-aerosol mixing thermodynamics, Atmos. Chem. Phys., 11, 3303–3318, https://doi.org/10.5194/acp-11-3303-2011, 2011.

Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., and Adams, R. P.: Convolutional Networks on Graphs for Learning Molecular Fingerprints, https://arxiv.org/abs/1509.09292, 2015.

Gadaleta, D., Lombardo, A., Toma, C., and Benfenati, E.: A new semi-automated workflow for chemical data retrieval and quality checking for modeling applications, J Cheminform, 10, 60, https://doi.org/10.1186/s13321-018-0315-6, 2018.

https://doi.org/10.5194/egusphere-2025-1191
Preprint. Discussion started: 20 March 2025

Galeazzo, T. and Shiraiwa, M.: Predicting glass transition temperature and melting point of organic compounds *via* machine learning and molecular embeddings, Environ. Sci.: Atmos., 2, 362–374, https://doi.org/10.1039/D1EA00090J, 2022.

Gani, R.: Group contribution-based property estimation methods: advances and perspectives, Current Opinion in Chemical Engineering, 23, 184–196, https://doi.org/10.1016/j.coche.2019.04.007, 2019.

Geerlings, P., De Proft, F., and Langenaeker, W.: Conceptual Density Functional Theory, Chem. Rev., 103, 1793–1874, https://doi.org/10.1021/cr990029p, 2003.

Goldstein, A. H. and Galbally, I. E.: Known and unexplored organic constituents in the earth's atmosphere, Environmental science & technology, 41, 1514–1521, 2007.

Huisman, A. J., Krieger, U. K., Zuend, A., Marcolli, C., and Peter, T.: Vapor pressures of substituted polycarboxylic acids are much lower than previously reported, Atmos. Chem. Phys., 13, 6647–6662, https://doi.org/10.5194/acp-13-6647-2013, 2013.

Hyttinen, N., Pullinen, I., Nissinen, A., Schobesberger, S., Virtanen, A., and Yli-Juuti, T.: Comparison of saturation vapor pressures of α-pinene + O_3 oxidation products derived from COSMO-RS computations and thermal desorption experiments, Atmos. Chem. Phys., 22, 1195–1208, https://doi.org/10.5194/acp-22-1195-2022, 2022.

Intergovernmental Panel on Climate Change: Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge University Press, 1 edn., ISBN 978-1-00-915789-6, https://doi.org/10.1017/9781009157896, 2023.

Isaacman-VanWertz, G. and Aumont, B.: Impact of organic molecular structure on the estimation of atmospherically relevant physicochemical parameters, Atmos. Chem. Phys., 21, 6541–6563, https://doi.org/10.5194/acp-21-6541-2021, 2021.

Jimenez, J. L., Canagaratna, M. R., Donahue, N. M., Prevot, A. S. H., Zhang, Q., Kroll, J. H., DeCarlo, P. F., Allan, J. D., Coe, H., Ng, N. L., Aiken, A. C., Docherty, K. S., Ulbrich, I. M., Grieshop, A. P., Robinson, A. L., Duplissy, J., Smith, J. D., Wilson, K. R., Lanz, V. A., Hueglin, C., Sun, Y. L., Tian, J., Laaksonen, A., Raatikainen, T., Rautiainen, J., Vaattovaara, P., Ehn, M., Kulmala, M., Tomlinson, J. M., Collins, D. R., Cubison, M. J., E., Dunlea, J., Huffman, J. A., Onasch, T. B., Alfarra, M. R., Williams, P. I., Bower, K., Kondo, Y., Schneider, J., Drewnick, F., Borrmann, S., Weimer, S., Demerjian, K., Salcedo, D., Cottrell, L., Griffin, R., Takami, A., Miyoshi, T., Hatakeyama, S., Shimono, A., Sun, J. Y., Zhang, Y. M., Dzepina, K., Kimmel, J. R., Sueper, D., Jayne, J. T., Herndon, S. C., Trimborn, A. M., Williams, L. R., Wood, E. C., Middlebrook, A. M., Kolb, C. E., Baltensperger, U., and Worsnop, D. R.: Evolution of Organic Aerosols in the Atmosphere, Science, 326, 1525–1529, https://doi.org/10.1126/science.1180353, 2009.

Kanakidou, M., Seinfeld, J. H., Pandis, S. N., Barnes, I., Dentener, F. J., Facchini, M. C., Van Dingenen, R., Ervens, B., Nenes, A., Nielsen, C. J., Swietlicki, E., Putaud, J. P., Balkanski, Y., Fuzzi, S., Horth, J., Moortgat, G. K., Winterhalter, R., Myhre, C. E. L., Tsigaridis, K., Vignati, E., Stephanou, E. G., and Wilson, J.: Organic aerosol and global climate modelling: a review, Atmos. Chem. Phys., 5, 1053–1123, https://doi.org/10.5194/acp-5-1053-2005, 2005.

