RSC Advances

PAPER

View Article Online
View Journal | View Issue

Cite this: *RSC Adv.*, 2024, **14**, 20048

Reparameterization of GFN1-xTB for atmospheric molecular clusters: applications to multi-acid–multi-base systems†

Yosef Knattrup, Jakub Kubečka, Haide Wu, Frank Jensen and Jonas Elm *

Atmospheric molecular clusters, the onset of secondary aerosol formation, are a major part of the current uncertainty in modern climate models. Quantum chemical (QC) methods are usually employed in a funneling approach to identify the lowest free energy cluster structures. However, the funneling approach highly depends on the accuracy of low-cost methods to ensure that important low-lying minima are not missed. Here we present a reparameterized GFN1-xTB model based on the clusteromics I–V datasets for studying atmospheric molecular clusters (AMC), denoted AMC-xTB. The AMC-xTB model reduces the mean of electronic binding energy errors from 7–11.8 kcal mol^{-1} to roughly 0 kcal mol^{-1} and the root mean square deviation from 7.6–12.3 kcal mol^{-1} to 0.81–1.45 kcal mol^{-1}. In addition, the minimum structures obtained with AMC-xTB are closer to the ωB97X-D/6-31++G(d,p) level of theory compared to GFN1-xTB. We employ the new parameterization in two new configurational sampling workflows that include an additional meta-dynamics sampling step using CREST with the AMC-xTB model. The first workflow, denoted the "independent workflow", is a commonly used funneling approach with an additional CREST step, and the second, the "improvement workflow", is where the best configuration currently known in the literature is improved with a CREST + AMC-xTB step. Testing the new workflow we find configurations lower in free energy for all the literature clusters with the largest improvement being up to 21 kcal mol^{-1}. Lastly, by employing the improvement workflow we massively screened 288 new multi-acid–multi-base clusters containing up to 8 different species. For these new multi-acid–multi-base cluster systems we observe that the improvement workflow finds configurations lower in free energy for 245 out of 288 (85.1%) cluster structures. Most of the improvements are within 2 kcal mol^{-1}, but we see improvements up to 8.3 kcal mol^{-1}. Hence, we can recommend this new workflow based on the AMC-xTB model for future studies on atmospheric molecular clusters.

Received 23rd April 2024
Accepted 16th June 2024

DOI: 10.1039/d4ra03021d

rsc.li/rsc-advances

1 Introduction

Molecular clusters, formed through the aggregation of various atmospheric species, play a central role in aerosol particle formation.[1] Aerosols are liquid or solid fine particles suspended in air that can act as cloud condensation nuclei (CCN) if they reach sizes at or above 50–100 nm.[2] Roughly 50% of CCN are believed to be initially formed as clusters.[3] CCN acts as nucleation cores for water uptake and then further growth into clouds meaning there is a direct correlation between aerosols and cloud number/properties and hence the climate. The biggest uncertainty in modern climate forcing predictions is due to the uncertainties from aerosol–cloud interactions.[1]

Sulfuric acid has been shown to be the main driver of cluster formation. Other key species are believed to be, bases (ammonia and amines), acids (methanesulfonic acid, nitric acid, iodine acids or organic acids), highly oxygenated organic molecules and water.[4–7] It is extremely difficult to experimentally measure the composition and formation mechanism of the initial clusters due to their small size and neutral charge. Mass spectrometry techniques can measure the cluster compositions of the charged cluster, however, it is unknown if the ionization of neutral clusters significantly changes the cluster composition/structure and it is also believed that fragmentation may happen in the instruments.[8–10] This leaves theoretical studies as the only way to elucidate the thermodynamics, kinetics, and molecular interactions governing cluster formation and its evolution. The main challenge for studying atmospheric molecular clusters is their complex configurational spaces, which require advanced configurational sampling techniques and computationally demanding quantum chemistry methods to evaluate the cluster properties

Department of Chemistry, Aarhus University, Langelandsgade 140, Aarhus C, 8000, Denmark. E-mail: jelm@chem.au.dk; Tel: +45 28938085

† Electronic supplementary information (ESI) available: Figure showing the conformer index at the AMC-xTB level of theory out of the 50 conformers optimized at the DFT level. Figure showing the comparison of the lowest free energy conformer found by the improvement configurational workflow compared to the original workflow if only 10 conformers are selected. Guide to downloading the AMC-xTB parameter file and the structures/calculations generated in this work. See DOI: https://doi.org/10.1039/d4ra03021d

accurately.[5] Furthermore, atmospheric clustering is believed to be a multi-species process,[11] adding another dimension of chemical complexity.

Thoroughly exploring the configurational space of atmospheric molecular clusters using, for instance, metadynamics simulations[12,13] or genetic algorithms[14–16] at a high level of theory is extremely computationally demanding. Hence, usually, a funneling approach[5,7,17] is applied, where the configurational space is initially explored at a low level of theory such as force-field or semiempirical methods, and only a subset of low energy structures is selected, reoptimized, and reexamined at a higher level of theory. This process is repeated with an increasing level of theory until only a few structures remain for evaluation at the desired high level. Schematically, the process can be given as:

(1) Generate initial cluster configurations:

ABCluster/OGOLEM/Basin hopping or similar.

(2) Semi-empirical calculations:

Optimization at the PM6/PM7/GFN1-xTB/GFN2-xTB or similar level.

(3) DFT calculations:

DFT optimization and vibrational frequency calculations.

(4) Single point energy refinement:

Single-point energy calculation at coupled cluster level on the DFT optimized geometry.

Between each step in the funneling approach, filtering can be applied to reduce the number of structures that need to be handled. This can either be based on an energy threshold or a set number of cluster structures. Eventually, we end up with a handful of structures at the highest obtainable level.

The first step in the funneling procedure is the generation of a large number of configuration candidates. The key idea is sampling a large part of the potential energy surfaces at a low level of theory to get estimates for the global free energy minimum. This is usually carried out using force-field methods in combination with genetic algorithms such as in ABCluster[15,16,18–21] and OGOLEM,[14,22–27] by random/manual sampling or using dynamic methods such as basin hopping.[28–32] The major issue at this step is that most force-field methods are unable to describe bond-breaking, such as proton transfer reactions, which are important for atmospherically relevant molecular clusters, requiring the sampling to include monomers where the hydrogens have been transferred to get adequate sampling. Furthermore, the accuracy of force-field methods is also insufficient to determine a subsample of the conformer candidates and all the candidates have to be taken to the higher level of theory.

