AARHUS UNIVERSITY

Dear Editor and Referee 2,

We sincerely thank referee 2 for their clear, constructive, and valuable feedback. Their insightful comments helped us to recognize the specific areas where our manuscript could be strengthened and clarified. We have taken all of their constructive critiques to heart and have revised the manuscript accordingly.

In the following, we provide point-by-point responses to all their comments. Referee comments are given in ***bold italic***, while our responses are given in roman (non-bold, non-italic). Excerpts from the revised manuscript to support our responses are highlighted using yellow highlight. The line and page number to which a response refers is indicated by (L### P#).

We believe that the 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

Neefjes et al. evaluate machine-learned interatomic potentials for molecular dynamics simulations of collisions between atmospherically relevant molecules. The study compares PaiNN, AIMNet2, and Δ-learning approaches trained on GFN1-xTB and ωB97X3c data. The authors show that models relying purely on local atomic environments can struggle to capture long-range interactions relevant for collision dynamics, whereas AIMNet2 performs well due to its explicit long-range treatment. While the importance of long-range interactions in MLIPs is not a new observation, the manuscript provides a useful and well-executed demonstration in the context of atmospheric chemistry. I recommend publication after the following revisions.

We thank the referee for their positive assessment and accurate summary of our study. We agree that the limitations of purely local machine learning interatomic potentials are well established within the broader computational chemistry community. However, as the referee notes, the specific consequences of these limitations on atmospheric particle formation modeling had not been previously investigated. We view this work as the necessary validation before these models can be deployed for large-scale nucleation simulations, demonstrating that they generally reproduce the underlying level of theory with low errors, while specifically confirming their ability to accurately capture the long-range forces governing collisions.

Page 5, line 114. "While it may not match the data efficiency or accuracy of PaiNN for geometry-sensitive properties..." Please provide a citation or quantitative evidence supporting this claim.

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

Page 4. Similarly, the discussion of PaiNN emphasizes its equivariant representation of vectorial properties. Does AIMNet2 include a comparable treatment of directional features, or does it rely on something different? Clarifying this would help readers compare the architectures.

We thank the referee for this question. It is indeed important to contrast PaiNN's use of equivariant directional features with the approach taken by AIMNet2. Unlike PaiNN, AIMNet2 does not enforce explicit equivariant representations for vectorial properties. Instead, it relies on invariant radial symmetry functions, capturing directional effects implicitly through its atom-centered environment representations and iterative message passing. We have now clarified this architectural distinction in the AIMNet2 method section:

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. While directional dependencies are not enforced through explicit equivariance, AIMNet2 captures these effects implicitly through its atom-centered representations and iterative message passing. 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.

L115 P5

Page 5, line 143. "Instead, it was assumed that the GFN1-xTB PES sufficiently overlaps with the relevant regions of the ωB97X-3c PES." I understand that MD with ωB97X-3c would be expensive. However, the manuscript would be improved by adding the implications of this decision.

We agree with the referee that the implications of this approach warrants proper discussion in the manuscript.

This dataset generation approach assumes that the topology of the GFN1-xTB PES qualitatively resembles that of the higher-level target method. Under this condition, minor structural discrepancies, such as slight differences in the bond lengths of local minima, are not problematic. The higher-level nuclear gradient calculations provide a net force directing the geometry toward the true ωB97X-3c minimum. Furthermore, while GFN1-xTB might incorrectly rank the relative energies of specific configurations, the ωB97X-3c calculations will correctly reorder them, provided all relevant local minima are sufficiently sampled during the semi-empirical dynamics.

The main limitation of this approach occurs if important regions of the ωB97X-3c configurational space are entirely missing from the GFN1-xTB PES. If the semi-empirical dynamics fails to visit specific critical conformations, the machine learning model will never be trained on them. In such cases, the training dataset would require explicit augmentation by sampling those missing configurations directly at the target level of theory.

For the relatively simple collision systems discussed here, this will most likely not be an issue, but for more complex clusters of flexible molecules, it is important to validate this assumption.

We have revised the manuscript to explicitly discuss these implications.

