nature portfolio

# Peer Review File

## Breaking Timescales with Generative Sampling of Conformational Transitions

Corresponding Author: Dr Christophe Chipot

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

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

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)
The authors present a newly developed approach to generating transition pathways using a generative diffusion model combined with committor-based filtering. The workflow, called Gen-COMPAS, is tested on three proteins, Trp cage, ribose-binding protein, and the mitochondrial carrier AAC, demonstrating its ability to map the relevant free-energy landscape efficiently at a fraction of the computational cost compared to traditional approaches.

I do appreciate that the Gen-COMPAS is different from a lot of recent approach that rely on large training data sets. However, the novelty is limited as other groups have developed related approaches, going all the way back to 2019's Boltzmann generators. Furthermore, most of the workflow combines elements published by many of the same authors here. While there is certainly synergy in the combination here, it does mean that the conceptual advance is incremental rather than transformative.

While the workflow does not require pre-defined CVs, it does require at least two known end states. It also utilizes targeted MD between these initial states. Both of these effectively pre-determine the relevant part of conformational space to sample, which may or may not be of particular interest for a given system. Furthermore, it is well known that TMD can introduce significant biases that cannot be relaxed in the time scale of typical simulations. This may be a fatal flaw in the method for more complex transitions than the ones explored here.

The examples presented involve either very small systems (Trp cage, NANMA) or limited conformational changes (RBP, AAC). All have clearly defined A/B states and a relatively straightforward transition between them. Thus, they do not definitively demonstrate the generalizability of the workflow.

Additionally, the workflow lacks experimental validation. While the generated intermediates and free-energy maps are consistent with prior simulations, this does not prove that they are accurate.

For all the above reasons, while I find the workflow to be quite promising, I do not think it rises to the level of novelty and validation necessary for publication in Nature.

(Remarks on code availability)

Referee #2

(Remarks to the Author)
The authors of this manuscript introduce a computational framework designed to address the challenging problem of

sampling rare molecular events. Rare events are fundamentally important and characteristic of many molecular (and broader) phenomena. Despite significant advancements, the reliable characterization of these events remains an open challenge in both computational and experimental domains.

The proposed framework integrates molecular dynamics (MD) simulations, enhanced sampling techniques, generative AI models, and transition path theory to effectively sample these rare transitions. The authors initially introduce the framework and provide performance benchmarks on small molecular systems (detailed in the Supporting Information). They then showcase its utility on three larger, biologically relevant systems: the folding of the Trp-cage mini-protein, the binding upon folding of the ribose-binding protein, and a conformational transition of the mitochondrial ADP/ATP carrier embedded within a membrane environment. For each system, the framework successfully samples the rare transition between two alternative metastable states. The analysis includes the identification of structures belonging to the putative transition state ensemble (TSE), the determination of dominant transition paths, and the calculation of free energy surfaces projected onto appropriate collective variables (CVs).

The manuscript presents several intriguing ideas. I particularly appreciate the concept of accelerating the sampling of short, unbiased trajectories by utilizing machine learning tools to iteratively and judiciously initialize new trajectories.

However, it is important to note that many of the core concepts employed are not entirely new. The use of the committor to enhance sampling efficiency and the application of neural networks to learn the committor have been previously pioneered by researchers such as Peters and Troup, and Ma and Dinner, respectively. More recently, several publications have appeared that share key ingredients with the current approach.

* The combination of enhanced sampling to generate a configuration training set and the subsequent use of short trajectories to train a committor model has been prominently featured in the work of the Parrinello group, who have combined enhanced sampling with a variational principle for committor learning.
* Methods proposed by Hummer, Bolhuis, and Dellago have focused on learning the committor iteratively and on-the-fly from short unbiased trajectories.
* Rotskoff proposed a self-consistent method for learning the committor from short unbiased trajectories exploring outwards from a metastable state.
* The concept of merging generative AI with committor learning and path sampling has also been demonstrated in prior work, for instance, by Dellago and co-authors. The machine learning community is also actively developing new methods to utilize generative techniques for path sampling.

Overall, while the manuscript features an interesting and valuable integration of various established techniques and concepts, it may not represent a significant conceptual leap forward.

Regarding significance, this manuscript certainly constitutes a valuable addition to the growing body of sampling algorithms and should be of interest to the computational community. Nonetheless, I am not fully convinced that this new method offers the kind of groundbreaking or enabling novelty expected. The benchmark systems are interesting and challenging, but the presented results are not truly unique to this technique. None of the test cases appear to be beyond the reach of alternative, established methods. While the folding of a mini-protein is a standard benchmark, and the two larger, biologically significant systems demonstrate scalability, their conformational changes are relatively contained. These systems could likely be sampled using many existing methods. For acceptance in this journal, a method should ideally demonstrate a unique edge by enabling the study of a system or process that is demonstrably difficult or impossible to reproduce with alternative techniques.

The paper is undeniably of high technical quality. However, I have two main conceptual concerns:

1. Dependence on Targeted MD with RMSD

The authors correctly highlight the inherent difficulty of biased sampling along poor CVs, which can lead to inaccurate estimates of mechanisms, free energy, and rates. Therefore, any new algorithm that extracts useful information from short unbiased trajectories is welcome. However, the core of the sampling strategy relies on targeted MD using the root-mean-square deviation (RMSD) as a CV. I find this choice surprising. It is widely recognized that RMSD is often an inadequate CV for steering system dynamics, as targeted MD is generally effective only when the free energy landscape is not highly curved, typically only for transitions between structures that are already quite close. This dependence suggests that reaching structures close to the putative TSE that are significantly different from the starting state structures might be challenging.

Furthermore, the authors do not discuss the scenario where targeted MD fails to converge to reasonable structures near those proposed by the generative model. I also question the fundamental necessity of the targeted MD step. If the structures produced by the generative models are of sufficient quality, a brief energy minimization might be enough to render them physically viable. The manuscript would benefit significantly from a plot illustrating the structural distance between the generative model outputs and the structures actually converged to by the targeted MD. If the former are high-quality, the utility of the latter is unclear. Conversely, if the former are poor, the success of the entire sampling procedure becomes critically dependent on the quality of the targeted MD itself.

2. Lack of Committor Validation

While the entire method is founded on the principled use of the committor, the manuscript lacks any committor validation

plot. The authors claim accurate identification of the committor and the separatrix (the isosurface where the committor value is 1/2). Estimating the committor is a difficult task, and any claim of accuracy must be supported by a rigorous validation. This is standardly done by "shooting" trajectories from structures on the separatrix and counting the fraction that ends up in one metastable state versus the other. The resulting histogram of counts must be sharply peaked around the value 1/2. The absence of this validation makes it impossible to assess the accuracy of the learned committor.

Additional issues

* Generative AI for Membrane Proteins: As generative AI models for complete membrane-protein systems are not yet widely available, the authors should elaborate on the specific implementation used to generate viable structures for the membrane-protein system featured in the manuscript.
* Kinetic Observables: Lines 350-352 state that "Knowledge of the committor enables the reconstruction of key kinetic observables, including the transition-state ensemble (TSE), transition-path ensemble (TPE)." It should be noted that the TSE and TPE are not strictly kinetic *observables*. Furthermore, while the manuscript claims that the sampling scheme enables the retrieval of kinetic observables, the authors never estimate any kinetic rates and compare them to benchmark values, which would substantially strengthen this claim.
* Lag Time Dependence: Equations 2 and 4 show that the variational principle and the committor estimate explicitly depend on the choice of a lag time. The optimization process and how the accuracy of the estimate depends on this crucial parameter should be discussed and analyzed.
* Molecular Representation: While the Z-matrix representation is reasonable and was standard until recently, it is becoming obsolete, particularly as it focuses only on the protein and is less suitable for materials or accurately capturing the environment of a complex molecular system. Given that the authors have also proposed GNN schemes for committor learning, it would be beneficial to explain the choice not to use them here.
* Multiple States: The committor is formally defined for a transition between two states. The authors should clarify their practical approach for handling systems with multiple metastable states.
* Convergence Test: The convergence test detailed in the Supporting Information, which assumes accuracy if repeated sampling iterations fail to add new structures, is theoretically questionable. While operationally acceptable, there is no formal guarantee that this lack of new structures ensures committor accuracy. The only reliable and straightforward way to confirm accuracy is through the committor validation tests mentioned previously.

A significant weakness is the complete absence of appropriate uncertainty quantification. There are no error bars or confidence intervals reported for any of the estimated quantities. For instance, demonstrating the error in estimating the free energy difference between metastable states as a function of cumulative sampling would be highly valuable.

In conclusion, the paper is well-written and holds interest for the rare event sampling community. However, it does not currently demonstrate sufficient novelty and significance to recommend publication in a top-tier journal.

From a technical standpoint, the lack of any committor validation and the central reliance on targeted MD along the RMSD, which is known to be an inadequate CV, are major points that must be addressed and clarified for publication in any reputable journal.

Unfortunately, a clear, demonstrable conceptual breakthrough is absent, and the manuscript does not convincingly show that the proposed method uniquely enables the sampling of systems otherwise inaccessible by standard, alternative techniques.

(Remarks on code availability)
I did not review the code.

Referee #3

(Remarks to the Author)
The work addresses an important challenge in molecular dynamics simulation of biomolecules by efficiently constructing transition paths between defined metastable states. The main innovation is the use of a denoising diffusion probabilistic model (DDPM, “diffusion model”) to navigate the configuration space intermediate between starting and end states A and B. A more technical aspect is the use of a variational committor network to estimate the commitment probability (“committor”) to B, also from trajectories that do not quite reach A or B. Finally, the internal representation of configurations in terms of Z-matrices is quite general, avoiding the need to pre-define collective variables, even if solvent water and ions appear not to be included explicitly. The work includes applications of increasing complexity, from the folding of the mini-protein “Trp-cage” over the binding of ribose to the ribose-binding protein RBP to the access transition of the membrane-bound ATP/ADP carrier protein AAC. Overall, I find this to be a very promising direction with quite impressive results. However, a number of major points are unclear and should be addressed in a revision, as listed in the following.

(1) It was not quite clear to me what the impact of the diffusion model in the whole procedure was. As I understand, it was used only to create perturbed initial configurations for further sampling (Fig. 1A,B). Wouldn’t it have been easier to simply start new simulations from configurations visited in the short runs, chosen according to their committor value as predicted by the variational committor network?
(2) I could not find a demonstration for the different applications (including the small peptides in SI) that the learned committor has predictive power. There are various ways of testing the committor, the simplest one being repeated initializations from a given starting configuration (the so-called “committor test”). In such a test (as sketched in Fig. 1D), do

individual configurations taken on the isocommittor surface individually have 50 % chance of ending up in A and B? I.e., is Gen-COMPAS able to reliably identify transition states? And how do the committor predictions perform for configurations away from the isocommittor surface, e.g., at the 30% and 70% isosurface?
(3) It was not clear to me if/how simulation-quality intermediate configurations were “generated” by the diffusion model (i.e., with proper stereochemistry and clash statistics) and how solvent was treated in their construction. As I understand, starting structures for the MD runs are not actually created by the diffusion model. Instead, the structures of the diffusion model serve as targets for “targeted MD” (TMD) simulations. I would imagine that this creates some uncertainty and dependence on details (hyperparameters) of the TMD simulations. For instance, if TMD is repeated with the same target but from different starting configurations, is the committor value of the respective endpoints the same (and as predicted) or dependent on the starting structure? I.e., how consistent is the TMD-based construction of intermediates?
(4) The quality of the free energy surfaces is quite impressive for Trp-cage (Fig. 2C), but it is not really assessed against a reference for the more challenging cases of RBP and AAC. Are these surfaces rough estimates? Can we trust the relative populations of A and B? I am particularly concerned about the free energy estimate, because it relies on some rather strong assumptions (eq S10-S11) to get a trajectory weight factor. Where do these assumptions break down? How would one know?
(5) Related to the preceding points, how reliably are commitors and free energies estimated in repeated, fully independent Gen-COMPAS runs? In such repeated Gen-COMPAS runs, how much do the results depend on hyperparameter choices (such as the length of TMD)?
(6) The statistics of the sampling is not quite clear. In particular, for step 7 in Figure 1A (“short simulations from transition states”), how many of those runs from the commitor=1/2 surface were done for each of the systems studied and where did they end up for given initial configurations? Also, how much computer time goes into which part of the Gen-COMPAS work flow?
(7) A qualitative and quantitative comparison to related established methods pursuing the same goal with similar procedures would be important to assess the advance of this new method. As points of reference, one should consider standard transition path sampling (as pioneered by Dellago, Bolhuis, Chandler et al.), AIMMD (ref 7), and the string methods of E, Vanden Eijnden et al. (refs 22-24) and by Pan, Roux et al. (ref 25). Of these, AIMMD in particular uses a quite similar work flow of path sampling and committor learning and appears to achieve high sampling efficiency in terms of finding transition paths and transition states (e.g., Horvath et al., DOI 10.1073/pnas.2506516122).

Minor:
(8) As a minor historical note, transition path sampling with targeted initialization has been used to sample complete transition paths between the inward and outward access states of a membrane transporter (Okazaki et al., DOI 10.1038/s41467-019-09739-0) similar to AAC here.
(9) As a minor technical point, do the Z-matrices include all atoms of the protein or only a subset (e.g., backbone)? How does this impact the TMD calculations? I can see advantages (e.g., by getting side chain rotamer resolution) and disadvantages (e.g., by TMD becoming trapped or driven into high-energy structures).

(Remarks on code availability)

Version 2:

Reviewer comments:

Referee #1

(Remarks to the Author)
I am impressed by the authors' thorough response to the reviewers issues with the previous version of the manuscript. In particular, the newly added applications go a long ways towards demonstrating generalizability of the approach.

I still want to bring up the question of novelty. I appreciate the authors' point that while the methods employed already exist, it's their combination that is novel. However, I take some issue with this statement in the response:
"This shift is, in our view, substantive. Existing approaches typically rely either on massive data generation through long-timescale simulations or on carefully engineered collective variables and prior mechanistic assumptions."

and the corresponding statement in the manuscript:
"Accordingly, a framework that can autonomously construct thermodynamically and kinetically consistent transition landscapes from minimal prior information remains lacking notwithstanding substantial progress at the interface of machine learning (ML) and molecular simulation."

This field is moving extremely quickly, and what may have been true even a year ago is no longer. There are other groups out there developing "physics-informed" and/or "MD-guided" approaches as well. I think the authors could do a little more to look into this and cite a few examples, which would also help to validate the authors' choice to go in this direction.

Finally, I would like for the authors to add a section (probably to the SI and to the README on the github) on (hyper)parameter tuning. At many stages, parameters have to be chosen that will affect the success of the methodology. These range from the TMD force constant to the VCN lag time to the diffusion model dimensions, and more. How sensitive is the method to these? What recommendations can you offer for choosing them?

A few typos I noticed:
Page 1: "LLaboratoire" - Laboratoire
Page 2: "trasition path theory" - transition
Page 4: "Gen-COPMAS" - Gen-COMPAS
Page 11: "probablistic" - probabilistic
Page 15: "consequentially sampled" - subsequently?
Page 16: "pore activating" - activation

(Remarks on code availability)
I briefly looked over the code. I believe it does include sufficient instructions for others to run it, although I did not try myself. As I noted in my review, it would be helpful to have more guidance on parameter tuning.

Referee #2

(Remarks to the Author)
I acknowledge that the authors have undertaken a substantial amount of work for these revisions, and I appreciate their efforts. However, several crucial issues remain unresolved, and in my view these must be carefully addressed before I could recommend this manuscript for publication in any journal, let alone Nature. I outline these concerns below.

# Quality of the committor models
The committor validation presented in Figures S7 and S8 is notably noisy, indicating sub-par committor learning compared to recent studies from other groups. For instance, Bolhuis and coworkers (arXiv:2507.04052) achieve substantially better predicted-vs-sampled correlation for a host–guest system that is arguably more challenging than vacuum NANMA isomerization (see their Fig. 17). The lower panel of Figure S8 shows that the validated committor values for configurations predicted at $q \approx 0.5$ are distributed broadly between 0 and 1, indicating that the model is nearly unpredictive in the transition region.
Furthermore, the contrast with Figure S9 is striking: that plot looks considerably cleaner. A well-trained committor model should yield comparably good agreement between learned and sampled committors regardless of whether the source structures originate from an enhanced sampling campaign or an unbiased trajectory. The fact that this is clearly not the case admits at least two interpretations:
1. Sampling problem. The TMD pipeline may produce structures that lie far from physical unbiased transition paths in phase space, leading to noisier correlations. TMD is known to often result in incomplete transitions along slow orthogonal degrees of freedom (hysteresis). Although the details of how the committor validation set is constructed are not entirely clear to me, if it includes TMD-generated structures initiated from both end-states, the resulting frames would be biased toward the basin neighborhoods and potentially occupy regions far from unbiased transition paths. This interpretation is consistent with the clearly bimodal distribution of sampled committors at $q = 0.5$ in Figure S7, rather than the expected unimodal distribution centered near 0.5. If correct, this would represent a fundamental limitation of the diffusion-model-guided TMD workflow—one that can be expected to worsen for more complex systems. I suggest a concrete test for this below.
2. Learning problem. The committor model may not have learned sufficiently general representations across the relevant regions of phase space.
That such difficulties already arise for one of the simplest systems studied raises serious concerns about the quality of the committor models for the more complex systems.
As an aside, I respectfully disagree with the authors' assertion that committor validation on the more complex systems would not have been computationally feasible:
"We believe that the committor validation test is not computationally addressable for most of the systems"
The barrier crossing time of Trp-cage, which represents an intermediate level of complexity among the systems studied, is approximately 3 ns (Qiu et al., PNAS 2006). A reasonable committor validation—say, 100 points with 100 shooting trajectories each—would require on the order of 30 μs of aggregate sampling. On a modern GPU node, this amounts to only a few days of wall-clock time, which should be well within reach for a manuscript submitted to a journal of this caliber.

#Possible hysteresis due to TMD simulations
As discussed above, there are substantive concerns regarding hysteresis introduced by TMD. The analysis presented in Section 2.6.2 of the supplement, which is intended to demonstrate that steering along RMSD does not introduce artifacts, is not fully convincing. It shows only that bidirectional steered trajectories approach configurations with similar predicted committor values. However, the poor predicted-vs-sampled validation in Figures S7 and S8 suggests that the predicted committor alone is not a sufficiently sensitive metric for detecting such hysteresis. An alternative interpretation, considering Figures S7, S8, S9, and S11 together, is that RMSD-based TMD does exhibit a hysteresis problem, but that the learned committor model fails to capture it. This would explain why the discrepancy appears only when comparing to sampled committors, and why the predicted-vs-predicted validation along an unbiased path looks considerably better.
A straightforward test would be to repeat the analysis of Figure S11 using sampled committors rather than predicted ones. If this substantially degrades the agreement, it would indicate that RMSD-based TMD is not suitable for this workflow. If the agreement remains as clean as in the current Figure S11, then TMD may indeed be appropriate, but the authors would need to investigate—including on Trp-cage—other explanations for the divergence in committor validation quality between Gen-COMPAS-generated structures and unbiased paths.

#Poor convergence of free energy profiles
Figure S6 presents the RMSE of the free energy landscape as a function of simulation effort across several systems and simulation campaigns. I thank the authors for including this analysis. However, it is concerning that, with the exception of

trialanine—a system that is trivial to converge—none of the other systems exhibits a converged free energy surface. I also wish to emphasize that this test only probes statistical convergence, and says nothing about the accuracy of the resulting landscape. I appreciate that some of these systems are inherently difficult to simulate, but one should be transparent about this limitation rather than presenting these profiles as converged free energy surfaces. Strictly speaking, a free energy is only well-defined when the underlying statistical distribution has been sampled to convergence, which does not appear to be the case here.

#Novelty
Finally, regarding the question of novelty. In their rebuttal, the authors argue that:
"The central contribution of this work is not the introduction of isolated methodological components, but the demonstration of a unified, end-to-end framework…"
"This shift is, in our view, substantive…a new route to mechanistic insight in regimes that would otherwise be computationally inaccessible."
I agree that the combination of existing elements into a unified workflow carries some novelty. However, integrative committor learning and sampling pipelines already exist, notably from the Parrinello and Hummer groups. These approaches, with various differences between them, can provide similar mechanistic insight to Gen-COMPAS, including transition mechanisms, committor surfaces, free energy landscapes, and kinetics. Importantly, these existing approaches are not dramatically less efficient computationally. While no direct comparison is available, the Parrinello group's 2024 study (Ray et al., Nat. Comput. Sci. 2024) investigated chignolin, which folds somewhat faster than Trp-cage but represents a comparable level of molecular complexity. Their protocol required approximately 800 ns of total sampling across 10 iterations, which is similar to the 594 ns the authors report for Trp-cage. Although the systems and methods differ, this comparison suggests that sub-microsecond committor-coupled sampling of transition mechanisms and free energy landscapes for fast-folding proteins is already achievable with current state-of-the-art methods.
The additional integration of steering simulations toward the output of a diffusion model is distinct from these existing workflows and represents a genuine point of novelty—even though the diffusion models themselves are not new. However, this alone does not, in my assessment, constitute sufficient novelty for publication in a journal of this standing.

In summary, while the scope and ambition of this work are commendable, the issues outlined above—concerning committor model quality, potential TMD-induced hysteresis, free energy convergence, and the degree of novelty over existing methods—lead me to recommend that this manuscript not be accepted in its current form.

(Remarks on code availability)

Referee #3

(Remarks to the Author)
As stated in my previous report, the work addresses an important challenge in molecular dynamics simulation of biomolecules: to extend the sampling to the transition region between metastable states. Key innovations are the use of a denoising diffusion probabilistic model to propose configurations, of a committor model, and of z-matrices as representation of internal coordinates. In the revision, the authors have substantially improved the manuscript and added further interesting and challenging applications to Vo-ring rotation in a V-type ATPase and acetylcholine receptor activation. They have also made a serious effort to address my earlier concerns, and those of the other reviewers. As stated before, I find this to be a very promising direction with quite impressive results. However, in my view a number of points deserve further attention. In the following I will first briefly review key responses to my earlier questions, keeping the old numbering. I will then ask for some clarifications for the new results.
Review of response to earlier questions:
(1) Does the diffusion model adding value?
The response and revision make the purpose of the diffusion-based generative model clearer and also its use: to generate reference structures for targeted MD, thereby avoid inaccuracies in the diffusion model while still harnessing some of its power. While this may not be the most elegant approach, it appears to a practicable solution to a major problem: how to create meaningful structures away from sampled structures. There is now also an effort to quantify the effects of this mixed procedure, in response also to reviewer #2. While I am still not entirely convinced that much value is added by using a diffusion model, I am overall satisfied with the effort and response.
However, as one minor point,
(2) Is the learned committor model predictive?
The authors now provide tests of their committor model for alanine dipeptide (NANMA) the model most easily sampled. The new results in SI Figs. S7 and S8 are decent, even if some concerns remain. In particular, it appears that the transition region is relatively poorly captured. The bimodal distribution of the “true” committor of presumed transition states in SI Figs. S7 center and S8 right indicates that the actual conformations cluster on either side of the barrier but not quite on top. The results in SI Fig. S9 look more reassuring. As I understand, they differ in the way test structures were chosen? In any case, this is a challenging problem, even for alanine dipeptide, and testing for the more challenging systems would in my view be important (e.g., for Trp-cage), also as a basis for comparison with other methods.
(3) How are diffusion model predictions turned into MD-ready configurations?
This question has been clarified.
(4) Can we trust the estimated free energy surfaces?
The authors now use a different method (RiteWeight). This is a rather new procedure, though at its core it looks to be closely

related to traditional Markov state modeling. Here it might be useful to discuss the potential shortcomings a bit more clearly in the text. As indicated in my earlier question, I can see that one can get “local” free energy maps quite right, but to get global maps (e.g., the A-to-B equilibrium) right is challenging because disparate and poorly connected regions in phase space need to be connected somehow.
(7) How does Gen-COMPAS compare to other methods?
I understand the hesitancy to compare directly to established methods and I appreciate that the discussion has been expanded and includes additional references. As a minor point, in ref 56 (Okazaki et al.) complete transition paths of a membrane transporter were not “inferred” but actually sampled in transition path MD simulations, as I understand.
Following are questions to the new results:
A. The added results for the acetylcholine receptor look promising. Here, it was not quite clear if the ligand acetylcholine exchanged spontaneously or if exchange was somehow driven. As I understand from the protocol, no excess ligand was present in the solution, but ligand position/s were included in the CVs.
a. Was excess ligand present in the solution?
b. If not, how was binding and unbinding modeled (by distances to all or a subset of ligands in the CVs)?
c. Was acetylcholine binding and unbinding reversible?
d. Did the mono-liganded form of the receptor form both from the di-liganded form and the unliganded form by release and binding of ligand? (The text suggests the former.)
e. Are the free energies in Fig. 4F at standard concentration of ligand (1M), as the conformational transition is coupled to binding?
B. The results for Vo ring rotation are also encouraging. Here, I wonder how the calculated free energies (Fig. S4B) compare to the results from “conventional” sampling in Roh et al. (SI Ref. 15), which served as starting point here, and to the results for F-type ATP synthase c-ring rotation (Blanc PNAS 2024; DOI: 10.1073/pnas.2314199121).