The next step is semi-empirical calculations as these are a better description of the chemistry and filtering can be applied. Of the common semiempirical methods, GFN1-xTB,[33] GFN2-xTB,[34] PM6,[35] and PM7 [36] are the most used in configurational sampling procedures for atmospheric molecular clusters.[5–7] GFN stands for geometries, frequencies and non-covalent interactions, which are the main target properties for the method. PM stands for parameterization method indicating the model version. Of these methods, GFN1-xTB has shown to have the highest correlation with electronic binding energies at a higher level of theories[37,38] and have been shown to have a higher correlation with DFT trajectories for molecular dynamics than GFN2-xTB.[39] The reason GFN2-xTB performs worse than GFN1-xTB for atmospheric molecular clusters (often involving sulfuric acid) is that there is a decrease in the number of d-functions for sulfur in the basis set for the newer GFN2-xTB model.

The third step is the subsequent optimization and vibrational frequency calculation of the structures with DFT. This is the main bottleneck in the sampling methodology as limited computational resources only allow a fixed number of DFT structures to be optimized. Therefore some form of filtering is required, often based on structural properties or electronic energies from the semiempirical calculations. To circumvent the inaccuracies of semiempirical methods an intermediate step can be included, involving single-point energy calculations at the DFT level on as many structures as possible. Another option for an intermediate step is the utilization of machine learning (ML) methods. One can calculate a subset of the structures at a desired DFT level and train an ML model to predict the energies of the remaining structures.[39,40] However, to mimic accurate DFT energies, kernel-based ML methods become computationally demanding[40–42] and neural-networks will require an extensive set of training data and hyper-parameter optimization.[43] Moreover, ML methods often fail when predicting on structures different from the training set.

Overall, the funneling approach is never more efficient than its weakest link given by the semiempirical step in 2, in which accuracy determines the number of structures that have to be optimized/have single points calculated at the DFT level. In this paper, we focus on reparameterizing the GFN1-xTB method based on DFT energies of atmospherically relevant molecular clusters yielding a GFN1-xTB model reparametrized based on ωB97X-D/6-31++G(d,p) for 'atmospheric molecular clusters' denoted AMC-xTB. This new parameterization is used to sample 288 large multi-acid–multi-base clusters containing AM equivalent to the clusters studied by Knattrup *et al.*[44]

2 Methodology

2.1 Computational details

Single-point energies, gradients, and geometries for the reparameterization, configurational sampling, and comparisons were calculated using the xtb 6.4.0 program using the GFN1-xTB[33] and AMC-xTB parameterizations. A modified version of ArbAlign[45] available in the JKCS program[46] was used to calculate the root-mean-square differences (RMSD) between molecular structures. Gaussian 16, version B.01 [47] was used for the DFT calculations. CREST 2.12 [12,13] with an energy window of 15 kcal mol^{-1} and in noncovalent interaction mode and ABCluster 2.0 [15,16] with a population of SN = 3000, number of generations of $g_{max} = 200$, and gen. survival of $g_{limit} = 4$ were used for additional configurational sampling.

2.2 Cluster data sets

For reparameterization of GFN1-xTB, we used the clusteromics I–V data sets[48–52] containing $(acid)_{0-2}(base)_{0-2}$ clusters of the following atmospherically relevant species: sulfuric acid (SA), methanesulfonic acid (MSA), nitric acid (NA), formic acid (FA),

ammonia (AM), methylamine (MA), dimethylamine (DMA), trimethylamine (TMA) and ethylenediamine (EDA). All structures are optimized at the ωB97X-D/6-31++G(d,p) level of theory, as benchmark studies[37,53] show this to be a good compromise between accuracy and speed, and we used up to the three lowest electronic energy configurations found per each cluster as the optimization set for GFN1-xTB reparametrization. This leads to an optimization set comprising of a total of 1073 clusters. The GFN1-xTB reparameterization based on this optimization set will be denoted as the AMC-xTB model.

All new data calculated is freely available in the Atmospheric Cluster DataBase[54] along with the new AMC-xTB parameter file (see Section S2).†

2.3 Optimization strategy

The GFN1-xTB model contains 15 global parameters, 2 element-pair-specific parameters, and 32 element-specific parameters of relevance to the species present in the optimization sets (H, C, N, O and S atoms). Initially, the Hessian was generated to probe the sensitivity of the parameters, however, we found it computationally feasible to employ a similar optimization strategy to the original GFN1-xTB paper,[33] where we optimize all relevant parameters simultaneously. We utilize a modified version of an in-house pseudo-Newton–Raphson optimizer by Jensen *et al.*[55] for the optimization of a target function (T) containing a linear combination of the difference in electronic binding energies (ΔE^{b}) in kcal mol^{-1} and gradient norms (g^{norm}) in hartree bohr^{-1} radius at the current GFN1-xTB parameterization and the target ωB97X-D/6-31++G(d,p) level of theory:

$$T = \frac{1}{N_{\mathrm{conf}}} \sum_{i}^{N_{\mathrm{conf}}} \left(\frac{\Delta E_i^{\mathrm{b}}}{N_i^{\mathrm{atoms}}} + g_i^{\mathrm{norm}} \right). \quad (1)$$

here, N_{conf} is the total number of structures in the optimization set, N_i^{atoms} is the number of atoms in the i-th structure. We normalized by the number of atoms in each structure to prevent "overfitting" to the larger clusters.

We chose the electronic binding energies as the target properties to get a better tool for filtering based on energies in configurational sampling procedures. The gradients were included directly in the target function. We use equilibrium structures at the given level of theory, which are supposed to have near-zero gradients. However, the upper limit is set by the accuracy threshold within the xTB program during the optimization, which makes the gradients non-zero in the calculations.

Including only the electronic bindings energies in the target function yields a much better fit for the energies but causes the gradients to "explode", effectively rendering the optimization functionality of AMC-xTB useless for our target species. Giving higher weight to the gradient norms in the target function makes the optimized structures more similar to the target ωB97X-D/6-31++G(d,p) level of theory, however, we found that it causes problems in the configurational sampling procedure where the decreased accuracy in determining the binding energies causes our configurational sampling to yield high-energy conformers at the DFT level. Overall, we found that including the gradient norms and difference in electronic bindings energies in a 1 : 1 ratio as the best compromise between the two properties.

