THE JOURNAL OF PHYSICAL CHEMISTRY B
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCB

Perspective

Advancing Molecular Dynamics: Toward Standardization, Integration, and Data Accessibility in Structural Biology

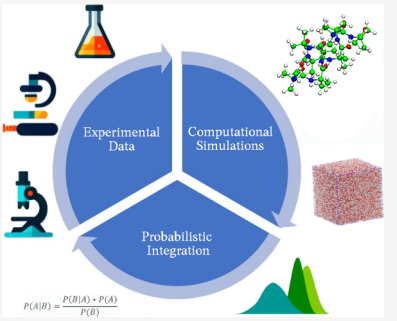
Marcelo Caparotta and Alberto Perez*

Cite This: https://doi.org/10.1021/acs.jpcb.3c04823

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Molecular dynamics (MD) simulations have become a valuable tool in structural biology, offering insights into complex biological systems that are difficult to obtain through experimental techniques alone. The lack of available data sets and structures in most published computational work has limited other researchers' use of these models. In recent years, the emergence of online sharing platforms and MD database initiatives favor the deposition of ensembles and structures to accompany publications, favoring reuse of the data sets. However, the lack of uniform metadata collection, formats, and what data are deposited limits the impact and its use by different communities that are not necessarily experts in MD. This Perspective highlights the need for standardization and better resource sharing for processing and interpreting MD simulation results, akin to efforts in other areas of structural biology. As the field moves forward, we will see an increase in popularity and benefits of MD-based integrative approaches combining experimental data and simulations through probabilistic reasoning, but these too are limited by uniformity in experimental data availability and choices on how the data are modeled that are not trivial to decipher from papers. Other fields have addressed similar challenges comprehensively by establishing task forces with different degrees of success. The large scope and number of communities to represent the breadth of types of MD simulations complicates a parallel approach that would fit all. Thus, each group typically decides what data and which format to upload on servers like Zenodo. Uploading data with FAIR (findable, accessible, interoperable, reusable) principles in mind including optimal metadata collection will make the data more accessible and actionable by the community. Such a wealth of simulation data will foster method development and infrastructure advancements, thus propelling the field forward.



INTRODUCTION

Since their development in the 1970s, molecular dynamics-based approaches have significantly advanced and gained credibility as a tool for hypothesis testing, experimental motivation and interpretation, and enhancing our understanding of biology.[1] These approaches, rooted in physical principles, are attractive for addressing various problems,[2] including evaluating free energy differences,[3] predicting structures,[4] unraveling mechanisms of action,[5,6] and providing high-resolution descriptions of systems on a picosecond-by-picosecond basis, which cannot be achieved through experimental techniques alone.

However, despite the achievements of these computational models, they encounter three primary limitations.[7] First, the accuracy of predictions relies on the quality of force fields employed.[8] Second, achieving adequate sampling can be computationally expensive.[9] Last, scalability becomes a challenge when dealing with the vast number of systems of interest in biology.[10] For instance, considering only natural amino acids, the protein sequence space grows exponentially as 20^N, where N is the sequence length, considering only natural amino acids.

In the early years of MD, the disparity between the time scales relevant to biological processes and those accessible to computer simulations and the limited accuracy of force fields limited the real-life application of this promising technique.[11,12] However, these limitations spurred the development of an impressive range of enhanced sampling strategies, which now form the basis of many successful approaches.[13] More recently, force field development—once reserved to a few expert groups—has seen an expansion both in terms of niche specific force field development and in the number of groups contributing new parameters.[14] With growing computer power, improved sampling techniques, and force fields, the systems of interest also grow in complexity, often requiring even longer simulations and more complex analysis tools.[15] Many such simulations are beyond what any group could sample on their own, using national and international supercomputer resources. Access of such simulations to the broader community,[16] as happened with SARS-CoV-2 related

Received: July 18, 2023
Revised: February 17, 2024
Accepted: February 21, 2024

simulations during the pandemic (e.g., https://bioexcel-cv19.bsc.es/, https://covid.bioexcel.eu/simulations/, or https://covid.molssi.org/), could spur development of new analysis methods and new biological insights. Such efforts have garnered the support of the community by contributing over 10,000 simulations and over 12.7 ms of simulation data[17] leading to an array of publications on the topic of SARS-CoV-2.

However, there are no clear standards for how to share simulation data (e.g., processed or raw data), what information (metadata) is provided to search and find such simulations, and whether processed data (e.g., structures of top clusters or analysis) are provided, or how to assess the quality and biological insights in different data sets, which limits its use by different communities. Online resources such as the open science framework,[18] Figshare,[19] or Zenodo[20] simplify the process of storing trajectories by uploading projects up to 50Gb that can be identified and cited through DOIs, foregoing some of the long-term limitations of storage requirements for individual groups or finding published trajectories once a student or postdoc moves to a new position.[21]

The validation of molecular dynamics-based approaches against experimental techniques has played a significant role in their development, resulting in their integration as routine tools in an expanding number of laboratories, including both computational and experimental settings.[22] Following George Box's notion that "All models are wrong, but some are useful",[23] the traditional MD paradigm is that if simulations reproduce some known experimental data for a particular system, then other conclusions derived from the simulation data might also shine new insights into the system. Though improved force fields, longer accessible time scales, and the ability to simulate larger and more complex systems, the paradigm is shifting for simulations to be more predictive and yielding testable hypothesis.[22] These developments are allowing truly integrative approaches that combine experiments and simulations[24−28] to solve problems that neither experiments nor simulation could solve on their own. Such integrative approaches typically employ probabilistic reasoning to combine experiments and simulations while maintaining a physical foundation and interpretability of the resulting ensembles.

To build on such findings, other members of the structural community need access to the models: as ensembles, trajectories, or, in some cases, single structures of the relevant biological states described by the trajectories (e.g., cluster centers). The heterogeneity in data reporting, data sources, and how the data are modeled gives rise to many different protocols[29−31]—each of which might answer different questions. While more groups are making information from simulations accessible there is a need for common standards[32] that will agree with FAIR principles (Findable, Accessible, Interoperable, and Reusable).[33]

Standardized formats and centralized databases have been crucial to the development in other domains of structural biology,[34,35] and more recently they have been the source for training the very AI models that have surpassed other tools in the field (e.g., in protein structure prediction).[36,37] Success of such initiatives often relies in the establishment of task forces[38−43] built from heterogeneous teams across the field that can represent different points of view and needs of the community and can drive enforcement of protocols (such as formats, validation, and metadata collection). Such task forces help in establishing standards and promoting best practices for reporting results and structures, which can adapt as the field requires.

Historically, such MD initiatives that tried to centralize MD trajectories[44−46] were dependent on the efforts of single groups (e.g., MODEL,[45] Dynameomics[44]). While this allowed homogeneous data collection, system setup, and running conditions, it required a large investment from a single group and updating trajectories to simulation lengths that are appropriate for current simulation lengths. More recent initiatives[32] are typically more focused on specific types of systems such as DNA,[46] GPCRs,[47] MemProtMD,[48] and NMRlipids[49] and often include simulations from multiple groups, programmatic access to data, standardized analysis, and metadata collection, while other recent initiatives, such as MDDB,[17] aim for an umbrella repository of simulation data. Alternatively, many individual groups upload their simulation data to servers like Zenodo, Figshare, or OSF, relying on those engines' searchable libraries. In such cases, there are no uniform standards for sharing the data.

