Massive Assessment of the Geometries of Atmospheric Molecular Clusters

Andreas Buchgraitz Jensen and Jonas Elm*

Department of Chemistry, Aarhus University, Langelandsgade 140, 8000 Aarhus C, Denmark

E-mail: jelm@chem.au.dk

Phone: +45 28938085

Abstract

Atmospheric molecular clusters are important for the formation of new aerosol particles in the air. However, current experimental techniques are not able to yield direct insight into the cluster geometries. This implies that to date there is limited information about how accurately the applied computational methods depict the cluster structures.

Here we massively benchmark the molecular geometries of atmospheric molecular clusters. We initially assess how well different DF-MP2 approaches reproduce the geometries of 45 dimer clusters obtained at a high DF-CCSD(T)-F12b/cc-pVDZ-F12 level of theory. Based on the results we find that the DF-MP2/aug-cc-pVQZ level of theory best resembles the DF-CCSD(T)-F12b/cc-pVDZ-F12 reference level. We subsequently optimize 1283 acid–base cluster structures (up to tetramers) at the DF-MP2/aug-cc-pVQZ level of theory and assess how more approximate methods reproduce the geometries. Out of the tested semi-empirical methods, we find that the newly parameterized atmospheric molecular cluster extended tight binding method (AMC-xTB) is most reliable for locating the correct lowest energy configuration and yield the lowest RMSD compared to the reference level. In addition, we find that the DFT-3c methods show similar performance as the usually employed ωB97X-D/6-31++G(d,p) level of theory at a potentially reduced computational cost. This suggests that these methods could prove valuable for large-scale screenings of cluster structures in the future.

1 Introduction

The formation of atmospheric molecular clusters is believed to be the initial step in atmospheric aerosol New Particle Formation (NPF).[1] Aerosol particles directly affect our global climate via aerosol-radiation interactions. This can either be scattering of sunlight by inorganic particles or absorption of radiation by black carbon.[2] Aerosol particles also indirectly affect the global climate via aerosol-cloud interactions.[3] For instance, aerosol particles around 50 nm or larger can act as nuclei for cloud droplet formation. Combined, these aerosol effects present the largest uncertainty in the understanding of our current climate and the prediction of climate change in the future.[4] NPF is a large source of this uncertainty, as it remains uncertain which atmospheric vapours that are important for the initial cluster formation and the subsequent growth.[5]

Atmospheric low-volatile acids such as Sulfuric Acid (SA),[6] Methanesulfonic Acid (MSA)[7,8] and Nitric Acid (NTA)[9] are believed to play a decisive role in the initial cluster formation and/or growth process. However, to facilitate the cluster formation process, the low-volatile acids must react with atmospheric bases such as Ammonia (AM),[10] alkylamines (Methylamine (MA), Dimethylamine (DMA), Trimethylamine (TMA))[11,12] and diamines (Ethylenediamine (EDA), Putrescine (PUT)).[13,14] At coastal regions iodine species emitted by algae are conceived to play a pivotal role in cluster formation.[15–17] Depending on the saturation vapour pressure of the compounds,[18,19] Highly Oxygenated organic Molecules (HOMs)[20,21] can also influence cluster formation[22,23] and growth.[24] Overall, this leads to strongly bound clusters held together by a combination of hydrogen-bonded interactions and electrostatic interactions.

Unfortunately, it is very difficult to measure the initial cluster formation using experimental techniques. State-of-the-art instruments such as the Chemical Ionization Atmospheric Pressure interface Time-Of-Flight mass spectrometer (CI-APi-TOF)[25] can measure the chemical composition of the clusters. However, the ionization efficiency by different CI reagent ions (nitrate[26]/bisulfate[27]/acetate[28]/iodide[29]) varies between individual atmo-

spheric species and fragmentation inside the instrument also affect the detection efficiency of the clusters.[30–33] Overall, CI-APi-TOF measurements only yield the chemical composition and at the moment there exist no experimental techniques to obtain direct insight into the exact cluster geometries.

Quantum Chemical (QC) calculations can yield explicit insight into the cluster geometries and based on statistical mechanics the thermochemistry can be calculated. The calculated binding free energies can be used to calculate the evaporation rates of the clusters,[34] which gives information about their corresponding stability under atmospheric conditions. Collision coefficients between monomers/clusters can be calculated using kinetic gas theory. Combined, the collision and evaporation rates can be used to calculate the cluster formation rate using, for instance, the Atmospheric Cluster Dynamics Code (ACDC).[35] This gives knowledge about the potential NPF rate of the species. However, the accuracy of the calculated NPF rates is very dependent on the applied QC methods. This implies that one can essentially get the NPF rate one is after, and thereby, accurate benchmarks are important to make an informed and unbiased decision on what QC level of theory to apply.

We have extensively benchmarked the required methodologies for obtaining accurate cluster electronic binding energies.[36–38] In the recent years we have curated the Clusteromics I–V datasets[39–43] composed of the acids SA, MSA, NTA and Formic Acid (FA) combined with the bases AM, MA, DMA, TMA and EDA. All combinations of the species have been considered, leading to a database of a total of 56,436 small acid–base clusters (up to 5 molecules). Based on previous benchmarks, the cluster geometries and vibrational frequencies were obtained at the ωB97X-D/6-31++G(d,p) level of theory[44,45] and all data is freely available in the Atmospheric Cluster DataBase (ACDB).[46,47] Recently, we applied the full Clusteromics I–III datasets to massively assess the electronic binding energies of 11,749 atmospheric molecular clusters.[48] Such a benchmark set can be used for detailed statistics to be carried out, allowing an informed decision on the best level of theory to apply. However, to date there exist no benchmark studies that have addressed the geometries of atmospheric molecular clusters.

In this paper we employ the Clusteromics I–V datasets to massively assess the geometries of atmospheric molecular clusters. We initially optimized the geometry of 45 dimer clusters at the DF-CCSD(T)-F12b/cc-pVDZ-F12 level of theory to probe how well cheaper wave function theory (MP2) can reproduce the cluster geometries. Based on the results we utilize DF-MP2/aug-cc-pVQZ to optimize the geometry of a subset of the Clusteromics I–V datasets. We optimized the geometry of the five lowest configurations of each cluster composition, yielding a total of 1283 cluster structures. Finally, we compare several faster and more approximate methods to this DF-MP2/aug-cc-pVQZ geometry dataset and apply detailed statistics to compare the geometries.

2 Methodology

2.1 Quantum Chemical Computational Details

For determining the reference method for our geometry benchmark we employed Molpro,[49–51] as analytical gradients are available for DF-MP2[52]/aug-cc-pVnZ (n = D, T, Q), DF-MP2-F12[53,54]/cc-pVnZ-F12[55] (n = D, T), and DF-CCSD(T)-F12b[56,57]/cc-pVDZ-F12. In all cases the default parameters, or as specified by their documentation, was used[a]. We used DF-CCSD(T)-F12b/cc-pVDZ-F12 as the highest level of theory to determine which variant of MP2 to use on the larger test set, as discussed in Section 2.1.1 and Section 3.1.

ORCA version 5.0.4[58] was used to optimize the geometries using r^{2}SCAN-3c,[59] B97-3c,[60] ωB97X-D3BJ,[61] PW91[62] and ωB97X-3c.[63] As ωB97X-3c is not a part of ORCA 5.0.4 the ORCA4wB97X-3c[64] script was used to set up the calculation. xTB version 6.4.0[65,66] was used for semi-empirical tight binding optimizations using GFN1-xTB,[67] GFN2-xTB[68] and the newly parameterized Atmospheric Molecular Cluster (AMC) extended tight binding method denoted AMC-xTB.[69] Gaussian16[70] was used for PM6[71] and PM7[72] geometry optimizations.

[a]Note that specific parameters are needed for employment of analytical gradients, see Molpro manual in the section "Explicitly correlated methods".

2.1.1 Reference Methods

To determine the reference method used for the full test set, we compare 5 levels of MP2 to DF-CCSD(T)-F12b/cc-pVDZ-F12 calculations. We have chosen the given coupled-cluster method, as CCSD(T) consistently has been shown to be an excellent reference method and is extensively used in various prominent benchmarks.[73–77] The explicit correlation approach (-F12) has been shown to approach the basis set limit faster than the conventional approach.[56,57,78]

In cases where CCSD(T) ($\mathcal{O}(n^7)$) is not feasible, MP2 ($\mathcal{O}(n^5)$) is often used as a cheaper alternative wave function-based approach. MP2 has been shown to give very reliable results[54,79,80] and should be a highly suitable reference for geometries.

We employ the correlation-consistent basis sets, with variants specifically optimized for the explicitly correlated methods.[81] For the coupled-cluster method "only" the double-zeta basis set was used as going to larger basis sets was not computationally feasible. However, because of the explicitly correlated corrections this is still a very decent reference method for geometry optimizations.

