Supplement of Geosci. Model Dev., 18, 7357–7371, 2025
https://doi.org/10.5194/gmd-18-7357-2025-supplement

Geoscientific Model Development
Open Access
EGU

Supplement of

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

Matteo Krüger et al.

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

Table S1. Atom features represented in the feature map linked to individual nodes of the graph representation. To obtain the required features from SMILES strings, we use the Python package RDKit (Landrum, 2013).

Feature name	Encoding	Possible values	Description
atom_type	OHE[1]	C, O, N, H, Cl, P, S, F, I, B, Br, Si	Element
n_heavy_neighbors	OHE[1]	0, 1, 2, 3, 4, MoreThanFour	Atom neighbors that are not H
formal_charge[2]	OHE[1]	-3, -2, -1, 0, 1, 2, 3, Extreme	Formal charge of atom
hybridisation_type	OHE[1]	S, SP, SP2, SP3, SP3D, SP3D2, OTHER	Atom hybridisation
is_in_a_ring	BOOL	0, 1	If atom is within ring structure
is_aromatic[2]	BOOL	0, 1	If atom is within conjugated structure
atomic_mass	FLOAT	-	Atomic mass in [u], scaled
vdw_radius	FLOAT	-	Van-der-Waals radius, scaled
covalent_radius	FLOAT	-	Covalent radius, scaled
chirality_type	OHE[1]	CHI_UNSPECIFIED, CHI_TETRAHEDRAL_CW, CHI_TETRAHEDRAL_CCW, CHI_OTHER	Chirality type
n_hydrogens	OHE[1]	0, 1, 2, 3, 4, MoreThanFour	Atom neighbors that are H

[1] One-hot-encoding [2] Omitted in confined data

Table S2. Bond features represented in the feature map linked to individual edges of the graph representation. To obtain the required features from SMILES strings, we use the Python package RDKit (Landrum, 2013).

Feature name	Encoding	Possible values	Description
bond_type	OHE[1]	SINGLE, DOUBLE, TRIPLE, AROMATIC	Type of bond
bond_is_in_ring	BOOL	0, 1	If bond is within ring structure
bond_is_conj	BOOL	0, 1	If bond is conjugated
stereo_type	OHE[1]	Z, E, ANY, NONE	Stereo type of bond

[1] One-hot-encoding

Table S3: List of molecular descriptors passed to the group contribution component of GC[2]NN models. To obtain the required molecular descriptors from SMILES strings, we use the Python package RDKit (Landrum, 2013).

Feature	Description	Present in model
mass	Molar mass	all
NumAtoms	Number of atoms	all
NumBonds	Number of bonds	all
NumSingleBonds[1]	Number of single bonds	all
NumDoubleBonds[1]	Number of double bonds	all
NumTripleBonds[1]	Number of triple bonds	all
NumAromBonds[1]	Number of aromatic bonds	all
AromC	Number of aromatic carbon atoms	all
Charge	Formal charge	all
BertzCT	Bertz complexity index	all
Ipc	Structural information content	all
NumHDonors	Number of hydrogen donors	all
TPSA	Topological polar surface area	all
NHOHCount	Number of -NH and -OH groups	all
MolMR	Molar refractivity	all
VSA_EState3	EState indices for 3rd bin of VSA	all
AvgIpc	Average information content per atom	all
OC-ratio	Oxygen-carbon ratio	all
C^2	Carbon atoms	all
O^2	Oxygen atoms	all
N^2	Nitrogen atoms	all
Cl^2	Chlorine atoms	broad
I^2	Iodine atoms	broad
S^2	Sulfur atoms	broad
F^2	Fluorine atoms	broad
P^2	Phosphorus atoms	broad
Si^2	Silicon atoms	broad
Br^2	Bromine atoms	broad
B^2	Boron atoms	broad
hydroxyl	Hydroxyl groups	all

Table S3: (continued)

Feature	Description	Present in model
ester	Ester groups	all
carbonyl	Carbonyl groups	all
carboxyle	Carboxyl groups	confined, broad
ketone	Ketone groups	GeckoQ
hydroperoxide	Hydroperoxide groups	GeckoQ
nitrate	Nitrate groups	GeckoQ
aldehyde	Aldehyde groups	GeckoQ
carbonic acid	Carbonic acid groups	GeckoQ
peroxide	Peroxide groups	GeckoQ
carbonylperoxynitrate	Carbonylperoxynitrate groups	GeckoQ
ether	Ether groups	GeckoQ
nitro	Nitro groups	broad, GeckoQ
nitroester	Nitroester groups	GeckoQ
amine	Amine groups	broad
amide	Amide groups	broad
sulfide	Sulfide groups	broad
nitrile	Nitrile groups	broad

[1] Normalized as fraction of all bonds in compound [2] Normalized as fraction of all atoms in compound

Table S4. GC^2NN hyperparameter description and tested ranges.

Hyperparameter	Description	Tested range
num_conv_layers	Number of graph conv. layers	[2, 8]
num_conv_nodes	Number of nodes in each conv. layer	[8, 128]
num_hidden_layers	Number of additional fully-connected hidden layers	[0, 2]
hidden_layer_nodes	Number of nodes in each additional fully-connected layer	[8, 128]
num_merging_layers	Number of merging layers	[0, 2]
merging_layer_nodes	Number of nodes in each merging layer	[8, 128]
learning_rate	Learning rate during training	$[1 \times 10^{-4}, 1 \times 10^{-2}]$
lr_decay	Learning rate decay in each training epoch	[0.97, 1.0]
weight_decay	L2 regularization to avoid large weights	0 or $[1 \times 10^{-5}, 1 \times 10^{-2}]$
acivations	Activation of each conv. layer	'ReLU', 'LeakyReLU', 'Tanh' or 'Sigmoid'
layer_types	Types of conv. layers	'GCN'[1] or 'GAT'[2]
heads	Number of attention heads in conv. layer[3]	[1, 8]
pass_edge_attr	If edge (bond) attributes are passed to a layer[3]	0 or 1
batch_size	Number of molecules in each training batch	[4, 32]

[1] Graph convolution layers (Zhang et al., 2019) [2] Graph attention layers (Veličković et al., 2017) [3] Only applicable to GAT layers

Table S5. Selected hyperparameters for $fdGC^2NN$ models.

