nature communications

Article

https://doi.org/10.1038/s41467-022-33783-y

Towards fully ab initio simulation of atmospheric aerosol nucleation

Received: 2 January 2022

Accepted: 29 September 2022

Published online: 14 October 2022

Check for updates

Shuai Jiang[1] ✉, Yi-Rong Liu[1], Teng Huang[2], Ya-Juan Feng[1], Chun-Yu Wang[1], Zhong-Quan Wang[2], Bin-Jing Ge[1], Quan-Sheng Liu[1], Wei-Ran Guang[1] & Wei Huang[1,2,3]

Atmospheric aerosol nucleation contributes to approximately half of the worldwide cloud condensation nuclei. Despite the importance of climate, detailed nucleation mechanisms are still poorly understood. Understanding aerosol nucleation dynamics is hindered by the nonreactivity of force fields (FFs) and high computational costs due to the rare event nature of aerosol nucleation. Developing reactive FFs for nucleation systems is even more challenging than developing covalently bonded materials because of the wide size range and high dimensional characteristics of noncovalent hydrogen bonding bridging clusters. Here, we propose a general workflow that is also applicable to other systems to train an accurate reactive FF based on a deep neural network (DNN) and further bridge DNN-FF-based molecular dynamics (MD) with a cluster kinetics model based on Poisson distributions of reactive events to overcome the high computational costs of direct MD. We found that previously reported acid-base formation rates tend to be significantly underestimated, especially in polluted environments, emphasizing that acid-base nucleation observed in multiple environments should be revisited.

The theoretical understanding of the nucleation mechanism largely relies on classical nucleation theory (CNT)[1], originally proposed in 1935, which gives a general mind map for nucleation thermodynamics and kinetics[2] even though the capillary assumption has been extensively criticized[3]. The theoretical model Atmospheric Cluster Dynamics Code (ACDC), which emerged[4] in 2011 and was subsequently broadly employed[5–10], surmounts the drawbacks of CNT through coupled quantum chemical thermodynamics[11] with birth–death equations[2]. In the framework of ACDC, collision rate constants and evaporation rates are the two most critical parameters, determining the accuracy of the prediction of macroparameters such as cluster concentrations and formation rates that can be directly determined with experiments for comparison[5]. Evaporation rates, derived from detailed balance and ab initio thermodynamics, can be very accurately obtained with sophisticated quantum chemical calculations[12]. However, collision rate constants, derived from a simple hard-sphere collision model, are still very rough, and the accuracy is far from that of ab initio-based evaporation rates. Moreover, determining accurate collision rate constants is extremely important, especially for collision-controlled systems such as sulfuric acid–dimethylamine systems, as evaporation rates are close to zero[13]. Pioneering work[14] investigated the collisions between sulfuric acid monomers; however, the force field (FF) utilized lacks reactivity, and the computational costs of extending the method to more collisions among molecules and/or clusters are enormous. Therefore, a highly accurate and inexpensive reactive FF for flexible nucleation clusters is urgently needed to simulate nucleation processes with full ab initio accuracy.

Here, we propose a general workflow to drive the aerosol nucleation simulation toward becoming fully ab initio. In the workflow, comprehensive data sets are first prepared through metadynamics coupled with active learning techniques. Then, a deep neural network-based force field (DNN-FF) is trained so that robust nucleation

[1]School of Information Science and Technology, University of Science and Technology of China, Hefei, Anhui 230026, China. [2]Laboratory of Atmospheric Physico-Chemistry, Anhui Institute of Optics & Fine Mechanics, Chinese Academy of Sciences, Hefei, Anhui 230031, China. [3]Center for Excellent in Urban Atmospheric Environment, Institute of Urban Environment, Chinese Academy of Sciences, Xiamen, Fujian 361021, China. ✉e-mail: shuaijiang@ustc.edu.cn

molecular dynamics (MD) simulations can be performed to derive the collision rate constants based on the Poisson distribution. Then, static quantum chemical thermodynamics-based evaporation rates are coupled with DNN-FF-based MD-derived collision rate constants into a cluster dynamics model to provide ab initio kinetics for simulating atmospheric aerosol nucleation.

Results

A general workflow for fully ab initio simulation of aerosol nucleation

The key modules in the workflow are shown in Fig. 1. The details in each module can be found in the Methods section, so here, the major points regarding the significance and correlation for each module are given. The initial data set is first prepared by metadynamics sampling in addition to subsequent screening and labeling. The screening is made by farthest point sampling (FPS), while the force and energy labeling is done by density function theory (DFT). Then, an active learning strategy with two force thresholds is utilized to supplement the initial data set to form the final data set to obtain the final force field. In each active learning iteration, DNN-FF-based MD based on the previously active learning iterations selected data set and metadynamics prepared data set is conducted. Then, inaccurate structures satisfying the threshold range are selected for labeling and added to the data set for the next iteration. Therefore, after finalizing the data set, multiple DNN-FF-based one nanosecond MD simulations can be performed. Finally, based on the Poisson distribution, collision rate constants are derived and combined with static quantum chemistry-based evaporation rates to obtain macroparameters such as the formation rate by a cluster kinetics model. The cluster size sampled by metadynamics is based on the cluster stability characteristic of acid-base clusters being mostly stable when the difference between an acid number and the base number is less than or equal to one[5]. Active learning not only supplements the structures for metadynamics sample size but also points to the cluster compositions with high evaporation rates, e.g., $(DMA)_4$, the cluster being composed of four dimethylamine molecules, as we can see from the active learning data set in Fig. 1c, which can normally be ignored through sampling under predefined cluster compositions. Notably, we will use $(SA)_m(DMA)_n$ to represent the cluster composed of m sulfuric acid molecules and n dimethylamine molecules. Collision rate constants, derived from MD simulations based on Poisson distribution reaction events (Fig. 1d), are essentially independent of cluster concentrations, making high-concentration MD simulations valuable for further cluster kinetics.