2.4 Updated configurational sampling workflows

The strength of the new AMC-xTB model is that it can be used directly in configurational sampling programs such as ABCluster and CREST. Here, we will test two different new workflows for applying the reparameterized models in cluster configurational sampling.

2.4.1 Original workflow. The workflow usually employed in studying atmospheric molecular clusters can be summarized as:

$$\mathrm{ABCluster} \rightarrow \mathrm{GFN1\text{-}xTB}_{\mathrm{all}}^{\mathrm{opt}} \rightarrow \mathrm{DFT}_{N\ \mathrm{lowest}}^{\mathrm{opt}}$$

here the number of configurations N that have to be optimized at the DFT level is a severe bottleneck in the number of new cluster systems that can be studied. Usually, 50–100 configurations are optimized at the DFT level.

2.4.2 Independent workflow. The independent workflow refers to configurational sampling from scratch using the well-established funneling approach using AMC-xTB instead of GFN1-xTB.[5,17] As the aim for the approach is to be generally applicable, we also included an additional CREST step, as it should be better at handling flexible organic compounds:

$$\mathrm{ABCluster} \rightarrow \mathrm{AMC\text{-}xTB}_{\mathrm{all}}^{\mathrm{opt}} \rightarrow \mathrm{CREST(AMC\text{-}xTB)}_{1\ \mathrm{lowest}} \rightarrow \mathrm{DFT}_{50\ \mathrm{lowest}}^{\mathrm{opt}}$$

here, ABCluster, a genetic algorithm for sampling clusters, is used for the initial sampling of all possible neutral/ionic combinations of monomers that yield overall neutral clusters. The xTB 6.4.0 program was then used to optimize all the configurations at the AMC-xTB level. The cluster lowest in electronic energy was then taken as the input structure for CREST in non-covalent interaction mode, again using our new AMC-xTB model. The initial ABCluster sampling is needed because we found CREST to be quite sensitive to the input structure and, therefore, needs a good guess for a starting structure. The 50 structures lowest in electronic energy are then optimized at the DFT level.

2.4.3 Improvement workflow. The improvement workflow refers to using the best structure currently known at the corresponding level of theory as the input structure for CREST using the AMC-xTB model.

$$\mathrm{Best\ structure} \rightarrow \mathrm{CREST(AMC\text{-}xTB)} \rightarrow \mathrm{DFT}_{50\ \mathrm{lowest}}^{\mathrm{opt}}$$

from here on, the workflow is the same as the independent workflow, where the 50 structures lowest in electronic energy are optimized at the corresponding DFT level.

3 Results and discussion

3.1 Extension of the multi-acid–multi-base dataset

To gain a more complete test set we extended the multi-acid–multi-base clusters systems by Knattrup *et al.*[44] using the same workflow for a total of 288 new AM-containing clusters. With the acids being SA, MSA, FA and NA and the bases being AM, MA,

DMA and TMA. This is the first sampling of multi-component clusters containing up to 8 different species yielding a data set where synergistic effects in cluster formation between different species of bases[41] and acids[44] and mixes thereof can be studied. Such clusters could be relevant for modeling polluted coastal environments. Fig. 1 presents the molecular structure of a newly sampled 8-component cluster. It is seen that all the acids have transferred a proton to all the bases.

The initial sampling yields binding free energies ranging from −28.43 kcal mol^{-1} [$(MSA)_1(NA)_1(FA)_1(AM)_2$] to −104.0 kcal mol^{-1} [$(SA)_3(NA)_1(AM)_1(MA)_1(DMA)_2$] for the cluster configurations lowest in free energy at the ωB97X-D/6-31++G(d,p) level of theory.

3.2 Assessment of the AMC-xTB binding energies

We reparameterized GFN1-xTB to obtain a new tight-binding semiempirical reparameterization denoted as AMC-xTB. Fig. 2 shows the error in electronic binding energies before (GFN1-xTB) and after (AMC-xTB) reparameterization. The models have been tested on the entire clusteromics I–V[48–52] data sets (56 436 data points), the sulfuric acid–multi-base $(SA)_{1-4}(AM/MA/DMA/TMA/EDA)_{1-4}$ cluster data set (684 data points) by Kubečka *et al.*[41] and the multi-acid–muti-base $(SA/FA/MSA/NA)_{1-4}(MA/DMA/TMA)_{1-4}$ by Knattrup *et al.*[44] including the new AM-containing clusters (1629 data points). All the tested structures are equilibrium structures at the ωB97X-D/6-31++G(d,p) level of theory. Although the Gaussian version and integration grid used for optimization differ for some structures, it was found to have a negligible effect on this comparison as we are studying the binding energies and not the absolute energies.

For all the data sets shown in Fig. 2 the reparameterization results in near-zero means of the energy errors. This is a reduction from error means of 3.7–11.8 kcal mol^{-1} for GFN1-xTB. In addition, the AMC-xTB model achieves a more narrow error distribution with the root mean square deviations decreasing from 7.6–12.3 kcal mol^{-1} to 0.81–1.45 kcal mol^{-1}, implying that there will be fewer outliers. The error span on the larger clusters for the Knattrup *et al.*[44] and Kubečka *et al.*[41] sets are similar to the error span for the smaller clusters in the optimization set. This shows that reparameterizing on smaller clusters is adequate for calculations on larger-sized clusters as the model gets some of the underlying chemistry correct and scales effectively with system size. The new AMC-xTB model reduces the number of structures needed to pass from the semiempirical step to the DFT step in configurational sampling. For atmospheric molecular clusters, this implies that the AMC-xTB model is unequivocally better to apply in the configurational sampling funneling workflow compared to GFN1-xTB.

