AARHUS UNIVERSITY

Dear Editor and Professor Patrick Rinke,

We sincerely thank the referee for their comprehensive feedback. Their comments have helped us improve the manuscript, particularly by enhancing the robustness of our machine learning methodology description and making the text more accessible to a broader audience. We have carefully considered all of their constructive suggestions and adjusted the manuscript accordingly.

Below, we provide a point-by-point response to each comment. The referee's comments are given in ***bold italics***, while our responses are provided in standard roman text. Excerpts from the revised manuscript are indicated with a yellow highlight to support our responses. The line and page number to which a response refers is denoted by (L### P#).

We believe these incorporated revisions and our detailed responses address the referee's feedback and strengthen the manuscript. We respectfully submit this revised version for your consideration for publication in *Atmospheric Chemistry and Physics*.

We look forward to hearing from you at your earliest convenience and thank you for considering our manuscript for publication.

Best regards,

Ivo Neefjes

AARHUS UNIVERSITY

1 Point-by-point responses

The manuscript evaluates machine-learning interatomic potentials for simulating gas-phase collisions and early clustering among key atmospheric precursors (sulfuric acid with itself, dimethylamine, and bisulfate) linking short-range reactive accuracy to long-range interaction fidelity that controls capture and collision rates. Using umbrella sampling and umbrella integration to reconstruct potentials of mean force (PMF) with explicit subtraction of the radial entropic term, the authors derive collision probabilities and rate coefficients from molecular dynamics and analyze how model architecture affects both kinetics and thermodynamics. The simulations capture reactive events such as proton transfer during acid-base encounters, while the conclusions stress that gains in short-range accuracy must not compromise the long-range forces that govern initial approach and rate enhancement; accordingly, the authors urge validation beyond scalar error metrics (MAE/RMSE) and propose deploying models with explicit long-range interactions where collision kinetics are targeted.

For Atmospheric Chemistry and Physics (ACP) readers, the study offers a practical framework to obtain physically faithful collision-sticking rates and PMFs for nucleation-relevant systems and provides actionable guidance on model selection: local, short-range-accurate networks for thermodynamic sampling and global or long-range–aware models for encounter kinetics, consistent with recent ACP advances on collision-sticking analysis. The data, trained models, and scripts are openly available via the Atmospheric Cluster Database, facilitating reproducibility and uptake in nucleation and early growth modelling. I therefore recommend publication once the more detailed comments below have been addressed.

We thank the referee for their positive assessment and for taking the time to provide such a detailed and accurate summary of our study.

"IPCC assessment report" - It would be good to spell out IPCC

We agree with the referee that while the abbreviation IPCC is well known in the atmospheric sciences community, we hope this paper will also be read by the machine learning and general quantum chemistry communities. As such, we have clarified this abbreviation:

According to the latest Intergovernmental Panel on Climate Change (IPCC) assessment report,...

L20 P1

"These calculations explicitly account for long-range interactions and provide a fully atomistic description. Furthermore, the resulting trajectories offer insight into the molecular-level dynamics governing collisions and the formation of stable clusters." There is a logical disconnect in the introduction. After this sentence all

seems fine, but I guess that is, because the paragraph leading up to this sentence did not specify how the MD was calculated. Maybe one can allude the reader already to the forthcoming problems otherwise the introduction meanders along for quite a while until the readers reaches the actual problem definition.

We thank the referee for their careful review and for pointing out this logical disconnect. We agree that our initial effort to cover the relevant MD literature inadvertently obscured the overarching logical thread of the introduction.

To improve the narrative flow, we have restructured this section to introduce molecular dynamics, and the specific challenges of using it to study particle formation, much earlier, specifically in the third paragraph. While the core content remains largely unchanged, we believe this revised structure establishes the study's primary focus much sooner and directly addresses the referee's concern:

To capture the dynamic nature of these initial steps, researchers increasingly rely on atomistic molecular dynamics (MD) simulations. These simulations provide a fully atomistic description and offer insight into the molecular-level dynamics governing collisions and the formation of stable clusters. However, accurately capturing the necessary physics at a reasonable computational cost remains challenging. The quality of MD simulations is in large part determined by the level of theory at which the interaction potential between the nuclei in the system is obtained. An ideal interatomic potential for particle formation MD must satisfy three competing requirements: it must accurately capture long-range attractive forces to model how molecules initially approach; it must describe short-range quantum effects, such as chemical reactions, to model cluster stabilization; and it must be computationally efficient enough to sample a statistically significant number of events.

