ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Clusterome: A Comprehensive Data Set of Atmospheric Molecular Clusters for Machine Learning Applications

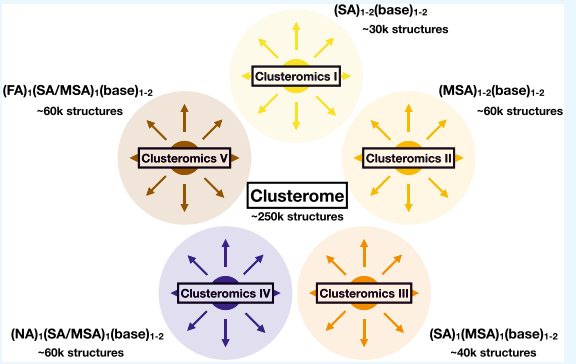
Yosef Knattrup, Jakub Kubečka, Daniel Ayoubi, and Jonas Elm*

Cite This: *ACS Omega* 2023, 8, 25155−25164

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Formation and growth of atmospheric molecular clusters into aerosol particles impact the global climate and contribute to the high uncertainty in modern climate models. Cluster formation is usually studied using quantum chemical methods, which quickly becomes computationally expensive when system sizes grow. In this work, we present a large database of ∼250k atmospheric relevant cluster structures, which can be applied for developing machine learning (ML) models. The database is used to train the ML model kernel ridge regression (KRR) with the FCHL19 representation. We test the ability of the model to extrapolate from smaller clusters to larger clusters, between different molecules, between equilibrium structures and out-of-equilibrium structures, and the transferability onto systems with new interactions. We show that KRR models can extrapolate to larger sizes and transfer acid and base interactions with mean absolute errors below 1 kcal/mol. We suggest introducing an iterative ML step in configurational sampling processes, which can reduce the computational expense. Such an approach would allow us to study significantly more cluster systems at higher accuracy than previously possible and thereby allow us to cover a much larger part of relevant atmospheric compounds.



1. INTRODUCTION

Aerosols are suspensions of solid and liquid particles in the air. They are the main contributors to uncertainties in modern climate models, as confirmed by the recent IPCC report.[1] Aerosols affect the climate by scattering and absorbing sunlight, which changes the global radiation balance,[2] and they act as seeds for cloud droplet formation, termed cloud condensation nuclei (CCN).[3] About half of the global CCN originate from atmospheric new particle formation (NPF[4]), i.e., gas-to-particle formation. NPF can be initialized via various gas-phase precursors bonding noncovalently into molecular clusters, which grow with further uptake of different vapors.[5] Hence, most climate-modeling uncertainty arises from the ambiguities of which specific compounds are involved in the initial clustering and the further growth into aerosol particles.[6]

Sulfuric acid has been shown to play a major part in NPF,[5,7−9] and atmospherically relevant acids such as methanesulfonic acid, nitric acid, and formic acid are capable of enhancing the cluster formation potential of sulfuric acid (SA)-based clusters.[10−14] However, the acids alone are not capable of forming strongly bound clusters under realistic atmospheric conditions. Instead, SA clusters are stabilized by highly abundant bases such as ammonia or bases with high basicity such as methylamine, dimethylamine, trimethylamine, and ethylenediamine.[8,15−25]

Recently, several quantum chemical studies have given insight into cluster thermodynamics and formation.[26,27] The cluster structures are usually examined using a funneling approach to efficiently explore the vast configurational space. In this approach, a multitude of structures are initially examined at a low level of theory, and only a few are taken for re-examination at higher levels of theory. Performing high-level (e.g., density functional theory (DFT)) calculations for numerous candidate structure candidates in order to find the global (free) energy minimum is currently an enormous bottleneck in atmospheric cluster calculations. Large clusters are especially problematic, as the number of minima in the configurational space scales exponentially with system size and the high-level DFT methods scale quartically.[28] As machine learning (ML) methods can be faster than DFT calculations, we examine how ML models could help to filter out high-energy structures and thus reduce the number of potential structures needed to be calculated using the computationally demanding quantum chemical methods.

Using ML models, one can train a property of interest, such as electronic energy, on a training set of molecular representations and then predict the same properties for a new structure.[29−31] However, several studies showed that the

Received: April 2, 2023
Accepted: June 16, 2023
Published: June 30, 2023

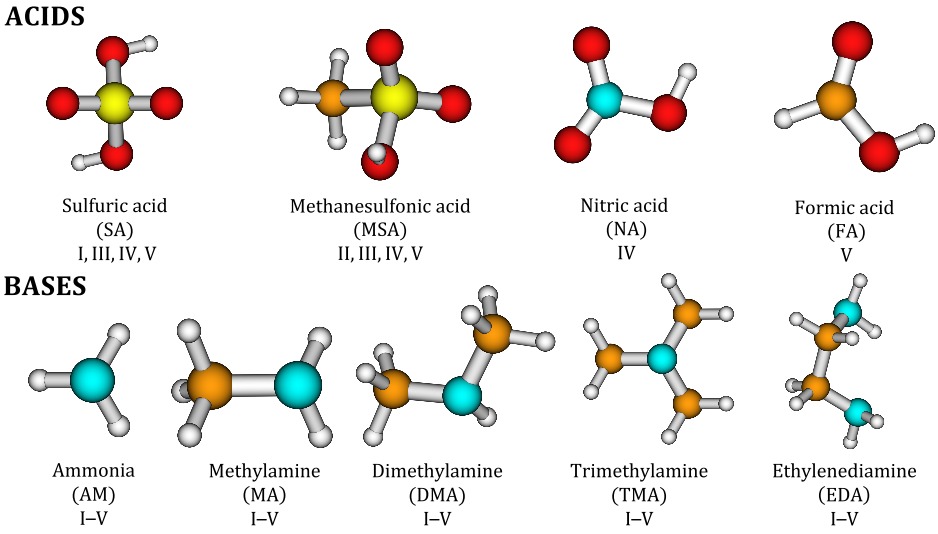


Figure 1. Monomers contained in the Clusteromics data sets. The Roman numerals refer to which Clusteromics data set the monomer is present in.

performance of Δ-ML models where the training and prediction are done on the difference between two methods consistently outperforms the 'direct-ML' models in predicting the property of interest.[32−34] For instance, Kubečka et al.[34] used Δ-ML between the GFN1-xTB and ωB97X-D/6-31++G(d,p) methods to predict sulfuric acid−water cluster binding energies with mean absolute errors down to 0.5 kcal/mol. To train a proper ML model, a representative database is required. In the Clusteromics series of papers,[10−14] our group has gathered a database of unique atmospheric acid−base cluster structures containing acids such as sulfuric acid (SA), methanesulfonic acid (MSA), formic acid (FA), and nitric acid (NA), and with bases such as ammonia (AM), methylamine (MA), dimethylamine (DMA), trimethylamine (TMA), and ethylenediamine (EDA). The acids and bases and their distribution in the Clusteromics series are depicted in Figure 1. This large data set consists of 22,870 equilibrium cluster structures at the ωB97X-D/6-31++G(d,p) level of theory.

In this work, we generate a massive data set of atmospheric molecular clusters based on the Clusteromics I−V data sets to test the applicability of ML methods on various multi-component cluster systems.

