nature portfolio

Peer Review File

Force-Free Molecular Dynamics Through Autoregressive Equivariant Networks

Corresponding Author: Dr Fabian Thiemann

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

A version of this paper was originally rejected for publication by Nature Machine Intelligence, however that decision was reconsidered after appeal by the authors.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Thiemann et al. present TrajCast, an autoregressive framework for accelerating molecular dynamics (MD) simulations by directly predicting atomic positions and velocities. The method is designed to enable efficient generation of long MD trajectories with significantly larger time steps, up to an order of magnitude beyond conventional approaches, while maintaining stability and fidelity in key physical observables such as the radial distribution functon(RDF) and vibrational density of states (VDOS). The paper is technically sound, and the results demonstrate clear advantages in terms of computational efficiency and trajectory smoothness compared to existing methods.

However, while the methodological contributions are solid, I do not find the work suitable for a broad, interdisciplinary journal such as Nature Machine Intelligence. Instead, I believe it would be more appropriately placed in a specialized journal focused on molecular simulations or computational chemistry. The core limitation lies in the scope of the impact. The use of longer time steps in MD is a known avenue for efficiency gains, and while a 10× speedup is noteworthy, it does not fundamentally address the deeper challenge in atomistic simulations: the exploration and discovery of new states or transitions across energy basins that are not present in the training data. In other words, the true bottleneck in many MD applications is not just generating longer trajectories within known regions of the potential energy surface, but in escaping from these regions to discover new, relevant configurations, something that inherently requires generalization beyond seen data. The model appears effective at interpolating within known basins but lacks evidence of robust generalization, which limits its applicability to the broader goals of data-driven molecular modelling. The method already requires a large training set from known molecular states, and there is no evidence that it can generalize to new states, which limits its use for replacing the robust computational methods to integrate Newton's equation of motion. Lastly, the simulations in the paper rely on classical force fields for defining the potential energy surface, and there is no demonstration that TrajCast can learn more complex many body potential energy surfaces, and whether learning those makes the model more data hungry.

In summary, while TrajCast is a well-executed method for trajectory acceleration with clear engineering merit, its scientific contribution may not meet the bar for a general AI journal. I recommend publication in a domain-specific journal where its technical depth and practical implications will be most appreciated.

(Remarks on code availability)
Code, model weights, and generated training data are published.

Reviewer #2

(Remarks to the Author)
Thiemann and co-workers have proposed a method TrajCast that can generate MD-like trajectories with significantly less computational resources. Their method employs equivariant message passing neural networks (MPNNs) to

autoregressively predict atomic positions and velocities thereby bypassing standard MD integration steps. While the work is interesting, there are several clarity issues and discrepancies that the author should consider addressing in their manuscript, before it is considered for publication.

In the “TrajCast Framework” section, the authors described that their equivariant MPNN model takes as input the atomic positions r(t)), velocities v(t), and atomic types Z. However, the accompanying training code appears to also rely on displacements and updated velocities. These quantities are not explicitly defined or discussed in the manuscript. For the sake of clarity and reproducibility, it would be beneficial if the authors could clarify what these additional inputs represent, how they are derived from molecular dynamics trajectories, and their specific role during training. The scripts or input files required to generate these data should be provided to comprehend the proposed workflow.

The update procedure for positions and velocities could be more clearly explained, including why such a correction is necessary and how it is implemented. This may also raise natural questions about whether modifying the system in this manner preserves consistency with the underlying potential energy surface and maintains correct Boltzmann sampling. A short discussion of how potential energy is treated during this process would help address such concerns. Moreover, it is unclear how introducing velocity rescaling via the CSVR thermostat improves the stability of TrajCast roll-outs. The phrase “autoregressive roll-outs tend to accumulate errors” could also benefit from a clearer explanation specifically, what kind of errors are accumulating, and why this leads to instability. Overall, restructuring this paragraph and expanding on the motivation for each design decision (possibly with a supporting schematic) would improve readability and help readers better understand the physical and computational reasoning behind the proposed methodology.

The authors state that their approach “ensures transferability across different chemical spaces.” It would be beneficial to clarify what is specifically meant by “chemical spaces” in this context. For example, does this refer to generalization across molecular conformations, different atomic species or bonding types, entirely new compounds, or even broader distinctions such as organic versus inorganic materials?

From the discussion, it appears that the model is trained on NVE trajectories but is capable of generating both NVE and NVT trajectories at inference time, depending on whether a thermostat is applied during roll-out. Please clarify this in the main text.

Including a schematic overview of the workflow—from training on NVE data to inference under either ensemble—could greatly improve clarity, particularly for readers less familiar with molecular dynamics protocols.