These effects include chemical reactions like proton transfers, which play a critical role in stabilizing atmospheric clusters. While semi-empirical quantum chemistry methods such as GFN1-xTB (Grimme et al., 2017) can model bond-breaking, they can exhibit significant errors for complex hydrogen-bonded systems, including quantitative inaccuracies in binding energies (Neefjes et al., 2026) and qualitative misidentifications of lowest-energy cluster configurations (Kubečka et al., 2019). Higher levels of theory, such as the DFT composite method ωB97X-3c (Müller et al., 2023), offer the necessary short- and long-range accuracy, but their computational cost makes even short MD simulations of small systems prohibitively expensive. Thus, neither classical nor conventional ab initio methods fully satisfy the requirements for large-scale cluster formation MD.

Recently, several machine learning (ML) architectures have been developed to construct accurate interatomic potentials for molecular systems. These machine learning interatomic potentials (MLIPs) offer a potential solution to this tradeoff, promising to reproduce the accuracy of high-level quantum theory at a reasonable computational cost. For instance, the polarizable atom interaction neural network (PaiNN) is an equivariant message-passing neural network capable of accelerating MD simulations while maintaining accuracy comparable to its reference training data (Schütt et al., 2021; Kubečka et al., 2024). Similarly, the second-generation atoms-in-molecules neural network (AIMNet2) has demonstrated high predictive accuracy across a wide range of molecular systems with remarkable efficiency, enabling

simulations of systems containing up to 10^5 atoms (Anstine et al., 2025).

However, MLIPs often rely on a local atomic environment approximation, in which the model encodes the environment around each atom up to a user-defined cutoff radius. This approximation improves transferability and computational efficiency but inherently limits the model to short-range interactions. The PaiNN model addresses this through a message-passing framework, where atoms exchange information with their neighbors via message and update blocks. Through multiple iterations, the effective interaction range grows, allowing atoms to indirectly access information from beyond the immediate cutoff. However, if all atoms in one subsystem (e.g., a molecule) lie beyond the cutoff radius of another, the interaction graph becomes disconnected. Consequently, no messages are exchanged, and the model treats the subsystems as non-interacting. AIMNet2 mitigates this by supplementing message passing with explicit long-range contributions. It predicts partial charges to model analytical Coulomb interactions and adds dispersion effects via the D3(BJ) correction scheme (Grimme et al., 2010, 2011).

The potential of MLIPs for atmospheric modeling has already been demonstrated by several recent studies that simulated the evolution of systems containing tens of particle-forming molecules to observe cluster formation dynamics (Jiang et al., 2022, 2023; Liu and Jiang, 2025). As the field increasingly adopts these methods for large-scale simulations, it is important to evaluate how well different model architectures capture long-range interactions alongside the necessary short-range accuracy and computational efficiency.

A rigorous metric for evaluating this long-range capability is the canonical collision rate coefficient. In cluster distribution dynamics models, such as the Atmospheric Cluster Dynamics Code (ACDC) (McGrath et al., 2012), cluster-forming collisions and cluster-removing evaporations are treated as independent processes, assuming that dissociation prior to thermalization from collisional excitation is negligible (Elm et al., 2020). This yields a pressure-independent collision rate coefficient, which represents the frequency of collisions per unit concentration. Traditionally, this coefficient is calculated using kinetic gas theory. In this framework, colliding partners are approximated as hard spheres, and intermolecular interactions are neglected entirely. While analytical approaches like the central field model can account for long-range forces, they require interaction parameters that are significantly more difficult to determine than standard hard-sphere radii (Neefjes et al., 2025).