2. METHODOLOGY

2.1. Computational Details. The stability and binding strength of atmospheric molecular clusters is typically evaluated at the ωB97X-D/6-31++G(d,p) level of theory, followed by high-level DLPNO−CCSD(T_0)/aug-cc-pVTZ single-point electronic energy corrections.[26,35] Recent benchmark work by Jensen et al.[33] revealed that empirically corrected DFT methods such as r^{2}SCAN-3c[36] are significantly faster than other DFT methods while performing with the same or better accuracy. In this work, we also examined the semiempirical methods AM1,[37] PM3,[38] GFN1-xTB,[39] and GFN2-xTB[40] for out-of-equilibrium structure generation using molecular dynamics (MD) simulations. The MD simulations were performed in ORCA 5.0.0.[41,42]

2.2. Machine Learning Model. In this work, we use the KRR method and the FCHL19 molecular representation,[43] which showed excellent results for water clusters[43] and atmospherically relevant clusters.[33,34,44] All mathematical scripts can be found within the quantum machine learning (QML) program,[45] which we interfaced with our own Jammy Key for Machine Learning (JKML[part of JKCS]) scripts that handle file management and automatizes the procedure. For kernel ridge regression with a local representation, it is assumed that a molecular property $\mathcal{M}(i)$ (energy in our case) of a molecular system target i can be written as a sum of atomic contributions.

$$\mathcal{M}(i) = \sum_{B\in i} \mathcal{M}_{\text{local}}(\mathbf{q}_B) = \sum_{B\in i}\sum_{j}\sum_{A\in j} \mathcal{K}(\mathbf{q}_A, \mathbf{q}_B)\alpha_j \tag{1}$$

where A and B are atoms in the target i and the trained system j, respectively, $\mathbf{q}$ is the molecular representation of the system, α_j is the regression coefficients for the training system j, and $\mathcal{K}(\mathbf{q}_A, \mathbf{q}_B)$ are the pairwise kernels between the atoms in the two molecules. We use a Gaussian kernel with the L_2-norm.

$$\mathcal{K}(\mathbf{q}_A, \mathbf{q}_B) = \delta_{Z_A Z_B} \exp\left(-\frac{\|\mathbf{q}_A - \mathbf{q}_B\|_2^2}{2\sigma^2}\right) \tag{2}$$

where σ is the kernel width and Z_A and Z_B are the nuclear charges of atoms A and B, respectively. These equations can be written in matrix notation as

$$\vec{E} = \mathbf{K}\cdot\vec{\alpha} \tag{3}$$

where $\vec{\alpha}$ is the collection of regression coefficients α_j and the elements of the kernel matrix $\mathbf{K}$ are given as a sum over pairwise kernels between atoms of two molecules A and B

$$\mathbf{K}_{ij} = \sum_{B\in i}\sum_{A\in j} \mathcal{K}(\mathbf{q}_A, \mathbf{q}_B) \tag{4}$$

The α coefficients, which can be obtained by minimizing the cost function, have the following analytical solution:

$$\boldsymbol{\alpha} = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{M} \tag{5}$$

Here, λ is a small constant added to the diagonal of the matrix to ensure numerical stability when inverting the kernel matrix.

https://doi.org/10.1021/acsomega.3c02203
ACS Omega 2023, 8, 25155−25164

Since the $(\mathbf{K} + \lambda\mathbf{I})$ is Hermitian and positive-definite, Cholesky decomposition is used to solve for the coefficients.[46]

In this work, the r^{2}SCAN-3c level of theory was chosen as the target level for the ML model. Although r^{2}SCAN-3c is fast and its substitution with ML is perhaps unnecessary, we use it as a proof of concept, and the trends should be applicable to any quantum chemical method. The model is based on Δ-ML and is thus trained on the difference of electronic binding energies between GFN1-xTB (XTB) and r^{2}SCAN-3c (DFT)

$$\Delta E_{\text{binding}}^{\text{DFT}-\text{XTB}} = \left(E_{\text{DFT}}^{\text{cluster}} - \sum E_{\text{DFT}}^{\text{monomers}}\right) - \left(E_{\text{XTB}}^{\text{cluster}} - \sum E_{\text{XTB}}^{\text{monomers}}\right) \quad (6)$$

where each bracket corresponds to cluster electronic binding energy at the level of theory given by the subscripts. Such a Δ-ML model could be applicable in the funneling configurational sampling approach commonly used for atmospheric clusters, as it could identify energetically high-lying structures and thus reduce the number of calculations at a DFT step. We used the hyperparameters $\sigma = 1$ and $\lambda = 10^{-4}$ found by Kubečka et al.[34] for the KRR model, as these seem to work well for acid−water and acid−base clusters.

2.2.1. Model Scaling. A disadvantage of KRR is that the inversion of the kernel matrix scales as $O(N^3)$, where N is the number of structures. However, for small data sets with less than $\sim 10^5$ structures, the pre-factor is quite small, and thus the training and evaluation procedure scales as $O(N^2)$ hindered by the kernel matrix construction. The matrix construction can be computationally demanding, and using the "normal" approach for QML, we were only able to train on ∼30,000 structures in a reasonable timeframe (∼3 days/48 CPU). Therefore, we implemented kernel matrix construction per part. Since the kernel matrix is symmetric, only the upper half of the matrix has to be calculated, and since the elements in the matrix are independent of each other, the calculation of the matrix elements can be split into an arbitrary number of smaller calculations. With the matrix split on a grid of N-times smaller matrices (follow Figure 2), the $(N^2 - N)/2$ asymmetric and N symmetric sub-matrices can be calculated within parallel independent jobs.

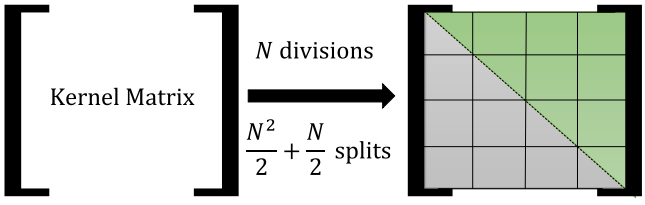


Figure 2. Illustration of kernel matrix being divided $N = 4$ times yielding 10 sub-calculations.

Afterward, the kernel matrix is reconstructed. Nevertheless, the $O(N^3)$-scaling Cholesky decomposition must still be performed and becomes the main bottleneck for large databases. This procedure allowed us to train on databases with up to 150k structures. The kernel−split procedure is implemented in JKML (use the -split <int> keyword). If the model is applied to a smaller data set, the computational times for model tests/evaluations, which would otherwise scale similarly to the training, are not demanding and the kernel matrix split will not be required.

3. RESULTS AND DISCUSSION

3.1. ML Model Validation. To show that the KRR method with FCHL19 works well for our systems, we did 5-fold cross-validation of each of the Clusteromics I−V data sets. We used the Δ-ML model to train on the difference in binding energies of the GFN1-xTB and r^{2}SCAN-3c methods. Finally, we predicted the r^{2}SCAN-3c binding energies achieving mean absolute errors (MAEs) lower than 0.5 kcal/mol even with a small training database of 125 structures as seen in Figure 3.

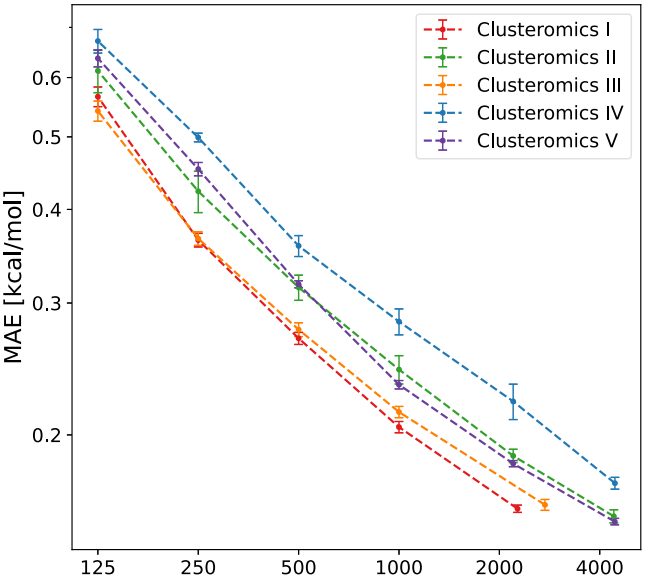


Figure 3. Mean absolute error (MAE) for 5-fold cross-validation of the Clusteromics data sets using increasing training set sizes. The error bars correspond to the standard deviation of the sample mean. Note the logarithmic axis.

The learning curves are almost linearly decreasing in this figure (note the logarithmic axes), validating the choice of hyperparameters and the suitability of the applied Δ-ML model for these systems. The predictions for Clusteromics III−V have slightly higher MAEs than the Clusteromics I−II predictions. The III−V databases contain more compounds, and thus the chemistry of the database is more diverse, leading to ∼10% worse results. However, when applying the full data sets very low errors of ∼0.1 kcal/mol are achieved. These results show the direct applicability of the Clusteromics I−V data sets for Δ-ML modeling.