2.2 Molecular Clusters Test Set

We employed the Clusteromics I–V[39–43] datasets for the benchmark. These datasets consist of clusters of the type $(\text{acid})_{1-2}(\text{base})_{1-2}$ and includes mixtures of different acids (SA, MSA, FA, and NTA) and bases (AM, MA, DMA, TMA, and EDA). Overall, the Clusteromics I–V datasets contain up to 56,436 cluster structures optimized at the ωB97X-D/6-31++G(d,p) level of theory. This is an insurmountable number of clusters to further optimize at the MP2 level. Hence, we have chosen to only use the 5 lowest energy structures for each cluster composition, which in total yields a dataset of 1283 cluster structures.

The goal of this paper is to compare the results of geometry optimizations done at different levels of theory. However, different methods might end up in different optimized geometries/configurations and this would lead to a skewed comparison. Hence, in the fol-

lowing sections (Section 2.3 and Section 2.4) we will give a description of the methods used to distinguish between different configurations, such that we in Section 3 can directly compare clusters which have actually been optimized to the same minimum geometry.

2.3 Statistical Tests

2.3.1 The Empirical Distribution Function

In Section 2.4.2 we are interested in determining whether a given dataset follows a specific statistical distribution. Here, a short introduction into statistical tests based on the Empirical Distribution Function (EDF) is given.[82] Lets consider a random variable X. The propability that X will give a value less than or equal to the value x is computed by evaluating the Cumulative Distribution Function (CDF) at x. Assuming that X follows a distribution given by the Probability Distribution Function (PDF) $f_X(t)$ the probability of X can be computed giving a value less than or equal to x, that is evaluate the CDF, by integration of the PDF as follows:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt. \tag{1}$$

Given a random ordered sample of data, denoted as $X_1, ..., X_n$, drawn from a given CDF the EDF is defined as:

$$F_n(x) = \frac{\text{number of observations} \leq x}{n}, \tag{2}$$

that is, the EDF goes up in value by $1/n$ for every value, X_i, in the data set as x goes from $-\infty$ to ∞. This EDF is an approximation of the theoretical CDF, converging as $n \to \infty$.

As $F_n(x)$ approaches the true $F_X(x)$ with increasing n, a natural measure to test if the sample data follows a proposed distribution would involve the difference of these. One such measure is the Anderson-Darling test which will be employed in the following sections. The Anderson-Darling statistic is a member of the quadratic statistics defined commonly as:

$$Q = n \int_{-\infty}^{\infty} \{F_n(x) - F_X(x)\}^2 \Omega(x) dF(x). \tag{3}$$

When $\Omega(x) = [F_X(x)\{1 - F_X(x)\}]^{-1}$ the Anderson-Darling statistic is obtained, which will be denoted A^2. This integral can be transformed into a much more manageable expression:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\log(Z_i) + \log(Z_{n+1-i})\right], \tag{4}$$

where $Z_i = F_X(X_i)$.

2.3.2 Unknown parameters of $F_X(x)$

To evaluate $F_X(x)$ the parameters, Γ, are needed for the given distribution, or at least an estimate of these $\hat{\Gamma}$. This leads to a set of scenarios (commonly also called cases) where all the parameters or only a (potentially empty) subset of them are known. The present work will operate from the scenario where all parameters are unknown. This entails an estimate of the parameters, which means that $Z_i = F_X(X_i; \hat{\Gamma})$. Estimation of the parameters is often carried out with methods like maximization of the log-likelihood[83] or in the case of the Cauchy distribution by means of sums of weighted order statistics.[84]

For the case of completely unknown parameters the critcal values used to reject the hypothesis of a given distribution depends on the distribution being tested and the size of the data set. Tables of critical values exists, we have used the ones provided by Ref. 82 (Table 4.26).

2.4 Algorithm for Comparing Configurations

To compare how similar two molecular structures are, an easy method is to simply compute the Root Mean Square Deviation (RMSD) between the two geometries. This can be done by finding the optimal rotation matrix between the two structures using fx. the Kabsch algorithm[85] before computing the actual RMSD. This works well for small systems, but for large systems, such as molecular clusters, this can quickly "hide" small differences. An example of this could be an ammonia molecule inside a multi-component cluster: If all

other monomers are largely the same, but the ammonia sits in a different orientation it is technically a different configuration, but as there are only 4 atoms in ammonia this deviation contributes very little to the overall RMSD.

We derived an algorithm (Algorithm 1) which overcomes this problem, which we have used as a tool in the benchmarking presented in this paper. The algorithm is implemented and freely available at: `https://gitlab.com/AndreasBuchgraitz/clusteranalysis`.

Algorithm 1: Divide clusters as SameAsReference and DifferentFromReference

Data: XYZ-files of optimized geometries
Result: Configurations labeled as the same or different configuration compared to the reference configuration

```
1  Cluster := specific configuration optimized at a given level of theory.
2  foreach cluster do
3      Identify monomers in cluster.

4  cluster_ref := Cluster optimized at reference theory.
5  cluster_vec := Vector of clusters optimized (from identical configuration as
     cluster_ref) at level of theory to be compared to reference theory.
6  foreach cluster ∈ cluster_vec do
7      Monomer pair := 2 equivalent (same relative position in cluster) monomers, 1
         from cluster and 1 from cluster_ref.
8      foreach monomer pair, p1 do
9          Compute rotation matrix for p1, rot-mat-p1.
10         foreach monomer pair, p2 do
11             Apply rot-mat-p1 to p2.
12             Compute discrepancy of rotated monomers in p2.

13 foreach cluster ∈ cluster_vec do
14     Analyse above computed deviation of monomers to distinguish between
         configurations of different geometry.
```

The division of molecular clusters into its consisting monomers (Line 3) is done as a "connectivity map". This means that for all atoms, A_i, the distance to all other atoms, A_j, is computed and these are connected, if they are reasonably close to each other. We have found that the maximum allowed distance between atoms computed as $min(s + s \cdot 0.3, 2.0\text{Å})$ with s being the smallest distance found between A_i and any of the other atoms A_j, to be

a good measure for "reasonably close". This is done ignoring all hydrogen atoms, which are added at the end to the monomer for which they are in closest proximity to.

The computation of the rotation matrix (Line 9) is performed using the Kabsch algorithm. In the main part of the algorithm (Line 4 to Line 12) a loop over all the optimized cluster geometries for all the computational methods which should be compared to the reference method is performed (Line 6). Now, this cluster geometry should compared to the reference geometry. This is done in such a way that, if the two structures are identical, a lot of zeroes are computed, such that there are as many numbers as possible to analyse in the later part of the algorithm. The idea is that if two clusters are identical then the rotation matrix from any pair of equivalent monomers should be equal and equal to the rotation matrix for the entire cluster. So if this is assumed and the rotation matrix for all monomer pairs (Line 9) is computed and applied to all monomers (Line 11) the discrepancy can be computed for all the M^2 rotated monomer pairs (M being the number of monomers) (Line 12).

The "discrepancy" computed for the rotated monomers (Line 12) is based on two different measurements: 1) the difference between all position coordinates ($3N$ numbers for N atoms), 2) $\cos(\theta)$ with θ being the angle between estimations of the dipole-moments of the two monomers. The estimation of the dipole-moments is based on the electronegativity and number of electrons in the outer shell of the atoms. This gives a decent estimate of the orientation of the dipole-moment though the magnitude is likely off by many magnitudes. However, we are only interested in analysing the orientation of the monomers.

Analysis of the discrepancy (Line 13) is performed in two ways for the two different measurements of deviation, as seen in the following two sections.

2.4.1 Analysis 1: Angles of the Dipole Moments

The analysis of the angle deviation can return three conclusions: "DefinitelyTheSameAsReference", "SameAsReference", and "DifferentFromReference". The conclusion is based on the value of $\cos(\theta)$ compared to two chosen values: 0.96 and 0.76, which corresponds

to approximately 16.26° and 40.54° respectively. If all computed values of $\cos(\theta)$ are larger than 0.96 "DefinitelyTheSameAsReference" is returned, else if all values are larger then 0.76 then "SameAsReference" is returned, else "DifferentFromReference" is returned. The check values are chosen as a rather tight value and a rather loose value, respectively. If the cluster pass the tight check it is very likely that it is the same configuration as the cluster computed at the reference level of theory. If the cluster only pass the loose check it also needs to pass the coordinate difference analysis (see below) for the cluster to be labeled as "SameAsReference" otherwise it is labeled as "DifferentFromReference".

