ACS OMEGA

Open Access

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

Article

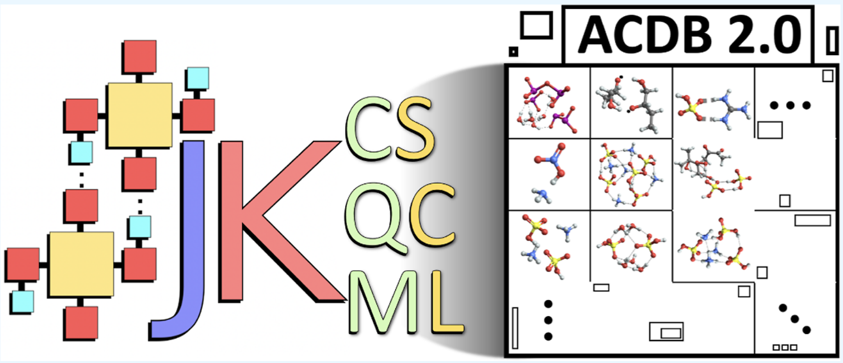
Computational Tools for Handling Molecular Clusters: Configurational Sampling, Storage, Analysis, and Machine Learning

Jakub Kubečka,* Vitus Besel, Ivo Neefjes, Yosef Knattrup, Theo Kurtén, Hanna Vehkamäki, and Jonas Elm

Cite This: *ACS Omega* 2023, 8, 45115−45128

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Computational modeling of atmospheric molecular clusters requires a comprehensive understanding of their complex configurational spaces, interaction patterns, stabilities against fragmentation, and even dynamic behaviors. To address these needs, we introduce the Jammy Key framework, a collection of automated scripts that facilitate and streamline molecular cluster modeling workflows. Jammy Key handles file manipulations between varieties of integrated third-party programs. The framework is divided into three main functionalities: (1) Jammy Key for configurational sampling (JKCS) to perform systematic configurational sampling of molecular clusters, (2) Jammy Key for quantum chemistry (JKQC) to analyze commonly used quantum chemistry output files and facilitate database construction, handling, and analysis, and (3) Jammy Key for machine learning (JKML) to manage machine learning methods in optimizing molecular cluster modeling. This automation and machine learning utilization significantly reduces manual labor, greatly speeds up the search for molecular cluster configurations, and thus increases the number of systems that can be studied. Following the example of the Atmospheric Cluster Database (ACDB) of Elm (ACS Omega, 4, 10965−10984, 2019), the molecular clusters modeled in our group using the Jammy Key framework have been stored in an improved online GitHub repository named ACDB 2.0. In this work, we present the Jammy Key package alongside its assorted applications, which underline its versatility. Using several illustrative examples, we discuss how to choose appropriate combinations of methodologies for treating particular cluster types, including reactive, multicomponent, charged, or radical clusters, as well as clusters containing flexible or multiconformer monomers or heavy atoms. Finally, we present a detailed example of using the tools for atmospheric acid−base clusters.

1. INTRODUCTION

Studying the formation and stability of molecular clusters has been of interest in many scientific domains, such as atmospheric chemistry,[1−6] biology,[7,8] astronomy,[9,10] and material science.[11,12] A particularly illustrative example is the study of molecular clusters formed in the atmosphere via gas-to-particle conversion, which is the first step in the formation of secondary atmospheric aerosol particles.[13−20] Once formed, these aerosols have a significant impact on climate and air quality and thus also on human health.[21−23]

Molecular clusters can be experimentally detected,[24−26] but cluster observations are generally complicated due to a variety of issues such as the typically very low concentrations (often below the detection limit of instruments), the changes a cluster undergoes inside the instruments,[27,28] and clusters being too small to detect. Many key properties strongly depend on the cluster configuration, which is a priori unknown and is seldom directly measurable. Hence, computational chemistry is an important additional tool to study clusters. Typically, quantum chemical calculations are required to obtain accurate cluster

Received: September 25, 2023
Revised: October 25, 2023
Accepted: October 26, 2023
Published: November 14, 2023

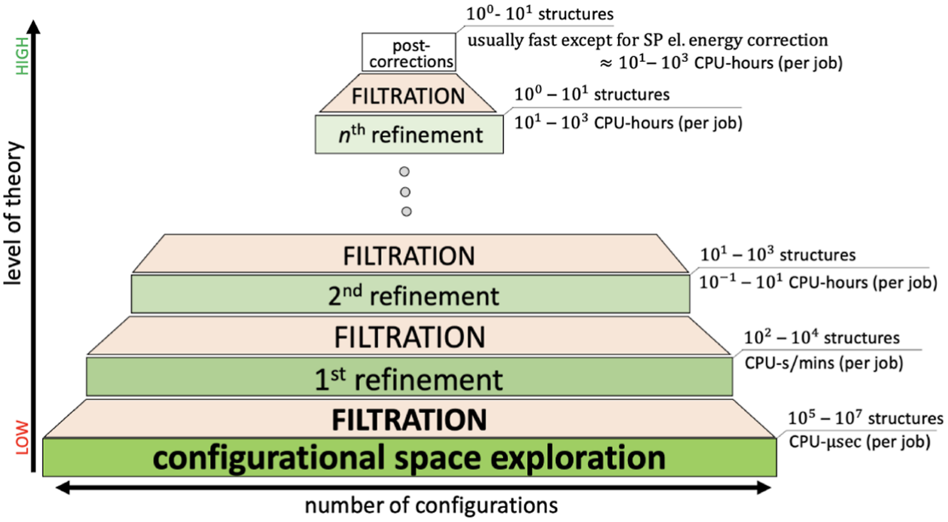


Figure 1. Schematic diagram of the bottom-up approach for conformational sampling.[50] This figure has been adapted from the dissertation of Kubečka.[51]

geometries, energies, charge distribution, and so on, which are used to understand the cluster's stability against fragmentation/evaporation or its potential growth into larger clusters/particles. Unfortunately, quantum chemical (QC) calculations are computationally expensive and the cost steeply grows with cluster size. Configurational sampling, the process of searching for the most relevant cluster configurations, is another bottleneck in molecular cluster studies, as the configurational space quickly grows in complexity with the size of the cluster and the flexibility of its monomers. Several programs exist to explore the vast configurational space of clusters (e.g., ABCluster,[29,30] OGOLEM,[31] and CREST[32]). Additionally, many programs for performing different types of QC calculations are available (e.g., Gaussian,[33] ORCA,[34,35] XTB,[36−39] and Turbomole[40]). Configurations can be manually passed from one program to another, but this is cumbersome and error-prone. Therefore, we present the Jammy Key for configurational sampling (JKCS) script that interfaces with the most commonly used third-party programs in the molecular cluster community, manages job submissions to computer clusters via the SLURM job scheduler, and handles the manipulation of the large number of files produced during the process. It further offers tools for data storage and analysis such as filtering, extraction of cluster properties from the output of the QC programs, and QC postcorrection calculations.

Machine learning (ML) methodologies have proven to be highly advantageous due to their capacity to replace time-intensive QC calculations. Several recent studies have harnessed ML techniques to investigate molecular clusters.[41−48] To assist in the creation of ML models for molecular cluster studies, we introduce Jammy Key for machine learning (JKML). JKML streamlines the training of ML models and their various applications such as predictions of energy/forces and even the creation of ML-based calculators. These calculators can effectively replace QC programs, enabling swift geometry optimization and molecular dynamics simulations, closely replicating the potential energy surface that was used to train the ML calculator. To train ML models, a considerable database of training data is necessary. We have, therefore, upgraded the Atmospheric Cluster Database (ACDB), originally introduced by Elm.,[49] to ACDB 2.0. In the new ACDB 2.0, cluster properties are stored in a single compressed file, which is easily manipulated by JKQC to use within JKML. The combination of JKCS, JKQC, JKML, and ACDB 2.0 provides the necessary tools to efficiently model a large variety of molecular clusters.

2. METHODOLOGY

We summarize the working principles of the three main functionalities of the Jammy Key framework: Jammy Key is used for configurational sampling (JKCS), quantum chemistry (JKQC), and machine learning (JKML). Additionally, we discuss how the Jammy Key framework is used to create an improved database for atmospheric clusters (ACDB 2.0).

2.1. Configurational Sampling. Direct examination of complex cluster configurational spaces at a desired high level of theory costs an immense amount of computational resources. Hence, the most common general strategy for configurational sampling (CS) is the so-called bottom-up approach, as illustrated in (Figure 1).[50] This approach is sometimes also referred to as the "building up" principle, but should be distinguished from strategies where the properties of the N-cluster need to be known to calculate the properties of the $(N + 1)$-cluster (an example of such a method is that introduced by Kildgaard et al.[52,53] to study hydration of dry molecular clusters). The bottom-up approach explores the configurational space using fast but less accurate methods, where only promising candidates are carried through several steps of reoptimization at higher levels of theory and filtering up to the desired level of theory. This methodology determines the workflow of the JKCS scripts.

2.1.1. System Setup. System setup involves preparation of the input file (JKCS0_copy copies the default input file into the working directory), where the cluster composition, charge, and spin multiplicity are defined. Typically, individual molecules, denoted as monomers, are used as the initial building blocks. Depending on the type of exploration, either flexible or rigid monomers are used. In the former case, only one conformer for each species needs to be supplied. In the latter case, including the lowest energy conformer of each monomer is a good starting point but adding other conformers and/or assorted protonation states improves the exploration.[54] Figure 2 depicts such a combination of rigid monomers to