3.2. Expansion of Clusteromics. Kubečka et al.[34] showed that if ML models need to extrapolate outside of the training database such as to larger clusters, the addition of out-of-equilibrium structures to the training database is needed. These out-of-equilibrium structures can be generated through 3 main procedures: (1) normal mode sampling,[47−50] (2) extraction from geometry optimization, where the intermediate structures saved in the quantum chemistry output file during optimization are used, and (3) using molecular dynamics (MD) simulations around each/some structure(s) within the initial database.[51,52] The Clusteromics databases are equilibrium structures at the ωB97X-D/6-31++G(d,p) level of theory. Here, we examine which method and how many out-of-equilibrium structures are required to expand the Clusteromics database. We validate our choice on the $(SA)_1(AM/DMA)_1$ and $(SA)_2(AM/DMA)_2$ clusters (the '/'

https://doi.org/10.1021/acsomega.3c02203
ACS Omega 2023, 8, 25155−25164

sign refers to only AM or only DMA molecules), all within the Clusteromics I database, giving an equilibrium data set of 106 structures. Further, we define a new test set by generating 20 new distinct out-of-equilibrium configurations for each equilibrium structure (using MD simulation, see below). When only equilibrium structures are used as the training database, the Δ-ML validated on the test set achieves MAEs of 16.5 kcal/mol as seen in Figure 4, which illustrates the need for out-of-equilibrium structures.

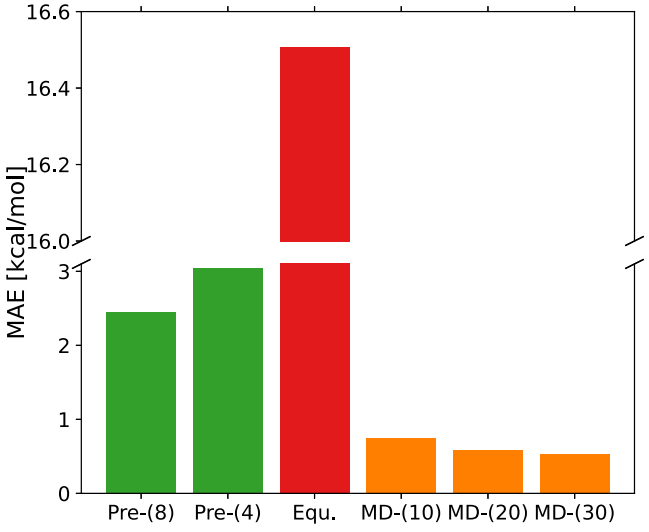


Figure 4. Mean absolute error (MAE) of the Δ-ML model for predicting binding energies of the random out-of-equilibrium structures of $(SA)_1(AM/DMA)_1$ and $(SA)_2(AM/DMA)_2$. The training database is composed of the equilibrium structures and: nothing else (red), up to 4 and 8 pre-equilibrium structures per equilibrium structures (green), and using 10, 20, and 30 MD generated structures per equilibrium structures (orange).

3.2.1. Pre-Equilibrium Extraction Method. From the QC output files in the Clusteromics data sets, we extracted all of the intermediate structures and energies from the ωB97X-D/6-31++G(d,p) geometry optimization. The extraction is limited to the number of steps during the geometry optimization. These optimizations typically require from a few to hundred(s) optimization steps, while the first few steps correspond to the largest change in geometry. Thus, we take up to the 4 or 8 first optimization structures, which are essentially out-of-equilibrium, to enlarge the Clusteromics database. Figure 4 shows that the Δ-ML MAE drops down to 3.04 or even to 2.44 kcal/mol when the first 4 or 8 pre-equilibrium structures, respectively, are also included in the training of the ML model. The new out-of-equilibrium structures improved our model, but clearly, its improvement is only limited to a few structures because there is no/minor gain when similar structures are added to the training database.

3.2.2. MD Simulation Method. The AM1 and PM3 methods do not accurately describe the interaction for simple systems such as the $(SA)_1(AM)_1$ cluster (see also the results in Jensen et al.[33]) and MD simulations with these methods lead to dissociation within a few nanoseconds for these clusters (see Figure S1). Hence, we discarded the AM1 and PM3 methods from further examination.

Figure 5 shows the MD simulation for the $(SA)_2(DMA)_2$ cluster. The other $(SA)_{1-2}(AM/DMA)_{1-2}$ clusters are shown in Section S5 together with the technical details in Section S3. The simulations are performed at three levels of theory: GFN1-xTB, GFN2-xTB, and r^{2}SCAN-3c. However, all single-point energies are calculated at r^{2}SCAN-3c. The correlations between the GFN*i*-xTB and r^{2}SCAN-3c//GFN*i*-xTB are also shown in Figure 5b,c. The GFN1-xTB and r^{2}SCAN-3c trajectories appear to have similar features in terms of a visual examination of the geometry evolution and the span of the trajectory energies. The GFN2-xTB method yields an erratic trajectory and a larger span of energies than the r^{2}SCAN-3c simulation. Additionally, when inspecting the cluster structures, the GFN2-xTB clusters are much more tightly bound, i.e., the equilibrium bond lengths and average cluster radius of GFN2-xTB geometries significantly differ from the r^{2}SCAN-3c geometries. The correlation of GFN2-xTB with r^{2}SCAN-3c//GFN2-xTB is also worse than in the case of GFN1-xTB. The poor performance of GFN2-xTB for clusters containing sulfur-based acids is consistent with the benchmark by Jensen et al.[33] and is attributed to the decrease in the number of d-functions for sulfur for the basis set compared to GFN1-xTB. Although r^{2}SCAN-3c is extremely fast compared to other DFT methods, using it for MD simulations for all clusters, which contain up to 42 atoms, is still computationally expensive and slow. Therefore, in this work, we use GFN1-xTB for further MD simulations, and we define the r^{2}SCAN-3c method as the high level of theory, as suggested by Jensen et al.[33]

We performed MD simulations, at the GFN1-xTB level, around each $(SA)_1(AM/DMA)_1$ and $(SA)_2(AM/DMA)_2$ equilibrium structure to generate up to 30 out-of-equilibrium structures for each cluster. For technical details, see Section S4. Figure 4 also shows the MAEs of the Δ-ML model when the equilibrium structure training database is expanded with up to 10, 20, and 30 out-of-equilibrium structures from the MD simulations. Already, the addition of 10 structures leads to an MAE of 0.75 kcal/mol. Adding more structures lowers the MAE but also increases the computational time for the ML training and evaluation (see Section 2.2.1).

Based on the comparison of the expansion methods, we chose the MD-(10) expansion scheme as a good compromise between the ML model accuracy and computational time. Subsequently, all Clusteromics databases were expanded with up to 10 out-of-equilibrium structures using the MD simulations at the GFN1-xTB level of theory. The simulations were run for 2 ps, with a 0.5 fs timestep, and saving the geometry at 0.2 ps intervals. After the expansion, the full Clusteromics data set contained ∼250k structures. For all structures, we calculated the single-point energies at the GFN1-xTB and r^{2}SCAN-3c levels. We furthermore define the 'standardized Clusteromics' term, where all monomers and homo-/hetero-dimers from the full Clustermics I−V data sets are added to the individual data set. All of the generated data sets are freely available in the Atmospheric Cluster DataBase (ACDB).[53]

3.3. ML Model Extrapolation. An ML model can easily fail when tested on structures different from those in the training data set. In the previous section, we demonstrated that training on the binding energies of equilibrium configurations yields large errors when predicting binding energies for out-of-equilibrium structures. In Figure 6, we show how an ML model can be tested on its transferability and extrapolation,[54] and we further examine these options in the following sections.