Hyperparameter	$fdGC^2NN$-confined	$fdGC^2NN$-broad	$fdGC^2NN$-GeckoQ
num_conv_layers	5	5	5
num_conv_nodes	[32, 64, 32, 32, 32]	[256, 128, 32, 256, 64]	[64, 32, 128, 32, 16]
num_hidden_layers	1	1	1
hidden_layer_nodes	32	32	32
learning_rate	1.94×10^{-3}	9×10^{-4}	4×10^{-3}
lr_decay	0.986	0.989	0.988
weight_decay	0	0	0
acivations	['Tanh', 'LeakyReLU', 'ReLU', 'Tanh', 'Tanh']	['Tanh', 'ReLU', 'ReLU', 'LeakyReLU', 'LeakyReLU']	['Tanh', 'Tanh', 'ReLU', 'ReLU', 'Tanh']
layer_types	[GAT, GAT, GCN, GAT, GCN]	[GCN, GAT, GCN, GAT, GAT]	[GAT, GCN, GCN, GCN, GAT]
heads	[4, 7, 0, 1, 0]	[0, 3, 0, 6, 5]	[5, 0, 0, 0, 3]
pass_edge_attr	[0, 1, 0, 1, 0]	[0, 0, 0, 0, 1]	[0, 0, 0, 0, 1]
batch_norm_layers	[1, 0, 0, 0, 0]	[0, 1, 0, 1, 0]	[1, 0, 1, 0, 0]
batch_size	32	16	64

Table S6. Selected hyperparameters for adGC^{2}NN models.

Hyperparameter	adGC^{2}NN
num_conv_layers	5
num_conv_nodes	[32, 16, 64, 16, 32]
num_hidden_layers	2
hidden_layer_nodes	32, 32
num_merging_layers	1
merging_layer_nodes	8
learning_rate	6.25×10^{-4}
lr_decay	0.985
weight_decay	0
acivations	['LeakyReLU', 'LeakyReLU', 'ReLU', 'ReLU', 'LeakyReLU']
layer_types	[GCN, GAT, GCN, GCN, GAT]
heads	[0, 6, 0, 0, 6]
pass_edge_attr	[0, 1, 0, 0, 1]
batch_norm_layers	[0, 0, 0, 0, 0]
batch_size	4

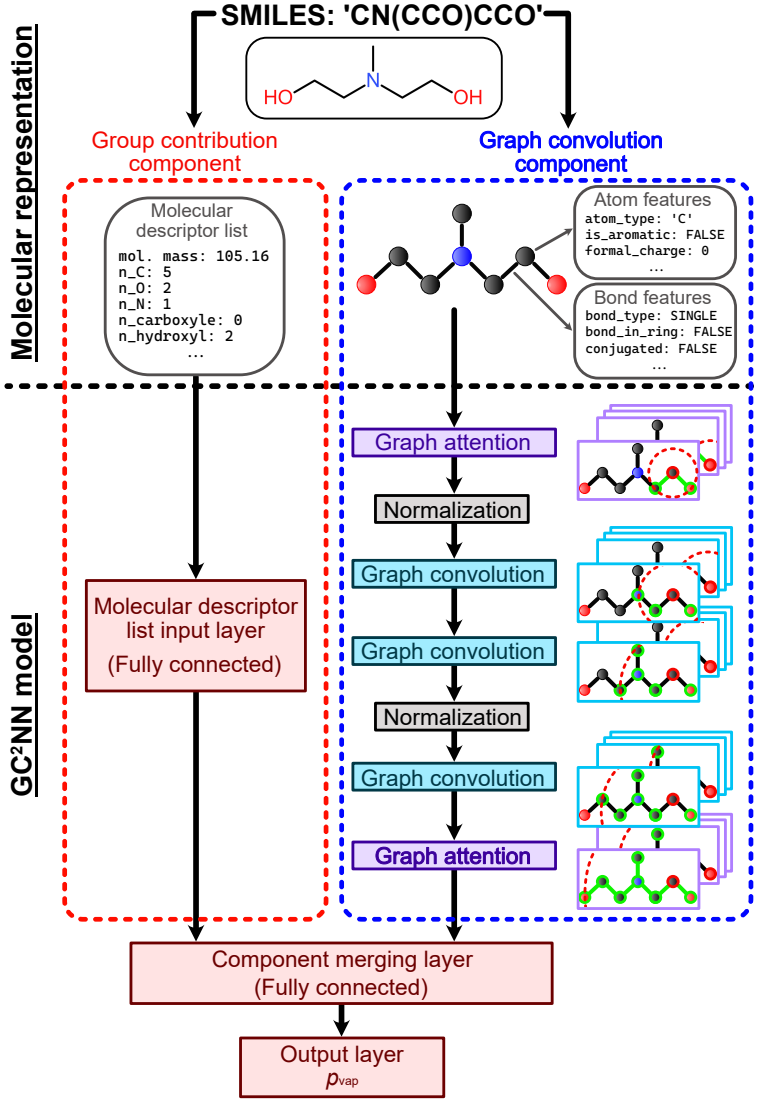

Figure S1. Schematic overview of molecular representation and model functionality of the fixed-depth GC^{2}NN (fdGC^{2}NN) model proposed in this work. Left: for the group contribution component, Simplified Molecular Input Line Entry System (SMILES) strings are used to derive holistic information on the molecule, such as its molar mass and the presence of atoms and functional groups (Tab. S3). Right: 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, graph convolution and batch normalization layers that normalize node or edge features across a batch, potentially stabilizing and accelerating the training. A fully-connected merging layer processes information from both model components and maps them to the single-node output layer, the p_{vap} prediction. Note that the displayed architecture represents model hyperparameters that were found optimal for a specific data set and model (fdGC^{2}NN-GeckoQ); the hyperparameters and thus architectures of other models presented in this study may deviate slightly in the type and order of layers in the graph convolution component (Tab. S5).