Atomistic MD collision trajectory simulations in the free molecular regime provide a powerful alternative to calculate these coefficients directly from the underlying physical interactions (Halonen et al., 2019; Neefjes et al., 2022; Yang et al., 2023; Knattrup et al., 2025; Tikkanen et al., 2025). As demonstrated by Halonen et al. (2019), explicitly capturing long-range interactions via MD using the classical OPLS-AA force field resulted in an enhancement factor of 2.7 relative to kinetic gas theory for sulfuric acid dimerization. Accurately reproducing these enhanced collision rates serves as a robust metric for evaluating the long-range behavior of MLIPs in atmospheric applications.

In this methodological study, we assess the ability of the PaiNN and AIMNet2 architectures to describe collisions governed by long-range interactions. We sampled training configurations using GFN1-xTB dynamics, subsequently computing energies and forces at both the GFN1-xTB and ωB97X-3c levels. Additionally, we employed delta-learning to upscale GFN1-xTB

simulations with PaiNN corrections to the ωB97X-3c level of theory. Since sulfuric acid is a key contributor to particle formation (Sipilä et al., 2010), we studied the sulfuric acid dimer, the sulfuric acid–dimethylamine system (to investigate stabilizing proton transfers), and the sulfuric acid–bisulfate system (to examine strong ionic long-range contributions). Following hyperparameter tuning, we evaluated model performance by comparing electronic energy and force predictions against independent test sets. Furthermore, we calculated the potential of mean force (PMF) through umbrella sampling to compare against GFN1-xTB reference data. Finally, we derived collision rate coefficients from MD collision trajectory simulations to evaluate the long-range dynamics of the models and examine how they vary across different levels of theory. By validating these ML models in the context of atmospheric particle formation, this study establishes the necessary groundwork for large-scale MD simulations in this domain.

L29 P2

PaiNN section: Is PaiNN actually charge aware? How does PaiNN handle the anionic system in the study?

We thank the referee for this insightful question, which directly relates to the limitations observed in our results. The standard PaiNN architecture lacks explicit, physics-based charge awareness. While the total molecular charge can be provided to PaiNN as an additional descriptor during training, the model treats it merely as a learned feature. Unlike AIMNet2, which uses charge to compute explicit long-range Coulombic interactions, PaiNN does not contain the built-in physics to propagate electrostatic effects beyond their defined cutoff radius.

PaiNN handles the charged system by implicitly learning the local effects of the charge directly from the training data. Because the reference semi-empirical and DFT calculations inherently capture the stronger interactions, polarization, and altered local geometries of the charged state, PaiNN successfully reproduces the complex short-range potential energy surface of the charged system. However, because it lacks explicit charge awareness and long-range Coulombic terms, it is blind to electrostatic interactions that extend beyond its 10 Å cutoff.

We have updated the methodology section to discuss this charge-unawareness in more detail:

The standard PaiNN architecture is not explicitly charge-aware and lacks long-range electrostatic or dispersion corrections. Instead, the model handles charged systems by implicitly learning the local, short-range effects of the charge directly from the energies and forces provided in the training data. However, a limitation of this purely local approach in the context of gas-phase collisions is that interactions cannot be transmitted between atoms separated by more than the cutoff distance without intermediate atoms to mediate the message passing. Increasing the cutoff can capture these long-range effects, but at the expense of higher computational cost and potentially reduced accuracy, as the model must learn to generalize over a significantly larger spatial domain.

L107 P4

AIMNet2 section: "The model combines local atomic environments with learned "atom-in-molecule" (AIM) embeddings" — I don't understand why these two aspects are singled out for AIMNet2. All message passing graph neural networks do this. Also PaiNN uses a sophisticated input representation of the local atomic environments and then builds up atomic embeddings during the training.

AIMNet2 section: "These embeddings, available for 14 elements..." — Are the authors speaking of an AIMNet2 foundation or a pre-trained model? Otherwise it shouldn't matter what the elements are, because a pristine AIMNet2 architecture could be trained on the data at hand and then contain the elements present in that dataset.