2.4.2 Analysis 2: Coordinate Differences

The analysis of the coordinate differences is performed using the methods described in Section 2.3. The targeted distribution is a Cauchy distribution, with PDF

$$f(x) = \frac{1}{\pi} \frac{\gamma}{\gamma^2 + (x - x_0)^2}, \tag{5}$$

as it approaches the Dirac delta function when the scale parameter, γ, approaches zero.[86] This is an appropriate limit, because if the two clusters are exactly the same geometry then the coordinate difference distribution would exactly be the Dirac-delta function. In reality the two geometries are not exactly the same when optimized at two different levels of theory, but assuming the differences between the two compared geometries is small, the differences should be able to be described well using the Cauchy distribution. Whereas, if the differences between the two geometries are large, fx. if one or more monomers have been significantly rotated, the data should not follow the cauchy distribution.

As an example, Figure 1 shows the distribution of the coordinate differences together with the proposed PDF, and the EDF with the proposed CDF. This have been plotted for two clusters where the first, $(\mathrm{NTA})_1(\mathrm{SA})_1(\mathrm{EDA})_2$, has optimized to the same minimum as the reference, and the second, $(\mathrm{FA})_1(\mathrm{MSA})_1(\mathrm{TMA})_1$, has optimized to a minimum different

from the reference method. It is here quite clear to see that the first cluster decently follows the proposed distribution, while the second cluster deviates significantly.

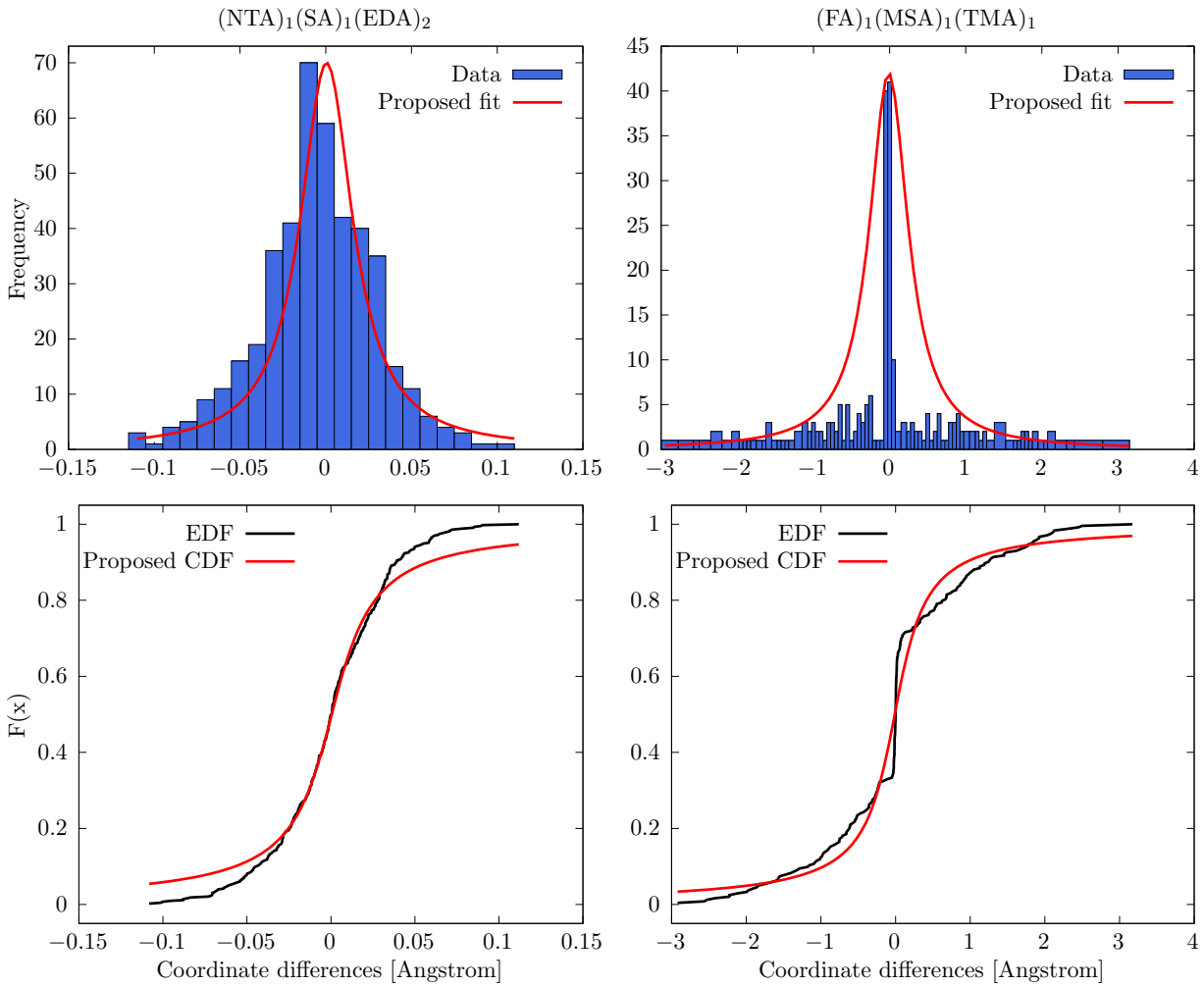

Figure 1: The two upper panels contain the histogram/distribution of the differences of the atomic positions for two molecular clusters. The geometries compared are optimized at the reference level (DF-MP2/aug-cc-pVQZ) and r^{2}SCAN-3c. The proposed fit is obtained through sums of weighted order statistics[84] for the cauchy PDF. In the two lower panels the EDF for the same data is plotted alongside the CDF for a cauchy distribution (using the same parameters as the PDF).

To test whether the coordinate differences comes from a Cauchy distribution the Anderson-Darling statistic is utilised as described in Section 2.3. If the Anderson-Darling statistics does not reject the Cauchy distribution as a distribution for the coordinate differences this part of the analysis label the given cluster as "SameAsReference" otherwise "DifferentFrom-

Reference".

2.4.3 Final Configuration Assignment

The final part of the analysis is to compare the results from the two sub-analyses in the previous sections. As shortly described above, there are three scenarios leading to two results: 1) the angle deviation analysis labels the cluster as "DefinitelyTheSameAsReference" which leads to the cluster overall being labeled as "SameAsReference", 2) if both the angle deviation and coordinate difference analyses label the cluster as "SameAsReference" the cluster is saved as "SameAsReference", 3) if only one or neither of the sub-analyses labels the cluster as "SameAsReference" it is saved as "DifferentFromReference".

3 Results and Discussion

3.1 Geometry Benchmark - Dimer Clusters

Before comparing the geometry optimization of different computational methods a reference method needs to be established. We employed the previously used set of 45 small dimer clusters found in Ref. 38. When choosing the reference method two concepts should be kept in mind, accuracy and computational wall time. When initially investigating the possible reference method on this small set of dimers a level of theory can be used which would be impossible to apply for the entire 1283 cluster data set described in Section 2.2.

To jugde the accuracy of the potential reference methods two measures are used, the electronic dissociation energy and the RMSD. The RMSD value is an obvious choice when comparing the geometry of small molecules or clusters. The electronic dissociation energy gives a decent estimate of the electronic potential without requiring specific information about "shape" of the potential.

Figure 2 compares the dissociation energy of five different levels of DF-MP2 to DF-CCSD(T)-F12b/cc-pVDZ-F12. The dissociation energy in question is from the bottom of

the electronic potential, D_e, computed as the optimized energy for the cluster from which the optimized energy of the isolated monomers is subtracted:

$$D_e = E^{\text{min}}_{\text{Cluster}} - \sum_{m \in \text{monomers}} E^{\text{min}}_{m}. \tag{6}$$