(Remarks on code availability)

Version 3:

Reviewer comments:

Referee #2

(Remarks to the Author)
We sincerely thank the authors for addressing our concerns, primarily on committor quality. The new committor validation plots are of satisfying quality and reassuring on the correctness of the sampling.

However, we sadly cannot recommend publication of this manuscript in Nature.

From a methodological point of view, the manuscript describes a pipeline that chains together existing components and applies it to standard benchmarks and more challenging systems. The method itself is not transformative or radically novel.

The paper contains ambitious applications of the method to complex biological systems. Still, they are assessed only for qualitative agreement with existing results, with which they broadly agree but quantitatively diverge (with biological consequences). This is true progress, given that their pipeline can handle complex systems. But these applications remain qualitative/confirmatory (where they do agree) and tentative/insufficiently explored regarding their biological implications (where they do not).

The progress is incremental, and – in good part – self-referential. Committor learned in full/internal coordinates "without CVs" is already in the authors' own Arredondo et al. (Nat Comput Sci 2026); the committor-consistent string and VCN are Megías et al. (Nat Comput Sci 2025) and Chen et al. (JCTC 2026). Generative models for related problems (such as coupling with TPS) are already established (MDGen, NeurIPS 2024; Raja et al., OM-TPS, ICML 2025; etc.). BioEmu (Lewis et al., Science 2025) and Boltzmann generators (Noé et al., Science 2019) already deliver one-shot equilibrium ensembles and ~1 kcal/mol free energies.

Putting these workflows, including those from the author's own group, together to push beyond benchmarks and produce results that are broadly qualitatively consistent with prior work is an engineering achievement, but not of the kind that would warrant publication in flagship Nature without a substantial discovery arising from it. That would not be consistent with the previous publishing practices in this field.

A last remark. In this iteration, we inspected and ran the code to reproduce some of the results presented in the manuscript. We managed to reproduce the committor convergence on NANMA, but we could not get the RMSD-based TMD fix to work properly in the generator output. Regrettably, the code doesn't actually give a harness for a complete run, only scripts for individual steps. We had to use Claude AI to make this harness. We would be happy to see this manuscript published in Nature Computational Science, provided that the authors submit rewritten code as a package with an end-to-end reproduction script that runs and produces reliable results.

(Remarks on code availability)
We would like to see the code as a comprehensive package, or at least as harnessed scripts that could be run end-to-end to reproduce some of the results in the manuscript.

Referee #3

(Remarks to the Author)
The authors have made a serious effort to address the reviewer concerns. I am particularly impressed by the thorough analysis of the learned committor model for NANMA and Trp-cage. The results are convincing. The usefulness of the diffusion-based generative model and its integration into the workflow are now much clearer and more compelling. I also appreciate the more nuanced discussion of the results and the expanded discussion of related results. My technical questions have been satisfactorily addressed. As previously stated, I find this to be a promising methodological direction with impressive results on diverse, challenging, and biologically relevant systems. I therefore recommend publication in Nature.

A very minor issue remains: On lines 183-184, a 100x speedup in sampling over conventional simulation is claimed. However, such a claim seems overly bold, as it is not quantified in terms of the asymptotic decay of the statistical error. 10 us of regular MD might have yielded comparable results, which would not diminish the value of the present work.

Additionally, there appears to be a minor typo on line 256: 'bound' should be 'bound to'."

(Remarks on code availability)

**Open Access** This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.
In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.
The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.
To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/

# Response to the comments of the Referees

Major additions and validations in the revised manuscript

**New biological applications.** We expanded the scope of the manuscript by incorporating two complex membrane protein systems: (i) the rotary transition of the V-ATPase $V_o$ domain and (ii) the activation of the muscle-type nicotinic acetylcholine receptor (nAChR). For the $V_o$ rotor, Gen-COMPAS reconstructs the complete rotary pathway and identifies intermediate metastable states that are consistent with previous experimental and computational studies. For nAChR, Gen-COMPAS uncovers an asymmetric activation pathway involving a mono-liganded intermediate, state in which one acetylcholine molecule binds first at a specific intersubunit interface, triggering a pre-active pore configuration before the second ligand binds. This sequential activation mechanism is in close agreement with recent experimental observations.

**Free-energy estimation and statistical uncertainty analysis.** We incorporated the recent RiteWeight framework to reweight ensembles of short unbiased trajectories and reconstruct free-energy landscapes without requiring extensive equilibrium sampling. To ensure reliability, we performed systematic convergence analyses by progressively increasing the number of trajectories, and estimated uncertainties using bootstrap resampling of the resulting free-energy landscapes. In addition, the uncertainty in committor predictions was quantified through ensemble estimates obtained from multiple VCN models.

**Validation against unbiased shooting.** To validate the Gen-COMPAS sampling strategy, we performed conventional unbiased shooting simulations for the NANMA system and compared the resulting transition statistics to those obtained from Gen-COMPAS-generated trajectories. The agreement confirms that the method recovers physically consistent transition behavior while substantially improving sampling efficiency.

**Systematic validation of TMD parameters.** We conducted a systematic parameter study for the targeted MD (TMD) stage using the Trp-cage system. A grid of four TMD durations and five force constants was explored during the initial iteration, and the distribution of RMSD relative to the target structures was analyzed for the first and last two percent of each trajectory. Comparable diagnostics were also generated for all other systems to ensure that the chosen parameters promote stable convergence toward physically plausible intermediate conformations.

**Reproducibility tests.** To assess robustness, we conducted five independent Gen-COMPAS runs for the Trp-cage system, i.e., three using the same parameters reported in the manuscript, and two using modified TMD force constants to evaluate the sensitivity of the method to parameter choices. These experiments show that Gen-COMPAS reproducibly reconstructs the thermodynamic and kinetic observables when reasonable parameter values are employed.

**Justification of the TMD refinement step.** To evaluate the necessity of the TMD refinement stage, we computed the Kullback–Leibler divergence (KLD) between transition-state ensembles (TSEs) directly generated by the diffusion model and those obtained after TMD refinement followed by equilibration. The reference distribution was constructed from DESRES trajectories within $\Delta q = 0.1$ of the separatrix, as determined by a VCN model trained on the DESRES dataset. The analysis shows that TMD refinement markedly improves the agreement with the reference distribution, thereby demonstrating the importance of this step prior to initiating unbiased shooting trajectories.

We thank the Referees for the time and effort devoted to evaluating our manuscript. We appreciate the range of perspectives offered in the reports, raising specific questions regarding novelty relative to existing approaches, the role and validity of the targeted MD (TMD) refinement step, the reliability of committor and free-energy estimates, and the generality of the framework beyond standard benchmark systems. In response, we have added new biological applications involving large membrane-protein assemblies, explicit committor-validation analyses, uncertainty quantification for free-energy estimates, reproducibility tests, and a systematic assessment of the TMD protocol. We have also clarified the scope and limitations of the method.

More broadly, we emphasize that the framework provides a direct route to reconstruct dynamical and mechanistic information from a small number of known states. Based solely on endpoint structures, it enables the identification of transition pathways, intermediate states or new metastable states, and free-energy landscapes, without relying on large training datasets or specialized HPC resources. This capability addresses a practical limitation in current computational approaches, where mechanistic insight is often inaccessible without either extensive sampling or strong prior assumptions.

By removing this dependence, the approach shifts the practical paradigm toward structure-to-mechanism inference, in which dynamical information can be systematically reconstructed from limited structural inputs. In this sense, the method provides a general and accessible route to mechanistic understanding, changing the way complex molecular processes are investigated computationally.

We respond below to each Referee in a point-by-point manner and indicate the corresponding modifications made in the revised manuscript and Supplementary Information.

## Response to Referee 1

The authors present a newly developed approach to generating transition pathways using a generative diffusion model combined with committor-based filtering. The workflow, called Gen-COMPAS, is tested on three proteins, Trp cage, ribose-binding protein, and the mitochondrial carrier AAC, demonstrating its ability to map the relevant free-energy landscape efficiently at a fraction of the computational cost compared to traditional approaches.

**Response**: We thank the Referee for their careful reading of the manuscript and for recognizing both the distinct character of the workflow relative to approaches that rely on large training datasets and its potential promise. We also appreciate the opportunity to clarify several points concerning novelty, scope, and validation.

I do appreciate that the Gen-COMPAS is different from a lot of recent approach that rely on large training data sets. However, the novelty is limited as other groups have developed related approaches, going all the way back to 2019's Boltzmann generators. Furthermore, most of the workflow combines elements published by many of the same authors here. While there is certainly synergy in the combination here, it does mean that the conceptual advance is incremental rather than transformative.

**Response**: The Referee raises the important question of how novelty should be assessed for a methodology that builds on existing ideas, including committor theory, generative modeling, and path sampling. We agree that individual elements of the workflow have precedents in the literature, including prior work from some of the present authors. However, we respectfully disagree with the characterization of the advance as incremental.

The central contribution of this work is not the introduction of isolated methodological components, but the demonstration of a unified, end-to-end framework that fundamentally changes how rare-event dynamics can be accessed in practice. In particular, Gen-COMPAS enables the reconstruction

of transition pathways, metastable or intermediate states, committor structure, and free-energy landscapes directly from a small number of endpoint configurations, without requiring large training datasets or extensive equilibrium sampling.

This shift is, in our view, substantive. Existing approaches typically rely either on massive data generation through long-timescale simulations or on carefully engineered collective variables and prior mechanistic assumptions. In contrast, the present framework removes these requirements by coupling generative modeling with committor-consistent sampling in a self-consistent manner. The resulting capability is not simply a more convenient implementation of existing ideas, but a new route to mechanistic insight in regimes that would otherwise be computationally inaccessible.

To our knowledge, no prior approach has been shown to achieve this combination of data efficiency, scalability, and quantitative accuracy for biomolecular transitions of comparable size and complexity. In this sense, the advance lies in enabling a qualitatively different mode of simulation, where dynamical information can be inferred from limited structural input rather than from exhaustive sampling.

To clarify this point, we have revised the Abstract and Introduction to more explicitly distinguish the present contribution from related earlier developments and to emphasize the capabilities enabled by the full workflow.

While the workflow does not require pre-defined CVs, it does require at least two known end states. It also utilizes targeted MD between these initial states. Both of these effectively pre-determine the relevant part of conformational space to sample, which may or may not be of particular interest for a given system. Furthermore, it is well known that TMD can introduce significant biases that cannot be relaxed in the time scale of typical simulations. This may be a fatal flaw in the method for more complex transitions than the ones explored here.

**Response**: The Referee notes that the workflow requires two known end states and employs TMD, and suggests that these elements may predefine the relevant region of conformational space. We respectfully view this differently. The specification of two metastable states is a standard assumption in committor theory and transition path sampling, and is not specific to the present method. Importantly, Gen-COMPAS does not prescribe intermediate states, transition mechanisms, or reaction coordinates. These quantities emerge from the learned committor and the iterative sampling procedure.

The role of targeted MD (TMD) in the present framework is also different from that implied in the comment. In Gen-COMPAS, TMD does not serve as a mechanism for sampling or for inferring thermodynamic or kinetic quantities. Rather, it acts as a physically grounded embedding step, in which configurations proposed by the diffusion model are relaxed and embedded into the full atomistic environment through Langevin dynamics. This step ensures that generated structures are dynamically accessible and consistent with the underlying Hamiltonian, while preserving the exploration capability of the generative model.

In this sense, TMD complements the diffusion model by correcting for limitations associated with finite training data and model uncertainty, rather than introducing systematic bias into the transition mechanism. All physically meaningful observables—such as committor values, transition pathways, and free-energy landscapes—are subsequently inferred from unbiased trajectories initiated from these embedded configurations. If the TMD stage were to introduce persistent bias, this would manifest as discrepancies in these observables when compared to reference simulations, which is not observed for the systems considered.

To address this point explicitly, we have performed additional analyses, including parameter sensitivity tests, convergence diagnostics, reproducibility studies, and quantitative comparisons of transition-

state ensembles before and after TMD refinement. These results are presented in pages R10–R12 of this response and in the new Supplementary Information section entitled "TMD validation and reproducibility."

Overall, we have revised the manuscript to clarify the role of the TMD step as a physics-based embedding layer within the workflow, and to document its effect through systematic validation. These analyses support the conclusion that TMD enables physically consistent initialization of trajectories without determining the mechanistic outcomes of the method.

The examples presented involve either very small systems (Trp cage, NANMA) or limited conformational changes (RBP, AAC). All have clearly defined A/B states and a relatively straightforward transition between them. Thus, they do not definitively demonstrate the generalizability of the workflow.

**Response**: We appreciate that the Referee raises a question about whether the systems studied are sufficiently challenging to support claims of generality. However, we respectfully disagree with the Referee's note of "relatively straightforward transition", particularly for the mitochondrial ADP/ATP carrier embedded in a membrane environment. The transition is coupled by conformational transition of the membrane protein and the movement of the ADP substrate, which is difficult to capture using conventional computational methods. Nevertheless, we agree that generality is best assessed through application to systems of increasing complexity. In the revised manuscript, we expand the scope of the study by including two additional membrane-protein examples: the rotary transition of the $V_o$ domain of a V-ATPase and the activation of the muscle-type nicotinic acetylcholine receptor (nAChR).

Both systems represent challenging cases of large-scale, highly cooperative conformational transitions in membrane proteins involving complex allosteric coupling and multiple metastable intermediates. The $V_o$ rotor transition requires coordinated rearrangements of a large membrane-embedded assembly mediated by sequential formation and rupture of interfacial salt bridges, while nAChR activation involves asynchronous subunit transitions coupled to ligand binding and pore activation in a pentameric ion channel. These processes occur on rugged free-energy landscapes and involve high-dimensional structural rearrangements that are difficult to capture with conventional enhanced-sampling approaches relying on predefined collective variables. Importantly, Gen-COMPAS reconstructs the transition pathways and identifies intermediate states consistent with experimental observations in both systems without imposing explicit reaction coordinates or symmetry assumptions. Together, these results demonstrate that the proposed framework can handle highly complex biomolecular transitions involving multiple coupled degrees of freedom within large structural assemblies, thereby illustrating its practical advantages beyond standard benchmark systems. (see Figs. R1 and R2)

It is especially worth noticing that, in the case of nAChR, the framework recovers the experimentally suggested mono-liganded/pre-active intermediate while starting only from the **two end-point structures** and **without introducing prior mechanistic information** into the model, achieving this **at a very moderate computational cost.**

In the revised manuscript, we have added a new subsection in the Results section: "Gen-COMPAS reveals an asymmetric activation mechanism of nAChR via a mono-liganded intermediate." The section starts:

The muscle-type nicotinic acetylcholine receptor (nAChR) is a prototypical pentameric ligand-gated ion channel (pLGIC) (. . . )

In addition, in the SI, we have added a new subsection under the section "Further Examples and Extended Data" that begins:

Vacuolar-type ATPases (V-ATPases) are rotary proton pumps that couple ATP hydrolysis in the soluble $V_1$ motor to proton translocation through the membrane-embedded $V_o$ domain. (. . . )

Additionally, the workflow lacks experimental validation. While the generated intermediates and free-energy maps are consistent with prior simulations, this does not prove that they are accurate.

**Response**: The Referee raises the point that the workflow is not validated directly against experiment. We agree that this is an important issue to address carefully. At the same time, for transition paths and committor functions, direct experimental observables are generally limited or unavailable. In this area, validation is therefore typically based on comparison with high-quality reference simulations and with experimentally supported mechanistic interpretations.

In the revised manuscript, we have strengthened the aspect of the experimental validation. In particular, for nAChR, Gen-COMPAS identifies an asymmetric activation pathway involving a mono-liganded intermediate state, in which one acetylcholine molecule binds first at a specific intersubunit interface, leading to a pre-active pore configuration before binding of the second ligand. This mechanism, as well as the associated free-energy differences between states, is in close agreement with recent experimental observations (see Fig. R2).

For all the above reasons, while I find the workflow to be quite promising, I do not think it rises to the level of novelty and validation necessary for publication in Nature.

**Response**: We thank the Referee for recognizing the promise of the workflow, while expressing concerns regarding novelty and validation. In the revised version, we have sought to address these concerns directly by adding new biological applications, committor-validation analyses, free-energy uncertainty estimates, reproducibility tests, and explicit assessment of the TMD refinement step. These additions clarify the scope of the contribution and provide a more complete validation of the approach.

## Response to Referee 2

The authors of this manuscript introduce a computational framework designed to address the challenging problem of sampling rare molecular events. Rare events are fundamentally important and characteristic of many molecular (and broader) phenomena. Despite significant advancements, the reliable characterization of these events remains an open challenge in both computational and experimental domains.

The proposed framework integrates molecular dynamics (MD) simulations, enhanced sampling techniques, generative AI models, and transition path theory to effectively sample these rare transitions. The authors initially introduce the framework and provide performance benchmarks on small molecular systems (detailed in the Supporting Information). They then showcase its utility on three larger, biologically relevant systems: the folding of the Trp-cage mini-protein, the binding upon folding of the ribose-binding protein, and a conformational transition of the mitochondrial ADP/ATP carrier embedded within a membrane environment. For each system, the framework successfully samples the rare transition between two alternative metastable states. The analysis includes the identification of structures belonging to the putative transition state ensemble (TSE), the determination of dominant transition paths, and the calculation of free energy surfaces projected onto appropriate collective variables (CVs).

The manuscript presents several intriguing ideas. I particularly appreciate the concept of accelerating the sampling of short, unbiased trajectories by utilizing machine learning tools to iteratively and judiciously initialize new trajectories.

**Response**: We thank the Referee for their careful reading of the manuscript and for recognizing both the technical quality of the work and the importance of the rare-event sampling problem addressed here. We also appreciate the positive assessment of the central idea of iteratively initializing short unbiased trajectories using machine-learning-guided proposals. Below, we respond to the concerns regarding novelty, significance, validation, and methodological choices.

However, it is important to note that many of the core concepts employed are not entirely new. The use of the committor to enhance sampling efficiency and the application of neural networks to learn the committor have been previously pioneered by researchers such as Peters and Troup, and Ma and Dinner, respectively. More recently, several publications have appeared that share key ingredients with the current approach.

* The combination of enhanced sampling to generate a configuration training set and the subsequent use of short trajectories to train a committor model has been prominently featured in the work of the Parrinello group, who have combined enhanced sampling with a variational principle for committor learning. * Methods proposed by Hummer, Bolhuis, and Dellago have focused on learning the committor iteratively and on-the-fly from short unbiased trajectories. * Rotskoff proposed a self-consistent method for learning the committor from short unbiased trajectories exploring outwards from a metastable state. * The concept of merging generative AI with committor learning and path sampling has also been demonstrated in prior work, for instance, by Dellago and co-authors. The machine learning community is also actively developing new methods to utilize generative techniques for path sampling.

**Response**: The Referee correctly notes that several of the conceptual ingredients used in the present work have important precedents in the literature, including prior work on committor learning and iterative path-sampling strategies. We fully acknowledge this broader context and have revised the Abstract and the Introduction to better position the manuscript relative to these earlier contributions.

Our central claim is not that the underlying concepts are individually new, but that their integra-

tion into a self-consistent and scalable workflow yields a combination of capabilities that, to our knowledge, have not previously been demonstrated for biomolecular transitions of this complexity. In particular, the present framework couples generative diffusion models, committor learning from short unbiased trajectories, and iterative path sampling in a way that enables quantitative characterization of free-energy landscapes, transition-state regions, and dominant pathways at substantially reduced computational cost.

We have revised the Introduction accordingly to make clearer both the continuity with earlier work and the specific practical advance represented by the present method.

Overall, while the manuscript features an interesting and valuable integration of various established techniques and concepts, it may not represent a significant conceptual leap forward.

**Response**: The Referee characterizes the work as an interesting integration of established ideas, while questioning whether it constitutes a substantial conceptual advance. We appreciate this distinction. Our view is that, in computational molecular science, advances frequently arise from the development of operational frameworks that combine existing concepts in ways that enable new applications or substantially improved efficiency. In this sense, the key issue is whether the integrated workflow provides a capability that was previously difficult to realize in practice.

For the systems examined here, Gen-COMPAS provides access to transition mechanisms, committor structure, and free-energy information at a computational cost that is substantially lower than that associated with long-timescale brute-force simulations. We have revised the manuscript to explain this point more clearly and to focus the discussion of novelty on demonstrated capability rather than on the novelty of the individual ingredients in isolation.

Regarding significance, this manuscript certainly constitutes a valuable addition to the growing body of sampling algorithms and should be of interest to the computational community. Nonetheless, I am not fully convinced that this new method offers the kind of groundbreaking or enabling novelty expected. The benchmark systems are interesting and challenging, but the presented results are not truly unique to this technique. None of the test cases appear to be beyond the reach of alternative, established methods. While the folding of a mini-protein is a standard benchmark, and the two larger, biologically significant systems demonstrate scalability, their conformational changes are relatively contained. These systems could likely be sampled using many existing methods. For acceptance in this journal, a method should ideally demonstrate a unique edge by enabling the study of a system or process that is demonstrably difficult or impossible to reproduce with alternative techniques.

**Response**: We acknowledge that the Referee is concerned about whether the benchmark systems demonstrate a sufficiently distinctive advantage over established alternatives. We agree that this is a central question. Our aim is not to argue that the studied transitions are impossible to access by other methods, in principle, but rather that obtaining comparable mechanistic and thermodynamic information for such systems can require substantial simulation effort with conventional approaches.

To strengthen this point, we have expanded the manuscript to include two additional and substantially more complex biomolecular systems: the rotary transition of the $V_o$ domain of a V-ATPase and the activation of the muscle-type nicotinic acetylcholine receptor (nAChR). Both systems involve large-scale and highly cooperative conformational transitions in membrane proteins, with multiple metastable intermediates and complex allosteric coupling. In both cases, Gen-COMPAS reconstructs pathways and identifies intermediate states that are consistent with available experimental or computational evidence, without requiring predefined reaction coordinates or symmetry constraints (see Figs. R1 and R2).

It is especially worth noticing that, in the case of nAChR, the framework recovers the experimen-

tally suggested mono-liganded/pre-active intermediate while starting only from the **two end-point structures** and **without introducing prior mechanistic information** into the model, achieving this **at a very moderate computational cost.** This mechanism, as well as the associated free-energy differences between states, is in close agreement with recent experimental observations (Fig. R2).

In the manuscript, we have added a new subsection in the Results section : "Gen-COMPAS reveals an asymmetric activation mechanism of nAChR via a mono-liganded intermediate." The section starts:

The muscle-type nicotinic acetylcholine receptor (nAChR) is a prototypical pentameric ligand-gated ion channel (pLGIC) (. . . )

Additionally, in the SI, we have added a new subsection under the section "Further Examples and Extended Data" that begins:

Vacuolar-type ATPases (V-ATPases) are rotary proton pumps that couple ATP hydrolysis in the soluble $V_1$ motor to proton translocation through the membrane-embedded $V_o$ domain. (. . . )

We believe that these additions better illustrate the practical scope of the method and clarify the types of systems for which the workflow may be particularly valuable.