The statement that "direct comparison with MD is infeasible, as trajectories will inevitably depart due to the thermostat-induced randomness" may be somewhat misleading. In standard MD practice, frame-by-frame trajectory alignment is generally not meaningful, and comparisons are instead based on ensemble-averaged quantities.

Finally, there appears to be a potential inconsistency in the following two statements. On one hand, the manuscript notes that TrajCast “tends to accumulate errors” during autoregressive roll-out and that applying a CSVR thermostat helps stabilize predictions. On the other hand, it is stated here that TrajCast can accurately reproduce NVE trajectories without a thermostat. These two claims may appear contradictory or at least not fully reconciled. It would be beneficial for the authors to clarify under which conditions TrajCast can produce stable NVE trajectories, and when the use of a thermostat becomes necessary for maintaining stability during long roll-outs.

The symbol α is used both as an index over vector components in the loss function and as a rescaling factor in the context of the CSVR thermostat. To avoid ambiguity, the authors may consider using a distinct notation for the thermostat-related rescaling factor, or disambiguate the contexts in which α appears.

The authors stated that TrajCast-generated trajectories closely match the potential energy distribution obtained from classical MD simulations (e.g., Fig. 2B), and that this agreement demonstrates the model’s ability to sample from the Boltzmann (canonical) distribution. Since TrajCast is trained solely on atomic positions and velocities—and does not receive explicit information about potential energy or forces—it is not immediately clear how it can reproduce the correct energy distribution, or for that matter any observable property. Given that potential energy is not an input, output, or training target, and is not directly derivable from velocity alone, the reported agreement with the Boltzmann distribution is surprising! A more detailed explanation of how this level of consistency is achieved—particularly in the absence of energy-driven learning—would strengthen the physical justification of the model’s behavior. This issue becomes even more pertinent when considering ensemble properties such as radial distribution functions and free energy surfaces (e.g., Fig. 2C). These observables are generally obtained through statistical sampling over configuration space, where the underlying distribution is weighted by potential energy (i.e., Boltzmann factors). Free energy landscapes, in particular, are sensitive to energy barriers and metastable configurations, yet TrajCast appears to reproduce them with high fidelity. It would be helpful if the authors could clarify how a model trained only on position-velocity data—without any explicit knowledge of energetic landscapes—can yield accurate ensemble observables and thermodynamic properties.

(Remarks on code availability)
In the “TrajCast Framework” section, the authors described that their equivariant MPNN model takes as input the atomic positions r(t)), velocities v(t), and atomic types Z. However, the accompanying training code appears to also rely on displacements and updated velocities. These quantities are not explicitly defined or discussed in the manuscript.

I could not check the codes due to the missing information and or scripts.
Detailed 'README' with instructions on the running the codes provided would be useful.

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
I read the authors response. I am very happy with the extra analysis they performed, showcasing how TrajCast can accelerate MD beyond known energy basins and training set. the authors have fully addressed my concerns and I recommend this paper to be published in Nature Machine Intelligence.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors have addressed most of my concerns/queries. The clarity issues have been resolved. However, I am still skeptical in the application and immediate impact of this work for the following reasons,

-- The trajectories predicted by TrajCast are limited to sampling relatively short timescales.
-- The conformational transitions of paracetamol in the gas phase is not really a rare event; one gets to sample the states with long unbiased simulations and then calculate the free energy surface from the equilibrium probabilities using lower-dimensional collective variables.
-- The prediction of an ensemble of configurations from metastable states separated by large (free)energy barriers is a challenge, which I am not so sure if the proposed method has convincingly addressed.
-- Bulk crystal and liquid/glass states usually do not have large configurational space, which makes TrajCast predict the configurations which are not far from the existing distribution. This might not be the case when one simulates rare events in the condensed phase.
– The computational framework is based on the NVE ensemble, which has been extrapolated to NVT by implementing a thermostat. While it works for now (short MD trajectories), obtaining a canonical distribution would still be challenging. In applications such as phase transitions, one needs to switch to the NPT, which would be a non-trivial extension.

(Remarks on code availability)

(Dated: 23 September 2025)

Response to Referees' Comments

We repeat all reports in full length *in italics*, respond in a point-by-point manner to all questions and suggestions, we refer **in bold** to changes made to the manuscript. In the revised manuscript, we show in red the text added or modified.