3.3.1. Transferability of Interactions. Even if all tested clusters contain molecules that are in the training data set, a specific type of intermolecular interaction might be missing.

https://doi.org/10.1021/acsomega.3c02203
ACS Omega 2023, 8, 25155−25164

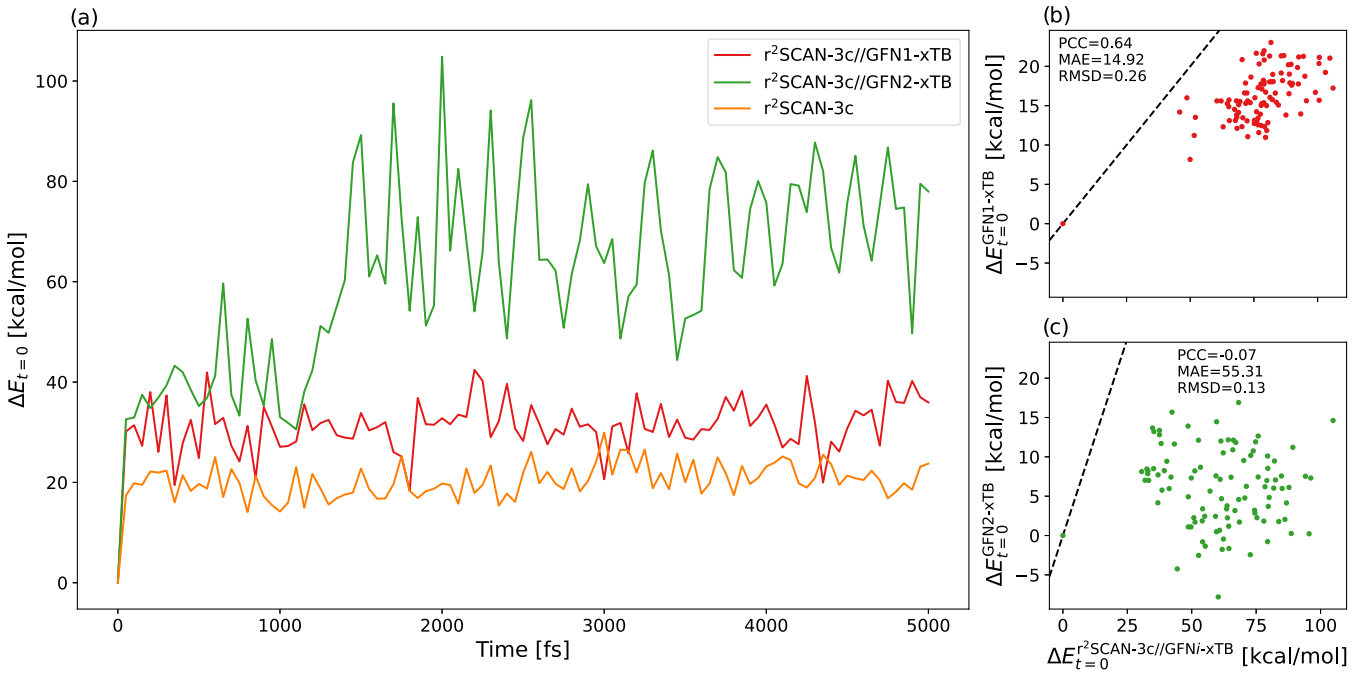


Figure 5. (a) Energy evolution of the $(SA)_2(DMA)_2$ cluster at r²SCAN-3c//GFN1-xTB (red), r²SCAN-3c//GFN2-xTB (green), and r²SCAN-3c (orange) level of theory relative to the energy of the initial ($t = 0$) structure. (b, c) Correlation of the relative energies $\Delta E_{t=0}^{\text{Method}}$ calculated with the GFN*i*-xTB and r²SCAN-3c//GFN*i*-xTB methods. PCC = Pearson correlation coefficient, MAE = mean absolute error [kcal/mol], RMSD = root-mean-squared displacement [kcal/mol].

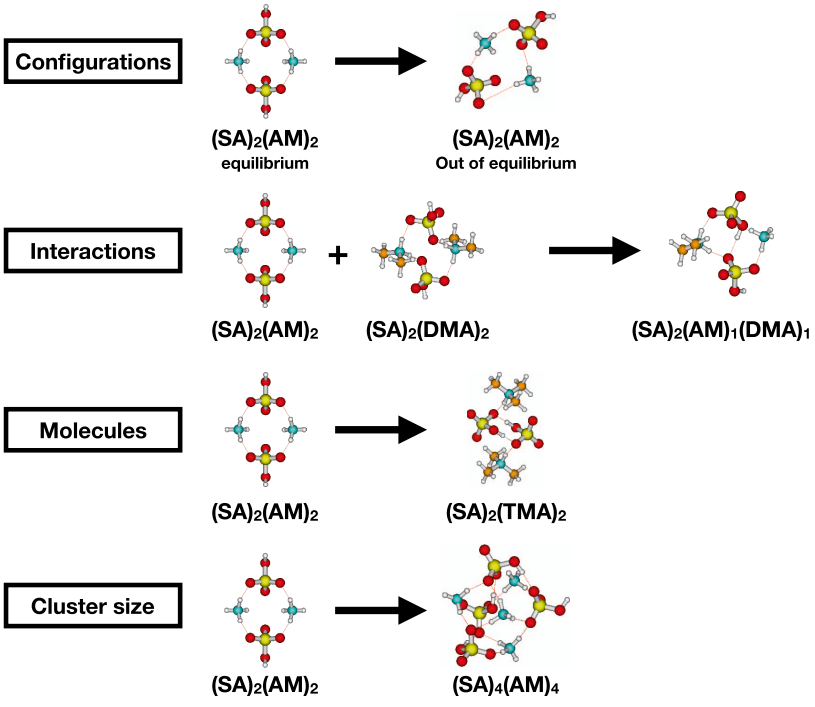


Figure 6. Different types of ML extrapolations tested in this work, as suggested by Kubečka et al.[54]

Here, we train the ML model on the equilibrium and out-of-equilibrium $(SA)_{0-2}(base)_{0-2}$ and $(SA)_{0-2}(base')_{0-2}$ clusters and test on the mixed $(SA)_{1-2}(base)_1(base')_1$ equilibrium and out-of-equilibrium clusters. Figure 7 shows the results for all base/base′ combinations. The MAEs are below 1 kcal/mol for all of the systems, and they all have a similar span of errors (3.1−6.5 kcal/mol), which indicates that there are some outliers. The systems containing EDA have the largest span of errors (3.8−6.5 kcal/mol). This is most likely due to EDA being a too "flexible" molecule with more complex configurational space compared to the other bases (see Figure 1). Overall, we use 1903−8459 training structures, and the MAEs are only slightly worse than in Figure 3, even though we also include out-of-equilibrium structures and extrapolate the ML model out of the training set. In general, we only tested what happens if indirect base−base' interactions are missing in the