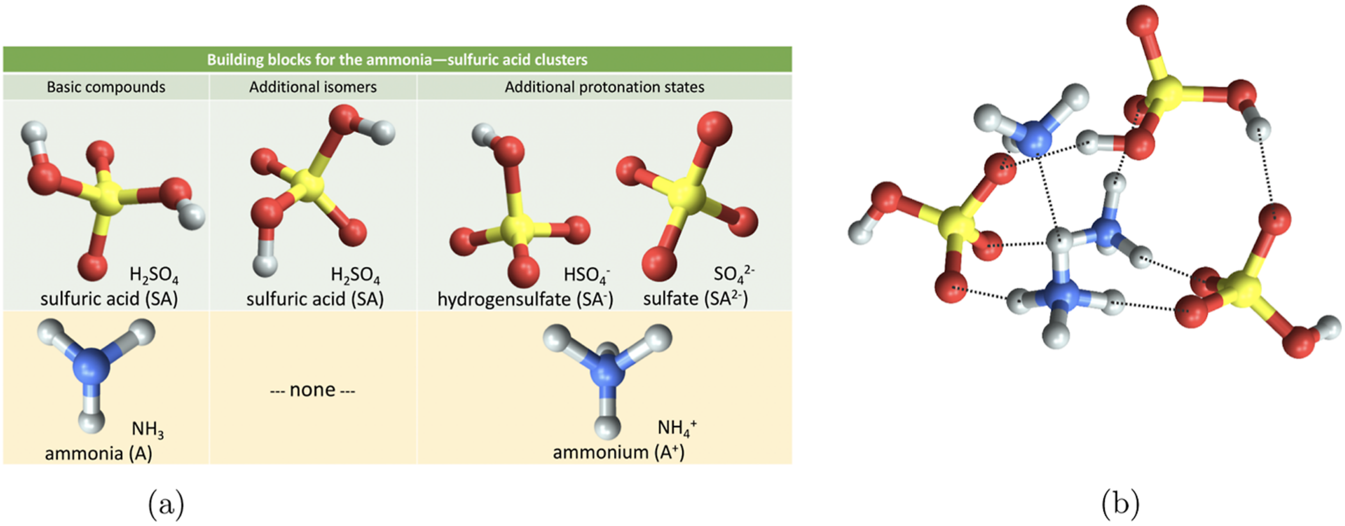


Figure 2. Illustration of all possible building block structures when constructing any ammonia—sulfuric acid clusters. Color coding: H (white), S (yellow), O (red), and N (blue).

construct the $(H_2SO_4)_3(NH_3)_3$ cluster. JKCS contains these building blocks for the most common atmospheric cluster forming acid and base molecules and takes care of combining all feasible monomer conformations and protonation states while adhering to stipulated criteria regarding cluster size (number of molecules) and charge. Further, JKCS1_prepare creates one folder for the CS of each cluster type.

2.1.2. Exploration. Configurational space exploration is generally performed at a low level of theory, that is, molecular mechanics or semiempirical methods (e.g., GFN-xTB[36−39]). JKCS (JKCS2_explore) currently communicates with two commonly used configuration space exploration programs: ABCluster[29,30] and CREST.[32] ABCluster employs the genetic artificial bee colony algorithm[55] and can be used with either rigid or flexible monomers. Rigid monomer exploration, often done at the molecular mechanics level, allows for inexpensive and fast exploration of large configurational spaces. The monomer rigidity guarantees that unwanted reactions do not take place (at least at the exploration stage) but also prevents proton transfers or conformational changes essential for cluster stability. Hence, introducing a combination of various building blocks in the system setup allows for a more thorough exploration. Starting from flexible monomers, ABCluster offers a slower exploration combining both cluster configuration and monomer conformation, and this search is typically performed using the GFN-xTB method. In CREST, the configurational space is explored through metadynamics simulations, again often using the GFN-xTB method. The choice of method is highly dependent on the studied cluster. At the end of the exploration, the energetically lowest-lying minimum structures are saved for further refinement.

2.1.3. Refinement. The JKCS3_run script allows communication with QC programs to refine the cluster geometries and energies. Since the number of trial structures obtained from the exploration step can be enormous, the first optimization step should ideally be performed using a computationally affordable method. One of the extended tight-binding (xTB) semiempirical methods implemented in the XTB program[36−39] is a robust choice for many systems, though caution is needed for reactive and radical systems. The PM6[56] and PM7[57] methods offer similar functionality. Subsequent single-point energy refinement or geometry reoptimization can be performed using composite electronic methods such as B97-3c[58] and r^{2}SCAN-3c.[59] For instance, Engsvang et al.[60,61] and Wu et al.[62] showed the applicability of these methods to large (up to 30 molecules) sulfuric acid and ammonia clusters. Nevertheless, a higher level of theory is often required to obtain accurate cluster geometries. For instance, density functional theory (DFT) methods, such as ωB97X-D/6-31++G(d,p),[63] have been successfully used to describe inorganic molecular clusters.[6] JKCS communicates with third-party programs such as XTB,[36−39] ORCA,[34,35] and Gaussian[33] and manages the calculation communication. The Jammy Key framework allows us to perform jobs on the user's local/login computer or on a computer cluster via SLURM job scheduler submission while offering various ways of job distributions, serializations, and parallelizations.

2.1.4. Data Filtering. Filtering is needed to reduce the set of structures passed on by each step as the computational cost per structure can increase by many orders of magnitude from one step to the next. Filtering should, at a minimum, remove redundant (identical or nearly identical configurations) and energetically high-lying configurations, as well as obviously unphysical structures (e.g., clusters that have fragmented or undergone unwanted reactions). Caution is advised when applying energetic filtering using specific cutoff values. For instance, low-level QC methods may predict the relative energies incorrectly, and a good filtering algorithm should take this into account. Therefore, the choice of appropriate filtering criteria is as important as the choice of methods at different levels of the bottom-up approach.

Here, the Jammy Key framework uses the JKQC script (as further described in the next chapter). JKQC allows users to filter structures based on both energetic and structural criteria, for example, the radius of gyration can be used to filter out molecular clusters that have fragmented during optimization. One can, for instance, filter out all structures that have a radius of gyration greater than 10 Å, and energy of x kcal/mol higher than the lowest energy found. Appropriate values for x require a benchmark examination, as they depend on the type of system, the cluster size, as well as the level of theory used at the step preceding the filtering. For example, Kubečka et al.[54] used a threshold of 5*M* kcal/mol for filtering after the XTB optimization step in their study on very strongly bound sulfuric acid−guanidine clusters, where *M* is the number of molecules in the cluster. However, lower values (e.g., 2.5*M* kcal/mol) may be appropriate for more weakly bound clusters. The energy threshold should use a high enough cutoff to account for energy reordering due to differences between the different QC methods and possible reordering due to postcorrections.

If two clusters have identical/similar properties, then only one of the configurations should be passed to the next step. This duplicate check can be done by comparing the chosen set of cluster properties such as the radius of gyration R_g, cluster energy *E*, or dipole moment μ. A slower but more accurate option is to compare root-mean-squared displacement (RMSD) between two identically oriented structures, which utilizes a modified version of the ArbAlign program.[64] Finally, we provide the *selection* method introduced by Kubečka et al.,[54] which selects a subset of the most distinct configurations from a large data set. This representative subset of configurations might result in not finding exactly the global minimum at the desired final level of theory; however, the best of the selected structures will be close to the real global minimum structure. Such a method is especially useful for large clusters, for which many low-lying energy minima are thermodynamically populated due to small energy gaps between them.

2.1.5. Postcorrections. Another category of data analysis performed via the JKQC script is the post-QC corrections. These can be separated into several subgroups:

- Thermal corrections, performed using the same method as used for geometry optimization, involve mainly vibrational frequency calculations and their contribution to the partition function along with the translational and rotation partition function calculation. Here, JKQC offers to check or correct for:
  (1) Imaginary vibrational frequencies as they indicate that the geometry was not optimized to a minimum. In that case, we attempt several geometry reoptimizations and, if unsuccessful, discard the structure.
  (2) Low-vibrational frequencies, as showed by Grimme,[65] are caused by treating the vibrations as harmonic, and can lead to unrealistically low free energies. The quasi-harmonic approximation (QHA) corrects the rigid-rotor harmonic oscillator (RRHO) by replacing low-vibrational contributions to entropy with internal rotational modes.
  (3) Vibrational anharmonicity is often just corrected by a scaling factor typical for each QC method applied to all vibrations.[66] We note that actual anharmonic vibrational frequencies can be calculated with many QC programs, but this is accompanied by numerical stability issues and is rarely feasible or cost-effective for molecular clusters.
  (4) Temperature in QC programs is by default set to 298.15 K (adjustable). JKQC swiftly recalculates thermal corrections at any temperature.
  (5) The rotational symmetry number is often not correctly recognized by QC programs due to computer precision or too high symmetry quantifier thresholds. Although often less important for clusters, the thermodynamic properties of monomers should be either calculated at the correct symmetry or corrected for the rotational symmetry error after QC programs. Here, we recommend the SYMMOL[67] program, which suggests the maximum symmetry group at a given tolerance.
- Electronic energy may need to be corrected by a single-point electronic energy calculation. This is not needed if a high-level method with a large enough basis set is used. Although the DFT method of our choice, ωB97X-D/6-31++G(d,p)[63], provides accurate geometries and thermal contributions, the electronic energy must be corrected at a higher level of theory. For instance, cost-effective domain-based local pair natural orbital (DLPNO) variants of coupled cluster methods such as DLPNO-CCSD(T_0)/aug-cc-pVTZ[68−70] with Normal-PNO criteria can be utilized to calculate Gibbs free energies as[52,66]

$$\Delta G = G^{\mathrm{DFT}} - E_{\mathrm{el}}^{\mathrm{DFT}} + E_{\mathrm{el}}^{\mathrm{DLPNO}} \tag{1}$$

$$\Delta G = -k_{\mathrm{B}}T\ln\left(\sum_{i\in\mathrm{minima}}\exp\left(\frac{-\Delta G_i}{k_{\mathrm{B}}T}\right)\right) \tag{2}$$

- Several energetically close but distinct configuration minima may be populated. In such a case, the lowest free energy minimum does not always sufficiently represent the average cluster structure or its properties. The entropy contribution of all energetically low-lying minima can be accounted for using the Boltzmann distribution, resulting in an average Gibbs free energy.[71] Nevertheless, here we assume that clusters have a crystal-like behavior and presume that the transition between different minima has a minor effect on the free energies. The problem of clusters with liquid-like behavior, populating many different low-lying free energy minima, should be addressed in future studies.

2.2. QC Data Handling, Storing, and Analysis. We present the Jammy Key for Quantum Chemistry (JKQC) Python script designed to store essential molecular cluster information into a Pandas[72,73] data frame. Structure parsing is accomplished with ASE,[74] while pertinent molecular properties are directly extracted from QC output files. This approach replaces numerous QC output files, potentially consuming multiple gigabytes of memory, with a single compact file of a few megabytes. Additionally, all the data filtering and postcorrections elaborated in the previous section can be easily performed using JKQC. JKQC is also automated to create input for other programs such as the Ion Mobility Software Suite (IMoS[75]) used to calculate collision cross sections and ion mobilities and the Atmospheric Cluster Dynamics Code (ACDC[76,77]) used to calculate cluster/particle formation rates based on cluster population dynamics. This automation reduces the human errors that accompany the manual construction of these files.