RESPONSE TO THE COMMENTS OF REFEREE #1

> *Thiemann et al. present TrajCast, an autoregressive framework for accelerating molecular dynamics (MD) simulations by directly predicting atomic positions and velocities. The method is designed to enable efficient generation of long MD trajectories with significantly larger time steps, up to an order of magnitude beyond conventional approaches, while maintaining stability and fidelity in key physical observables such as the radial distribution function(RDF) and vibrational density of states (VDOS). The paper is technically sound, and the results demonstrate clear advantages in terms of computational efficiency and trajectory smoothness compared to existing methods.*
>
> *However, while the methodological contributions are solid, I do not find the work suitable for a broad, interdisciplinary journal such as Nature Machine Intelligence. Instead, I believe it would be more appropriately placed in a specialized journal focused on molecular simulations or computational chemistry. The core limitation lies in the scope of the impact.*

We thank Referee 1 for carefully reviewing our manuscript and their positive assessment of its quality and advances. We also greatly appreciate their constructive feedback and comments. In response, we have undertaken a major revision of the manuscript, which we believe has significantly improved the overall impact of our work.

> ***Remark 1.1:*** *The use of longer time steps in MD is a known avenue for efficiency gains, and while a 10× speedup is noteworthy, it does not fundamentally address the deeper challenge in atomistic simulations: the exploration and discovery of new states or transitions across energy basins that are not present in the training data. In other words, the true bottleneck in many MD applications is not just generating longer trajectories within known regions of the potential energy surface, but in escaping from these regions to discover new, relevant configurations, something that inherently requires generalization beyond seen data.*

We are grateful for this insightful comment from Referee 1 and agree that the ability to explore the phase space beyond the training distribution is essential for the broader utility of data-driven trajectory generation. **In our major revision, we have performed several important changes and provide additional details to address the point of generalizability.**

In particular, we revisited the paracetamol experiment and its free energy surface defined by the hydroxyl and amine torsional dihedrals. In both MD and *TrajCast* NVT trajectories, the amine flip

($|\psi_2| > \pi/2$) is observed, but occurs very rarely, corresponding to a free energy barrier $\gtrsim 7\ k_BT$. **To determine whether *TrajCast* generates this transition from prior examples or can escape local minima by producing novel configurations, we have examined the coverage of the torsion angles in the training data, as detailed in the new section C.1 of the SI.** This analysis shows that the amine dihedral, ψ_2, is confined to the range of $[-\pi/2, \pi/2]$ across the entire training set. In fact, the amine flip does not occur throughout MD NVE trajectories from which we sampled training, validation, and test sets. These results provide strong evidence that *TrajCast* can escape local minima, generate configurations not seen during training, and accurately reproduce their relative significance in the free energy landscape. **We have added a short discussion of this analysis in the subsection on paracetamol in the Results and highlighted these capabilities in Introduction and Conclusion.**

> ***Remark 1.2:*** *The model appears effective at interpolating within known basins but lacks evidence of robust generalization, which limits its applicability to the broader goals of data-driven molecular modelling. The method already requires a large training set from known molecular states, and there is no evidence that it can generalize to new states, which limits its use for replacing the robust computational methods to integrate Newton's equation of motion.*

We thank Referee 1 for this important point. While our analysis of the dihedral coverage in paracetamol demonstrates that *TrajCast* can escape free-energy minima and generate configurations outside the training distribution, we agree that an additional benchmark is needed to more directly assess its generalizability.

To address this point, we have performed an additional experiment demonstrating that *TrajCast* achieves quantitative accuracy with MD in a zero-shot setting. Specifically, we apply the model trained on liquid water at 300 K to simulate the cooling and the amorphization of water into the hyperquenched glassy state. To this end, we adapted our CSVR thermostat implementation to perform temperature ramps, cooling the system at 10 K/ns down to 180 K.

This benchmark is particularly demanding due to the glassy phase transition inducing profound changes in both the dynamics and structure of the system which are entirely absent from the training distribution. The low cooling rate requires long trajectories, resulting in our longest roll-out of 2.4 million autoregressive steps (12 ns). To further demonstrate scalability and reduce finite-size effects that could blur the comparison between reference and prediction, we performed these simulations on a system of 216 water molecules, more than three times larger than the 64 molecules used during training. Remarkably, despite these challenges, *TrajCast* remains fully stable across four independent runs and quantitatively reproduces the dynamical (diffusion coefficient, new Fig. 5A) and structural changes (RDFs and order parameters, new Fig. 5B and 5C) observed in MD, demonstrating that, by learning how atoms move conditioned on their local environment, the model can generalize beyond the training distribution and capture emergent physical phenomena. Unlike a recent MLIP study (*R.J. Szukalo et al., Proc. Natl. Acad. Sci. U.S.A. 122 (32) e2509609122 (2025)*) that incorporated crystalline phases of ice in their training set, *TrajCast* achieves very good agreement with its reference while being trained solely on ambient liquid data, highlighting its data efficiency, robustness, and true generalizability across thermodynamic regimes.