Instead, it was assumed that the GFN1-xTB PES sufficiently overlaps with the relevant regions of the ωB97X-3c PES. This assumption holds as long as the GFN1-xTB PES has

the same topological features as the ωB97X-3c PES in the relevant regions. Small structural discrepancies are corrected, as the higher-level nuclear gradients provide a net force directing the geometry toward the true ωB97X-3c minimum. The assumption, however, breaks down if important regions of the ωB97X-3c PES are entirely missing from the GFN1-xTB PES. While this is unlikely for the relatively simple collision systems studied here, more complex clusters may require the dataset to be augmented with unvisited structures.

L157 P6

In Figure 2 and Table 3, please clarify what the errors are relative to in the captions.

This is a fair point. We agree that it is easy for a reader to become confused about what the errors signify when the reference is not explicitly stated. The values shown in Figure 2 and Table 3 represent reproduction errors, i.e., the error of a machine learning model, trained on a specific level of theory, relative to calculations performed directly at that level of theory. To make this unambiguous, we have revised the captions as follows:

Table 3. Mean absolute errors (MAE) of the machine learning models relative to the level of theory they were trained on for electronic energies (E_{el}) and component-wise forces (F) across the three studied systems: H_2SO_4–H_2SO_4, H_2SO_4–$NH(CH_3)_2$, and H_2SO_4–HSO_4^-. Results are reported for AIMNet2, PaiNN, and Δ-PaiNN trained on GFN1-xTB and ωB97X-3c training data. Units: E_{el} in kcal mol^{-1} and F in kcal mol^{-1} Å^{-1}.

Figure 2. Electronic energy reproduction errors for machine learning models trained on ωB97X-3c data relative to the ωB97X-3c level of theory, shown as a function of the center-of-mass (COM) distance across the three studied systems: H_2SO_4–H_2SO_4 (a), H_2SO_4–$NH(CH_3)_2$ (b), and H_2SO_4–HSO_4^- (c). Results are shown for AIMNet2, PaiNN, and Δ-PaiNN.

I believe that Figure 2 isn't introduced in the text until after Table 3 is introduced. The authors should adjust the ordering of their manuscript.

We thank the referee for pointing this out. The inverted placement in the compiled manuscript was an artifact of LaTeX maintaining independent floating queues for figures and tables, which inadvertently allowed the table to bypass the earlier figure despite being placed correctly in our source code. We have implemented a float barrier in the revised manuscript to ensure they now appear in the correct order in the current single-column format. Should the manuscript be accepted, we will communicate this intended order to the typesetter during the proofing stage of the two-column layout.

Page 10, line 259. "For the H_2SO_4-H_2SO_4 system, performance is consistently low across the entire coordinate." This implies that performance is bad, when I think you mean to say that the errors are low across the entire coordinate.

This is indeed badly phrased. We have changed "performance is" to "the errors are":

For the H_2SO_4-H_2SO_4 system, the errors are consistently low across the entire coordinate.

L286 P11

Page 13, Table 4. It is my understanding that the Δ-PaiNN model here has been trained on GFN1-xTB as the baseline and target reference. I agree with the authors note on page 10 that this should yield an essentially-zero correction. Why then is the full RMSE of PaiNN 0.053 kcal/mol and the full RMSE of Δ-PaiNN 3.4 kcal/mol? I would expect these to be closer. If this difference is due to the "inherent uncertainty associated with finite umbrella sampling" (page 12), then should this have been properly mitigated or evaluated separately? Does this actually justify such a large error?

We thank the referee for raising this issue. It is indeed surprising that the Δ-PaiNN method does not reproduce the GFN1-xTB curve within a smaller margin of error, and an RMSE of 3.4 kcal/mol cannot be ascribed to uncertainties in the umbrella sampling simulations alone. To investigate this, we plotted the short-distance repulsive wall of the potential of mean force (PMF) for H_2SO_4–H_2SO_4 (Fig. R1). At energies above ∼30 kcal/mol, the PaiNN model perfectly follows the GFN1-xTB reference, while AIMNet2 and Δ-PaiNN erroneously bend toward a slower increase in energy. We have rigorously verified our data to ensure this is not an artifact of mislabeling the PaiNN and Δ-PaiNN models.