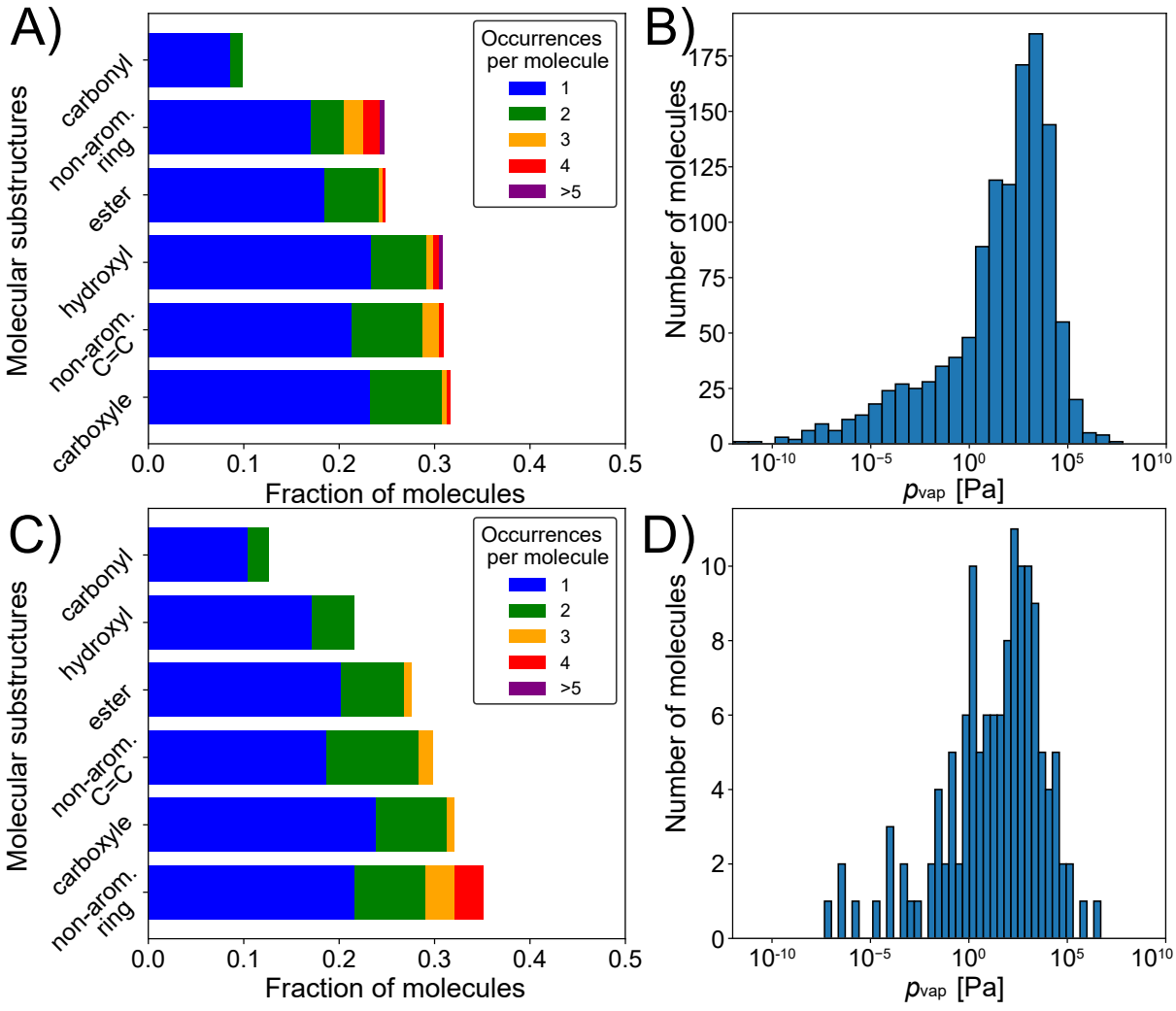

Figure S2. Occurrences of molecular substructures and vapor pressure measurements in the confined training plus validation (n = 1215; A, B) and test data set (n = 134; C, D), suitable for EVAPORATION (Compernolle et al., 2011). Panels A, and C show all substructures which are present in more than 1% of molecules in the respective data set. Panels B and D display histograms of experimental vapor pressure measurements in each data set. The distributions of molecular substructures and experimental vapor pressures are dissimilar, as 474 compounds present in the EVAPORATION training data are excluded from the test set.