We have added a new subsection on these zero-shot experiments to the Results section, linked it in the Introduction, Conclusion, and Abstract, and included a section in the SI presenting all diffusion coefficients and order parameters resolved by seed, as well as the RDFs for each temperature and element combination.

We believe that these changes significantly enhance the impact of our work, providing strong evidence of *TrajCast*'s ability to generalize beyond interpolation and demonstrating its usefulness well beyond the efficient generation of configurations similar to those encountered during training.

> ***Remark 1.3:*** *Lastly, the simulations in the paper rely on classical force fields for defining the potential energy surface, and there is no demonstration that TrajCast can learn more complex many body potential energy surfaces, and whether learning those makes the model more data hungry.*

We thank the referee for this important point. Indeed, in this work we have restricted ourselves to simulations based on classical force fields, which provide a well-controlled setting to demonstrate and analyze the capabilities of *TrajCast*. Our goal here was to introduce the architecture and establish its ability to learn from trajectories, rather than to benchmark it against the most complex potential energy surfaces available.

Further, the design of *TrajCast* is not tied to classical force fields, and in principle it can be trained on trajectories generated from a more accurate many-body PES such as MLIPs or *ab initio* MD. Regarding the concern that a more complex PES may require significantly larger datasets, we note that NequIP has demonstrated that training on CCSD/CCSD(T)-level data can achieve, on average, identical or even lower MAEs for energies and forces compared to training on DFT data, when using the same number of configurations (*S. Batzner et al., Nat. Commun. 13, 2453 (2022)*). Further, recent investigations (*A. E. A. Allen et al., npj Comput. Mater. 10, 154 (2024)*) combining different levels of theory have shown that even data at lower levels can be useful, for example through pretraining or meta-learning, to enable a model to subsequently reach higher levels of accuracy without requiring all reference data to be generated at the most expensive *ab initio* quality. While these results have so far been established for MLIPs, we believe that, in light of the architectural similarities, the same principles could be transferable to trajectory forecasting with *TrajCast*.

We added three sentences in the Conclusion addressing this point.

> ***Remark code availability****: Code, model weights, and generated training data are published.*

We appreciate the reviewer noting the availability of the code, model weights, and training data.

> *In summary, while TrajCast is a well-executed method for trajectory acceleration with clear engineering merit, its scientific contribution may not meet the bar for a general AI journal. I recommend publication in a domain-specific journal where its technical depth and practical implications will be most appreciated.*

We thank Referee 1 again for their thoughtful and constructive comments. We believe that our major revisions to address their points have significantly improved the impact and clarity of our manuscript.

RESPONSE TO THE COMMENTS OF REFEREE #2

> *Thiemann and co-workers have proposed a method TrajCast that can generate MD-like trajectories with significantly less computational resources. Their method employs equivariant message passing neural networks (MPNNs) to autoregressively predict atomic positions and velocities thereby bypassing standard MD integration steps. While the work is interesting, there are several clarity issues and discrepancies that the author should consider addressing in their manuscript, before it is considered for publication.*

We are grateful to Referee 2 for reviewing our work and for their valuable feedback and comments. In response, we have substantially revised the manuscript to improve clarity and address the points raised. We believe these changes significantly strengthen the presentation of our method and hope they satisfactorily resolve the referee's concerns.

> ***Remark 2.1:*** *In the "TrajCast Framework" section, the authors described that their equivariant MPNN model takes as input the atomic positions r(t)), velocities v(t), and atomic types Z. However, the accompanying training code appears to also rely on displacements and updated velocities. These quantities are not explicitly defined or discussed in the manuscript. For the sake of clarity and reproducibility, it would be beneficial if the authors could clarify what these additional inputs represent, how they are derived from molecular dynamics trajectories, and their specific role during training. The scripts or input files required to generate these data should be provided to comprehend the proposed workflow.*

We thank the referee for this comment and agree that the description can be clarified. *TrajCast* only requires atomic positions, atomic types, and current velocities as inputs. It directly predicts the atomic displacements (from which the new positions are obtained via simple addition) and the updated velocities.

The training code requires (via `reference_fields`) the labels pointing to the displacements and next-step velocities in the XYZ files comprising the training, validation, and test data. This accommodates different naming conventions (e.g., `new_velocities` versus `updated_velocities`). However, these quantities themselves are not used as model inputs. A detailed walkthrough of training a *TrajCast* model explaining all keywords and arguments is provided in our accompanying example notebook (`examples/training/training.ipynb`) in our repository on GitHub.