DATABASES IMPROVE COMMUNITY REUSE OF MD SIMULATIONS

A common concern among our structural biology collaborators is the perceived lack of actionable output in MD papers, often providing descriptive simulations without offering usable data, such as coordinates for other communities. The great success of the PDB[50] lies in providing a common starting point for various modeling approaches irrespective of what technique (NMR, X-ray, CryoEM) was used to derive the structure. The metadata collected in the PDB usually provide insights as to what considerations to take into place when modeling the system, for example, capturing the effect of crystal contacts in system stability, generating missing residues, or removing purification tags before using the structure. Different modeling communities (e.g., Docking, MD, ...) learn to use and modify those initial structures as starting points to answer biological questions. Even beyond its original intent, the PDB has provided the collective data to develop AI-based technologies that predict protein structures and even complexes.[36,37]

Different communities have varied expectations of MD simulations. Some seek practical insights to advance understanding, while others focus on system setup, interoperability, reproducibility, and method development. Structural biologists, for instance, might prioritize obtaining representative structures of the most relevant states (e.g., open/close states, active/inactive states) and emphasize experimental validation of those states before analyzing and obtaining biological data, rather than studying the whole ensemble. In this sense, knowing the purpose of the simulation informs users about what the ensembles might capture. For example, simulations using adaptive sampling strategies might unveil kinetic relations between different states,[9] long equilibrium trajectories at melting temperature might give information about folding pathways and intermediates,[51] while MD simulations starting from sequence alone (e.g., from blind competitions such as CASP[52]) are limited at best to predict the major (native) state.

The MD community, with greater access to computational resources, is keen on comparing ensembles from different force fields, developing enhanced sampling approaches that can be benchmarked against state-of-the-art simulations, and developing analysis tools. Storage requirements, crucial for ensembles, can be managed by storing dry trajectories, down

https://doi.org/10.1021/acs.jpcb.3c04823
J. Phys. Chem. B XXXX, XXX, XXX−XXX

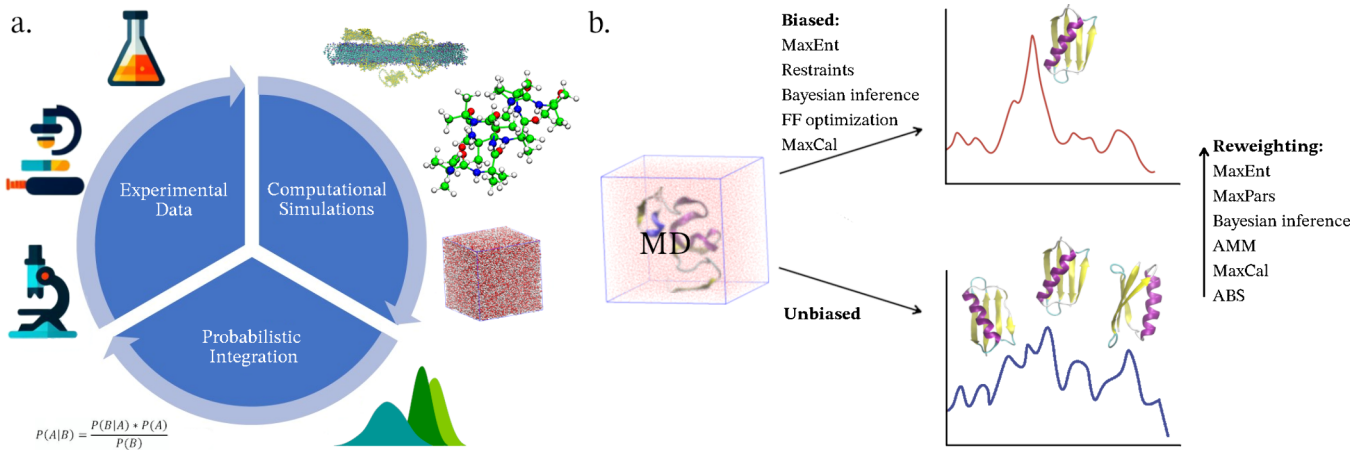


Figure 1. Complementary nature of computational simulations and biophysical experiments in integrative structural biology. a) Experimental techniques are utilized to refine computational simulations through the application of probabilistic reasoning. This integration of experimental data enhances the accuracy and reliability of the simulations. b) Biasing or reweighting MD simulations with experimental data. When using biases (top), the system directly samples a distribution that aligns with the experimental data. Alternatively (bottom), unbiased simulations lead to a population distribution that is later reweighted against experimental data.[73]

sampling, or maintaining key structures with velocities for trajectory restarts. With current computer power and deterministic algorithms,[53] it is possible to reproduce a trajectory given n intermediate steps, with the caveat that some MD software will require the same computer architecture for true reproducibility. However, even in this scenario factors like software and library versions or the effect of load balancing can produce different results. Since simulations are stochastic, most of the time a user might not need the exact trajectory but the conclusions that can be extracted from the ensemble. For example, some initiatives like Folding@home[54] opt to share a Markov state model of the data where the metastable states and kinetic relations between them are given.

Often, papers addressing new analysis tools will be tested against *in house* simulations run by the group developing the analysis, and often reviewers will ask for longer simulations on more systems. This is a particular scenario where reusing community tested and verified trajectories might save a lot of time in choosing interesting systems and running for long enough.

LACK OF INTEROPERABILITY ACROSS MD PLATFORMS REQUIRES MORE INFORMATIVE METADATA

Bioinformatics tools have historically had a large appeal to the structural biology community because they are fast, easy to setup, and provide an interpretable output. More recently, AlphaFold (AF)[37] and other AI programs[36] have taken the scientific community by surprise with its spectacular success for protein structure determination.[36,37] Beyond its success, it has been rapidly embraced by the structural biology community thanks to its simple input, concise output, and easy-to-gray confidence score for the predictions. AF is useful for many aspects of molecular modeling and structural biology and has already been incorporated into many experimental pipelines.[55−58] Beyond protein structure prediction, AF also has some degree of transferability, with great potential for predicting protein−protein[59] and protein−peptide[60] complexes—and even ranking peptides by binding affinity.[61]

However, while tools like AlphaFold excel at a specific task (structure prediction), MD simulation protocols undergo significant changes depending on the research question—such as sampling around a native protein state, describing processes like binding or folding, or predicting free energy differences between different drugs. All simulations will typically require three steps: 1) system setup, 2) equilibration, 3) production runs.[62] The first step helps introduce the physics model, solvents, and other molecules that are going to define the system. Tools such as Charmm-GUI[63] and others[64,65] facilitate this interconversion[66,67] from an initial structure to the generation of the simulation system. However, not everything is accounted for; for example, SARS-CoV-2 exposed many expert groups the challenges in modeling glycans into proteins, which lead to new protocols introduced in these pipelines. The equilibration process refers to reaching the simulation temperature, equilibrating the system to the correct density, and relaxing the solvent/solute system such that the initial conditions do not artificially drive the solute away from the initial starting state. The production is what generates the ensemble that will be analyzed for insights. Depending on the purpose of the simulations, conventional or enhanced sampling approach, the form of the potential energy,[68] and simulation package, the conditions of the simulation will be different, and it is still frequent to find suboptimal parameter choices in the literature (e.g., Berendsen thermostat/barostat). Furthermore, not all simulation packages can run all force fields, and some of the simulation parameters do not have the exact counterparts in other packages. Thus, reproducing or reanalyzing simulation data with a different package than the one used to generate the data is not possible (e.g., recalculate the energy for every frame in an ensemble).[68] This lack of interoperability[68,69] requires specific attention when deciding what metadata users will need to reproduce or reanalyze ensembles.