The referee is completely correct in questioning these statements. Coming from a computational chemistry background, our original text did not properly describe the machine learning architecture. We inadvertently conflated the fundamental mechanics of standard message-passing networks, as well as the specific parameterization of the pre-trained models, with the architectural specifics of AIMNet2. As the referee rightly points out, iteratively building environment-aware embeddings is a feature of all message-passing neural network potentials (including PaiNN), and the 14-element parameterization refers to the pre-trained weights rather than a limitation of the architecture. To correct this, we have thoroughly revised the AIMNet2 methodology section. We removed the confusing statements and instead focused on the true architectural distinctions of AIMNet2 that are relevant to our study: its generalized embedding strategy, native charge awareness, and reliance on radial symmetry functions:

AIMNet2 is the second generation of the Atoms-In-Molecules Neural Network developed by Anstine et al. (2025). Using a message-passing architecture, the model iteratively refines invariant representations of local atomic environments, defined by radial symmetry functions, to build complex "atom-in-molecule" (AIM) embeddings. A key feature of AIMNet2 is its generalized embedding strategy, which avoids element-specific subnetworks and allows the model to flexibly represent highly diverse chemical compositions.

Beyond its local representations, AIMNet2 explicitly incorporates electronic and long-range physical effects. The architecture is charge-aware, using the total molecular charge as an input parameter to dynamically infer atom-centered partial charges during the message-passing phase. These partial charges are iteratively updated through a neural charge equilibration (NQE) scheme. The total potential energy is then calculated as the sum of the local configurational energy, explicit Coulombic electrostatic interactions derived from the learned partial charges, and a D3(BJ) dispersion correction.

Designed for generalizability, AIMNet2 natively supports systems with different charge states and spin multiplicities. By explicitly accounting for these varying electronic states and long-range interactions, the model is well-suited for a wide range of chemically complex systems.

L115 P5

AIMNet2 section: "While it may not match the data efficiency or accuracy of PaiNN for geometry-sensitive properties..." — Citations would be warranted to

back up this statement, unless it refers to the conclusions of this study, in which case this should be stated.

The referee is entirely justified in questioning this statement. Upon reflection, this claim was based on anecdotal observations during our group's work with these models, rather than a rigorous quantitative benchmark. Because we have not formally compared the relative data efficiencies of PaiNN and AIMNet2, and are unaware of published literature that provides this specific comparison, we agree that the statement lacks the necessary supporting evidence. Consequently, we have removed this sentence from the revised manuscript:

Designed for generalizability, AIMNet2 natively supports systems with different charge states and spin multiplicities. By explicitly accounting for these varying electronic states and long-range interactions, the model is well-suited for a wide range of chemically complex systems.

L126 P5

There are no citations in the Delta learning section. At least this statement "When the two levels of theory are correlated, this delta-learning approach can substantially reduce model errors." could do with a citation.

We agree with the referee that this statement warrants a citation. We have added a reference to the work by Ramakrishnan et al. (2015), which demonstrated that because the baseline method already captures the underlying physical chemistry, the remaining energy difference is much easier for a machine learning model to predict than learning the target value from scratch. Furthermore, to highlight earlier applications of Δ-learning specifically for molecular dynamics and atomic forces, we have also cited Bogojeski et al. (2020). The revised manuscript now reads:

In this framework, molecular dynamics (MD) simulations are performed at the lower level of theory but are corrected to approximate the high level of theory (Bogojeski et al., 2020). When the two levels of theory are correlated, this delta-learning approach can substantially reduce model errors (Ramakrishnan et al., 2015).

L131 P5

Delta learning section: "The main drawback is that the overall efficiency is fundamentally bounded by the cost of the lower-level baseline." The efficiency of what?