To improve clarity and address the referee's comment, we have restructured the *TrajCast Framework* section to clearly list the inputs (atom types, positions, current velocities) and emphasize that the MPNN predicts atomic displacements and new velocities. We also highlight this in the Methods section under Reference Datasets and explicitly state how these targets are defined and computed.

> ***Remark 2.2:*** *The update procedure for positions and velocities could be more clearly explained, including why such a correction is necessary and how it is implemented. This may also raise natural questions about whether modifying the system in this manner preserves consistency with the underlying potential energy surface and maintains correct Boltzmann sampling. A short discussion of how potential energy is treated during this process would help address such concerns. Moreover, it is unclear how introducing velocity rescaling via the CSVR thermostat improves the stability of TrajCast roll-outs. The phrase "autoregressive roll-outs tend to accumulate errors" could also benefit from a clearer explanation*

specifically, what kind of errors are accumulating, and why this leads to instability. Overall, restructuring this paragraph and expanding on the motivation for each design decision (possibly with a supporting schematic) would improve readability and help readers better understand the physical and computational reasoning behind the proposed methodology.

We appreciate this feedback from Referee 2 regarding the momentum conservation and error propagation within *TrajCast*.

Focusing on the physical constraints first, we refine the output of the readout layer for displacements and velocities to enforce conservation of total linear and angular momentum in the NVE ensemble. Importantly, these constraints are incorporated directly within the model architecture and consistently during both training and roll-out. This ensures that the network weights are optimized with these corrections enforced, rather than applied *post hoc*, and preserves consistency with the reference dynamics without modifying the underlying potential energy surface. **To improve clarity, we have restructured the paragraph on incorporating momentum conservation explaining into the architecture and now explicitly mention that the constraint layer is applied consistently during both training and roll-out.**

Regarding stabilizing the model predictions in long roll-out simulations, small stepwise prediction errors can accumulate over time, causing drift in positions and velocities and potentially leading to instability or unphysical behavior (such as artificial bond dissociation or molecule explosion). Similar to successes in autoregressive models for turbulence modeling, we find that the noise injected by canonical rescaling of velocities using the CSVR thermostat helps mitigate this issue. While this does not result in the exact NVE trajectory, this allows to maintain a physically meaningful distribution of particle velocities around the target temperature providing overall stabilization of the trajectory generation process. **We have rephrased and extended the paragraph focusing on NVT roll-outs in the TrajCast framework section and clarified the reasoning behind the ensemble choice.**

Remark 2.3: *The authors state that their approach "ensures transferability across different chemical spaces." It would be beneficial to clarify what is specifically meant by "chemical spaces" in this context. For example, does this refer to generalization across molecular conformations, different atomic species or bonding types, entirely new compounds, or even broader distinctions such as organic versus inorganic materials?*

We thank the referee for this comment and agree that the term "transferability across chemical space" can be ambiguous. In particular, we note that the link between interpolating between similar molecules or materials and predicting atomic velocities or displacements based on local environments is not strictly correct. **We have accordingly removed the respective phrase from the paragraph.**

To clarify, when referring to *interpolation across chemical space* we mean the concept of foundation model, recently explored for machine learning interatomic potentials (MLIPs). In this context, models trained on large databases of inorganic crystals have been shown, within accuracy limits, to generalize to chemically diverse systems such as organic molecules even if these were not explicitly included in the training data. The only requirement is the presence of the relevant chemical species in the training set albeit they may appear in different phases and types of systems. Such foundation models can provide an excellent starting point and can then be fine-tuned in a data-efficient manner to model the specific systems of interest.

While we trained individual *TrajCast* models for the systems explored in our experiments, we emphasize that the architecture itself is, in principle, capable of treating the same chemical element across different environments, for instance carbon in a diamond lattice and in an organic molecule, assuming it has seen sufficient training data. **We have added a sentence in the conclusion clarifying this idea and connecting it with universal MLIPs.**

> ***Remark 2.4:*** *From the discussion, it appears that the model is trained on NVE trajectories but is capable of generating both NVE and NVT trajectories at inference time, depending on whether a thermostat is applied during roll-out. Please clarify this in the main text.*

We thank Referee 2 for this comment. Indeed, TrajCast is trained exclusively on NVE data, while the trajectories for our experiments are generated in the NVT ensemble by rescaling after each MPNN forward pass the predicted velocities using the CSVR thermostat. Although this was mentioned in the original manuscript (*TrajCast* Framework and Methods sections), we agree it can be stated more explicitly.

We have therefore modified and extended the description of trajectory forecasting in the NVT ensemble in the *TrajCast* Framework section and added a clarifying statement in the final paragraph of the same section, emphasizing that the model is always trained on NVE trajectories, regardless of whether trajectories are later generated in the canonical ensemble.