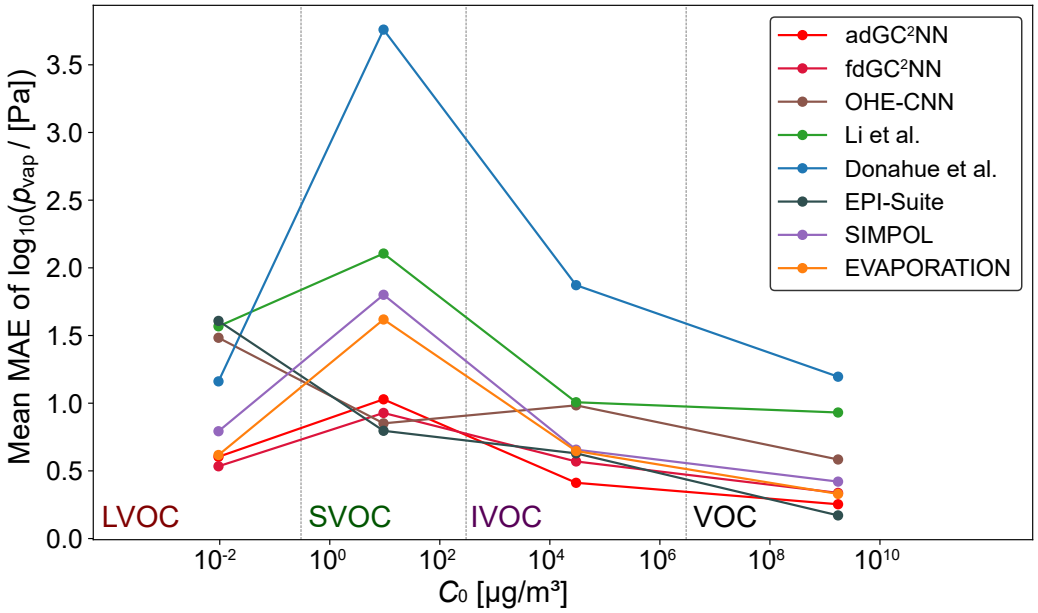

Figure S3. Mean confined test set prediction errors of four volatility bins as a function of experimental saturation concentration (C_0). Vertical dashed lines indicate interval borders of volatility bins. The number of compounds in each bin in the test set is ELVOC: 0, LVOC: 4, SVOC: 8, IVOC: 52, VOC: 70.

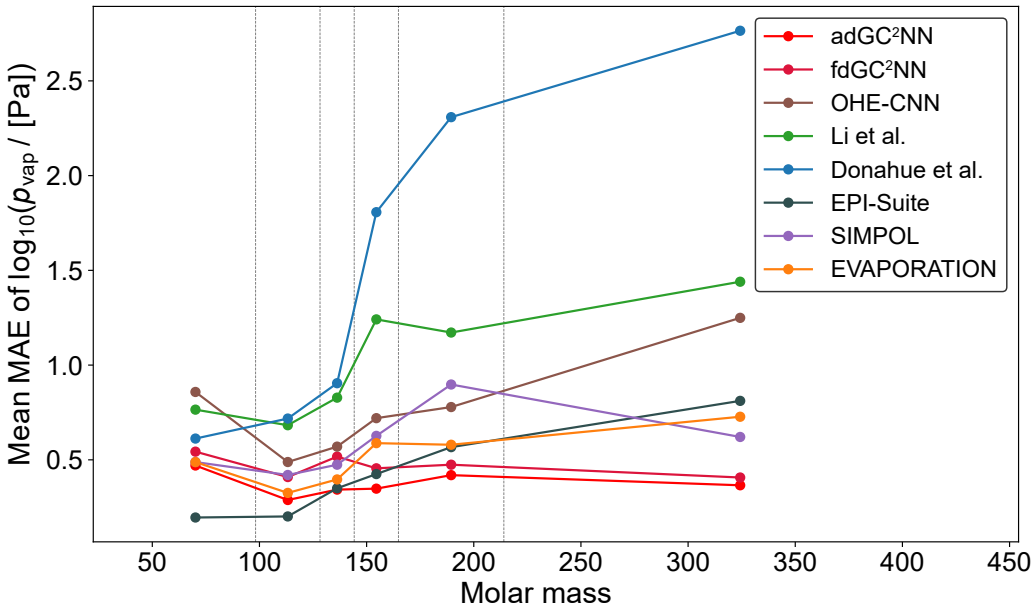

Figure S4. Mean confined test set prediction errors as a function of binned molecular masses. Vertical dashed lines indicate interval borders of mass bins. Bin intervals are selected so that each bin contains roughly 20 compounds from the test data set (22, 22, 21, 24, 22, 23).

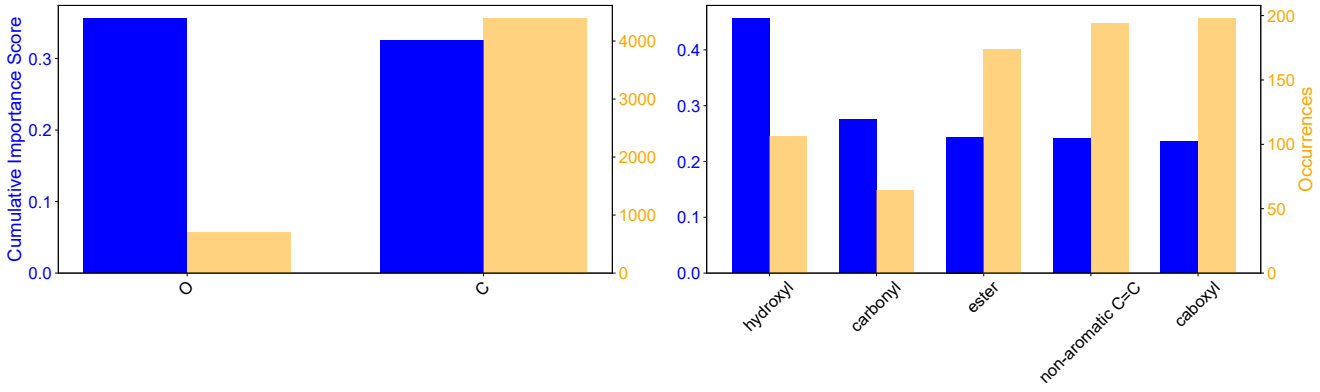

Figure S5. Cumulative importance scores and occurrences of atoms and functional groups in the confined test set (organic compounds with a limited set of functional groups), calculated in the second layer (graph attention layer) in the graph component of the trained T+V adGC^{2}NN-confined. Specifically, self-loop importances of the nodes attributed to various elements or functional groups are averaged to determine their relative importance among all neighboring nodes they are convoluted with.