Due to the interpolative nature of DNNs, high accuracy could be maintained for clusters up to $(SA)_{10}(DMA)_{10}$. The accuracy for clusters beyond $(SA)_{10}(DMA)_{10}$ is unknown, but we expect a further decrease in

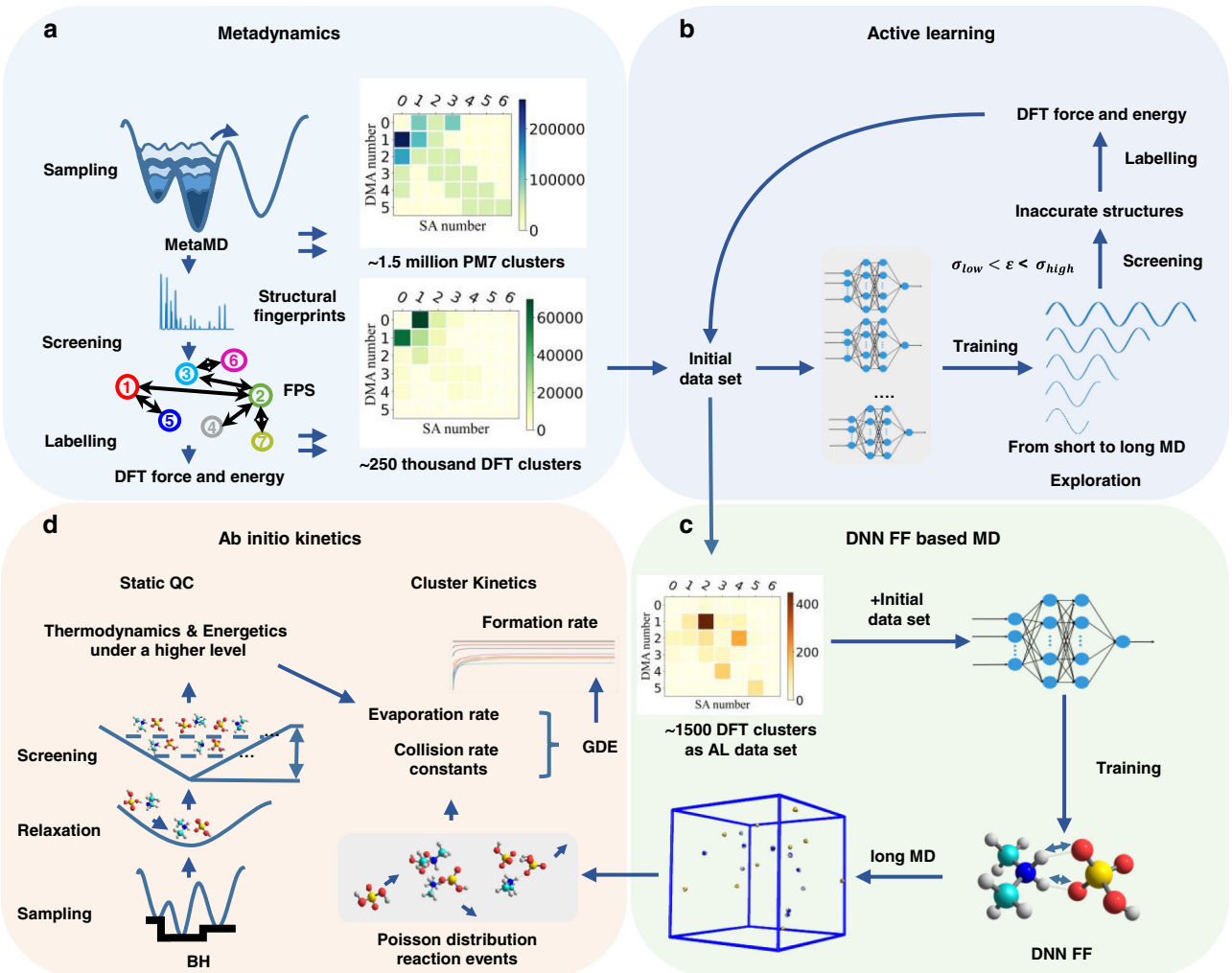

Fig. 1 | A general workflow towards fully ab initio simulation of atmospheric aerosol nucleation. It includes the steps to prepare the data set for training a deep neural network-based force field (DNN-FF), to apply DNN-FF by molecular dynamics (MD), to derive collision rate constants from MD, and to couple collision rate constants with cluster kinetics model for studying atmospheric aerosol nucleation. **a**, **b** show the metadynamics and active learning techniques used to prepare a data set for the deep neural network, respectively. **c** DNN-FF-driven MD. **d** Cluster kinetics simulation based on MD-derived collision rate constants and static quantum chemistry (QC) calculation-derived evaporation rates.