> ***Remark 2.5:*** *Including a schematic overview of the workflow—from training on NVE data to inference under either ensemble—could greatly improve clarity, particularly for readers less familiar with molecular dynamics protocols.*

We thank the referee for this suggestion. **We have updated Figure 1 by adding labels that explicitly indicate training on NVE data and inference under the NVT ensemble. The figure caption has also been adjusted to clarify this distinction.**

> ***Remark 2.6:*** *The statement that "direct comparison with MD is infeasible, as trajectories will inevitably depart due to the thermostat-induced randomness" may be somewhat misleading. In standard MD practice, frame-by-frame trajectory alignment is generally not meaningful, and comparisons are instead based on ensemble-averaged quantities.*

We thank Referee 2 for this comment and fully agree that MD trajectories should be compared based on ensemble averages. This is the approach we have taken throughout the manuscript and had already stated in the original version. **To make this point clearer, we have revised the first paragraph in the Experiments section to explicitly note that comparison via ensemble averages is standard practice.**

> ***Remark 2.7:*** *Finally, there appears to be a potential inconsistency in the following two statements. On one hand, the manuscript notes that TrajCast "tends to accumulate errors" during autoregressive roll-out and that applying a CSVR thermostat helps stabilize predictions. On the other hand, it is stated here that TrajCast can accurately reproduce NVE trajectories without a thermostat. These two claims may appear contradictory or at least not fully reconciled. It would be beneficial for the authors to clarify under which conditions TrajCast can produce stable NVE trajectories, and when the use of a thermostat becomes necessary for maintaining stability during long roll-outs.*

We thank the referee for highlighting this point and agree that this statement could be interpreted as inconsistent. While TrajCast can reproduce NVE trajectories accurately based on a frame-by-frame comparison, this may be only reliable for relatively short roll-out periods, such as those shown in the SI (up to 1,000 steps). For longer timescales, as in the simulations presented in the manuscript involving up to 2.4 million autoregressive steps, small errors can accumulate, potentially leading to instability or unphysical behavior. Therefore, for such extended simulations we recommend using the CSVR thermostat to maintain stability and enable meaningful comparison with experimental results, which are typically obtained at constant temperature. **To reconcile this, we have added a sentence in the introductory paragraph of the Experiments section and at the end of section B.2 of the SI, noting that accurate forecasting in NVE is limited to short trajectories and that the CSVR thermostat is required for accessing longer timescales.**

> ***Remark 2.8:*** *The symbol α is used both as an index over vector components in the loss function and as a rescaling factor in the context of the CSVR thermostat. To avoid ambiguity, the authors may consider using a distinct notation for the thermostat-related rescaling factor, or disambiguate the contexts in which α appears.*

We thank the referee for pointing out this potentially confusing ambiguity. **We have renamed the Cartesian component index of displacements and new velocities in the loss function to from α to ν, while retaining α exclusively for the CSVR thermostat rescaling factor.**

> ***Remark 2.9:*** *The authors stated that TrajCast-generated trajectories closely match the potential energy distribution obtained from classical MD simulations (e.g., Fig. 2B), and that this agreement demonstrates the model's ability to sample from the Boltzmann (canonical) distribution. Since TrajCast is trained solely on atomic positions and velocities—and does not receive explicit information about potential energy or forces—it is not immediately clear how it can reproduce the correct energy distribution, or for that matter any observable property. Given that potential energy is not an input, output, or training target, and is not directly derivable from velocity alone, the reported agreement with the Boltzmann distribution is surprising! A more detailed explanation of how this level of consistency is achieved—particularly in the absence of energy-driven learning—would strengthen the physical justification of the model's behavior. This issue becomes even more pertinent when considering ensemble properties such as radial distribution functions and free energy surfaces (e.g., Fig. 2C). These observables are generally obtained through statistical sampling over configuration space, where the underlying distribution is weighted by potential energy (i.e., Boltzmann factors). Free energy landscapes, in particular, are sensitive to energy barriers and metastable configurations, yet TrajCast appears to reproduce them with high fidelity. It would be helpful if the authors could clarify how a model trained only on position-velocity data—without any explicit knowledge of energetic landscapes—can yield accurate ensemble observables and thermodynamic properties.*

We thank the reviewer for highlighting this point. *TrajCast* is trained to reproduce the evolution of atomic positions and velocities over short time intervals within the NVE ensemble. In doing so, it effectively learns the conditional probability distribution of how atoms move (their positions and velocities change) given their local environment, implicitly encoding energetic constraints such as steric repulsion, chemical bonds, and collective modes. While the potential energy is never used directly, these constraints are present in the training data generated from MD simulations.