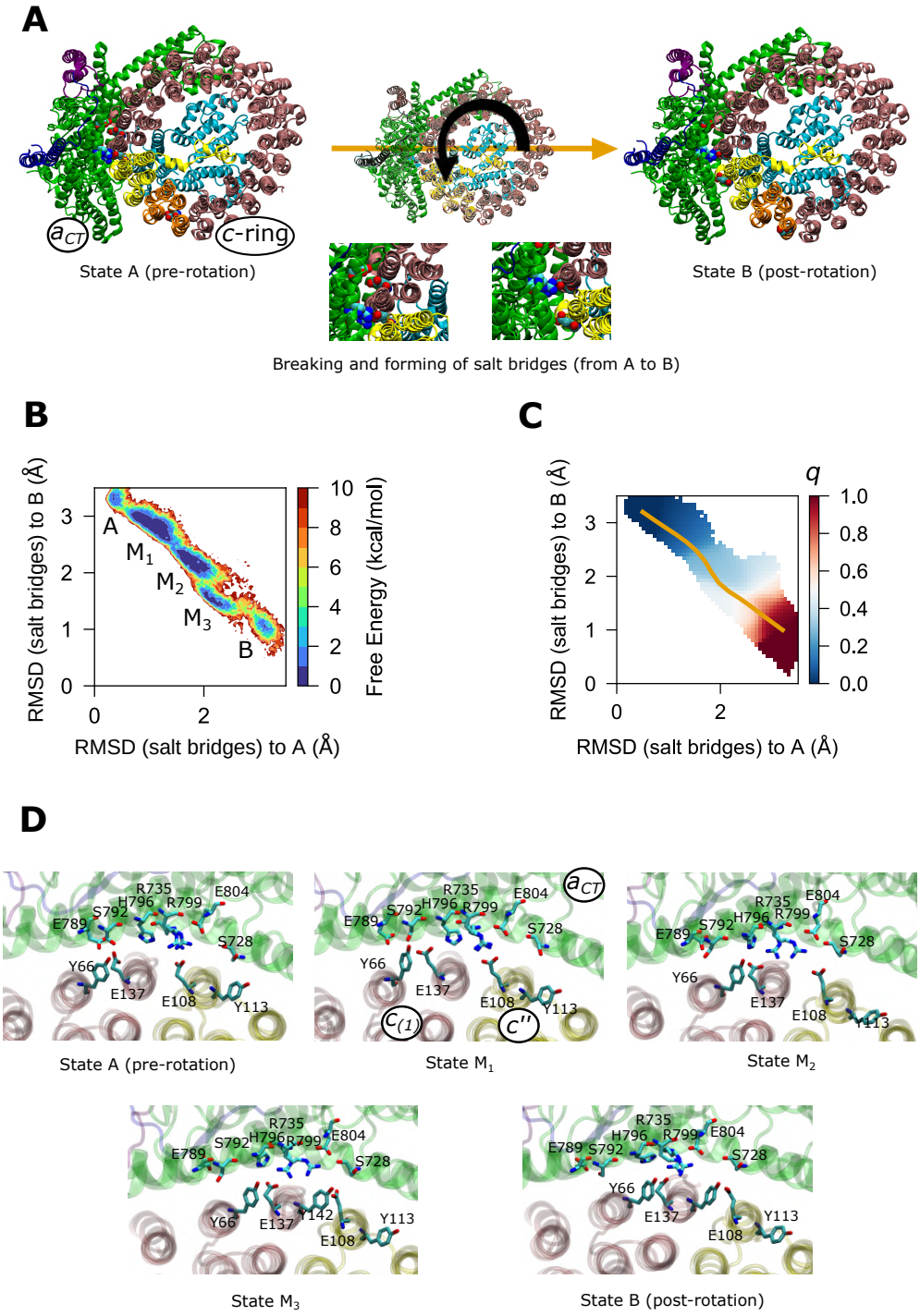


Figure R1: **Gen-COMPAS of V-ATPase $V_o$ domain rotation (A)** Representative structures and rotation of the V-ATPase $V_o$, and the forming and breaking of salt bridges controlling the transition. **(B)** Free-energy landscapes obtained by Gen-COMPAS showing the 5 metastable states. **(C)** The committor and committor-consistent pathways revealing the rotary mechanism by Gen-COMPAS. **(D)** Structures extracted from the Gen-COMPAS simulations along the rotary pathway, and some of the polar residues related to the rotation near the $a_{CT}$:c–ring interface.

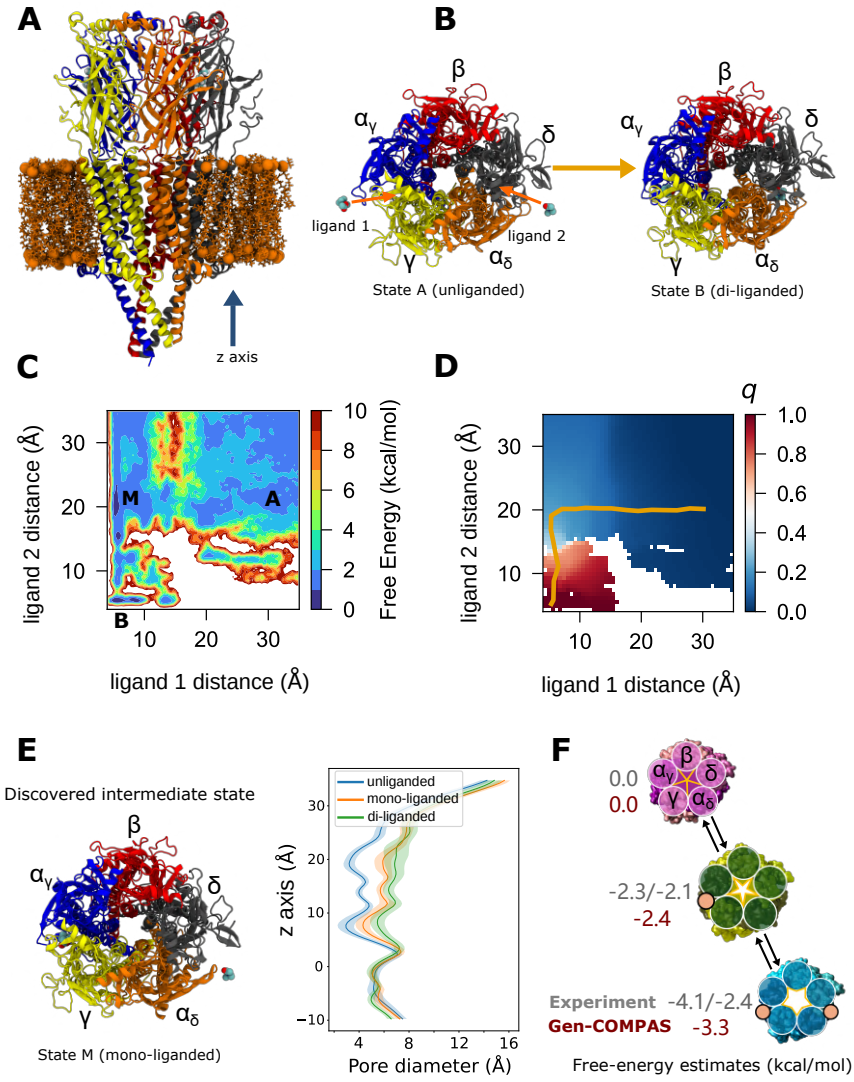


Figure R2: **Gen-COMPAS of the nicotinic acetylcholine receptor (nAChR) (A)** The di-liganded/pore active nAChR in membrane **(B)** Unliganded and pore close state of nAChR (state A) and di-liganded and pore active state of nAChR (state B). The activation is triggered by binding of the two acetylcholine molecules. **(C)** The FEL and **(D)** the committor and committor-consistant pathway projected onto two dimensions (the distance betwen ligands with respect to their binding sites). **(E)** The discovered intermediate mono-liganded/pore pre-active state clustered from Gen-COMPAS unbiased trajectories and the pore diameter along the $z$–axis of the simulation box for unliganded, mono-liganded, and di-liganded states. **(F)** Proposed gating mechanism. Binding of the first acetylcholine molecule at the $\alpha_\gamma$–$\gamma$ interface induces a pre-active configuration of the pore. Subsequent binding of the second acetylcholine molecule at the $\alpha_\delta$–$\delta$ interface completes the activation and stabilizes the active channel. The free-energy differences between these states were estimated by integrating the corresponding basins on the FELs and are consistent with experimental values, inferred from two different kinetic schemes, see M. Thompson *et al.*, Science **391**,eadw1264(2026).

The paper is undeniably of high technical quality. However, I have two main conceptual concerns:

1. Dependence on Targeted MD with RMSD

The authors correctly highlight the inherent difficulty of biased sampling along poor CVs, which can lead to inaccurate estimates of mechanisms, free energy, and rates. Therefore, any new algorithm that extracts useful information from short unbiased trajectories is welcome. However, the core of the sampling strategy relies on targeted MD using the root-mean-square deviation (RMSD) as a CV. I find this choice surprising. It is widely recognized that RMSD is often an inadequate CV for steering system dynamics, as targeted MD is generally effective only when the free energy landscape is not highly curved, typically only for transitions between structures that are already quite close. This dependence suggests that reaching structures close to the putative TSE that are significantly different from the starting state structures might be challenging.

Furthermore, the authors do not discuss the scenario where targeted MD fails to converge to reasonable structures near those proposed by the generative model. I also question the fundamental necessity of the targeted MD step. If the structures produced by the generative models are of sufficient quality, a brief energy minimization might be enough to render them physically viable. The manuscript would benefit significantly from a plot illustrating the structural distance between the generative model outputs and the structures actually converged to by the targeted MD. If the former are high-quality, the utility of the latter is unclear. Conversely, if the former are poor, the success of the entire sampling procedure becomes critically dependent on the quality of the targeted MD itself.

**Response**: We thank the Referee for raising this interesting question. We believe, however, that the comment slightly misinterprets the role of this step in the workflow. The purpose of TMD is not simply energetic relaxation, but to ensure dynamical continuity by connecting the structures generated by the diffusion model to regions of configurational space that are accessible under the true physical Hamiltonian. In this sense, whether alternative relaxation procedures could be used is largely an implementation detail rather than a conceptual limitation, and this part of the workflow does not affect the validity of the approach.

Regarding the Referee's question about possible failure modes of the TMD procedure, or whether TMD is strictly necessary, we emphasize that its function is better understood as dynamical embedding. The diffusion model generates molecular conformations without explicitly representing environmental degrees of freedom such as solvent, ions, or membrane lipids (which cannot be addressed by energy minimization). The TMD step therefore embeds these conformations into the full atomistic environment and connects them smoothly to dynamically accessible regions of phase space under the physical Hamiltonian.

To verify the robustness of this step, we tested the dependence on TMD hyperparameters using the Trp-cage fast-folding protein by varying the force constant and simulation length. We analyzed the RMSD distribution with respect to the target structures over the first and last 2% of the trajectories. Provided that the parameters are chosen within a reasonable range (i.e., avoiding very small force constants or very short simulations), the trajectories consistently converge toward the targets (Fig. R10). The same diagnostic analysis was performed for all systems studied in the manuscript to confirm the appropriateness of the chosen TMD parameters (Fig. R11). Details are added in the section "Validation of TMD parameters" in the SI.

To quantitatively assess the consistency of this procedure and the possible influence of the TMD protocol related to the actual transition, we analyzed the committor difference,

$$\delta q = q_A - q_B,$$

where $q_A$ and $q_B$ are the committor values of the structures obtained from the TMD trajectories

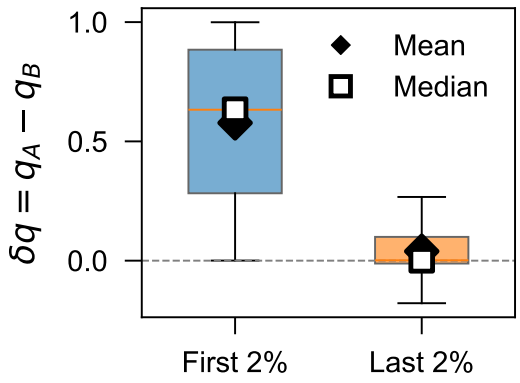


Figure R3: Distribution of the committor difference $\delta q = q_A - q_B$ for configurations obtained from TMD trajectories toward the same diffusion-model-generated targets in the Trp-cage system, accumulated over all Gen-COMPAS iterations. The first ∼2% of the trajectories (blue) shows a broad distribution centered around $\delta q \approx 0.5$, reflecting the memory of the initial metastable states. The last ∼2% (orange) is sharply peaked around $\delta q \approx 0$, indicating convergence to committor-consistent configurations independent of the starting state.

initialized from states $A$ and $B$, respectively, toward the same diffusion-model target.

We evaluated the committor difference, $\delta q$, at the beginning and at the end of the TMD trajectories for all the iterations of the Trp-cage system (see Fig. R3). At the initial stage of TMD (first ∼2% of the trajectory), the distribution of $\delta q$ is broad and centered around approximately 0.5. This behavior is expected because the trajectories originate from distinct metastable states, and therefore the committor values still retain the memory of their starting configurations.

In contrast, at the end of the TMD trajectories (last ∼2%), the distribution of $\delta q$ becomes sharply peaked around zero with very small variance. This indicates that the final configurations reached by TMD from the two different initial states converge to essentially identical committor values. In other words, the endpoint of the TMD trajectory is determined primarily by the diffusion-model-generated target rather than by the initial configuration used to initialize the trajectory.

These results demonstrate that, while the early part of the TMD trajectories naturally reflects the different starting states, the final structures obtained after TMD are highly consistent and largely independent of the initial state. The diffusion-model-generated targets, therefore, define the transition configurations robustly, while the TMD protocol serves primarily to embed these structures into a physically consistent atomistic environment.

In addition, to validate the generation of the TSE and to highlight the difference between the structures generated by the diffusion model and the TMD in such an ensemble, we have quantified the differences of these outcomes from the Gen-COMPAS for Trp-cage protein with respect to the DESRES simulation. For that, we computed an estimate of the Kullback–Leibler divergence (KLD) of the ensemble generated directly by the diffusion model and the one generated after the application of TMD with subsequent equilibration with the transition ensemble that we can identify from the DESRES simulation. The results can be seen in the following table.

| Iteration | Diffusion | TMD | # generated structures (Diffusion – TMD) |
|---|---|---|---|
| $0^{\text{th}}$ | 104.62 | 27.55 | 20 – 42 |
| $1^{\text{st}}$ | 106.56 | 28.64 | 139 – 281 |
| $2^{\text{nd}}$ | 71.06 | 14.37 | 297 – 597 |

Table R1: KLD of Trp-cage TSE directly generated from the Diffusion model and after the application of the TMD with the posterior equilibration. We selected as the reference distribution the set of structures of the DESRES trajectory within $\Delta q = 0.1$ around the separatrix from a VCN model learnt on this trajectory (402 structures). The estimate of the KLD is computed for $k = 3$.

The KLD estimator is built such that it avoids dimensional problems on histogram representation. For that, we have used the one proposed in Refs. F. Pérez-Cruz, IEEE Int. Symp. Inf. Theory, 1666–1670 (2008) and Q. Wang *et al.*, IEEE Trans. Inf. Theory **55** (5) 2392–2405 (2009), based on the $k$-th nearest neighbor (k-NN) algorithm,

$$\widehat{\mathscr{D}}_{\text{KL}}(P_{\text{GC}}\|P_{\text{DESRES}}) = \frac{d}{n}\sum_{i=1}^{n}\ln\frac{\nu_k(i)}{\rho_k(i)} + \ln\frac{m}{n-1}, \tag{R1}$$

where $d$ is the dimensionality of the Z-matrix set of coordinates, $n$ is the number of structures identified in the TSE estimated within the different iterations of the Gen-COMPAS methodology, $\Omega_{\text{GC}}$; $m$ the number of structures identified in the TSE estimated in the the DESRES trajectory, $\Omega_{\text{DESRES}}$; $\nu_k(i)$ is the $k$-NN of the $i$-th structure, $\mathbf{x}_i \in \Omega_{\text{GC}}$ with respect to the structures in $\Omega_{\text{DESRES}}$; and $\rho_k(i)$ is the $k$-NN of $\mathbf{x}_i \in \Omega_{\text{GC}}$ with respect to $\Omega_{\text{GC}} \setminus \{\mathbf{x}_i\}$.

These results are added in the SI in a new section: "Transition-state-ensemble validation for Trp-cage".

From the previous analysis, we believe that TMD creates a physically closer ensemble to the *real* one as the comparison with the diffusion structures highlights. However, this does not mean that the diffusion part is unnecessary. As we state in the manuscript, the diffusion model is able to generate plausible intermediates learned from previous iteration knowledge with a cheap computational cost compared to MD simulations even with enhanced sampling techniques. However, the role of TMD is not to discard the generated structures but to give them physical meaning. The former is the reason why TMD complements the diffusion generative part of the methodology.

2. Lack of Committor Validation

While the entire method is founded on the principled use of the committor, the manuscript lacks any committor validation plot. The authors claim accurate identification of the committor and the separatrix (the isosurface where the committor value is 1/2). Estimating the committor is a difficult task, and any claim of accuracy must be supported by a rigorous validation. This is standardly done by "shooting" trajectories from structures on the separatrix and counting the fraction that ends up in one metastable state versus the other. The resulting histogram of counts must be sharply peaked around the value 1/2. The absence of this validation makes it impossible to assess the accuracy of the learned committor.

**Response**:

We thank the Referee for pointing out the validation of the predictive power of the committor learning model. We have performed this validation it for NANMA, since the computational cost for performing a shooting analysis is considerably more affordable than for the larger and more complex systems studied here. Specifically, we selected configurations generated during the different iterations of the Gen-COMPAS procedure, and identified, according to the committor predicted by the VCN model trained on the heavy-atom Z-matrix representation, structures compatible with $q = 0.3$, 0.5, 0.7 within a tolerance of $\Delta q = 0.05$. For each selected configuration, we launched 1000

unbiased simulations with randomized velocities drawn from a Maxwell–Boltzmann distribution compatible with $T = 300$ K. The results for each set of configurations can be observed in Fig. R4. The resulting distributions of observed committor values are compared with the target committor intervals from which the structures were selected. The sampled committor distributions are centered around these theoretical ranges, and their mean values show good agreement with the expected committor values.

To further validate the results from the VCN model trained using the Z-matrix representation, we compared the predicted committor values with those obtained from a shooting analysis performed along an unbiased trajectory of NANMA comprising 460,000 steps. For each configuration sampled at every step of this trajectory, we performed the shooting analysis described above. The results are depicted at Fig. R5. A clear correlation is observed between the sampled committor probabilities and those predicted by the learned model. The correlation coefficient between sampled and learned values is of 0.9999668, indicating excellent agreement between the two estimates.

These results are added in the SI, in subsection "Committor Validation Analysis", with the new Figures S7–S9. We have added a new sentence in the main manuscript referring to this new section in the SI:

(...) A detailed account of the committor validation procedure is available in the "Committor Validation Analysis" section of the Supplementary Information (SI).

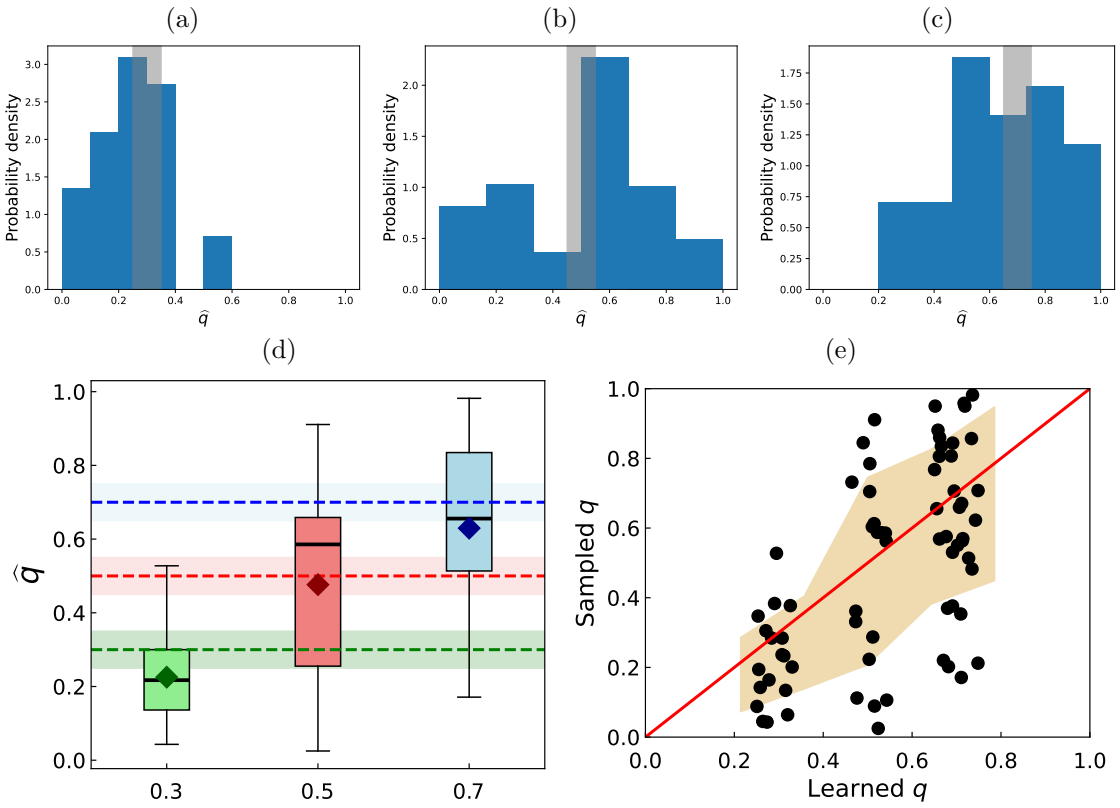


Figure R4: Histograms of the sampled committor values for selected configurations within $q = 0.3 \pm 0.05$ (a), $q = 0.5 \pm 0.05$ (b), $q = 0.7 \pm 0.05$ (c) according to the VCN learnings. The vertical grey band correspond to the learned committor interval of selection. In (d) the boxplots for each histogram is represented against the interval in which each set of structures are selected according to the VCN learnings. The mean values for each set are represented with diamonds. We also plot the Learned $q$ vs. Sampled $q$ plot (e) from which a clear correlation can be observed.

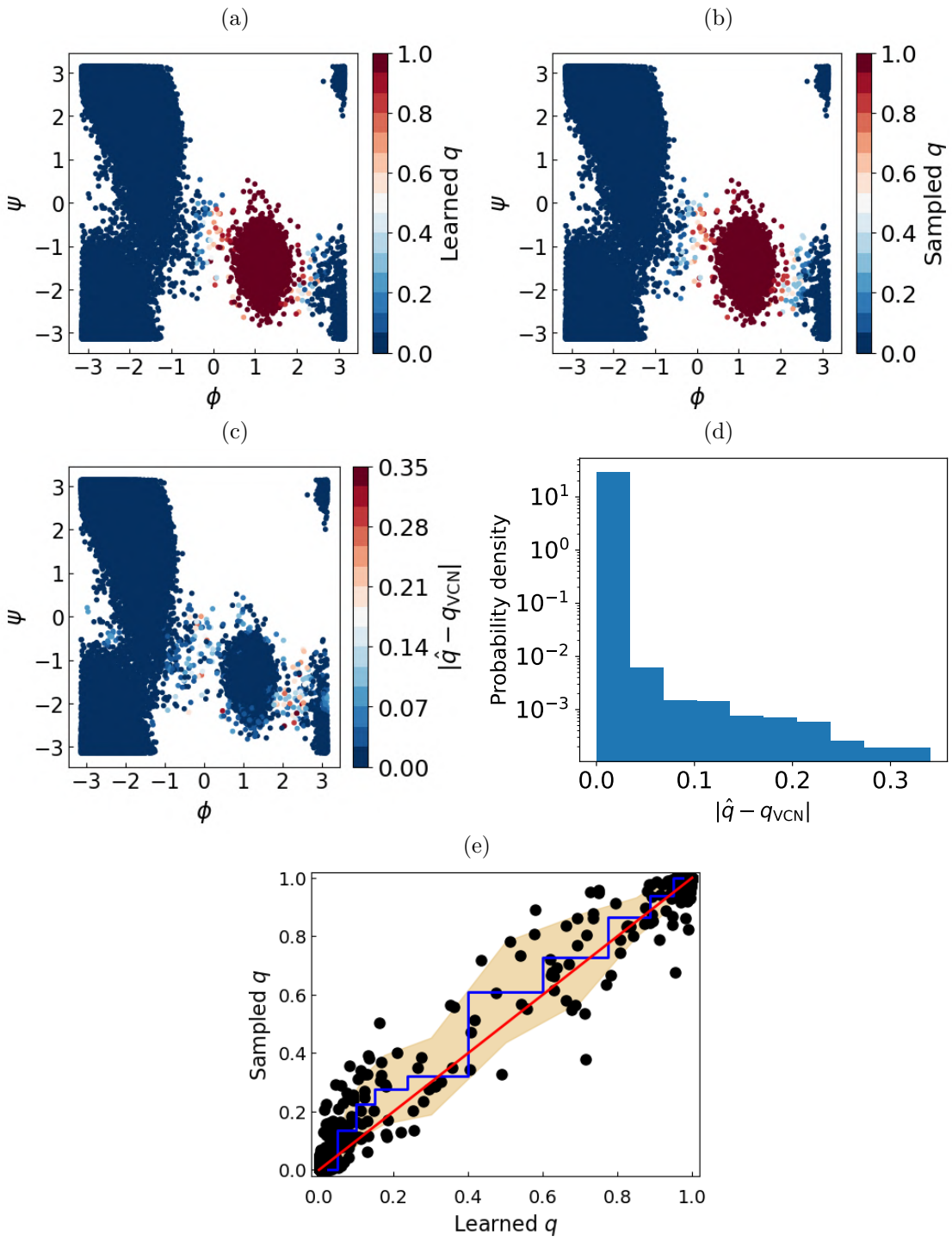


Figure R5: Analysis of the learned committor probability $q$ for NANMA showing (a) learned values with the VCN using the Z-matrix representation, $q_{\mathrm{VCN}}$ projected into the $(\phi, \psi)$ two-dimensional space in radians, (b) sampled committor values, $\hat{q}$, (c) the absolute error, $|\hat{q} - q_{\mathrm{VCN}}|$ projected into the $(\phi, \psi)$ two-dimensional space in radians, (d) the distribution of the absolute errors, and (e) learned vs. sampled values to infer the correlation between VCN outcomes and the estimates from the shooting analysis.