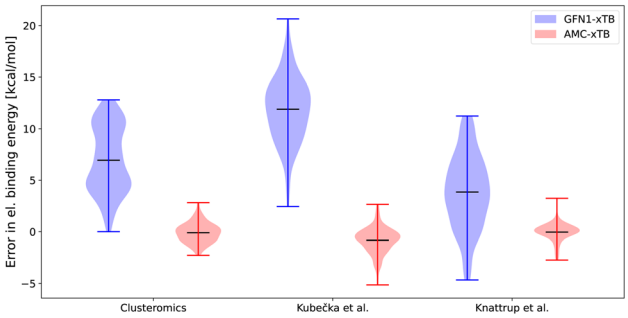
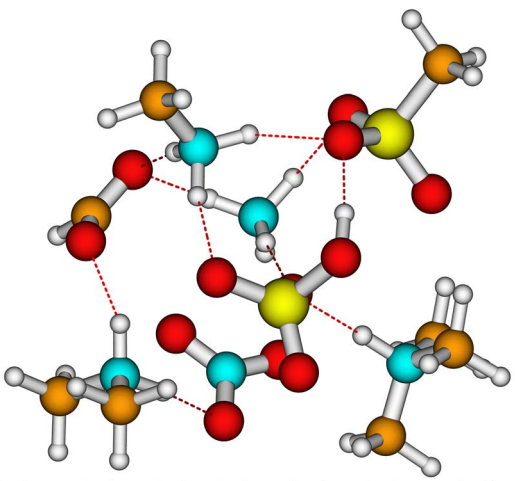


Fig. 2 Error in the electronic binding energies for the GFN1-xTB and AMC-xTB methods compared with the ωB97X-D/6-31++G(d,p) level of theory. The clusteromics I–V[48–52] sets have $(SA/FA/MSA/NA)_{0-2}(AM/MA/DMA/TMA/EDA)_{0-2}$ clusters, the Kubečka *et al.*[41] set has sulfuric acid–multi-base $(SA)_{1-4}(AM/MA/DMA/TMA/EDA)_{1-4}$ clusters, Knattrup *et al.*[44] set has the multi-acid–muti-base $(SA/FA/MSA/NA)_{1-4}(MA/DMA/TMA)_{1-4}$ clusters including the new AM-containing clusters sampled in this work.



$(SA)_1(MSA)_1(NA)_1(FA)_1(AM)_1(MA)_1(DMA)_1(TMA)_1$

Fig. 1 The $(SA)_1(MSA)_1(NA)_1(FA)_1(AM)_1(MA)_1(DMA)_1(TMA)_1$ cluster structure lowest in Gibbs free energy at the ωB97X-D/6-31G++(d,p) level of theory with the quasi-harmonic approximation (cutoff of 100 cm^{-1}) for the initial sampling. Yellow = sulfur, red = oxygen, cyan = nitrogen, brown = carbon and white = hydrogen.

3.3 Assessment of the AMC-xTB geometries and gradients

The gradient norms were also a part of the optimization scheme (eqn (1)). Fig. 3 shows the gradient norms given by the xtb program for the two parameterizations. The structures are equilibrium structures at the ωB97X-D/6-31++G(d,p) level of theory, so the ideal gradient norms should be below the default gradient convergence thresholds of 10^{-3} E_h/α. None of the methods manages to be below this threshold, but the AMC-xTB model is close. This does not directly mean the model is closer to the correct structure, as the new parameters might just have flattened the potential energy surface at this point without moving closer to the minimum. However, including the gradients in the target function avoids numerical instability, and yields reasonable optimized structures. To test if the structures are closer to a minimum at the DFT level, the initial DFT structures of all three datasets were optimized using the different parameterizations, and the RMSD was computed between the initial DFT structure and the GFN1-xTB/AMC-xTB optimized structures (see Table 1).

We find that the AMC-xTB model reduces the mean RMSD of the full clusteromics set from 0.484 Å to 0.355 Å and a similar reduction is seen for the Knattrup *et al.*[44] and Kubečka *et al.*[41]

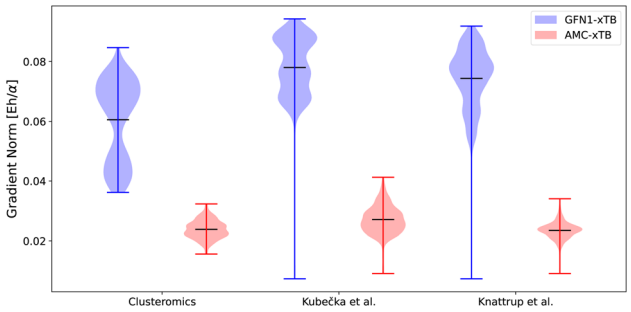


Fig. 3 The gradient norms for the GFN1-xTB and AMC-xTB methods. α is the Bohr radius. The clusteromics[48–52] set is $(SA/FA/MSA/NA)_{0-2}(AM/MA/DMA/TMA/EDA)_{0-2}$ clusters, the Kubečka *et al.*[41] set is sulfuric acid–multi-base $(SA)_{1-4}(AM/MA/DMA/TMA/EDA)_{1-4}$ clusters, Knattrup *et al.*[44] is the multi-acid–muti-base $(SA/FA/MSA/NA)_{1-4}(MA/DMA/TMA)_{1-4}$ clusters including the new AM-containing clusters sampled in this work. The structures are equilibrium structures at the ωB97X-D/6-31++G(d,p) level of theory.

Table 1 Comparison of the mean, median, and standard deviation (std) of the root-mean-squared differences (RMSD) between the initial DFT structure and the optimized structure at the given parameterization. The clusteromics I–V[48–52] sets includes the $(SA/FA/MSA/NA)_{0-2}(AM/MA/DMA/TMA/EDA)_{0-2}$ clusters, the Kubečka *et al.*[41] set comprise the sulfuric acid–multi-base $(SA)_{1-4}(AM/MA/DMA/TMA/EDA)_{1-4}$ clusters and Knattrup *et al.*[44] set has the multi-acid–muti-base $(SA/FA/MSA/NA)_{1-4}(MA/DMA/TMA)_{1-4}$ clusters including the new AM-containing clusters sampled in this work. The lowest errors are shown in bold

Method/data set	Mean	Median	Std
GFN1-xTB/clusteromics	0.484	0.387	0.330
AMC-xTB/clusteromics	**0.355**	**0.242**	**0.282**
GFN1-xTB/Knattrup *et al.*	0.378	0.367	0.138
AMC-xTB/Knattrup *et al.*	**0.235**	**0.193**	**0.125**
GFN1-xTB/Kubečka *et al.*	0.376	0.361	0.140
AMC-xTB/Kubečka *et al.*	**0.189**	**0.164**	**0.094**

sets with RMSDs being reduced from 0.378 Å to 0.235 Å and from 0.376 Å to 0.189 Å, respectively. This, coupled with the smaller gradients, suggests that the reparameterized model is closer to a minimum at the DFT level. This implies that the preoptimization step in a funneling approach with the AMC-xTB model compared to GFN1-xTB yields structures closer to the DFT structure and will likely reduce the subsequent optimization time at the DFT level.