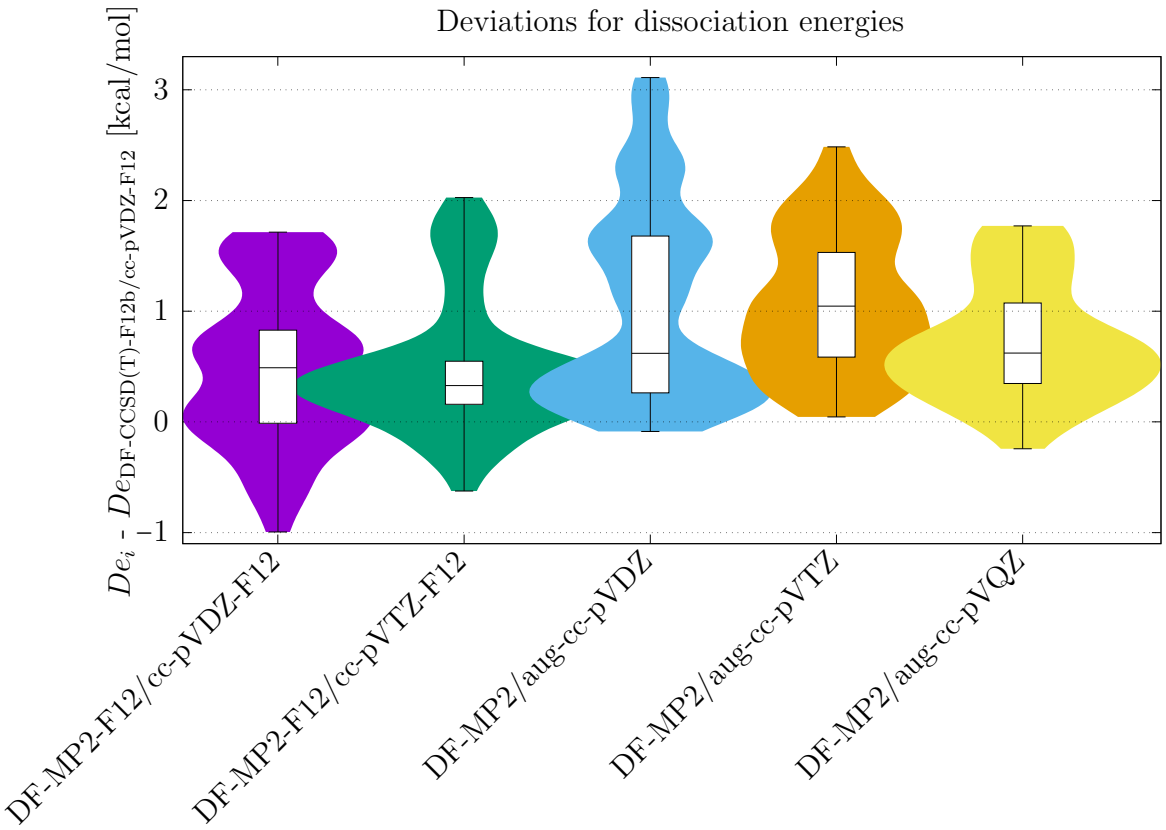

Figure 2: Dissociation energy deviations. Computed as the difference between the minimum electronic energy of the dimer and the minimum electronic energy of the isolated monomers.

The deviations from the high level of theory have been plotted as violinplots with a boxplot on top. These together give a good sense of the distribution of the data as the boxplots contain 25 % of the data in each of the four sections, and the violinplots highlighting the distribution making it a bit easier to see where most of the data points are located (where the violinplots are the widest).

In all cases there seem to be an offset of approximately 0.5 kcal/mol in the D_e-values compared to the reference DF-CCSD(T)-F12b/cc-pVDZ-F12 calculations. Overall, decent

agreement between most of the MP2 methods is seen, with DF-MP2/aug-cc-pVDZ deviating and being slightly worse. Accuracy-wise the DF-MP2-F12/cc-pVTZ-F12 calculations of the D_e best resembles the DF-CCSD(T)-F12b/cc-pVDZ-F12 reference.

Figure 3 shows the RMSD between the high level reference and the 5 tested levels of MP2 methods. Four of the violinplots seem to have an outlier, this is because they find a slightly different geometry compared to the coupled cluster geometry. In all cases there is a slight shift in the geometries with median RMSD differences between roughly 0.01 Å and 0.03 Å. Overall, good agreement between the two F12 methods and the DF-MP2/aug-cc-pVQZ is seen, and again DF-MP2/aug-cc-pVDZ seems to be quite off. Here it actually seems that DF-MP2-F12/cc-pVDZ-F12 would be the optimal choice for accuracy, but the difference between this and DF-MP2-F12/cc-pVTZ-F12 and DF-MP2/aug-cc-pVQZ is so small that we attribute this to a lucky cancellation of errors.

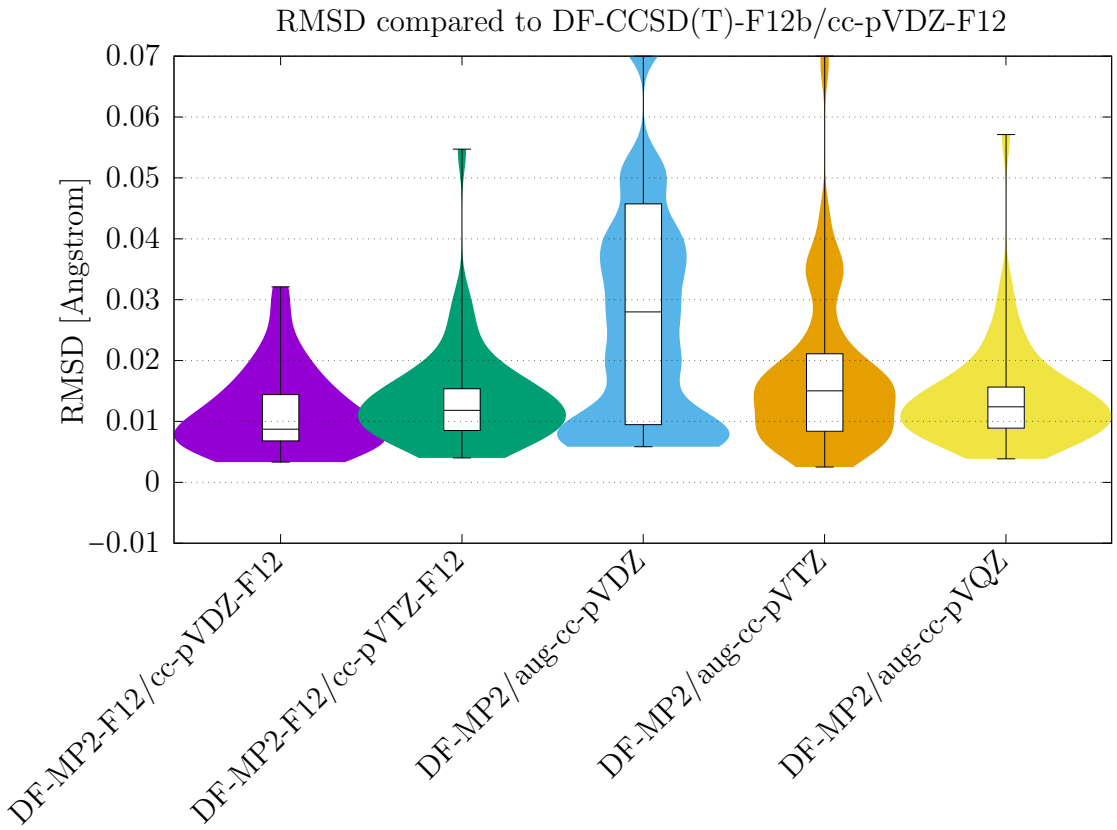

Figure 3: RMSD computed after rotation using the Kabsch algorithm.

Based on accuracy DF-MP2-F12/cc-pVTZ-F12 performs the best compared to the reference method, but as mentioned, computational wall time is also a significant factor. DF-MP2-F12/cc-pVTZ-F12 took about 200 hours to optimize one of the larger systems ($(MSA)_1(TMA)_1$) in the dimer cluster test set while DF-MP2/aug-cc-pVQZ only took about 4 hours. Even though DF-MP2-F12/cc-pVTZ-F12 would be the best choice, it is simply not feasible when going to larger cluster sizes containing up to four monomers. Hence, based on both the calculated D_e-values and the RMSD, we will employ the DF-MP2/aug-cc-pVQZ level of theory as the reference method for optimizing the geometries of the large 1283 cluster test set.

3.2 Geometry Benchmark - Full Test Set

3.2.1 Configurational Ordering

Using the DF-MP2/aug-cc-pVQZ level of theory as reference method, the full set of 1283 cluster geometries were optimized and compared to several approximate methods using the algorithm described in Section 2.4. These methods have been chosen based on previous benchmarks on atmospheric molecular clusters and their wide application in cluster formation studies. Table 1 presents the number of clusters labeled as "SameAsReference" and what fraction of the clusters this corresponds to for all the tested methods.

Table 1: Computational methods tested listed alongside with the number and fraction of clusters labeled as "SameAsReference".