Additional issues

* Generative AI for Membrane Proteins: As generative AI models for complete membrane-protein systems are not yet widely available, the authors should elaborate on the specific implementation used to generate viable structures for the membrane-protein system featured in the manuscript.

**Response**: We thank the Referee for raising this point and appreciate the opportunity to clarify our implementation. In this work, we train new generative diffusion models directly on the molecular dynamics trajectories produced for each system. Therefore, these models learn the conformational statistics of the protein as sampled in MD simulations where the protein is already embedded in an explicit membrane environment. Importantly, the generative model itself does not explicitly include membrane atoms. Instead, it learns the conformational space of the protein that is physically constrained by the surrounding membrane during the MD sampling used for training.

The structures proposed by the generative model are therefore interpreted as candidate intermediate protein conformations rather than fully equilibrated membrane–protein configurations. Any minor inconsistencies arising from the absence of the membrane in the generative stage are corrected in the subsequent TMD refinement step. In this step, TMD simulations are initiated from equilibrated endpoint structures where the protein is properly embedded in the membrane and solvated environment. The generated intermediate conformations serve only as structural targets, while the MD simulations enforce the full physical environment, allowing the membrane, solvent, and protein degrees of freedom to relax consistently.

This hybrid strategy that combines generative proposals with physics-based MD refinement allows Gen-COMPAS to treat large membrane proteins such as AAC, the $V_o$ rotor domain, and nAChR without requiring a specialized generative model that explicitly includes the membrane. The approach therefore leverages the physical realism of MD while retaining the exploration efficiency of generative models.

* Kinetic Observables: Lines 350-352 state that "Knowledge of the committor enables the reconstruction of key kinetic observables, including the transition-state ensemble (TSE), transition-path ensemble (TPE)." It should be noted that the TSE and TPE are not strictly kinetic *observables*. Furthermore, while the manuscript claims that the sampling scheme enables the retrieval of kinetic observables, the authors never estimate any kinetic rates and compare them to benchmark values, which would substantially strengthen this claim.

**Response**: We agree with the Referee that transition-state and transition-path ensembles are not kinetic observables, and we have changed throughout the manuscript "kinetic observables" by "kinetic information". With respect to the kinetic rates, the manuscript does not claim to compute transition rates but to reconstruct mechanistically meaningful ensembles, consistent with committor theory. Rate estimation would require additional assumptions and is orthogonal to the primary contribution, which concerns sampling efficiency and mechanistic resolution. To avoid any misunderstanding, we have amended the paragraph mentioned by the Referee. Before, it read:

(...) Knowledge of the committor enables the reconstruction of key kinetic observables, including the transition-state ensemble (TSE), transition-path ensemble (TPE), reactive currents, and transition rates between metastable states.

Now, it reads:

(...) Knowledge of the committor enables the reconstruction of key kinetic information, including the TSE, TPE, or reactive currents.

We believe that this is solved in the new version of the manuscript.

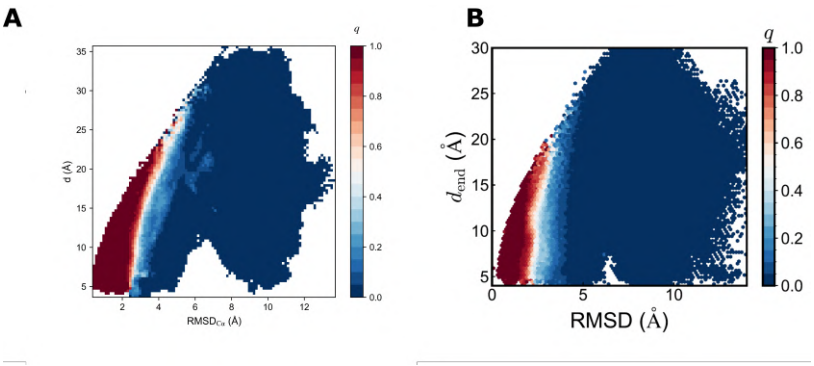


Figure R6: **(A)** The VCN with Z-matrix representation trained on Gen-COMPAS Trp-cage data projected onto RMSD of $C_\alpha$ atoms and end to end distance. **(B)** The qGNN trained on DESRES Trp-cage data, see S. Contreras Arredondo *et al.*, Nat. Comput. Sci. (2026).

* Lag Time Dependence: Equations 2 and 4 show that the variational principle and the committor estimate explicitly depend on the choice of a lag time. The optimization process and how the accuracy of the estimate depends on this crucial parameter should be discussed and analyzed.

**Response**: We thank the Referee for this comment. The dependence of the variational principle on the lag time has been addressed previously in other previous works, see for example: H. Chen *et al.*, J. Chem. Theory Comput. **19**, 14, 4414–4426 (2023); A. Megías *et al.*, Nat. Comput. Sci. **5**, 592-602 (2025); and S. Contreras Arredondo *et al.*, Nat. Comput. Sci. (2026). We have added the following sentence to the main text:

(. . . ) Moreover, the influence of the choice of the lag time $\tau$ on the learning process is discussed in detail in Refs. 26,27,29.

* Molecular Representation: While the Z-matrix representation is reasonable and was standard until recently, it is becoming obsolete, particularly as it focuses only on the protein and is less suitable for materials or accurately capturing the environment of a complex molecular system. Given that the authors have also proposed GNN schemes for committor learning, it would be beneficial to explain the choice not to use them here.

**Response**: We thank the Referee for making this thoughtful comment. While Z-matrix representations were historically standard and remain appropriate for describing internal molecular coordinates, particularly in well-defined conformational studies, it is fair to question their generality. Z-matrices primarily encode internal geometry and are less naturally suited to representing molecular systems with meaningful interaction with the environment. That said, describing the Z-matrix as "obsolete" may be somewhat overstated. It remains a valid and well-defined coordinate system for most classes of problems.

With respect to the comment of the Referee about the GNN recently developed by some of the authors of this manuscript, we want to clarify that the VCN used with the Z-matrix representation can be substituted by another committor learning model. The intention is to minimize computational cost and to keep the compatibility with committor-consistent pathway searching scheme. Still, systems exposed in this work are well described by the Z-matrix representation. By comparing, for example, the committor obtained from the Z-matrix VCN and the recently published GNN for the Trp-Cage folding process, we can observe good agreement (see Fig. R6).

* Multiple States: The committor is formally defined for a transition between two states. The authors should clarify their practical approach for handling systems with multiple metastable states.

**Response**: We thank the Referee for pointing this theoritical formality out. In our practical approach, we define the states $A$ and $B$ as the global initial and final endpoints of the process; any intermediate metastable states simply emerge naturally along the pathway as regions with intermediate committor values. To explicitly demonstrate this, we have added an analysis of the $V_o$ domain of a V-ATPase to the SI (see Fig. R1 or S4). By defining only the global endpoints of the rotary step, our method successfully resolved multiple intermediate metastable basins, the transition-state ensemble, and the sequential salt-bridge rearrangements driving the mechanism.

* Convergence Test: The convergence test detailed in the Supporting Information, which assumes accuracy if repeated sampling iterations fail to add new structures, is theoretically questionable. While operationally acceptable, there is no formal guarantee that this lack of new structures ensures committor accuracy. The only reliable and straightforward way to confirm accuracy is through the committor validation tests mentioned previously.

**Response**: We thank the Referee for addressing this point. We believe that the committor validation test is not computationally addressable for most of the systems, which is the trigger for bringing up machine-learning-guide methodologies. However, we have addressed this validation for NANMA as we have discussed before and for Referee 3 in their question (2). The validation for this system shows good agreement between learned and sampled committor values in terms of distributions and correlations. We also believe that there are other signatures of validity that strengthen our outcomes, such as, for example, the estimates of the FELs.

To check the FEL estimations from Gen-COMPAS, we have included a section in the SI, belonging to the section "Convergence and uncertainty analysis of FEL estimations", where we presented a way to verify the convergence of the FEL as a function of the number of trajectories included in the analysis. The text being added to the SI states as follows:

To assess the robustness of the FELs obtained from the Gen-COMPAS simulations, we performed systematic convergence and uncertainty analyses based on trajectory-level bootstrapping (. . . )

The results can be seen in Fig. R7.

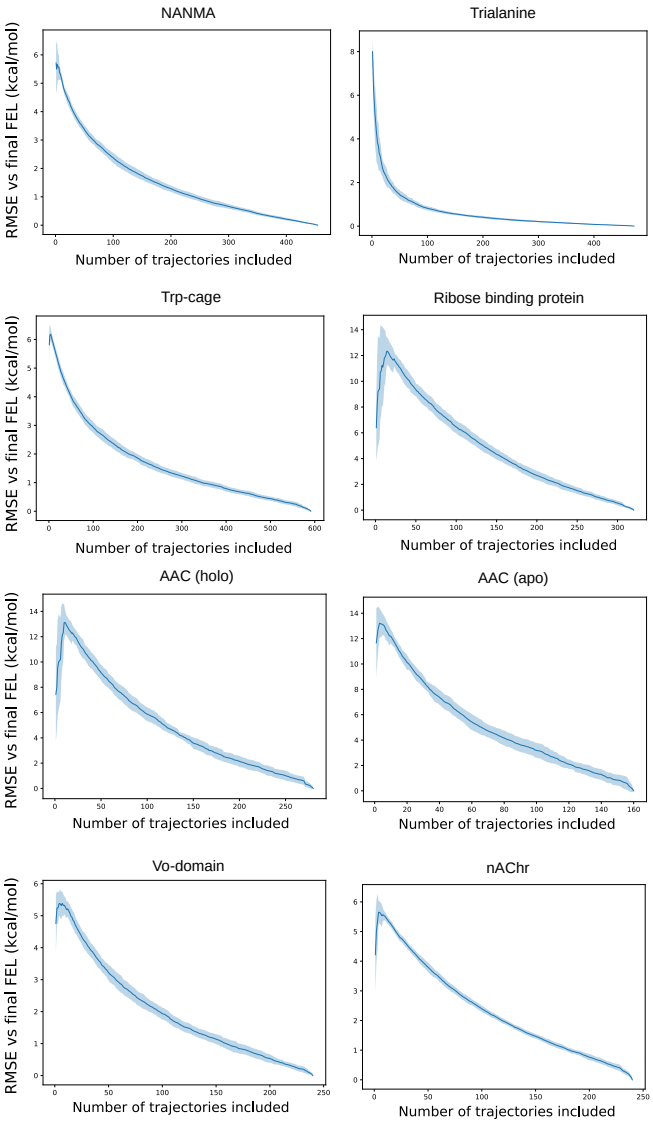


Figure R7: **Convergence of FEL estimation** The convergence check based on comparing the FEL estimation with respect to the final estimation as including more unbiased trajectories from Gen-COMPAS.

A significant weakness is the complete absence of appropriate uncertainty quantification. There are no error bars or confidence intervals reported for any of the estimated quantities. For instance, demonstrating the error in estimating the free energy difference between metastable states as a function of cumulative sampling would be highly valuable.

**Response**: We appreciate that the Referee highlights the absence of error and uncertainty estimates. We agree that uncertainty quantification is important, and we welcome the opportunity to include it. Therefore, the error estimation of the FEL for every system based on bootstrapping and the uncertainty check of the committor predictions based on ensemble training are included in the SI, belonging to the section "Convergence and uncertainty analysis of FEL estimations" and the section "Ensemble-based uncertainty estimation of the committor":

We estimated statistical uncertainties in the FEL using a bootstrap resampling procedure performed at the trajectory level (. . . )

and

As can be seen in the previous subsection, the validation of committor function through shooting unbiased trajectories, although numerically stable, is computationally costly and practically complicated and therefore inaccessible for complex systems. To assess the reliability of the learned committor function with reduced computational cost, we opted to estimate the prediction uncertainty using an ensemble of independently trained neural networks (. . . )

The newly included results can be seen in Fig. R8 and R9.

As with committor validation, we believe that this is a presentational and analytical extension, not a challenge to the soundness or novelty of the methodology. We believe that uncertainty estimation, as well as committor validation, have been solved with the analysis carried out in the new version of the work.

In conclusion, the paper is well-written and holds interest for the rare event sampling community. However, it does not currently demonstrate sufficient novelty and significance to recommend publication in a top-tier journal.

From a technical standpoint, the lack of any committor validation and the central reliance on targeted MD along the RMSD, which is known to be an inadequate CV, are major points that must be addressed and clarified for publication in any reputable journal.

Unfortunately, a clear, demonstrable conceptual breakthrough is absent, and the manuscript does not convincingly show that the proposed method uniquely enables the sampling of systems otherwise inaccessible by standard, alternative techniques.

**Response**: We thank the Referee for the detailed suggestions and for recognizing the technical quality and relevance of the work. In the revised manuscript, we have responded by adding explicit committor-validation analyses, uncertainty estimates, reproducibility tests, free-energy reweighting with RiteWeight, and two additional complex biological examples. While the methodology is directly relevant to the rare-event sampling community, we believe that its significance extends beyond this domain. The proposed framework provides a practical and scalable approach to extract dynamical and mechanistic insight from a limited number of static configurations, without requiring large training datasets or access to exceptional computational resources. By enabling the reconstruction of transition pathways, intermediate states, and free-energy landscapes in a data-efficient manner and without prior mechanistic bias, the method may serve as a broadly applicable tool for researchers aiming to understand complex molecular processes from limited structural information. We believe that the results presented in the revised version of the manuscript support this broader applicability.

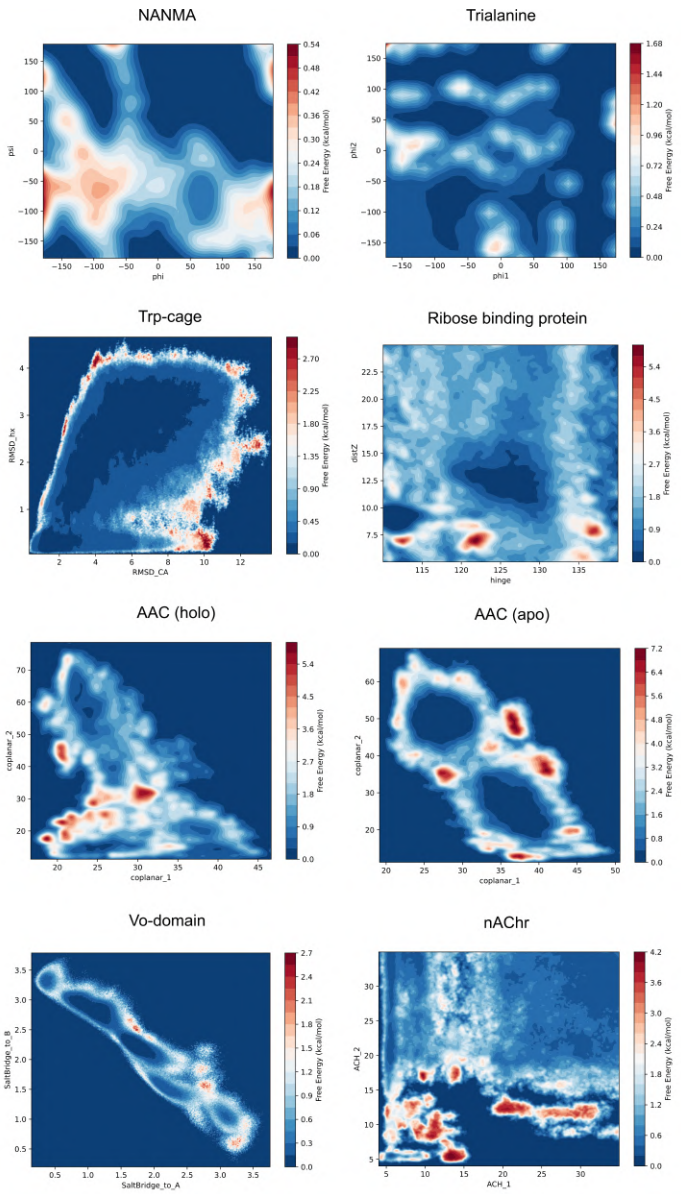


Figure R8: **Standard deviation of FEL estimation** The uncertainty estimation of the FEL estimation from Gen-COMPAS by checking the standard deviation of the weighted histogram projected on two dimensions.

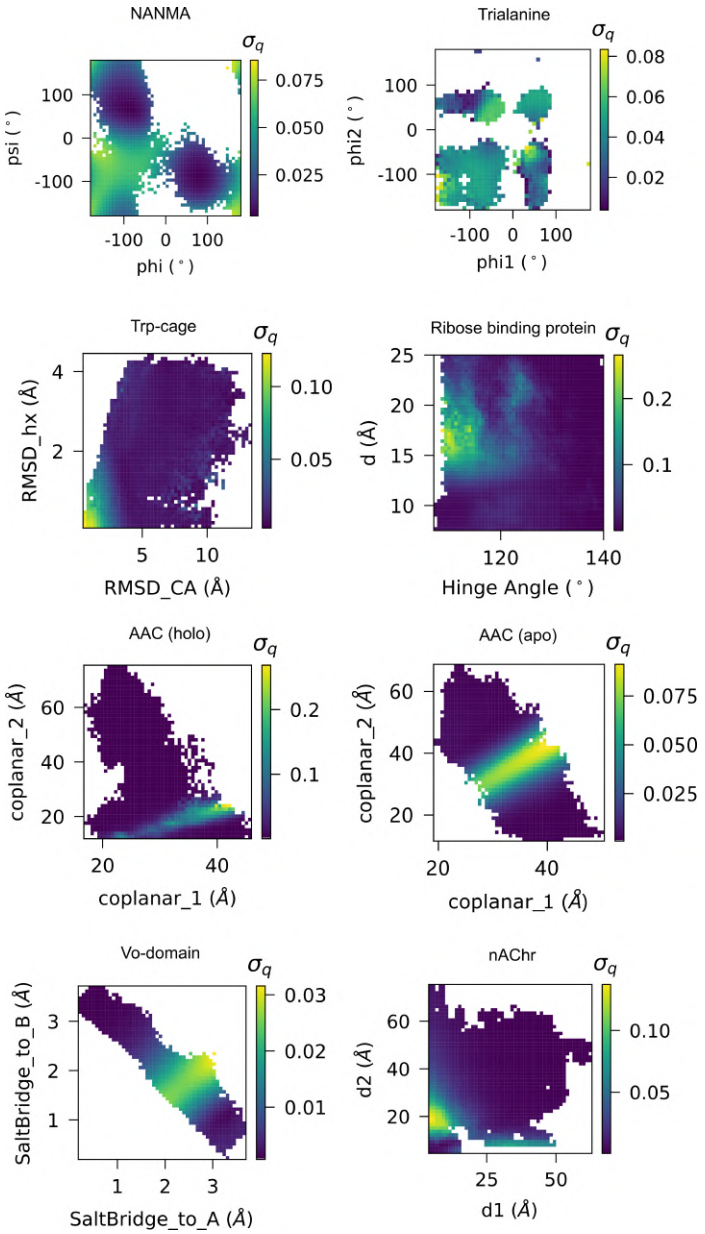


Figure R9: **Uncertainty estimation of committor function** The uncertainty estimation of the committor function from Gen-COMPAS by checking the standard deviation of the ensemble predictions projected on two dimensions.

## Response to Referee 3

The work addresses an important challenge in molecular dynamics simulation of biomolecules by efficiently constructing transition paths between defined metastable states. The main innovation is the use of a denoising diffusion probabilistic model (DDPM, "diffusion model") to navigate the configuration space intermediate between starting and end states A and B. A more technical aspect is the use of a variational committor network to estimate the commitment probability ("committor") to B, also from trajectories that do not quite reach A or B. Finally, the internal representation of configurations in terms of Z-matrices is quite general, avoiding the need to pre-define collective variables, even if solvent water and ions appear not to be included explicitly. The work includes applications of increasing complexity, from the folding of the mini-protein "Trp-cage" over the binding of ribose to the ribose-binding protein RBP to the access transition of the membrane-bound ATP/ADP carrier protein AAC. Overall, I find this to be a very promising direction with quite impressive results. However, a number of major points are unclear and should be addressed in a revision, as listed in the following.

**Response**: We sincerely thank the Referee for this thoughtful and encouraging assessment of our work. We are grateful for the clear summary of the main methodological components, i.e., the diffusion-based transition path construction, the committor learning, or the use of Z-matrix representations, as well as for recognizing the increasing complexity of the presented applications.

We especially appreciate the Referee's view that this approach represents a promising direction, and that the results are "impressive". At the same time, we fully acknowledge that several important aspects require clarification. We are grateful for the detailed comments that follow, and we hope that the revisions carefully address each of the major points raised.

(1) It was not quite clear to me what the impact of the diffusion model in the whole procedure was. As I understand, it was used only to create perturbed initial configurations for further sampling (Fig. 1A,B). Wouldn't it have been easier to simply start new simulations from configurations visited in the short runs, chosen according to their committor value as predicted by the variational committor network?

**Response**: We thank the Referee for raising this important point, and for prompting us to clarify the role of the diffusion model in the overall procedure.

While it is correct that the diffusion model is used to generate perturbed configurations that seed further sampling (Fig. 1A,B), its role is more fundamental than merely providing noise around already visited states. The key purpose of the diffusion model is to propose configurations across a broader region of configurational space, in particular within regions relevant to transitions between metastable states. In complex systems with significant timescale separation, short unbiased trajectories initiated within the basins typically remain confined there and may not reach configurations outside the metastable wells. Said differently, such transitions correspond to so-called rare events. As a result, sampling based solely on configurations visited in short runs—even when selected according to their committor values—may fail to explore configurations withing the transition region because of the underlying free-energy barriers. One possible way to alleviate this limitation is to employ enhanced-sampling methodologies; however, these approaches generally rely on the selection of collective variables, which does not guarantee adequate exploration of the transition-state ensemble.

The diffusion model alleviates this limitation by learning a generative representation of the configurational space and enabling proposals that extend beyond those directly observed in short trajectories. In this sense, it facilitates exploration between metastable regions by generating plausible intermediate configurations that would be extremely unlikely to appear in short-time dynamics. These proposals then serve as starting points for subsequent dynamical refinement and committor-based characterization of the transition process.

In order to clarify this point further, we have added the following text to the main manuscript:

(. . . ) The role of the intermediate outcomes generated by the diffusion model is to solve the exploration of the rare events characterizing the transition mechanism. (. . . )

(2) I could not find a demonstration for the different applications (including the small peptides in SI) that the learned committor has predictive power. There are various ways of testing the committor, the simplest one being repeated initializations from a given starting configuration (the so-called "committor test"). In such a test (as sketched in Fig. 1D), do individual configurations taken on the isocommittor surface individually have 50% chance of ending up in A and B? I.e., is Gen-COMPAS able to reliably identify transition states? And how do the committor predictions perform for configurations away from the isocommittor surface, e.g., at the 30% and 70% isosurface?

**Response**: We thank the Referee for raising the important point of validating the predictive power of the learned committor model. We have performed this validation it for NANMA, since the computational cost for performing a shooting analysis is considerably more affordable than for the larger and more complex systems studied here. Specifically, we selected configurations generated during the different iterations of the Gen-COMPAS procedure, and identified, according to the committor predicted by the VCN model trained on the heavy-atom Z-matrix representation, structures compatible with $q = 0.3$, 0.5, 0.7 within a tolerance of $\Delta q = 0.05$. For each selected configuration, we launched 1000 unbiased simulations with randomized velocities drawn from a Maxwell–Boltzmann distribution compatible with $T = 300$ K. The results for each set of configurations can be observed in Fig. R4. The resulting distributions of observed committor values are compared with the target committor intervals from which the structures were selected. The sampled committor distributions are centered around these theoretical ranges, and their mean values show good agreement with the expected committor values.

To further validate the results from the VCN model trained using the Z-matrix representation, we compared the predicted committor values with those obtained from a shooting analysis performed along an unbiased trajectory of NANMA comprising 460,000 steps. For each configuration sampled at every step of this trajectory, we performed the shooting analysis described above. The results are depicted at Fig. R5. A clear correlation is observed between the sampled committor probabilities and those predicted by the learned model. The correlation coefficient between sampled and learned values is of 0.9999668, indicating excellent agreement between the two estimates.

These results are added in the SI, in subsection "Committor Validation Analysis", with the new Figures S7–S9. We have added a new sentence in the main manuscript referring to this new section in the SI:

(. . . ) A detailed description of a committor validation procedure is provided in the section "Committor Validation Analysis" in the Supplementary Information (SI).

(3) It was not clear to me if/how simulation-quality intermediate configurations were "generated" by the diffusion model (i.e., with proper stereochemistry and clash statistics) and how solvent was treated in their construction. As I understand, starting structures for the MD runs are not actually created by the diffusion model. Instead, the structures of the diffusion model serve as targets for "targeted MD" (TMD) simulations. I would imagine that this creates some uncertainty and dependence on details (hyperparameters) of the TMD simulations. For instance, if TMD is repeated with the same target but from different starting configurations, is the committor value of the respective endpoints the same (and as predicted) or dependent on the starting structure? I.e., how consistent is the TMD-based construction of intermediates?