Given the wide range of questions that simulations can address, there is no unique protocol that will work for all systems. Here, the value of developing and maintaining up-to-date tutorials to exemplify good simulation practices becomes helpful. In particular, when these questions are posed, robust and transferable protocols can be established. A prime example is the use of alchemical free energy methods in drug discovery: how does a chemical change (ligand or protein mutant) affect the stability of the system ($\Delta G_{mutation}$)? Protocols for such

https://doi.org/10.1021/acs.jpcb.3c04823
J. Phys. Chem. B XXXX, XXX, XXX−XXX

methods are now well established to yield a free energy value with calculated error bars and have been incorporated into many standard computational chemistry programs, often geared toward big pharma.[70,71] However, tutorials cannot cover every possible scenarios. Including all metadata needed to reproduce a project in simulations will allow users to transfer simulation protocols from specific projects into their pipelines. This becomes especially crucial for new methods that have not been tested as thoroughly as more established ones or for which tutorials are not available.

POTENTIAL OF MD SIMULATIONS IN INTEGRATIVE STRUCTURAL BIOLOGY

In its evolution into a well-established technique, MD has traversed various stages,[1] progressing from using data to validate specific simulation aspects to becoming a predictive technique driving insights and hypotheses that are subsequently experimentally validated.[22] Concurrently, integrative and hybrid modeling approaches (Figure 1a) have emerged to solve problems that neither experiments nor computational approaches could resolve on their own. Integrative structural biology grapples with data fraught with inherent uncertainties and noise, such as experimental errors or limited resolution.[35,72] Probabilistic reasoning allows for a systematic and quantitative assessment of these uncertainties, providing a more accurate understanding of the reliability and limitations of the structural models generated (Figure 1b). However, determining the number of states represented by the data or reconciling different sources of data compatible with distinct states poses a challenge. MD, with its natural sampling of diverse states, provides a robust framework for modeling such ambiguous and noisy data. While MD has historically incorporated certain types of information (e.g., single interresidue distances), truly integrative MD approaches have only recently become commonplace.[72]

Some of the challenges in this field are the heterogeneity in experimental data reporting originating from different systems and laboratories, heterogeneity in how different modeler groups handle the data (e.g., some data might be ignored), and how the final models are reported (e.g., a structure, an ensemble, interpretation of the data, modeling of the errors, ...). Such heterogeneity ultimately leads to a diverse range of protocols and techniques, typically associated with different computational laboratories, using similar probabilistic integration techniques (e.g., Bayesian inference[74,75]) to answer a diverse set of biological questions.[26,27,73,76,77] Every new project requires a deep understanding of the data, how to interpret it and apply it to simulations, or how to use it to reweight MD ensembles. Users who access MD data need to understand exactly how the experimental information was modeled, which might be hard to obtain from a paper. This can be done by providing more verbose metadata where the original experimental information, how it is modeled, and with which software (and version) is provided.

The challenges increase when simulating dynamic and flexible systems, such as protein conformational changes, protein−ligand interactions, or integrating structural information across different scales, from atomic to cellular levels.[30,31,78] Below, we provide two case examples highlighting the challenges discussed in this context, focusing on structure determination using sparsely labeled NMR or CryoEM data.

In the case of sparsely labeled NMR,[79] the challenge lies in identifying possible interpretations of the NOESY spectra based on the chemical shifts of different atoms. Ambiguity and noise arise as a signal in the spectra can have multiple interpretations, some peaks may not be visible, and some regions along the sequence might be blind to the experiment (i.e., no label in those residues). Traditional NMR structure determination efforts encounter difficulties due to the numerous viable structures arising from data interpretations. By combining the sparsely labeled NMR data set with simulations, typically using Bayesian inference, structures that are both physically compatible with the force field and representative of the ensemble's Boltzmann distribution can be obtained.[80−83] However, this approach does not provide information about the uncertainty associated with each peak in the data set. A different use of Bayesian inference is often observed in CryoEM data analysis. Here, the initial state is already close to the native state, as revealed by the density map. However, the CryoEM data set does not possess uniform resolution, leading to uneven distribution of uncertainties throughout the structure. In this scenario, the objective is to evaluate the uncertainty ensemble based on the collected data and simulation framework.[24,84] Furthermore, the data might contain information about different conformers which can be rationalized in terms of different states and their associated populations.[85]

BRINGING TOGETHER MULTIPLE POINTS OF VIEW FOR BETTER USE OF SIMULATION DATA

From the early database initiatives championed by individual groups to the more contemporary community efforts that allow interactive and programmatic data access, the field has been steadily advancing methods to share and facilitate greater interaction with MD data. As computational power has grown, so have MD databases, expanding from under 10 terabytes in 2010 to hundreds of terabytes in recent years.[21] The infrastructure, storage capacity, and resources for interactive analysis of these databases is also improving. Early projects required physical shipment of data disks among consortium methods, while more recent and larger projects can exchange information rapidly (e.g., through globus). Maintaining storage requirements for such databases is increasingly challenging in a centralized manner, necessitating multiple distribution nodes. Projects related to COVID,[16] for instance, have found sponsorship of storage space through either molSSI or AWS (https://registry.opendata.aws/foldingathome-covid19/), this type of initiative does not scale to all possible projects and other sources of funding are needed.

To overcome these challenges, the community is now witnessing the emergence of new funding initiatives and consortiums that aim to build scalable systems for MD simulation data,[21] and possibly for on-demand analysis.[86] One of the challenges in depositing data is making sure credit is given where it is deserved: that is, the database should be citable but also the contributions from different users contributing trajectories.[87,88] For example, the set of protein folding trajectories for a small set of fast protein folders have been a standard in the field that have been reused and analyzed well beyond the initial paper.[51] Access to the trajectories was granted based on requests by individuals, and papers cited the original publication. In the new paradigm, such trajectories would already be available to the community, but there should be a mechanism by which using them would still give recognition to the original creators rather than just the

https://doi.org/10.1021/acs.jpcb.3c04823
J. Phys. Chem. B XXXX, XXX, XXX−XXX

database. Such recognition is critical when analyzing the impact of one's work or when requesting funding to agencies.[87]

Simultaneously, recent years have seen an increase of trajectories being deposited by individual users in services like Zenodo, Figshare, or the Open Science Framework (OSF), which provide storage capacities (up to 50Gb per project) and a citable DOI. A recent analysis[89] of data deposited in these servers revealed close to 2,000 MD data sets (March 2023) for a cumulative 14 Tb of data. While these data were accessible and downloadable, finding it is not trivial, which reinforces the need for better standardization of metadata to identify and classify data. These data sets also vary in how they store data: how often data are stored, if solvent is stored, and velocities. Some of the data sets thus allow us to restart trajectories while others do not. Despite current limitations, some studies are already making use of this wealth of data to reach conclusions that would be hard to replicate by the efforts of a single group.[90] Other approaches such as Folding@Home,[91] which generate a large number of trajectories for a project, have opted for depositing Markov State Models derived from the simulations in the Open Science Framework (OSF), which are much compressed with respect to the original trajectories, while keeping the original trajectories available on demand.

Throughout this Perspective, we have underscored the need for effective metadata that will enable users from different communities to search and retrieve data sets of interest, download relevant information for their communities (e.g., single structures, ensembles, or input data), and identify whether simulations have been compared with experiments and, if so, how they have been integrated. We have also addressed how data were modeled and the nature of the original data. These are just a few of the things that we believe are important. Often, the specifics of a simulation, including the type of simulation, the data format, and the modeling approach, are unique to a particular group and cannot typically be included in all detail in the manuscripts themselves. Therefore, data repositories play a critical role in bridging this gap by providing detailed descriptions (actual input) of how each simulation was conducted. Including such details would enable users from various communities—not just those who run simulations—to grasp the purpose and significance of these simulations and to use them according to their needs. This aligns with the FAIR guidelines.[33]

We can learn from existing communities (Figure 2) that have faced similar challenges related to data formatting and integration. For instance, the integrative/hybrid modeling community has established a database (PDB-dev) that endorses a federated system[92] linking separate pieces of information—meaning that not all information is stored in a single location. This could include, for example, both the models and experimental data. By synchronizing efforts, unifying formats, and fostering effective communication, the PDB-dev database offers capabilities that extend beyond those of the traditional PDB (e.g., multiscale resolution). Such initiatives have greatly benefitted from community workshops to identify needs/solutions and task forces that help oversee implementation (see Figure 2).