Method	Number of clusters	Fraction of clusters
PM6	489	0.3811
PM7	429	0.3344
GFN1-xTB	743	0.5791
GFN2-xTB	729	0.5682
AMC-xTB	938	0.7311
r^{2}SCAN-3c	1196	0.9322
B97-3c	1117	0.8706
ωB97X-3c	1187	0.9252
PW91/aug-cc-pVTZ	844	0.6578
ωB97X-D3BJ/6-31++G(d,p)	1194	0.9306
ωB97X-D3BJ/aug-cc-pVTZ	1182	0.9213

It is seen that r^{2}SCAN-3c, ωB97X-3c, ωB97X-D3BJ/6-31++G(d,p), and ωB97X-D3BJ/aug-cc-pVTZ finds the same minimum as the reference for more than 92% of the clusters indicating that these works very well for these types of systems. B97-3c also performs very well with 87% of the clusters as the same minimum geometry. This implies that the significantly cheaper r^{2}SCAN-3c and ωB97X-3c methods might be a good alternative to the usually employed ωB97X-D3BJ/6-31++G(d,p) level of theory for obtaining the cluster strutcures. Interestingly, there is seen little difference between increasing the basis set size from 6-31++G(d,p) to aug-cc-pVTZ for ωB97X-D3BJ. Based on the large aug-cc-pVTZ basis set it is seen that PW91 is performing significantly worse than the ωB97X-D3BJ functionals with only 66% of the clusters being assigned to the correct minimum. This is consistent with numerous previous benchmarks of the binding energies of atmospheric molecular clusters.[38,48]

The semi-empirical methods PM6 and PM7 find less than 40% of the clusters to be the same minimum. GFN1-xTB and GFN2-xTB are performing slightly better, but only finds the presumed correct minimum 56% and 57% of the time, respectively. This illustrates that these methods cannot reliable identify the correct configurations and should only be used as pre-screening tools in configurational sampling and not for the final geometries. This is consistent with the previous work by Kurfman et al.[87] for $(SA)_3$ cluster configurations and

the work by Wu et al.[88] on large $(SA)_n(AM/DMA)_n$ clusters, with $n = 1 - 20$. The newly parameterized AMC-xTB method is performing significantly better than the other semi-empirical methods. Hence, AMC-xTB should be the best method to use for pre-optimization of atmospheric molecular clusters during cluster configurational sampling. It should be noted that AMC-xTB was parameterized on the full clusteromics I–V data sets and thereby the improved performance is not surprising.

3.2.2 Geometry RMSD

Besides looking at the number of configurations which have been optimized to the same minimum as the reference method, the quality of these minima can also be investigated. Figure 4 shows the RMSD compared to the reference structures. Here, the RMSD for all the clusters which have been labeled as "SameAsReference" have been plotted using the same violinplot and boxplot style as previously. Very similar to the results listed in Table 1 it is seen that the 4 best performing methods are r^{2}SCAN-3c, ωB97X-3c, ωB97X-D3BJ/6-31++G(d,p), and ωB97X-D3BJ/aug-cc-pVTZ. This again indicates that the r^{2}SCAN-3c and ωB97X-3c functionals should yield reliable geometries of atmospheric molecular clusters. The B97-3c method is again also seen as a strong alternative. Similar to the previous section the semi-empirical methods PM6, PM7, GFN1-xTB, GFN2-xTB and AMC-xTB gives the worst results. PM6 and PM7 again performs worse than GFN1-xTB and GFN2-xTB. The newly parameterized AMC-xTB method has a slightly lower RMSD and more narrow distribution compared to other semi-empirical methods. This further illustrates the robustness of the AMC-xTB method.

Besides looking at the RMSD the two measurements used in Algorithm 1 could also be analysed. Figures similar to Figure 4 have been included in the supporting information where the plotted values are, the maximum deviation between the orientations of the dipole moments for the monomers in a given cluster, and the maximum value of a coordinate deviation. The trend is overall the same as seen for the RMSD in Figure 4. Figures showing the

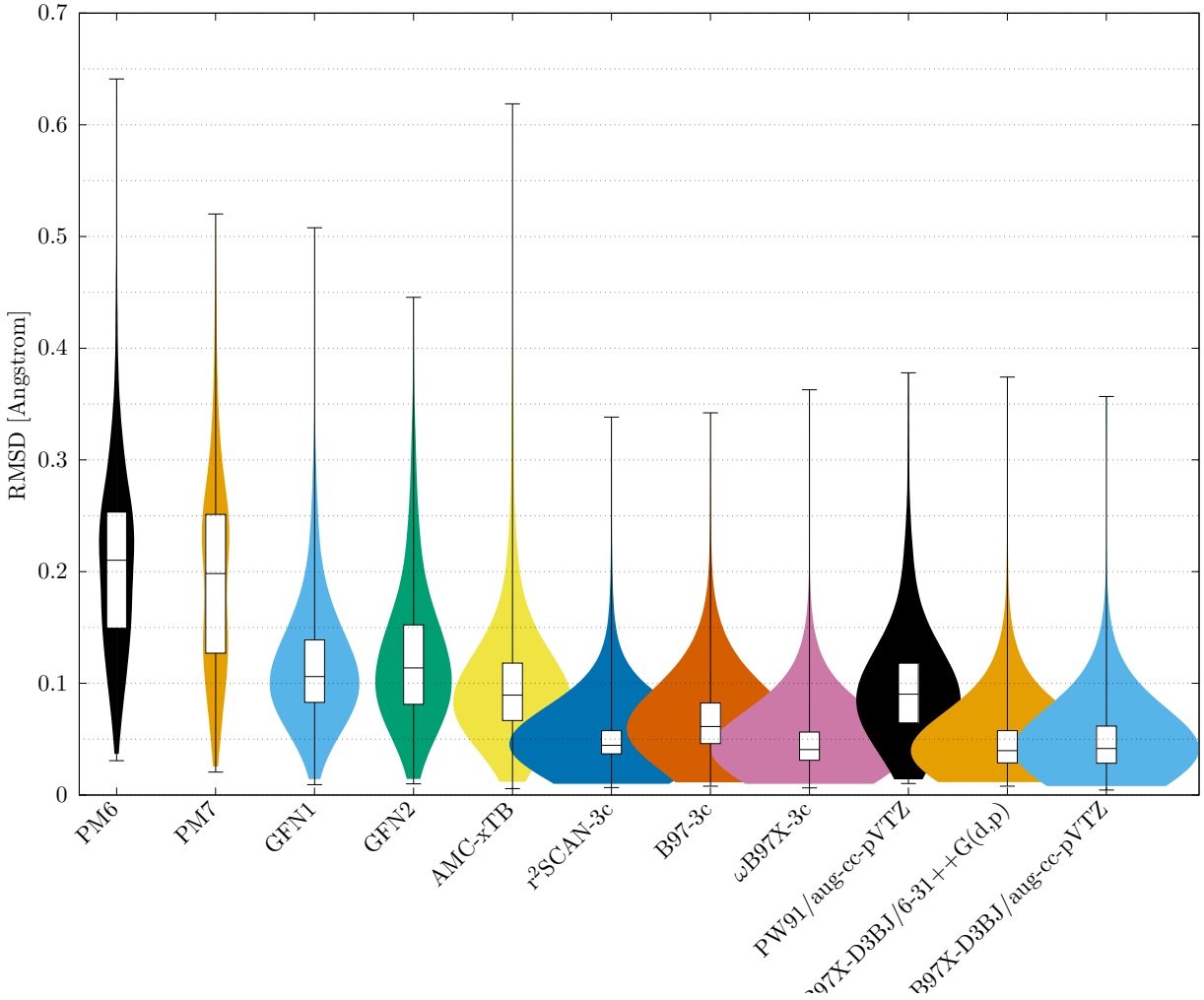

Figure 4: RMSD computed between listed methods and DF-MP2/aug-cc-pVQZ. This figure includes all the molecular clusters which have been labeled as "SameAsReference".

distributions for all three measurements for the clusters labeled as "DifferentFromReference" are also included in the supporting information.

Overall, the fact that ωB97X-D3BJ/6-31++G(d,p) performs well compared to higher level methods, both in regards to locating the correct lowest energy minimum and with respect to RMSD, further strengthens its wide utilization in cluster formation studies.

3.3 Timings

The actual computational wall time is a decisive factor in how many cluster configurations can be optimized at the final DFT level. We tested the computational timings for two of the largest clusters from our test set: $(FA)_1(MA)_2(MSA)_1$ which has 28 atoms and $(DMA)_2(FA)_2$ which has 30 atoms.

Figure 5 presents the (unparallelized) CPU times of the most promising methods relative to PW91/aug-cc-pVTZ. We have left out the semi-empirical methods in this figure as these are all extremely fast (very few optimizations taking more than a couple of seconds).

Non-surprisingly, it is seen that B97-3c and r^{2}SCAN-3c are very cheap while DF-MP2/aug-cc-pVQZ is the most expensive. We do not see a huge time-advantage of the ωB97X-3c method over the ωB97X-D3BJ/6-31++G(d,p) level of theory. If this is compared to Ref. 63, where they show some timings for ωB97C-3c, they have much larger system size (381 atoms) and they compare to the DFT functional ωB97X-V with a quadruple zeta basis set, where the largest basis set compared here is the triple zeta basis set.