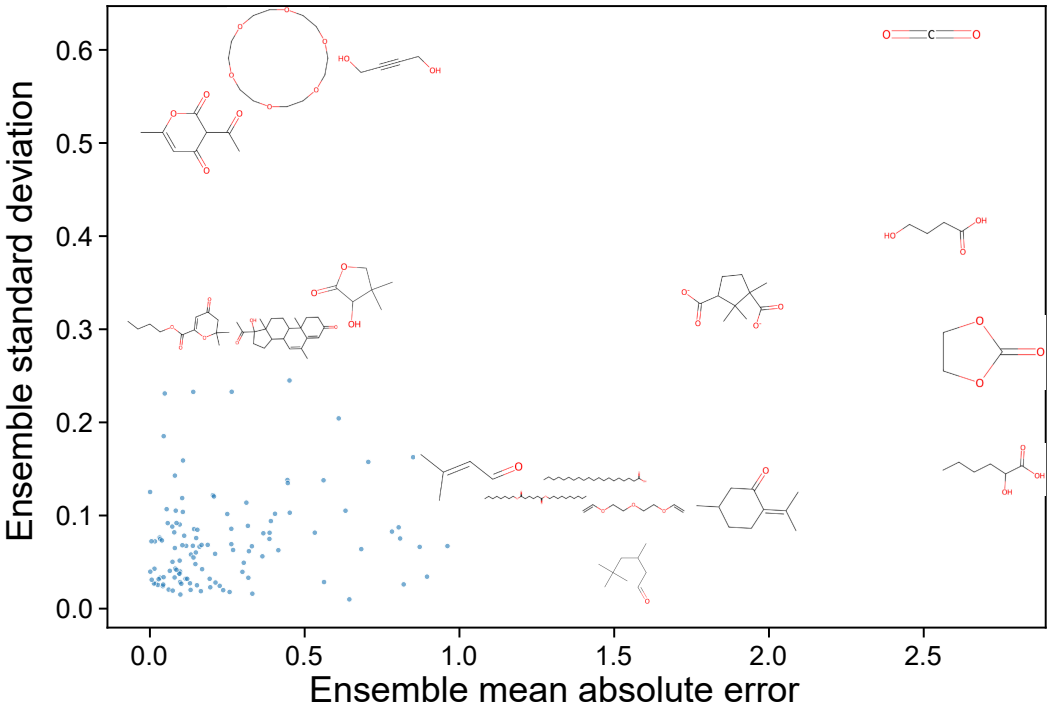

Figure S6. Ensemble standard deviation as a function of ensemble mean absolute error for the confined test set (organic compounds with a limited set of functional groups). Ensemble predictions originate from the confined adGC^{2}NN 5-fold cross validation models which are trained on different subsets of the training data. All compounds with an ensemble standard deviation larger than 0.3 or an ensemble mean absolute error larger than 1.0 are plotted as molecular structures. The compounds on the top of the figure are associated with large model uncertainty, while compounds in the bottom right have large errors despite small model uncertainty, a potential indicator for experimental uncertainty.

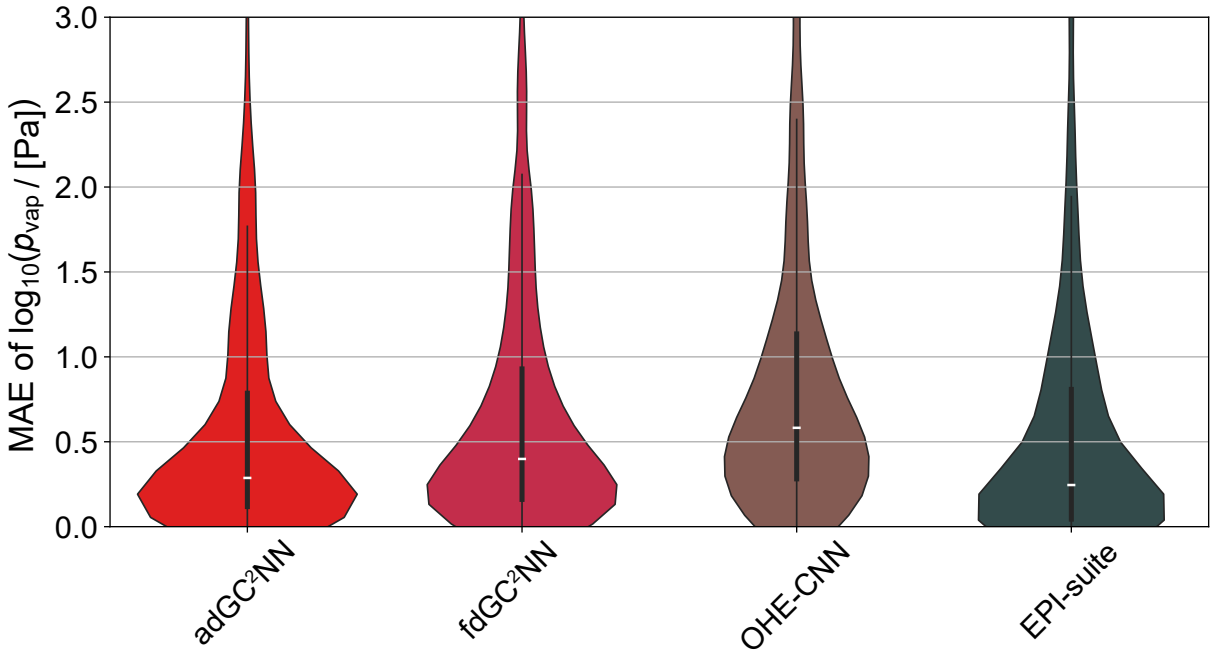

Figure S7. Violin plots representing broad test set error distribution of various models. Medians are shown as white markers, interquartile ranges as vertical wide black lines and 1.5 × interquartile ranges as narrow black lines.