3.4 Test of new configurational sampling workflows

To further test how the new AMC-xTB model fares in cluster configurational sampling, we tested the independent and improvement workflow for several previously studied $(acid)_4(base)_4$ cluster systems. Hence, the workflow is tested on clusters up to twice the size of those used in the reparameterization.

Fig. 4 shows the difference in binding free energy for the lowest free-energy configuration found by employing the independent and improvement workflows for the $(SA)_4(EDA)_4$, $(SA)_4(AM)_4$, and $(SA)_4(AM)_1(MA)_1(DMA)_1(TMA)_1$ clusters

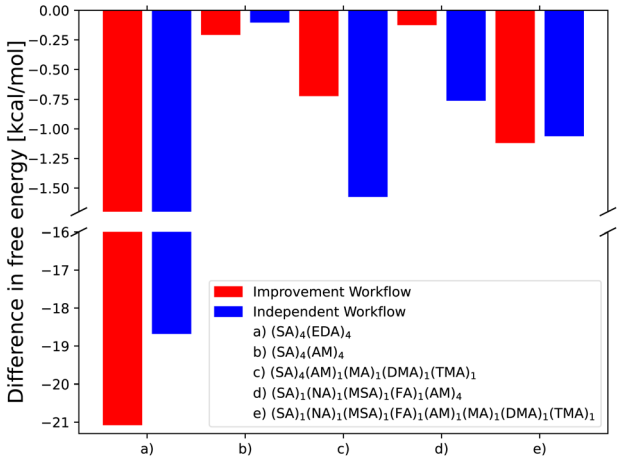


Fig. 4 Comparison of the lowest free energy conformer found by the independent and improvement configurational workflows compared to the configurations found by Kubečka *et al.*[41] (a–c) and two new multi-component AM clusters sampled in the same way as Knattrup *et al.*[44] (d and e). Gibbs free energies are calculated at the ωB97X-D/6-31++G(d,p) level of theory with the quasi-harmonic approximation (cutoff of 100 cm^{-1}) and vib. frequencies scaled by 0.996 in accordance with Kubečka *et al.*[41]

compared to Kubečka *et al.*[41] and the $(SA)_1(MSA)_1(NA)_1(FA)_1(AM)_4$ and the new $(SA)_1(MSA)_1(NA)_1(FA)_1(AM)_1(MA)_1(DMA)_1(TMA)_1$ clusters sampled in this work. The SA–AM clusters have been extensively studied previously[6,56–61] and are therefore believed to be well-sampled using the original configurational sampling procedure and thereby difficult to improve. Still, the new CREST + AMC-xTB methodology manages to find cluster structures lower in free energy by 0.21 kcal mol^{-1} compared to the previous works.

In the case of the $(SA)_4(AM)_4$ and $(SA)_1(MSA)_1(NA)_1(FA)_1(AM)_1(MA)_1(DMA)_1(TMA)_1$ clusters, the independent/improvement workflows perform similar and yield similar free energy improvements. However, for the $(SA)_4(AM)_1(MA)_1(DMA)_1(TMA)_1$ clusters, the independent workflow works slightly better, finding a cluster 0.85 kcal mol^{-1} lower in free energy compared to the improvement workflow. This illustrates that the sampling is very sensitive to the configuration used as input for CREST, although it might also be due to the randomness of the dynamic processes in CREST. The reason for the difference might be that the original work's configurational sampling was worse than the independent workflow, yielding a worse starting structure for the CREST sampling within the improvement workflow. We see a massive improvement in the configurational sampling of the $(SA)_4(EDA)_4$ clusters by −18/−21 kcal mol^{-1}. This is caused by the flexibility of the EDA molecule, as it is the only monomer that contains a C–C bond it can rotate around, making it difficult to sample the full configurational space using only ABCluster with rigid molecules. This improvement should primarily be attributed to the inclusion of metadynamics sampling in CREST and not purely the parameterization of AMC-xTB as it allows the EDA to rotate around its bonds and find a structure with more/better paired intermolecular interactions as seen in Fig. 5. It should also be noted that the main

improvements are the electronic binding energy and the thermal contribution varies very little between the clusters.

However, this shows the strength of the presented workflows as they can be used for clusters containing more flexible organic molecules.

3.5 Massive improvement test

Based on the previous sections, it is clear that the improvement workflow could locate more stable clusters. As the potential energy surface of multi-acid–multi-base clusters becomes very complicated, here we test this new approach for such systems. These 288 new AM-containing clusters were used as a massive test set for the improvement workflow using the newly parameterized AMC-xTB model. The improvement workflow manages to find configurations lower in free energy at the ωB97X-D/6-31++G(d,p) for 245 out of the 288 clusters (85.1%) as can be seen in Fig. 6.