We thank the referee for pointing out this ambiguity. By "efficiency," we were referring to the overall computational cost of the simulation. While the evaluation of the machine learning model is typically fast, a Δ-learning approach also requires evaluating the baseline method (e.g., GFN1-xTB) at every time step. Consequently, the total simulation time is fundamentally bottlenecked by this baseline calculation. This limitation becomes especially pronounced for larger systems, as the computational cost of GFN1-xTB scales with the number of (valence) electrons, whereas the machine learning models scale with the number of atoms. We have

updated the manuscript to clarify this distinction:

The main drawback is that while the evaluation of the machine learning model is typically fast, the overall simulation speed is fundamentally limited by the computational cost of the lower-level baseline.

L134 P5

Data set generation: "Gradient calculations" - What are gradient calculations?

By "gradient calculations," we refer to the computation of the gradients of the potential energy with respect to the nuclear coordinates, which correspond to the negative atomic forces. These force evaluations were executed using the `engrad` keyword in ORCA, as well as the TBLite calculator within the Atomic Simulation Environment (ASE). We have updated the manuscript to explicitly clarify this terminology for a broader audience:

Atomic forces of the selected structures were obtained through potential energy gradient calculations with respect to the nuclear coordinates using GFN1-xTB (TBLite, version 0.2.1) and ωB97X-3c (ORCA, version 6.0.1) (Neese, 2012).

L155 P6

We note that while ωB97X-3c is used in this study, atomic forces can be calculated using any quantum chemistry method on the GFN1-xTB structures to obtain a training set at that level of theory.

L164 P6

Hyperparameter tuning: "we assigned a great weight to the force loss" - Presumably "great" should read "greater" or simply "larger".

We agree that "great weight" is an awkward phrasing and have changed it to "greater weight" as suggested:

..., we assigned a greater weight to the force loss.

L240 P9

Hyperparameter tuning: "This makes sense, as the product of the batch size and batches per epoch determines the total number of samples seen in one epoch." - I am slightly confused. Isn't the definition of an epoch that it is a full run through the data? Once the batch size is determined, the number of batches is set and then simply determines how often the weights are updated in an epoch.

We completely agree that the standard textbook definition of an epoch is exactly one full pass

through the training data. However, the AIMNet2 training framework (built on PyTorch Ignite) operationally defines an "epoch" as a fixed, user-defined number of training steps, detaching it from the actual dataset size.

In our specific setup, our dataset contains 20,000 molecules. With a batch size of 16, a traditional epoch would consist of 1,250 batches. Because the batches per epoch hyperparameter was set to 4,000, the model simply processes 64,000 samples (looping through the dataset multiple times) between each validation step.

We have updated the manuscript to state the operational definition and clarify the exact number of samples processed:

For PaiNN, the batch size and number of features were identified as the most important hyperparameters, with a smaller batch size and a higher number of features correlating with improved performance. In contrast, for AIMNet2, the "batches per epoch" was the dominant hyperparameter. This difference stems from how each training framework defines an epoch. In PaiNN, an epoch follows the standard definition of a single, full pass through the training data. Thus, the training set size and batch size strictly determine the batches per epoch. However, the AIMNet2 framework decouples the definition of an epoch from the dataset size, operationally defining it as a user-specified number of steps before each validation check. Therefore, in AIMNet2, the product of the batch size and the batches per epoch determines the total number of samples processed per operational epoch. When this product is smaller than the training set size, a validation step is triggered before the model has seen all training samples. When the product exceeds the dataset size (e.g., using 1,000 batches of size 16 means the AIMNet2 model processes 16,000 samples per validation cycle for our 2,000-molecule dataset), the model sees data multiple times per epoch, which further aids convergence.

L247 P9

Hyperparameter tuning: "It is important to note that we did not necessarily identify the optimal hyperparameter combination for our systems." — Did you consider using hyperparameter tuning with sample efficient Bayesian optimization as shown e.g., by Stuke et al , Mach. Learn.: Sci. Technol. 2, 035022 (2021)?

We thank the referee for bringing this relevant study to our attention. We had not previously considered sample-efficient Bayesian optimization for hyperparameter tuning. Given the low test errors obtained in the present study, our random grid-search approach proved sufficient for these relatively simple collision systems. However, we completely agree that a more rigorous optimization strategy will be valuable in the future. As we move to more complex simulations, such as simulating multiple particle-forming precursors within a simulation box, we plan to explore this Bayesian optimization.