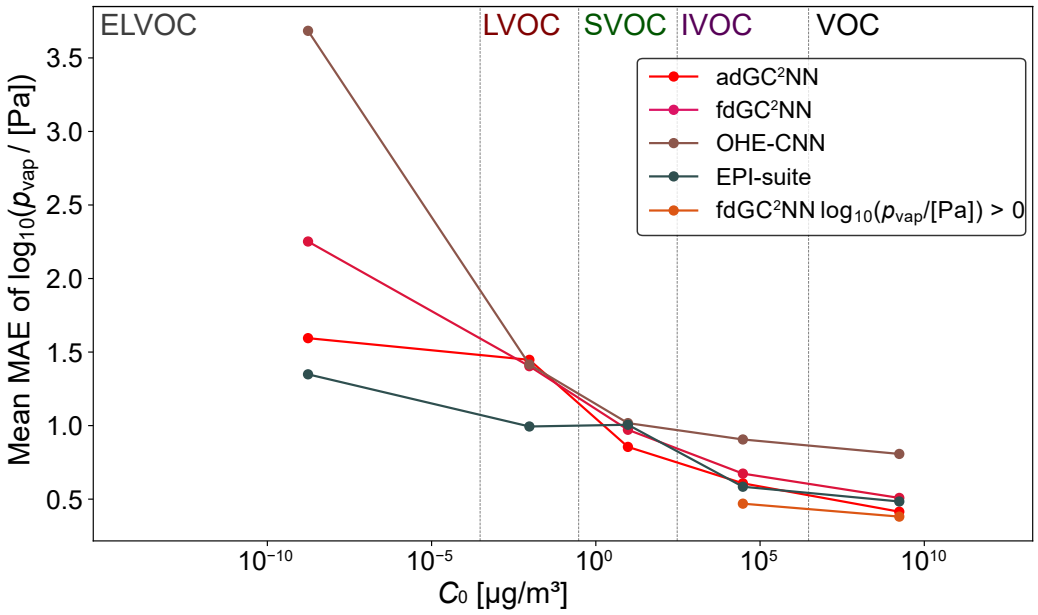

Figure S8. Mean broad test set prediction errors of five volatility bins as a function of experimental saturation concentration (C_0). Vertical dashed lines indicate interval borders of volatility bins. The number of compounds in each bin in the test set is ELVOC: 10, LVOC: 53, SVOC: 135, IVOC: 217, VOC: 202. An additional fdGC^{2}NN model is trained and tested on a subset of 3116 compounds (n_{train} = 2805, n_{test} = 311) with $\log_{10}(p_{vap}$ / [Pa]) > 0.

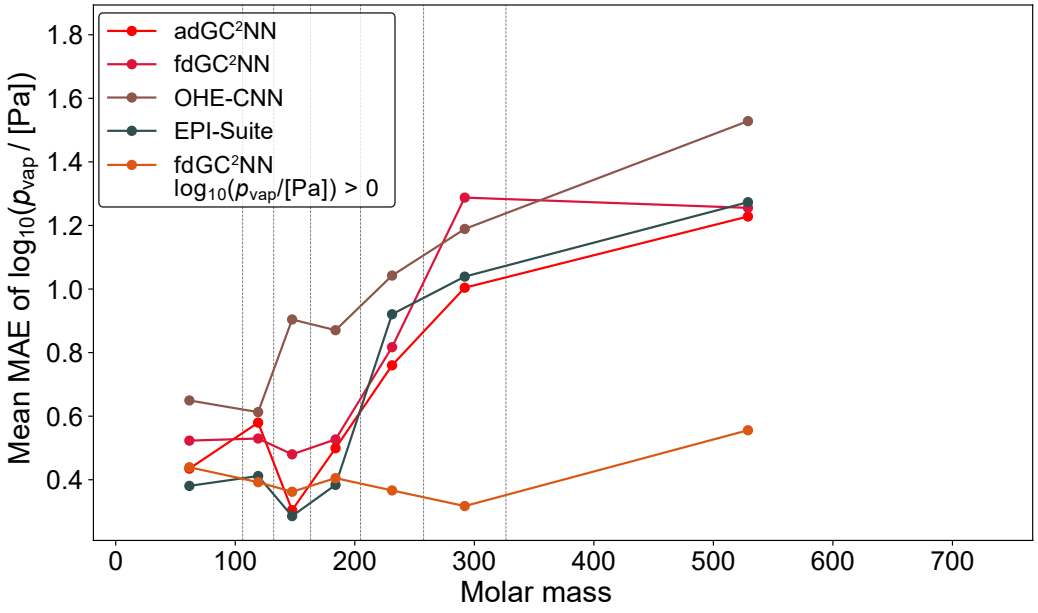

Figure S9. Mean broad test set prediction errors as a function of binned molecular masses. Vertical dashed lines indicate interval borders of mass bins. Bin intervals are selected so that each bin contains roughly 90 compounds from the test data set (88, 87, 88, 89, 88, 87, 89). An additional fdGC^{2}NN model is trained and tested on a subset of 3116 compounds (n_{train} = 2805, n_{test} = 311) with $\log_{10}(p_{vap}$ / [Pa]) > 0.

