JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

Perspective

Sharing Data from Molecular Simulations

Mark James Abraham, Rossen Pavlov Apostolov, Jonathan Barnoud, Paul Bauer, Christian Blau, Alexandre M.J.J. Bonvin, Matthieu Chavent, John Damon Chodera, Karmen #ondi#-Jurki#, Lucie Delemotte, Helmut Grubmüller, Rebecca J. Howard, E. Joseph Jordan, Erik Lindal, O.H. Samuli Ollila, Jana Selent, Daniel G. A. Smith, Phill James Stansfeld, Johanna K. S. Tiemann, Mikael Trellet, Christopher J. Woods, and Artem Zhmurov

J. Chem. Inf. Model., **Just Accepted Manuscript** • DOI: 10.1021/acs.jcim.9b00665 • Publication Date (Web): 17 Sep 2019

Just Accepted

is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Sharing Data from Molecular Simulations

Mark Abraham[1], Rossen Apostolov[2], Jonathan Barnoud[3,†], Paul Bauer[1], Christian Blau[1], Alexandre M.J.J. Bonvin[4], Matthieu Chavent[5,], John Chodera[6], Karmen Čondić-Jurkić[6,7], Lucie Delemotte[1], Helmut Grubmüller[8], Rebecca J. Howard[9], E. Joseph Jordan[9], Erik Lindahl[9], O. H. Samuli Ollila[10], Jana Selent[11], Daniel G. A. Smith[12], Phillip J. Stansfeld[13], Johanna K.S. Tiemann[14], Mikael Trellet[4], Christopher Woods[15], Artem Zhmurov[1]*

AUTHOR ADDRESS

1- Science for Life Laboratory, Department of Applied Physics, KTH Royal Institute of Technology, Box 1031, SE-171 21 Solna

2- PDC Center for High Performance Computing, School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden

3- University of Groningen, Netherlands

4- Utrecht University, Faculty of Science, Bijvoet Center, Utrecht, the Netherlands

5- IPBS, Université Paul Sabatier, Toulouse, France

6- Computational and Systems Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, USA

7- Open Force Field Consortium

8- Max Planck Institute for Biophysical Chemistry, Goettingen, Germany

9- Science for Life Laboratory, Department of Biophysics and Biochemistry, Stockholm University, Box 1031, SE-171 21 Solna

10- Institute of Biotechnology, University of Helsinki, Finland

11- Research Programme on Biomedical Informatics, Hospital del Mar Medical Research Institute (IMIM) & Department of Experimental and Health Sciences, Pompeu Fabra University, Barcelona, Spain

12- The Molecular Sciences Software Institute, Blacksburg, USA

13- Department of Biochemistry, University of Oxford, Oxford, UK

14- Institute of Medical Physics and Biophysics, Faculty of Medicine, University Leipzig, Leipzig 04107, Germany

15- University of Bristol, Bristol, UK

†- current address: Intangible Realities Laboratory, University of Bristol, UK

KEYWORDS Data Sharing, Open Science, Reproducibility, File Standard, Molecular Simulation

ABSTRACT Given the need for modern researchers to produce open, reproducible scientific output, the lack of standards and best practices for sharing data and workflows used to produce and analyze molecular dynamics (MD) simulations have become an important issue in the field. There are now multiple well-established packages to perform molecular dynamics simulations, often highly tuned for exploiting specific classes of hardware, and each with strong communities surrounding them, but with very limited interoperability/transferability options. Thus, the choice of the software package often dictates the workflow for both simulation production and analysis. The level of detail in documenting the workflows and analysis code varies greatly in published work, hindering reproducibility of the reported results and the ability for other researchers to build on these studies. An increasing number of researchers are motivated to make their data available, but many challenges remain in order to effectively share and reuse simulation data. To discuss these and other issues related to best practices in the field in general, we organized a workshop in November 2018 (https://bioexcel.eu/events/workshop-on-sharing-data-from-molecular-simulations/). Here, we present a brief overview of this workshop and topics discussed. We hope this effort will spark further conversation in the MD community to pave the way towards more open, interoperable and reproducible outputs coming from research studies using MD simulations.