Elm[49] recently established the ACDB. This database contains clusters composed of molecules responsible for atmospheric new particle formation (NPF): acids (e.g., sulfuric and nitric), bases (e.g., ammonia and dimethylamine), water,

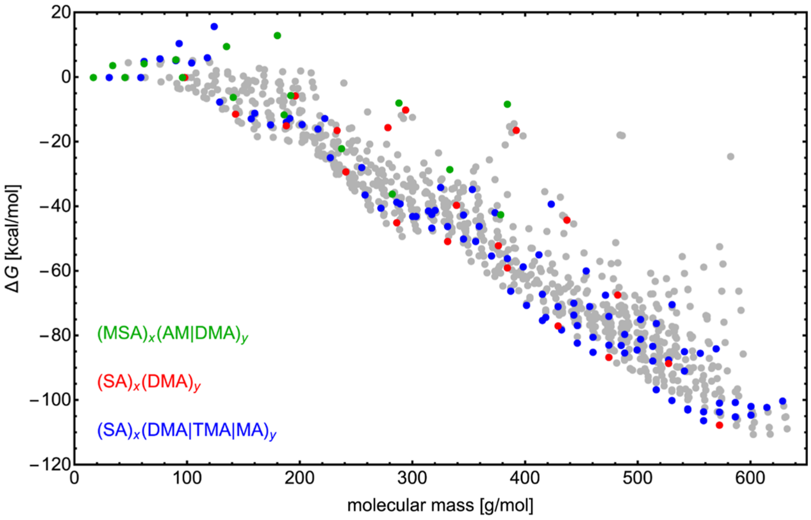


Figure 3. Lowest binding free energies of all cluster types stored in the Atmospheric Clusters Database 2.0 at DLPNO-CCSD(T_0)/aug-cc-pVTZ//ωB97X-D/6-31++G(d,p) level of theory. Three cluster types are highlighted with different colors: where MSA = methanesulfonic acid, SA = sulfuric acid, AM = ammonia, DMA = dimethylamine, MA, methylamine, and TMA = trimethylamine.

and, as of yet, only a few organic molecules. We gathered the ACDB database alongside additional cluster structures and properties from over 30 publications into a new database, ACDB 2.0. Rather than the SDF files of the original ACDB, ACDB 2.0 is constructed as compressed pickle files that contain a large number of cluster properties and are easily read and manipulated by JKQC. We will continuously update the database with the most recently published data. The database currently encompasses more than 1 million entries, spanning various levels of theory. For instance, ~100k single-point energies are now available at the ωB97X-D/6-31++G(d,p)[63] level of theory. ACDB 2.0 offers a comparison of different properties between a large set of atmospherically relevant clusters. Figure 3, for instance, depicts the lowest binding free energies of ~1.5k different cluster types at the DLPNO-CCSD(T_0)/aug-cc-pVTZ//ωB97X-D/6-31++G(d,p) level of theory with NormalPNO criteria and with QHA and anharmonicity corrections applied. Highlighting three specific cluster types, the figure underscores the potential stability of sulfuric acid−base molecular clusters, illustrating their central contribution to atmospheric NPF. ACDB 2.0 furthermore acts as a foundation for ML applications.

2.3. Machine Learning.

2.3.1. ML Model. The recent explosion of ML utilization in quantum chemistry has shown that ML potentials can mimic potential energy surfaces within the chemical accuracy of QC methods. There are several ML techniques for regression tasks, such as artificial neural networks (NN),[78] Gaussian process regression (GPR),[79] and kernel ridge regression (KRR),[80] each with their own strengths and weaknesses.[81] The first task in creating an ML model is choosing the molecular representation for the studied system. Such a representation should be invariant to transformations that do not change the particular property (translation, rotation, mirroring, or nuclear permutation). It should uniquely describe the system and be continuous and, ideally, differentiable.[82] Commonly used representations for molecular systems are the Coulomb Matrix (CM),[83] Bag of Bonds (BoB),[84] Many-Body Tensor Representation (MBTR),[85] Smooth Overlap of Atomic Positions (SOAP),[86] FCHL18/19,[87,88] and those integrated in NN architectures such as SchNet[89] and PaiNN.[90] Several ML studies have already been conducted for atmospherically relevant molecular systems. Jääskeläinen showed that ML approaches are useful to improve cluster structure selection and sampling in general.[48] NNs have been used to model large sulfuric acid−dimethylamine clusters[47] and the NN potential ANI-2x[91] has been benchmarked for small dimer clusters.[92] KRR/GPR has been used to predict cluster binding energies,[44−46,61] saturation vapor pressures of organic molecules,[41,42] and chemical potentials of organic molecules in atmospherically relevant solutions.[43]

Our ML-oriented subpackage, JKML, offers an interface between the JKQC-constructed database files (e.g., those stored in ACDB 2.0) and two ML programs, quantum machine learning (QML[93]) and SchNetPack.[94,95] In the procedure, XYZ coordinates are extracted and together with the property of interest (e.g., electronic energy, forces, or mobility) are stored in a database. Subsequently, JKML uses QML or SchNetPack to perform the training, validation, and testing of the predicted property or its difference from a reference state. In the case of energies, these can be atomization energies for molecules, or binding energies for molecular clusters

$$\Delta E = E_{\mathrm{cluster}} - \sum_{i \in \mathrm{monomers}} E_i \tag{3}$$

In our previous work,[44,46] we showed that utilizing Δ-ML,[96] that is, predicting the difference in binding energy between a low and high QC method, increases the accuracy of the model compared to direct-ML. For instance, the difference in electronic binding energy is calculated as

$$\Delta\Delta E^{\mathrm{HIGH|LOW}} = \Delta E^{\mathrm{HIGH}} - \Delta E^{\mathrm{LOW}} \tag{4}$$

KRR can potentially achieve higher accuracy than NN for small databases. On the other hand, NN is accurate and fast for large training databases. Since KRR becomes computationally demanding with increasing database size, training an ML model on energy gradient/forces, that is, 3*N*-times more variables, seems more suitable for NN. However, several KRR-based methods suitable for GPU/TPU also exist (e.g., QML-lightning[97] and sGDML[98]). For now, JKML allows for the training of the aforementioned NN-based potential utilizing the forces extracted from QC via JKQC. The trained model can be used for geometry optimizations and fast MD simulations. However, for accurate and fast modeling, a training database must be constructed for the system at hand. Additionally, the greater the number of atom types in the studied system, the greater the required training database.

2.3.2. Categorization Trick. JKML allows multinode parallelization for kernel construction within the KRR calculations. However, even with this parallelization, it becomes computationally demanding to train on more than 100,000 data points. For each test structure, we offer training data set reduction based on structure similarity, thus removing the need for training on redundant or unnecessary structures. Here, we utilize the MBTR[85] representation, as implemented in the DScribe[99] library, to calculate the distribution of atom-specific bonds and bond angles ρ_{MBTR} of each structure x_i. Furthermore, we define the similarity between two structures by calculating the overlap of the two distributions at given bond lengths ($r \in \langle 0.7, 2\rangle$ Å) and eventually also bond angles ($\alpha \in \langle 0, 2\pi\rangle$ rad)

$$\Delta(x_i, x_j) = \sum_{\omega}^{r,\cos(\alpha)} \int \left|\rho_{\mathrm{MBTR},\omega}(x_i) - \rho_{\mathrm{MBTR},\omega}(x_j)\right| \mathrm{d}\omega \qquad (5)$$

The MBTR representation is a discretized (we use a 100-unit grid for bonds and a five-unit grid for angles) function corresponding to a sum of Gaussian functions with a small deviation of $\sigma = 10^{-9}$, an exponential weighting function of 0.5, and a minimum threshold of 3×10^{-3}. Gaussian functions are situated around each bond length and/or bond angle for specific atoms. Thus, a value of $\Delta(x_i, x_j) = 0$ indicates that the structures are identical, and low values determine high similarity. Consequently, one can speed up the ML modeling of a target configuration by training only on a small data set.

3. APPLICATION AND DISCUSSION

3.1. Configurational Sampling Obstacles. Cluster formation inevitably involves both enthalpy (ΔH) and entropy (ΔS) changes, and hence the clustering free energy (e.g., $\Delta G = \Delta H - T\Delta S$) increases with increasing temperature (T). Weakly bonded clusters thus typically require a low temperature and/or high vapor concentrations to form in the gas phase, while strongly bonded clusters may also be formed at room temperature and trace concentrations of the vapors. Figure 4 illustrates several cluster systems ordered by their binding strength. Examples include noble gas atomic clusters (e.g., helium and argon) and molecular clusters like carbon dioxide (CO_2), which are stabilized by weak London dispersion interactions. Other clusters composed of methanol (CH_3OH), butanol (C_4H_7OH), or water (H_2O) are bound relatively more strongly, primarily via hydrogen bonds. On the opposite end, acid−base, NaCl salt, and ionic-liquid clusters are bonded by robust ion−ion Coulomb interactions.

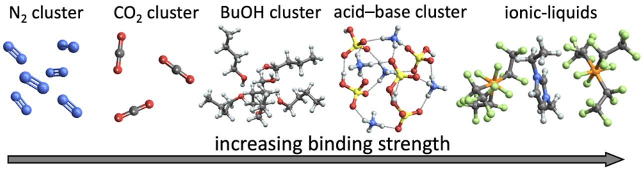


Figure 4. Several examples of molecular clusters sampled by JKCS sorted with respect to their binding energy. Color coding: H (white), O (red), C (gray), S (yellow), N (blue), P (orange), F (lime).

Neither the binding energy nor the reactivity (e.g., proton transfer) alone is sufficient to determine how straightforward the CS process is for a given clustering system. Over several years, JKCS has been used to study various clusters and their properties.[3,6,44,46,54,62,100−113] Based on these studies, we present typical examples of molecular cluster properties that complicate CS compared to a reference case of a nonreactive, one-component, crystal-like cluster with a single low-energy configuration, formed from rigid, closed-shell monomers with only one conformer. This reference case can be, for example, a water cluster at a very low temperature, that is, an ice crystal. Clusters fall into one or several of the following categories:

Multicomponent:

e.g., $(H_2SO_4)_x(NH_3)_y(CH_3NH_2)_z((CH_3)_2NH)_u(H_2O)_w$.

JKCS can construct clusters with an arbitrary number of components. However, the number of dimensions to study increases with the number of components, especially when various different combinations of $\{x, y, z, u, w\}$ need to be examined. There is no general solution to this problem. One could potentially lump some monomers[114] together based on similarity or use ML methods to accelerate the CS.[46,105] Another option would be to use a lower level of theory along, e.g., the water (w) axis.