As a result, long autoregressive roll-outs statistically reproduce the same configuration space with the correct Boltzmann weighting (when sampled in NVT), which in turn yields ensemble properties including RDFs, energy distributions, and free-energy surfaces in close agreement with MD, despite the absence of explicit energies or forces. Moreover, because the model captures the local dynamical rules rather than memorizing trajectories, it can generalize beyond its training distribution, enabling it to escape free energy minima (amine flip) and reproduce phenomena such as water amorphization in our additional zero-shot experiment. **We have added a note highlighting this point in the discussion of the amine flip at the end of the Paracetamol section.**

> ***Remark code availability**: In the "TrajCast Framework" section, the authors described that their equivariant MPNN model takes as input the atomic positions r(t)), velocities v(t), and atomic types Z. However, the accompanying training code appears to also rely on displacements and updated velocities. These quantities are not explicitly defined or discussed in the manuscript.*
>
> *I could not check the codes due to the missing information and or scripts. Detailed 'README' with instructions on the running the codes provided would be useful.*

We thank the referee for their comments regarding code availability. As noted in Remark 2.1, displacements and updated velocities are not additional inputs but the labels used during training. In our example training notebook, these labels can be easily computed via `compute_additional_fields` of our `ASETrajectory` class, with further details provided in the corresponding docstring.

We also note that the repository contains a README with links to two example notebooks that provide step-by-step walkthroughs for both training and trajectory generation. To our knowledge, these resources have enabled users to run the code successfully.

We would like to thank Referee 2 for their careful review and constructive feedback. We believe that addressing their comments has significantly improved the clarity of our manuscript.

ADDITIONAL CHANGES

Beyond the revisions prompted directly by the referee reports, we have also updated the manuscript to improve clarity and completeness. In particular, **we added recent references in the Introduction and the Results section, including work that appeared after our submission**, to better situate our contribution in the literature. These additions do not affect our conclusions but provide further context.

(Dated: 13 January 2026)

Response to Referees' Comments

We repeat all reports in full length *in italics*, respond in a point-by-point manner to all questions and suggestions, we refer **in bold** to changes made to the manuscript. In the revised manuscript, we show in red the text added or modified.

RESPONSE TO THE COMMENTS OF REFEREE #1

> *I read the authors response. I am very happy with the extra analysis they performed, showcasing how TrajCast can accelerate MD beyond known energy basins and training set. the authors have fully addressed my concerns and I recommend this paper to be published in Nature Machine Intelligence.*

We thank Referee 1 for their positive assessment and are happy that the additional analysis addressed their concerns. We appreciate the recommendation for publication and would like to thank the reviewer again for their efforts and constructive feedback which has allowed us to significantly improve the impact of our manuscript.

RESPONSE TO THE COMMENTS OF REFEREE #2

> *The authors have addressed most of my concerns/queries. The clarity issues have been resolved. However, I am still skeptical in the application and immediate impact of this work for the following reasons.*

We thank Referee 2 for reading the revised version of our manuscript and acknowledging that the clarity issues have been resolved.

> ***Remark 2.1:*** *The trajectories predicted by TrajCast are limited to sampling relatively short timescales.*

We thank Referee 2 for raising this point. TrajCast is not intrinsically limited to sampling short timescales. In this work, trajectories of up to 12 ns were generated. These trajectories were sufficient to fully converge all relevant structural and thermodynamic observables considered here. While the simulations could, in principle, be extended further, additional propagation did not provide new physical insight and was therefore unnecessary for the present analysis. **We have added a sentence in the section on paracetamol emphasising that the trajectories could have been extended further and were merely stopped due to the convergence of the properties of interest.**

> ***Remark 2.2:*** *The conformational transitions of paracetamol in the gas phase is not really a rare event; one gets to sample the states with long unbiased simulations and then calculate the free energy surface from the equilibrium probabilities using lower-dimensional collective variables.*

We thank the Referee for raising this point. We agree that the flipping of the amine group in paracetamol can be captured with unbiased simulations. **We have clarified this point in the revised manuscript.**

However, the purpose of this experiment is to demonstrate that TrajCast can capture states that are absent from the training set. In our experiment, TrajCast is trained only on configurations within a single basin obtained from NVE trajectories in which the conformational flip is not observed. Despite this, during inference in the NVT ensemble TrajCast generates trajectories in which transitions to previously unseen configurations arise naturally and remain physically meaningful, and with the correct energy barrier relative to the reference MD simulations in the canonical ensemble. This provides a stringent test of the model's ability to predict conformational changes without explicit prior exposure. We provide more information below, in our response to Remark 2.3.