Overall, it is seen that all the DFT-3c methods are performing well compared to the reference DF-MP2/aug-cc-pVQZ level of theory and can potentially be used as alternatives to the usually employed ωB97X-D3BJ/6-31++G(d,p) level for obtaining accurate cluster geometries, with a potential gain in computational efficiency. This gain in efficiency will be significantly improved for larger cluster systems, where the ωB97X-D3BJ/6-31++G(d,p) level of theory becomes prohibitively expensive. It should be noted that besides the geometries, the accuracy of the vibrational frequencies is also important as they are used to

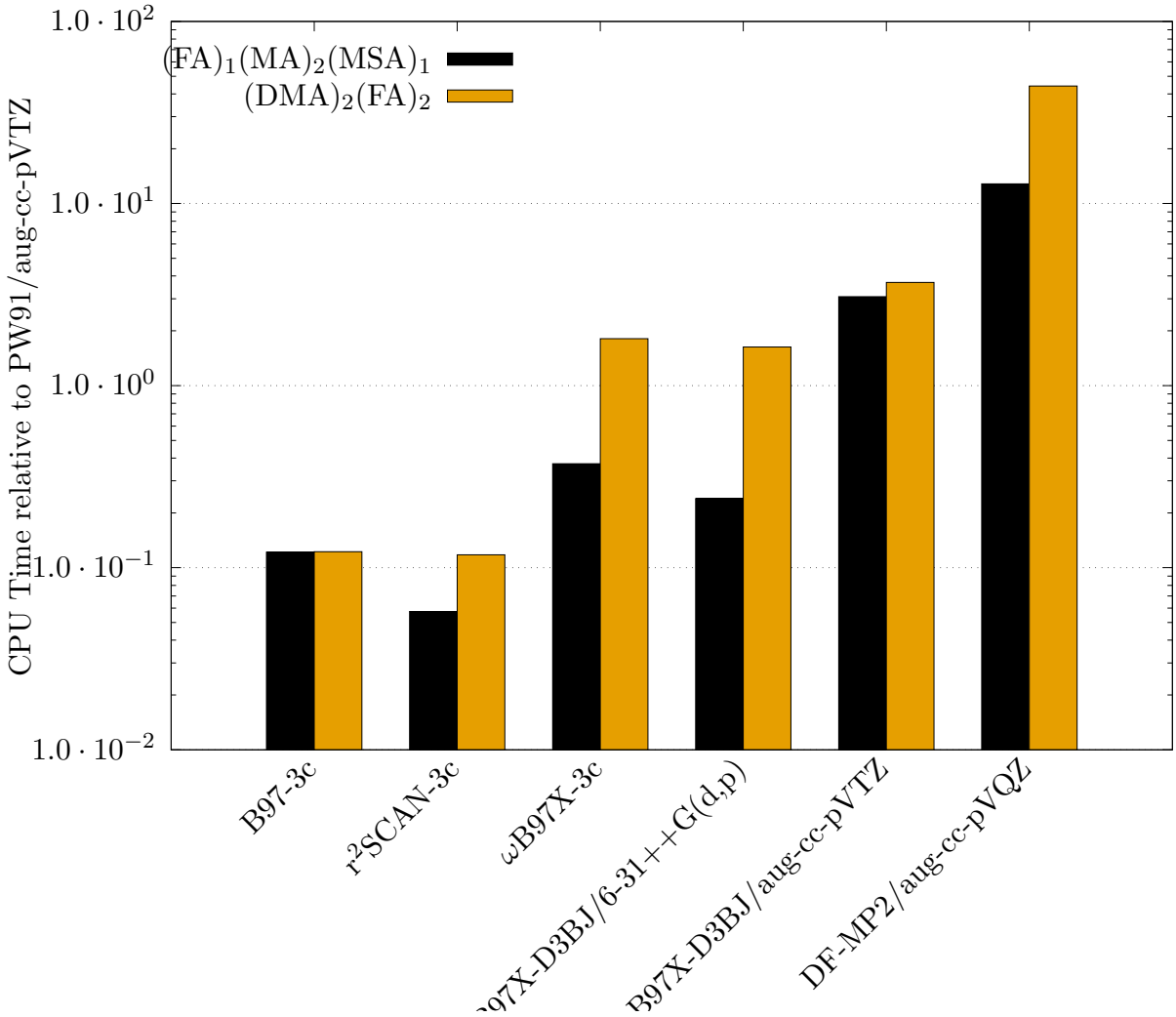

Figure 5: CPU time relative to PW91/aug-cc-pVTZ calculations for two clusters each with four monomers.

determine the zero-point vibrational energy and entropy contribution. Hence, to obtain more reliable free energies, a large scale investigation of the accuracy of the vibrational frequencies should be performed in the future.

4 Conclusions

We have massively assessed the accuracy of various approximate methodologies in yielding accurate geometries of atmospheric molecular clusters compared to a reference DF-MP2/aug-cc-pVQZ level of theory. Out of the tested semi-empirical methods (PM6, PM7, GFN1-xTB, GFN2-xTB and AMC-xTB) we find that the newly parameterized Atmospheric Molecular Cluster (AMC) extended tight-binding method is vastly superior. Hence, we can recommend that AMC-xTB is used for pre-optimization during configurational sampling of atmospheric molecular clusters.

We find that the DFT-3c methods and ωB97X-D3BJ/6-31++G(d,p) all perform well, both in regard to finding the correct lowest minimum cluster configuration and based on overall RMSD. The fact that ωB97X-D3BJ/6-31++G(d,p) performs well compared to higher level methods based on a large test set further strengthens its wide utilization in cluster formation studies. Based on computational timings the B97-3c and r^{2}SCAN-3c methods look very promising for obtaining accurate geometries at a reduced computational cost. For the cluster sizes investigated here ωB97X-3c and ωB97X-D3BJ/6-31++G(d,p) are comparable in accuracy and speed, but for larger systems ωB97X-3c might have an advantages in computational cost.

In addition to the geometries, the accuracy of the vibrational frequencies is also important for atmospheric cluster formation, as they are used to determine the thermal contribution to the free energy of the clusters. Therefore, we suggest that a study of the accuracy of the vibrational frequencies of a large representative test set should be performed in the future.

Acknowledgement

The authors thank Independent Research Fund Denmark grant number 9064-00001B 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/`.

Additionally Andreas Buchgraitz Jensen would like to thank Patrick Norman for the hospitality during a visit to KTH where a lot of the initial code used here was written and re-written.

Supporting Information Available

The following is available as supporting information:

- A table similar to Table 1, but showing the numbers for clusters labeled as "Different-FromReference".
- Figures similar to Figure 4 where we plot the two measurements used to distinguish cluster configurations. Plots are found both for clusters labeled as "SameAsReference" and "DifferentFromReference".
- A table showing the critical values used for the Anderson-Darling test for the Cauchy distribution, taken from Ref. 82.

Author Information

Corresponding Author

Jonas Elm - Department of Chemistry, Langelandsgade 140, Aarhus University, 8000 Aarhus C, Denmark; Phone: +45 28938085; Email: jelm@chem.au.dk

Authors

Andreas Buchgraitz Jensen - Department of Chemistry, Langelandsgade 140, Aarhus University, 8000 Aarhus C, Denmark; Email: buchgraitz@chem.au.dk

References

(1) 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.

(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) IPCC, 2021: 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., P. Zhai, A. Pirani, S.L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M.I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J.B.R. Matthews, T.K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, and B. Zhou (eds.)]. Cambridge University Press. In Press, doi:10.1017/9781009157896.

(5) Kirkby, J.; Amorim, A.; Baltensperger, U.; Carslaw, K. S.; Christoudias, T.; Curtius, J.; Donahue, N. M.; Haddad, I. E.; Flagan, R. C.; Gordon, H. et al. Atmospheric New Particle Formation from the CERN CLOUD Experiment. *Nat. Geosci.* **2023**, *16*, 948–957.

(6) 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.

(7) Dawson, M. L.; Varner, M. E.; Perraud, V.; Ezell, M. J.; Gerber, R. B.; Finlayson-Pitts, B. J. Simplified mechanism for new particle formation from methanesulfonic acid, amines, and water via experiments and ab initio calculations. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 18719–18724.

(8) Bork, N.; Elm, J.; Olenius, T.; Vehkamäki, H. Methane sulfonic acid-enhanced formation of molecular clusters of sulfuric acid and dimethyl amine. *Atmos. Chem. Phys.* **2014**, *14*, 12023–12030.

(9) Wang, M.; Kong, W.; Marten, R.; He, X.-C.; Chen, D.; Pfeifer, J.; Heitto, A.; Kontkanen, J.; Dada, L.; Kürten, A. et al. Rapid Growth of New Atmospheric Particles by Nitric Acid and Ammonia Condensation. *Nature* **2020**, *581*, 184–189.

(10) 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.

(11) 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.

(12) 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.

(13) 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.

(14) 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.

(15) Sipilä, M.; Sarnela, N.; Jokinen, T.; Henschel, H.; Junninen, H.; Kontkanen, J.; Richters, S.; Kangasluoma, J.; Franchin, A.; Peräkylä, O. et al. Molecular-scale Evidence of Aerosol Particle Formation Via Sequential Addition of HIO_3. *Nature* **2016**, *537*, 532–534.