We revised the manuscript as follows:

It is important to note that we did not necessarily identify the optimal hyperparameter combination for our systems. For instance, while our 100-epoch tuning procedure offers a reasonable indication of training behavior, some hyperparameters might converge slower

but dominate during longer training. Identifying the best training settings would require a systematic search over a broader range of values. While automated techniques such as Bayesian optimization (Stuke et al., 2021) could be explored for more complex systems in future work, the chosen hyperparameters provide sufficiently low test errors for the collision systems studied here, as discussed in the following subsection.

L264 P10

NN training: The accuracies reported in Table 3 are impressive. Already the PaiNN and AIMNet2 models achieve very accurate forces of only a few meV per Angstrom. The Delta-PaiNN model is then even better. What accuracy is actually required for the collision calculations?

We thank the referee for raising this highly relevant question.

As demonstrated in this study and by Knattrup et al. (2025), the long-range capture kinetics for these acid–base systems are not heavily dependent on atomistic specifics and can often be well approximated using simple electrostatic and dispersion potentials. However, our objective here is for the machine learning interatomic potential to accurately describe the entire cluster formation process: the initial approach, the collision, and the subsequent cluster formation through stabilizing proton transfers.

Capturing these events requires stable molecular dynamics trajectories. In the machine learning potential community, a mean absolute force error of 1 kcal/mol/Å is frequently used as a benchmark target for accuracy (e.g., Zaverkin and Kästner, 2021). However, as pointed out by Fu et al. (2022), mean absolute force errors only tell part of the story. The stability of molecular dynamics simulations is sensitive to outliers. Even if the mean errors are low, a single anomalous force prediction in a high-energy region can drive the dynamics off course, potentially leading to catastrophic failures such as unphysical fragmentation.

Because collision trajectories inherently explore these high-energy, non-equilibrium states, minimizing maximum force errors is critical. This is exactly why our study advocates moving beyond global scalar metrics (like mean absolute errors) and instead performing targeted validation against the physical properties of interest, such as spatially resolved errors (e.g., as a function of center-of-mass distance), potentials of mean force, and collision rate coefficients.

Finally, from a computational chemistry perspective, the standard target for thermodynamic properties is "chemical accuracy" (errors < 1 kcal/mol). It is important to distinguish between the error of the method compared to physical reality (the inherent error of the underlying level of theory) and the reproduction error of the machine learning model (which we report in Table 3). Because the total simulation error compounds both of these sources, the machine learning reproduction error must be negligible compared to the inherent error of the reference data. Therefore, achieving energy reproduction errors that fall well below 1 kcal/mol is necessary to ensure the physical outcomes are dictated by the electronic structure method rather than artifacts of the machine learning fit.

Line 265: "This error results from a mismatch between the training labels, which

include long-range stabilization, and the model's short-ranged (< 10 Å) representation. At these distances, the reference energies are significantly lowered by electrostatic interactions. However, due to the 10 Å cutoff, PaiNN interprets the collision partners as two non-interacting, free species. Consequently, during training, the model is forced to attribute the substantial stabilization energy of the interacting pair to the local atomic environments of the isolated monomers. In essence, the model erroneously learns that these structures, separated by more than the cutoff yet still interacting, are representative of the free molecular state, resulting in a fundamentally distorted PES." The analysis sounds interesting, although I would have expected it in the discussion section, but I got a little lost. Figure 2 reports the errors that the authors analyse, but where do we see a mismatch in training labels? And that PaiNN interprets the collision partners as two non-interacting, free species? I feel the reader is given insufficient information to follow the argument.

The referee makes a fair point. The concept we intended to describe is indeed nuanced, and our original phrasing was too abstract and lacked the necessary physical context to easily follow the argument.

The "mismatch" we intended to describe is between the potential energy of the interacting system in the training data and the physical limitations of the model's spatial cutoff. When the two collision partners are, for example, 12 Å apart, the electronic structure calculations in the training data already show a significantly lowered potential energy compared to the isolated monomers due to long-range ion–dipole interactions. These interactions also induce slight geometric distortions in the approaching monomers. To help visualize this, we have provided an illustrative potential of mean force in Fig. R1, showing how the system's energy has already dropped significantly at distances well beyond the 10 Å cutoff.