https://doi.org/10.1021/acsomega.3c02203
ACS Omega 2023, 8, 25155−25164

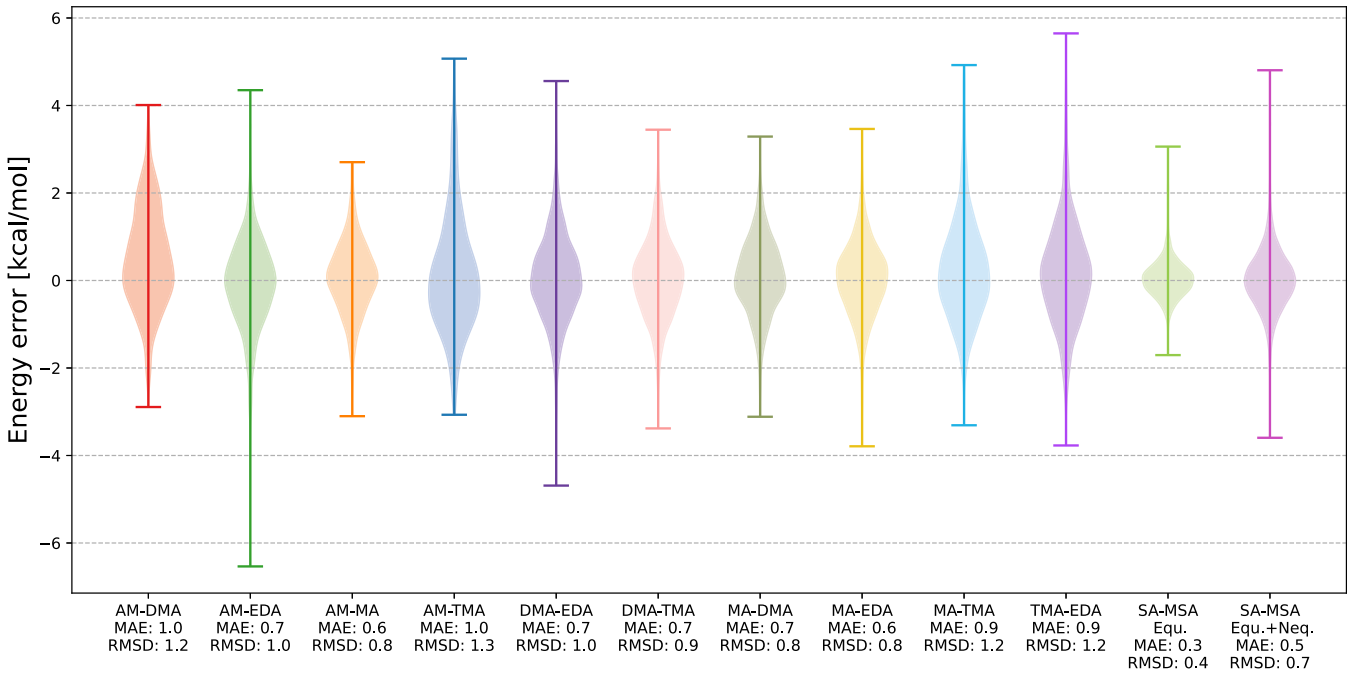


Figure 7. ML model error distribution for electronic binding energy prediction (y-axis) of systems with the indirect base−base' (x-axis) interaction missing in the training data set. The right part of the graph shows the same for both equilibrium (eq.) and also nonequilibrium (neq.) clusters with SA−MSA interaction missing in the training data set. MAE = mean absolute error [kcal/mol], RMSD = root-mean-squared deviation [kcal/mol].

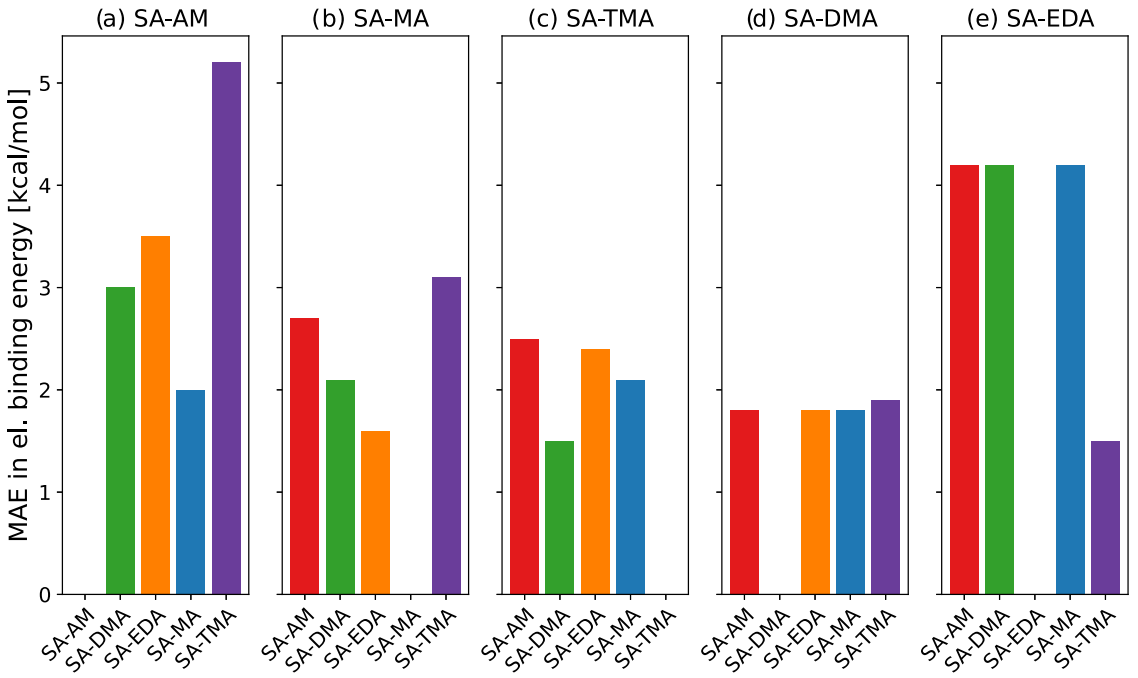


Figure 8. Mean absolute errors (MAEs) of ML modeled binding energies when extrapolating to different types of base molecules.

training set. However, the MAEs would most likely rise if a direct SA−base interaction was missing.

We performed a similar test for the acid interactions by training the ML model on the standardized Clusteromics I (only SA−base clusters) and standardized Clusteromics II (only MSA−base clusters) databases. The test is performed on Clusteromics III, which contains the mixed SA−MSA−base clusters. Figure 7 shows that we reach a low MAE of 0.3 kcal/mol for the equilibrium test data sets and 0.5 kcal/mol for the equilibrium and out-of-equilibrium data sets. Low MAEs are achieved as the SA−MSA interactions are similar to the SA−SA interactions because the MSA-methyl group does not play a significant role in the cluster binding energy. Most of the outliers outside the main distribution are again EDA-based clusters, and the ones with the largest error are the $(EDA)_2$-based clusters. Fortunately, these outliers are all high-energy conformers, and the model is capable of predicting the energetically lowest structures, which are the ones sought after during configurational sampling. In general, the MAE is approximately only twice as large compared to the test on the

https://doi.org/10.1021/acsomega.3c02203
ACS Omega 2023, 8, 25155−25164

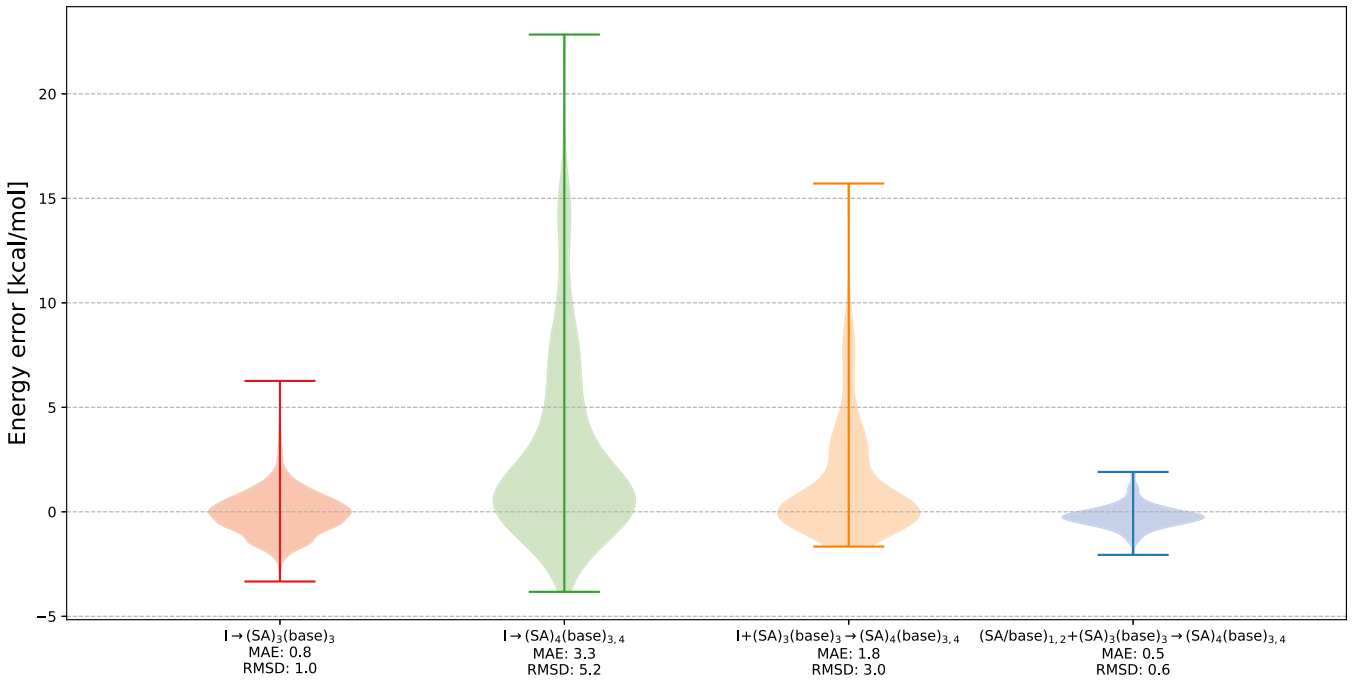


Figure 9. Error distribution for ML modeling of binding energies for large cluster sizes.

full equilibrium data sets shown in Figure 3. This shows that the model is easily capable of extrapolating the acid interactions with a low loss in accuracy and that the model is capable of handling out-of-equilibrium mixed acid clusters. To show that not accounting for the direct interactions can lead to larger errors, we trained on out-of-equilibrium and equilibrium $(SA)_{1-2}(DMA)_{1-2}$ and $(NA)_{1-2}(AM)_{1-2}$ clusters and monomers and predicted on the out-of-equilibrium and equilibrium $(SA)_{1-2}(AM)_{1-2}$ and $(NA)_{1-2}(DMA)_{1-2}$ clusters separately. This yielded MAEs of 1.1 kcal/mol for the SA-AM system and 3 kcal/mol for the NA-DMA system, both of which are several times larger than the errors given in Figure 3.