Reactive:

e.g., $(\text{organic})_x(H_2O)_y(O_3)$, $(R{=}O)_1(R{-}OO\cdot)_1$, $(H_2O)_x(NH_3)_y$.

Intra- and intermolecular reactions (e.g., oxidation, bond breaking, but also proton transfer) can occur within clusters if available through thermal fluctuations. Sampling potential reactants and products of the relevant reactions separately, utilizing reactive potentials within, for example, CREST,[32] or performing MD simulations are possible solutions. Transition state conformers could also be searched by fixing the reactive area.

Acid−base:

e.g., $(H_2SO_4)_x(NH_3)_y$, $(HNO_3)_x((CH_3)_2NH)_y(H_2O)_z$.

This is a subgroup of reactive clusters that undergo proton transfer within the cluster. Using combinations of conformer/protonation states in rigid monomer exploration enables a thorough CS while accounting for all possible proton transfers. An ABCluster[29,30] search with rigid monomers could also be followed by re-exploration around energetically low-lying structures within the reactive potential via CREST.[32]

Multiconformer/flexible monomers:

e.g., $(H_2SO_4)_x(C_2H_4(NH_2)_2)_y$, $(C_4H_9OH)_x$.

Some molecules can have a large number of conformers. Including all or as many conformers as possible in a rigid monomer exploration guarantees better exploration. We recommend more conformer combinations, that is, more parallel explorations, with short exploration times rather than one long thorough exploration with one conformer combination. Typically, the ABCluster[29,30] combined global-configuration and conformation search or CREST[32] search are suitable for this problem.

Metastable monomer conformers:

e.g., $(\text{organic})_x(H_2O)_y$, $(HIO_3)_x(HIO_2)_y$.

Monomers within clusters can take on configurations that are not stable minima in the gas phase but can exist and even dominate inside clusters, as they are stabilized by the cluster environment. These metastable monomer conformers should be manually constructed and included in the exploration step. They should not be preoptimized as they would only revert to

Table 1. Overview of Several Cluster Studies That Utilized the JKCS Package

single-component nonreactive clusters	
Carbon dioxide clusters at ∼40−90 K.[106] Achieving accurate binding energies required challenging QC calculations such as high-level theory (CCSD(T)) along with diverse corrections because lower methods struggled to precisely capture the weak dispersion interactions.	Figure 5A
Butanol (C_4H_7OH) clusters.[107] Butanol has multiple internal rotations, yielding multiple conformers. Clusters were formed through their random combinations. Modern DFT or even some semiempirical methods provide sufficiently accurate formation thermodynamics for these hydrogen-bonded clusters, eliminating the need for higher-level corrections.	Figure 5B
multicomponent nonreactive clusters	
Butanol and water condensing onto NaCl seed.[107,108] We used ABCluster conformation and configurational search. Note that CS of systems with more nonmixing physical phases (e.g., cluster formation on a surface) is currently available within ABCluster after introducing restrictions for phase mixing.[115]	Figure 5C
Diethylene glycol around ionic-liquid clusters.[109] Multiple conformers were introduced at the beginning of CS. With no proton transfer present, interactions were well described by DFT methods with large basis sets.	Figure 5D
Clusters of sulfuric acid and organic molecules.[110] A highly oxygenated organic molecule (HOM)[116] was only represented by one conformer to speed up the CS. Due to the inherent flexibility of organic molecules, employing multiple representative conformers (e.g., found via the Spartan program[117]) or using CREST[32] is advisable. Furthermore, assessing the reactivity within such clusters is crucial due to the presence of reactive functional groups, potentially participating in proton transfer.	Figure 5E
clusters containing heavy atom and metastable monomer	
Iodine-containing molecular clusters.[3,100] Iodic acid has a metastable conformation in clusters. We constructed an extra building block with varied proton orientations. Sequentially, advanced QC methods were used to account for relativistic effects. Conversely, a low level of theory was used for large clusters to obtain approximate configurations for collision cross-section estimation.	Figure 5F
reactive organic clusters	
Dimers of organic alkoxy radicals (formed in peroxy radical self-reactions).[111,112] ABCluster with multiple conformers per monomer was used to obtain trial structures further passed to higher-level theory, which is able to describe radical systems. Nevertheless, using CREST[32] is advisable for future studies.	Figure 5G
Accretion reactions on dust particle.[113] Dust particles were approximated by a representative molecule. Transition state modeling was performed at a high level of theory to describe the reaction energy barriers.	Figure 5H

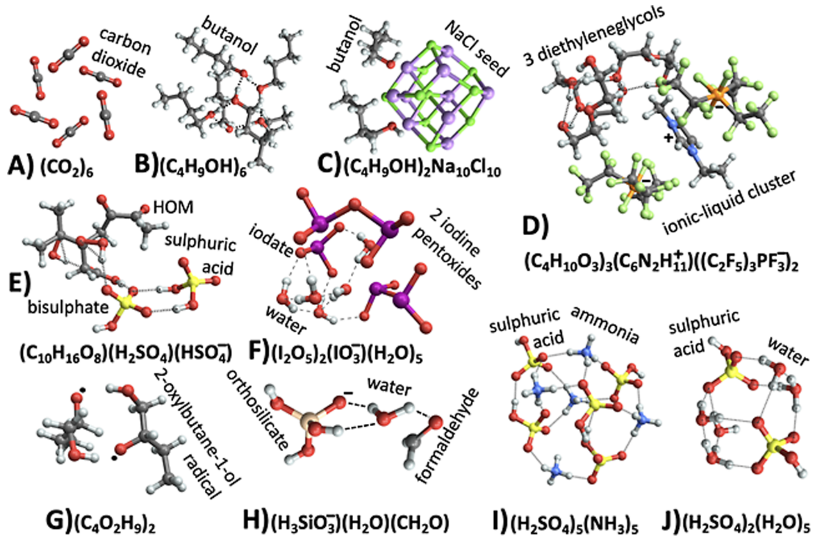


Figure 5. Examples of clusters corresponding to studies are presented in Table 1. Color coding: H (white), S (yellow), O (red), C (gray), I (purple), Na (mauve), Cl (green), N (blue), P (orange), F (lime), Si (cream).

the gas-phase minimum structure. CREST[32] might be, in some cases, more suitable for this problem.

Heavy atom(s):

e.g., $(HIO_3)_x(H_2O)_y$, $K_x[Pb(ligand)_y]$.

Low-level theory energy evaluations may fail for clusters containing heavy atoms (period five or higher in the periodic table) since they might lack a description of relativistic effects, polarization, or other heavy-element-related phenomena. Heavy-atom-related effects should be accounted for in the CS by at least including pseudopotentials during DFT calculations or scalar relativistic Hamiltonians.

Liquid-like:

e.g., $(H_2SO_4)_x(NH_3)_y(H_2O)_z$, or $(Ar)_x$ and $(CO_2)_x$ at low temperature.

The description of the thermodynamic properties of these clusters is difficult since we cannot use the superposition approximation of the lowest free energy minima. There is a (cluster-type-dependent) threshold temperature above which the energy barriers separating different local energy-minimum conformations become easy to overcome by thermal fluctuation. MD simulations can likely provide insight into this problem; however, further research on this topic is still required.

Charged clusters or ionized molecules:

e.g., $Cl^-(H_2O)_x$, $(C_xH_yO_z)\cdot NO_3^-$, $(H_2SO_4)_x(NH_3)_y$.

Clusters may be charged or contain ionic monomers even when the clusters or monomers are initially neutral (e.g., due to proton transfer; see acid−base clusters). Charges present in clusters cause inductive effects (electron flows) and charge delocalization over several neighboring molecules. Hence, additional polarization and diffuse basis set functions are needed. Moreover, if we require the charge to be localized on a specific cluster molecule (e.g., due to photoionization or charge transfer), additional QC techniques might be necessary such as constrained DFT and various fragmentation methods.

Unpaired electrons:

e.g., $[(CH_3)_3C{-}O\cdot]_2$, $(R{-}CO{-}OO\cdot)_1(H_2O)_x$.

Radical systems can usually be described only by QC methods (not force fields, FF). The potential energy surface can still be explored using FF to obtain a broad set of initial structures, but FF parameters should be carefully checked to prevent, for example, radical centers from being treated as ions. Nevertheless, exploration at, for example, the XTB level of theory is more suitable. Further, high-level QC methods should be used to describe these open-shell systems (freezing problematic parts might be helpful, especially during initial optimizations). The QC results should also be checked for spin contamination, where relevant.

We advise readers to have a look at the articles gathered in Table 1 and Figure 5. These articles used JKCS for various cluster systems, and the CS of these systems is presented along with technical details within the main text or the Supporting Information. However, JKCS has so far found the most applications for CS of atmospheric acid−base molecular clusters, which we focus on in detail in the next section.

3.2. JKCS Workflow for Acid−Base Clusters.

Under atmospheric conditions, clusters that grow to nanoparticles or aerosols from gas molecules often involve strong acids and bases. After collision, these molecules undergo proton transfer reactions forming strongly bound ion pairs (salt), exemplified by cases such as $(H_2SO_4)_2(NH_3)_2 \rightarrow (HSO_4^-)_2(NH_4^+)_2$. To accommodate the protonation states of monomers and multiconformer sulfuric acid during CS, the provided building blocks (Figure 2a) are employed. Also, an accurate QC examination of thermodynamic properties necessitates a high level of theory and extra-polarized and diffuse basis functions. These methodologies were used for studies of acid−base clusters such as sulfuric acid−ammonia clusters,[61,62,102] sulfuric acid−dimethylamine clusters,[62,104] systems involving trimethylamine oxide,[103] sulfuric acid−multibase clusters,[46] and even multiacid−multibase clusters.[105] For these cluster types, the CS procedure via a multistep funneling approach[6,54] has been well optimized. For instance, Knattrup et al.[105] recently used the workflow scheme depicted in Figure 6.