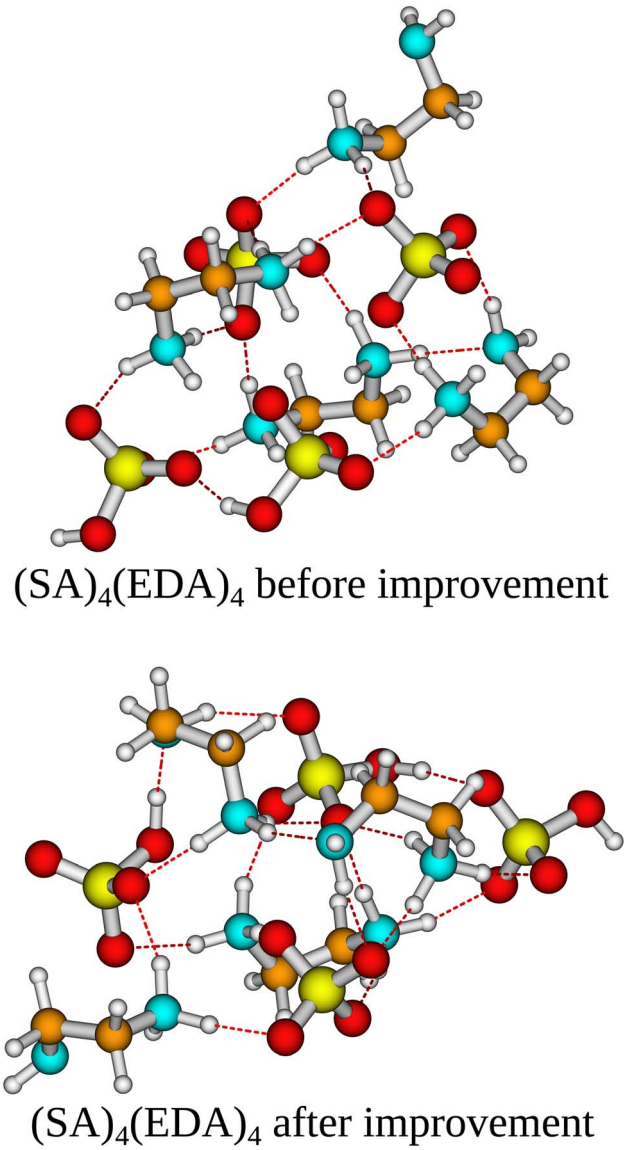


Fig. 5 The $(SA)_4(EDA)_4$ cluster structure lowest in Gibbs free energy at the ωB97X-D/6-31G++(d,p) level of theory with the quasi-harmonic approximation (cutoff of 100 cm^{-1}) before and after the improvement workflow. Yellow = sulfur, red = oxygen, cyan = nitrogen, brown = carbon, and white = hydrogen.

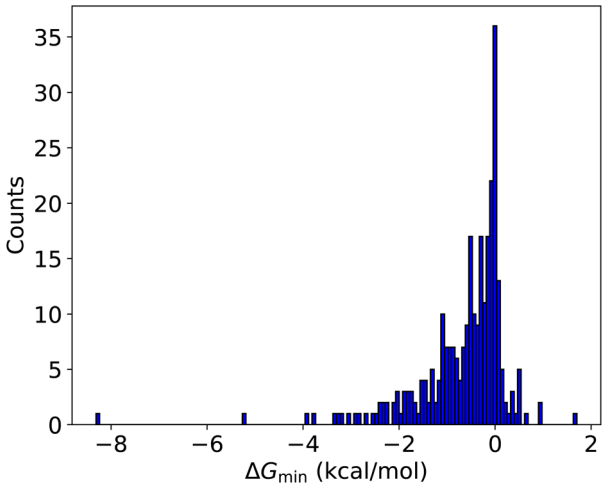


Fig. 6 Comparison of the lowest free energy conformer found by the improvement configurational workflow compared to the original workflow. Gibbs free energies are calculated at the ωB97X-D/6-31++G(d,p) level of theory with the quasi-harmonic approximation (cutoff of 100 cm^{-1}).

In most cases, the improvement is between 0–2 kcal mol^{-1}. However, for the $(SA)_1(MSA)_1(NA)_1(FA)_1(AM)_1(DMA)_2(TMA)_1$, $(SA)_3(NA)_1(AM)_1(MA)_3$, and $(SA)_3(FA)_1(AM)_1(MA)_1(DMA)_2$ clusters a massive improvement of 8.3, 5.2 and 3.9 kcal mol^{-1} is observed, respectively.

Comparing the conformer index at the AMC-xTB level of theory with the final ωB97X-D/6-31++G(d,p) level of theory with the quasi-harmonic approximation (cutoff of 100 cm^{-1}) the Gibbs free energy minimum at the DFT level is also the electronic energy minimum at the AMC-xTB level of theory for 66 of the clusters (see Fig. S1†). If 10 conformers are included from the AMC-xTB level of theory, the free energy minimum energy is captured for 155 out of the 288 clusters with improvements found for 209 (see Fig. S2†). Furthermore, the maximum error is 2 kcal mol^{-1} with a mean of 0.12 kcal mol^{-1} when reducing from 50 to 10 conformers.

This highlights the need for including dynamics-based sampling procedures for atmospheric clusters even though the system might seem fairly rigid. It can also be envisioned that the improvement workflow will be quite important when studying much larger $(SA)_{1-20}(base)_{1-20}$ clusters as recently done by Engsvang *et al.*[42,62] and Wu *et al.*[38] For large clusters, the global minimum is tricky to locate, and adding dynamics-based configurational sampling might aid in the process.

4 Conclusions

We have reparameterized the GFN1-xTB model to yield better binding electronic energies and gradient norms for atmospherically relevant clusters composed of the following species: sulfuric acid (SA), methanesulfonic acid (MSA), nitric acid (NA), formic acid (FA), ammonia (AM), methylamine (MA), dimethylamine (DMA), trimethylamine (TMA), and ethylenediamine

(EDA). The reparameterization, denoted AMC-xTB, for use in the xtb/CREST program, is based on the ωB97X-D/6-31++G(d,p) level of theory. The model shows a substantial decrease in the error of the binding electronic energies compared to the original GFN1-xTB parameterization and the gradient norms of the equilibrium structures are closer ωB97X-D/6-31++G(d,p) level of theory compared to GFN1-xTB. The reparameterization strategy is general and can be used to reparameterize other methods such as GFN2-xTB.

We tested two new configurational sampling procedures with the new parameterizations being employed in the xTB and CREST programs. The first workflow, denoted as "improvement workflow," is based on improving the best structure currently known in the literature with CREST and then doing the DFT calculations. The second workflow, denoted the "independent workflow," starts by configurational sampling using ABCluster, followed by xtb, CREST, and then DFT. Using the two workflows, we find cluster structures lower in free energy for the following $(SA)_4(EDA)_4$, $(SA)_4(AM)_4$,$(SA)_4(AM)_1(MA)_1(DMA)_1(TMA)_1$, $(SA)_1(MSA)_1(NA)_1(FA)_1(AM)_4$ and $(SA)_1(MSA)_1(NA)_1(FA)_1(AM)_1(MA)_1(DMA)_1(TMA)_1$ systems in all cases compared to the best-known value in the literature.