3.3.2. Transferability to Other Molecules. The ML model trains and predicts the sum of atomic energy contributions to the overall cluster binding energy (see eq 1). Here, we examine the model transferability to systems with the same type of atoms but with different types of molecules. Although the acid−base interaction is in nature 'similar,' the binding energy differs, and this might give rise to high errors. Using the databases expanded by out-of-equilibrium structures, we trained our ML model on the $(SA)_{0-2}(base)_{0-2}$ clusters but also include all monomers and all base dimer clusters and predicted the binding energies of the other $(SA)_{1-2}(base')_{1-2}$ clusters. Figure 8 shows five graphs, where each is for a different base used in the training set and the MAEs of the modeled binding energies of other systems (see Figure S9 for the span of energies). Note that the inclusion of monomers in the training data set lowers the MAEs but essentially does not have an effect on the span of energies. The extrapolation from the SA−DMA clusters to the other SA−base clusters might be viable as the MAEs are the lowest with values below 2 kcal/mol. However, the span of errors ranges from 10 kcal/mol up to 25 kcal/mol. The rest of the systems are even worse. Unsurprisingly, the extrapolation from the weak base AM, which does not contain any methyl group, to the TMA monomer leads to the largest errors. For a larger system size test, we trained on the standardized Clusteromics I data set with all MA-containing clusters removed (except the monomer and homodimer) and predicted on the removed MA clusters. This test yielded a low MAE of 0.52 kcal/mol, showing that significantly lower errors can be reached if several similar molecules are added; however, the error is still a few times greater than if the MA-containing clusters were included, as seen in Figure 3. Based on these results, we do not recommend prediction on molecules outside the training set, as the models could yield large errors.

3.3.3. Extrapolation to Larger Cluster Sizes. Kubečka et al.[34] previously presented that training on equilibrium and out-of-equilibrium $(SA)_{0-2}(W)_{0-5}$ clusters allowed prediction on larger $(SA)_{4-7}(W)_{0-10}$ clusters. To test the capabilities of the model to extrapolate to larger cluster sizes, we trained on the standardized Clusteromics I set (i.e., $(SA)_{0-2}(base)_{0-2}$, where base refers to all possible base combinations) and the predicted binding energies of the 5119 equilibrium $(SA)_3(base)_3$ clusters from Xie et al.[55] and the 315 equilibrium $(SA)_4(base)_3$ and $(SA)_4(base)_4$ clusters from Kubečka et al.[44]

Figure 9 shows the ML binding energy errors for the $(SA)_3(base)_3$ and $(SA)_4(base)_{3,4}$ clusters. The prediction on $(SA)_3(base)_3$ clusters gives a low MAE of 0.8 kcal/mol with an acceptable span of errors. We found that the configurations with errors above 5 kcal/mol are high-energy conformers, and thus they are less relevant. However, extrapolation to larger $(SA)_4(base)_{3,4}$ clusters gives an MAE of 3.3 kcal/mol and errors up to 23 kcal/mol. Furthermore, the distribution shows a tail extending upward, showing that there is a type of interaction/structure the model cannot describe. By inspection of the structures with the highest error, we find that most of the structures contain EDA, and in several structures, EDA is double protonated and SA appears completely deprotonated (i.e., forming a sulfate ion). Furthermore, some of the $(SA)_4(base)_{3,4}$ clusters also contain direct EDA−EDA interactions, which are not described in the smaller Clusteromics I data set. This suggests that the model is not capable of extrapolating beyond one additional acid−base pair

https://doi.org/10.1021/acsomega.3c02203
ACS Omega 2023, 8, 25155−25164

for mixed-base clusters and that an iterative procedure where the clusters with one additional acid−base pair are predicted and calculated, expanded with MD, and included in the database might be needed, as suggested by Elm.[56] Unfortunately, such a procedure will quickly lead to large databases, which are slow or even impossible to train and predict with. Furthermore, testing such a procedure, though without expanding the $(SA)_3(base)_3$ with MD, barely decreases the span, as seen in Figure 9; however, it manages to almost halve the MAE from 3.2 to 1.8 kcal/mol. Another approach is to include only monomers/dimers and structures with one less acid−base pair below the current size without expanding with MD. To test this approach, we trained on all of the monomers/dimers and the $(SA)_3(base)_3$ equilibrium structures and predicted on the equilibrium $(SA)_4(base)_{3,4}$ clusters. This step-wise approach leads to a low MAE of 0.5 kcal/mol with the maximal error of 2 kcal/mol (see Figure 9). Such an approach would work as long as the equilibrium structures from the smaller sizes do not differ drastically from the target structures. In addition, for application in cluster configurational sampling, expansion with nonequilibrium structures will be required where pre-equilibrium extraction can be utilized to minimize the number of calculations necessary.

3.4. ML-Based Configurational Sampling. Based on the findings in the previous sections, we suggest inserting an ML step in the funneling workflow as

$$\text{ABCluster} \rightarrow \text{GFN1} - \text{xTB} \rightarrow \text{filter} \rightarrow \text{ML(GFN1} - \text{xTB to DFT)} \rightarrow \text{DFT opt}$$

This case is only applicable when the DFT method is slow or there is no available method with well-correlated energies and low computation cost. Based on Section 3.3.3, we suggest the iterative process illustrated in Figure 10, where the training

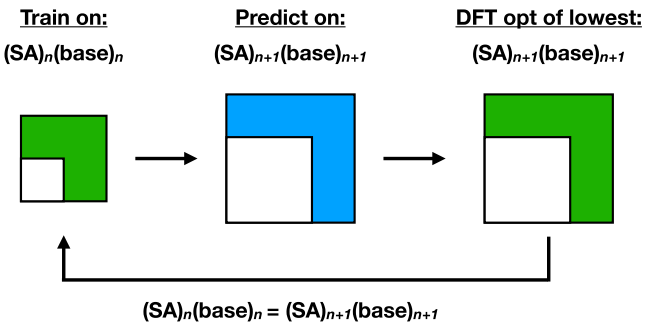


Figure 10. Illustration of the iterative ML-based funneling approach using the full ML model.

data are structures with one acid−base pair less $(SA)_n(base)_n$, monomers, and dimers and prediction is on $(SA)_{n+1}(base)_{n+1}$. The disadvantage of this method is the necessity to use the building-up approach for the clusters, i.e., sequentially building from smaller to larger clusters. Such an approach would, however, have a lower ML error, and computation-wise, the same amount of resources would be used. We expect that for larger clusters, the iterative step can go beyond one additional acid−base pair as the clusters become more similar at larger sizes.

4. CONCLUSIONS

We have created a large database of ∼250k atmospheric relevant structures for machine learning purposes. The database is based on the Clusteromics I−V $(acid)_{0-2}(base)_{0-2}$ equilibrium ωB97X-D/6-31++G(d,p) structures containing acids such as sulfuric acid, methanesulfonic acid, formic acid, and nitric acid, and with the bases ammonia, methylamine, dimethylamine, trimethylamine, and ethylenediamine. The equilibrium data was expanded with up to 10 out-of-equilibrium structures per equilibrium structure using GFN1-xTB MD trajectories, and all structures had their single-point energies calculated with r^2SCAN-3c. By testing the machine learning model kernel ridge regression with the FCHL19 representations, we find that the model can extrapolate to larger cluster sizes of one additional acid−base pair and transfer acid−acid and base−base interactions with mean absolute errors below 1 kcal/mol. Kernel ridge regression cannot extrapolate between different molecules even if they contain the same molecules as such an extrapolation yields a span of errors of ∼10−25 kcal/mol. We find that by parallelizing the kernel matrix construction, we can train on databases with up to ∼150k structures. We suggest introducing an iterative Δ-machine learning step in configurational sampling trained on the difference between the DFT level and the semiempirical level. The model should be trained on the monomers/dimers and structures with one less acid−base pair relative to the target size to yield the lowest mean absolute error.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c02203.