https://doi.org/10.1038/s41467-022-33783-y

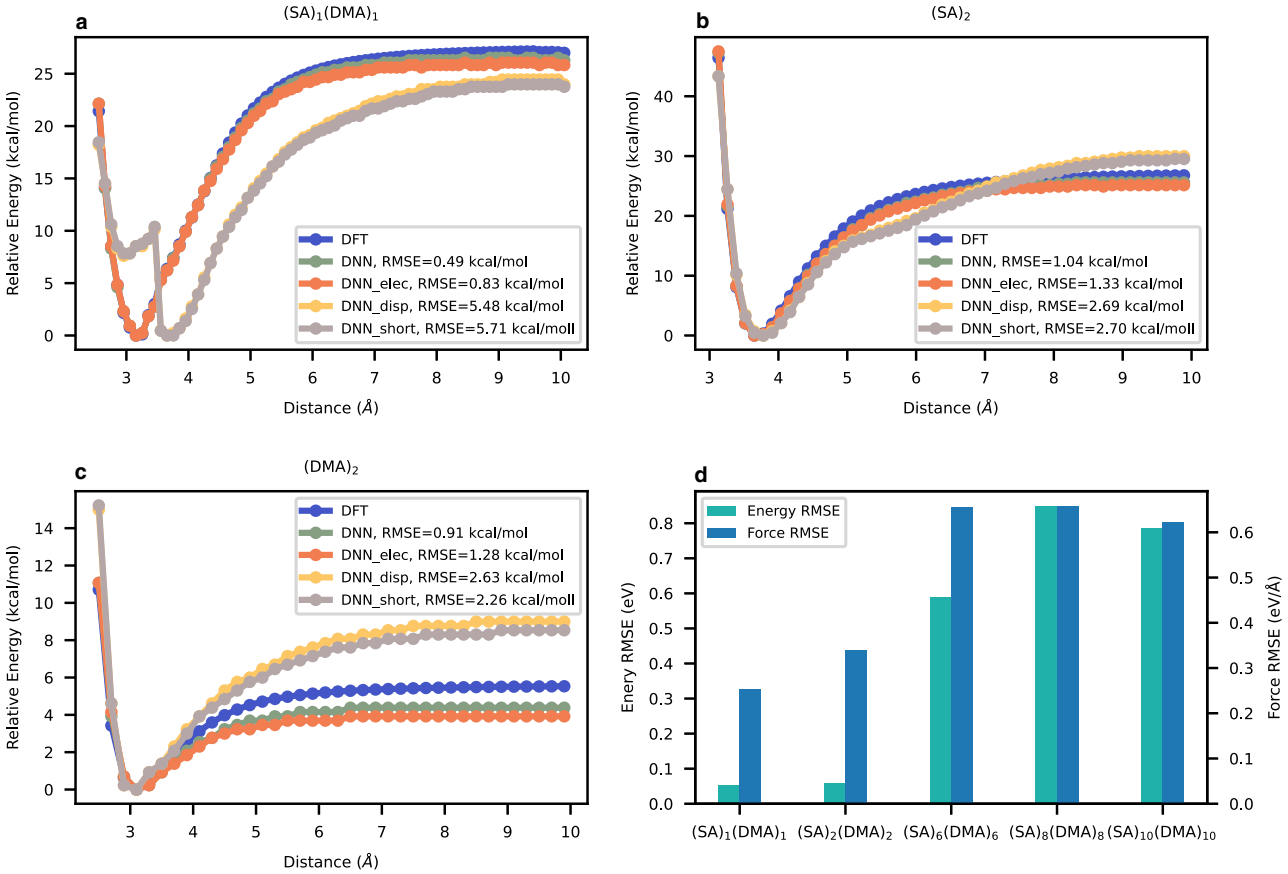

Fig. 2 | Deep neural network-based force field (DNN-FF) benchmark. a–c show the dimer detachment curves for $(SA)_1(DMA)_1$, $(SA)_2$, and $(DMA)_2$, respectively, where SA and DMA represent sulfuric acid and dimethylamine molecules, respectively. The relative energy is the isomer energy minus the energy of the most stable isomer. DNN, DNN_elec, DNN_disp, and DNN_short represent the model with electrostatics and dispersions, the model with electrostatics and without dispersions, the model without electrostatics and with dispersions, and the model without electrostatics and dispersions, respectively. **d** Energy and force root mean squared error (RMSE) values in the interpolation and extrapolation regimes of the test set. Source data are provided as a Source Data file.

accuracy. The ultimate goal of simulating atmospheric aerosol nucleation is to conduct MD under ambient or laboratory conditions, where the cluster size can easily go beyond the interpolation regime, so how to apply the DNN model with size-limited accuracy to atmospheric nucleation becomes a problem that needs to be addressed. We bridge the gap between microparameters and macroparameters by embedding MD-derived rate constants based on the Poisson distribution into a cluster kinetics model. These DNN-FF-based MD-derived constants coupled with static quantum chemistry (QC)-derived evaporation rates effectively drive the aerosol simulation towards full ab initio calculations.

The benchmark of the DNN-FF

The dimer detachment curves in Fig. 2a–c provide a basic picture of the performance of DNN-FFs significantly affected by long-range interactions. Generally, DNN-FF with long-range interactions performs very well on the investigated dimer systems, with the root mean squared error (RMSE) of the relative energy being close to 1 kcal/mol, the so-called chemical accuracy. Adding electrostatics and dispersion corrections into the short-range DNN model not only decreases the RMSE but also improves the curve smoothness. From the RMSE, electrostatic interactions are clearly more important than dispersion. The maximum cluster size within the training set is $(SA)_5(DMA)_6$, so the energy and force RMSE values in the extrapolation regime are larger than those in the interpolation regime, as expected. Even in the extrapolation regime, the DNN model still yields an encouraging accuracy close to that of a recently reported combustion reaction DNN[15]. In summary, the DNN model's superior performance in energy and force descriptions, in addition to its distinguished size scalability, lays a solid foundation for robust nucleation MD simulations.

Structural and energetic characteristics from DNN-FF-based MD

With the robust size scalability of the DNN model, nanosecond-scale MD simulations for cluster collision and evaporation can be performed, and a representative snapshot is shown in Fig. 3a. Furthermore, isolated clusters can be singled out to gain insights into their structural evolution. Here, $(SA)_6(DMA)_6$ is chosen since it is the largest cluster with the same number of acids and bases observed in the DNN-FF-based MD and is very close to the lowest experimentally detectable cluster size (~1.7 nm)[5]. For the most stable isomer (Fig. 3b), the sulfuric acid molecules are hydrogen bonded with each other, forming a shell with the cluster center of mass (COM) inside, while all dimethylamine molecules are protonated. The structural similarity can be seen through the closely connected points in the energy basin (Fig. 3d). In addition, here, the $(SA)_6(DMA)_6$ cluster emerges from the collision of $(SA)_4(DMA)_5$ and $(SA)_2(DMA)_1$; the high-energy isomers during a collision and subsequent rearrangement can also be seen in Fig. 3d (semitransparent red points in the upper left corner). Interestingly, none of the proton-transferred nitrogen-oxygen bonds break during the simulation (Supplementary Fig. 5), indicating quite strong bonding from proton transfer. Comparatively, we see that one of the two protons initially covalently bonded with the oxygen atom in the sulfuric acid molecule transfers to one dimethylamine molecule, while the other proton moves from one sulfuric acid molecule to another (Supplementary Figs. 4, 6). After the collision, the molecules in the cluster rearrange, finally making the proton number within sulfuric