Testing the improvement workflow on 288 large multi-acid–multi-base cluster systems, the workflow finds improvements for 85.1% of the clusters, showing the need for dynamics-based sampling.

The parameterization strategy given here is not specific to either GFN1-xTB or atmospheric clusters and could be used in general. For instance, one could imagine increasing the number of d-functions in the basis set for sulfur atoms in GFN2-xTB and then reparameterizing the new GFN2-xTB model or doing a reparameterization for much larger clusters.

Author contributions

Conceptualization: J. E.; methodology: Y. K., J. K., H. W., F. J. and J. E.; software: F. J.; formal analysis: Y. K. and J. K.; investigation: Y. K., J. K. and H. W.; resources: J. E.; writing – original draft: Y. K., J. K. and J. E.; writing – review & editing: Y. K., J. K., H. W., F. J. and J. E.; visualization: Y. K.; project administration: J. E.; funding acquisition: J. E.; supervision: F. J. and J. E.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

Funded by the European Union (ERC, ExploreFNP, project 101040353 and MSCA, HYDRO-CLUSTER, project 101105506). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union, the European Research Executive Agency, or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. The authors thank the Independent Research Fund Denmark grant number 9064-00001B and for financial support. This work was funded by the Danish National Research Foundation (DNRF172) through the Center of Excellence for Chemistry of Clouds. The numerical results presented in this work were obtained at the Centre for Scientific Computing, Aarhus https://phys.au.dk/forskning/faciliteter/cscaa/.

Notes and references

1 J. G. Canadell, P. M. S. Monteiro, M. H. Costa, L. Cotrim da Cunha, P. Cox, A. V. Eliseev, S. Henson, M. Ishii, S. Jaccard, C. Koven and *et al.*, in *Global Carbon and Other Biogeochemical Cycles and Feedbacks*, ed. V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis and et al., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2021, pp. 673–816.
2 O. Boucher and U. Lohmann, *Tellus B*, 1995, 281–300.
3 J. Merikanto, D. V. Spracklen, G. W. Mann, S. J. Pickering and K. S. Carslaw, *Atmos. Chem. Phys.*, 2009, **9**, 8601–8616.
4 J. Almeida, S. Schobesberger, A. Kürten, I. K. Ortega, O. Kupiainen-Määttä, A. P. Praplan, A. Adamov, A. Amorim, F. Bianchi, M. Breitenlechner, *et al.*, *Nature*, 2013, **502**, 359–363.
5 J. Elm, J. Kubečka, V. Besel, M. J. Jääskeläinen, R. Halonen, T. Kurtén and H. Vehkamäki, *J. Aerosol Sci.*, 2020, **149**, 105621.
6 M. Engsvang, H. Wu, Y. Knattrup, J. Kubečka, A. B. Jensen and J. Elm, *Chem. Phys. Rev.*, 2023, **4**, 031311.
7 J. Elm, D. Ayoubi, M. Engsvang, A. B. Jensen, Y. Knattrup, J. Kubečka, C. J. Bready, V. R. Fowler, S. E. Harold, O. M. Longsworth, et al., *Wiley Interdiscip. Rev.: Comput. Mol. Sci.*, 2023, **13**, e1662.
8 T. Jokinen, M. Sipilä, H. Junninen, M. Ehn, G. Lönn, J. Hakala, T. Petäjä, R. L. I. Mauldin, M. Kulmala and D. R. Worsnop, *Atmos. Chem. Phys.*, 2012, **12**, 4117–4125.
9 E. Zapadinsky, M. Passananti, N. Myllys, T. Kurtén and H. Vehkamäki, *J. Phys. Chem. A*, 2019, **123**, 611–624.
10 M. Passananti, E. Zapadinsky, T. Zanca, J. Kangasluoma, N. Myllys, M. P. Rissanen, T. Kurtén, M. Ehn, M. Attoui and H. Vehkamäki, *Chem. Commun.*, 2019, **55**, 5946–5949.
11 K. Lehtipalo, C. Yan, L. Dada, F. Bianchi, M. Xiao, R. Wagner, D. Stolzenburg, L. R. Ahonen, A. Amorim, A. Baccarini, *et al.*, *Sci. Adv.*, 2018, **4**, eaau5363.
12 P. Pracht, S. Grimme, C. Bannwarth, F. Bohle, S. Ehlert, G. Feldmann, J. Gorges, M. Müller, T. Neudecker, C. Plett, S. Spicher, P. Steinbach, P. A. Wesołowski and F. Zeller, *J. Chem. Phys.*, 2024, **160**, 114110.
13 P. Pracht, F. Bohle and S. Grimme, *Phys. Chem. Chem. Phys.*, 2020, **22**, 7169–7192.
14 J. M. Dieterich and B. Hartke, *Mol. Phys.*, 2010, 279–291.
15 J. Zhang and M. Dolg, *Phys. Chem. Chem. Phys.*, 2015, **17**, 24173–24181.
16 J. Zhang and M. Dolg, *Phys. Chem. Chem. Phys.*, 2016, **18**, 3003–3010.
17 J. Kubečka, V. Besel, T. Kurtén, N. Myllys and H. Vehkamäki, *J. Phys. Chem. A*, 2019, **123**, 6022–6033.