Visualization of the dissociation occurring for the PM3 and AM1 methods; figures of the MD energy trajectories for the $(SA)_1(AM)_1$, $(SA)_2(AM)_2$, $(SA)_1(DMA)_1$, and $(SA)_2(DMA)_2$ clusters; the settings for the MD trajectories; explanation of how the out-of-equilibrium data set test was done; figures of the Hybrid MD trajectories for the $(SA)_1(AM)_1$, $(SA)_2(AM)_2$, and $(SA)_1(DMA)_1$ clusters; figure showing the span of errors for the ML extrapolation of molecules; and location to the clusterome data set structures (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Jonas Elm − *Department of Chemistry, iClimate, Aarhus University, 8000 Aarhus C, Denmark;* orcid.org/0000-0003-3736-4329; Phone: +45 28938085; Email: jelm@chem.au.dk

Authors

Yosef Knattrup − *Department of Chemistry, Aarhus University, 8000 Aarhus C, Denmark;* orcid.org/0000-0003-3549-7494

Jakub Kubečka − *Department of Chemistry, Aarhus University, 8000 Aarhus C, Denmark*

Daniel Ayoubi − *Department of Chemistry, Aarhus University, 8000 Aarhus C, Denmark;* orcid.org/0000-0003-1972-6285

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c02203

Notes

The authors declare no competing financial interest.

https://doi.org/10.1021/acsomega.3c02203
ACS Omega 2023, 8, 25155−25164

■ ACKNOWLEDGMENTS

The authors thank the Independent Research Fund Denmark grant number 9064-00001B for financial support. The numerical results presented in this work were obtained at the Centre for Scientific Computing, Aarhus http://phys.au.dk/forskning/cscaa/

■ REFERENCES

(1) Canadell, J. G.; Monteiro, P. M. S.; Costa, M. H.; da Cunha, L. C.; Cox, P.; Eliseev, A. V.; Henson, S.; Ishii, M.; Jaccard, S.; Koven, C.et al. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*; Masson-Delmotte, V.; Zhai, P.; Pirani, A.; Connors, S. L.; Péan, C.; Berger, S.; Caud, N.; Chen, Y.; Goldfarb, L.; Gomis, M. I.et al., Eds.; Cambridge University Press: Cambridge, United Kingdom and New York, NY, USA, 2021; pp 673−816.

(2) Haywood, J.; Boucher, O. Estimates of the Direct and Indirect Radiative Forcing due to Tropospheric Aerosols: A Review. *Rev. Geophys.* **2000**, *38*, 513−543.

(3) Lohmann, U.; Feichter, J. Global indirect aerosol effects: A review. *Atmos. Phys. Chem* **2005**, *5*, 715−737.

(4) Merikanto, J.; Spracklen, D. V.; Mann, G. W.; Pickering, S. J.; Carslaw, K. S. Impact of nucleation on global CCN. *Atmos. Chem. Phys.* **2009**, *9*, 8601−8616.

(5) Kulmala, M.; Kontkanen, J.; Junninen, H.; Lehtipalo, K.; Manninen, H. E.; Nieminen, T.; Petäjä, T.; Sipilä, M.; Schobesberger, S.; Rantala, P.; et al. Direct Observations of Atmospheric Aerosol Nucleation. *Science* **2013**, *339*, 943−946.

(6) Tröstl, J.; Chuang, W.; Gordon, H.; Heinritzi, M.; Yan, C.; Molteni, U.; Ahlm, L.; Frege, C.; Bianchi, F.; Wagner, R.; et al. The role of low-volatility organic compounds in initial particle growth in the atmosphere. *Nature* **2016**, *533*, 527−531.

(7) Sipilää, M.; Berndt, T.; Petäjä, T.; Brus, D.; Vanhanen, J.; Stratmann, F.; Patokoski, J.; Mauldin, R. L.; Hyvärinen, A.-P.; Lihavainen, H.; et al. The Role of Sulfuric Acid in Atmospheric Nucleation. *Science* **2010**, *327*, 1243−1246.

(8) Kirkby, J.; Curtius, J.; Almeida, J.; Dunne, E.; Duplissy, J.; Ehrhart, S.; Franchin, A.; Gagne, S.; Ickes, L.; Kürten, A.; et al. Role of Sulphuric Acid, Ammonia and Galactic Cosmic Rays in Atmospheric Aerosol Nucleation. *Nature* **2011**, *476*, 429−433.

(9) Schobesberger, S.; Junninen, H.; Bianchi, F.; Lönn, G.; Ehn, M.; Lehtipalo, K.; Dommen, J.; Ehrhart, S.; Ortega, I. K.; Franchin, A.; et al. Molecular Understanding of Atmospheric Particle Formation from Sulfuric Acid and Large Oxidized Organic Molecules. *Proc. Natl. Acad. Sci. U.S.A.* **2013**, *110*, 17223−17228.

(10) Elm, J. Clusteromics I: Principles, Protocols and Applications to Sulfuric Acid - Base Cluster Formation. *ACS Omega* **2021**, *6*, 7804−7814.

(11) Elm, J. Clusteromics II: Methanesulfonic Acid-Base Cluster Formation. *ACS Omega* **2021**, *6*, 17035−17044.

(12) Elm, J. Clusteromics III: Acid Synergy in Sulfuric Acid-Methanesulfonic Acid-Base Cluster Formation. *ACS Omega* **2022**, *7*, 15206−15214.

(13) Knattrup, Y.; Elm, J. Clusteromics IV: The Role of Nitric Acid in Atmospheric Cluster Formation. *ACS Omega* **2022**, *7*, 31551−31560.

(14) Ayoubi, D.; Knattrup, Y.; Elm, J. Clusteromics V: Organic Enhanced Atmospheric Cluster Formation. *ACS Omega* **2023**, *8*, 9621−9629.

(15) Weber, R. J.; Marti, J. J.; McMurry, P. H.; Eisele, F. L.; Tanner, D. J.; Jefferson, A. Measured Atmospheric New Particle Formation Rates: Implications for Nucleation Mechanisms. *Chem. Eng. Comm.* **1996**, *151*, 53−64.

(16) Kurtén, T.; Loukonen, V.; Vehkamäki, H.; Kulmala, M. Amines are Likely to Enhance Neutral and Ion-induced Sulfuric Acid-water Nucleation in the Atmosphere More Effectively than Ammonia. *Atmos. Chem. Phys.* **2008**, *8*, 4095−4103.

(17) Loukonen, V.; Kurtén, T.; Ortega, I. K.; Vehkamäki, H.; Pádua, A. A. H.; Sellegri, K.; Kulmala, M. Enhancing Effect of Dimethylamine in Sulfuric Acid Nucleation in the Presence of Water - A Computational Study. *Atmos. Chem. Phys.* **2010**, *10*, 4961−4974.

(18) Nadykto, A. B.; Yu, F.; Jakovleva, M. V.; Herb, J.; Xu, Y. Amines in the Earthas Atmosphere: A Density Functional Theory Study of the Thermochemistry of Pre-Nucleation Clusters. *Entropy* **2011**, *13*, 554−569.

(19) Nadykto, A. B.; Herb, J.; Yu, F.; Xu, Y. Enhancement in the Production of Nucleating Clusters due to Dimethylamine and Large Uncertainties in the Thermochemistry of Amine-Enhanced Nucleation. *Chem. Phys. Lett.* **2014**, *609*, 42−49.

(20) Jen, C. N.; McMurry, P. H.; Hanson, D. R. Stabilization of Sulfuric acid Dimers by Ammonia, Methylamine, Dimethylamine, and Trimethylamine. *J Geophys. Res.: Atmos.* **2014**, *119*, 7502−7514.

