Available online **www.ijpras.com**

International Journal of Pharmaceutical Research & Allied Sciences, 2025, 14(1):15-26
https://doi.org/10.51847/zxBEdTo5M1

Original Article

ISSN : 2277-3657
CODEN(USA) : IJPRPM

Reproducible Molecular-Dynamics Evidence Requires More Than Longer Trajectories, Larger Systems, and Visually Persuasive Conformational Movies That Obscure Uncertainty

Hans Müller[1]*, Anna Schmidt[1], Thomas Weber[2]

[1]Department of Reproducible Molecular Dynamics and Uncertainty, Faculty of Pharmaceutical Sciences, University of Bonn, Bonn, Germany.
[2]Department of MD Simulation and Conformational Analysis, Faculty of Pharmacy, LMU Munich, Munich, Germany.

***Email:** hans.mueller@uni-bonn.de*

ABSTRACT

Molecular-dynamics simulation can reveal motions, interactions, and state transitions that are difficult to observe directly, but its evidential strength is frequently inferred from computational scale rather than demonstrated through reproducibility. Longer trajectories, larger systems, elaborate enhanced-sampling calculations, and visually persuasive conformational movies may increase descriptive richness without resolving uncertainty arising from system preparation, force-field selection, initial conditions, incomplete sampling, collective-variable construction, analytical flexibility, or disagreement with external evidence. This methodological framework article proposes the Molecular-Dynamics Evidential Reproducibility Framework, an original non-empirical synthesis that connects claim specification, system provenance, model adequacy, independent sampling, estimator-specific convergence, uncertainty decomposition, analytical robustness, negative controls, external validation, and bounded claim formation. The framework is organized as a gated reproducibility chain in which no single component can substitute for a missing evidential function elsewhere. A complementary Claim-Strength Matrix classifies simulation claims from visual illustration to externally validated mechanistic inference according to cumulative evidence from replication, convergence, robustness, and validation. Eight testable propositions are derived, including the non-substitution, estimator-first, replication, analysis-invariance, model-perturbation, visualization-boundary, reproducibility–validity distinction, and cumulative-claim propositions. Reporting requirements and governance recommendations are proposed for authors, reviewers, and journals. The framework does not prescribe universal trajectory durations, system sizes, replica numbers, or numerical convergence thresholds. It is intended to make uncertainty and decision boundaries visible, not to certify a simulation as correct. Prospective evaluation is required before the framework can be treated as a validated reporting or editorial standard.

Key words: *Molecular dynamics, Reproducibility, Trajectory convergence, Uncertainty quantification, Conformational ensemble, Force field*

Received: 01 November 2024; **Revised:** 24 January 2025; **Accepted:** 26 January 2025

INTRODUCTION

Molecular-dynamics simulation, kinetic state modeling, and increasingly complex computational representations have expanded the range and apparent resolution of claims that can be made about molecular motion, conformational ensembles, and biological mechanisms [1–3]. Contemporary simulations can generate trajectories extending across multiple temporal scales, represent systems containing millions of interaction sites, reconstruct

projected free-energy landscapes, estimate association or dissociation kinetics, and produce visually compelling animations of molecular transitions. These advances are scientifically valuable, but they also create a methodological asymmetry: the apparent realism of an output can increase more rapidly than the evidence supporting its interpretation.

A trajectory is not itself a conformational ensemble, and an observed transition is not automatically a prevalent pathway. A projected energy landscape is not the complete physical landscape, a stable numerical estimate is not necessarily a physically accurate estimate, and reproducible execution is not equivalent to mechanistic validation. Molecular dynamics generates time-ordered configurations under a defined model and protocol. Evidential claims arise only after decisions about preparation, parameters, sampling, representation, estimation, uncertainty, robustness, and validation have been made. The distinction is central because a trajectory can be computed correctly while representing an incorrectly prepared system, an unsuitable force field, an incompletely sampled ensemble, or an analytically fragile interpretation.

Visual persuasion intensifies this problem. A movie may show a ligand leaving a pocket, a helix bending, a domain opening, or a communication pathway propagating through a protein. Such observations can be useful for hypothesis generation. They do not, by themselves, establish the probability of the event, its characteristic timescale, its causal role, or its robustness to alternative modeling choices. The eye is drawn to coherent movement and may interpret continuity as mechanism even when the selected frames represent one realization among many, a biased trajectory, an uncommon fluctuation, or a projection determined by the analyst.

This article develops a proposed Molecular-Dynamics Evidential Reproducibility Framework. The framework treats reproducibility as a chain of connected evidential responsibilities rather than as a single property of code, trajectory duration, system size, or data availability. It further proposes a Claim-Strength Matrix through which the maximum defensible interpretation is conditioned on independent replication, estimator-specific convergence, uncertainty characterization, robustness to scientifically reasonable alternatives, negative controls, and external validation.

The contribution is methodological and non-empirical. It does not present a new simulation dataset, benchmark, software implementation, or validated scoring rule. Its purpose is to integrate established concerns that are often addressed separately and to derive testable propositions, reporting expectations, decision boundaries, and governance mechanisms. The central argument is that computational scale and visual sophistication may strengthen a study only when they serve a clearly defined evidential function. They cannot compensate for missing replication, unknown convergence, model misspecification, analytical instability, or absent validation.