The energy deviation occurs at center-of-mass (COM) distances below 2.8 Å. Out of our 20,000-structure dataset, only 63 structures fall below this distance, and only 22 are below 2.5 Å. Although we sampled windows down to a COM distance of 2.0 Å, the large repulsive forces in this region counteract the bias potential. Consequently, the molecules rarely adopt configurations with COM distances below 2.8 Å, leaving the training set overwhelmingly biased toward larger distances. This bias is intentional, as our focus is on the system's behavior near the minimum and at long ranges. Accurately modeling the highly repulsive region would require a data generation scheme that explicitly targets this region.

To demonstrate that this region of the potential energy surface is physically negligible for our purposes, we evaluated the maximum kinetic energy available in our simulations for scaling the repulsive wall. We performed collision trajectory simulations with initial relative velocities up to 800 m/s, which corresponds to a kinetic energy of approximately 3.8 kcal/mol for the H_2SO_4–H_2SO_4 system. Even if this entire kinetic energy were converted to potential energy to climb the repulsive wall, the system would not reach the energy threshold where the curves begin to diverge.

Furthermore, we can calculate the standard-state (Helmholtz) binding free energy, $\Delta F^{\ominus}$, using the Δ-PaiNN curve as is versus a curve where the tail below 2.8 Å is corrected to perfectly follow the GFN1-xTB reference:

$$\Delta F^{\ominus} = -k_{\mathrm{B}}T \ln \frac{4\pi}{V^{\ominus}} \int_0^{r_{\max}} r^2 \exp\left(-\beta w(r)\right) \mathrm{d}r \,, \tag{1}$$

where k_B is the Boltzmann constant, T is the temperature, $V^\ominus$ is the standard volume, r is the COM distance, $\beta = 1/k_B T$, and $w(r)$ is the PMF. Integrating up to $r_{max} = 10$ Å yields -5.6154008 kcal/mol for the uncorrected Δ-PaiNN curve and -5.6154012 kcal/mol for the corrected curve. Thus, the deviation between GFN1-xTB and Δ-PaiNN in the highly repulsive region is entirely negligible for both equilibrium and collision properties.

The question remains as to why Δ-PaiNN deviates in the repulsive region rather than applying the expected near-zero correction to the GFN1-xTB baseline. Because of the extreme sparsity of training data at low COM distances, the model lacks the constraints necessary to uniquely identify the physically correct potential energy surface. Consequently, Δ-PaiNN overfits to an erroneous path in this unconstrained region, inappropriately adjusting the baseline GFN1-xTB results. While this artifact could be resolved by generating additional training data in the repulsive regime, we have refrained from retraining the models since the dynamics of this inaccessible region are outside the scope of our study.

Given the referee's valid criticism of the full RMSE, we propose removing the distinction between "full" and "shoulder" RMSE in the manuscript. Instead, we will only report the RMSE between the two points where the PMF crosses zero (i.e., in the repulsive and long-range regions). This modification does not alter our conclusions. Rather, it streamlines Table 4 and the accompanying text, ensuring that we only evaluate and discuss the regions of the potential energy surface for which the models were rigorously trained and which are physically accessible.

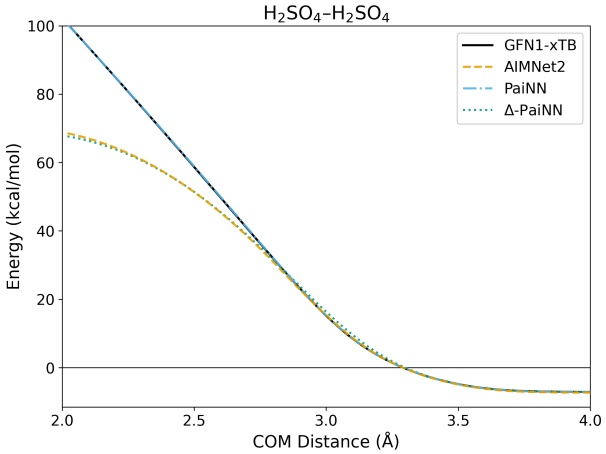

Figure R1: The repulsive region of the potential of mean force along the center-of-mass (COM) distance obtained from umbrella sampling for the H_2SO_4–H_2SO_4 collision system. Results are shown for GFN1-xTB and the AIMNet2, PaiNN, and Δ-PaiNN machine learning models trained on GFN1-xTB.