(21) Nadykto, A. B.; Herb, J.; Yu, F.; Xu, Y.; Nazarenko, E. S. Estimating the Lower Limit of the Impact of Amines on Nucleation in the Earth's Atmosphere. *Entropy* **2015**, *17*, 2764−2780.

(22) Glasoe, W. A.; Volz, K.; Panta, B.; Freshour, N.; Bachman, R.; Hanson, D. R.; McMurry, P. H.; Jen, C. Sulfuric Acid Nucleation: An Experimental Study of the Effect of Seven Bases. *J. Geophys. Res.: Atmos.* **2015**, *120*, 1933−1950.

(23) Jen, C. N.; Bachman, R.; Zhao, J.; McMurry, P. H.; Hanson, D. R. Diamine-Sulfuric Acid Reactions are a Potent Source of New Particle Formation. *Geophys. Res. Lett.* **2016**, *43*, 867−873.

(24) Elm, J.; Jen, C. N.; Kurtén, T.; Vehkamäki, H. Strong Hydrogen Bonded Molecular Interactions between Atmospheric Diamines and Sulfuric Acid. *J. Phys. Chem. A* **2016**, *120*, 3693−3700.

(25) Elm, J.; Passananti, M.; Kurtén, T.; Vehkamäki, H. Diamines Can Initiate New Particle Formation in the Atmosphere. *J. Phys. Chem. A* **2017**, *121*, 6155−6164.

(26) Elm, J.; Kubečka, J.; Besel, V.; Jääskeläinen, M. J.; Halonen, R.; Kurtén, T.; Vehkamäki, H. Modeling the Formation and Growth of Atmospheric Molecular Clusters: A Review. *J. Aerosol Sci.* **2020**, *149*, No. 105621.

(27) Zhang, X.; Tan, S.; Chen, X.; Yin, S. Computational Chemistry of Cluster: Understanding the Mechanism of Atmospheric New Particle Formation at the Molecular Level. *Chemosphere* **2022**, *308*, No. 136109.

(28) Kubečka, J. Developing Efficient Configurational Sampling: Structure, Formation, and Stability of Atmospheric Molecular Clusters, Ph.D. Thesis; University of Helsinki: Finland, 2021.

(29) Dral, P. O. *Chemical Physics and Quantum Chemistry*; Ruud, K.; Brändas, E. J., Eds.; Academic Press, 2020; Vol. *81*, pp 291−324.

(30) Bender, A.; Schneider, N.; Segler, M.; Walters, W.; Engkvist, O.; Rodrigues, T. Evaluation guidelines for machine learning tools in the chemical sciences. *Nat. Rev. Chem.* **2022**, *6*, 428−442.

(31) Noé, F.; Tkatchenko, A.; Müller, K.-R.; Clementi, C. Machine Learning for Molecular Simulation. *Annu. Rev. Phys. Chem.* **2020**, *71*, 361−390.

(32) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; von Lilienfeld, O. A. Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *J. Chem. Theory Comput.* **2015**, *11*, 2087−2096.

(33) Jensen, A. B.; Kubečka, J.; Schmitz, G.; Christiansen, O.; Elm, J. Massive Assessment of the Binding Energies of Atmospheric Molecular Clusters. *J. Chem. Theory Comput.* **2022**, *18*, 7373−7383.

(34) Kubečka, J.; Rasmussen, F. R.; Christensen, A. S.; Elm, J. Quantum Machine Learning Approach for Studying Atmospheric Cluster Formation. *Environ. Sci. Technol. Lett* **2022**, *9*, 239−244.

(35) Smith, J. N.; Draper, D. C.; Chee, S.; Dam, M.; Glicker, H.; Myers, D.; Thomas, A. E.; Lawler, M. J.; Myllys, N. Atmospheric clusters to nanoparticles: Recent progress and challenges in closing the gap in chemical composition. *J. Aerosol Sci.* **2021**, *153*, No. 105733.

(36) Grimme, S.; Hansen, A.; Ehlert, S.; Mewes, J.-M. r2SCAN-3c: A "Swiss army knife" composite electronic-structure method. *J. Chem. Phys.* **2021**, *154*, No. 064103.

https://doi.org/10.1021/acsomega.3c02203
ACS Omega 2023, 8, 25155−25164

(37) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(38) Stewart, J. J. P. Optimization of parameters for semiempirical methods II. Applications. *J. Comput. Chem.* **1989**, *10*, 221−264.
(39) Grimme, S.; Bannwarth, C.; Shushkov, P. A Robust and Accurate Tight-Binding Quantum Chemical Method for Structures, Vibrational Frequencies, and Noncovalent Interactions of Large Molecular Systems Parametrized for All spd-Block Elements (Z = 1-86). *J. Chem. Theory Comput.* **2017**, *13*, 1989−2009.
(40) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB−An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652−1671.
(41) Neese, F. The ORCA program system. *WIREs Comput. Mol. Sci* **2012**, *2*, 73−78.
(42) Neese, F. Software update: The ORCA program systemVersion 5.0. *WIREs Comput. Mol. Sci.* **2022**, *12*, No. e1606.
(43) Christensen, A. S.; Bratholm, L. A.; Faber, F. A.; von Lilienfeld, O. A. FCHL revisited: Faster and more accurate quantum machine learning. *J. Chem. Phys.* **2020**, *152*, No. 044107.
(44) Kubečka, J.; Neefjes, I.; Besel, V.; Qiao, F.; Xie, H.-B.; Elm, J. Atmospheric Sulfuric Acid-Multi-Base New Particle Formation Revealed through Quantum Chemistry Enhanced by Machine Learning. *J. Phys. Chem. A* **2023**, *127*, 2091−2103.
(45) Christensen, A. S.; Faber, F. A.; Huang, B.; Bratholm, L. A.; Tkatchenko, A.; Muller, K. R.; von Lilienfeld, O. A. "QML: A Python Toolkit for Quantum Machine Learning", 2017, https://github.com/qmlcode/qml.
(46) Rasmussen, C. E.; Williams, C. K. I. *Gaussian Processes for Machine Learning*; The MIT Press, 2005.
(47) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **2017**, *8*, 3192−3203.
(48) Smith, J. S.; Isayev, O.; Roitberg, A. ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules. *Sci. Data* **2017**, *4*, No. 170193.
(49) Smith, J. S.; Zubatyuk, R.; Nebgen, B.; Lubbers, N.; Barros, K.; Roitberg, A.; Isayev, O.; Tretiak, S. The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Sci. Data* **2020**, *7*, No. 134.
(50) Devereux, C.; Smith, J. S.; Huddleston, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E. Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *J. Chem. Theory Comput.* **2020**, *16*, 4192−4202.
(51) Bhalla, S.; Melnekoff, D. T.; Aleman, A.; Leshchenko, V.; Restrepo, P.; Keats, J.; Onel, K.; Sawyer, J. R.; Madduri, D.; Richter, J.; et al. Patient similarity network of newly diagnosed multiple myeloma identifies patient subgroups with distinct genetic features and clinical implications. *Sci. Adv.* **2021**, *7*, No. eabg9551.
(52) Schütt, K.; Arbabzadah, F.; Chmiela, S.; Müller, K.-R.; Tkatchenko, A. Quantum-Chemical Insights from Deep Tensor Neural Networks. *Nat. Commun.* **2017**, *8*, No. 13890.
(53) Elm, J. An Atmospheric Cluster Database Consisting of Sulfuric Acid, Bases, Organics, and Water. *ACS Omega* **2019**, *4*, 10965−10974.
(54) Kubečka, J.; Knattrup, Y.; Engsvang, M.; Jensen, A. B.; Ayoubi, D.; Wu, H.; Christiansen, O.; Elm, J. Current and Future Machine Learning Approaches for Modelling Atmospheric Cluster Formation. *Nat. Comput. Sci.* **2023**, *3*, 495−503.
(55) Xie, H. B.; Elm, J. Tri-Base Synergy in Sulfuric Acid-Base Clusters. *Atmosphere* **2021**, *12*, No. 1260.
(56) Elm, J. Toward a Holistic Understanding of the Formation and Growth of Atmospheric Molecular Clusters: A Quantum Machine Learning Perspective. *J. Phys. Chem. A* **2021**, *125*, 895−902.

https://doi.org/10.1021/acsomega.3c02203
ACS Omega 2023, 8, 25155−25164