**Response**: We are grateful to the Referee for raising this important point regarding the role of the TMD protocol and the potential dependence of the generated intermediates on simulation details.

As the Referee correctly notes, the structures produced by the diffusion model are not directly used as simulation starting points. Instead, they serve as targets for targeted molecular dynamics (TMD) simulations that are performed within the full atomistic environment.

In our implementation, the diffusion models are trained using the coordinates of all heavy atoms of the protein, allowing the model to capture internal structural correlations and stereochemistry of the solute. However, explicitly including solvent molecules, membrane components, or ions in the generative model is currently impractical due to the very large and variable number of environmental degrees of freedom at play. For this reason, diffusion-generated structures are embedded into a physically realistic environment through TMD simulations performed in explicit solvent.

In the TMD protocol, simulations are always initialized from the same endpoint structures corresponding to the metastable states $A$ and $B$. This choice ensures that the generated intermediates are not constructed sequentially from previously generated structures, which could otherwise lead to accumulation of structural errors. Instead, each trajectory begins from a well-equilibrated configuration of the physical system, and is then guided toward the structure proposed by the diffusion model.

To quantitatively assess the consistency of this procedure and the possible influence of the TMD protocol, we analyzed the committor difference,

$$\delta q = q_A - q_B,$$

where $q_A$ and $q_B$ are the committor values of the structures obtained from the TMD trajectories initialized from states $A$ and $B$, respectively, toward the same diffusion-model target.

We evaluated the committor difference, $\delta q$, at the beginning and at the end of the TMD trajectories for all the iterations of the Trp-cage system (see Fig. R3). At the initial stage of TMD (first $\sim$2% of the trajectory), the distribution of $\delta q$ is broad and centered around approximately 0.5. This behavior is expected because the trajectories originate from distinct metastable states, and therefore the committor values still retain the memory of their starting configurations.

In contrast, at the end of the TMD trajectories (last $\sim$2%), the distribution of $\delta q$ becomes sharply peaked around zero with very small variance. This indicates that the final configurations reached by TMD from the two different initial states converge to essentially identical committor values. In other words, the endpoint of the TMD trajectory is determined primarily by the diffusion-model-generated target rather than by the initial configuration used to initialize the trajectory.

These results demonstrate that, while the early part of the TMD trajectories naturally reflects the different starting states, the final structures obtained after TMD are highly consistent and largely independent of the initial state. The diffusion-model-generated targets, therefore, define the transition configurations robustly, while the TMD protocol serves primarily to embed these structures into a physically consistent atomistic environment.

Further, to quantify the quality of the structures generated by the diffusion model and those obtained after the TMD-based initialization, we performed an additional analysis using the Trp-cage system. Specifically, we compared the transition-state ensemble produced by Gen-COMPAS with a reference ensemble obtained from long-timescale DESRES simulations. We quantified the difference between ensembles by computing an estimate of the Kullback–Leibler divergence (KLD) between (i) the ensemble generated directly by the diffusion model and (ii) the ensemble obtained after applying TMD followed by equilibration. The reference distribution was derived from the transition ensemble identified in the DESRES simulations.

The results of this analysis are summarized in Table. R1 of this response and are also reported in the SI as Table. S5. These results indicate that the ensembles obtained after TMD initialization

remain consistent with the transition ensemble derived from the reference simulations and show that the TMD step does not introduce significant bias, while enabling the diffusion-generated structures to be embedded in a physically realistic atomistic environment.

(4) The quality of the free energy surfaces is quite impressive for Trp-cage (Fig. 2C), but it is not really assessed against a reference for the more challenging cases of RBP and AAC. Are these surfaces rough estimates? Can we trust the relative populations of A and B? I am particularly concerned about the free energy estimate, because it relies on some rather strong assumptions (eq S10-S11) to get a trajectory weight factor. Where do these assumptions break down? How would one know?

**Response**: The comment of Referee is well taken, and the free-energy surfaces reported in the manuscript should be regarded as estimates that can serve as starting points for further simulations when the highest level of quantitative precision is required. Because the assumptions underlying the previous procedure were not sufficiently robust mathematically and could lead to numerical instabilities, we have replaced the previous algorithm with the RiteWeight algorithm S. Kania *et al.*, Zenodo (2026). A new section entitled "Free-energy estimation using RiteWeight" has been added to the SI, where the algorithm is described in detail, and the free-energy estimations in the manuscript have been updated accordingly.

Using weights derived from RiteWeight, we additionally quantify the uncertainty of the free-energy estimates through bootstrap resampling and assess convergence of the free-energy landscapes (FELs) by monitoring the RMSE with respect to the final estimate as increasing numbers of trajectories are incorporated into the weighted histogram analysis (see Fig. R8 and R7). These analyses are presented in a new subsection of the SI entitled "Convergence and uncertainty analysis of FEL estimations":

To assess the robustness of the FELs obtained from the Gen-COMPAS simulations, we performed systematic convergence and uncertainty analyses based on trajectory-level bootstrapping. (. . . )

We have changed the main text to add the reference of the RiteWeight algorithm. Before, it read:

(. . . ), and (4) approximation of FELs (see Supplementary Information for further details).

Now, it reads:

(. . . ), and (4) estimates of FELs using RiteWeight algorithm[36] (see SI for further details)

We have added the new reference 36:

[36] Kania, S., Webber, R. J., Simpson, G., Aristoff, D. & Zuckerman, D. M. RiteWeight: Randomized Iterative Trajectory Reweighting for Steady-State Distributions Without Discretization Error (2026).

(5) Related to the preceding points, how reliably are commitors and free energies estimated in repeated, fully independent Gen-COMPAS runs? In such repeated Gen-COMPAS runs, how much do the results depend on hyperparameter choices (such as the length of TMD)?

**Response**: We thank the Referee for raising this point. To assess the dependence of the results on the TMD hyperparameters, we performed independent TMD simulations on the Trp-cage fast-folding protein using different force constants and simulation durations. We analyzed the distribution of the RMSD (with respect to the targets) for the first and last 2% of the TMD trajectories. The analysis shows that, provided the parameters are chosen within a reasonable range, i.e., avoiding excessively small force constants or overly short simulation times, the TMD trajectories converge well toward the target structures. The results are gathered in Fig. R10. For all the systems investigated in

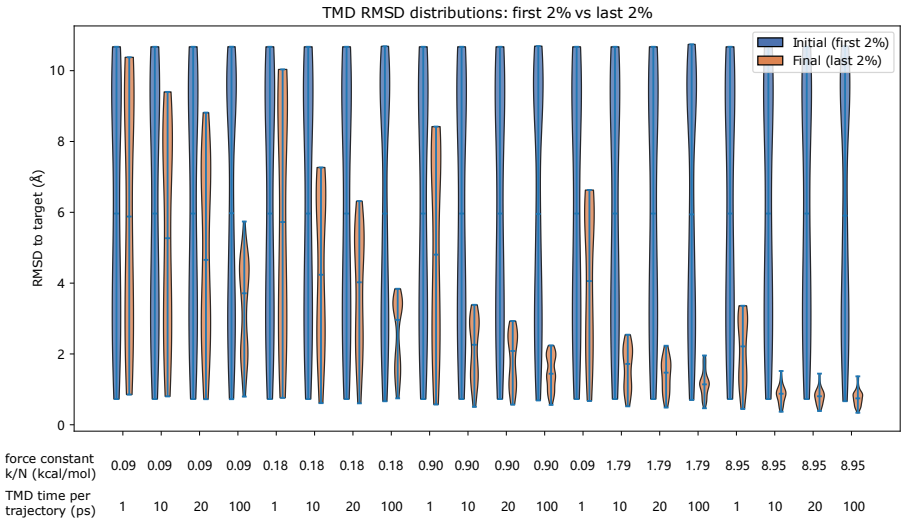


Figure R10: Dependency on TMD parameters tested on Trp-cage. The TMD is replicated for the $0^{\text{th}}$-iteration of the Gen-COMPAS.

the manuscript, we also performed the same diagnostoc analysis to verify the appropriateness of the chosen TMD parameters (see Fig. R11).

Furthermore, to assess reproducibility, we performed five independent Gen-COMPAS runs on the Trp-cage fast-folding protein (see Fig.R12), three of which used identical parameters. The FEL estimation and committor predictions are in good agreement across the independent runs. When the TMD parameters are chosen poorly (for example, in **Run 1**, where the force constant is set too small), the sampling of transition states becomes suboptimal. However, this issue can be readily avoided by verifying the distribution of RMSDs relative to the targets at the end of the TMD simulations, as illustrated in Fig. R10 and R11.

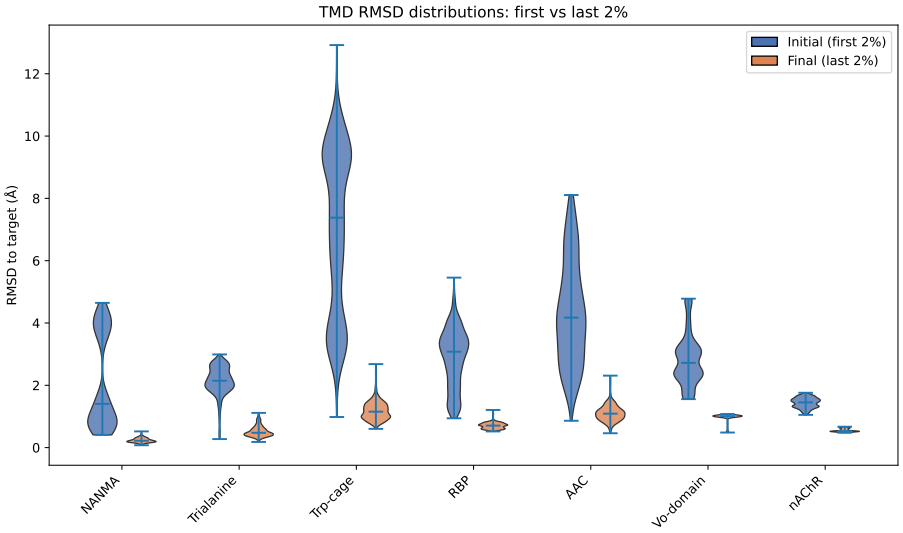


Figure R11: TMD convergence check on all systems.

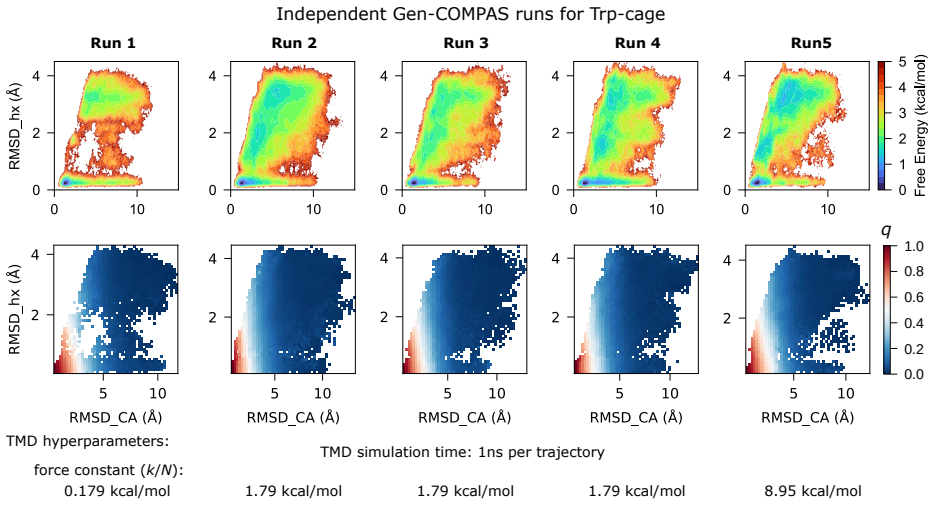


Figure R12: Independent runs of Gen-COMPAS on Trp-cage

(6) The statistics of the sampling is not quite clear. In particular, for step 7 in Figure 1A ("short simulations from transition states"), how many of those runs from the commitor=1/2 surface were done for each of the systems studied and where did they end up for given initial configurations? Also, how much computer time goes into which part of the Gen-COMPAS work flow?

**Response**: We refer to the Referee to the subsection "Numerical details" in the SI where all the computational details of Gen-COMPAS work flow in the difference systems are discussed. Moreover, in Table S4, all the details about the unbiased simulations can be found. We thank the Referee for bringing up the computational cost and have summarized the GPU hours for each step of the Gen-COMPAS framework that are spent on the molecular systems other than the two toy models in Table S6 in the SI.

(7) A qualitative and quantitative comparison to related established methods pursuing the same goal with similar procedures would be important to assess the advance of this new method. As points of reference, one should consider standard transition path sampling (as pioneered by Dellago, Bolhuis, Chandler et al.), AIMMD (ref 7), and the string methods of E, Vanden Eijnden et al. (refs 22-24) and by Pan, Roux et al. (ref 25). Of these, AIMMD in particular uses a quite similar work flow of path sampling and committor learning and appears to achieve high sampling efficiency in terms of finding transition paths and transition states (e.g., Horvath et al., DOI 10.1073/pnas.2506516122).

**Response**: We are grateful to the Referee for this constructive suggestion. However, the application of these methodologies to the highly complex systems that most clearly illustrate the advantages of Gen-COMPAS is beyond the scope of the present study, as it would require substantial additional computational resources. Nevertheless, to acknowledge the point raised by the Referee, we have added the following text in the introduction:

(...) Deep learning techniques based on trasition path theory (TPT)[7] recently allowed the study of large conformational transitions[25].

We have added the new reference 25:

[25] Horvath, F. et al. STIM1 transmembrane helix dimerization captured by AI-guided transition path sampling. *Proc. Natl. Acad. Sci. U.S.A.* 122, e2506516122 (2025).

Minor:

(8) As a minor historical note, transition path sampling with targeted initialization has been used to sample complete transition paths between the inward and outward access states of a membrane transporter (Okazaki et al., DOI 10.1038/s41467-019-09739-0) similar to AAC here.

**Response**: We thank the Referee for bringing this reference to our attention. We have incorporated it into the revised manuscript as follows:

(...), while bypassing extensive MD simulations previously required to infer transition pathways in membrane transporters[56].

We have added the new reference 56:

[56] Okazaki, K.-i. *et al.* Mechanism of the electroneutral sodium/proton antiporter PaNhaP from transition-path shooting. *Nat. Commun.* 10, 1742 (2019).

(9) As a minor technical point, do the Z-matrices include all atoms of the protein or only a subset (e.g., backbone)? How does this impact the TMD calculations? I can see advantages (e.g., by getting side chain rotamer resolution) and disadvantages (e.g., by TMD becoming trapped or driven into high-energy structures).

Response: We are grateful to the reviewer for raising this point. The set of atoms used to define the internal coordinates (Z-matrix) and the corresponding targeted MD (TMD) restraints varies depending on the size of the system and the nature of the conformational transition under investigation. For the small benchmark peptides (NANMA and trialanine), we used all heavy atoms to fully resolve the local conformational changes associated with backbone and side-chain rearrangements.

For the Trp-cage mini-protein, we used all backbone atoms together with side-chain carbon atoms (e.g., $C_\alpha$, $C_\beta$, etc.), which provides sufficient structural resolution while maintaining stable and efficient TMD steering during the folding transition. For the larger biomolecular systems (RBP, AAC, $V_o$ domain, and nAChR), we instead used $C_\alpha$ atoms for the protein along with a small set of heavy atoms for the ligands when present. This choice balances structural resolution with computational tractability for large systems and avoids excessive restraints that could otherwise introduce artificial distortions during TMD.

Importantly, because the purpose of TMD in the Gen-COMPAS workflow is only to guide the system toward physically plausible conformations near generated target structures, the detailed relaxation of side chains and other degrees of freedom occurs subsequently during the unbiased MD shooting trajectories. The exact atom selections used for each system in the TMD are reported in Table S4 of the SI, and the atom selections used for the Z-matrix featurization are provided in Table S5 of the SI.

# Response to the comments of the referees of Nature manuscript 2025-10-28431B

We thank the referees for their careful evaluation of our revised manuscript, and for their very constructive comments. We are grateful to the referees for recognizing the substantial additional work, the newly added challenging applications, and the improved presentation of the Gen-COMPAS workflow. We are encouraged by the overall positive assessment and understand that the remaining concerns primarily relate to validation, possible hysteresis in targeted molecular dynamics (TMD)-generated configurations, free-energy convergence, and practical parameter guidance. We address these points below, and have revised the manuscript and the Supporting Information (SI) accordingly.

## Response to Referee 1

I am impressed by the authors' thorough response to the reviewers issues with the previous version of the manuscript. In particular, the newly added applications go a long ways towards demonstrating generalizability of the approach.

**Response**: We thank the Referee for this positive assessment of our revised manuscript. We are pleased that the additional applications helped demonstrate the generalizability of the Gen-COMPAS workflow.

I still want to bring up the question of novelty. I appreciate the authors' point that while the methods employed already exist, it's their combination that is novel. However, I take some issue with this statement in the response: "This shift is, in our view, substantive. Existing approaches typically rely either on massive data generation through long-timescale simulations or on carefully engineered collective variables and prior mechanistic assumptions".

and the corresponding statement in the manuscript: "Accordingly, a framework that can autonomously construct thermodynamically and kinetically consistent transition landscapes from minimal prior information remains lacking notwithstanding substantial progress at the interface of machine learning (ML) and molecular simulation".

This field is moving extremely quickly, and what may have been true even a year ago is no longer. There are other groups out there developing "physics-informed" and/or "MD-guided" approaches as well. I think the authors could do a little more to look into this and cite a few examples, which would also help to validate the authors' choice to go in this direction.

**Response**: We thank the Referee for this important suggestion. We agree that the field is moving at a very rapid pace, and that our previous wording was too broad. We have revised our statement of novelty, and now explicitly acknowledge recent physics-informed, committor-coupled, and MD-guided approaches. These additions clarify the relationship between our method and existing approaches, while highlighting the specific advances introduced in this work.

Specifically, we have revised the manuscript as follows:

In the Introduction, it read:

Generative models, including Boltzmann generators[30] and recent frameworks like MDGen[31] and BioEmu[32] aim to produce equilibrium structures directly through learned latent-variable representations. However, these models require extensive pretraining datasets, often comprising hundreds of milliseconds of prior MD simulation. For new rare events, such intermediate structural data are intrinsically unavailable. Because configurations are generated within learned latent spaces rather

than under the explicit molecular Hamiltonian, these models remain susceptible to mode imbalance and structural artifacts, compromising strict thermodynamic and kinetic consistency. Accordingly, a framework that can autonomously construct thermodynamically and kinetically consistent transition landscapes from minimal prior information remains lacking notwithstanding substantial progress at the interface of machine learning (ML) and molecular simulation.

and it now reads:

Generative and MD-guided models, including Boltzmann generators[30], MDGen[31], BioEmu[32], force-free trajectory predictors[33], and recent diffusion/action-based transition-path samplers[34,35], have rapidly expanded the scope of molecular simulation by learning equilibrium ensembles, surrogate dynamics, or transition-path proposals. This progress also clarifies the need for approaches that combine data-driven proposal with explicit-Hamiltonian dynamics. Many such models still depend on extensive pretraining or learned/bias-induced dynamics, and, for a new rare event, intermediate transition data are precisely what is unavailable. Thus, from a few structures alone, obtaining both kinetic observables, such as committors, transition states, and dynamic pathways, and thermodynamic observables, such as reweighted FELs, remains challenging. Either class of observables remains demanding to capture: Kinetic characterization often requires extensive transition statistics, whereas thermodynamic reconstruction requires sufficient equilibrium sampling.

These difficulties become especially acute for large, heterogeneous biomolecular assemblies, where membrane coupling, ligand binding, domain rearrangements, and multiple slow degrees of freedom make exhaustive transition sampling, or direct learned-force biasing, difficult to apply robustly. In practice, even a semi-quantitative mechanistic picture can still require substantial computational resources and considerable human intervention.

Finally, I would like for the authors to add a section (probably to the SI and to the README on the github) on (hyper)parameter tuning. At many stages, parameters have to be chosen that will affect the success of the methodology. These range from the TMD force constant to the VCN lag time to the diffusion model dimensions, and more. How sensitive is the method to these? What recommendations can you offer for choosing them?

**Response**: We agree with the Referee that practical guidance on parameter choices is key for reproducibility and for applying the method to qualitatively new problems. We have, therefore, added a parameter-guidance section to the Supporting Information, and expanded the GitHub README to include the same practical recommendations.

We emphasize that, in our applications, most machine-learning hyperparameters were not tuned separately for each system. The generative model architecture, training protocol, optimizer settings, and diffusion-related hyperparameters were kept essentially fixed across the complex systems studied here, with the batch size being the parameter adjusted when required by GPU-memory limitations and the noise the parameter for adjusting the exploration for the generative sampling. The exact parameter values used in each case are reported in the SI tables (Tables. S3–S5). Similarly, the variational committor network (VCN) time lag is not a strongly sensitive parameter within a broad physically reasonable range, as was also observed in our previous work. Training is affected only when the time lag is chosen at pathological limits, either too short to remove non-Markovian vibrational correlations, or too long, such that transition-region correlations have largely decayed.

The parameters that require the most careful attention are the TMD bias strength and the noise level used during generative sampling. The former was discussed in detail in the previous revision, and is now connected explicitly to the TMD-parameter scans and structural diagnostics reported in the SI. For the latter, we now recommend a noise range of approximately 0–50. Larger values can be useful in early iterations, especially when only a few short trajectories are available, and additional structural

diversity is needed to initialize intermediate-state sampling. Once additional transition data have been accumulated, values in the range of 0–2 are usually sufficient. Increasing the noise generates structures lying far from the accumulated data distribution, and, thereby, enhances diversity, whereas lowering it keeps generated structures closer to known configurations at the expense of reducing exploration. These recommendations, together with the precise values used in this work, have been added to the SI and to the GitHub README (see the newly added section entitled 2.8.1 Practical guidance for parameter tuning).

A few typos I noticed: Page 1: "LLaboratoire" - Laboratoire Page 2: "trasition path theory" - transition Page 4: "Gen-COPMAS" - Gen-COMPAS Page 11: "probablistic" - probabilistic Page 15: "consequentially sampled" - subsequently? Page 16: "pore activating" - activation

**Response**: We thank the Referee for pointing out these typographical errors. We have corrected all of them in the revised manuscript.

I briefly looked over the code. I believe it does include sufficient instructions for others to run it, although I did not try myself. As I noted in my review, it would be helpful to have more guidance on parameter tuning.

**Response**: We thank the Referee for checking the code repository. As noted above, we have expanded the SI guidance on parameter tuning, and have mirrored the same practical recommendations in the README in the repository update.

## Response to Referee 2

I acknowledge that the authors have undertaken a substantial amount of work for these revisions, and I appreciate their efforts. However, several crucial issues remain unresolved, and in my view these must be carefully addressed before I could recommend this manuscript for publication in any journal, let alone Nature. I outline these concerns below.

**Response**: We thank the Referee for acknowledging the substantial revisions of our manuscript, and for clearly stating their remaining concerns. In the revised SI, we have addressed these concerns in four concrete ways.

First, we have clarified that the raw diffusion-generated structures were used for the stringent tests presented in Fig. S7 and Fig. S8, and we have added new committor-validation data for NANMA, including aimless shooting from TMD-extracted structures, and a sampled-committor version of the bidirectional TMD hysteresis test. Second, we have endeavored at the behest of the Referee an extensive validation of our Trp-cage committor predictions based on aimless shooting. Third, we have also added a layered validation consisting of an independent consistency check on Trp-cage, using the long, unbiased D. E. Shaw Research (DESRES) trajectory. Fourth, we have clarified the scope of the novelty claim, namely Gen-COMPAS is not merely intended to be yet another method for small-peptide folding benchmarks, such as chignolin, but a workflow for exploring transition pathways in biologically realistic systems that are far larger, involving molecular processes strongly coupled to very slow degrees of freedom, including the mitochondrial ADP/ATP carrier (AAC) inward-facing/outward-facing conformational transitions, the ATPase $V_o$-domain rotation, and the asymmetric, muscle-type nicotinic acetylcholine receptor (nAChR) activation induced by dual ACh binding.

A number of points also warrant clarification. First, we could not locate two of the references quoted by the Referee, namely Qiu et al. PNAS 2006 and Ray et al. Nat. Comput. Sci. 2024. We suspect that these citations might in fact refer to Juraszek and Bolhuis, PNAS, 2006, and Kang et al. Nat. Comput. Sci. 2024, respectively, but could not ascertain it. We hope that the Referee will agree that in the absence of the correct references, assessing that a fair comparison was made against the suggested studies becomes impossible.