> ***Remark 2.3:*** *The prediction of an ensemble of configurations from metastable states separated by large (free)energy barriers is a challenge, which I am not so sure if the proposed method has convincingly addressed.*

We thank Referee 2 for their comment. In our work, we show how TrajCast probes metastability in two complementary regimes. First, for paracetamol, the method predicts the flipping of the amine group, corresponding to two metastable configurations separated by an activation free-energy barrier $\gtrsim 7\ k_B T$. This barrier height is sufficient to define distinct metastable basins separated by activated dynamics. Importantly, configurations corresponding to the flipped state and the transition region are absent from the training set. Despite this, the reconstructed free-energy profile is in quantitative agreement with classical molecular dynamics simulations, indicating that the method captures both metastable states and the barrier separating them.

Second, we demonstrate the generation of hyperquenched glassy water that is structurally indistinguishable from that obtained using classical molecular dynamics at the same cooling rate. Glass formation corresponds to a kinetically arrested non-equilibrium state accessed under rapid cooling (we provide more details in the answer to Remark 2.4). In this case, the model is trained exclusively on liquid water configurations at room temperature and is never exposed to glassy states. The ability to reproduce the glassy ensemble, therefore, demonstrates generalization into metastable regions of configuration space reached under strong non-equilibrium driving and absent from the training data.

Our results show that TrajCast can generate physically meaningful ensembles across metastable regimes, encompassing both activated transitions and kinetically trapped states, within the scope demonstrated in the manuscript. **To avoid ambiguity, we have revised the text to clarify the method's intended scope and to avoid implying a general solution to rare-event sampling.**

> ***Remark 2.4:*** *Bulk crystal and liquid/glass states usually do not have large configurational space, which makes TrajCast predict the configurations which are not far from the existing distribution. This might not be the case when one simulates rare events in the condensed phase.*

We thank the reviewer for raising this point. Glassy water constitutes a non-trivial example which motivated our choice of this system for the present study.

Hyperquenched glassy water (HGW) is known to exhibit a highly complex potential energy landscape that spans an ample configurational space, with detectable structural differences. Different cooling rates can steer the system toward distinct regions of this landscape, resulting in configurations with different structural properties (see, for example, Giovambattista et al., *Phys. Rev. E* **71**, 061505 (2005); *Phys. Rev. E* **72**, 031510 (2005); Handle et al., *J. Chem. Phys.* **150**, 244506 (2019)).

In our numerical experiments, TrajCast is trained exclusively on configurations of liquid water at room temperature. Nevertheless, the glassy state produced by TrajCast after 12 ns, during which TrajCast is blind to inputs from classical molecular dynamics, is structurally equivalent to that obtained from conventional molecular dynamics simulations. This indicates that TrajCast effectively recovers a physically meaningful cooling protocol, despite having no explicit exposure to low-temperature or glassy configurations. We also note that the initial and final states are separated by 12 ns of simulation, of which only the last 2 ns are simulated in the glassy state. Given the well-documented sensitivity of glassy water structure to cooling rate, such agreement is unlikely to arise from overfitting or trivial structural convergence, and instead provides evidence that the model captures the essential physical mechanisms governing glass formation. Additionally, we would like to emphasise that in addition to the positions, TrajCast also relies on the atomic velocities which are distinctly different during the cooling process compared to training data.

We have added a discussion in the revised manuscript to clarify this point.

> ***Remark 2.5:*** *The computational framework is based on the NVE ensemble, which has been extrapolated to NVT by implementing a thermostat. While it works for now (short MD trajectories), obtaining a canonical distribution would still be challenging. In applications such as phase transitions, one needs to switch to the NPT, which would be a non-trivial extension.*

We thank Referee 2 for raising this important point and agree that extending TrajCast to the NPT ensemble is a natural and interesting direction for future work. While TrajCast itself propagates positions and velocities and does not explicitly access forces or stresses, this does not require a fundamental change to the framework. As in the NVT case, where a thermostat is applied only at inference time, an NPT simulation could be realized by coupling TrajCast to an external barostat, with instantaneous pressure obtained from the same classical or ML force field used to generate the training data. Although this introduces some additional overhead due to force evaluations, these would be required far less frequently than in standard MD, preserving the overall computational efficiency enabled by the large effective time step of TrajCast. Indeed, recent work (Bigi *et al.*, NeurIPS 2025) has shown that such a strategy can be practical providing reassurance that it could be implemented without modifying the core model. **We have modified the discussion on different ensembles in the Conclusion highlighting that a promising avenue for forecasting in NPT is the coupling with a force field of choice.**

We thank Referee 2 again for their reviewing our manuscript and providing this constructive feedback. We believe implementing these suggestions have improved the quality and clarity of our manuscript.