Considering the larger volumes of data generated by simulations, a federated system, where data trajectories are not centralized in one repository, appears to be a more practical solution. Some initiatives are already exploring the use of MD simulation data in artificial intelligence.[93,94] The availability of informative metadata will be essential in building

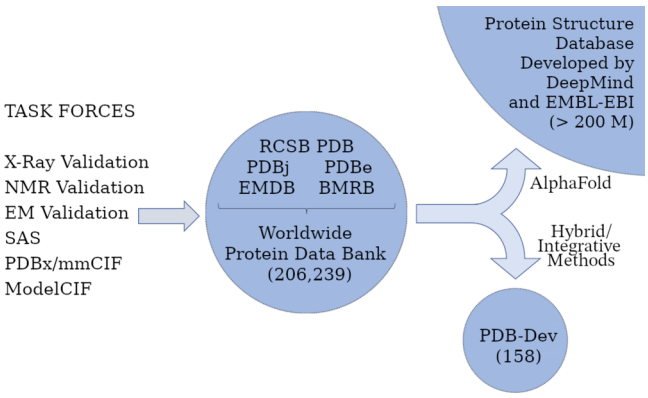


Figure 2. Task forces play a crucial role in directing, unifying, and standardizing data acquisition and structure representation efforts, leading to the success of the Protein Data Bank (PDB). The establishment of the PDB has subsequently paved the way for recent advancements in artificial intelligence, exemplified by the remarkable achievements of AlphaFold.[95] Additionally, the task force for hybrid/integrative structural biology has been instrumental in expanding the PDB-Dev archiving system, further enhancing the collaborative and interdisciplinary nature of structural biology research.[38] Numbers in parentheses indicate the total count of structures deposited as of June 20, 2023.

robust data sets. This will involve filtering simulations based on various criteria, such as the type of MD (such as conventional MD), the force fields used, the duration of simulations, and the specific simulation conditions. More advanced features will include programmatic access to databases, requiring access to software capable of processing trajectories on the fly. This type of initiative is, for example, done in CASP where all the models submitted are analyzed by an ever-expanding list of tested computational tools.[96]

CONCLUSION AND OUTLOOK

In conclusion, the advancements in molecular dynamics (MD) simulations have paved the way for new insights and discoveries in structural biology. However, the field faces challenges that hinder its full potential. The limitations of MD simulations, such as accuracy, computational expense, and data interpretation, necessitate the need for standardization, integration, and data accessibility. Promoting data sharing and reproducibility can drive the development of better interconversion tools to overcome formatting differences and new protocols for compiling and sharing metadata. This will lead to simulation data that is easier to find and more reusable by the community, thus adhering to FAIR principles. Moreover, user-friendly software tools, up-to-date tutorials, and MD databases will empower researchers to implement protocols effectively and analyze simulations efficiently. By embracing Open Science principles and fostering collaboration, the structural biology community can overcome these challenges, leading to more robust integrative approaches and a deeper understanding of complex biological systems.

AUTHOR INFORMATION

Corresponding Author

Alberto Perez − *Department of Chemistry and Quantum Theory Project, University of Florida, Gainesville, Florida 32611, United States;* orcid.org/0000-0002-5054-5338; Email: perez@chem.ufl.edu

https://doi.org/10.1021/acs.jpcb.3c04823
J. Phys. Chem. B XXXX, XXX, XXX−XXX

Author

Marcelo Caparotta − *Department of Chemistry and Quantum Theory Project, University of Florida, Gainesville, Florida 32611, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpcb.3c04823

Author Contributions

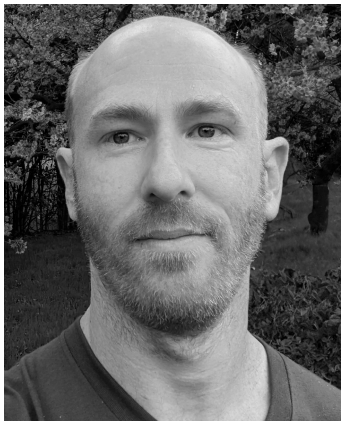
The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

Biographies

Marcelo Caparotta is a Postdoc Associate in the Department of Chemistry and Quantum Theory Project at University of Florida, USA. He obtained his Ph.D. from Universidad Nacional de Cuyo (UNCUYO), Argentina, in 2022. His research interests are currently focused on enhanced sampling techniques in molecular dynamics, protein−ligand binding, and deep learning.



Alberto Perez obtained his Ph.D. from the University of Barcelona in 2008, supported by a Generalitat de Catalunya Fellowship. He conducted his research on molecular simulations of nucleic acids under the guidance of Modesto Orozco and F. Javier Luque. Subsequently, he was awarded an EMBO fellowship for his postdoctoral work on protein folding with Ken Dill, first at UCSF and later at Stony Brook University, where he became a Laufer Center Junior Fellow. Alberto joined the Chemistry department at the University of Florida in 2018 as an assistant Professor to work in advancing molecular simulations of proteins and nucleic acids. Since the start of his independent career, he has received prestigious awards including the NSF CAREER Award in 2022 and the ACS COMP Cadence/OpenEye Outstanding Junior Faculty Award in 2023.

■ ACKNOWLEDGMENTS

A.P. thanks support from the NSF CAREER award CHE-2235785 and University of Florida startup funds.

■ REFERENCES

(1) Schlick, T.; Portillo-Ledesma, S. Biomolecular Modeling Thrives in the Age of Technology. *Nat. Comput. Sci.* **2021**, *1* (5), 321−331.

(2) Brini, E.; Simmerling, C.; Dill, K. Protein Storytelling through Physics. *Science* **2020**, *370* (6520), No. eaaz3041.

(3) Wang, L.; Wu, Y. J.; Deng, Y. Q.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.* **2015**, *137* (7), 2695−2703.

(4) Perez, A.; Morrone, J. A.; Simmerling, C.; Dill, K. Advances in Free-Energy-Based Simulations of Protein Folding and Ligand Binding. *Curr. Opin Struc Biol.* **2016**, *36*, 25−31.

(5) Di Bartolo, A. L.; Caparotta, M.; Masone, D. Intrinsic Disorder in A-Synuclein Regulates the Exocytotic Fusion Pore Transition. *ACS Chem. Neurosci.* **2023**, *14* (11), 2049−2059.

(6) De Vivo, M.; Masetti, M.; Bottegoni, G.; Cavalli, A. Role of Molecular Dynamics and Related Methods in Drug Discovery. *J. Med. Chem.* **2016**, *59* (9), 4035−4061.

(7) Schlick, T.; Portillo-Ledesma, S.; Myers, C. G.; Beljak, L.; Chen, J.; Dakhel, S.; Darling, D.; Ghosh, S.; Hall, J.; Jan, M.; Liang, E.; Saju, S.; Vohr, M.; Wu, C.; Xu, Y.; Xue, E. Biomolecular Modeling and Simulation: A Prospering Multidisciplinary Field. *Annu. Rev. Biophys* **2021**, *50* (1), 267.