Second, a transition-path duration, or transition-path-sampling (TPS) path length, is not an estimate of the cost of an unbiased committor shooting test. The estimate of the Referee based on a few-nanosecond Trp-cage transition-path time is, therefore, not applicable to equilibrium committor validation. In our DESRES first-arrival analysis, separatrix-region configurations require commitment times that are roughly two orders of magnitude longer, that is, on the order of hundreds of nanoseconds rather than a handful of nanoseconds. Thus, the suggested $\sim 30$ $\mu$s estimate appears to underestimate the required amount of sampling.

Nevertheless, leveraging our entire computational resources, we performed the aimless unbiased shooting validation requested by the Referee for Trp-cage. Specifically, we selected 100 configurations, the Gen-COMPAS-predicted committor values of which span the full interval $q_{\mathrm{pred}} \in [0, 1]$. These configurations were taken from an independent Gen-COMPAS run that was not used in the training (Run 2 in Fig. S17 in the SI). From each configuration, we launched 20 independent unbiased trajectories, each propagated for up to 50 ns, or until it first reached basin $A$, or basin $B$. This validation effort corresponds to a maximum aggregate sampling time of, $100 \times 20 \times 50$ ns $= 100$ $\mu$s, which already exceeds the Referee's estimate, and, in practice, required over 400 NVIDIA L40s GPU-days. Most importantly, however substantive, this calculation should not be regarded as an exact asymptotic committor validation. Because trajectories are terminated after a finite time, configurations with long commitment times introduce finite-horizon/censoring bias. The calculation, therefore, provides a stringent, albeit still time-limited shooting validation, rather than a perfectly unbiased estimate of the true equilibrium committor.

To further support our committor predictions, we performed a series of independent checks based on a long DESRES trajectory. It is noteworthy that the DESRES data are completely independent from the Gen-COMPAS training data, and were generated with different force-field parameters. Nevertheless, the Gen-COMPAS-predicted committor map, $q_{\text{pred}}$, agrees closely with the DESRES first-arrival committor map in $\text{RMSD}_{\text{CA}}$–$\text{RMSD}_{\text{hx}}$ space, with minimal deviation. In addition, first-arrival analysis from the DESRES frames, with $q_{\text{pred}} \in [0.45, 0.55]$ according to the Gen-COMPAS predictions, gives nearly equal commitment probabilities to the two basins. For $n = 2028$ resolved frames, $\langle q_{\text{pred}} \rangle = 0.4998$ with 95% CI $[0.4985, 0.5010]$, while $q_{\text{sample}} = 0.5335$ with 95% Wilson CI $[0.5118, 0.5552]$ ($N_A = 946$, $N_B = 1082$). These results provide direct evidence that the learned Trp-cage separatrix captures transferable dynamical structure rather than a force-field-specific artifact.

As regards novelty, we respectfully disagree with the Referee that performance on chignolin-scale systems by other methodologies is a fair basis for dismissing the merits of Gen-COMPAS. Chignolin is a ten-residue $\beta$-hairpin, and, while remaining an important benchmark for method development, it certainly does not compare to the systems targeted here. Gen-COMPAS has been applied to conformational transitions involving hundreds to thousands of residues, and strongly coupled collective motions, like in the AAC inward-facing/outward-facing switch, the $V_o$-domain rotation, and the asymmetric nAChR activation upon dual ACh binding. These systems involve many correlated fast-slow degrees of freedom, and are exponentially more difficult to explore than a small $\beta$-hairpin. The true relevant comparison is, therefore, not only whether a method performs well on chignolin-like systems, but also whether it can generate physically meaningful transition pathways and committor-informed transition-state ensembles in substantially larger biomolecular objects.

We have also revised the manuscript and SI to clarify the interpretation of the free-energy landscapes (FELs). These landscapes are useful mechanistic projections of the sampled regions, but they should not be interpreted as fully converged and accurate global equilibrium free-energy surfaces. We would, nevertheless, like to stress in light of a recent cross-method computational investigation across biological complexity that, given empirically chosen collective variables, apparent convergence of enhanced-sampling free-energy calculations is not a guarantee of correctness (Kang et al. Nat. Comm. 2026, in press).

We highlight hereafter the changes brought to the SI:

(i) In 2.6 Committor Validation Analysis, we moved the committor-validation material out of *Convergence of Gen-COMPAS*, and made it an independent subsection.

(ii) In 2.6.1 Committor validation for NANMA, we separated the raw-diffusion shooting stress test from the workflow-relevant TMD-extracted shooting validation, we added Fig. R1, and we defined both predicted and sampled bidirectional $\delta q$ tests.

(iii) In 2.6.2 Committor validation for Trp-cage, we added the aimless shooting results, the DESRES committor-map comparison, the empirical separatrix-shooting analysis, and the TMD $\delta q$ tests.

(iv) In 2.4 Free-energy estimation using RiteWeight and Convergence and uncertainty analysis of FEL estimations, we clarified the distinction between useful mechanistic FEL estimates and strict global equilibrium convergence.

# Quality of the committor models The committor validation presented in Figures S7 and S8 is notably noisy, indicating sub-par committor learning compared to recent studies from other groups. For instance, Bolhuis and coworkers (arXiv:2507.04052) achieve substantially better predicted-vs-sampled correlation for a host–guest system that is arguably more challenging than vacuum NANMA isomerization (see their Fig. 17). The lower panel of Figure S8 shows that the validated committor values for configurations predicted at $q \approx 0.5$ are distributed broadly between 0 and 1, indicating that the model is nearly unpredictive in the transition region.

**Response**: We thank the Referee for this comment. We agree that the original presentation of the NANMA validation was somewhat insufficiently clear, but we respectfully disagree that Figs. S7-S8 show that the final committor model is "nearly unpredictive in the transition region". The key point is that Figs. S7-S8 are not a validation of the physical structures used in the Gen-COMPAS production workflow. They are a deliberately stringent stress test, wherein shooting was initiated directly from raw diffusion-generated structures, including intermediate diffusion outputs, to test the generative component before MD relaxation. We have now made this distinction clear in SI Section 2.6. We have also made the two other more relevant validations explicit, namely (i) a comparison against real unbiased NANMA simulations in Fig. S10 (Fig. S9 in the previous SI), and (ii) a new shooting validation using structures extracted from TMD trajectories, shown in Fig. R1. The latter directly tests the workflow-relevant transition-region structures, while avoiding data leakage, because the TMD trajectories themselves were not used for training. Specifically, the revised SI features a new paragraph entitled ***Direct shooting from raw diffusion-generated structures. . .***, and we clarify that these configurations are raw diffusion outputs, not structures to which Gen-COMPAS assigns direct kinetic or thermodynamic meaning. This distinction is crucial, because raw diffusion-generated structures may contain local geometric inconsistencies or fast-relaxing degrees of freedom that have not yet been propagated under the molecular-mechanics force field. The TMD step is precisely included in Gen-COMPAS to connect generated candidate geometries to physically propagated configurations.

To evaluate the predictive power on physically meaningful structures, we first compared the learned committor against an independent, unbiased NANMA trajectory in Fig. S10 (Fig. S9 in the previous SI). This test uses real dynamics outside the Gen-COMPAS training data, and, therefore, directly addresses whether or not the learned committor generalizes beyond the training workflow.

In addition, to address the concern of the Referee regarding transition-region structures, we have added a new shooting validation from TMD-extracted configurations. We selected structures from TMD trajectories, excluding the final TMD frame, according to predicted iso-committor slices centered at $q_{\text{pred}} \approx 0.1$, 0.3, 0.5, 0.7, and 0.9, and subsequently performed unbiased shooting from those selected configurations. As shown in Fig. R1, the committor-sample means are 0.174, 0.359, 0.485, 0.617, and 0.799, respectively. The pointwise predicted-versus-sampled comparison gives a Pearson correlation coefficient of 0.848. Specifically, the shooting estimates in the $q_{\text{pred}} \approx 0.5$ slice obeys a single modal distribution centered at $q_{\text{sample}} = 0.5$.

This TMD-extracted validation also avoids data leakage. The committor model was not trained on the TMD trajectories. It is important to note that the training set consisted only of unbiased trajectories initiated from the final TMD configurations, which were deliberately excluded from the new TMD-extracted validation. Thus, the model did not see the validation structures during training.

We also note that the comparison with arXiv:2507.04052 (already published in R. S. Breebaart et al., Phys. Rev. Lett. 2026) is not a like-for-like assessment of committor-model quality. The relevant comparison should be made to the panel showing the actual committor, rather than the logit committor. In that representation, the region near $q \approx 0.4$–$0.6$ is also broad and visibly undersampled. It is not surprising as committor values are well-known to be extremely difficult to sample or validate. Therefore, the claim that the cited work demonstrates substantially better predicted-versus-sampled agreement is not supported in the simple way implied by the comment of the Referee.

We have revised SI Section 2.6 accordingly. The new narrative now distinguishes raw diffusion-generated stress tests from validations on physically propagated structures, explains the difference between Figs. S7–S8, Fig. S9 (now Fig. S10), and Fig. R1 (now Fig. S9 in SI), and reports the new TMD-extracted shooting results. These results do not remove all statistical uncertainty, which is

expected for sampled-committor tests, but we cannot overstate that they certainly do not support the conclusion that the final committor model is "nearly unpredictive in the transition region".

Furthermore, the contrast with Figure S9 is striking: that plot looks considerably cleaner. A well-trained committor model should yield comparably good agreement between learned and sampled committors regardless of whether the source structures originate from an enhanced sampling campaign or an unbiased trajectory. The fact that this is clearly not the case admits at least two interpretations: 1. Sampling problem. The TMD pipeline may produce structures that lie far from physical unbiased transition paths in phase space, leading to noisier correlations. TMD is known to often result in incomplete transitions along slow orthogonal degrees of freedom (hysteresis). Although the details of how the committor validation set is constructed are not entirely clear to me, if it includes TMD-generated structures initiated from both end-states, the resulting frames would be biased toward the basin neighborhoods and potentially occupy regions far from unbiased transition paths. This interpretation is consistent with the clearly bimodal distribution of sampled committors at q = 0.5 in Figure S7, rather than the expected unimodal distribution centered near 0.5. If correct, this would represent a fundamental limitation of the diffusion-model-guided TMD workflow—one that can be expected to worsen for more complex systems. I suggest a concrete test for this below. 2. Learning problem. The committor model may not have learned sufficiently general representations across the relevant regions of phase space. That such difficulties already arise for one of the simplest systems studied raises serious concerns about the quality of the committor models for the more complex systems.

**Response**: We thank the Referee for pointing out that the contrast between Figs. S7–S8 and Fig. S9 (now Fig. S10) requires a clearer explanation. The validation sets in these figures refer to different objects. The original Figs. S7 and S8 tested direct shooting from raw diffusion-generated structures before atomistic propagation through the TMD protocol. Fig. S9 (now Fig. S10) probes an independent, long, unbiased trajectory. These figures, therefore, do not reflect equivalent validation ensembles.

This distinction warrants the introduction of the TMD step. If TMD-induced hysteresis were the dominant source of the discrepancy, one would expect structures extracted from TMD trajectories to reflect degraded sampled-committor consistency. Instead, the new validation in Fig. R1 shows that shooting from TMD-driven structures yields substantially more consistent committor behavior than direct shooting from raw generated structures.

We agree that TMD can in principle introduce nonequilibrium or hysteretic artifacts, and that a learned committor alone may not fully reflect them. For this reason, we have added a direct sampled-committor test, as per the request of the Referee. The sampled-committor differences between the two targeting directions are centered near zero across the predicted iso-committor slices. This result does not prove that RMSD-based TMD is universally artifact-free, and we would certainly refrain from making such a claim. It does, however, show that, for NANMA and Trp-cage (see Fig. R4) under the present protocol, endpoints from opposite TMD steering directions are kinetically consistent within the statistical uncertainty of the shooting calculations.

As an aside, I respectfully disagree with the authors' assertion that committor validation on the more complex systems would not have been computationally feasible: "We believe that the committor validation test is not computationally addressable for most of the systems" The barrier crossing time of Trp-cage, which represents an intermediate level of complexity among the systems studied, is approximately 3 ns (Qiu et al., PNAS 2006). A reasonable committor validation—say, 100 points with 100 shooting trajectories each—would require on the order of 30 $\mu$s of aggregate sampling. On a modern GPU node, this amounts to only a few days of wall-clock time, which should be well within reach for a manuscript submitted to a journal of this caliber.

**Response**:

We thank the Referee for raising this point. We have reexamined the estimate in question carefully, and we agree that direct shooting validation is, indeed, valuable. However, we respectfully disagree with the premise that a full-scale Trp-cage committor validation would require only $\sim$ 30 $\mu$s of aggregate sampling. As noted previously, the cited "Qiu et al., PNAS 2006" reference appears to be incorrect. Qiu et al. reported experimental Trp-cage folding kinetics in J. Am. Chem. Soc. 2002, with a folding time of approximately 4 $\mu$s. We suspect that the relevant reference hinted by the Referee is the 2006 Proc. Natl. Acad. Sci. USA paper by Juraszek and Bolhuis. The $\sim$ 3 ns timescale in that work refers to transition-path/barrier-crossing segments sampled by TPS—not to the average commitment time required for unbiased committor shooting from arbitrary separatrix configurations. A second important distinction is the definition of the metastable basins. In the aforementioned 2006 TPS study, relatively broad basin definitions were used to facilitate TPS. In stark contrast, here, we use substantially narrower basin definitions, because our goal is to characterize transitions between a true random coil and the fully folded state. As a consequence, unbiased shooting trajectories initiated near the separatrix can require much longer times to commit to either basin.

***Re-examination of the cost of unbiased shooting validation.*** The equilibrium committor is an infinite-horizon first-hitting probability,

$$q(\mathbf{x}) = \mathbb{P}(\tau_B(\mathbf{x}) < \tau_A(\mathbf{x})),$$

where $\tau_S(\mathbf{x})$ is the first hitting time of basin $S$ for initial condition $\mathbf{x}$. Therefore, an asymptotically exact unbiased shooting estimate requires each shooting trajectory to be propagated until it reaches either basin $A$, or basin $B$. Using our narrower unfolded/folded basin definitions, the DESRES first-arrival analysis provides a more direct estimate of this cost. As shown in Fig. R2 C,D, configurations near the separatrix require about 1,600 saved frames on average before first arrival to either basin. With the DESRES saving interval of 0.2 ns, this corresponds to an average commitment time of roughly 320 ns. Thus, a shooting calculation with 100 configurations and 100 shots per configuration would require an aggregate simulation time on the order of milliseconds, rather than the $\sim$ 30 $\mu$s estimate of the Referee from the barrier-crossing duration. This estimate also neglects the substantial wall-clock cost of running many short, independent GPU-based simulations, including initialization, I/O, and scheduling overhead.

***Aimless shooting validation for Trp-cage.*** Nevertheless, in response to the Referee's concern, we performed a direct aimless-shooting validation for Trp-cage. Specifically, we selected 100 configurations spanning the full range of Gen-COMPAS-predicted committor values, $q_{\text{pred}} \in [0, 1]$, from an independent Gen-COMPAS run that was not used for training (Run 2 in Fig. S17). From each configuration, we launched 20 independent unbiased shooting trajectories, each propagated for up to 50 ns, or until it first reached either basin $A$, or basin $B$. This corresponds to a maximum aggregate sampling effort of $100 \times 20 \times 50$ ns $= 100$ $\mu$s, which already exceeds the estimated cost of the Referee, requiring substantive computational resources.

Because each trajectory was terminated after a finite time, this calculation should be interpreted as a finite-horizon shooting test rather than an exact infinite-time committor validation. To reduce the influence of finite-time censoring, we retained only initial configurations for which more than 50% of the launched trajectories committed to either basin within the simulation window. This resulted in $n = 91$ validation points. The sampled committor, $q_{\text{sample}}$, obtained from the fraction of committed trajectories reaching basin $B$, strongly agrees with the Gen-COMPAS prediction over the full committor range, with a Pearson correlation of $r = 0.898$, and $R^2 = 0.806$ (Fig. R3). The error bars denote a 90% Wilson confidence derived from the aimless-shooting strategy. These results provide direct evidence that the Gen-COMPAS-predicted committor is quantitatively consistent with independent, unbiased shooting outcomes, while also illustrating why an asymptotically

exact committor validation with longer trajectories and many more shots per configuration remains substantially more prohibitive than the optimistic estimate of the Referee.

***Checks of Gen-COMPAS committor on the DESRES trajectory.*** In lieu of relying solely on finite-time aimless shooting as committor validation, or pursuing computationally infeasible exhaustive unbiased shooting, we performed a series of DESRES-based checks of the Gen-COMPAS Trp-cage committor.

In SI Section 2.6.2, we first compare the Gen-COMPAS-predicted committor map, $q_{\text{pred}}$, evaluated on DESRES configurations, with a DESRES first-arrival committor map in the same $\text{RMSD}_{\text{CA}}$–$\text{RMSD}_{\text{hx}}$ space. The Gen-COMPAS-predicted separatrix coincides with the DESRES intermediate-committor region, with minimal deviation and no displacement toward either stable basin.

We then performed empirical shooting from the Gen-COMPAS-predicted separatrix using the unbiased DESRES trajectory. Specifically, we selected DESRES frames satisfying $q_{\text{pred}} \in [0.45, 0.55]$, and followed the trajectory forward until first arrival to basin $A$, or basin $B$. For $n = 2028$ resolved frames, the mean predicted committor is

$$\langle q_{\text{pred}} \rangle = 0.4998 \quad \text{with 95\% CI } [0.4985, 0.5010],$$

and the empirical first-arrival committor is

$$q_{\text{sample}} = 0.5335 \quad \text{with 95\% Wilson CI } [0.5118, 0.5552],$$

with $N_A = 946$ and $N_B = 1082$. Thus, configurations predicted by Gen-COMPAS to lie on the $q = 0.5$ separatrix commit to the two basins with nearly equal probability in an entirely independent unbiased trajectory. The absolute deviation from the ideal value is only 0.0335, which is minimal given the temporal correlations in the DESRES trajectory, the finite number of statistically independent transition-region samples, and the different force-field parameters.

Taken together, these results provide compelling evidence for the predictive quality of the Trp-cage committor. The agreement is particularly remarkable because Gen-COMPAS and DESRES use different force-field parameterizations, suggesting that the learned separatrix captures transferable dynamical structure rather than force-field-specific artifacts.

#Possible hysteresis due to TMD simulations As discussed above, there are substantive concerns regarding hysteresis introduced by TMD. The analysis presented in Section 2.6.2 of the supplement, which is intended to demonstrate that steering along RMSD does not introduce artifacts, is not fully convincing. It shows only that bidirectional steered trajectories approach configurations with similar predicted committor values. However, the poor predicted-vs-sampled validation in Figures S7 and S8 suggests that the predicted committor alone is not a sufficiently sensitive metric for detecting such hysteresis. An alternative interpretation, considering Figures S7, S8, S9, and S11 together, is that RMSD-based TMD does exhibit a hysteresis problem, but that the learned committor model fails to capture it. This would explain why the discrepancy appears only when comparing to sampled committors, and why the predicted-vs-predicted validation along an unbiased path looks considerably better. A straightforward test would be to repeat the analysis of Figure S11 using sampled committors rather than predicted ones. If this substantially degrades the agreement, it would indicate that RMSD-based TMD is not suitable for this workflow. If the agreement remains as clean as in the current Figure S11, then TMD may indeed be appropriate, but the authors would need to investigate—including on Trp-cage—other explanations for the divergence in committor validation quality between Gen-COMPAS-generated structures and unbiased paths.

**Response**: We acknowledge that predicted committor values alone should be treated as a diagnostic rather than conclusive evidence that TMD introduces no hysteresis or steering artifacts. We, there-

fore, performed the suggested aimless-shooting sampled-committor test for NANMA, and reported the result in Fig. R1H–I.

Let $x_A^{\mathrm{TMD}}$ and $x_B^{\mathrm{TMD}}$ denote the endpoint structures obtained by steering from the two opposite directions toward the same generated target. We computed both the neural-network-predicted and the directly sampled committor differences,

$$\delta q_{\mathrm{pred}} = q_{\mathrm{pred}}(\mathbf{x}_A^{\mathrm{TMD}}) - q_{\mathrm{pred}}(\mathbf{x}_B^{\mathrm{TMD}}), \tag{R1}$$

and

$$\delta q_{\mathrm{sample}} = q_{\mathrm{sample}}(\mathbf{x}_A^{\mathrm{TMD}}) - q_{\mathrm{sample}}(\mathbf{x}_B^{\mathrm{TMD}}). \tag{R2}$$

If the predicted committor were concealing a TMD hysteresis, one would expect the $\delta q_{\mathrm{pred}}$ distribution to remain narrow while the directly sampled $\delta q_{\mathrm{sample}}$ distribution would become broad, shifted, or otherwise degraded relative to the predicted-committor result. This is not what we typically observed. As shown in Fig. R1H–I, both the predicted- and sampled-committor $\delta q$ distributions are in fact centered near zero across the predicted committor slices. Thus, the agreement does not substantially degrade when sampled committors are used, indicating that for NANMA, from-$A$ and from-$B$ TMD steering produces structures with consistent commitment probabilities within statistical uncertainty.

For Trp-cage, as suggested by the Referee, we examined whether the TMD-based sampling protocol introduces a systematic perturbation into the committor estimate. Toward this end, we randomly selected ten pairs of TMD endpoint structures (with each pair having the same target, starting from $A$ and from $B$) and performed both Gen-COMPAS committor prediction and aimless-shooting analysis (20 trajectories each, lasting 50–ns, or until commitment) from these structures. The resulting distributions of $\delta q$ from both tests exhibit narrow distribution and median values around zero. This result indicates that, under our TMD setup and protocol, the sampling procedure does not introduce any perceptible systematic shift in the committor, and that any additional noise from TMD is minimal compared with the scale of the committor variations analyzed here. These results are gathered in Fig. R4 and have been added to Fig. S13B in the SI.

#Poor convergence of free energy profiles Figure S6 presents the RMSE of the free energy landscape as a function of simulation effort across several systems and simulation campaigns. I thank the authors for including this analysis. However, it is concerning that, with the exception of trialanine—a system that is trivial to converge—none of the other systems exhibits a converged free energy surface. I also wish to emphasize that this test only probes statistical convergence, and says nothing about the accuracy of the resulting landscape. I appreciate that some of these systems are inherently difficult to simulate, but one should be transparent about this limitation rather than presenting these profiles as converged free energy surfaces. Strictly speaking, a free energy is only well-defined when the underlying statistical distribution has been sampled to convergence, which does not appear to be the case here.

**Response**: It is true that the root mean squared error (RMSE) analysis probes statistical stabilization relative to the available sampling, and does not per se prove thermodynamic accuracy of the underlying equilibrium distribution. This is also one of the bottlenecks for traditional, collective-variable (CV)-based enhanced-sampling approaches, and even for them, reaching accurate statistical convergence is extremely difficult. Here, our primary objective is not to showcase that we have perfectly accurate FELs projected onto a posteriori CVs, as we have already emphasized in the Discussion section of the manuscript.

Nevertheless, we have revised the manuscript and the SI to make this point even clearer. In the complex systems that we have examined, our results may not satisfy the strictest criteria for fully converged global equilibrium FELs, but the residual degradation relative to the final estimates falls

within the kcal/mol threshold to allow mechanistic interpretation, and the uncertainty on the basin topology, local barriers, and transition regions is bounded. We, therefore, continue to present the FELs as a powerful tool for mechanistic discovery, while distinguishing this use from a claim of unassailable global equilibrium convergence.

The original phrasing was:

The generated FELs represent a powerful tool for identifying key states and barriers, but for applications demanding the highest quantitative precision, they should serve as a well-informed starting point.

and the new phrasing is:

The reweighted FELs remain a powerful tool for identifying metastable basins, barriers, and transition regions. Although not every complex-system FEL satisfies the strictest criteria for global equilibrium convergence, the analysis of the FEL estimates reveals that the remnant fluctuations are sufficiently limited to allow mechanistic interpretation (see SI). We, therefore, use these FELs to support local and pathway-level observations, while absolute global $A$-to-$B$ free-energy differences should be viewed cautiously as estimates.

#Novelty Finally, regarding the question of novelty. In their rebuttal, the authors argue that: "The central contribution of this work is not the introduction of isolated methodological components, but the demonstration of a unified, end-to-end framework..." "This shift is, in our view, substantive...a new route to mechanistic insight in regimes that would otherwise be computationally inaccessible." I agree that the combination of existing elements into a unified workflow carries some novelty. However, integrative committor learning and sampling pipelines already exist, notably from the Parrinello and Hummer groups. These approaches, with various differences between them, can provide similar mechanistic insight to Gen-COMPAS, including transition mechanisms, committor surfaces, free energy landscapes, and kinetics. Importantly, these existing approaches are not dramatically less efficient computationally. While no direct comparison is available, the Parrinello group's 2024 study (Ray et al., Nat. Comput. Sci. 2024) investigated chignolin, which folds somewhat faster than Trp-cage but represents a comparable level of molecular complexity. Their protocol required approximately 800 ns of total sampling across 10 iterations, which is similar to the 594 ns the authors report for Trp-cage. Although the systems and methods differ, this comparison suggests that sub-microsecond committor-coupled sampling of transition mechanisms and free energy landscapes for fast-folding proteins is already achievable with current state-of-the-art methods. The additional integration of steering simulations toward the output of a diffusion model is distinct from these existing workflows and represents a genuine point of novelty—even though the diffusion models themselves are not new. However, this alone does not, in my assessment, constitute sufficient novelty for publication in a journal of this standing.