Why simulation duration alone does not establish evidential strength

A long individual trajectory does not quantify sensitivity to initial conditions, replica variability, or parameter uncertainty, and reliance on a small number of realizations can support apparently coherent but non-reproducible conclusions [4]. Duration increases the number of integration steps and may increase the probability that a particular transition is encountered. It does not reveal whether another initial velocity assignment, starting conformation, protonation state, or force-field parameterization would produce a materially different distribution of outcomes.

The evidential target must therefore be distinguished from computational elapsed time. If the claim concerns an equilibrium population, the relevant question is whether the estimated population is stable across independent samples and appropriate diagnostics. If the claim concerns a transition rate, the relevant question is whether state definitions, kinetic assumptions, rare-event treatment, and available transition statistics support that rate. If the claim concerns a mechanism, the relevant question is whether the proposed sequence is recurrent, perturbationally informative, robust to analysis choices, and consistent with evidence not used to construct it. None of these questions is answered simply by reporting microseconds or milliseconds.

Length and independence are also not interchangeable. Several trajectories initiated from the same equilibrated structure with minimally varied conditions may be technically separate yet remain informationally dependent. Conversely, one continuous trajectory may contain several transitions without providing independent evidence if the system remains confined to a narrow region of phase space or repeatedly revisits states through the same local route. Replicate design must therefore specify what independence means for the target claim [5].

System size creates a related misconception. Larger systems can reduce some artifacts by including additional solvent, membrane, nucleic-acid, protein, or environmental context. They may also introduce more slow coordinates, more poorly parameterized interactions, and greater sampling demands. A large simulation can be

more realistic in composition while being less complete in ensemble coverage. Atom count is consequently a property of model scope, not an independent measure of evidential strength.

Enhanced sampling does not remove this distinction. Biasing methods can accelerate exploration, but their conclusions remain conditional on collective variables, bias construction, reweighting assumptions, and the relationship between accelerated exploration and the physical observable of interest. A highly explored low-dimensional projection may obscure unsampled orthogonal coordinates. Computational effort is meaningful only when it is connected to a prespecified inferential target and accompanied by diagnostics capable of evaluating that target [6].

The proposed non-substitution proposition follows: no increase in trajectory duration, system size, computational cost, visual complexity, or methodological sophistication can substitute for an evidential function that it does not perform. Longer trajectories may improve sampling; they do not automatically establish independence. Larger systems may improve contextual representation; they do not establish parameter validity. More elaborate graphics may communicate motion; they do not establish probability or causality.

Sources of variability in system preparation, force fields, sampling, and analysis

Nominally similar simulation tasks can produce different outputs across groups, software environments, and force-field constructions, while alternative parameterization strategies can change the modeled balance among conformational states [7–9]. Such variation is not necessarily evidence that one implementation is erroneous. It demonstrates that an MD result is conditional on a network of decisions that must be exposed before reproducibility or validity can be evaluated.

System preparation

System preparation defines the molecular object being simulated. Relevant choices include source structure, biological assembly, unresolved residues, side-chain reconstruction, alternate conformations, protonation states, tautomers, disulfide bonds, post-translational modifications, cofactors, metal coordination, ligand stereochemistry, membrane composition, solvent model, ion identity, ionic strength, box geometry, restraints, minimization, and equilibration.

These decisions can affect electrostatic networks, local packing, hydration, binding geometry, and accessible conformational states. A precisely reproduced trajectory based on an undocumented or chemically implausible preparation remains evidentially weak. Provenance should therefore include not only the final coordinate file but also the transformations through which that file was produced.

Force fields and environmental models

A force field is an approximate model of interaction energetics. Its adequacy is property-, system-, and condition-dependent. A parameter set developed to improve folded-protein behavior may not provide equivalent accuracy for intrinsically disordered states, ligand chemistry, membrane interfaces, metal centers, or highly charged environments. Water models, long-range electrostatics, cutoff schemes, combining rules, and boundary conditions similarly influence the modeled energy landscape.

Force-field agreement cannot be assumed merely because two models produce structurally plausible trajectories. Nor does disagreement necessarily reveal which model is closer to physical reality. It instead identifies model-form uncertainty that should constrain the claim or motivate comparison with external observables.

Sampling variability

Sampling variability arises from initial coordinates, velocity assignments, stochastic thermostats, replica exchange histories, bias deposition, adaptive decisions, and transitions among metastable states. Chaotic divergence means that microscopic trajectories separate rapidly even under small changes, while statistically meaningful observables may or may not remain stable.

Replication should therefore target the distribution of a quantity of interest rather than frame-by-frame trajectory identity. Exact trajectory reproduction may be useful for software testing, but scientific replicability generally asks whether independent calculations support compatible ensemble, thermodynamic, kinetic, or mechanistic conclusions.

Analytical variability

Analytical choices can be as consequential as simulation choices. Structural alignment, atom selection, distance definitions, feature construction, dimensionality reduction, clustering, state boundaries, lag times, smoothing, histogram bins, reweighting, uncertainty estimators, and movie-frame selection can each influence the apparent result.

Analytical flexibility is especially important when the same trajectory supports several plausible narratives. One collective variable may show a two-state transition, another a continuous deformation, and a third no separation. Confirmatory inference requires the analytical path to be justified in relation to the claim rather than selected because it produces the clearest visual or numerical separation.