18 G.-L. Hou, J. Zhang, M. Valiev and X.-B. Wang, *Phys. Chem. Chem. Phys.*, 2017, **19**, 10676–10684.
19 H. Li, O. Kupiainen-Määttä, H. Zhang, X. Zhang and M. Ge, *Atmos. Environ.*, 2017, **166**, 479–487.
20 J. Ling, X. Ding, Z. Li and J. Yang, *J. Phys. Chem. A*, 2017, **121**, 661–668.
21 H. Zhang, O. Kupiainen-Määttä, X. Zhang, V. Molinero, Y. Zhang and Z. Li, *J. Chem. Phys.*, 2017, **146**, 184308.
22 T. T. Odbadrakh, A. G. Gale, B. T. Ball, B. Temelso and G. C. Shields, *JoVE*, 2020, e60964.
23 B. T. Ball, S. Vanovac, T. T. Odbadrakh and G. C. Shields, *J. Phys. Chem. A*, 2021, **125**, 8454–8467.
24 S. E. Harold, C. J. Bready, L. A. Juechter, L. A. Kurfman, S. Vanovac, V. R. Fowler, G. E. Mazaleski, T. T. Odbadrakh and G. C. Shields, *J. Phys. Chem. A*, 2022, **126**, 1718–1728.
25 C. J. Bready, S. Vanovac, T. T. Odbadrakh and G. C. Shields, *J. Phys. Chem. A*, 2022, **126**, 5195–5206.
26 O. M. Longsworth, C. J. Bready and G. C. Shields, *Environ. Sci.: Atmos.*, 2023, **3**, 1335–1351.
27 O. M. Longsworth, C. J. Bready, M. S. Joines and G. C. Shields, *Environ. Sci.: Atmos.*, 2023, **3**, 1585–1600.
28 D. J. Wales and J. P. K. Doye, *J. Phys. Chem. A*, 1997, **101**, 5111–5116.
29 Y.-P. Zhu, Y.-R. Liu, T. Huang, S. Jiang, K.-M. Xu, H. Wen, W.-J. Zhang and W. Huang, *J. Phys. Chem. A*, 2014, **118**, 7959–7974.
30 X.-Q. Peng, Y.-R. Liu, T. Huang, S. Jiang and W. Huang, *Phys. Chem. Chem. Phys.*, 2015, **17**, 9552–9563.
31 S.-K. Miao, S. Jiang, J. Chen, Y. Ma, Y.-P. Zhu, Y. Wen, M.-M. Zhang and W. Huang, *RSC Adv.*, 2015, **5**, 48638–48646.
32 S.-S. Lv, S.-K. Miao, Y. Ma, M.-M. Zhang, Y. Wen, C.-Y. Wang, Y.-P. Zhu and W. Huang, *J. Phys. Chem. A*, 2015, **119**, 8657–8666.
33 S. Grimme, C. Bannwarth and P. Shushkov, *J. Chem. Theory Comput.*, 2017, **13**, 1989–2009.
34 C. Bannwarth, S. Ehlert and S. Grimme, *J. Chem. Theory Comput.*, 2019, **15**, 1652–1671.
35 J. J. P. Stewart, *J. Mol. Model.*, 2007, **13**, 1173–1213.
36 J. J. P. Stewart, *J. Mol. Model.*, 2013, **19**, 1–32.
37 A. B. Jensen, J. Kubečka, G. Schmitz, O. Christiansen and J. Elm, *J. Chem. Theory Comput.*, 2022, **18**, 7373–7383.
38 H. Wu, M. Engsvang, Y. Knattrup, J. Kubečka and J. Elm, *ACS Omega*, 2023, **8**, 45065–45077.
39 Y. Knattrup, J. Kubečka, D. Ayoubi and J. Elm, *ACS Omega*, 2023, **8**, 25155–25164.
40 J. Kubečka, F. R. Rasmussen, A. S. Christensen and J. Elm, *Environ. Sci. Technol. Lett.*, 2022, **9**, 239–244.
41 J. Kubečka, I. Neefjes, V. Besel, F. Qiao, H.-B. Xie and J. Elm, *J. Phys. Chem. A*, 2023, **127**, 2091–2103.
42 M. Engsvang, J. Kubečka and J. Elm, *ACS Omega*, 2023, **8**, 34597–34609.
43 J. Kubečka, Y. Knattrup, M. Engsvang, A. B. Jensen, D. Ayoubi, H. Wu, O. Christiansen and J. Elm, *Nat. Comput. Sci.*, 2023, **3**, 495–503.
44 Y. Knattrup, J. Kubečka and J. Elm, *J. Phys. Chem. A*, 2023, **127**, 7568–7578.
45 B. Temelso, J. M. Mabey, T. Kubota, N. Appiah-Padi and G. C. Shields, *J. Chem. Inf. Model.*, 2017, **57**, 1045–1054.
46 J. Kubečka, V. Besel, I. Neefjes, Y. Knattrup, T. Kurtén, H. Vehkamäki and J. Elm, *ACS Omega*, 2023, **8**, 45115–45128.
47 M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji and et al., *Gaussian 16*, Revision A 03, Gaussian, Inc., Wallingford CT, 2016.
48 J. Elm, *ACS Omega*, 2021, **6**, 7804–7814.
49 J. Elm, *ACS Omega*, 2021, **6**, 17035–17044.
50 J. Elm, *ACS Omega*, 2022, **7**, 15206–15214.
51 Y. Knattrup and J. Elm, *ACS Omega*, 2022, **7**, 31551–31560.
52 D. Ayoubi, Y. Knattrup and J. Elm, *ACS Omega*, 2023, **8**, 9621–9629.
53 N. Myllys, J. Elm and T. Kurtén, *Comput. Theor. Chem.*, 2016, **1098**, 1–12.
54 J. Elm, *ACS Omega*, 2019, **4**, 10965–10974.
55 F. Jensen, *J. Chem. Phys.*, 2001, **115**, 9113–9125.
56 V. Besel, J. Kubečka, T. Kurtén and H. Vehkamäki, *J. Phys. Chem. A*, 2020, **124**, 5931–5943.
57 H. R. Leverentz, J. I. Siepmann, D. G. Truhlar, V. Loukonen and H. Vehkamäki, *J. Phys. Chem. A*, 2013, **117**, 3819–3825.
58 I. K. Ortega, O. Kupiainen, T. Kurtén, T. Olenius, O. Wilkman, M. J. McGrath, V. Loukonen and H. Vehkamäki, *Atmos. Chem. Phys.*, 2012, **12**, 225–235.
59 N. Myllys, J. Kubečka, V. Besel, D. Alfaouri, T. Olenius, J. N. Smith and M. Passananti, *Atmos. Chem. Phys.*, 2019, **19**, 9753–9768.
60 A. B. Nadykto and F. Yu, *Chem. Phys. Lett.*, 2007, **435**, 14–18.
61 T. Kurtén, L. Torpo, M. R. Sundberg, V. Kerminen, H. Vehkamäki and M. Kulmala, *Atmos. Chem. Phys.*, 2007, **7**, 2765–2773.
62 M. Engsvang and J. Elm, *ACS Omega*, 2022, **7**, 8077–8083.