(8) Chmiela, S.; Sauceda, H. E.; Müller, K.-R.; Tkatchenko, A. Towards Exact Molecular Dynamics Simulations with Machine-Learned Force Fields. *Nat. Commun.* **2018**, *9* (1), 3887.

(9) Bowman, G. R. Accurately Modeling Nanosecond Protein Dynamics Requires at Least Microseconds of Simulation. *J. Comput. Chem.* **2016**, *37* (6), 558−566.

(10) Dryden, D. T. F.; Thomson, A. R.; White, J. H. How Much of Protein Sequence Space Has Been Explored by Life on Earth? *J. R. Soc. Interface* **2008**, *5* (25), 953−956.

(11) McCammon, J. A.; Gelin, B. R.; Karplus, M. Dynamics of Folded Proteins. *Nature* **1977**, *267* (5612), 585−590.

(12) Levitt, M.; Warshel, A. Computer-Simulation of Protein Folding. *Nature* **1975**, *253* (5494), 694−698.

(13) Hénin, J.; Lelièvre, T.; Shirts, M. R.; Valsson, O.; Delemotte, L. Enhanced Sampling Methods for Molecular Dynamics Simulations [Article v1.0]. *Living J. Comput. Mol. Sci.* **2022**, *4* (1), 1583.

(14) Mondal, A.; Chang, L.; Perez, A. Modelling Peptide−Protein Complexes: Docking, Simulations and Machine Learning. *Qrb Discov* **2022**, *3*, 1−54.

(15) Dommer, A.; Casalino, L.; Kearns, F.; Rosenfeld, M.; Wauer, N.; Ahn, S.-H.; Russo, J.; Oliveira, S.; Morris, C.; Bogetti, A.; Trifan, A.; Brace, A.; Sztain, T.; Clyde, A.; Ma, H.; Chennubhotla, C.; Lee, H.; Turilli, M.; Khalid, S.; Tamayo-Mendoza, T.; Welborn, M.; Christensen, A.; Smith, D. G.; Qiao, Z.; Sirumalla, S. K.; O'Connor, M.; Manby, F.; Anandkumar, A.; Hardy, D.; Phillips, J.; Stern, A.; Romero, J.; Clark, D.; Dorrell, M.; Maiden, T.; Huang, L.; McCalpin, J.; Woods, C.; Gray, A.; Williams, M.; Barker, B.; Rajapaksha, H.; Pitts, R.; Gibbs, T.; Stone, J.; Zuckerman, D. M.; Mulholland, A. J.; Miller, T.; Jha, S.; Ramanathan, A.; Chong, L.; Amaro, R. E. #COVIDisAirborne: AI-Enabled Multiscale Computational Microscopy of Delta SARS-CoV-2 in a Respiratory Aerosol. *Int. J. High Perform. Comput. Appl.* **2023**, *37* (1), 28−44.

(16) Amaro, R. E.; Mulholland, A. J. A Community Letter Regarding Sharing Biomolecular Simulation Data for COVID-19. *J. Chem. Inf. Model.* **2020**, *60* (6), 2653−2656.
(17) Beltrán, D.; Hospital, A.; Gelpí, J. L.; Orozco, M. A New Paradigm for Molecular Dynamics Databases: The COVID-19 Database, the Legacy of a Titanic Community Effort. *Nucleic Acids Res.* **2024**, *52*, D393−D403.
(18) Foster, E. D.; Deardorff, A. Open Science Framework (OSF). *J. Méd. Libr. Assoc.* **2017**, *105* (2), 203−206.
(19) Singh, J. FigShare. *J. Pharmacol. Pharmacother.* **2011**, *2* (2), 138−139.
(20) Sicilia, M.-A.; García-Barriocanal, E.; Sánchez-Alonso, S. Community Curation in Open Dataset Repositories: Insights from Zenodo. *Procedia Comput. Sci.* **2017**, *106*, 54−60.
(21) Hospital, A.; Battistini, F.; Soliva, R.; Gelpí, J. L.; Orozco, M. Surviving the Deluge of Biosimulation Data. *WIREs Comput. Mol. Sci.* **2020**, *10* (3), e1449.
(22) Hollingsworth, S. A.; Dror, R. O. Molecular Dynamics Simulation for All. *Neuron* **2018**, *99* (6), 1129−1143.
(23) Box, G. E. P. Science and Statistics. *J. Am. Stat. Assoc.* **1976**, *71* (356), 791−799.
(24) Bonomi, M.; Camilloni, C.; Cavalli, A.; Vendruscolo, M. Metainference: A Bayesian Inference Method for Heterogeneous Systems. *Sci. Adv.* **2016**, *2* (1), e1501177−e1501177.
(25) Boomsma, W.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K. Combining Experiments and Simulations Using the Maximum Entropy Principle. *Plos Comput. Biol.* **2014**, *10* (2), No. e1003406.
(26) Crehuet, R.; Buigues, P. J.; Salvatella, X.; Lindorff-Larsen, K. Bayesian-Maximum-Entropy Reweighting of IDP Ensembles Based on NMR Chemical Shifts. *Entropy* **2019**, *21* (9), 898.
(27) Tang, W. S.; Silva-Sánchez, D.; Giraldo-Barreto, J.; Carpenter, B.; Hanson, S. M.; Barnett, A. H.; Thiede, E. H.; Cossio, P. Ensemble Reweighting Using Cryo-EM Particle Images. *J. Phys. Chem. B* **2023**, *127* (24), 5410−5421.
(28) MacCallum, J. L.; Perez, A.; Dill, K. Determining Protein Structures by Combining Semireliable Data with Atomistic Physical Models by Bayesian Inference. *Proc. National Acad. Sci.* **2015**, *112* (22), 6985−6990.
(29) Raveh, B.; Sun, L.; White, K. L.; Sanyal, T.; Tempkin, J.; Zheng, D.; Bharath, K.; Singla, J.; Wang, C.; Zhao, J.; Li, A.; Graham, N. A.; Kesselman, C.; Stevens, R. C.; Sali, A. Bayesian Metamodeling of Complex Biological Systems across Varying Representations. *Proc. National Acad. Sci.* **2021**, *118* (35), No. e2104559118.