Proposed molecular-dynamics reproducibility framework

Transparent enhanced-sampling practice requires access to the inputs, bias definitions, and methodological provenance needed to understand how the reported exploration was produced [10]. Building on this principle, the proposed Molecular-Dynamics Evidential Reproducibility Framework connects the full path from system construction to bounded claim formation. To prevent any single simulation feature from serving as a proxy for evidential strength, the proposed framework connects system preparation, model specification, sampling, uncertainty assessment, robustness testing, validation, and bounded claim formation in a single reproducibility chain (**Figure 1**).

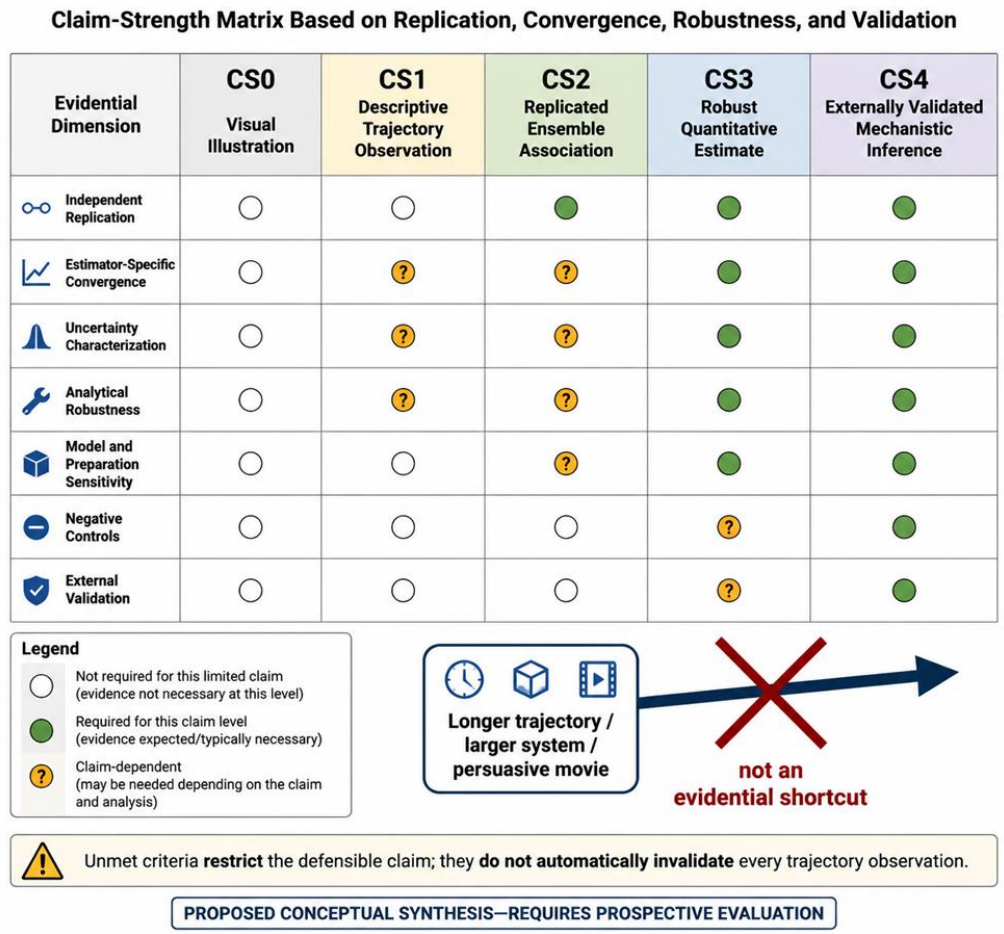

Figure 1. Molecular-Dynamics Reproducibility Chain from System Preparation to Claim Formation

The first stage is claim specification. Authors should state whether the simulation is intended to illustrate a possible motion, describe an ensemble, estimate a quantitative observable, predict a kinetic quantity, or support a mechanistic explanation. This declaration determines which later evidential functions are required.

The second stage is system provenance and preparation. Every structural and chemical transformation that creates the simulated system should be traceable. The third stage is model and force-field adequacy, which asks whether the selected representation is appropriate for the interactions and observables central to the claim.

The fourth stage is sampling plurality. Independent trajectories, replicas, walkers, windows, or biased histories should be designed around the target estimator. The fifth stage is estimator-specific convergence, because no trajectory is converged in the abstract. Only a defined observable, distribution, free-energy difference, rate estimate, or state model can be assessed for convergence.

Metadynamics illustrates why this distinction matters: the scientific object of interpretation is a reconstructed probability distribution or associated free-energy estimate, not the visual course of the evolving bias alone [11]. More broadly, projected free-energy landscapes depend on collective-variable adequacy, exploration, reweighting, and convergence and should not be treated as direct maps of complete molecular mechanism [12].

The sixth stage is uncertainty decomposition, followed by analytical robustness, negative controls, and external validation. The final stage is bounded claim formation, in which interpretation is restricted to the highest level supported by the weakest unresolved evidential requirement.