https://doi.org/10.1038/s41467-022-33783-y

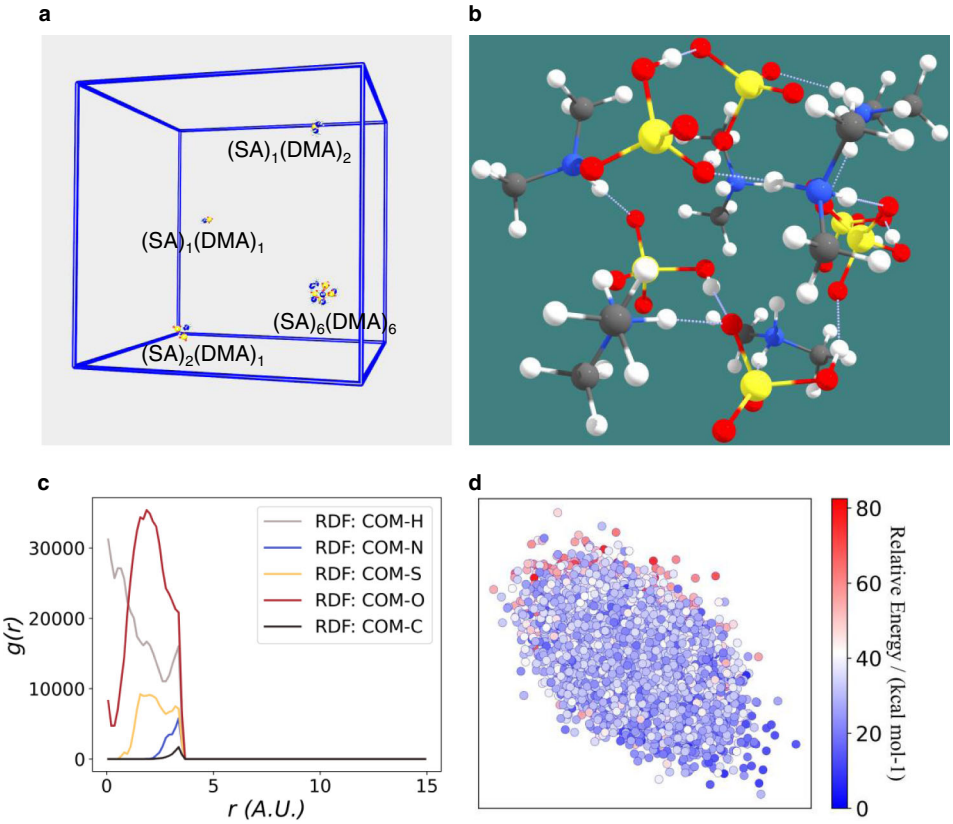

Fig. 3 | Structural distribution for $(SA)_6(DMA)_6$ isomers derived from deep neural network-based force field (DNN-FF)-based molecular dynamics (MD). **a** shows a snapshot of MD at 1 ns. The cyan, white, red, blue, and yellow circles represent C, H, O, N, and S atoms, respectively. The N and S atom radii are increased threefold for clarity. **b** The most stable isomer in the trajectory. **c** Radial distribution function (RDF) between the cluster center of mass (COM) and the five elements. **d** Kernel principal component analysis (KPCA)[58] maps of isomers using a global Smooth Overlap of Atomic Positions (SOAP)[59] kernel. Source data are provided as a Source Data file.

acid molecule one (Supplementary Fig. 6). From the above analysis, essential structural insights can be obtained by collecting clusters with the same composition from the MD trajectory.

Aerosol nucleation kinetics

DNN-FF-driven MD simulation of molecular cluster collisions and evaporations follow a Poisson distribution[16], so the so-called collision enhancement factor (CEF) can be derived. This term is the quotient of MD-derived collision rate constants divided by hard-sphere collision model-derived collision rate constants. The CEF is typically more than one due to long-range intermolecular forces[14]. However, for the collision between dimethylamine monomers, the CEF is below one at 300 K, mainly because of the intermolecular repulsion of dimethylamine molecules (Fig. 4a). From Supplementary Table 1, for the same reaction, the CEF at 278 K is slightly larger than the CEF at 300 K. Replacing hard-sphere collision rate constants with MD-derived constants in a cluster kinetics model paves the way for fully ab initio simulation of aerosol nucleation. The derived formation rates divided by those based on hard-sphere collision rate constants give the formation rate enhancement factor (FREF) (Fig. 4c, d). In representative clean (Fig. 4c) and polluted (Fig. 4d) environments, FREF has a strong negative correlation with the DMA concentration and a weak negative correlation with the SA concentration. Under the typical SA and DMA concentrations with different CS values, FREF in the polluted environment is much larger than FREF in the clean environment (yellow points in Fig. 4c, d). This indicates that aerosol particle formation in a polluted environment is much more underestimated than that in a clean environment. Generally, here, FREF ranges from one to several hundreds; however, this is the lower bound value, as there are still many collision rate constants that need to be replaced. Therefore, obtaining larger formation rates than expected before challenges the idea that SA-DMA nucleation is collision-limited with zero evaporation rates[17], providing the alternative scenario that collision rate constants are underestimated when evaporation rates are low but not zero. Further studies regarding this topic are definitely needed when fully ab initio kinetics are available in the future.

Discussion

Currently, the coupling between machine learning (ML) and chemistry is on the rise, so training a DNN model with good performance on training and test data sets is becoming increasingly routine[18–26]. However, training a DNN model with good size scalability and applicability to reactive MD simulations, especially for flexible molecules, as in this case, is still very challenging[27]. Here, robust MD simulations prove the high quality of data sets given by metadynamics and active learning, as well as the excellent performance of descriptors combined with the neural network framework and parameters. The workflow we propose here works for strong acid and strong base nucleation systems, which are the most significant systems in the aerosol nucleation field due to their strong nucleation ability. However, for systems with apparent barriers, such as the sulfuric acid-ammonia system, it is probably vital to utilize the strategy[28] to introduce transition state configurations to obtain a uniformly accurate model.

Future work can be first conducted on how to produce a compact data set, probably heavily relying on active learning. Another aspect that should be investigated is improving the DNN model accuracy, possibly through transfer learning[29] or Δ-learning[30]. For nucleation applications, more diverse box size and initial monomer spatial