(16) He, X.-C.; Tham, Y. J.; Dada, L.; Wang, M.; Finkenzeller, H.; Stolzenburg, D.; Iyer, S.; Simon, M.; Kürten, A.; Shen, J. et al. Role of Iodine Oxoacids in Atmospheric Aerosol Nucleation. *Science* **2021**, *371*, 589–595.

(17) He, X.-C.; Simon, M.; Iyer, S.; Xie, H.-B.; Rörup, B.; Shen, J.; Finkenzeller, H.; Stolzenburg, D.; Zhang, R.; Baccarini, A. Iodine Oxoacids Enhance Nucleation of Sulfuric Acid Particles in the Atmosphere. *Science* **2023**, *382*, 1308–1314.

(18) Schervish, M.; Donahue, N. M. Peroxy Radical Chemistry and the Volatility Basis Set. *Atmos. Chem. Phys.* **2020**, *20*, 1183–1199.

(19) Simon, M.; Dada, L.; Heinritzi, M.; Scholz, W.; Stolzenburg, D.; Fischer, L.; Wagner, A. C.; Kürten, A.; Rörup, B.; He, X.-C. et al. Molecular Understanding of New-particle Formation from α-pinene between −50 and +25 °C. *Atmos. Chem. Phys.* **2020**, *20*, 9183–9207.

(20) Ehn, M.; Thornton, J. A.; Kleist, E.; Sipilä, M.; Junninen, H.; Pullinen, I.; Springer, M.; Rubach, F.; Tillmann, R.; Lee, B. et al. A Large Source of Low-Volatility Secondary Organic Aerosol. *Nature* **2014**, *506*, 476–479.

(21) Bianchi, F.; Kurtén, T.; Riva, M.; Mohr, C.; Rissanen, M. P.; Roldin, P.; Berndt, T.; Crounse, J. D.; Wennberg, P. O.; Mentel, T. F. et al. Highly Oxygenated Organic Molecules (HOM) from Gas-Phase Autoxidation Involving Peroxy Radicals: A Key Contributor to Atmospheric Aerosol. *Chem. Rev.* **2019**, *119*, 3472–3509.

(22) Riccobono, F.; Schobesberger, S.; Scott, C. E.; Dommen, J.; Ortega, I. K.; Rondo, L.; Almeida, J.; Amorim, A.; Bianchi, F.; Breitenlechner, M. et al. Oxidation Products of

Biogenic Emissions Contribute to Nucleation of Atmospheric Particles. *Science* **2014**, *344*, 717–721.

(23) 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.

(24) 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.

(25) Jokinen, T.; Sipilä, M.; Junninen, H.; Ehn, M.; Lönn, G.; Petäjä, J. H. T.; Mauldin III, R. L.; Kulmala, M.; Worsnop, D. R. Atmospheric Sulphuric Acid and Neutral Cluster Measurements using CI-APi-TOF. *Atmos. Chem. Phys.* **2012**, *12*, 4117–4125.

(26) Ehn, M.; Kleist, E.; Junninen, H.; Petäjä, T.; Lönn, G.; Schobesberger, S.; Dal, M. M.; Trimborn, A.; Kulmala, M.; Worsnop, D. R. et al. Gas Phase Formation of Extremely Oxidized Pinene Reaction Products in Chamber and Ambient Air. *Atmos. Chem. Phys.* **2012**, *12*, 5113–5127.

(27) Sipilä, M.; Sarnela, N.; Jokinen, T.; Junninen, H.; Hakala, J.; Rissanen, M. P.; Praplan, A.; Simon, M.; Kürten, A.; Bianchi, F. et al. Bisulfate - Cluster Based Atmospheric Pressure Chemical Ionization Mass Spectrometer for High-Sensitivity (< 100 ppqV) Detection of Atmospheric Dimethyl Amine: Proof-of-concept and First Ambient Data from Boreal Forest. *Atmos. Meas. Tech.* **2015**, *8*, 4001–4011.

(28) Berndt, T.; Richters, S.; Jokinen, T.; Hyttinen, N.; Kurtén, T.; Otkjær, R. V.; Kjaergaard, H. G.; Stratmann, F.; Herrmann, H.; Sipilä, M. et al. Hydroxyl Radical-induced Formation of Highly Oxidized Organic Compounds. *Nat. Commun.* **2016**, *7*, 13677.

(29) Lee, B. H.; Lopez-Hilfiker, F. D.; Mohr, C.; Kurtén, T.; Worsnop, D. R.; Thornton, J. A. An Iodide-adduct High-resolution Time-of-flight Chemical-ionization Mass Spectrometer: Application to Atmospheric Inorganic and Organic Compounds. *Environ. Sci. Technol.* **2014**, *48*, 6309–6317.

(30) Zapadinsky, E.; Passananti, M.; Myllys, N.; Kurtén, T.; Vehkamäki, H. Modeling on Fragmentation of Clusters inside a Mass Spectrometer. *J. Phys. Chem. A* **2019**, *123*, 611–624.

(31) Passananti, M.; Zapadinsky, E.; Zanca, T.; Kangasluoma, J.; Myllys, N.; Rissanen, M. P.; Kurtén, T.; Ehn, M.; Attouid, M.; Vehkamäki, H. How Well Can We Predict Cluster Fragmentation Inside a Mass Spectrometer? *Chem. Commun.* **2019**, *55*, 5946–5949.

(32) Zanca, T.; Kubečka, J.; Zapadinsky, E.; Passananti, M.; Kurtén, T.; Vehkamäki, H. Highly Oxygenated Organic Molecule Cluster Decomposition in Atmospheric Pressure Interface Time-of-flight Mass Spectrometers. *Atmos. Meas. Tech.* **2020**, *13*, 3581–3593.

(33) Alfaouri, D.; Passananti, M.; Zanca, T.; Ahonen, L.; Kangasluoma, J.; Kubečka, J.; Myllys, N.; Vehkamäki, H. A Study on the Fragmentation of Sulfuric Acid and Dimethylamine Clusters Inside an Atmospheric Pressure Interface Time-of-flight Mass Spectrometer. *Atmos. Meas. Tech.* **2022**, *15*, 11–19.

(34) Ortega, I. K.; Kupiainen-Määttä, O.; Kurtén, T.; Olenius, T.; Wilkman, O.; McGrath, M. J.; Loukonen, V.; Vehkamäki, H. From quantum chemical formation free energies to evaporation rates. *Atmos. Chem. Phys.* **2012**, *12*, 225–235.

(35) McGrath, M. J.; Olenius, T.; Ortega, I. K.; Loukonen, V.; Paasonen, P.; Kurtén, T.; Kulmala, M.; Vehkamäki, H. Atmospheric Cluster Dynamics Code: A Flexible Method for Solution of the Birth-Death Equations. *Atmos. Chem. Phys.* **2012**, *12*, 2345–2355.

(36) Elm, J.; Bilde, M.; Mikkelsen, K. V. Assessment of Binding Energies of Atmopsheric Clusters. *Phys. Chem. Chem. Phys* **2013**, *15*, 16442–16445.

(37) Elm, J.; Kristensen, K. Basis Set Convergence of the Binding Energies of Strongly Hydrogen-Bonded Atmospheric Clusters. *Phys. Chem. Chem. Phys* **2017**, *19*, 1122–1133.

(38) Schmitz, G.; Elm, J. Assessment of the DLPNO Binding Energies of Strongly Noncovalent Bonded Atmospheric Molecular Clusters. *ACS Omega* **2020**, *5*, 7601–7612, PMID: 32280904.

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

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

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

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

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

(44) Elm, J.; Mikkelsen, K. V. Computational Approaches for Efficiently Modelling of Small Atmospheric Clusters. *Chem. Phys. Lett.* **2014**, *615*, 26–29.

(45) Myllys, N.; Elm, J.; Kurtén, T. Density Functional Theory Basis Set Convergence of Sulfuric Acid-Containing Molecular Clusters. *Comp. Theor. Chem.* **2016**, *1098*, 1–12.

(46) Elm, J. An Atmospheric Cluster Database Consisting of Sulfuric Acid, Bases, Organics, and Water. *ACS Omega* **2019**, *4*, 10965–10974.

(47) Kubečka, J.; Besel, V.; Neefjes, I.; Knattrup, Y.; Kurtén, T.; Vehkamäki, H.; Elm, J. Computational Tools for Handling Molecular Clusters: Configurational Sampling, Storage, Analysis, and Machine Learning. *ACS Omega* **2023**, *8*, 45115–45128.

(48) 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.

(49) Werner, H.-J.; Knowles, P. J.; Knizia, G.; Manby, F. R.; Schütz, M. Molpro: a general-purpose quantum chemistry program package. *WIREs Computational Molecular Science* **2012**, *2*, 242–253.