The distinctions required to prevent trajectory occurrence, ensemble support, computational reproducibility, and mechanistic validation from being treated as interchangeable are summarized in **Table 1.**

Table 1. Definitions, components, relationships, and intended uses in Reproducible Molecular-Dynamics Evidence Requires More Than Longer Trajectories, Larger

Construct	Operational definition	Evidential question answered	Relationship to adjacent constructs	Intended methodological use	Evidence basis	Status in this article	Decision boundary
Molecular-dynamics trajectory	A time-ordered sequence of configurations generated under a specified model, integrator, protocol, and initial condition	What occurred in this realization?	A trajectory contributes samples to, but is not identical to, an ensemble	Describe simulated temporal evolution	Trajectory-level interpretation [1].	Established construct	Occurrence does not establish prevalence, kinetics, or causality
Conformational ensemble	A probability-weighted set of configurations or states under defined conditions	Which structures and populations are supported under the model?	Requires sampling and statistical interpretation of trajectories	Support population and distribution claims	Ensemble interpretation [1].	Established construct	An incompletely sampled ensemble may appear structurally plausible
Energy landscape	A representation relating configurations or reduced coordinates to energetic or free-energy organization	How are modeled states and barriers organized?	Depends on representation, coordinates, and sampling	Organize thermodynamic hypotheses	State-model interpretation [2].	Established construct	A projection is not the complete high-dimensional landscape
Enhanced sampling	A family of methods that modifies exploration or weighting to access otherwise rare regions	Has exploration been accelerated under stated assumptions?	Requires reweighting, collective-variable justification, or method-specific interpretation	Improve access to slow coordinates	Transparency requirements [10].	Established construct	Faster exploration does not automatically establish physical kinetics
Force field	A parameterized approximation of molecular interaction energetics	Under which model are structures and motions generated?	Defines the modeled energy landscape and interacts with preparation choices	Make model assumptions explicit	Force-field dependence [8].	Established construct	Benchmark success in one domain does not imply universal adequacy

Free energy	A thermodynamic quantity relating state probability or reversible work under defined conditions	What thermodynamic preference is estimated?	Distinct from transition rate, pathway, and residence time	Support thermodynamic comparison	Proposed operational distinction	Established concept, operationalized here	A stable free-energy estimate can remain model-dependent
Binding kinetics	Association and dissociation behavior expressed through rates or time-dependent state transitions	How rapidly does binding or unbinding occur?	Distinct from binding affinity and pathway visualization	Support kinetic inference	Proposed operational distinction	Established concept, operationalized here	Observed binding or unbinding is not itself a rate estimate
Residence time	A kinetic measure related to persistence of a bound state under defined conditions	How long does a specified bound state persist statistically?	Depends on state definition and kinetic model	Support drug–target kinetic interpretation	Proposed operational distinction	Established concept, operationalized here	A single long bound interval is not a residence-time estimate
Allostery	Coupling through which perturbation at one region alters states, energetics, or function elsewhere	Does the evidence support linked response across sites?	May involve population shifts, dynamics, structural pathways, or energetic coupling	Organize mechanistic hypotheses	Proposed operational distinction	Established concept, bounded here	Correlated motion or a movie alone does not establish causal allostery
Trajectory convergence	Stability of a prespecified estimator under relevant diagnostic, sampling, and model conditions	Has the claimed quantity stabilized sufficiently for its intended use?	Depends on the target observable and does not establish model accuracy	Evaluate sampling adequacy	Uncertainty framework [5].	Established methodological concept	There is no universal convergence of an entire trajectory
Computational reproducibility	Ability to regenerate compatible outputs from sufficiently specified inputs, code, software, and workflow	Can the calculation be independently rerun or reconstructed?	Supports auditability but differs from replicability and validity	Enable inspection and reuse	Cross-implementation variability [7].	Established construct	Reproducible execution can reproduce systematic error
Replicability	Compatibility of conclusions across independent realizations or implementations	Does the result persist beyond one realization?	Supports estimation of initial-condition and sampling variability	Test stability of conclusions	Replica dependence [4].	Established construct	Repetition under identical bias does not establish physical validity
Robustness	Stability of the claim under scientifically reasonable changes in preparation, model,	Does the interpretation depend critically on one defensible choice?	Extends replication beyond random variation	Identify model and analytical fragility	Proposed integration	Proposed framework component	Passing selected tests does not establish invariance to every alternative

	estimator, or analysis						
Mechanistic validation	Support for a causal or explanatory claim from orthogonal, prospective, or perturbational evidence	Does evidence beyond the generating analysis support the proposed mechanism?	Requires more than computational reproducibility or descriptive agreement	Bound high-strength mechanistic claims	Proposed integration	Proposed framework component	Agreement with one observable may not validate the full mechanism
Claim strength	Maximum defensible interpretation given replication, convergence, robustness, and validation	What can responsibly be concluded?	Integrates the complete reproducibility chain	Calibrate language and decision use	Original synthesis	Proposed construct	The level applies to a specific claim, not automatically to the entire study

Note. Evidence-derived entries are limited to the methodological function supported by the cited source. Rows marked "proposed" represent the article's original conceptual synthesis. The table supplies no universal thresholds, validated rankings, or regulatory criteria.

Replicate design, convergence assessment, and uncertainty decomposition

Ensemble calculations can represent quantities of interest as distributions across independent realizations and thereby improve evaluation of reliability beyond a single trajectory [13]. Replication should be designed around the scientific claim rather than adopted as a ceremonial number of runs. A study of equilibrium populations may require independent exploration of relevant basins; a kinetic study may require independent transition information; an enhanced-sampling calculation may require independent bias histories or walkers; and a mechanistic study may require recurrence across distinct preparations or perturbations.