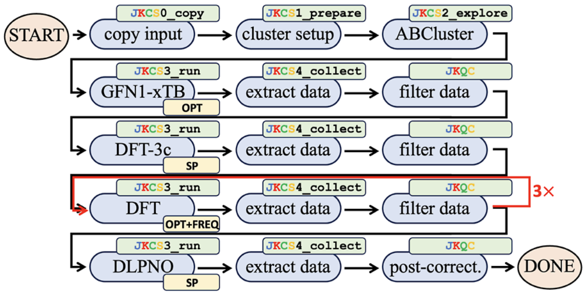


Figure 6. CS workflow used by Knattrup et al.[105]

Clearly, this multistep procedure would be cumbersome to do manually. Consequently, its full automation allowed us to explore hundreds of cluster systems. Nevertheless, we recommend that new users perform each command separately and examine its outcome first. Several molecular cluster benchmark studies[45,66,118−121] can be used to choose a suitable QC method. For instance, we currently use GFN1-xTB implemented in the XTB program,[36−39] DFT-3c such as B97-3c or r^{2}SCAN-3c implemented in ORCA,[34,35] DFT such as ωB97X-D/6-31++G(d,p)[63] implemented in the Gaussian[33] and ORCA programs, and DLPNO such as DLPNO-CCSD(T$_0$)/aug-cc-pVTZ[68−70] with NormalPNO criteria also implemented in the ORCA program. Note that we typically need to restart some DFT calculations, as the minimum was not found. This can be caused by the calculations not converging or by the presence of an imaginary vibrational frequency. The filtering of redundant structures (nonunique, fragmented, reacted, and with too high energy) and all subsequent postcorrections is implemented within JKQC and described in the manual.

Atmospheric molecular clusters often become quickly solvated by a few water molecules due to high air humidity. Therefore, the consideration of hydration becomes essential in the study of atmospheric clusters, which introduce water as an extra component, increasing the CS complexity. Water can also function as both an acid and a base, potentially introducing reactivity through proton transfer reactions. For simplicity, hydration was often omitted in studies involving atmospheric acid−base clusters. However, this introduces an additional source of error in such studies, as the impact of water can enhance cluster formation by up to 2 orders of magnitude.[122]

Rasmussen et al.[101] demonstrated that for water-containing clusters, the approach of Kildgaard et al.[52,53] can outperform the JKCS method presented here. Their technique requires knowledge of low-energy structures for “dry” clusters and involves sequentially inserting water between existing bonds or around the cluster, exploring only a fraction of configurational space. While both methods are able to find the global minimum, JKCS demands significantly more computational resources due to its exploration of a larger configurational space. For sizable clusters with distinct predictable bonding patterns, like large hydrate clusters, alternative approaches such as Kildgaard’s method could offer faster CS. We encourage future studies to incorporate water.

3.3. ML Potential.

In our recent Clusterome paper,[123] we presented a large (∼250k), multiacid−multibase, atmospherically relevant, molecular cluster database (available in the ACDB 2.0 repository, see the Supporting Information). Here, we extract only ∼32k structures of the $(H_2SO_4{-}SA)_{0-2}(bases)_{0-2}$ clusters (termed Clusteromics I), where bases correspond to ammonia (AM), methylamine (MA), dimethylamine (DMA), trimethylamine (TMA), and ethylenediamine (EDA). We trained the KRR potential with the FCHL19[88] molecular representation (via JKML) to examine whether Δ-ML r^{2}SCAN-3c||GFN1-xTB can substitute the single-point r^{2}SCAN-3c in the JKCS workflow mentioned in the previous section. Hence, we selected 10 random equilibrium $(H_2SO_4)_3(base)_3$ clusters from Xie et al.[124] and predicted their binding electronic energies with the ML model. Figure 7 shows that the learning curve (orange line) rapidly converges to a mean absolute error (MAE) of 1.11 kcal/mol, corresponding to the model trained on the full Clusteromics I (black dotted line). Hence, already 1k structures randomly selected from the Clusteromics I database are enough to train an accurate ML potential and replace the r^{2}SCAN-3c step in the ML workflow. The training on 1000 structures with a subsequent test on 10 structures takes ∼5 CPU hours. However, performing r^{2}SCAN-3c itself takes only ∼1 CPU hour. Since this QC method is computationally quite fast, using ML does not speed up the process. More useful is, for instance, Δ-ML$^{\omega\mathrm{B97X\text{-}D}\|\mathrm{GFN1\text{-}xTB}}$ as used in our previous work on the CS of SA-multibase clusters[46] or even Δ-ML$^{\mathrm{DLPNO}\|\omega\mathrm{B97X\text{-}D}}$ as suggested in our recent perspective.[81] However, as a proof of concept, we will use r^{2}SCAN-3c in the

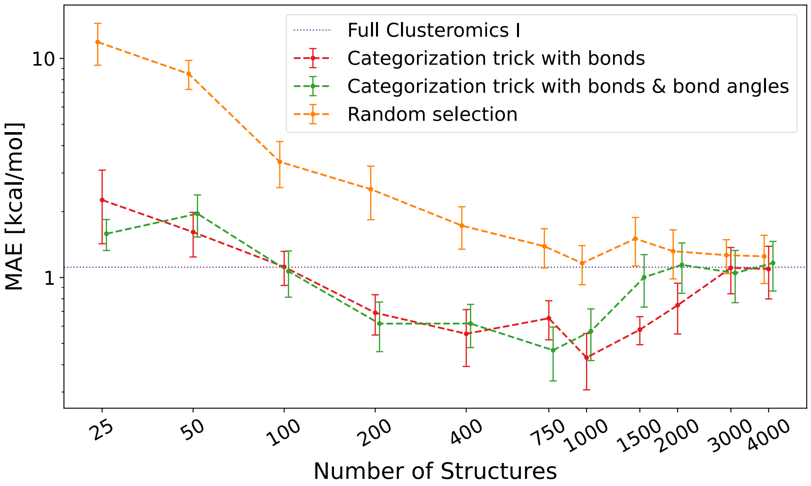


Figure 7. Learning curves, that is, mean absolute error (MAE) functions of the training set size, for the electronic binding energy predictions of 10 $(H_2SO_4)_3(\text{base})_3$ clusters from Xie et al.[124] The learning curves with different training subdatabase selection approaches (red, green, orange) converge to the MAE of training on the full (∼32k) database (black dotted line). Error bars correspond to the standard deviation of the sample mean.

next section to show the potential ML speedup by the categorization trick.

3.3.1. Fast ML. JKML offers two types of in-house algorithms to speed up KRR modeling. The first, kernel splitting (*-split* ⟨*int*⟩), employs multinode parallelization of kernel matrix constructions. Hence, the same results are obtained with the same computational resources but in a shorter wall-clock time. Second, the "categorization trick" (*-categorize* ⟨*int*⟩), compares the similarity between a test structure and all training and predicts using an ML model trained only on a subset of similar structures, where the similarity is performed via bond (and bond angle) comparison. This leads to faster predictions with an expected minor decrease in accuracy. With more test structures, a new ML model is trained for each tested structure. Another option would be to combine the training subsets into one reduced subset.

Figure 7 shows the learning curves for the selection via the categorization trick (green and red curves). This smart selection from the full database leads to approximately fivefold lower MAEs compared to a random selection. It can even reach lower values compared to the prediction on the full database, as structures that do not resemble the test structure(s) do not bias the model. When the training set size grows, all learning curves converge to the MAE of the full database. Further, we only use the categorization trick based on bonds, as including angles does not seem to improve the categorization trick.

Finally, to provide statistically accountable proof, we again used the ∼32k $(SA)_{0-2}(\text{bases})_{0-2}$ clusters (Clusteromics I) for training and tested it on 5k $(SA)_3(\text{bases})_3$ and 1k $(SA)_4(\text{bases})_4$ clusters from Kubečka et al.[46] Table 2 shows that the MAE of the predicted binding energies using the model trained on the full database is similar to the errors of the categorization trick with 100 and 200 trained structures for each target structure. Here, it is worth noting that the MAE of the ML predictions is less important for CS as we sort and filter configurations based on relative energies. Hence, the RMSD (i.e., span of errors) defines the quality of the ML model. As an example of computational times, the training on the full Clusteromics I took ∼152 CPU days and the prediction of the $(SA)_3(\text{base})_3$ set took an additional 103 CPU days (overall 255 CPU days). The equivalent process using the selection trick with 100 and 200 structures took overall only ∼13 and ∼43 CPU days, respectively. Clearly, selecting only 200 training structures in the $(SA)_3(\text{base})_3$ modeling is fast but we could reach even greater accuracy by using ∼800 structures (based on Figure 7). However, this would lead to ∼16 times slower modeling.

Table 2. Test of the Categorization Trick (*-categorize*) Method with 100 and 200 Most Similar Structures Selected from the Clusteromics I and Errors in the Prediction of Binding Energies of the Equilibrium $(SA)_3(\text{base})_3$ and $(SA)_4(\text{base})_4$ Clusters (Where SA Is Sulfuric Acid)

train	test	methods	MAE ± RMSD [kcal/mol]
Clusteromics I	$(SA)_3(\text{base})_3$	**full database**	**0.8 ± 1.0**
		-categorize 100	1.4 ± 1.7
		-categorize 200	1.2 ± 1.5
	$(SA)_4(\text{base})_4$	**full database**	**3.3 ± 5.2**
		-categorize 100	3.6 ± 4.7
		-categorize 200	2.6 ± 3.7

To summarize, Δ-ML is able to substitute the r^2SCAN-3c step in the CS workflow. This could be followed by filtering 10% of the lowest Δ-ML energies to the next step (e.g., DFT optimization). Figure 8 shows the correlation between Δ-ML predicted energies and r^2SCAN-3c energies for the $(SA)_3(\text{DMA})_3$ cluster, which supports the use of a 200-structure categorization trick. With this single-point energy approach, fewer structures need to be taken to computationally demanding DFT calculation as opposed to filtering straight on the semiempirical energy ordering. With very large diverse databases or training more parameters (e.g., forces), we also recommend using other training data set reduction methods[125] and/or using JKML coupled with SchNetPack[94,95] for training a NN potential.

4. CONCLUSIONS

In this work, we introduced the Jammy Key framework, a collection of scripts designed for systematic CS of molecular

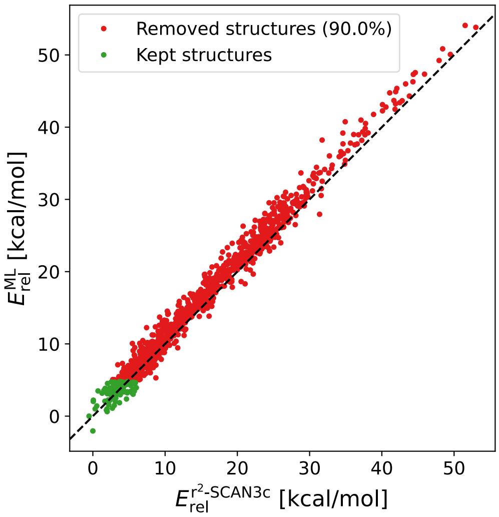


Figure 8. Correlation between Δ-ML and r^{2}SCAN-3c relative electronic energies (compared to the lowest conformer) for the equilibrium $(SA)_3(DMA)_3$ clusters (SA = sulfuric acid and DMA = dimethylamine). Here, the Clusteromics I data set was used for training the ML model with the 200-structure categorization trick. Filtering out 90% of redundant structures is highlighted in red color.

clusters as well as their handling, storing, and subsequent analysis. Notably, its core strengths are organized processing and automated file administration. The toolkit interfaces with commonly used third-party software such as ABCluster, CREST, XTB, Gaussian, and ORCA for executing quantum chemistry calculations while retaining adaptability for integration of alternative third-party programs. Its ultimate aim is to identify a representative set of structures corresponding to the lowest free energy minima. We demonstrated the application of the JKCS to various systems, primarily focusing on atmospheric molecular clusters, although the underlying principles are universally applicable.