(50) Werner, H.-J.; Knowles, P. J.; Manby, F. R.; Black, J. A.; Doll, K.; Heßelmann, A.; Kats, D.; Köhn, A.; Korona, T.; Kreplin, D. A. et al. The Molpro quantum chemistry package. *The Journal of Chemical Physics* **2020**, *152*, 144107.

(51) Werner, H.-J.; Knowles, P. J.; Celani, P.; Györffy, W.; Hesselmann, A.; Kats, D.; Knizia, G.; Köhn, A.; Korona, T.; Kreplin, D. et al. MOLPRO, version , a package of ab initio programs. see `https://www.molpro.net`.

(52) Werner, H.-J.; Manby, F. R.; Knowles, P. J. Fast linear scaling second-order Møller-Plesset perturbation theory (MP2) using local and density fitting approximations. *The Journal of Chemical Physics* **2003**, *118*, 8149–8160.

(53) Werner, H.-J.; Adler, T. B.; Manby, F. R. General orbital invariant MP2-F12 theory. *The Journal of Chemical Physics* **2007**, *126*, 164102.

(54) Győrffy, W.; Knizia, G.; Werner, H.-J. Analytical energy gradients for explicitly correlated wave functions. I. Explicitly correlated second-order Møller-Plesset perturbation theory. *The Journal of Chemical Physics* **2017**, *147*, 214101.

(55) Peterson, K. A.; Adler, T. B.; Werner, H.-J. Systematically convergent basis sets for explicitly correlated wavefunctions: The atoms H, He, B–Ne, and Al–Ar. *The Journal of Chemical Physics* **2008**, *128*, 084102.

(56) Adler, T. B.; Knizia, G.; Werner, H.-J. A simple and efficient CCSD(T)-F12 approximation. *The Journal of Chemical Physics* **2007**, *127*, 221106.

(57) Győrffy, W.; Werner, H.-J. Analytical energy gradients for explicitly correlated wave functions. II. Explicitly correlated coupled cluster singles and doubles with perturbative triples corrections: CCSD(T)-F12. *The Journal of Chemical Physics* **2018**, *148*, 114104.

(58) Neese, F. Software update: The ORCA program system—Version 5.0. *WIREs Comput. Mol. Sci.* **2022**, *12*, e1606.

(59) Grimme, S.; Hansen, A.; Ehlert, S.; Mewes, J.-M. r2SCAN-3c: A "Swiss army knife" composite electronic-structure method. *J. Chem. Phys.* **2021**, *154*, 064103.

(60) Brandenburg, J. G.; Bannwarth, C.; Hansen, A.; Grimme, S. B97-3c: A Revised Low-cost Variant of the B97-D Density Functional Method. *J. Chem. Phys.* **2018**, *148*, 064104.

(61) Najibi, A.; Goerigk, L. The Nonlocal Kernel in van der Waals Density Functionals as an Additive Correction: An Extensive Analysis with Special Emphasis on the B97M-V and ωB97M-V Approaches. *J. Chem. Theory Comput.* **2018**, *14*, 5725–5738, PMID: 30299953.

(62) Perdew, J. P.; Chevary, J. A.; Vosko, S. H.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. Atoms, Molecules, Solids, and Surfaces: Applications of the Generalized Gradient Approximation for Exchange and Correlation. *Phys. Rev. B* **1992**, *46*, 6671–6687.

(63) Müller, M.; Hansen, A.; Grimme, S. wB97X-3c: A composite range-separated hybrid DFT method with a molecule-optimized polarized valence double-zeta basis set. *The Journal of Chemical Physics* **2023**, *158*, 014103.

(64) Müller, M. ORCA4wB97X-3c. `https://github.com/grimme-lab/ORCA4wB97X-3c`. Accessed July 3, 2024.

(65) Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. Extended tight-binding quantum chemistry methods. *WIREs Comput. Mol. Sci.* **2020**, *11*, e01493.

(66) Semiempirical extended tight-binding program package xtb. `https://github.com/grimme-lab/xtb`. Accessed June 27, 2024.

(67) 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.

(68) 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.

(69) Knattrup, Y.; Kubečka, J.; Wu, H.; Jensen, F.; Elm, J. Reparameterization of GFN1-xTB for atmospheric molecular clusters: applications to multi-acid–multi-base systems. *RSC Adv.* **2024**, *14*, 20048–20055.

(70) Gaussian 16, Revision A.03, 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, et al., Gaussian, Inc., Wallingford CT, 2016.

(71) Stewart, J. Optimization of Parameters for Semiempirical Methods V: Modification of NDDO Approximations and Application to 70 Elements. *J. Mol. Model.* **2007**, *13*, 1173–1213.

(72) Stewart, J. Optimization of Parameters for Semiempirical Methods VI: More Modifications to the NDDO Approximations and Re-optimization of Parameters. *J. Mol. Model.* **2013**, *19*, 1–32.

(73) Goerigk, L.; Hansen, A.; Bauer, C.; Ehrlich, S.; Najibi, A.; Grimme, S. A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions. *Phys. Chem. Chem. Phys.* **2017**, *19*, 32184–32215.

(74) Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. Benchmark database of accurate (MP2 and CCSD(T) complete basis set limit) interaction energies of small model complexes, DNA base pairs, and amino acid pairs. *Phys. Chem. Chem. Phys.* **2006**, *8*, 1985–1993.

(75) Donchev, A. G.; Taube, A. G.; Decolvenaere, E.; Hargus, C.; McGibbon, R. T.; Law, K.-H.; Gregersen, B. A.; Li, J.-L.; Palmo, K.; Siva, K. et al. Quantum chemical benchmark databases of gold-standard dimer interaction energie. *Sci. Data* **2021**, *8*.

(76) Burns, L. A.; Faver, J. C.; Zheng, Z.; Marshall, M. S.; Smith, D. G. A.; Vanommeslaeghe, K.; MacKerell, J., Alexander D.; Merz, J., Kenneth M.; Sherrill, C. D. The BioFragment Database (BFDb): An open-data platform for computational chemistry analysis of noncovalent interactions. *The Journal of Chemical Physics* **2017**, *147*, 161727.

(77) Řezáč, J.; Riley, K. E.; Hobza, P. S66: A Well-balanced Database of Benchmark Interaction Energies Relevant to Biomolecular Structures. *Journal of Chemical Theory and Computation* **2011**, *7*, 2427–2438.

(78) Hättig, C.; Tew, D. P.; Köhn, A. Communications: Accurate and efficient approximations to explicitly correlated coupled-cluster singles and doubles, CCSD-F12. *The Journal of Chemical Physics* **2010**, *132*, 231102.

(79) Helgaket, T.; Jørgensen, P.; Olsen, J. *Molecular electronic-structure theory*; John Wiley & Sons, Ltd., 2000; ISBN: 9781118531471.

(80) Riley, K. E.; Platts, J. A.; Řezáč, J.; Hobza, P.; Hill, J. G. Assessment of the Performance of MP2 and MP2 Variants for the Treatment of Noncovalent Interactions. *The Journal of Physical Chemistry A* **2012**, *116*, 4159–4169.

(81) Yousaf, K. E.; Peterson, K. A. Optimized auxiliary basis sets for explicitly correlated methods. *The Journal of Chemical Physics* **2008**, *129*, 184108.

(82) Stephens, M. A. In *Goodness-of-fit techniques*; D'Agostino, R. B., Stephens, M. A., Eds.; Marcel Dekker, 1986; ISBN: 0-8247-8705-6.

(83) Barlow, R. J. In *Statistics - A Guide to the Use of Statistical Methods in the Physical Sciences*; D. J. Sandiford, A. C. P., F. Mandl, Ed.; Wiley, 1989; ISBN: 0-471-92294-3.

(84) Chernoff, H.; Gastwirth, J. L.; Johns, M. V. Asymptotic Distribution of Linear Combinations of Functions of Order Statistics with Applications to Estimation. *The Annals of Mathematical Statistics* **1967**, *38*, 52–72.

(85) Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A* **1976**, *32*, 922–923.

(86) Katz, M. G.; Tall, D. A Cauchy-Dirac Delta Function. *Found Sci* **2013**, *18*, 107–123.

(87) Kurfman, L. A.; Odbadrakh, T. T.; Shields, G. C. Calculating Reliable Gibbs Free Energies for Formation of Gas-PhaseClusters that Are Critical for Atmospheric Chemistry: $(H_2SO_4)_3$. *J. Phys. Chem. A* **2021**, *15*, 3169–3176.

(88) Wu, H.; Engsvang, M.; Knattrup, Y.; Kubecka, J.; Elm, J. Improved Configurational Sampling Protocol for Large Atmospheric Molecular Clusters. *ACS Omega* **2023**, *8*, 45065–45077.