The relationship between independent replication, estimator-specific convergence, robustness to plausible alternatives, external validation, and the maximum defensible claim is organized in the proposed Claim-Strength Matrix (**Figure 2**).

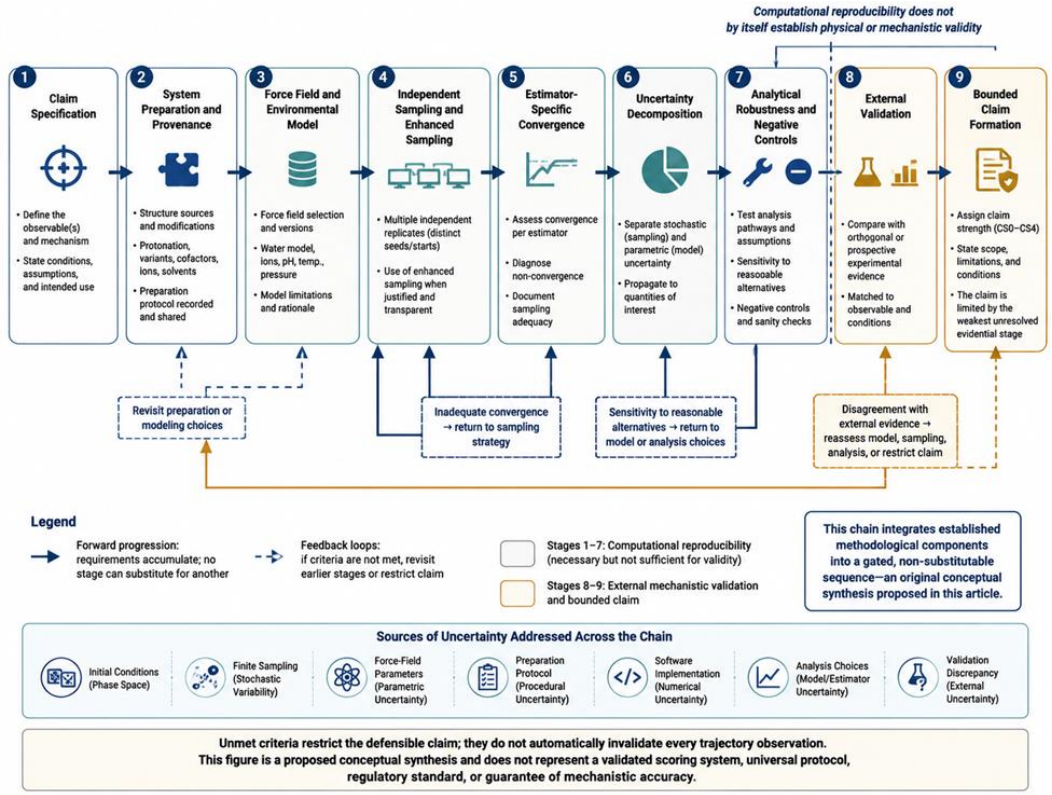

Figure 2. Claim-Strength Matrix Based on Replication, Convergence, Robustness, and Validation

Convergence should be evaluated for a prespecified estimator. Independent biased simulations can support convergence assessment of free-energy surfaces when the diagnostic is constructed around the estimated mean force or related target quantity [14]. Other claims may require state-population stability, kinetic-timescale consistency, block analysis, between-replica agreement, uncertainty intervals, or stability under additional sampling. No single diagnostic is universally sufficient.

Simulation duration and ensemble size may influence different components of accuracy, convergence, and precision [15]. A longer trajectory may permit slower decorrelation or additional transitions, whereas additional independent trajectories may improve estimation of between-realization variability. Their relative importance depends on the observable, correlation structure, metastability, and inference procedure. The framework therefore rejects universal prescriptions such as a minimum duration or fixed number of replicas.

Distinguishing exploratory visualization from confirmatory inference

Kinetic interpretation depends on feature selection, dimensionality reduction, and state-model construction, and integrated learning approaches do not eliminate the need to validate the representation used for inference [16]. A trajectory can be analyzed in Cartesian coordinates, internal coordinates, contacts, distances, torsions, hydration variables, or learned latent representations. Each representation privileges some distinctions and suppresses others.

Exploratory analysis is not methodologically inferior. It is appropriate when the purpose is to generate hypotheses, identify candidate coordinates, inspect unexpected interactions, or develop experimental questions. The problem arises when an exploratory representation is presented as confirmatory without independent testing.

Strong kinetic inference can emerge when simulation-derived estimates are confronted with independent kinetic measurements or prospectively testable perturbations [17]. The evidential value arises not from the visual plausibility of a transition but from correspondence between a prespecified computational quantity and information not used to construct the narrative.

Binding free energy, association or dissociation rates, transition pathways, and residence time are distinct quantities that require different estimators, assumptions, and validation strategies [18]. A ligand that remains bound throughout a trajectory has not thereby been shown to have a long residence time. A ligand that dissociates once has not yielded a reliable dissociation rate. A frequently observed pathway under a biased coordinate is not necessarily the dominant unbiased pathway.

The proposed visualization-boundary rule is therefore explicit: a conformational movie can demonstrate that a motion occurred within a selected realization and representation. Without additional evidence, it cannot establish how often the motion occurs, whether it is kinetically dominant, whether it is causally necessary, or whether it persists under alternative preparations, force fields, replicas, and analytical choices.