Because PaiNN has a strict 10 Å spatial cutoff, it evaluates this 12 Å system as two completely isolated, non-interacting monomers. During training, the model must map the significantly lowered potential energy (due to the long-range attractive interactions) to these effectively isolated geometries. Consequently, PaiNN erroneously learns that the slight internal geometric distortion is the cause of the lowered potential energy. It maps the interaction energy onto the isolated monomers, treating their distorted structures as new, highly stable global minima for the free monomers.

We have completely rewritten the paragraph in the manuscript to remove the confusing phrasing and explicitly explain the mechanism of this failure:

In contrast, the ionic H_2SO_4–HSO_4^- system illustrates the inherent limitations of applying a purely local representation to systems with strong long-range interactions. Although the electronic energy MAE for PaiNN appears relatively low, the distance-resolved error plot (Fig. 2c) reveals significant deviations at separations $r > 10$ Å. At these large distances, the training data already capture substantial stabilization energy due to long-range ion–dipole interactions, which also induce slight geometric distortions in the approaching collision partners. However, because PaiNN employs a strict 10 Å spatial cutoff, it evaluates the system as two completely isolated, non-interacting collision partners. During training, the model must map the significantly lowered potential energy of the interacting system to these

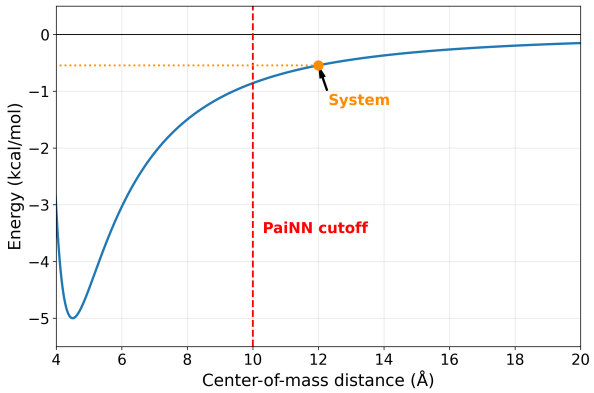

Figure R1: Illustrative potential of mean force highlighting the spatial cutoff limitation. At a center-of-mass distance of 12 Å (orange dot), the interacting system already experiences a lowered potential energy compared to the isolated monomers (Energy = 0 kcal/mol). However, because this distance exceeds PaiNN's 10 Å local atomic environment cutoff (red line), the model evaluates the system as two completely isolated, non-interacting monomers.

isolated, slightly distorted molecular structures. Consequently, PaiNN erroneously learns to associate the long-range electrostatic stabilization entirely with these slight internal structural changes. In essence, the model is forced to view this distorted geometry as the lowest energy conformer of the isolated molecule, creating an artificial global minimum that fundamentally distorts the PES.

L291 P11

Potential of mean force: ***"The potential of mean force (PMF) along the center-of-mass distance represents the effective free energy averaged over all collision orientations accessed during the simulations, showing how the system's stability changes as the collision partners approach. The well depth and shape provide information on the binding strength, while the shoulder towards larger distances reflects the strength of long-range interactions.´´ — Shouldn't one show a PMF as an example or a sketch of a typical PMF so that the reader can follow the statements here? Otherwise it is hard to imagine what the well depth and the shoulder refer to. (I see now that I have read on that PMFs are shown a little later in the section; maybe refer to them here already.)***

We thank the referee for this helpful suggestion. We agree that pointing the reader to a visual example earlier in the text makes the physical interpretation of the potential of mean force features much clearer. We have adjusted the manuscript to explicitly reference the potentials of mean force shown in Fig. 3 when introducing the concepts of the well depth and the shoulder:

The potential of mean force (PMF) along the center-of-mass distance represents the effective free energy averaged over all collision orientations accessed during the simulations, showing how the system's stability changes as the collision partners approach (see e.g., Fig. 3). The

well depth and shape provide information on the binding strength, while the shoulder towards larger distances reflects the strength of long-range interactions.