JKQC is another powerful tool that allows the extraction of coordinates, forces, and other properties from QC programs into a compressed file. JKQC offers further manipulation of the stored data frame including sorting, filtering, specific data printing, applying QC corrections, calculation of binding (and atomization) properties, and producing input files for, for example, the ACDC or the Ion Mobility Software suite (IMoS). Consequently, the analysis of large data sets with JKQC becomes more automated and faster. We used these advantages to upgrade the architecture of the ACDB, conceived by J. Elm. Hence, the new version, ACDB 2.0, contains more cluster descriptions and takes less memory space, and the manipulation of the database is significantly improved. As a result, ACDB 2.0 presently encompasses over 1 million entries of molecular cluster configurations/properties that are suitable for ML applications.

Finally, we introduce the JKML that interfaces with the QML and SchNetPack packages. This interface facilitates the application of ML techniques from the KRR and NN families. We outline the potential integration of ML into the CS procedure and deliberate on their applicability with respect to the training data set size. We believe that incorporating ML techniques holds substantial promise in reducing the computational expenses associated with future investigations of molecular clusters.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c07412.

Package availability (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Jakub Kubečka − *Aarhus University, Department of Chemistry, Aarhus 8000, Denmark;* orcid.org/0000-0002-8002-0911; Phone: +420 724946622; Email: ja-kubecka@chem.au.dk

Authors

Vitus Besel − *University of Helsinki, Institute for Atmospheric and Earth System Research/Physics, Faculty of Science, Helsinki 00140, Finland;* orcid.org/0000-0003-4535-5422

Ivo Neefjes − *University of Helsinki, Institute for Atmospheric and Earth System Research/Physics, Faculty of Science, Helsinki 00140, Finland;* orcid.org/0000-0003-4549-0114

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

Theo Kurtén − *University of Helsinki, Institute for Atmospheric and Earth System Research/Chemistry, Faculty of Science, Helsinki 00140, Finland;* orcid.org/0000-0002-6416-4931

Hanna Vehkamäki − *University of Helsinki, Institute for Atmospheric and Earth System Research/Physics, Faculty of Science, Helsinki 00140, Finland;* orcid.org/0000-0002-5018-1255

Jonas Elm − *Aarhus University, Department of Chemistry, Aarhus 8000, Denmark;* orcid.org/0000-0003-3736-4329

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c07412

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research has been financially supported by the European Research Council project 692891-DAMOCLES, Academy of Finland Centers of Excellence Program (CoE VILMA, grant no. 346368), the Danish National Research Foundation (DNRF172) through the Center of Excellence for Chemistry of Clouds, the Independent Research Fund Denmark grant number 9064-00001B, and the European Union (MSCA, HYDROCLUSTER, project 101105506). J. Kubečka also gratefully acknowledges discussions with F. Keshavarz concerning QC calculations. Funded by the European Union. Views and opinions expressed are however those of the author(s) 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. We thank the CSC—Finnish IT Centre and the Centre for Scientific Computing, Aarhus (https://phys.au.dk/forskning/faciliteter/cscaa) for access to computer clusters and also computer capacity from the Finnish Grid and

Cloud Infrastructure (persistent identifier urn:nbn:fi:research-infras-2016072533).

REFERENCES

(1) Chen, J.; Jiang, S.; Liu, Y.-R.; Huang, T.; Wang, C.-Y.; Miao, S.-K.; Wang, Z.-Q.; Zhang, Y.; Huang, W. Interaction of Oxalic Acid with Dimethylamine and its Atmospheric Implications. *RSC Adv.* **2017**, *7*, 6374−6388.

(2) Chen, D.; Wang, W.; Li, D.; Wang, W. Atmospheric implication of synergy in methanesulfonic acid−base trimers: a theoretical investigation. *RSC Adv.* **2020**, *10*, 5173−5182.

(3) Ahonen, L.; Li, C.; Kubečka, J.; Iyer, S.; Vehkamäki, H.; Petäjä, T.; Kulmala, M.; Hogan, C. J., Jr Ion mobility-mass spectrometry of iodine pentoxide−iodic acid hybrid cluster anions in dry and humidified atmospheres. *J. Phys. Chem. Lett.* **2019**, *10*, 1935−1941.

(4) Kumar, M.; Li, H.; Zhang, X.; Zeng, X. C.; Francisco, J. S. Nitric acid−amine chemistry in the gas phase and at the air−water interface. *J. Am. Chem. Soc.* **2018**, *140*, 6456−6466.

(5) Odbadrakh, T. T.; Gale, A. G.; Ball, B. T.; Temelso, B.; Shields, G. C. Computation of atmospheric concentrations of molecular clusters from ab initio thermochemistry. *J. Vis. Exp.* **2020**, *158*, No. e60964.

(6) Elm, J.; Kubečka, J.; Besel, V.; Jääskeläinen, M. J.; Halonen, R.; Kurtén, T.; Vehkamäki, H. Modeling the formation and growth of atmospheric molecular clusters: A review. *J. Aerosol Sci.* **2020**, *149*, 105621.

(7) Jilkine, A.; Angenent, S. B.; Wu, L. F.; Altschuler, S. J. A Density-Dependent Switch Drives Stochastic Clustering and Polarization of Signaling Molecules. *PLoS Comput. Biol.* **2011**, *7*, No. e10022711.

(8) Mondello, P.; Paludo, J.; Novak, J.; Wenzl, K.; Yang, Z. Z.; Jalali, S.; Krull, J. E.; Braggio, E.; Dasari, S.; Manske, M. K.; et al. Molecular Clusters and Tumor-Immune Drivers of IgM Monoclonal Gammopathies. *Clin. Cancer Res.* **2023**, *29*, 957−970.

(9) Song, W.; Hu, Y.; Xie, M.; Li, Y.; Zhang, Z.; Jiang, N.; Liu, F.; Shan, X.; Sheng, L. Proton transfer processes in interstellar molecular clusters under synchrotron VUV radiation: Taking acrylonitrile and acetonitrile dimers as example. *J. Electron Spectrosc. Relat. Phenom.* **2020**, *242*, 146954.

(10) Zhang, Z.; Nie, W.; Sun, F.; Zhang, Y.; Xie, M.; Hu, Y. Conformational Landscapes and Infrared Spectra of Gas-phase Interstellar Molecular Clusters [$(C_3H_3N)(CH_3OH)_n$, n = 1−4]. *J. Phys. Chem. A* **2020**, *124*, 2398−2407.

(11) Evstrop'ev, S. K.; Nikonorov, N.; Saratovskii, A.; Danilovich, D. The effect of UV irradiation on the formation of silver molecular clusters and their stabilization in solutions and composite and oxide coatings. *Opt. Spectrosc.* **2020**, *128*, 707−712.

(12) Falaise, C.; Ivanov, A. A.; Molard, Y.; Amela Cortes, M.; Shestopalov, M. A.; Haouas, M.; Cadot, E.; Cordier, S. From supramolecular to solid state chemistry: crystal engineering of luminescent materials by trapping molecular clusters in an aluminium-based host matrix. *Mater. Horiz.* **2020**, *7*, 2399−2406.

(13) Brock, C. A.; Hamill, P.; Wilson, J. C.; Jonsson, H. H.; Chan, K. R. Particle formation in the upper tropical troposphere: A source of nuclei for the stratospheric aerosol. *Science* **1995**, *270*, 1650−1653.

(14) Rose, C.; Sellegri, K.; Asmi, E.; Hervo, M.; Freney, E.; Colomb, A.; Junninen, H.; Duplissy, J.; Sipilä, M.; Kontkanen, J.; et al. Major contribution of neutral clusters to new particle formation at the interface between the boundary layer and the free troposphere. *Atmos. Chem. Phys.* **2015**, *15*, 3413−3428.

(15) Mäkelä, J. M.; Aalto, P.; Jokinen, V.; Pohja, T.; Nissinen, A.; Palmroth, S.; Markkanen, T.; Seitsonen, K.; Lihavainen, H.; Kulmala, M. Observations of ultrafine aerosol particle formation and growth in boreal forest. *Geophys. Res. Lett.* **1997**, *24*, 1219−1222.

(16) Metzger, A.; Verheggen, B.; Dommen, J.; Duplissy, J.; Prevot, A. S. H.; Weingartner, E.; Riipinen, I.; Kulmala, M.; Spracklen, D. V.; Carslaw, K. S.; et al. Evidence for the role of organics in aerosol particle formation under atmospheric conditions. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 6646−6651.

(17) 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.

(18) Schobesberger, S.; Franchin, A.; Bianchi, F.; Rondo, L.; Duplissy, J.; Kürten, A.; Ortega, I. K.; Metzger, A.; Schnitzhofer, R.; Almeida, J.; et al. On the composition of ammonia−sulfuric-acid ion clusters during aerosol particle formation. *Atmos. Chem. Phys.* **2015**, *15*, 55−78.

(19) Liu, L.; Li, H.; Zhang, H.; Zhong, J.; Bai, Y.; Ge, M.; Li, Z.; Chen, Y.; Zhang, X. The role of nitric acid in atmospheric new particle formation. *Phys. Chem. Chem. Phys.* **2018**, *20*, 17406−17414.

(20) Temelso, B.; Morrison, E. F.; Speer, D. L.; Cao, B. C.; Appiah-Padi, N.; Kim, G.; Shields, G. C. Effect of mixing ammonia and alkylamines on sulfate aerosol formation. *J. Phys. Chem. A* **2018**, *122*, 1612−1622.

(21) Falcon-Rodriguez, C. I.; Osornio-Vargas, A.; Sada-Ovalle, I.; Segura-Medina, P. Aeroparticles, composition, and lung diseases. *Front. Immunol.* **2016**, *7*, 1−9.