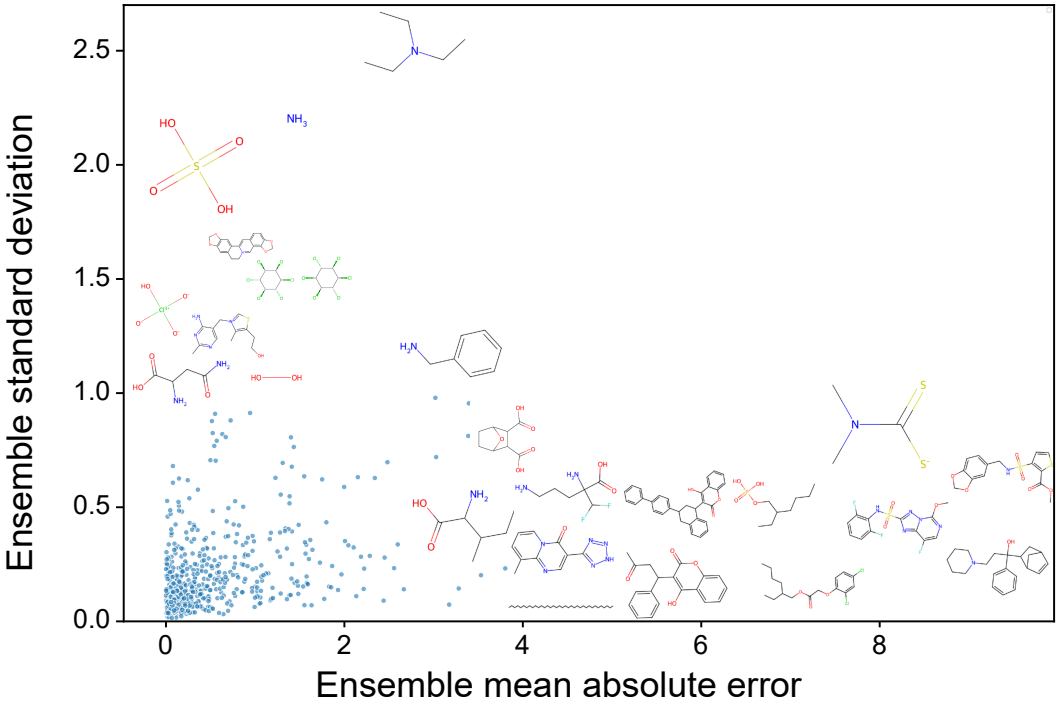

Figure S10. Ensemble standard deviation as a function of ensemble mean absolute error for the broad test set. Ensemble predictions originate from the broad adGC^{2}NN 5-fold cross validation models which are trained on different subsets of the training data. All compounds with an ensemble standard deviation larger than 1.0 or an ensemble mean absolute error larger than 4.0 are plotted as molecular structures. The compounds on the top of the figure are associated with large model uncertainty, while compounds in the bottom right have large errors despite small model uncertainty, a potential indicator for experimental uncertainty.

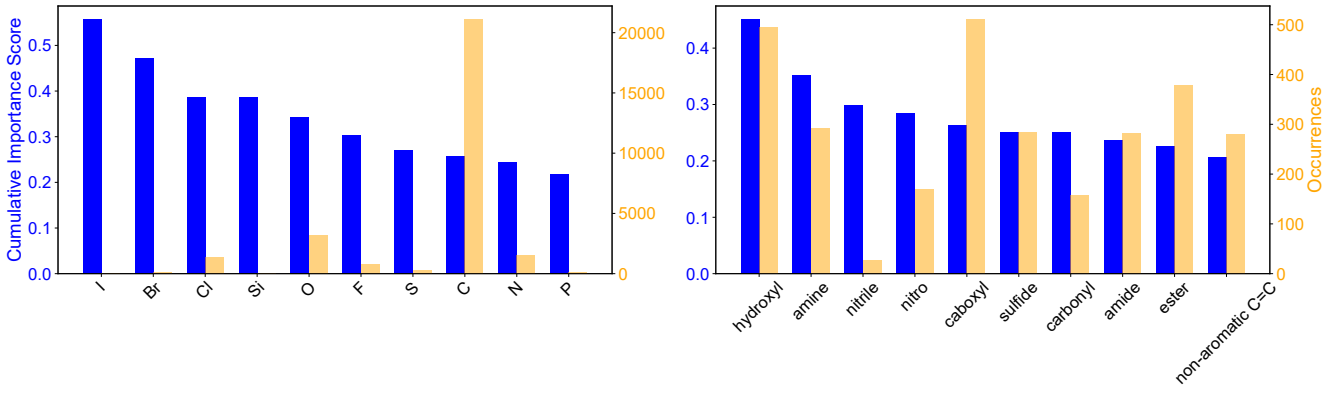

Figure S11. Cumulative importance scores and occurrences of atoms and functional groups in the broad test set (including inorganic compounds), calculated in the second layer (graph attention layer) in the graph component of the trained T+V adGC^{2}NN-broad. Specifically, self-loop importances of the nodes attributed to various elements or functional groups are averaged to determine their relative importance among all neighboring nodes they are convoluted with.