https://doi.org/10.1038/s41467-022-33783-y

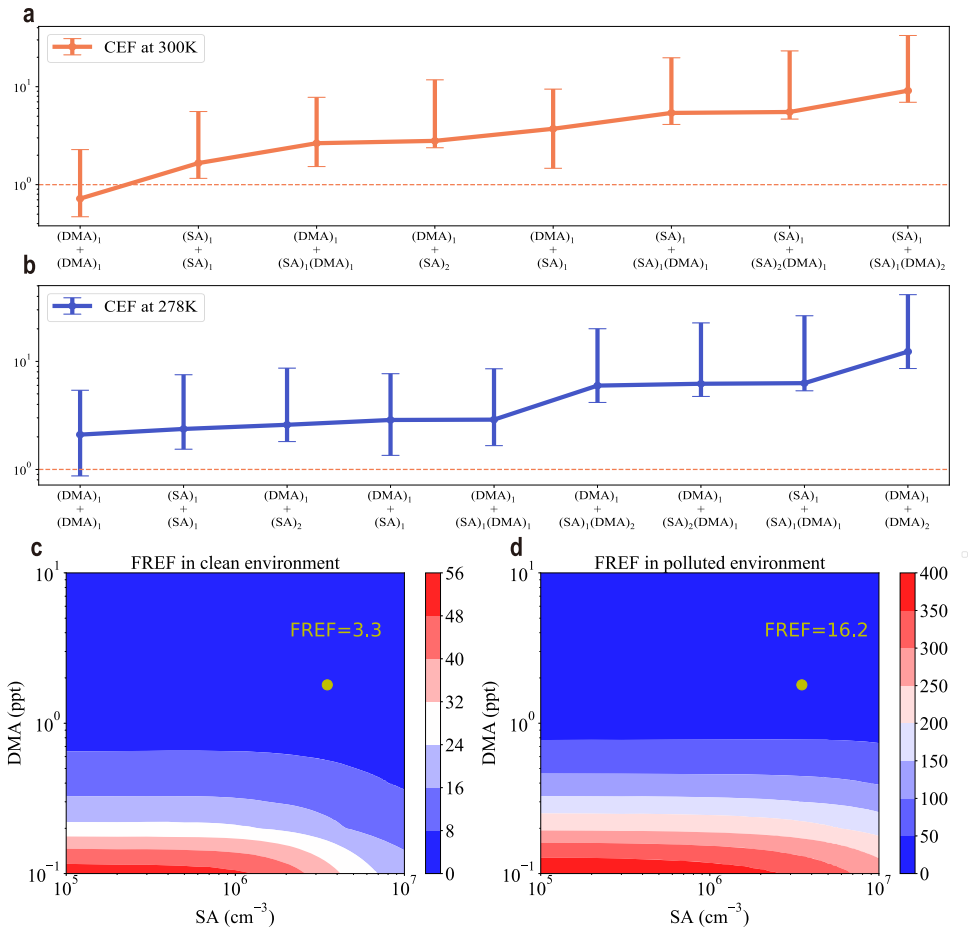

Fig. 4 | Collision and formation rate enhancement factor (CEF and FREF). **a**, **b** show the collision dependence of the CEF at 300 and 278 K, respectively. The reaction list can be found in Supplementary Table 1. The error bars shown in **a**, **b**, which represent the upper and lower bound of CEF, result from a 95% confidence level of Poisson-distributed reactions. Events **c**, **d** give the FREFs for representative clean ($T = 278$ K, $CS = 2.6 \times 10^{-3}$ s^{-1}) and polluted environments ($T = 278$ K, $CS = 2.7 \times 10^{-2}$ s^{-1}), respectively, where T and CS are the temperature and the condensation sink coefficient, respectively. A temperature of 300 K is chosen to resemble the conditions of standard temperature and pressure (STP), while 278 K represents typical spring-time conditions at the boreal forest site in Hyytiälä[60], the flagship observatory in the new particle formation field and winter-time conditions for observation in Beijing[61]. Yellow points in **c**, **d** show FREFs under typical clean and polluted environments, respectively, where $[SA] = 3.5 \times 10^6$ cm^{-3} and $[DMA] = 1.8$ ppt. Source data are provided as a Source Data file.

distribution MD simulations need to be conducted to lower the uncertainties[16] to estimate collision rate constants. In addition, a larger box-size simulation with more molecules inside is needed to include more types of collisions so that the simulation can be fully ab initio. Afterward, many ab initio-derived collision rates could tentatively be predicted purely by molecular physical chemistry properties to reduce simulation time costs.

The complexity and variety of nucleation precursors in the ambient environment, especially in polluted environments, necessitate new theoretical methods in addition to static QC calculations to unravel the complicated associated mechanisms. We believe the workflow proposed here, with the introduction of the DNN-FF and the bridging between MD-derived rate constants with cluster kinetics, paves the way toward the full ab initio simulation of atmospheric aerosol nucleation. The highly accurate formation rate derived here can be further parametrized into a climate model to improve climate prediction on global and local scales.

Methods

Metadynamics

Instead of being sampled by basin-hopping[31], which has been widely applied in atmospheric noncovalent interaction clusters[32], the potential energy surfaces of nucleation clusters are sampled by metadynamics (MetaMD)[33,34] due to its remarkable ability to sample high-energy isomers to prepare the initial data set for further active learning iterations. The bump perturbation can be calculated as[33]

$$V_{\text{bump}}\left(\vec{\mathbf{R}}\right) = \sum_{\alpha} \lambda e^{-\sum_{ij}\left(D_{ij}\left(\vec{\mathbf{R}}\right)-D_{ij}^{\alpha}\right)^2/(2\sigma^2)} \quad (1)$$

Here, $\vec{\mathbf{R}}$ represents the atomic coordinates; α sums over snapshots of geometries where the matrix of atomic distances at a given point during the trajectory: $D_{ij} = 1/|\vec{r}_{ij}|$ is the collective variable, where $\vec{r}_{ij}$ are the atomic distances; i and j loop over all the atoms in the frame; D_{ij}^{α} are the previous distance matrices, which we accumulate every τ femtoseconds (fs); the bumps with bump width σ and bump height λ are applied to all elements of the contact matrix. The bump width σ, bump height λ, and bump time τ are set to 2.0, 1.0, and 10, respectively, while the MD temperature, time step, and thermostat for the NVT ensemble are set to 600 K, 0.5 fs, and Anderson, respectively. The cluster whose size is within the range of $(SA)_m(DMA)_n$ ($m = 0$–4, $n = 0$–4) is sampled with MetaMD in the Tensormol[35] package interfaced with the PM7 semiempirical method in Gaussian16[36]. To save computational costs, not all cluster sizes within the range are sampled. According to the experimental and theoretical predictions, the sulfuric acid–dimethylamine system tends to grow with a similar number of molecules within the cluster[5]; therefore, for large clusters, those with a difference between the number of acid and base molecules less than or

equal to one are included. For each cluster size, ~50,000 structures are sampled and subsequently selected by the farthest point sampling (FPS) method based on the many-body tensor representation (MBTR) descriptor[37] for further DFT (ωB97XD/6−31++G(d,p)) energy and force labeling. ωB97XD/6−31++G(d,p) is chosen because the systematic benchmark[38] for aerosol nucleation clusters proves its good balance between accuracy and cost. The detailed MetaMD sampling and subsequent DFT calculation procedures are listed in Supplementary Table 2.