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., et al.: PubChem substance and compound databases, Nucleic acids research, 44, D1202–D1213, 2016.

Kipf, T. N. and Welling, M.: Semi-Supervised Classification with Graph Convolutional Networks, https://arxiv.org/abs/1609.02907, 2017.

Kirkby, J., Amorim, A., Baltensperger, U., Carslaw, K. S., Christoudias, T., Curtius, J., Donahue, N. M., Haddad, I. E., Flagan, R. C., Gordon, H., Hansel, A., Harder, H., Junninen, H., Kulmala, M., Kürten, A., Laaksonen, A., Lehtipalo, K., Lelieveld, J., Möhler, O., Riipinen, I., Stratmann, F., Tomé, A., Virtanen, A., Volkamer, R., Winkler, P. M., and Worsnop, D. R.: Atmospheric new particle formation from the CERN CLOUD experiment, Nat. Geosci., 16, 948–957, https://doi.org/10.1038/s41561-023-01305-0, 2023.

https://doi.org/10.5194/egusphere-2025-1191
Preprint. Discussion started: 20 March 2025

Krüger, M., Wilson, J., Wietzoreck, M., Bandowe, B. A. M., Lammel, G., Schmidt, B., Pöschl, U., and Berkemeier, T.: Convolutional neural network prediction of molecular properties for aerosol chemistry and health effects, Natural Sciences, 2, e20220016, https://doi.org/10.1002/ntls.20220016, 2022.

Kulmala, M. and Kerminen, V.-M.: On the formation and growth of atmospheric nanoparticles, Atmospheric Research, 90, 132–150, https://doi.org/10.1016/j.atmosres.2008.01.005, 2008.

Kurtén, T., Hyttinen, N., D'Ambro, E. L., Thornton, J., and Prisle, N. L.: Estimating the saturation vapor pressures of isoprene oxidation products $C_5H_{12}O_6$ and $C_5H_{10}O_6$ using COSMO-RS, Atmos. Chem. Phys., 18, 17 589–17 600, https://doi.org/10.5194/acp-18-17589-2018, 2018.

Landrum, G.: RDKit: Open-source cheminformatics, Release, 1, 4, https://www.rdkit.org, 2013.

Li, Y., Pöschl, U., and Shiraiwa, M.: Molecular corridors and parameterizations of volatility in the chemical evolution of organic aerosols, Atmos. Chem. Phys., 16, 3327–3344, https://doi.org/10.5194/acp-16-3327-2016, 2016.

Longnecker, E. R., Bakker-Arkema, J. G., and Ziemann, P. J.: Comparison of Vapor Pressure Estimation Methods Used to Model Secondary Organic Aerosol Formation from Reactions of Linear and Branched Alkenes with OH/NO_x, ACS Earth Space Chem., p. acsearthspacechem.4c00285, https://doi.org/10.1021/acsearthspacechem.4c00285, 2025.

Lumiaro, E., Todorović, M., Kurten, T., Vehkamäki, H., and Rinke, P.: Predicting gas–particle partitioning coefficients of atmospheric molecules with machine learning, Atmos. Chem. Phys., 21, 13 227–13 246, https://doi.org/10.5194/acp-21-13227-2021, 2021.

Moller, B., Rarey, J., and Ramjugernath, D.: Estimation of the vapour pressure of non-electrolyte organic compounds via group contributions and group interactions, Journal of Molecular Liquids, 143, 52–63, https://doi.org/10.1016/j.molliq.2008.04.020, 2008.

Naef, R. and Acree, W. E.: Calculation of the Vapour Pressure of Organic Molecules by Means of a Group-Additivity Method and Their Resultant Gibbs Free Energy and Entropy of Vaporization at 298.15 K, Molecules, 26, 1045, https://doi.org/10.3390/molecules26041045, 2021.

Nannoolal, Y., Rarey, J., Ramjugernath, D., and Cordes, W.: Estimation of pure component properties, Fluid Phase Equilibria, 226, 45–63, https://doi.org/10.1016/j.fluid.2004.09.001, 2004.

Pankow, J. F.: Review and comparative analysis of the theories on partitioning between the gas and aerosol particulate phases in the atmosphere, Atmospheric Environment (1967), 21, 2275–2283, https://doi.org/10.1016/0004-6981(87)90363-5, 1987.

Pankow, J. F. and Asher, W. E.: SIMPOL.1: a simple group contribution method for predicting vapor pressures and enthalpies of vaporization of multifunctional organic compounds, Atmos. Chem. Phys., 8, 2773–2796, https://doi.org/10.5194/acp-8-2773-2008, 2008.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library, Advances in neural information processing systems, 32, 2019.