**Response**: We appreciate the discussion of the Referee on existing committor-coupled sampling frameworks. However, before addressing the methodological comparison, we must first respectfully correct a factual point. The Referee refers to "Ray et al. Nat. Comput. Sci. 2024" as the basis for their comparison. We were unfortunately unable to identify this publication. We can only surmise that the Referee in fact meant the related work of Kang et al. Nat. Comput. Sci. 2024. This distinction is, however, not merely bibliographic. It matters because the argument of the Referee about novelty and computational feasibility is built on a specific comparison to a specific method and benchmark system, and that comparison must be made accurately.

For clarity, we have never claimed that Gen-COMPAS was the first method in the committor field—or that of transition path theory (TPT), nor have we ever claimed that diffusion models, per se, constitute a new molecular-simulation methodology. The point of novelty is of a different nature:

Gen-COMPAS integrates diffusion-based seeding, MD-based propagation through TMD, unbiased shooting diagnostics, unsupervised committor learning, and thermodynamic/kinetic reconstruction, melded into a workflow designed to reduce dependence on predefined CVs, early sampled transition channels, and online committor-force biasing. The Referee will surely agree that each of these components represents a bottleneck in rare-event sampling. With our method, we endeavor to meet the challenge organizing them in a single framework. This distinction is absolutely central here, and it is not captured by simply comparing aggregate nanoseconds on small fast-folding mini-proteins.

First, the comparison to the Parrinello-group chignolin study is not like-for-like. Chignolin has been a useful benchmark in method development, but it remains a toy protein system compared with the biological objects tackled in this work. It is a small, ten-residue $\beta$-hairpin. Trp-cage is already more complex with its twenty amino acids, involving an $\alpha$-helix, a $3_{10}$-helix, hydrophobic-core packing, and partially folded or misfolded kinetic traps. More importantly, Trp-cage is not the unsurpassable horizon of our demonstration. In our manuscript, Gen-COMPAS is applied not only to small proteins, for which some benchmarking may be available in the literature, but also to substantially larger and more fragile biomolecular systems, including complex, integral membrane proteins and molecular machines, featuring membrane interactions and ligand-induced conformational transitions. The associated biological processes span regimes where the practical limitations of directly biasing a learned committor become severe. Without mentioning any of the other systems in the comment, the comparison, and, thus, the claim of "insufficient novelty" appears partial and unwarranted.

We would like to respectfully quote the Referee from their comments in the first round of reviews: "Regarding significance, this manuscript certainly constitutes a valuable addition to the growing body of sampling algorithms and should be of interest to the computational community. Nonetheless, I am not fully convinced that this new method offers the kind of groundbreaking or enabling novelty expected. The benchmark systems are interesting and challenging, but the presented results are not truly unique to this technique. None of the test cases appear to be beyond the reach of alternative, established methods. While the folding of a mini-protein is a standard benchmark, and the two larger, biologically significant systems demonstrate scalability, their conformational changes are relatively contained. These systems could likely be sampled using many existing methods. For acceptance in this journal, a method should ideally demonstrate a unique edge by enabling the study of a system or process that is demonstrably difficult or impossible to reproduce with alternative techniques".

We believe that the systems we are presenting in our manuscript, especially, the AAC conformational transition coupled with ADP swiveling and the nAChR asymmetric conformational transition coupled with dual ACh binding, are unique and demonstrably difficult to reproduce with any other methods currently available. The fact that we are able to capture reasonable mechanisms, estimate FELs on whatever post-selected CVs, and compare them directly with up-to-date experimental observations with merely hundreds of nanoseconds of simulation cost cannot be ignored and certainly does not qualify for "insufficient novelty".

Second, there is a fundamental methodological difference at play. In committor-biased, or logit-committor-biased dynamics, the learned model is not used merely as an analysis tool; it enters the equations of motion through a biasing potential and its corresponding forces. For neural-network committors, this necessarily involves nonlinear activation functions and high-dimensional gradients. In fragile systems such as membrane proteins, ligand-activated machines, or complexes with delicate inter-domain and protein-membrane interactions, such non-conservative forces can easily drive the system into unphysical distortions if the bias is too strong. Conversely, reducing the bias magnitude to avoid destabilizing the system renders convergence extremely difficult. This is not a minor implementation detail, but a practical bottleneck for applying force-driven committor-bias strategies to large, heterogeneous biomolecular assemblies.

Gen-COMPAS deliberately avoids making the learned committor an online biasing tool that must safely steer every atomistic degree of freedom during MD. Instead, diffusion-generated proposals are leveraged to broaden the exploration of candidate transition regions, and TMD is used to physically propagate atomistic configurations under controlled steering. The committor model is then used for diagnosis, organization, and thermodynamic/kinetic reconstruction, rather than as a high-gain, nonlinear biasing force acting continuously on a fragile system. This design choice is one of the reasons why Gen-COMPAS can be applied to complex systems for which direct committor-force-driven sampling would be highly problematic in practice.

Thus, the relevant comparison is not whether another method can obtain transition-state information for chignolin with sub-microsecond aggregate sampling. The relevant question is in fact whether the same strategy remains practical when the system contains a membrane, ligand-coupled motions, multiple metastable basins, fragile collective rearrangements, and reaction channels that may not be represented in the earliest sampling iterations. The present work demonstrates that Gen-COMPAS can generate physically reasonable reaction pathways and free-energy projections for such systems using only hundreds of nanoseconds of atomistic propagation. We contend that this capability is not established by any existing chignolin-scale benchmarks.

We have, therefore, revised the manuscript to position Gen-COMPAS more precisely. Existing committor-biased and shooting-informed methods are important and complementary, but they solve a different problem under different assumptions. Benchmarks on an already sampled small-protein ensemble does not imply practicality to never-sampled channels, fragile coupled degrees of freedom, or large membrane-associated molecular machines. Gen-COMPAS contributes a different route—it uses generative proposals to expand transition-region exploration, and then, validates and organizes these candidates through MD propagation and committor-based analysis, without relying primarily on direct online committor-force biasing. We believe this is a substantive methodological contribution, particularly because the manuscript demonstrates it beyond toy, fast-folding proteins, and in systems where conventional force-driven committor-bias workflows are not realistically expected to be stable or efficient.

The Introduction used to read:

Generative models, including Boltzmann generators[30] and recent frameworks like MDGen[31] and BioEmu[32] aim to produce equilibrium structures directly through learned latent-variable representations. However, these models require extensive pretraining datasets, often comprising hundreds of milliseconds of prior MD simulation. For new rare events, such intermediate structural data are intrinsically unavailable. Because configurations are generated within learned latent spaces rather than under the explicit molecular Hamiltonian, these models remain susceptible to mode imbalance and structural artifacts, compromising strict thermodynamic and kinetic consistency. Accordingly, a framework that can autonomously construct thermodynamically and kinetically consistent transition landscapes from minimal prior information remains lacking notwithstanding substantial progress at the interface of machine learning (ML) and molecular simulation.

and now reads:

Generative and MD-guided models, including Boltzmann generators[30], MDGen[31], BioEmu[32], force-free trajectory predictors[33], and recent diffusion/action-based transition-path samplers[34,35], have rapidly expanded the scope of molecular simulation by learning equilibrium ensembles, surrogate dynamics, or transition-path proposals. This progress also clarifies the need for approaches that combine data-driven proposal with explicit-Hamiltonian dynamics. Many such models still depend on extensive pretraining or learned/bias-induced dynamics, and, for a new rare event, intermediate transition data are precisely what is unavailable. Thus, from a few structures alone, obtaining both kinetic observables, such as committors, transition states, and dynamic pathways, and thermodynamic observables, such as reweighted FELs, remains challenging. Either class of observables

remains demanding to capture: Kinetic characterization often requires extensive transition statistics, whereas thermodynamic reconstruction requires sufficient equilibrium sampling.

These difficulties become especially acute for large, heterogeneous biomolecular assemblies, where membrane coupling, ligand binding, domain rearrangements, and multiple slow degrees of freedom make exhaustive transition sampling, or direct learned-force biasing, difficult to apply robustly. In practice, even a semi-quantitative mechanistic picture can still require substantial computational resources and considerable human intervention.

In summary, while the scope and ambition of this work are commendable, the issues outlined above—concerning committor model quality, potential TMD-induced hysteresis, free energy convergence, and the degree of novelty over existing methods—lead me to recommend that this manuscript not be accepted in its current form.

**Response**: We sincerely thank the Referee for the continued and detailed evaluation of our manuscript, and believe that the revised manuscript now answers the main technical concerns raised in their review. In particular, the different validations performed address explicitly three distinct issues that were previously conflated, namely (i) direct shooting from raw diffusion-generated configurations, (ii) sampled-committor validation of physically propagated TMD-selected structures, and (iii) hysteresis diagnostics based on bidirectional TMD endpoints. The new shooting and bidirectional sampled-committor simulations clearly demonstrate that the learned committor provides a meaningful organization of transition-region configurations, and that the TMD protocol does not introduce any detectable systematic hysteresis under our working conditions.

We have also clarified the interpretation of the Trp-cage validations and of the reconstructed FELs. The Trp-cage shooting and DESRES-based comparisons are presented as finite-horizon and semi-quantitative checks rather than as exhaustive, infinite-time committor calculations, while the FELs are now explicitly framed as mechanistic projections, the absolute free-energy differences of which should be interpreted with care. Our revisions are intended to avoid overstatement, while still documenting that the model-selected transition regions and pathway-level interpretations are dynamically consistent.

Finally, we respectfully reassert that Gen-COMPAS makes a substantive methodological contribution in the current research landscape. Its novelty does not lie in any single component taken individually, but in integrating generative proposal, controlled atomistic propagation, committor-based organization, and thermodynamic/kinetic reconstruction into a practical workflow for systems where exhaustive committor sampling or continuous committor-force biasing is difficult—not to say outright impossible. The revised manuscript now provides additional evidence lending support to this workflow, and we are grateful to the Referee for contributing to this improvement.

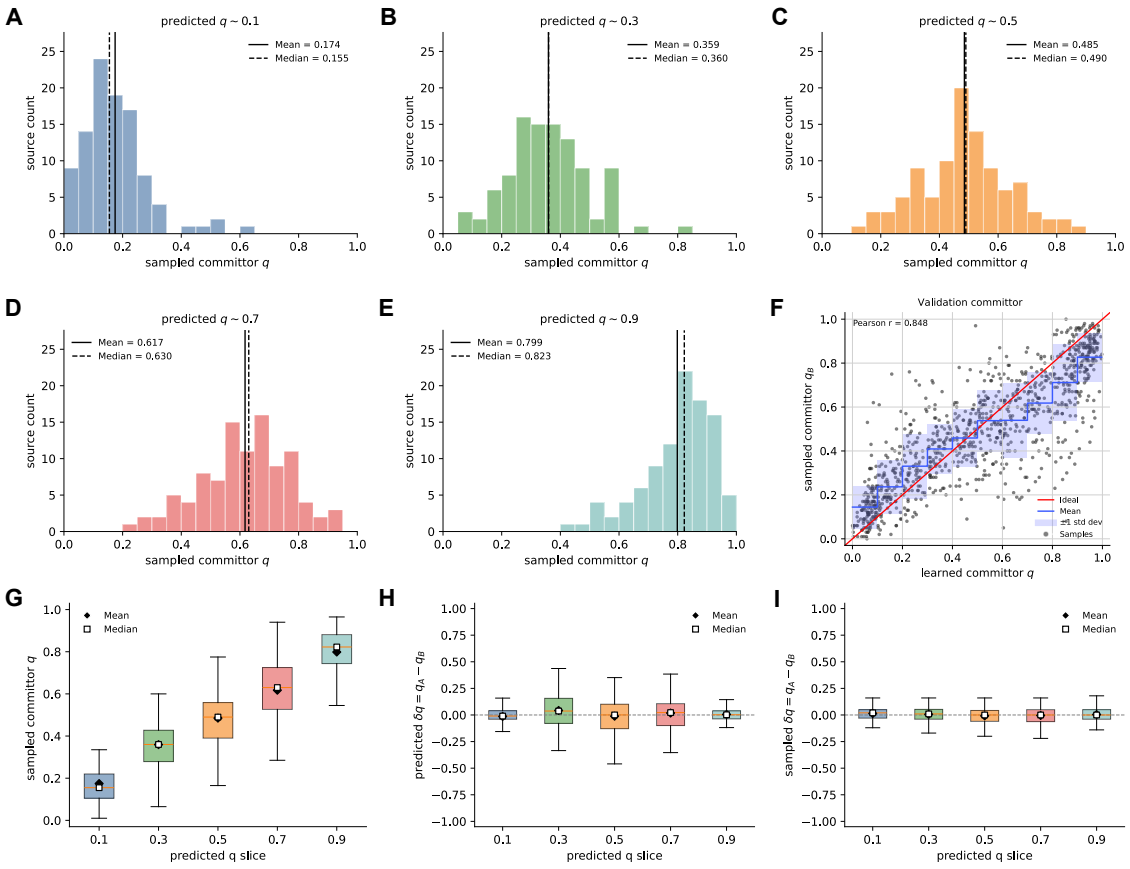


Figure R1: **Committor validation for NANMA using unbiased shooting trajectories.** (A–E) Distributions of sampled committor values for configurations selected from predicted committor slices centered at $q \approx 0.1$, 0.3, 0.5, 0.7, and 0.9, respectively, with configurations extracted from TMD trajectories. Solid and dashed vertical lines indicate the mean and median sampled committors. The corresponding means are 0.174, 0.359, 0.485, 0.617, and 0.799, showing that the sampled committors increase consistently with the predicted committor slices. (F) Pointwise comparison between the predicted committor and the sampled committor $p_B$ obtained from unbiased shooting. The red line denotes the ideal relation $p_B = q$, while the blue line and shaded region show the binned mean and standard deviation. The Pearson correlation coefficient is 0.848. (G) Box plots summarizing the sampled committor distributions for each predicted committor slice. Diamonds and open squares denote the mean and median, respectively. (H) Distribution of the bidirectional predicted committor difference, $\delta q = q_A - q_B$, across predicted committor slices. (I) Distribution of the bidirectional sampled committor difference, $\delta q$, across predicted committor slices. The distributions are centered near zero, indicating that the two steering directions yield consistent sampled committor estimates within statistical uncertainty.

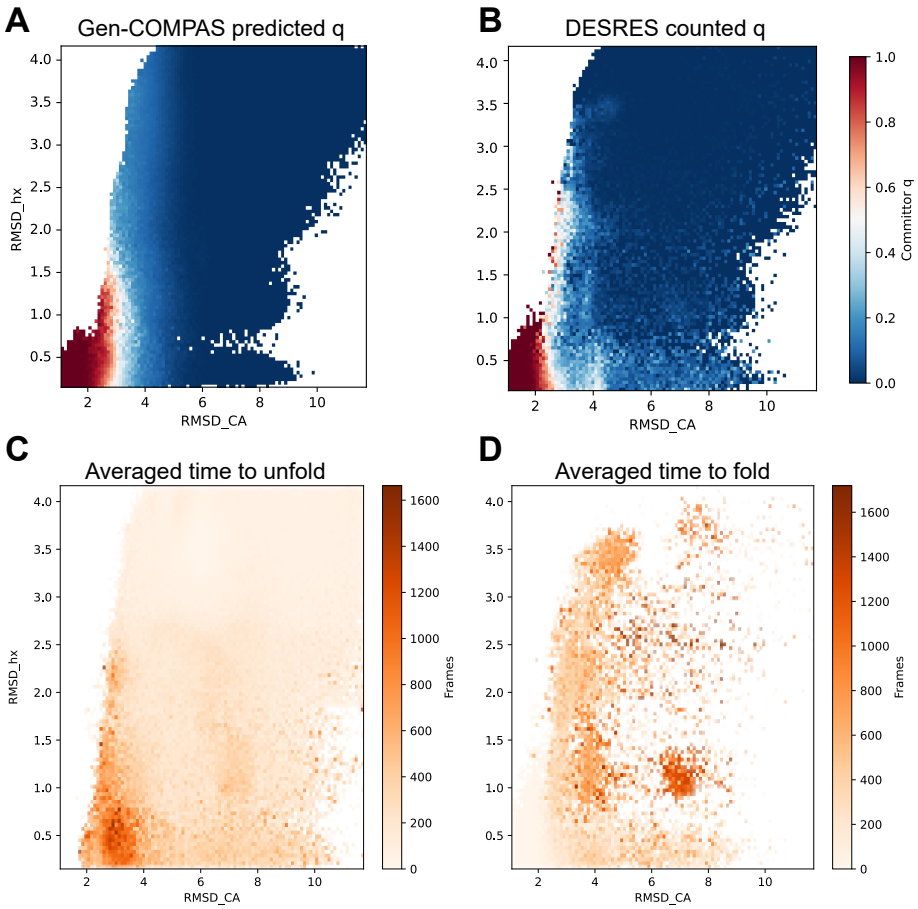


Figure R2: **Semi-quantitative comparison between the Gen-COMPAS-predicted Trp-cage committor and a DESRES-based first-arrival reference.** (A) Gen-COMPAS-predicted committor values evaluated on DESRES configurations and projected onto the two-dimensional CV space defined by $\mathrm{RMSD_{CA}}$ and $\mathrm{RMSD_{hx}}$. (B) DESRES-based first-arrival committor estimated in the same CV space by assigning each frame according to the first basin reached by its future trajectory. The reference value in each bin is computed as $q_{\mathrm{AB}} = N_B/(N_A + N_B)$, where $N_B$ and $N_A$ denote the numbers of frames that first reach the folded and unfolded basins, respectively. (C,D) Mean first-arrival times, in saved frames, from each bin to the unfolded basin (C) and folded basin (D), respectively. Although the DESRES-based committor is noisy because frames within each bin are temporally correlated and the DESRES force field is not identical to that used in our simulations, the comparison shows that the Gen-COMPAS-predicted transition-state region lies in the broad DESRES separatrix region rather than being displaced into either stable basin. The arrival-time maps further show that configurations near the separatrix can require hundreds to more than one thousand saved frames, with 0.2 ns saving frequency, before commitment.

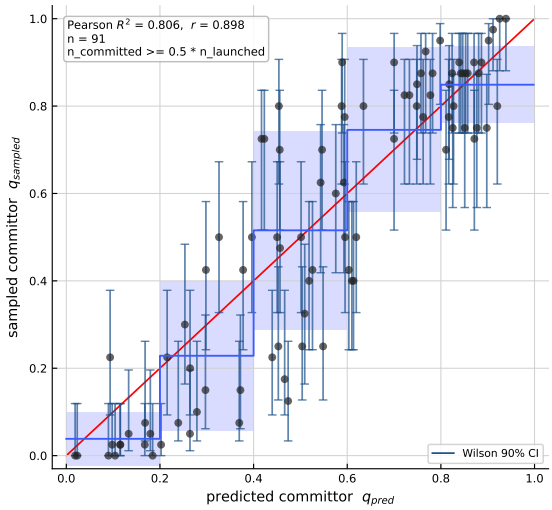


Figure R3: **Aimless-shooting validation of the Gen-COMPAS-predicted committor for Trp-cage.** Initial structures were selected across the full range of predicted committor values, $q_{\mathrm{pred}} \in [0, 1]$, from an independent Gen-COMPAS run not used for training. From each structure, unbiased shooting trajectories were launched and propagated until reaching basin A or basin B, or until the finite simulation-time cutoff was reached (50 ns). The sampled committor, $q_{\mathrm{sample}}$, was estimated from the fraction of committed trajectories reaching basin B. To reduce finite-time censoring effects, only structures for which more than 50% of the launched trajectories committed to either basin were included, giving $n = 91$ validation points. Error bars denote 90% Wilson confidence intervals. The strong agreement with the diagonal, with Pearson $r = 0.898$ and $R^2 = 0.806$, supports the acceptable accuracy of the predicted committor.

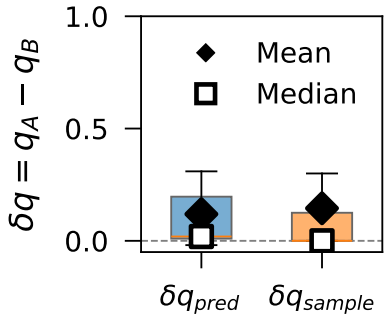


Figure R4: **Boxplot of the committor perturbation induced by the TMD-based sampling protocol for Trp-cage.** Twenty TMD endpoint structures (towards 10 targets from A and B) were randomly selected, and for each structure the committor was evaluated using both Gen-COMPAS prediction and aimless shooting. The distributions of $\delta q$ have median value 0.0, indicating that the TMD setup does not introduce a detectable systematic shift in the committor estimate and adds only minimal noise to the sampling procedure.

## Response to Referee 3

As stated in my previous report, the work addresses an important challenge in molecular dynamics simulation of biomolecules: to extend the sampling to the transition region between metastable states. Key innovations are the use of a denoising diffusion probabilistic model to propose configurations, of a committor model, and of z-matrices as representation of internal coordinates. In the revision, the authors have substantially improved the manuscript and added further interesting and challenging applications to Vo-ring rotation in a V-type ATPase and acetylcholine receptor activation. They have also made a serious effort to address my earlier concerns, and those of the other reviewers. As stated before, I find this to be a very promising direction with quite impressive results. However, in my view a number of points deserve further attention. In the following I will first briefly review key responses to my earlier questions, keeping the old numbering. I will then ask for some clarifications for the new results.

**Response**: We thank the Referee for this positive assessment and for recognizing the substantial revisions, the added applications, and our effort to address the previous concerns. We are encouraged that the Referee finds the direction promising, and we respond to the remaining points below.

Review of response to earlier questions: (1) Does the diffusion model adding value? The response and revision make the purpose of the diffusion-based generative model clearer and also its use: to generate reference structures for targeted MD, thereby avoid inaccuracies in the diffusion model while still harnessing some of its power. While this may not be the most elegant approach, it appears to a practicable solution to a major problem: how to create meaningful structures away from sampled structures. There is now also an effort to quantify the effects of this mixed procedure, in response also to reviewer #2. While I am still not entirely convinced that much value is added by using a diffusion model, I am overall satisfied with the effort and response.

**Response**: We thank the Referee for this assessment. We are pleased that the role of the diffusion model and its connection to the subsequent TMD procedure are now clearer. We agree that the diffusion model should be viewed primarily as a proposal mechanism: it generates structurally plausible reference configurations, which are then relaxed, filtered, and tested through MD-based procedures rather than being used directly as physical transition paths.

We want to suggest that the main reason for introducing the denoising diffusion probabilistic model (DDPM) is not merely to generate structures along the dominant progress coordinate from A to B, but also to broaden the exploration of degrees of freedom that are not captured by a single committor-like coordinate. In complex chemical and biological systems, there is generally no guarantee that the transition between two metastable states proceeds through a single reaction channel. Multiple reactive tubes may coexist, and configurations with similar committor values can still differ along slow orthogonal degrees of freedom. Sampling only one such channel can therefore bias both the reconstructed thermodynamics and the inferred kinetics. And sampling all of the reaction channels is a challenging task.

From this perspective, the DDPM provides a way to diversify the structural proposals before MD relaxation. Heuristically, the denoising process can be viewed as an annealing procedure in latent/configurational space: it generates configurations that retain learned structural correlations while allowing exploration beyond the immediate neighborhood of previously sampled trajectories.

We have clarified this point in the revised manuscript. The intended value of the DDPM is to improve the balance between exploration and physical consistency: it helps propose diverse candidate structures, including variations within similar committor slices, while the following TMD, shooting, and committor-learning steps determine which of these proposals are dynamically meaningful. This design is motivated by the need to avoid over-committing the sampling procedure to a single reaction

tube discovered in the earliest iterations.

However, as one minor point, (2) Is the learned committor model predictive? The authors now provide tests of their committor model for alanine dipeptide (NANMA) the model most easily sampled. The new results in SI Figs. S7 and S8 are decent, even if some concerns remain. In particular, it appears that the transition region is relatively poorly captured. The bimodal distribution of the "true" committor of presumed transition states in SI Figs. S7 center and S8 right indicates that the actual conformations cluster on either side of the barrier but not quite on top. The results in SI Fig. S9 look more reassuring. As I understand, they differ in the way test structures were chosen? In any case, this is a challenging problem, even for alanine dipeptide, and testing for the more challenging systems would in my view be important (e.g., for Trp-cage), also as a basis for comparison with other methods.

**Response**:

We sincerely thank the Referee for mentioning this important matter. We agree that the committor prediction and therefore its intrinsic power to select transition regions are vital for the validation, and we have substantially clarified and expanded the analysis in the SI Section 2.6. We have performed additional extensive committor validations for NANMA and Trp-cage.