Active learning

Based on the initial data set prepared by MetaMD, an active learning or an on-the-fly strategy[39] is utilized. The MetaMD sampling subset after screening is the initial data set to kick off the active learning iterations. In each iteration, first, 400,000 steps of training with different seeds are conducted to generate four DNN models. Then, the constant-temperature, constant-volume ensemble (NVT) MD simulations are performed in LAMMPS[40] based on trained DNN models. During the MD simulations, four DNN models are utilized to pinpoint the candidate clusters whose error indicators satisfy the threshold range. The error indicator is the maximal standard deviation of the atomic force predicted by the model ensemble. The upper and lower threshold values are 0.50 and 0.35 eV/Å, respectively, indicating that those whose error indicator is below 0.35 eV/Å are regarded as accurate and those whose error indicator is above 0.50 eV/Å are regarded as physically unreasonable. Finally, the energies and forces of the candidate clusters are obtained by ωB97XD/6−31++G(d,p) in the Gaussian16[36] package for training in the next iteration. Notably, the candidate clusters are carved out from the MD trajectory according to the interatomic distance cut-off of 3.5 Å. Here, the clusters are obtained when the shortest interatomic distance between molecules is shorter than the interatomic distance cut-off value to maintain the integrity of the molecular cluster, which is different from the strategy used in a similar work conducted for combustion reactions[15]. The detailed iteration processes are listed in Supplementary Table 3.

DNN model

The smooth version of the deep potential[41,42] model is conducted in active learning. In deep potential, the potential energy of a molecular cluster is a sum of "atomic energies" $E = \sum_i E_i$, where E_i is determined by the local environment of atom i within a cut-off radius. The model includes two networks: the embedding network and the fitting network. The embedding network is of size (25, 50, 100) and the fitting network is of size (240, 240, 240). The fitting network uses ResNet architecture[43]. The cut-off radius is set to 6.0 Å and the descriptors decay smoothly from 5.8 to 6.0 Å. The learning rate starts at 1.0×10^{-3} and exponentially decays every 2000 steps in 400,000 training steps in each active learning iteration. The loss function is defined as a sum of different mean square errors of the DNN predictions for energy and force. The long-range DNN model, Physnet[44], is utilized to train on the final data set for 10,000,000 steps. DeePMD with strictly local descriptors is integrated with LAMMPS, which guarantees high efficiency of MD exploration, so DeePMD is utilized in the active learning iterations. Despite the recent appearance of the long-range version DeePMD (DPLR)[45], we switched to Physnet for production, as preparing maximally localized Wannier centers for DPLR requires a large cell where the molecular electron density decreases to zero on the faces of the cell, which is computationally expensive. Because D3BJ instead of D3 is fitted in Physnet, the final data set is further calculated at the level of ωB97X-D3BJ/6−31++G(d,p) through ORCA 5.0[46] to label structures with the energies and forces as well as the dipole moments. The width of the neural network is controlled by setting the feature space dimensionality and radial basis function number to 128 and 64, respectively, while the neural network depth is controlled by setting the stacked modular building blocks number, residual block number for atom-wise refinements, residual block number for refinements of proto-message and residual block number in output blocks to 5, 2, 3, and 1, respectively. The cut-off radius for interactions in the neural network is set to 10 Å and long-range interactions are explicitly included by electrostatics and dispersion corrections.

Molecular dynamics

The collision and evaporation simulations of molecular clusters are conducted under the NVT ensemble at 278 and 300 K through the Atomic Simulation Environment (ASE) with ten SA molecules and ten DMA molecules initially randomly placed in the cubic box with a length of 85 Å. The random positions are given by the packmol[47] package with the stable SA and DMA monomers being the input structures. In each run, MD has performed 100 ps with the COM for each molecule being fixed for equilibration and subsequently 1 ns for production. The cluster positions in the production stage are recorded every 10 fs. The snapshot in Fig. 3a is plotted by VMD[48], while the structure in Fig. 3b is plotted by Chemcraft[49]. The RDF and structural clustering analysis is conducted by freud[50] and ASAP[51], respectively. The proton transfer distance threshold between O and H is set to 1.23 Å. Collison rate constants are derived according to the Poisson distribution feature of the reactive (collision and evaporation) events[16] using the ChemTraYzer software package[52]. The Poisson-based collision rate constant k can be calculated according to

$$k = \frac{\sum_j N_j}{V \sum_j \left(\sum_i^M C_i \Delta t_i \right)_j} \tag{2}$$

Here, N_j is the collision event number in MD run j, V is the MD box volume, i is the subsimulation number in MD run j separated by reactive events, C is the product of reactant concentrations, and Δt is the interval between reactive events. The detailed derivation for rate constants and confidence interval of Poisson-based reaction events can be found in the literature[16].

Cluster kinetics

The molecular cluster kinetics simulations are performed by the home-built Python version[53] of the Atmospheric Cluster Dynamics Code (ACDC)[4] to solve the ordinary differential equations. The collision rate constants are partially replaced by the MD-observed collision event-derived constants, and the remaining collision rate constants are calculated by a hard-sphere collision model. The evaporation rates are calculated assuming a detailed balance based on quantum chemical thermodynamics[11] from the literature[54]. The condensation sink (CS) is set to be $2.6 \times 10^{-3}\ s^{-1}$ and $2.7 \times 10^{-2}\ s^{-1}$ to mimic condensation under clean[55] and polluted[56] environments, respectively. In ACDC, the formation rate can be calculated by

$$J = \sum_{i=0}^{4} \sum_{j=0}^{4} \sum_{k=0}^{4} \sum_{l=0}^{4} \beta_{ik,jl} c_{ik} c_{jl} (i + j \geq 4, k + l > 4) \tag{3}$$

Here, i and j refer to the number of SA molecules in each binary collision molecular cluster, and k and l refer to the number of DMA molecules in each binary collision molecular cluster. The time evolution of the cluster concentration c_i can be obtained by solving the birth and death equations given by

$$\frac{dC_i}{dt} = \frac{1}{2} \sum_{j<i} \beta_{j,(i-j)} c_j c_{i-j} + \sum_j \gamma_{(i+j)\to i} c_{(i+j)} - \sum_j \beta_{i,j} c_i c_j - \frac{1}{2} \sum_{j<i} \gamma_{i\to j} c_i + CS \tag{4}$$

Here, CS represents the condensation sink. $\beta_{i,j}$ represents the collision rate constants obtained from the hard-sphere collision model

and is calculated by

$$\beta_{i,j} = \left(\frac{3}{4\pi}\right)^{1/6} \left(\frac{6k_bT}{m_i} + \frac{6k_bT}{m_j}\right)^{1/2} \left(V_i^{1/3} + V_j^{1/3}\right)^2 \quad (5)$$

Here, T represents the temperature, k_b represents the Boltzmann constant, and m_i and V_i represent the mass and volume of cluster i, respectively. The evaporation coefficient $\gamma_{(i+j)\to i}$ is calculated by

$$\gamma_{(i+j)\to i,j} \frac{c_i^e c_j^e}{c_{i+j}^e} = \beta_{i,j} c_{ref} \exp\left(\frac{\triangle G_{i+j} - \triangle G_i - \triangle G_j}{k_b T}\right) \quad (6)$$

Here, i and j are the daughter clusters, $\beta_{i,j}$ is the collision rate constant between i and j, c_i^e is the equilibrium concentration of cluster i, ΔG_i is the free energy of formation of cluster i from the constituent monomers, and c_{ref} is the monomer concentration of the reference vapor for which the free energies were calculated.