Pöschl, U. and Shiraiwa, M.: Multiphase Chemistry at the Atmosphere–Biosphere Interface Influencing Climate and Public Health in the Anthropocene, Chem. Rev., 115, 4440–4475, https://doi.org/10.1021/cr500487s, 2015.

Ratcliff, L. E., Mohr, S., Huhs, G., Deutsch, T., Masella, M., and Genovese, L.: Challenges in large scale quantum mechanical calculations, WIREs Comput Mol Sci, 7, e1290, https://doi.org/10.1002/wcms.1290, 2017.

Reiser, P., Neubert, M., Eberhard, A., Torresi, L., Zhou, C., Shao, C., Metni, H., van Hoesel, C., Schopmans, H., Sommer, T., and Friederich, P.: Graph neural networks for materials science and chemistry., Communications Materials, 3, 241 722, 10.1038/s43246-022-00315-6, 2022.

https://doi.org/10.5194/egusphere-2025-1191
Preprint. Discussion started: 20 March 2025

Shiraiwa, M., Berkemeier, T., Schilling-Fahnestock, K. A., Seinfeld, J. H., and Pöschl, U.: Molecular corridors and kinetic regimes in the multiphase chemical evolution of secondary organic aerosol, Atmos. Chem. Phys., 14, 8323–8341, https://doi.org/10.5194/acp-14-8323-2014, 2014.

Shrivastava, M., Cappa, C. D., Fan, J., Goldstein, A. H., Guenther, A. B., Jimenez, J. L., Kuang, C., Laskin, A., Martin, S. T., Ng, N. L., Petaja, T., Pierce, J. R., Rasch, P. J., Roldin, P., Seinfeld, J. H., Shilling, J., Smith, J. N., Thornton, J. A., Volkamer, R., Wang, J., Worsnop, D. R., Zaveri, R. A., Zelenyuk, A., and Zhang, Q.: Recent advances in understanding secondary organic aerosol: Implications for global climate forcing, Reviews of Geophysics, 55, 509–559, https://doi.org/10.1002/2016RG000540, 2017.

Tabor, D. P., Gómez-Bombarelli, R., Tong, L., Gordon, R. G., Aziz, M. J., and Aspuru-Guzik, A.: Mapping the frontiers of quinone stability in aqueous media: implications for organic aqueous redox flow batteries, J. Mater. Chem. A, 7, 12 833–12 841, https://doi.org/10.1039/C9TA03219C, 2019.

Tahami, S., Movagharnejad, K., and Ghasemitabar, H.: Estimation of the critical constants of organic compounds via a new group contribution method, Fluid Phase Equilibria, 494, 45–60, https://doi.org/10.1016/j.fluid.2019.04.022, 2019.

Tang, B., Kramer, S. T., Fang, M., Qiu, Y., Wu, Z., and Xu, D.: A self-attention based message passing neural network for predicting molecular lipophilicity and aqueous solubility, Journal of cheminformatics, 12, 1–9, 2020.

Ulrich, N., Goss, K.-U., and Ebert, A.: Exploring the octanol–water partition coefficient dataset using deep learning techniques and data augmentation, Commun Chem, 4, 90, https://doi.org/10.1038/s42004-021-00528-9, 2021.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., and Bengio, Y.: Graph Attention Networks, https://doi.org/10.48550/ARXIV.1710.10903, version Number: 3, 2017.

von Lilienfeld, O. A. and Burke, K.: Retrospective on a decade of machine learning for chemical discovery, Nature Communications, 11, https://api.semanticscholar.org/CorpusID:222163935, 2020.

Wang, C., Yuan, T., Wood, S. A., Goss, K.-U., Li, J., Ying, Q., and Wania, F.: Uncertain Henry's law constants compromise equilibrium partitioning calculations of atmospheric oxidation products, Atmos. Chem. Phys., 17, 7529–7540, https://doi.org/10.5194/acp-17-7529-2017, 2017.

Withnall, M., Lindelöf, E., Engkvist, O., and Chen, H.: Building attention and edge message passing neural networks for bioactivity and physical–chemical property prediction, Journal of cheminformatics, 12, 1, 2020.

Xiong, G., Wu, Z., Yi, J., Fu, L., Yang, Z., Hsieh, C., Yin, M., Zeng, X., Wu, C., Lu, A., Chen, X., Hou, T., and Cao, D.: ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties, Nucleic Acids Research, 49, W5–W14, https://doi.org/10.1093/nar/gkab255, 2021.

Zhang, S., Tong, H., Xu, J., and Maciejewski, R.: Graph convolutional networks: a comprehensive review, Comput Soc Netw, 6, 11, https://doi.org/10.1186/s40649-019-0069-y, 2019.