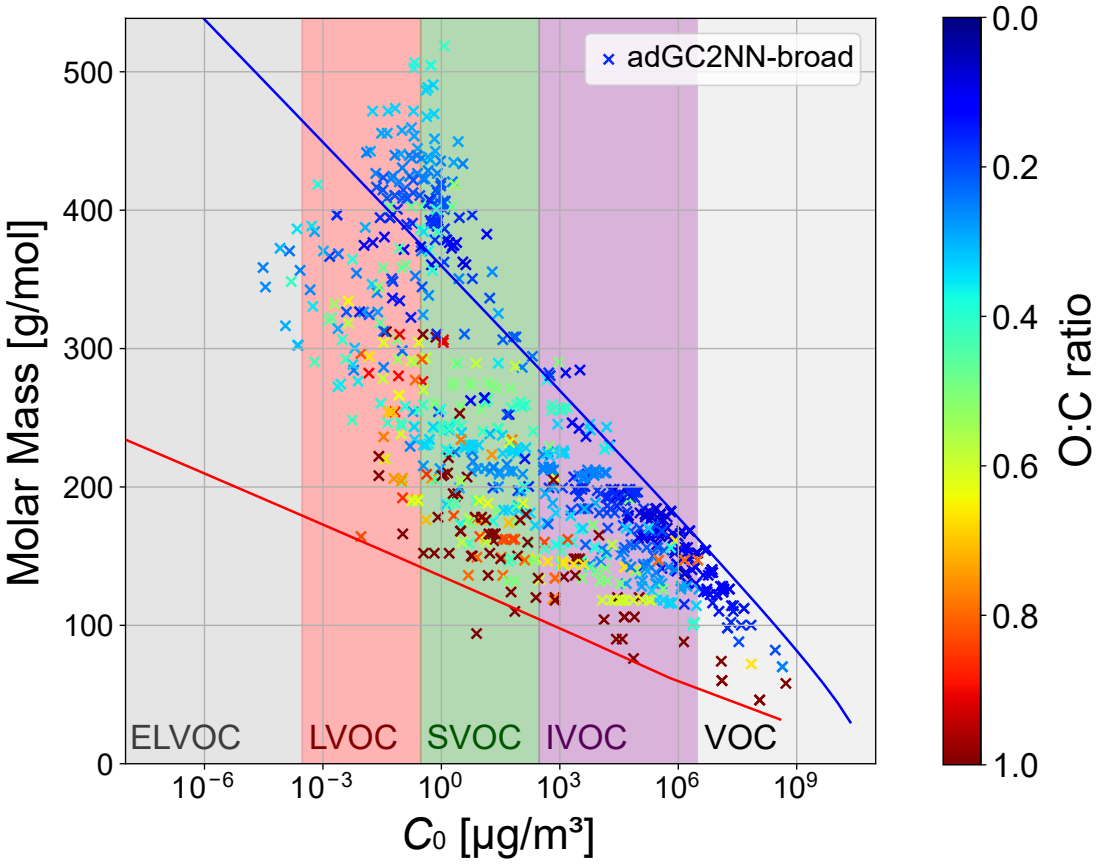

Figure S12. Molecular corridor plots following Shiraiwa et al. (2014). Application of the adGC^{2}NN-broad model to a data set of atmospherically relevant compounds (Shiraiwa et al., 2014). Blue and red boundary lines correspond to the volatility of n-alkanes and sugar alcohols (as determined by EVAPORATION), respectively.

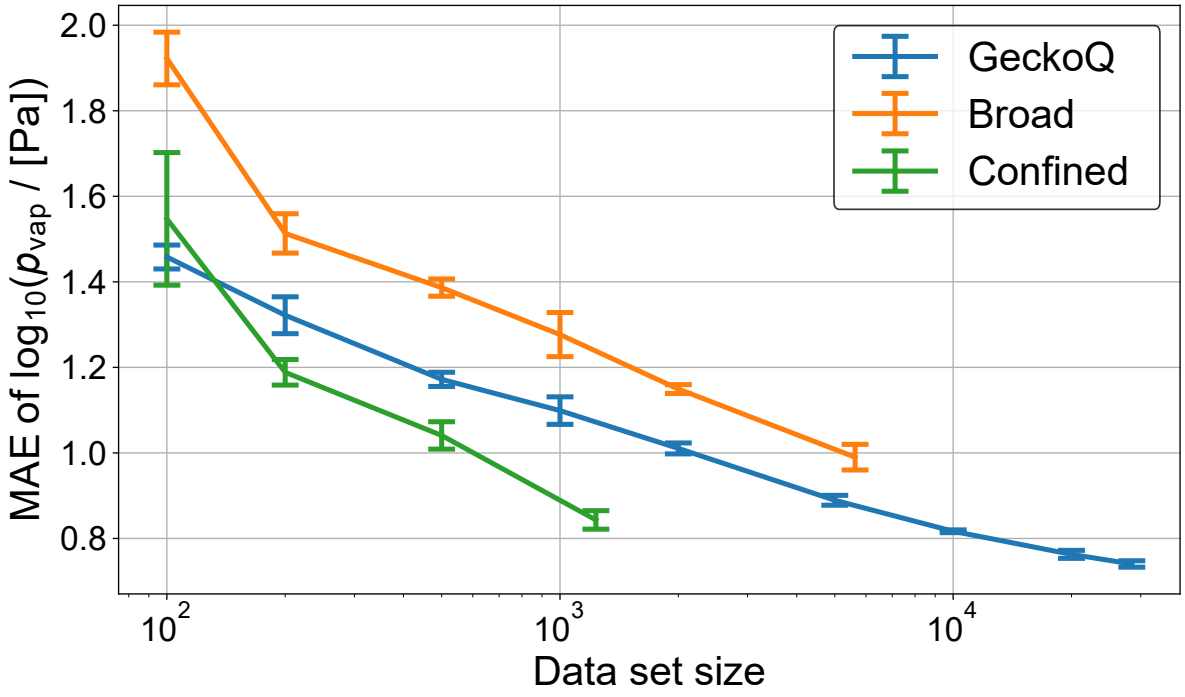

Figure S13. Mean absolute error (MAE) for independent test sets (confined: n = 137; broad: n = 625; GeckoQ: n = 3,163), as a function of training data set size of graph-only GCNN models trained on subsets of the three data sets. The experiment is performed by sampling subsets of various size from each of the respective data sets and training GCNN models on these. Hyperparameter 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 among the cross-validation folds.

References

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.

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

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.

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.

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.