Data availability

The training and test data set for DNN, the DNN-FF model, and molecular dynamics trajectories based on DNN-FF are available on figshare (https://doi.org/10.6084/m9.figshare.20968156.v1)[57]. Source data are provided with this paper.

Code availability

The codes, including metadynamics sampling, active learning, DNN training, molecular dynamics, and cluster kinetics, in addition to the data and scripts to reproduce all the figures in the manuscript and supplementary materials, are available on figshare (https://doi.org/10.6084/m9.figshare.20968156.v1)[57].

References

1. Frenkel, J. Statistical theory of condensation phenomena. *J. Chem. Phys.* **7**, 200–201 (1939).
2. Vehkamäki, H. & Riipinen, I. Thermodynamics and kinetics of atmospheric aerosol particle formation and growth. *Chem. Soc. Rev.* **41**, 5160–5173 (2012).
3. Merikanto, J., Zapadinsky, E., Lauri, A. & Vehkamäki, H. Origin of the failure of classical nucleation theory: incorrect description of the smallest clusters. *Phys. Rev. Lett.* **98**, 145702 (2007).
4. McGrath, M. J. et al. Atmospheric cluster dynamics code: a flexible method for solution of the birth-death equations. *Atmos. Chem. Phys.* **12**, 2345–2355 (2011).
5. Almeida, J. et al. Molecular understanding of sulphuric acid-amine particle nucleation in the atmosphere. *Nature* **502**, 359–363 (2013).
6. Yao, L. et al. Atmospheric new particle formation from sulfuric acid and amines in a Chinese megacity. *Science* **361**, 278–281 (2018).
7. Lehtipalo, K. et al. The effect of acid–base clustering and ions on the growth of atmospheric nano-particles. *Nat. Commun.* **7**, 11594 (2016).
8. Liu, L. et al. Unexpected quenching effect on new particle formation from the atmospheric reaction of methanol with SO<sub>3</sub&gt. *Proc. Natl Acad. Sci. USA* **116**, 24966–24971 (2019).
9. 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.* **140**, 6456–6466 (2018).
10. Li, H. et al. Self-catalytic reaction of SO3 and NH3 to produce sulfamic acid and its implication to atmospheric particle formation. *J. Am. Chem. Soc.* **140**, 11020–11028 (2018).
11. Ortega, I. K. et al. From quantum chemical formation free energies to evaporation rates. *Atmos. Chem. Phys.* **12**, 225–235 (2011).
12. Elm, J. et al. Modeling the formation and growth of atmospheric molecular clusters: a review. *J. Aerosol Sci.* **149**, 105621 (2020).
13. Kuerten, A. et al. Neutral molecular cluster formation of sulfuric acid-dimethylamine observed in real time under atmospheric conditions. *Proc. Natl Acad. Sci. USA* **111**, 15019–15024 (2014).
14. Halonen, R., Zapadinsky, E., Kurtén, T., Vehkamäki, H. & Reischl, B. Rate enhancement in collisions of sulfuric acid molecules due to long-range intermolecular forces. *Atmos. Chem. Phys.* **19**, 13355–13366 (2019).
15. Zeng, J., Cao, L., Xu, M., Zhu, T. & Zhang, J. Z. H. Complex reaction processes in combustion unraveled by neural network-based molecular dynamics simulation. *Nat. Commun.* **11**, 5713 (2020).
16. Kröger, L. C., Kopp, W. A., Döntgen, M. & Leonhard, K. Assessing statistical uncertainties of rare events in reactive molecular dynamics simulations. *J. Chem. Theory Comput.* **13**, 3955–3960 (2017).
17. Kürten, A. et al. New particle formation in the sulfuric acid–dimethylamine–water system: reevaluation of CLOUD chamber measurements and comparison to an aerosol nucleation and growth model. *Atmos. Chem. Phys.* **18**, 845–863 (2017).
18. Dral, P. O. Quantum chemistry in the age of machine learning. *J. Phys. Chem. Lett.* **11**, 2336–2347 (2020).
19. Noé, F., Tkatchenko, A., Müller, K. R. & Clementi, C. Machine learning for molecular simulation. *Annu. Rev. Phys. Chem.* **71**, 361–390 (2020).
20. Mueller, T., Hernandez, A. & Wang, C. Machine learning for interatomic potential models. *J. Chem. Phys.* **152**, 050902 (2020).
21. Unke, O. T. et al. Machine learning force fields. *Chem. Rev.* **121**, 10142–10186 (2021).
22. Behler, J. Four generations of high-dimensional neural network potentials. *Chem. Rev.* **121**, 10037–10072 (2021).
23. Meuwly, M. Machine learning for chemical reactions. *Chem. Rev.* **121**, 10218–10239 (2021).
24. Keith, J. A. et al. Combining machine learning and computational chemistry for predictive insights into chemical systems. *Chem. Rev.* **121**, 9816–9872 (2021).
25. Kulichenko, M. et al. The rise of neural networks for materials and chemical dynamics. *J. Phys. Chem. Lett.* **12**, 6227–6243 (2021).
26. Westermayr, J., Gastegger, M., Schütt, K. T. & Maurer, R. J. Perspective on integrating machine learning into computational chemistry and materials science. *J. Chem. Phys.* **154**, 230903 (2021).
27. Schran, C., Brieuc, F. & Marx, D. Transferability of machine learning potentials: protonated water neural network potential applied to the protonated water hexamer. *J. Chem. Phys.* **154**, 051101 (2021).
28. Yang, M., Bonati, L., Polino, D. & Parrinello, M. Using metadynamics to build neural network potentials for reactive events: the case of urea decomposition in water. *Catal. Today* **387**, 143–149 (2022).
29. Smith, J. S. et al. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* **10**, 2903 (2019).
30. 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.* **11**, 2087–2096 (2015).
31. Jiang, S. et al. Study of Cl–(H2O)n (n = 1–4) using basin-hopping method coupled with density functional theory. *J. Comput. Chem.* **35**, 159–165 (2014).
32. Zhang, J. & Glezakou, V.-A. Global optimization of chemical cluster structures: methods, applications, and challenges. *Int. J. Quantum Chem.* **121**, e26553 (2021).
33. Herr, J. E., Yao, K., McIntyre, R., Toth, D. W. & Parkhill, J. Metadynamics for training neural network model chemistries: a competitive assessment. *J. Chem. Phys.* **148**, 241710 (2018).
34. Laio, A. & Parrinello, M. Escaping free-energy minima. *Proc. Natl Acad. Sci. USA* **99**, 12562–12566 (2002).
35. Yao, K., Herr, J. E., Toth, D. W., McKintyre, R. & Parkhill, J. The TensorMol-0.1 model chemistry: a neural network augmented with long-range physics. *Chem. Sci.* **9**, 2261–2269 (2018).