Confirmatory use of visualization should be linked to a defined statistical object. Examples include the fraction of replicas exhibiting a transition, the estimated occupancy of a state, a transition-path ensemble, a committor-related analysis, a distribution of pathway features, or a prospectively tested perturbational prediction. The movie then illustrates evidence rather than functioning as the evidence itself.

Reporting requirements for trajectories, parameters, code, and analytical choices

Shared coordinates are insufficient for reproducibility when topology, force-field definitions, preparation records, software information, and analytical metadata are unavailable [19]. A trajectory file without provenance may be impossible to interpret because the chemical identities, periodic box, virtual sites, constraints, atom ordering, or parameter assignments are unclear.

Standardization and accessible data structures can improve interoperability, inspection, and reuse of MD outputs [20]. The proposed minimum reporting bundle contains the following elements:

- Intended claim type and quantity of interest.
- Starting structures and biological-assembly decisions.
- Preparation transformations and chemical-state assignments.
- Force field, water model, ligand parameters, and parameter-generation procedures.
- Software, version, compiler or hardware-sensitive settings where relevant.
- Integration, thermostat, barostat, constraints, cutoffs, electrostatics, and boundary conditions.
- Minimization, equilibration, restraint, and production protocols.
- Definition and generation of independent replicas.
- Enhanced-sampling variables, bias parameters, walkers, windows, and reweighting procedures.

- Primary convergence diagnostic and its rationale.
- Uncertainty sources examined and unresolved.
- Primary analytical pipeline and reasonable alternatives.
- Negative controls and model-perturbation tests.
- Validation observable, matching conditions, and discrepancy treatment.
- Code, input files, processed data, and sufficient trajectory information for reuse.
- Deviations from the prespecified analytical plan.

Not every project can deposit every full-precision coordinate at every timestep. Storage and privacy or proprietary constraints may require reduced representations, curated subsets, or controlled access. Such limitations should be declared, and authors should provide the highest-value reproducibility objects: complete inputs, transformation scripts, analysis code, metadata, random seeds where meaningful, and enough data to verify the main claims.
Reporting should also preserve failed or contradictory replicas. Excluding trajectories because they do not support the preferred narrative can convert natural variability into apparent certainty. Exclusion criteria should be prespecified or justified through identifiable technical failure rather than inferential inconvenience.

Robustness tests and negative-control simulations
Cross-software comparison can reveal implementation dependence even when calculations are intended to represent the same free-energy problem [21]. Agreement across engines strengthens numerical reproducibility, whereas disagreement identifies protocol, parameter-conversion, estimator, or implementation choices requiring investigation. Neither outcome alone establishes physical accuracy.
Force-field selection can materially alter predictions and rankings and should be treated as model-choice uncertainty rather than as an invisible background decision [22]. In biomolecular settings, analogous perturbation tests may include alternative protein or ligand parameters, protonation states, water models, ion conditions, starting conformations, membrane compositions, or restraint-release procedures.
The proposed framework distinguishes four classes of robustness tests:

1. Preparation perturbations: plausible changes in protonation, tautomerism, unresolved-region modeling, cofactors, ions, or environmental composition.
2. Model perturbations**:** alternative force fields, parameter sets, water models, or boundary-condition implementations.
3. Sampling perturbations: new seeds, independent starting states, alternative enhanced-sampling schedules, or extended sampling targeted to unresolved regions.
4. Analytical perturbations: alternative features, clustering methods, state definitions, lag times, reweighting choices, smoothing procedures, or uncertainty estimators.

Negative controls ask whether the workflow can generate an apparently meaningful result when the proposed signal should be absent or altered. Depending on the system, controls may include neutral mutations, nonbinding analogues, scrambled labels, inactive collective variables, null perturbations, unrelated structural regions, or synthetic trajectories with known properties. A negative control is useful only when it challenges the same analytical or mechanistic logic used for the primary claim.
Failure of a robustness test does not automatically invalidate the simulation. It reveals conditionality. A force-field-dependent population estimate may still be reported as a model-specific result. An analysis-dependent pathway may remain an exploratory hypothesis. The required response is calibrated interpretation, not automatic suppression.

Claim-strength levels for molecular-dynamics studies
Binding-kinetics predictions require benchmark systems, appropriate experimental rate information, and evaluation strategies aligned with the kinetic quantity being claimed [23]. The absence of such evidence does not make kinetic simulation uninformative, but it limits whether the result should be described as illustrative, comparative, predictive, or mechanistically validated.
Allostery similarly includes multiple theoretical descriptions involving population shifts, structural coupling, dynamic redistribution, and energetic communication [24]. A pathway of correlated residues, a moving domain, or a change in fluctuation amplitude can support an allosteric hypothesis. It does not alone establish that the pathway is causal, necessary, sufficient, or functionally relevant.

The proposed Claim-Strength Matrix defines five levels:
CS0: Visual illustration
A selected trajectory segment or structural sequence illustrates a possible configuration or motion. No claim is made regarding population, rate, recurrence, or mechanism.

CS1: Descriptive trajectory observation
A feature is reported as having occurred under a fully documented model and protocol. The description remains conditional on that realization and avoids ensemble language.