(22) Gan, W. Q.; FitzGerald, J. M.; Carlsten, C.; Sadatsafavi, M.; Brauer, M. Associations of ambient air pollution with chronic obstructive pulmonary disease hospitalization and mortality. *Am. J. Respir. Crit. Care Med.* **2013**, *187*, 721−727.

(23) Mei, M.; Song, H.; Chen, L.; Hu, B.; Bai, R.; Xu, D.; Liu, Y.; Zhao, Y.; Chen, C. Early-life exposure to three size-fractionated ultrafine and fine atmospheric particulates in Beijing exacerbates asthma development in mature mice. *Part. Fibre Toxicol.* **2018**, *15*, 13.

(24) Eisele, F. L.; Hanson, D. R. First Measurement of Prenucleation Molecular Clusters. *J. Phys. Chem. A* **2000**, *104*, 830−836.

(25) Keutsch, F. N.; Saykally, R. J. Water clusters: Untangling the mysteries of the liquid, one molecule at a time. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 10533−10540.

(26) Sorokin, A.; Arnold, F.; Wiedner, D. Formation and growth of sulfuric acid−water cluster ions: Experiments, modelling, and implications for ion-induced aerosol formation. *Atmos. Environ.* **2006**, *40*, 2030−2045.

(27) 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.

(28) 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. Technol.* **2022**, *15*, 11−19.

(29) Zhang, J.; Dolg, M. ABCluster: the artificial bee colony algorithm for cluster global optimization. *Phys. Chem. Chem. Phys.* **2015**, *17*, 24173−24181.

(30) Zhang, J.; Dolg, M. Global optimization of clusters of rigid molecules using the artificial bee colony algorithm. *Phys. Chem. Chem. Phys.* **2016**, *18*, 3003−3010.

(31) Dieterich, J. M.; Hartke, B. OGOLEM: Global cluster structure optimisation for arbitrary mixtures of flexible molecules. A multiscaling, object-oriented approach. *Mol. Phys.* **2010**, *108*, 279−291.

(32) Pracht, P.; Bohle, F.; Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Phys. Chem. Chem. Phys.* **2020**, *22*, 7169−7192.

(33) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; et al. *Gaussian 16* Revision A.03; Gaussian Inc.: Wallingford CT, 2016.

(34) Neese, F. The ORCA program system. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2*, 73−78.

(35) Neese, F. Software update: The ORCA program system, version 4.0. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2018**, *8*, No. e1327.

(36) Bannwarth, C.; Caldeweyher, E.; Ehlert, S.; Hansen, A.; Pracht, P.; Seibert, J.; Spicher, S.; Grimme, S. Extended tight-binding

quantum chemistry methods. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2020**, *11*, 1−49.

(37) 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.

(38) 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.

(39) Pracht, P.; Caldeweyher, E.; Ehlert, S.; Grimme, S. A robust non-self-consistent tight-binding quantum chemistry method for large molecules. 2019, https://chemrxiv.org/engage/chemrxiv/article-details/60c742abbdbb890c7ba3851a, Preprint, 1−19.

(40) TURBOMOLE. *TURBOMOLE V7.2017, a development of University of Karlsruhe and Forschungszentrum Karlsruhe GmbH, 1989−2007*; TURBOMOLE GmbH, 2017;.

(41) Lumiaro, E.; Todorović, M.; Kurten, T.; Vehkamäki, H.; Rinke, P. Predicting gas−particle partitioning coefficients of atmospheric molecules with machine learning. *Atmos. Chem. Phys.* **2021**, *21*, 13227−13246.

(42) Besel, V.; Todorović, M.; Kurtén, T.; Rinke, P.; Vehkamäki, H. Atomic structures, conformers and thermodynamic properties of 32k atmospheric molecules. *Sci. Data* **2023**, *10*, 450.

(43) Hyttinen, N.; Pihlajamäki, A.; Häkkinen, H. Machine Learning for Predicting Chemical Potentials of Multifunctional Organic Compounds in Atmospherically Relevant Solutions. *J. Phys. Chem. Lett.* **2022**, *13*, 9928−9933.

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

(45) 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.

(46) Kubečka, J.; Neefjes, I.; Besel, V.; Qiao, F.; Xie, H. B.; Elm, J. Atmospheric Sulfuric Acid-Multi-Base New Particle Formation Revealed through Quantum Chemistry Enhanced by Machine Learning. *J. Phys. Chem. A* **2023**, *127*, 2091−2103.

(47) Jiang, S.; Liu, Y.-R.; Huang, T.; Feng, Y.-J.; Wang, C.-Y.; Wang, Z.-Q.; Ge, B.-J.; Liu, Q.-S.; Guang, W.-R.; Huang, W. Towards fully ab initio simulation of atmospheric aerosol nucleation. *Nat. Commun.* **2022**, *13*, 6067.

(48) Jääskeläinen, M. Comparing descriptors for molecular clusters in unsupervised learning. M.Sc. Thesis, University of Helsinki, 2020.

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

(50) Jensen, F. *Introduction to computational chemistry*; John Wiley & Sons, Inc.: USA, 2006.

(51) Kubečka, J. Developing efficient configurational sampling: structure, formation, and stability of atmospheric molecular clusters. Ph.D. Thesis, University of Helsinki, 2021.

(52) Kildgaard, J. V.; Mikkelsen, K. V.; Bilde, M.; Elm, J. Hydration of atmospheric molecular clusters: A new method for systematic configurational sampling. *J. Phys. Chem. A* **2018**, *122*, 5026−5036.

(53) Kildgaard, J. V.; Mikkelsen, K. V.; Bilde, M.; Elm, J. Hydration of Atmospheric Molecular Clusters II: Organic Acid−Water Clusters. *J. Phys. Chem. A* **2018**, *122*, 8549−8556.

(54) Kubečka, J.; Besel, V.; Kurtén, T.; Myllys, N.; Vehkamäki, H. Configurational sampling of noncovalent (atmospheric) molecular clusters: Sulfuric acid and guanidine. *J. Phys. Chem. A* **2019**, *123*, 6022−6033.

(55) Karaboga, D.; Basturk, B. On the performance of artificial bee colony (ABC) algorithm. *Appl. Soft Comput.* **2008**, *8*, 687−697.

(56) Stewart, J. J. P. Optimization of parameters for semiempirical methods V. Modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* **2007**, *13*, 1173−1213.

(57) Stewart, J. J. P. 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.

(58) 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.

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

(60) Engsvang, M.; Elm, J. Modeling the Binding Free Energy of Large Atmospheric Sulfuric Acid−Ammonia Clusters. *ACS Omega* **2022**, *7*, 8077−8083.

(61) Engsvang, M.; Kubečka, J.; Elm, J. Toward Modeling the Growth of Large Atmospheric Sulfuric Acid−Ammonia Clusters. *ACS Omega* **2023**, *8*, 34597.

(62) Wu, H.; Engsvang, M.; Knattrup, Y.; Kubečka, J.; Elm, J. Improved Configurational Sampling Protocol for Large Atmospheric Molecular Clusters. ChemRxiv. 2010, https://chemrxiv.org/engage/chemrxiv/article-details/64f365883fdae147fa55f05e (accessed Sep 04, 2023).

(63) Chai, J.-D.; Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom-atom dispersion corrections. *Phys. Chem. Chem. Phys.* **2008**, *10*, 6615−6620.

(64) Temelso, B.; Mabey, J. M.; Kubota, T.; Appiah-Padi, N.; Shields, G. C. ArbAlign: A tool for optimal alignment of arbitrarily ordered isomers using the Kuhn-Munkres algorithm. *J. Chem. Inf. Model.* **2017**, *57*, 1045−1054.

(65) Grimme, S. Supramolecular binding thermodynamics by dispersion-corrected density functional theory. *Chem.—Eur. J.* **2012**, *18*, 9955−9964.

(66) Myllys, N.; Elm, J.; Kurtén, T. Density functional theory basis set convergence of sulfuric acid-containing molecular clusters. *Comput. Theor. Chem.* **2016**, *1098*, 1−12.

(67) Pilati, T.; Forni, A. SYMMOL: A program to find the maximum symmetry group in an atom cluster, given a prefixed tolerance. *J. Appl. Crystallogr.* **1998**, *31*, 503−504.

(68) Riplinger, C.; Neese, F. An efficient and near linear scaling pair natural orbital based local coupled cluster method. *J. Chem. Phys.* **2013**, *138*, 034106.

(69) Riplinger, C.; Sandhoefer, B.; Hansen, A.; Neese, F. Natural triple excitations in local coupled cluster calculations with pair natural orbitals. *J. Chem. Phys.* **2013**, *139*, 134101.

(70) Riplinger, C.; Pinski, P.; Becker, U.; Valeev, E. F.; Neese, F. Sparse maps − A systematic infrastructure for reduced-scaling electronic structure methods. II. Linear scaling domain based pair natural orbital coupled cluster theory. *J. Chem. Phys.* **2016**, *144*, 024109.

(71) Partanen, L.; Vehkamäki, H.; Hansen, K.; Elm, J.; Henschel, H.; Kurtén, T.; Halonen, R.; Zapadinsky, E. Effect of conformers on free energies of atmospheric complexes. *J. Phys. Chem. A* **2016**, *120*, 8613−8624.

(72) McKinney, W. Data Structures for Statistical Computing in Python *Proceedings of the 9th Python in Science Conference*; SCIPY, 2010; pp 56−61.

(73) The pandas development team. *pandas-dev/pandas: Pandas (v2.1.0)*; Zenodo, 2023;.

(74) Hjorth Larsen, A.; Jørgen Mortensen, J.; Blomqvist, J.; Castelli, I. E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M. N.; Hammer, B.; Hargus, C.; et al. The atomic simulation environment—a Python library for working with atoms. *J. Condens. Matter Phys.* **2017**, *29*, 273002.

(75) Coots, J.; Gandhi, V.; Onakoya, T.; Chen, X.; Larriba-Andaluz, C. A parallelized tool to calculate the electrical mobility of charged aerosol nanoparticles and ions in the gas phase. *J. Aerosol Sci.* **2020**, *147*, 105570.

(76) Olenius, T.; Kupiainen-Määttä, O.; Ortega, I. K.; Kurtén, T.; Vehkamäki, H. Free energy barrier in the growth of sulfuric acid−ammonia and sulfuric acid−dimethylamine clusters. *J. Chem. Phys.* **2013**, *139*, 084312.

(77) Olenius, T. Atmospheric Cluster Dynamics Code. 2023, https://github.com/olenius/ACDC accessed (Oct 29, 2020).