36. Frisch, M. J. et al. Gaussian 16 revision a. 03. (Gaussian inc., 2016).
37. Huo, H. & Rupp, M. J. a. p. a. Unified representation of molecules and crystals for machine learning. Preprint at https://arxiv.org/abs/1704.06439 (2017).
38. Elm, J. & Mikkelsen, K. V. Computational approaches for efficiently modelling of small atmospheric clusters. *Chem. Phys. Lett.* **615**, 26–29 (2014).
39. Zhang, Y. et al. DP-GEN: a concurrent learning platform for the generation of reliable deep learning based potential energy models. *Comput. Phys. Commun.* **253**, 107206 (2020).
40. Aktulga, H. M., Fogarty, J. C., Pandit, S. A. & Grama, A. Y. Parallel reactive molecular dynamics: numerical methods and algorithmic techniques. *Parallel Comput.* **38**, 245–259 (2012).
41. Wang, H., Zhang, L., Han, J. & E, W. DeePMD-kit: a deep learning package for many-body potential energy representation and molecular dynamics. *Comput. Phys. Commun.* **228**, 178–184 (2018).
42. Zhang, L., Han, J., Wang, H., Car, R. & E, W. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Phys. Rev. Lett.* **120**, 143001 (2018).
43. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* (eds Tuytelaars, T. et al.) (IEEE, 2016).
44. Unke, O. T. & Meuwly, M. PhysNet: a neural network for predicting energies, forces, dipole moments, and partial charges. *J. Chem. Theory Comput.* **15**, 3678–3693 (2019).
45. Zhang, L. et al. A deep potential model with long-range electrostatic interactions. *J. Chem. Phys.* **156**, 124107 (2022).
46. Neese, F., Wennmohs, F., Becker, U. & Riplinger, C. The ORCA quantum chemistry program package. *J. Chem. Phys.* **152**, 224108 (2020).
47. Martínez, L., Andrade, R., Birgin, E. G. & Martínez, J. M. PACKMOL: a package for building initial configurations for molecular dynamics simulations. *J. Comput. Chem.* **30**, 2157–2164 (2009).
48. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph.* **14**, 33–38 (1996).
49. Andrienko, G. J. U. Chemcraft - graphical software for visualization of quantum chemistry computations. https://www.chemcraftprog.com (2005).
50. Ramasubramani, V. et al. freud: a software suite for high throughput analysis of particle simulation data. *Comput. Phys. Commun.* **254**, 107275 (2020).
51. Cheng, B. et al. Mapping materials and molecules. *Acc. Chem. Res.* **53**, 1981–1991 (2020).
52. Döntgen, M. et al. Automated discovery of reaction pathways, rate constants, and transition states using reactive molecular dynamics simulations. *J. Chem. Theory Comput.* **11**, 2517–2524 (2015).
53. Xu, C.-X. et al. Formation of atmospheric molecular clusters of methanesulfonic acid–diethylamine complex and its atmospheric significance. *Atmos. Environ.* **226**, 117404 (2020).
54. 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.* **139**, 084312 (2013).
55. Maso, M. D. et al. Annual and interannual variation in boreal forest aerosol particle number and volume concentration and their connection to particle formation. *Tellus, Ser. B* **60**, 495–508 (2008).
56. Wu, Z. et al. New particle formation in Beijing, China: Statistical analysis of a 1-year data set. *J. Geophys. Res.* **112**, D09209 (2007).
57. Jiang, S. et al. Towards fully ab initio simulation of atmospheric aerosol nucleation. figshare https://doi.org/10.6084/m9.figshare.20968156.v1 (2022).
58. Schölkopf, B., Smola, A. & Müller, K.-R. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.* **10**, 1299–1319 (1998).
59. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
60. Kulmala, M. et al. Direct observations of atmospheric aerosol nucleation. *Science* **339**, 943–946 (2013).
61. Guo, S. et al. Remarkable nucleation and growth of ultrafine particles from vehicular exhaust. *Proc. Natl Acad. Sci. USA* **117**, 3427–3432 (2020).

Acknowledgements

We thank Linfeng Zhang, Jinzhe Zeng, and other deep potential community contributors for the diligent support of DeePMD and DP-GEN. We thank Linfeng Zhang for the very helpful comments on the manuscript. We thank John E. Herr for the help with the metadynamics test and analysis. We thank Roope Halonen and Bernhard Reischl for sharing with us the input files to reproduce sulfuric acid monomer collision simulations. We acknowledge the support of the GPU cluster built by the MCC Lab of Information Science and Technology Institution, USTC. This work was supported by the National Natural Science Foundation of China (Grant No. 41877305).

Author contributions

S.J. proposed the workflow and conducted metadynamics, active learning, DNN training, as well as molecular dynamics and analysis. Y.-R.L., T.H., Y.-J.F., C.-Y.W., and Z.-Q.W. helped write and edit the paper. B.-J.G., Q.-S.L., and W.-R.G. improved the terminology expressions of the manuscript. W.H. helped with the design of the workflow. All authors commented on the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-022-33783-y.

Correspondence and requests for materials should be addressed to Shuai Jiang.

Peer review information *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.