L303 P12

Table 5: How accurate would the collision rates need to be to be useful in downstream modelling? Presumably the small difference between GFN1-xTB and ωB97X-3c for the neutral dimers is not noticeable, but how about the difference of ∼ 1 for H_2SO_4–H_2SO_4? What I am looking for is some form of contextualisation.

We thank the referee for this question. It is indeed important to contextualize the required accuracy for collision rate coefficients in downstream applications.

To our knowledge, there is currently no community-wide standard for collision rate accuracy analogous to the "chemical accuracy" threshold (1 kcal/mol) used for quantum chemistry energies. Collision rate coefficients are primarily used in cluster distribution dynamics models (such as the Atmospheric Cluster Dynamics Code). In these models, the particle formation rate scales linearly with the collision rate coefficient.

However, formation rates also depend on evaporation rate coefficients, which are calculated from quantum chemically derived binding free energies. Because evaporation rates depend exponentially on these binding energies, errors in the underlying thermochemistry dominate the overall uncertainty. For instance, approaching chemical accuracy (an error of 1 kcal/mol) in binding free energies still introduces a factor of roughly 5 uncertainty in the evaporation rate at room temperature.

A reasonable overarching goal for cluster dynamics simulations is to maintain total particle formation rate errors below one order of magnitude. Allowing for a factor of 5 uncertainty from the evaporation rates, a conservative acceptable error margin for the collision rate coefficients would be a factor of 1.5. Keeping in mind additional sources of error (e.g., improperly defined sources/losses and finite simulation sizes), this factor of 1.5 bounds the compounding error within an order of magnitude.

Evaluated against this threshold, the reproduction errors of all models for all systems fall safely within this threshold, except when the PaiNN model is applied to the charged H_2SO_4–HSO_4^- system, where the error approaches 50% (a factor of ∼2).

We have updated the corresponding paragraph in the manuscript to provide this context:

To contextualize these results, it is important to note that the accuracy of particle formation rates in cluster distribution dynamics simulations depends on both the collision and evaporation rate coefficients. Because evaporation rates depend exponentially on binding free energies, these errors typically outweigh errors in collision rate coefficients. An error of just 1 kcal mol^{-1} in binding free energies introduces a factor of ∼5 uncertainty in the evaporation rate. As such, we consider an error of a factor of 1.5 in the collision rate coefficients acceptable. In the worst-case scenario where collision and evaporation rate coefficient errors compound in the same direction, a factor of 1.5 collision rate coefficient error would still result in an overall uncertainty in the particle formation rates of less than an order of magnitude.

Evaluated against this threshold, PaiNN yields notably lower rate coefficients for the charged H_2SO_4–HSO_4^- system. For both GFN1-xTB and ωB97X-3c training data, the model underestimates the rates by nearly 50% (roughly a factor of 2) relative to the GFN1-xTB reference. As discussed in Sec. 3.5, this substantial deviation stems from the model's inability to detect collisions beyond its 10 Å cutoff, effectively neglecting the significant contribution of the long-range tail.

L390 P18

References

Fu, X., Wu, Z., Wang, W., Xie, T., Keten, S., Gomez-Bombarelli, R., and Jaakkola, T.: Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations, arXiv preprint arXiv:2210.07237, https://doi.org/10.48550/arXiv.2210.07237, 2022.

Knattrup, Y., Neefjes, I., Kubečka, J., and Elm, J.: Growth of atmospheric freshly nucleated particles: a semi-empirical molecular dynamics study, Aerosol Research, 3, 237–251, https://doi.org/10.5194/ar-3-237-2025, 2025.

Zaverkin, V. and Kästner, J.: Exploration of transferable and uniformly accurate neural network interatomic potentials using optimal experimental design, Machine Learning: Science and Technology, 2, 035 009, https://doi.org/10.1088/2632-2153/abe294, 2021.