(30) Sali, A. From Integrative Structural Biology to Cell Biology. *J. Biol. Chem.* **2021**, *296*, No. 100743.
(31) Rout, M. P.; Sali, A. Principles for Integrative Structural Biology Studies. *Cell* **2019**, *177* (6), 1384−1403.
(32) Abraham, M.; Apostolov, R.; Barnoud, J.; Bauer, P.; Blau, C.; Bonvin, A. M. J. J.; Chavent, M.; Chodera, J.; Čondić-Jurkić, K.; Delemotte, L.; Grubmüller, H.; Howard, R. J.; Jordan, E. J.; Lindahl, E.; Ollila, O. H. S.; Selent, J.; Smith, D. G. A.; Stansfeld, P. J.; Tiemann, J. K. S.; Trellet, M.; Woods, C.; Zhmurov, A. Sharing Data from Molecular Simulations. *J. Chem. Inf. Model.* **2019**, *59* (10), 4093−4099.
(33) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, Ij. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; Santos, L. B.; da, S.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; Hoen, P. A. C. 't; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; Schaik, R. van; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; Lei, J. van der; Mulligen, E. van; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **2016**, *3* (1), No. 160018.
(34) Leitner, A.; Bonvin, A. M. J. J.; Borchers, C. H.; Chalkley, R. J.; Chamot-Rooke, J.; Combe, C. W.; Cox, J.; Dong, M.-Q.; Fischer, L.; Götze, M.; Gozzo, F. C.; Heck, A. J. R.; Hoopmann, M. R.; Huang, L.; Ishihama, Y.; Jones, A. R.; Kalisman, N.; Kohlbacher, O.; Mechtler, K.; Moritz, R. L.; Netz, E.; Novak, P.; Petrotchenko, E.; Sali, A.; Scheltema, R. A.; Schmidt, C.; Schriemer, D.; Sinz, A.; Sobott, F.; Stengel, F.; Thalassinos, K.; Urlaub, H.; Viner, R.; Vizcaíno, J. A.; Wilkins, M. R.; Rappsilber, J. Toward Increased Reliability, Transparency, and Accessibility in Cross-Linking Mass Spectrometry. *Structure* **2020**, *28* (11), 1259−1268.
(35) Schneidman-Duhovny, D.; Pellarin, R.; Sali, A. Uncertainty in Integrative Structural Modeling. *Curr. Opin Struc Biol.* **2014**, *28*, 96−104.
(36) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; Dijk, A. A. van; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science* **2021**, *373* (6557), 871−876.
(37) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583−589.
(38) Burley, S. K.; Kurisu, G.; Markley, J. L.; Nakamura, H.; Velankar, S.; Berman, H. M.; Sali, A.; Schwede, T.; Trewhella, J. PDB-Dev: A Prototype System for Depositing Integrative/Hybrid Structural Models. *Structure* **2017**, *25* (9), 1317−1318.
(39) Trewhella, J.; Hendrickson, W. A.; Kleywegt, G. J.; Sali, A.; Sato, M.; Schwede, T.; Svergun, D. I.; Tainer, J. A.; Westbrook, J.; Berman, H. M. Report of the WwPDB Small-Angle Scattering Task Force: Data Requirements for Biomolecular Modeling and the PDB. *Structure* **2013**, *21* (6), 875−881.
(40) Read, R. J.; Adams, P. D.; Arendall, W. B.; Brunger, A. T.; Emsley, P.; Joosten, R. P.; Kleywegt, G. J.; Krissinel, E. B.; Lütteke, T.; Otwinowski, Z.; Perrakis, A.; Richardson, J. S.; Sheffler, W. H.; Smith, J. L.; Tickle, I. J.; Vriend, G.; Zwart, P. H. A New Generation of Crystallographic Validation Tools for the Protein Data Bank. *Structure* **2011**, *19* (10), 1395−1412.
(41) Montelione, G. T.; Nilges, M.; Bax, A.; Güntert, P.; Herrmann, T.; Richardson, J. S.; Schwieters, C. D.; Vranken, W. F.; Vuister, G. W.; Wishart, D. S.; Berman, H. M.; Kleywegt, G. J.; Markley, J. L. Recommendations of the WwPDB NMR Validation Task Force. *Structure* **2013**, *21* (9), 1563−1570.
(42) Sali, A.; Berman, H. M.; Schwede, T.; Trewhella, J.; Kleywegt, G.; Burley, S. K.; Markley, J.; Nakamura, H.; Adams, P.; Bonvin, A. M. J. J.; Chiu, W.; Peraro, M. D.; Di Maio, F.; Ferrin, T. E.; Grünewald, K.; Gutmanas, A.; Henderson, R.; Hummer, G.; Iwasaki, K.; Johnson, G.; Lawson, C. L.; Meiler, J.; Marti-Renom, M. A.; Montelione, G. T.; Nilges, M.; Nussinov, R.; Patwardhan, A.; Rappsilber, J.; Read, R. J.; Saibil, H.; Schröder, G. F.; Schwieters, C. D.; Seidel, C. A. M.; Svergun, D.; Topf, M.; Ulrich, E. L.; Velankar, S.; Westbrook, J. D. Outcome of the First WwPDB Hybrid/Integrative Methods Task Force Workshop. *Structure* **2015**, *23* (7), 1156−1167.
(43) Henderson, R.; Sali, A.; Baker, M. L.; Carragher, B.; Devkota, B.; Downing, K. H.; Egelman, E. H.; Feng, Z.; Frank, J.; Grigorieff, N.; Jiang, W.; Ludtke, S. J.; Medalia, O.; Penczek, P. A.; Rosenthal, P. B.; Rossmann, M. G.; Schmid, M. F.; Schröder, G. F.; Steven, A. C.; Stokes, D. L.; Westbrook, J. D.; Wriggers, W.; Yang, H.; Young, J.; Berman, H. M.; Chiu, W.; Kleywegt, G. J.; Lawson, C. L. Outcome of the First Electron Microscopy Validation Task Force Meeting. *Structure* **2012**, *20* (2), 205−214.