The apparent discrepancy between Figs. S7–S8 and Fig. S9 (now Fig. S10) arises from the different purposes of the tests. In Figs. S7–S8, we performed aimless shooting directly from raw diffusion-generated structures as a stringent stress test of the generative model, without assigning these raw structures direct physical or dynamical meaning (which, we concur, was ambiguous in the previous narrative, and we do appreciate that the Referee brought this specific point to our attention towards clarification). In Fig. S9 (now Fig. S10), we instead compared the learned committor against independent, long unbiased NANMA simulations. To further verify the predictive power of the committor model, we added an additional shooting validation using structures extracted from TMD trajectories. This test does not introduce data leakage, because the TMD trajectory data were not included in the model training. In fact, the training set contains only unbiased trajectories initiated from the endpoints of the TMD trajectories, and those endpoint configurations were deliberately excluded from the new TMD-extracted validation.

For NANMA, this TMD-extracted validation is the test most directly relevant to the Gen-COMPAS production workflow. As shown in Fig. R1, structures were selected from predicted iso-committor slices centered at $q_{\text{pred}} \approx 0.1$, 0.3, 0.5, 0.7, and 0.9. The corresponding sampled committor means are 0.174, 0.359, 0.485, 0.617, and 0.799, and the predicted-versus-sampled comparison gives a Pearson correlation coefficient of 0.848. In particular, in the $q_{\text{pred}} \approx 0.5$ slice, a clear single-modal distribution of $q_{\text{sample}}$ emerges, peaked at 0.5. This clarifies the difference between the stricter raw-generation stress test in Figs. S7–S8 and the workflow-relevant TMD-extracted validation in Fig. R1. We also added a sampled-committor bidirectional TMD test, for which paired endpoints driven toward the same generated target give $\delta q_{\text{pred}} = q_A - q_B$ and $\delta q_{\text{sample}}$ centered near zero within statistical uncertainty.

We agree in general with the Referee that validation on a more complex system is important and warranted in the present work. In response to this concern, we have now performed a direct aimless-shooting validation for Trp-cage, in addition to a DESRES-based layered validation. Specifically, we have selected 100 configurations spanning the full range of Gen-COMPAS-predicted committor values, $q_{\text{pred}} \in [0, 1]$, from an independent Gen-COMPAS run that was not used for training (Run 2 in Fig. S17 in the SI). From each configuration, we launched 20 independent unbiased shooting trajectories, each propagated for up to 50 ns, or until it first reached basin $A$, or basin $B$. This corresponds to a maximum aggregate sampling effort of $100 \times 20 \times 50$ ns $= 100$ $\mu$s. Because the trajectories are necessarily terminated after a finite time, this validation should be interpreted as a

finite-horizon shooting test rather than an exact infinite-time committor estimate. To reduce the influence of uncommitted/censored trajectories, we have included only structures for which more than 50% of the launched trajectories committed to either basin within the simulation window. This resulted in $n = 91$ validation points.

The sampled committor, $q_{\text{sample}}$, was estimated from the fraction of committed trajectories reaching basin $B$. As shown in Fig. R3, $q_{\text{sample}}$ agrees well with the Gen-COMPAS prediction over the full committor range, with a Pearson correlation coefficient of $r = 0.898$, and $R^2 = 0.806$. The error bars denote a 90% Wilson confidence. These results provide direct evidence that the Gen-COMPAS-predicted committor is quantitatively consistent with independent unbiased shooting outcomes for Trp-cage, while also making explicit the finite-time nature of the validation.

As an additional test for the Trp-cage system, we examined whether the TMD-based sampling protocol introduces a systematic perturbation into the committor estimate. We randomly selected ten pairs of TMD endpoint structures (each pair with the same target starting from $A$ and $B$), and evaluated the committor using both Gen-COMPAS prediction and aimless shooting. The resulting distributions of $\Delta q$ from both tests have median value 0.0, indicating that under our TMD setup, the sampling procedure does not introduce a detectable systematic shift in the committor, and adds only minimal noise to the validation protocol.

We further performed a DESRES-based independent validation. The DESRES data are completely independent of the Gen-COMPAS training data, and were generated with different force-field parameters, making this a stringent transferability test. First, we evaluated the Gen-COMPAS committor on DESRES configurations, and compared the resulting $q_{\text{pred}}$ map with a DESRES first-arrival committor map in the same $\text{RMSD}_{\text{CA}}$–$\text{RMSD}_{\text{hx}}$ space. The Gen-COMPAS-predicted separatrix coincides closely with the DESRES intermediate-committor region, with minimal deviation, and no displacement toward either stable basin (see Fig. R2).

Second, we took all DESRES frames satisfying $q_{\text{pred}} \in [0.45, 0.55]$, and followed the DESRES trajectory forward from each frame until first arrival to basin $A$ or $B$. For the $n = 2028$ resolved frames, the mean predicted committor is $\langle q_{\text{pred}} \rangle = 0.4998$ with 95% CI $[0.4985, 0.5010]$, while the empirical first-arrival committor is $q_{\text{sample}} = 0.5335$ with 95% Wilson CI $[0.5118, 0.5552]$ ($N_A = 946$, $N_B = 1082$). Thus, configurations predicted by Gen-COMPAS to lie on the $q = 0.5$ separatrix commit to the two basins with nearly equal probability in an entirely independent unbiased trajectory. The absolute deviation from the ideal value is only 0.0335, which is marginal given the temporal correlations in the DESRES trajectory, the finite number of statistically independent transition-region samples, and the different force-field parameters.

Taken together, the NANMA shooting tests, the direct Trp-cage aimless-shooting validation, and the DESRES-based Trp-cage analysis support the predictive quality of the learned committor. In particular, the consistency between Gen-COMPAS and DESRES for Trp-cage is noteworthy because the two datasets use different force-field parameterizations, suggesting that the learned separatrix captures transferable dynamical structure rather than a force-field-specific artifact.

(3) How are diffusion model predictions turned into MD-ready configurations? This question has been clarified.

**Response**: We thank the Referee for confirming that this point has been clarified.

(4) Can we trust the estimated free energy surfaces? The authors now use a different method (RiteWeight). This is a rather new procedure, though at its core it looks to be closely related to traditional Markov state modeling. Here it might be useful to discuss the potential shortcomings a bit more clearly in the text. As indicated in my earlier question, I can see that one can get

"local" free energy maps quite right, but to get global maps (e.g., the A-to-B equilibrium) right is challenging because disparate and poorly connected regions in phase space need to be connected somehow.

**Response**: We agree with the Referee. We have added a more substantive, and hopefully clearer discussion on the limitations of RiteWeight and related reweighting/Markov-state-model-like approaches. In particular, we now distinguish the reliable mechanistic use of the FELs from the stricter claim of globally converged equilibrium free-energy differences. The RMSE/degradation analysis shows that the remnant fluctuations are within a reasonable margin, germane for interpreting local basins, barriers, and transition regions. We, therefore, retain the statement that the FELs constitute a powerful tool for mechanistic analysis, while making clear that global $A$-to-$B$ free-energy differences ought to be viewed with the appropriate caution—just like any free-energy difference determined using enhanced-sampling strategies.

The main manuscript now states:

The generated FELs represent a powerful tool for identifying key states and barriers, but for applications demanding the highest quantitative precision, they should serve as a well-informed starting point.

The reweighted FELs remain a powerful tool for identifying metastable basins, barriers, and transition regions. Although not every complex-system FEL satisfies the strictest criteria for global equilibrium convergence, the analysis of the FEL estimates reveals that the remnant fluctuations are sufficiently limited to allow mechanistic interpretation (see SI). We, therefore, use these FELs to support local and pathway-level observations, while absolute global $A$-to-$B$ free-energy differences should be viewed cautiously as estimates.

(7) How does Gen-COMPAS compare to other methods? I understand the hesitancy to compare directly to established methods and I appreciate that the discussion has been expanded and includes additional references. As a minor point, in ref 56 (Okazaki et al.) complete transition paths of a membrane transporter were not "inferred" but actually sampled in transition path MD simulations, as I understand.

**Response**: We thank the Referee for pointing this misconstruction out. We have corrected the phrasing and now describe the work by Okazaki et al. as sampling complete transition paths rather than inferring them.

Following are questions to the new results: A. The added results for the acetylcholine receptor look promising. Here, it was not quite clear if the ligand acetylcholine exchanged spontaneously or if exchange was somehow driven. As I understand from the protocol, no excess ligand was present in the solution, but ligand position/s were included in the CVs. a. Was excess ligand present in the solution? b. If not, how was binding and unbinding modeled (by distances to all or a subset of ligands in the CVs)? c. Was acetylcholine binding and unbinding reversible? d. Did the mono-liganded form of the receptor form both from the di-liganded form and the unliganded form by release and binding of ligand? (The text suggests the former.) e. Are the free energies in Fig. 4F at standard concentration of ligand (1M), as the conformational transition is coupled to binding?

**Response**: We thank the Referee for identifying this ambiguity. We have revised the acetylcholine-receptor section and simulation protocol to elaborate how the ligands were handled in our simulations.

(a) Only two ACh molecules were present in the solution, i.e., the same two molecules observed in the di-liganded cryoEM structure. They were placed in the water phase, roughly 2 nm away from

the two binding interfaces in order to keep the molecular composition identical between the two endpoints. The simulated transition connects an unliganded closed endpoint to a di-liganded active endpoint in a finite-size system.

(b) Binding and unbinding were not imposed through ligand-distance CVs during generation. The protein and the two ACh molecules were represented explicitly and generated jointly by the diffusion model as one molecular system. In this full heavy-atom representation, local graph-neural message passing captures short-range protein-ligand contacts, while residue-level attention captures longer-range conformational coupling, which is consistent throughout all the systems examined in our manuscript. The ligand distances shown in Fig. 4 are post-simulation projection variables and basin descriptors, rather than empirically predefined variables used to drive the generative model or the committor training.

(c) Within the finite-size-system Gen-COMPAS trajectories, ligand release and reuptake relative to the two binding interfaces can occur as part of the generated and MD-propagated transition ensemble. As the Referee kindly pointed out, this should not be interpreted as exchange with a bulk-ligand reservoir, because no excess ligand bath was included. We have revised the manuscript section **Gen-COMPAS captures an asymmetric activation mechanism of nAChR via a mono-liganded intermediate** accordingly, and it now reads:

. . . Only two acetylcholine (ACh) molecules were modeled in the solution, i.e., the same two molecules observed in the di-liganded cryoEM structure. The receptor and the two ACh molecules were represented and generated jointly as one molecular system. The ligand-protein distance coordinates were used only after sampling to project the FEL and define interpretable basins. This explicit joint protein–ligand generation allows ligand occupancy and receptor rearrangement to be learned as coupled degrees of freedom. . .

(d) The mono-liganded basin is observed from the Gen-COMPAS trajectory along the finite-size-system transition network connecting the unliganded and the di-liganded endpoints. Operationally, it can be reached from di-liganded-like configurations by release of one explicit ACh molecule, and from unliganded-like configurations through binding of one of the explicitly represented ACh molecules to one interface.

(e) The free energies in Fig. 4F are not standard-state binding free energies at 1 M ACh. They are apparent basin free-energy differences in the simulated finite-size system, obtained from the reweighted Gen-COMPAS ensemble. We should compare them semi-quantitatively with experimental kinetic-model-based estimates, but we did not apply a 1-M standard-state ligand correction. We thank the Referee for bringing up this issue that evidently was not clear enough in the previous version of the manuscript. We have modified the phrasing accordingly for transparency and reproducibility.

The narrative used to read:

Integrating the FEL basins corresponding to the unliganded, mono-liganded, and di-liganded states yields free-energy differences consistent with experimental estimates derived from two kinetic schemes, with agreement within 1 kcal/mol.

and now reads:

Integrating the FEL basins corresponding to the unliganded, mono-liganded, and di-liganded states yields apparent basin free-energy differences for the simulated finite-size-system setup. These values fall within the range suggested by experimental kinetic schemes, albeit no 1-M standard-state ligand correction was applied. We, therefore, interpret the comparison as semi-quantitative rather than as a difference of standard binding free energies.

B. The results for Vo ring rotation are also encouraging. Here, I wonder how the calculated free energies (Fig. S4B) compare to the results from "conventional" sampling in Roh et al. (SI Ref. 15), which served as starting point here, and to the results for F-type ATP synthase c-ring rotation (Blanc PNAS 2024; DOI: 10.1073/pnas.2314199121).

**Response**: We thank the Referee for this useful suggestion. We have revised the section on $V_o$-domain rotation in the SI to clarify both the quantitative comparison that can be made with Roh et al. and the more qualitative comparison that can be made with the F-type ATP synthase study of Blanc et al.

Roh et al. provide the most direct reference because they studied the same $V_o$ system, and their structures define the adjacent rotor–stator states used to initiate our calculation. At the semi-quantitative level, our Gen-COMPAS FEL is broadly consistent with their conventional-sampling results obtained from a combination of a string-method-with-swarms-of-trajectories (SMwST) path optimization followed by enhanced-sampling along a path-CV. Roh et al. report a main rotary barrier of approximately 8 kcal/mol, corresponding to the disruption of the interaction of $c''$E108 with arginines R735 and R799 borne by the stator, and a small end-state free-energy difference, $\Delta G_{AB} < 1$ kcal/mol, whereas our estimated main barrier is approximately 6 kcal/mol with the same qualitative conclusion that $\Delta G_{AB} < 1$ kcal/mol. Thus, both calculations support a nearly isoenergetic adjacent-state transition separated by a moderate rotary barrier, although Gen-COMPAS yields a somewhat lower apparent barrier.

A difference with the results of Roh et al. ought to be noted here. In our Gen-COMPAS FEL, the basin labeled M3 has a free energy close to that of the initial state $A$, whereas the analogous region in Roh et al. lies approximately 5 kcal/mol above $A$. This discrepancy, as well as the lower apparent barrier at the transition state in our FEL, may reflect the high degeneracy of the projected CV space, and the presence of multiple reaction channels. For example, different combinations of interfacial hydrogen-bond and salt-bridge rearrangements may lead to similar values of the low-dimensional rotary CVs, while remaining partially distinct mechanistically. Roh et al. followed a more restricted pathway supplied by the SMwST—which ignores the possibility of branching—whereas Gen-COMPAS samples concomitantly multiple transition pathways. If these alternate channels overlap in the projected CVs, the resulting free-energy profile may average over several routes and lower the apparent barrier, or stabilize intermediate basins such as M3. This characterization evidently assumes that our estimated FEL is sufficiently accurate. We, nevertheless, feel that the agreement in the barrier scale and $\Delta G_{AB}$ is encouraging.

The comparison with the work of Blanc et al. must necessarily be done at a qualitative level. Their study examines the rotation of the $c$-ring in an F-type ATP synthase, in stark contrast with the V-type ATPase used in our own work. It ought to be recalled that unlike the isolated membrane sector, $F_o$, of an F-type ATP synthase, which acts as a passive proton pore, a free $V_o$ domain, like the one we investigated in its auto-inhibited state, does not conduct protons along a pH gradient under physiological conditions. We also recall here that the molecular origin of this difference in activity between the two related proton channel subcomplexes remains unknown.

In addition, the FELs of Blanc et al. are constructed for different protonation states, and explicitly analyzed in relation to proton transfer. In that study, the $c$-ring rotation FEL is strongly shaped by protonation-state changes, water-wire-mediated proton transfer, and electrostatic stabilization of the forward-rotated state, whereas the calculations performed for the $V_o$ domain focus on the rotary transition between structurally defined adjacent states. Instead of using the Blanc et al. free energies in a like-for-like numerical benchmark, we prefer to quote these results as an important point of comparison, underscoring that rotary molecular motors—either ATP-ases or synthases—can exhibit discrete metastable substates, barrier-separated angular transitions, and strong coupling between interfacial electrostatics and rotary directionality. We have added this clarification to the SI,

and we agree with the Referee that applying Gen-COMPAS to related rotary systems in the future will provide a useful way to compare how similar functional architectures can produce different kinetic and thermodynamic landscapes.

To make the comparison explicit for the reader, we have revised the $V_o$ subsection of the SI in 1.3 Rotary mechanism of the V-ATPase $V_o$ domain, and added the reference to both the SI and the manuscript.

# Response to the comments of the referees of Nature manuscript 2025-10-28431B

We sincerely thank the referees for their careful evaluation of the revised manuscript. We are particularly grateful that both referees found the additional committor validation convincing and that Referee #3 regarded the technical questions raised in the previous round as satisfactorily addressed. In response to the remaining comments, we have substantially rewritten and reorganized the open-source implementation to provide a fully documented end-to-end workflow, clarified the methodological provenance of Gen-COMPAS, and corrected the two minor points identified by Referee #3.

## Response to Referee #2

We sincerely thank the authors for addressing our concerns, primarily on committor quality. The new committor validation plots are of satisfying quality and reassuring on the correctness of the sampling.

However, we sadly cannot recommend publication of this manuscript in Nature.

From a methodological point of view, the manuscript describes a pipeline that chains together existing components and applies it to standard benchmarks and more challenging systems. The method itself is not transformative or radically novel.

The paper contains ambitious applications of the method to complex biological systems. Still, they are assessed only for qualitative agreement with existing results, with which they broadly agree but quantitatively diverge (with biological consequences). This is true progress, given that their pipeline can handle complex systems. But these applications remain qualitative/confirmatory (where they do agree) and tentative/insufficiently explored regarding their biological implications (where they do not).

The progress is incremental, and – in good part – self-referential. Committor learned in full/internal coordinates "without CVs" is already in the authors' own Arredondo et al. (Nat Comput Sci 2026); the committor-consistent string and VCN are Megías et al. (Nat Comput Sci 2025) and Chen et al. (JCTC 2026). Generative models for related problems (such as coupling with TPS) are already established (MDGen, NeurIPS 2024; Raja et al., OM-TPS, ICML 2025; etc.). BioEmu (Lewis et al., Science 2025) and Boltzmann generators (Noé et al., Science 2019) already deliver one-shot equilibrium ensembles and 1 kcal/mol free energies.

Putting these workflows, including those from the author's own group, together to push beyond benchmarks and produce results that are broadly qualitatively consistent with prior work is an engineering achievement, but not of the kind that would warrant publication in flagship Nature without a substantial discovery arising from it. That would not be consistent with the previous publishing practices in this field.

**Response**: We thank the Referee for this assessment. We have now made the provenance of every component explicit in the new Methods subsection entitled *Gen-COMPAS overall workflow*. This subsection clearly states that Gen-COMPAS integrates and adapts established concepts rather than introducing each component de novo. It identifies and cites the foundations of the workflow, including denoising diffusion models, variational committor learning, committor-consistent path construction, targeted molecular dynamics, and trajectory reweighting with RiteWeight. Then the section defines precisely how these components are organized in the Gen-COMPAS feedback loop. We believe that this revised presentation gives readers a transparent account of what is established, where each component originates, and what is implemented in the present workflow.

A last remark. In this iteration, we inspected and ran the code to reproduce some of the results presented in the manuscript. We managed to reproduce the committor convergence on NANMA, but

we could not get the RMSD-based TMD fix to work properly in the generator output. Regrettably, the code doesn't actually give a harness for a complete run, only scripts for individual steps. We had to use Claude AI to make this harness. We would be happy to see this manuscript published in Nature Computational Science, provided that the authors submit rewritten code as a package with an end-to-end reproduction script that runs and produces reliable results.

Referee #2 (Remarks on code availability): We would like to see the code as a comprehensive package, or at least as harnessed scripts that could be run end-to-end to reproduce some of the results in the manuscript.

**Response**: We fully agree that the software should allow users to execute the workflow without manually connecting a collection of independent scripts. We have, therefore, rewritten Gen-COMPAS as an installable package with a unified command-line interface that organizes the complete workflow, including data preparation, generative-model training and inference, RMSD-based TMD refinement, committor learning and selection, and downstream analysis. Still, we recommend that users execute Gen-COMPAS in a stepwise manner to avoid failed simulations resulting from unsuitable parameter choices. This notice is included in the README of the open-sourced package.

We have also reorganized the examples distributed through the GitHub repository. Ready-to-use configuration files and the corresponding molecular-simulation inputs are now provided for representative systems in the `examples` directory. These examples document the required parameters and provide concrete starting points that users can adapt to new systems. The hyperparameters detailed in the SI are also double-checked. In addition, we developed a graphical user interface that guides users through the available options and generates a valid configuration file, reducing the need for manual editing and helping to prevent formatting or parameter-name errors.

The revised package, installation instructions, workflow documentation, example configurations, and graphical configuration-file generator are available at https://github.com/Tangcyu/Gen-COMPAS. The Code availability statement in the manuscript has also been updated to describe the installable package, unified entry point, configuration files, and example inputs.

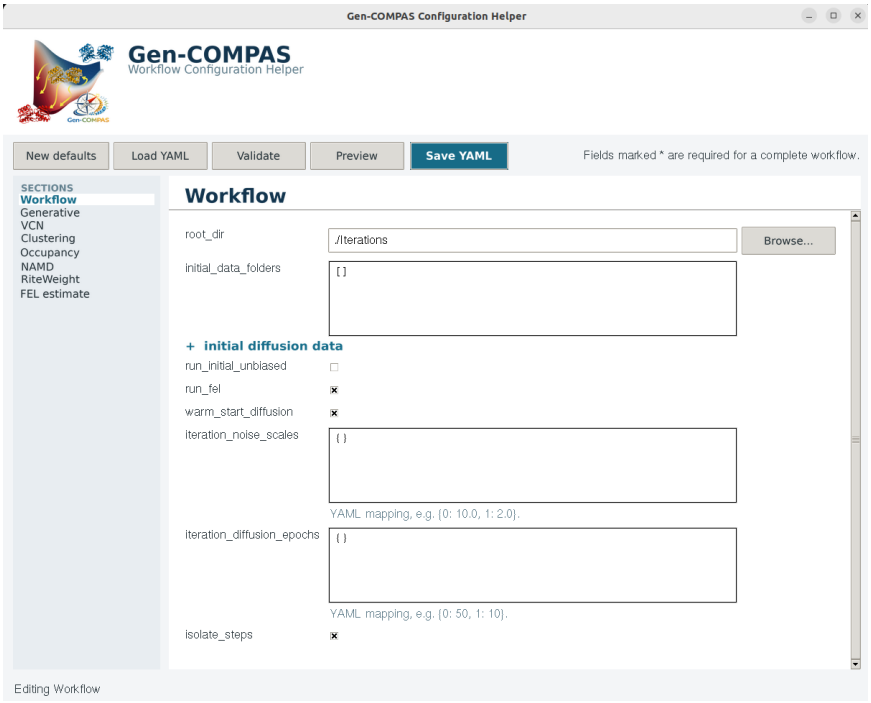


**Figure R1. Graphical user interface for Gen-COMPAS.** The interface guides users through the available options and generates a valid configuration file.

## Response to Referee #3

The authors have made a serious effort to address the reviewer concerns. I am particularly impressed by the thorough analysis of the learned committor model for NANMA and Trp-cage. The results are convincing. The usefulness of the diffusion-based generative model and its integration into the workflow are now much clearer and more compelling. I also appreciate the more nuanced discussion of the results and the expanded discussion of related results. My technical questions have been satisfactorily addressed. As previously stated, I find this to be a promising methodological direction with impressive results on diverse, challenging, and biologically relevant systems. I therefore recommend publication in Nature.

**Response**: We sincerely thank the Referee for this positive assessment and for recognizing the additional committor validation, the clearer presentation of the integrated workflow, and the broader range of applications.

A very minor issue remains: On lines 183–184, a 100× speedup in sampling over conventional simulation is claimed. However, such a claim seems overly bold, as it is not quantified in terms of the asymptotic decay of the statistical error. 10 $\mu$s of regular MD might have yielded comparable results, which would not diminish the value of the present work.

**Response**: We agree and thank the Referee for this important qualification. We have removed the speedup claim and now report only the difference between the aggregate trajectory lengths, explicitly stating that it is a comparison of sampling budgets rather than a measured acceleration in statistical convergence. The revised text reads:

> For Trp-cage, Gen-COMPAS used 594 ns of aggregate unbiased sampling, whereas the conventional MD trajectory used for semi-quantitative validation was 208 $\mu$s long. This approximately 350-fold difference in trajectory length reflects a comparison of sampling budgets rather than the asymptotic decay of statistical error or wall-clock time, since a shorter conventional MD simulation could yield comparable observables. The learned committor, thereby, estimates the probability of folding before unfolding without requiring a folding event to occur in a single continuous trajectory.

Additionally, there appears to be a minor typo on line 256: "bound" should be "bound to".

**Response**: We thank the Referee for noticing this typo. We have changed the wording in the revised manuscript.

We thank the referees again for their careful reading and constructive comments, which have helped improve the clarity and usability of Gen-COMPAS. We believe that the revised methodological attribution, the fully documented software package, and the textual revisions have improved both the transparency and the practical usability of Gen-COMPAS.