(78) Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. U.S.A.* **1982**, *79*, 2554−2558.

(79) Chen, Z.; Wang, B.; Gorban, A. N. Multivariate Gaussian and Student-t process regression for multi-output prediction. *Neural. Comput. Appl.* **2020**, *32*, 3005−3028.

(80) Hofmann, T.; Schölkopf, B.; Smola, A. J. Kernel methods in machine learning. *Ann. Stat.* **2008**, *36*, 1171−1220.

(81) Kubečka, J.; Knattrup, Y.; Engsvang, M.; Jensen, A. B.; Ayoubi, D.; Wu, H.; Christiansen, O.; Elm, J. Current and future machine learning approaches for modeling atmospheric cluster formation. *Nat. Comput. Sci.* **2023**, *3*, 495−503.

(82) Rupp, M. Machine learning for quantum mechanics in a nutshell. *Int. J. Quantum Chem.* **2015**, *115*, 1058−1073.

(83) Rupp, M.; Tkatchenko, A.; Müller, K. R.; von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(84) Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K. R.; Tkatchenko, A. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett.* **2015**, *6*, 2326−2331.

(85) Huo, H.; Rupp, M. Unified representation of molecules and crystals for machine learning. *Mach. Learn.: Sci. Technol.* **2022**, *3*, 045017.

(86) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *87*, 184115.

(87) Faber, F. A.; Christensen, A. S.; Huang, B.; von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *J. Chem. Phys.* **2018**, *148*, 241717.

(88) Christensen, A. S.; Bratholm, L. A.; Faber, F. A.; Anatole von Lilienfeld, O. FCHL revisited: Faster and more accurate quantum machine learning. *J. Chem. Phys.* **2020**, *152*, 044107.

(89) Schütt, K. T.; Kindermans, P.; Sauceda, H. E.; Chmiela, S.; Tkatchenko, A.; Müller, K. *Advances in Neural Information Processing Systems*; NIPS, 2017.SchNet: A continuous-filter convolutional neural network for modeling quantum interactions

(90) Schütt, K. T.; Unke, O. T.; Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra *Proceedings of the 38th International Conference on Machine Learning*; PMLR, 2021;.

(91) Devereux, C.; Smith, J. S.; Huddleston, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E. Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *J. Chem. Theory Comput.* **2020**, *16*, 4192−4202.

(92) Jiang, S.; Liu, Y.-R.; Wang, C.-Y.; Huang, T. Benchmarking general neural network potential ANI-2x on aerosol nucleation molecular clusters. *Int. J. Quantum Chem.* **2023**, *123*, No. e27087.

(93) Christensen, A. S.; Faber, F. A.; Huang, B.; Bratholm, L. A.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. L. QML: A Python toolkit for quantum machine learning. 2017, https://github.com/qmlcode/qml (accessed Sep 10, 2018).

(94) Schütt, K. T.; Kessel, P.; Gastegger, M.; Nicoli, K. A.; Tkatchenko, A.; Müller, K. R. SchNetPack: A Deep Learning Toolbox For Atomistic Systems. *J. Chem. Theory Comput.* **2019**, *15*, 448−455.

(95) Schütt, K. T.; Hessmann, S. S. P.; Gebauer, N. W. A.; Lederer, J.; Gastegger, M. SchNetPack 2.0: A neural network toolbox for atomistic machine learning. *J. Chem. Phys.* **2023**, *158*, 144801.

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

(97) Browning, N. J.; Faber, F. A.; Anatole von Lilienfeld, O. GPU-Accelerated Approximate Kernel Method for Quantum Machine Learning. *J. Chem. Phys.* **2022**, *157*, 214801.

(98) Chmiela, S.; Sauceda, H. E.; Poltavsky, I.; Müller, K. R.; Tkatchenko, A. sGDML: Constructing accurate and data efficient molecular force fields using machine learning. *Comput. Phys. Commun.* **2019**, *240*, 38−45.

(99) Himanen, L.; Jäger, M. O.; Morooka, E. V.; Federici Canova, F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **2020**, *247*, 106949.

(100) He, X.; Iyer, S.; Sipilä, M.; Ylisirniö, A.; Peltola, M.; Kontkanen, J.; Baalbaki, R.; Simon, M.; Kürten, A.; Tham, Y. J.; et al. Determination of the collision rate coefficient between Charged iodic acid clusters and iodic acid using the appearance time method. *Aerosol Sci. Technol.* **2020**, *55*, 231−242.

(101) Rasmussen, F. R.; Kubečka, J.; Besel, V.; Vehkamäki, H.; Mikkelsen, K. V.; Bilde, M.; Elm, J. Hydration of atmospheric molecular clusters III: Procedure for efficient free energy surface exploration of large hydrated clusters. *J. Phys. Chem. A* **2020**, *124*, 5253−5261.

(102) Besel, V.; Kubečka, J.; Kurtén, T.; Vehkamäki, H. Impact of quantum chemistry parameter choices and cluster distribution model settings on modeled atmospheric particle formation rates. *J. Phys. Chem. A* **2020**, *124*, 5931−5943.

(103) Myllys, N.; Ponkkonen, T.; Chee, S.; Smith, J. Enhancing potential of trimethylamine oxide on atmospheric particle formation. *Atmosphere* **2019**, *11*, 35.

(104) Myllys, N.; Kubečka, J.; Besel, V.; Alfaouri, D.; Olenius, T.; Smith, J. N.; Passananti, M. Role of base strength, cluster structure and charge in sulfuric-acid-driven particle formation. *Atmos. Chem. Phys.* **2019**, *19*, 9753−9768.

(105) Knattrup, Y.; Kubečka, J.; Elm, J. Nitric Acid and Organic Acids Suppress the Role of Methanesulfonic Acid in Atmospheric New Particle Formation. *J. Phys. Chem. A* **2023**, *127*, 7568−7578.

(106) Dingilian, K. K.; Lippe, M.; Kubečka, J.; Krohn, J.; Li, C.; Halonen, R.; Keshavarz, F.; Reischl, B.; Kurtén, T.; Vehkamäki, H.; et al. New Particle Formation from the Vapor Phase: From Barrier-Controlled Nucleation to the Collisional Limit. *J. Phys. Chem. Lett.* **2021**, *12*, 4593−4599.

(107) Toropainen, A.; Kangasluoma, J.; Kurtén, T.; Vehkamäki, H.; Keshavarz, F.; Kubečka, J. Heterogeneous Nucleation of Butanol on NaCl: A Computational Study of Temperature, Humidity, Seed Charge, and Seed Size Effects. *J. Phys. Chem. A* **2021**, *125*, 3025−3036.

(108) Toropainen, A.; Kangasluoma, J.; Vehkamäki, H.; Kubečka, J. Heterogeneous Ion-Induced Nucleation of Water and Butanol Vapors Studied via Computational Quantum Chemistry beyond Prenucleation and Critical Cluster Sizes. *J. Phys. Chem. A* **2023**, *127* (18), 3976.

(109) Keshavarz, F.; Kubečka, J.; Attoui, M.; Vehkamäki, H.; Kurtén, T.; Kangasluoma, J. Molecular origin of the sign preference of ion-induced heterogeneous nucleation in a complex ionic liquid—diethylene glycol system. *J. Phys. Chem. C* **2020**, *124*, 26944−26952.

(110) 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. Technol.* **2020**, *13*, 3581−3593.

(111) Valiev, R. R.; Hasan, G.; Salo, V.-T.; Kubečka, J.; Kurten, T. Intersystem Crossings Drive Atmospheric Gas-Phase Dimer Formation. *J. Phys. Chem. A* **2019**, *123*, 6596−6604.

(112) Hasan, G.; Salo, V.-T.; Valiev, R.; Kubečka, J.; Kurtén, T. Comparing Reaction Routes for ^{3}(RO···OR′) Intermediates Formed in Peroxy Radical Self- and Cross-Reactions. *J. Phys. Chem. A* **2020**, *124*, 8305−8320.

(113) Keshavarz, F.; Shcherbacheva, A.; Kubečka, J.; Vehkamäki, H.; Kurtén, T. Computational study of the effect of mineral dust on secondary organic aerosol formation by accretion reactions of closed-shell organic compounds. *J. Phys. Chem. A* **2019**, *123*, 9008−9018.

(114) Olenius, T.; Bergström, R.; Kubečka, J.; Myllys, N.; Elm, J. Reducing chemical complexity in representation of new-particle formation: evaluation of simplification approaches. *Environ. Sci.: Atmos.* **2023**, *3*, 552−567.

(115) Zhang, J.; Glezakou, V.-A.; Rousseau, R.; Nguyen, M.-T. NWPEsSe: An adaptive-learning global optimization algorithm for nanosized cluster systems. *J. Chem. Theory Comput.* **2020**, *16*, 3947−3958.

(116) 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.

(117) Wavefunction, Inc. *Spartan'18*; Wavefunction, Inc., 2020;. https://store.wavefun.com/.

(118) Elm, J.; Bilde, M.; Mikkelsen, K. V. Assessment of density functional theory in predicting structures and free energies of reaction of atmospheric prenucleation clusters. *J. Chem. Theory Comput.* **2012**, *8*, 2071−2077.

(119) Elm, J.; Bilde, M.; Mikkelsen, K. V. Assessment of binding energies of atmospherically relevant clusters. *Phys. Chem. Chem. Phys.* **2013**, *15*, 16442−16445.

(120) 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.

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

(122) Henschel, H.; Navarro, J. C. A.; Yli-Juuti, T.; Kupiainen-Määttä, O.; Olenius, T.; Ortega, I. K.; Clegg, S. L.; Kurtén, T.; Riipinen, I.; Vehkamäki, H. Hydration of Atmospherically Relevant Molecular Clusters: Computational Chemistry and Classical Thermodynamics. *J. Phys. Chem. A* **2014**, *118*, 2599−2611.

(123) Knattrup, Y.; Kubečka, J.; Ayoubi, D.; Elm, J. Clusterome: A Comprehensive Data Set of Atmospheric Molecular Clusters for Machine Learning Applications. *ACS Omega* **2023**, *8*, 25155−25164.

(124) Xie, L.; Cheng, H.; Fang, D.; Chen, Z.-N.; Yang, M. Enhanced QM/MM sampling for free energy calculation of chemical reactions: A case study of double proton transfer. *J. Chem. Phys.* **2019**, *150*, 044111.

(125) Smith, J. S.; Nebgen, B.; Lubbers, N.; Isayev, O.; Roitberg, A. E. Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* **2018**, *148*, 241733.