https://doi.org/10.1021/acs.jpcb.3c04823
J. Phys. Chem. B XXXX, XXX, XXX−XXX

(44) van der Kamp, M. W.; Schaeffer, R. D.; Jonsson, A. L.; Scouras, A. D.; Simms, A. M.; Toofanny, R. D.; Benson, N. C.; Anderson, P. C.; Merkley, E. D.; Rysavy, S.; Bromley, D.; Beck, D. A.C.; Daggett, V. Dynameomics: A Comprehensive Database of Protein Dynamics. *Structure* **2010**, *18* (4), 423−435.
(45) Meyer, T.; D'Abramo, M.; Hospital, A.; Rueda, M.; Ferrer-Costa, C.; Perez, A.; Carrillo, O.; Camps, J.; Fenollosa, C.; Repchevsky, D.; Gelpí, J. L.; Orozco, M. MoDEL (Molecular Dynamics Extended Library): A Database of Atomistic Molecular Dynamics Trajectories. *Structure* **2010**, *18* (11), 1399−1409.
(46) Hospital, A.; Andrio, P.; Cugnasco, C.; Codo, L.; Becerra, Y.; Dans, P. D.; Battistini, F.; Torres, J.; Goñi, R.; Orozco, M.; Gelpí, J. Ll. BIGNASim: A NoSQL Database Structure and Analysis Portal for Nucleic Acids Simulation Data. *Nucleic Acids Res.* **2016**, *44* (D1), D272−D278.
(47) Rodríguez-Espigares, I.; Torrens-Fontanals, M.; Tiemann, J. K. S.; Aranda-García, D.; Ramírez-Anguita, J. M.; Stepniewski, T. M.; Worp, N.; Varela-Rial, A.; Morales-Pastor, A.; Medel-Lacruz, B.; Pándy-Szekeres, G.; Mayol, E.; Giorgino, T.; Carlsson, J.; Deupi, X.; Filipek, S.; Filizola, M.; Gómez-Tamayo, J. C.; Gonzalez, A.; Gutiérrez-de-Terán, H.; Jiménez-Rosés, M.; Jespers, W.; Kapla, J.; Khelashvili, G.; Kolb, P.; Latek, D.; Marti-Solano, M.; Matricon, P.; Matsoukas, M.-T.; Miszta, P.; Olivella, M.; Perez-Benito, L.; Provasi, D.; Ríos, S.; Torrecillas, I. R.; Sallander, J.; Sztyler, A.; Vasile, S.; Weinstein, H.; Zachariae, U.; Hildebrand, P. W.; Fabritiis, G. D.; Sanz, F.; Gloriam, D. E.; Cordomi, A.; Guixà-González, R.; Selent, J. GPCRmd Uncovers the Dynamics of the 3D-GPCRome. *Nat. Methods* **2020**, *17* (8), 777−787.
(48) Newport, T. D.; Sansom, M. S. P.; Stansfeld, P. J. The MemProtMD Database: A Resource for Membrane-Embedded Protein Structures and Their Lipid Interactions. *Nucleic Acids Res.* **2019**, *47*, No. D390.
(49) Bacle, A.; Buslaev, P.; Garcia-Fandino, R.; Favela-Rosales, F.; Ferreira, T. M.; Fuchs, P. F. J.; Gushchin, I.; Javanainen, M.; Kiirikki, A. M.; Madsen, J. J.; Melcr, J.; Rodríguez, P. M.; Miettinen, M. S.; Ollila, O. H. S.; Papadopoulos, C. G.; Peón, A.; Piggot, T. J.; Piñeiro, A.; Virtanen, S. I. Inverse Conformational Selection in Lipid−Protein Binding. *J. Am. Chem. Soc.* **2021**, *143* (34), 13701−13709.
(50) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235−242.
(51) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How Fast-Folding Proteins Fold. *Science* **2011**, *334* (6055), 517−520.
(52) Kryshtafovych, A.; Schwede, T.; Topf, M.; Fidelis, K.; Moult, J. Critical Assessment of Methods of Protein Structure Prediction (CASP)—Round XV. *Proteins: Struct., Funct., Bioinform.* **2023**, *91* (12), 1539−1549.
(53) Bowers, K. J.; Chow, E.; Xu, H.; Dror, R. O.; Eastwood, M. P.; Gregersen, B. A.; Klepeis, J. L.; Kolossvary, I.; Moraes, M. A.; Sacerdoti, F. D.; Salmon, J. K.; Shan, Y.; Shaw, D. E. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. *IEEE Explore* **2006**, 43.
(54) Volez, V. A.; Pande, V. S.; Bowman, G. R. Folding@home: Achievements from over 20 years of citizen science herald the exascale era. *Biophys. J.* **2023**, *122* (14), 2852−2863.
(55) Tejero, R.; Huang, Y. J.; Ramelot, T. A.; Montelione, G. T. AlphaFold Models of Small Proteins Rival the Accuracy of Solution NMR Structures. *Frontiers Mol. Biosci* **2022**, *9*, No. 877000.
(56) Terwilliger, T. C.; Poon, B. K.; Afonine, P. V.; Schlicksup, C. J.; Croll, T. I.; Millan, C.; Richardson, J. S.; Read, R. J.; Adams, P. D. Improved AlphaFold Modeling with Implicit Experimental Information. *Nat. Methods* **2022**, *19* (11), 1376−1382.
(57) Stahl, K.; Graziadei, A.; Dau, T.; Brock, O.; Rappsilber, J. Protein Structure Prediction with In-Cell Photo-Crosslinking Mass Spectrometry and Deep Learning. *Nat. Biotechnol.* **2023**, *41*, 1810.
(58) Drake, Z. C.; Seffernick, J. T.; Lindert, S. Protein Complex Prediction Using Rosetta, AlphaFold, and Mass Spectrometry Covalent Labeling. *Nat. Commun.* **2022**, *13* (1), 7846.
(59) Evans, R.; O'Neill, M.; Pritzel, A.; Antropova, N.; Senior, A.; Green, T.; Žídek, A.; Bates, R.; Blackwell, S.; Yim, J.; Ronneberger, O.; Bodenstein, S.; Zielinski, M.; Bridgland, A.; Potapenko, A.; Cowie, A.; Tunyasuvunakool, K.; Jain, R.; Clancy, E.; Kohli, P.; Jumper, J.; Hassabis, D. Protein Complex Prediction with AlphaFold-Multimer. *bioRxiv* **2022**, DOI: 10.1101/2021.10.04.463034.
(60) Tsaban, T.; Varga, J. K.; Avraham, O.; Ben-Aharon, Z.; Khramushin, A.; Schueler-Furman, O. Harnessing Protein Folding Neural Networks for Peptide−Protein Docking. *Nat. Commun.* **2022**, *13* (1), 176.
(61) Chang, L.; Perez, A. Ranking Peptide Binders by Affinity with AlphaFold**. *Angew. Chem., Int. Ed.* **2023**, *62* (7), No. e202213362.
(62) Braun, E.; Gilmer, J.; Mayes, H. B.; Mobley, D. L.; Monroe, J. I. Best Practices for Foundations in Molecular Simulations. *Living J Comput Mol Sci.* **2019**, *1* (11), 5957.
(63) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: A Web-based Graphical User Interface for CHARMM. *J. Comput. Chem.* **2008**, *29* (11), 1859−1865.
(64) Żaczek, S. MDMS: Software Facilitating Performing Molecular Dynamics Simulations. *J. Comput. Chem.* **2020**, *41* (3), 266−271.
(65) Parkman, J. A.; Barksdale, C. A.; Michaelis, D. J. CAN: A New Program to Streamline Preparation of Molecular Coordinate Files for Molecular Dynamics Simulations. *J. Comput. Chem.* **2021**, *42* (28), 2031−2035.
(66) Shirts, M. R.; Klein, C.; Swails, J. M.; Yin, J.; Gilson, M. K.; Mobley, D. L.; Case, D. A.; Zhong, E. D. Lessons Learned from Comparing Molecular Dynamics Engines on the SAMPL5 Dataset. *J. Comput.-Aided Mol. Des.* **2017**, *31* (1), 147−161.
(67) Thompson, M. W.; Gilmer, J. B.; Matsumoto, R. A.; Quach, C. D.; Shamaprasad, P.; Yang, A. H.; Iacovella, C. R.; McCabe, C.; Cummings, P. T. Towards Molecular Simulations That Are Transparent, Reproducible, Usable by Others, and Extensible (TRUE). *Mol. Phys.* **2020**, *118* (9−10), No. e1742938.
(68) Kanagalingam, G.; Schmitt, S.; Fleckenstein, F.; Stephan, S. Data Scheme and Data Format for Transferable Force Fields for Molecular Simulation. *Sci. Data* **2023**, *10* (1), 495.
(69) Hancock, M.; Peulen, T.-O.; Webb, B.; Poon, B.; Fraser, J. S.; Adams, P.; Sali, A. Integration of Software Tools for Integrative Modeling of Biomolecular Systems. *J. Struct Biol.* **2022**, *214* (1), No. 107841.
(70) Schindler, C. E. M.; Baumann, H.; Blum, A.; Böse, D.; Buchstaller, H.-P.; Burgdorf, L.; Cappel, D.; Chekler, E.; Czodrowski, P.; Dorsch, D.; Eguida, M. K. I.; Follows, B.; Fuchß, T.; Grädler, U.; Gunera, J.; Johnson, T.; Lebrun, C. J.; Karra, S.; Klein, M.; Knehans, T.; Koetzner, L.; Krier, M.; Leiendecker, M.; Leuthner, B.; Li, L.; Mochalkin, I.; Musil, D.; Neagu, C.; Rippmann, F.; Schiemann, K.; Schulz, R.; Steinbrecher, T.; Tanzer, E.-M.; Lopez, A. U.; Follis, A. V.; Wegener, A.; Kuhn, D. Large-Scale Assessment of Binding Free Energy Calculations in Active Drug Discovery Projects. *J. Chem. Inf Model* **2020**, *60* (11), 5457−5474.
(71) Abel, R.; Wang, L.; Mobley, D. L.; Friesner, R. A. A Critical Review of Validation, Blind Testing, and Real- World Use of Alchemical Protein-Ligand Binding Free Energy Calculations. *Curr. Top Med. Chem.* **2017**, *17* (23), 2577−2585.
(72) Orioli, S.; Larsen, A. H.; Bottaro, S.; Lindorff-Larsen, K. How to Learn from Inconsistencies: Integrating Molecular Simulations with Experimental Data. *Prog. Mol. Biol. Transl* **2020**, *170*, 123−176.
(73) Mondal, A.; Lenz, S.; MacCallum, J. L.; Perez, A. Hybrid Computational Methods Combining Experimental Information with Molecular Dynamics. *Curr. Opin. Struct. Biol.* **2023**, *81*, No. 102609.
(74) Rieping, W.; Habeck, M.; Nilges, M. Inferential Structure Determination. *Science* **2005**, *309* (5732), 303−306.
(75) Habeck, M.; Nilges, M.; Rieping, W. Bayesian Inference Applied to Macromolecular Structure Determination. *Phys. Rev. E* **2005**, *72* (3), 031912.
(76) Bonomi, M.; Vendruscolo, M. Determination of Protein Structural Ensembles Using Cryo-Electron Microscopy. *Curr. Opin Struc Biol.* **2019**, *56*, 37−45.