The revised table:

Table 1: Root-mean-square errors (RMSE) in kcal mol^{-1} for the potentials of mean force (PMFs) predicted by the machine learning models relative to the GFN1-xTB reference. The RMSE is evaluated between the point where the PMF drops below zero in the short-range repulsive region and the first point it returns to zero in the long-range non-interacting region.

System	Method	Evaluation Range (Å)	RMSE (kcal mol^{-1})
H_2SO_4–H_2SO_4	AIMNet2	3.28–11.06	0.15
	PaiNN		0.058
	Δ-PaiNN		0.042
H_2SO_4–$NH(CH_3)_2$	AIMNet2	2.72–8.78	0.046
	PaiNN		0.17
	Δ-PaiNN		—
H_2SO_4–HSO_4^-	AIMNet2	2.90–15.00	0.20
	PaiNN		0.090
	Δ-PaiNN		—

The revised accompanying text:

Figure 3 shows the PMFs for all three systems calculated using AIMNet2 and PaiNN (trained on GFN1-xTB), compared against the GFN1-xTB reference. For the H_2SO_4–H_2SO_4 system, the Δ-PaiNN model trained on GFN1-xTB was again included as a sanity check. While this model should theoretically reproduce the GFN1-xTB reference PMF, small deviations are nonetheless expected due to the inherent uncertainty associated with finite umbrella sampling. Additionally, because the machine learning models lack training data in the highly repulsive, physically inaccessible regimes at short distances, much larger deviations can occur in these regions (see Fig. S3 in the Supporting Information).

Table 4 lists the root-mean-square errors (RMSEs) of the predicted PMFs relative to the reference. We report the RMSE only between the point where the PMF drops below zero in the short-range repulsive region and the first point it returns to zero in the long-range non-interacting region. At shorter distances, the steep energies of the repulsive wall lead to sparse training data coverage, which can result in localized high errors. However, because these configurations are physically inaccessible at atmospheric temperatures, excluding this region ensures the reported RMSE reflects the model's performance in the region relevant to the clustering dynamics. Conversely, we exclude the asymptotic long-range tail because the collision partners are essentially non-interacting here. Including an extensive non-interacting region, which is well-sampled and exhibits minimal energy variation, would disproportionately lower the average error, masking the model's performance in the interaction region. While the PMF should theoretically approach zero asymptotically, sampling noise causes the long-range zero-crossing to occur at finite distances (< 20 Å) for the systems studied here.

L318 P12

Page 17, Table 5. The text regarding this figure would benefit from identifying what magnitude of error in the collision rate coefficients is considered acceptable for atmospheric modeling applications.

We thank the referee for this comment. It is indeed important to contextualize the obtained values. The accuracy of particle formation rates obtained from cluster distribution dynamics simulations depends on the accuracy of the collision rate coefficients and the cluster binding free energies used to calculate the evaporation rate coefficients. Because evaporation rate coefficients depend exponentially on the cluster binding free energy, an error of just 1 kcal/mol already results in a discrepancy of a factor of ∼5. As such, errors in binding free energies typically outweigh errors in collision rate coefficients. To keep errors in particle formation rates below an order of magnitude in the worst-case scenario where the evaporation and collision errors compound in the same direction, and assuming a binding free energy error of around chemical accuracy (1 kcal/mol), we consider an error of up to a factor of 1.5 in the collision rate coefficient to be acceptable.

We have added a paragraph explaining this context to the manuscript.

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 $\mathrm{kcal\,mol^{-1}}$ 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.

Conversely, for the neutral systems (H_2SO_4–H_2SO_4 and H_2SO_4–$NH(CH_3)_2$), the ML models trained on GFN1-xTB data exhibit excellent agreement with the reference calculations. All three architectures reproduce the GFN1-xTB reference rate coefficients closely, with the largest deviation observed for AIMNet2 applied to the H_2SO_4–$NH(CH_3)_2$ system (∼10% discrepancy).

L390 P18