CS2: Replicated ensemble association
Independent simulations support a reproducible distribution, association, state occupancy, or recurring qualitative behavior under the specified model. Uncertainty across realizations is reported.

CS3: Robust quantitative estimate
A prespecified quantitative target satisfies relevant convergence diagnostics and remains sufficiently stable under defensible analytical, preparation, and model perturbations. Residual uncertainty is explicitly bounded.

CS4: Externally validated mechanistic inference
A replicated and robust computational inference is supported by orthogonal or prospective evidence matched to the claimed observable, conditions, and causal interpretation.

The levels are cumulative and claim-specific. One article may contain a CS3 free-energy estimate, a CS2 conformational association, and a CS0 mechanistic animation. The highest level achieved by one claim should not be transferred automatically to every conclusion in the paper.
The matrix is not a numerical score. It does not assign weights, universal thresholds, or pass–fail cutoffs. Its function is linguistic and evidential: to align the strength of wording with the demonstrated inferential support.

Limitations and implementation barriers
Robust molecular simulation is limited by weaknesses in question formulation, system preparation, control design, data generation, analysis, and curation [25]. The proposed framework makes these dependencies visible but cannot eliminate them. It may also create additional burdens, including computational replication, storage, metadata preparation, alternative analyses, and reviewer expertise.
Open force-field development improves inspectability, communal testing, and the traceability of parameter-generation decisions but does not remove domain-of-applicability limitations [26]. Similarly, open code may contain systematic errors, automated workflows may reproduce flawed assumptions efficiently, and FAIR data may preserve an invalid model with excellent metadata.
The framework also faces proportionality challenges. A qualitative exploratory study should not be required to perform every validation step expected of a quantitative kinetic prediction. Small research groups may lack access to multiple software engines, specialized experimental collaborators, or storage for extensive trajectories. Journal policies that ignore these asymmetries could privilege computational scale over methodological reasoning—the very substitution error the framework is intended to prevent.
Checklist behavior is another risk. Authors may report replicas without demonstrating independence, include convergence plots unrelated to the claimed estimator, provide a repository without sufficient provenance, or perform alternative analyses chosen to preserve the same narrative. Compliance must therefore be evaluated functionally: does the supplied evidence address the uncertainty or failure mode for which it is presented?
The framework cannot specify universally optimal negative controls. A neutral mutation may not be chemically neutral, an alternative force field may differ in several dimensions simultaneously, and disagreement with an experimental observable may reflect condition mismatch rather than simulation failure. Validation itself contains uncertainty and model dependence.

CONCLUSION

Reproducible molecular-dynamics evidence cannot be reduced to trajectory duration, atom count, computational expense, or visual persuasiveness. These features may contribute to a rigorous study, but only when their evidential role is explicit. A long trajectory does not automatically establish independence, a large system does

not establish model adequacy, a stable estimate does not establish physical accuracy, and a compelling movie does not establish ensemble prevalence or causal mechanism.

The proposed Molecular-Dynamics Evidential Reproducibility Framework organizes system preparation, model choice, sampling, convergence, uncertainty, robustness, reporting, negative controls, validation, and claim formation as a connected chain. Its accompanying Claim-Strength Matrix restricts conclusions according to the evidence actually demonstrated. Both constructs are proposed conceptual syntheses rather than validated standards.

The framework's central decision rule is non-substitution: no single component can compensate for an unresolved evidential function elsewhere. Its practical objective is not to demand maximal computation from every study, but to align evidential burden with claim strength. Exploratory trajectories remain valuable when described as exploratory. Quantitative and mechanistic claims require progressively stronger evidence.

Future evaluation should determine whether the proposed framework improves reporting clarity, reviewer consistency, reproducibility, and the calibration of molecular-dynamics claims without creating disproportionate barriers. Until such evaluation is available, the framework should be treated as a structured basis for methodological reasoning rather than as a certification system.