https://doi.org/10.1021/acs.jpcb.3c04823
J. Phys. Chem. B XXXX, XXX, XXX−XXX

(77) Bonomi, M.; Heller, G. T.; Camilloni, C.; Vendruscolo, M. Principles of Protein Structural Ensemble Determination. *Curr. Opin Struc Biol.* **2017**, *42*, 106−116.
(78) Saltzberg, D. J.; Viswanath, S.; Echeverria, I.; Chemmama, I. E.; Webb, B.; Sali, A. Using Integrative Modeling Platform to Compute, Validate, and Archive a Model of a Protein Complex Structure. *Protein Sci.* **2021**, *30* (1), 250−261.
(79) Sala, D.; Huang, Y. J.; Cole, C. A.; Snyder, D. A.; Liu, G.; Ishida, Y.; Swapna, G. V. T.; Brock, K. P.; Sander, C.; Fidelis, K.; Kryshtafovych, A.; Inouye, M.; Tejero, R.; Valafar, H.; Rosato, A.; Montelione, G. T. Protein Structure Prediction Assisted with Sparse NMR Data in CASP13. *Proteins Struct Funct Bioinform* **2019**, *87* (12), 1315−1332.
(80) Mondal, A.; Perez, A. Simultaneous Assignment and Structure Determination of Proteins From Sparsely Labeled NMR Datasets. *Frontiers Mol. Biosci* **2021**, *8*, No. 774394.
(81) Perez, A.; Morrone, J. A.; Brini, E.; MacCallum, J. L.; Dill, K. A. Blind Protein Structure Prediction Using Accelerated Free-Energy Simulations. *Sci. Adv.* **2016**, *2* (11), No. e1601274.
(82) Kuenze, G.; Bonneau, R.; Leman, J. K.; Meiler, J. Integrative Protein Modeling in RosettaNMR from Sparse Paramagnetic Restraints. *Structure* **2019**, *27* (11), 1721−1734.e5.
(83) Kuenze, G.; Meiler, J. Protein Structure Prediction Using Sparse NOE and RDC Restraints with Rosetta in CASP13. *Proteins Struct Funct Bioinform* **2019**, *87* (12), 1341−1350.
(84) Bonomi, M.; Hanot, S.; Greenberg, C. H.; Sali, A.; Nilges, M.; Vendruscolo, M.; Pellarin, R. Bayesian Weighing of Electron Cryo-Microscopy Data for Integrative Structural Modeling. *Structure* **2019**, *27* (1), 175−188.e6.
(85) Giraldo-Barreto, J.; Ortiz, S.; Thiede, E. H.; Palacio-Rodriguez, K.; Carpenter, B.; Barnett, A. H.; Cossio, P. A Bayesian Approach to Extracting Free-Energy Profiles from Cryo-Electron Microscopy Experiments. *Sci. Rep-uk* **2021**, *11* (1), No. 13657.
(86) Bayarri, G.; Andrio, P.; Hospital, A.; Orozco, M.; Gelpí, J. L. BioExcel Building Blocks Workflows (BioBB-Wfs), an Integrated Web-Based Platform for Biomolecular Simulations. *Nucleic Acids Res.* **2022**, *50* (W1), W99−W107.
(87) Bierer, B. E.; Crosas, M.; Pierce, H. H. Data Authorship as an Incentive to Data Sharing. *N. Engl. J. Med.* **2017**, *377* (4), 402.
(88) Pierce, H. H.; Dev, A.; Statham, E.; Bierer, B. E. Credit Data Generators for Data Reuse. *Nature* **2019**, *570* (7759), 30−32.
(89) Tiemann, J. K. S.; Szczuka, M.; Bouarroudj, L.; Oussaren, M.; Garcia, S.; Howard, R. J.; Delemotte, L.; Lindahl, E.; Baaden, M.; Lindorff-Larsen, K.; Chavent, M.; Poulain, P. MDverse: Shedding Light on the Dark Matter of Molecular Dynamics Simulations. *eLife* **2023**, *12*, RP90061.
(90) Antila, H. S.; Ferreira, T. M.; Ollila, O. H. S.; Miettinen, M. S. Using Open Data to Rapidly Benchmark Biomolecular Simulations: Phospholipid Conformational Dynamics. *J. Chem. Inf. Model.* **2021**, *61* (2), 938−949.
(91) Voelz, V. A.; Pande, V. S.; Bowman, G. R. Folding@home: Achievements from over Twenty Years of Citizen Science Herald the Exascale Era. *Biophys. J.* **2023**, *122*, 2852.
(92) Berman, H. M.; Adams, P. D.; Bonvin, A. A.; Burley, S. K.; Carragher, B.; Chiu, W.; DiMaio, F.; Ferrin, T. E.; Gabanyi, M. J.; Goddard, T. D.; Griffin, P. R.; Haas, J.; Hanke, C. A.; Hoch, J. C.; Hummer, G.; Kurisu, G.; Lawson, C. L.; Leitner, A.; Markley, J. L.; Meiler, J.; Montelione, G. T.; Phillips, G. N.; Prisner, T.; Rappsilber, J.; Schriemer, D. C.; Schwede, T.; Seidel, C. A. M.; Strutzenberg, T. S.; Svergun, D. I.; Tajkhorshid, E.; Trewhella, J.; Vallat, B.; Velankar, S.; Vuister, G. W.; Webb, B.; Westbrook, J. D.; White, K. L.; Sali, A. Federating Structural Models and Data: Outcomes from A Workshop on Archiving Integrative Structures. *Structure* **2019**, *27* (12), 1745−1759.
(93) Arts, M.; Satorras, V. G.; Huang, C.-W.; Zügner, D.; Federici, M.; Clementi, C.; Noé, F.; Pinsler, R.; Berg, R. van den Two for One: Diffusion Models and Force Fields for Coarse-Grained Molecular Dynamics. *J. Chem. Theory Comput.* **2023**, *19* (18), 6151−6159.
(94) Noé, F.; Fabritiis, G. D.; Clementi, C. Machine Learning for Protein Folding and Dynamics. *Curr. Opin Struc Biol.* **2020**, *60*, 77−84.
(95) Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; Žídek, A.; Green, T.; Tunyasuvunakool, K.; Petersen, S.; Jumper, J.; Clancy, E.; Green, R.; Vora, A.; Lutfi, M.; Figurnov, M.; Cowie, A.; Hobbs, N.; Kohli, P.; Kleywegt, G.; Birney, E.; Hassabis, D.; Velankar, S. AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models. *Nucleic Acids Res.* **2022**, *50* (D1), D439−D444.
(96) Simpkin, A. J.; Mesdaghi, S.; Rodríguez, F. S.; Elliott, L.; Murphy, D. L.; Kryshtafovych, A.; Keegan, R. M.; Rigden, D. J. Tertiary Structure Assessment at CASP15. *Proteins: Struct., Funct., Bioinform.* **2023**, *91* (12), 1616−1635.

https://doi.org/10.1021/acs.jpcb.3c04823
J. Phys. Chem. B XXXX, XXX, XXX−XXX