ACKNOWLEDGMENTS: None

CONFLICT OF INTEREST: None

FINANCIAL SUPPORT: None

ETHICS STATEMENT: None

REFERENCES

1. Hollingsworth SA, Dror RO. Molecular dynamics simulation for all. Neuron. 2018;99(6):1129-43. doi:10.1016/j.neuron.2018.08.011
2. Husic BE, Pande VS. Markov state models: From an art to a science. J Am Chem Soc. 2018;140(7):2386-96. doi:10.1021/jacs.7b12191
3. Noé F, Tkatchenko A, Müller KR, Clementi C. Machine learning for molecular simulation. Annu Rev Phys Chem. 2020;71:361-90. doi:10.1146/annurev-physchem-042018-052331
4. Knapp B, Ospina L, Deane CM. Avoiding false positive conclusions in molecular simulation: The importance of replicas. J Chem Theory Comput. 2018;14(12):6127-38. doi:10.1021/acs.jctc.8b00391
5. Wan S, Sinclair RC, Coveney PV. Uncertainty quantification in classical molecular dynamics. Philos Trans A Math Phys Eng Sci. 2021;379(2197):20200082. doi:10.1098/rsta.2020.0082
6. Vassaux M, Wan S, Edeling W, Coveney PV. Ensembles are required to handle aleatoric and parametric uncertainty in molecular dynamics simulation. J Chem Theory Comput. 2021;17(8):5187-97. doi:10.1021/acs.jctc.1c00526
7. Schappals M, Mecklenfeld A, Kröger L, Botan V, Köster A, Stephan S, et al. Round robin study: Molecular simulation of thermodynamic properties from models with internal degrees of freedom. J Chem Theory Comput. 2017;13(9):4270-80. doi:10.1021/acs.jctc.7b00489
8. Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot BL, et al. CHARMM36m: An improved force field for folded and intrinsically disordered proteins. Nat Methods. 2017;14(1):71-3. doi:10.1038/nmeth.4067
9. Tian C, Kasavajhala K, Belfon KAA, Raguette L, Huang H, Migues AN, et al. ff19SB: Amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution. J Chem Theory Comput. 2020;16(1):528-52. doi:10.1021/acs.jctc.9b00591
10. PLUMED consortium. Promoting transparency and reproducibility in enhanced molecular simulations. Nat Methods. 2019;16(8):670-3. doi:10.1038/s41592-019-0506-8
11. Invernizzi M, Parrinello M. Rethinking metadynamics: From bias potentials to probability distributions. J Phys Chem Lett. 2020;11(7):2731-6. doi:10.1021/acs.jpclett.0c00497
12. Bussi G, Laio A. Using metadynamics to explore complex free-energy landscapes. Nat Rev Phys. 2020;2(4):200-12. doi:10.1038/s42254-020-0153-0
13. Wan S, Bhati AP, Wade AD, Coveney PV. Ensemble-based approaches ensure reliability and reproducibility. J Chem Inf Model. 2023;63(22):6959-63. doi:10.1021/acs.jcim.3c01654

14. Bjola A, Salvalaglio M. Estimating free-energy surfaces and their convergence from multiple, independent static and history-dependent biased molecular-dynamics simulations with mean force integration. J Chem Theory Comput. 2024;20(13):5418-27. doi:10.1021/acs.jctc.4c00091
15. Lai TT, Brooks CL III. Accuracy and reproducibility of Lipari-Szabo order parameters from molecular dynamics. J Phys Chem B. 2024;128(44):10813-22. doi:10.1021/acs.jpcb.4c04895
16. Mardt A, Pasquali L, Wu H, Noé F. VAMPnets for deep learning of molecular kinetics. Nat Commun. 2018;9(1):5. doi:10.1038/s41467-017-02388-1
17. Paul F, Wehmeyer C, Abualrous ET, Wu H, Crabtree MD, Schöneberg J, et al. Protein-peptide association kinetics beyond the seconds timescale from atomistic simulations. Nat Commun. 2017;8(1):1095. doi:10.1038/s41467-017-01163-6
18. Decherchi S, Cavalli A. Thermodynamics and kinetics of drug-target binding by molecular simulation. Chem Rev. 2020;120(23):12788-833. doi:10.1021/acs.chemrev.0c00534
19. Abraham MJ, Apostolov RP, Barnoud J, Bauer P, Blau C, Bonvin AMJJ, et al. Sharing data from molecular simulations. J Chem Inf Model. 2019;59(10):4093-9. doi:10.1021/acs.jcim.9b00665
20. Caparotta M, Perez A. Advancing molecular dynamics: Toward standardization, integration, and data accessibility in structural biology. J Phys Chem B. 2024;128(10):2219-27. doi:10.1021/acs.jpcb.3c04823
21. Loeffler HH, Bosisio S, Duarte Ramos Matos G, Suh D, Roux B, Mobley DL, et al. Reproducibility of free energy calculations across different molecular simulation software packages. J Chem Theory Comput. 2018;14(11):5567-82. doi:10.1021/acs.jctc.8b00544
22. McCready C, Sladekova K, Conroy S, Gomes JRB, Fletcher AJ, Jorge M. Quantifying the uncertainty of force field selection on adsorption predictions in MOFs. J Chem Theory Comput. 2024;20(11):4869-84. doi:10.1021/acs.jctc.4c00287
23. Nunes-Alves A, Kokh DB, Wade RC. Recent progress in molecular simulation methods for drug binding kinetics. Curr Opin Struct Biol. 2020;64:126-33. doi:10.1016/j.sbi.2020.06.022
24. Wodak SJ, Paci E, Dokholyan NV, Berezovsky IN, Horovitz A, Li J, et al. Allostery in its many disguises: From theory to applications. Structure. 2019;27(4):566-78. doi:10.1016/j.str.2019.01.003
25. Brown AM, Lemkul JA. Robustness in biomolecular simulations: Addressing challenges in data generation, analysis, and curation. Cell Rep Phys Sci. 2025;6(5):102566. doi:10.1016/j.xcrp.2025.102566
26. Wang L, Behara PK, Thompson MW, Gokey T, Wang Y, Wagner JR, et al. The Open Force Field Initiative: Open software and open science for molecular modeling. J Phys Chem B. 2024;128(29):7043-67. doi:10.1021/acs.jpcb.4c01558
27. Gygli G, Pleiss J. Simulation Foundry: Automated and F.A.I.R. molecular modeling. J Chem Inf Model. 2020;60(4):1922-7. doi:10.1021/acs.jcim.0c00018
28. Amaro RE, Åqvist J, Bahar I, Battistini F, Bellaiche A, Beltran D, et al. The need to implement FAIR principles in biomolecular simulations. Nat Methods. 2025;22(4):641-5. doi:10.1038/s41592-025-02635-0