Introduction

Molecular simulations have become increasingly powerful and accessible in recent years, due in part to the rise of HPC[1-3] and GPU-powered clusters and powerful desktop computers[4] as well as the development of user-friendly software to set-up simulations[5,6]. The underlying physical models and methods have also improved over the years to address ever more complex biological and chemical questions[7,8]. Finally, the number of users and available tools is continuously increasing, as is the amount and complexity of workflows and produced outputs[9,10]. In this context, defining best practices related to documentation of protocols and code used to generate and/or analyze Molecular Dynamics (MD) simulations is becoming more important than ever[11]. A set of guidelines for reporting results obtained using molecular dynamics techniques and an opportunity to share data, similar to what structural biologists have achieved with the world-wide Protein Data Bank[12] (wwPDB), should generally help to improve the quality, reproducibility, statistics, and re-use of the published results.

Here, we would like to focus on the term reproducibility. The definition of reproducibility and its distinction from replicability can vary between disciplines[13-15], but in this context, we will broadly define reproducibility as the ability to re-implement the workflows of published work and obtain similar behavior for observables of interest as well as define the appropriate way to measure/calculate and report these observables[16]. Reproducibility is a long-standing issue for molecular modeling[17] and a key step toward better reproducibility and improved collaboration is making data more accessible and workflows interoperable. This can help reduce the entry barrier for the newcomers, but it could also help the existing practitioners to focus on answering scientific questions rather than wasting time in redeveloping existing sets of parameters or translating files formats to pass from one software to another. To reach this goal, it is now necessary to overcome several difficulties:

- First, there is now a multitude of package-specific file formats and object models. This variety, although increasing the efficiency for each package, introduces limitations in the

interoperability and creates friction for users juggling with various software to generate and analyze their data.

- Second, there is still a lack of exhaustive documentation related to new software development. The proliferation of various libraries and toolkits definitely opens up new avenues of research, but documenting the entire workflow from building a molecular model and parameterization to data analysis and visualization has become more complex. The method sections in publications often lack sufficient details to successfully re-implement the protocol or repeat the study from scratch, and default parameters to run a simulation may vary from one software version to another.
- Last but not least, there is no consensus to share data. The recent years have seen developments of different open data platforms, but the (ever-increasing) size of the generated trajectories makes it difficult to share simulation data efficiently. The absence of appropriate infrastructure, guidelines, and incentives further complicate the situation[18,19].

In general, we are witnessing a growing effort to make science more open by researchers themselves and increasingly so by funders and journals[20,21]. Soon, it may be mandatory to share data and deposit models obtained from hybrid/integrative approaches combining molecular modeling and experimental results[22]. Finding a way to consistently share data, workflows, and protocols will be thus necessary to ensure an efficient information exchange. Defining best practices and coming up with solutions should be a community effort to achieve the best outcome for everyone involved. In an effort to start a discussion around these questions, we organized a BioExcel workshop on *Sharing Data from Molecular Simulations* (SDMS) in Stockholm, November 2018. In this paper, we present a summary of discussions broadly focused on 4 topics:

- Standardization of file formats
- Streamlining molecular simulations data
- Tools for trajectory file sharing
- Reproducibility of molecular simulations

Each topic was introduced by 2 researchers and then openly discussed by all participants. All the presentations and the discussions were recorded and are accessible here: https://bioexcel.eu/sdms18-recordings/. The slides for the majority of the talks can be found here: https://doi.org/10.5281/zenodo.2652703 .

Standardization of file formats

While in structural biology the established PDB file format was stable for decades[12], the MD simulations field has a tendency to produce a multitude of input/output formats each related to one MD package[1,23-27]. With the rapid growth in complexity, size, and number of macromolecular structures led by advances in experimental techniques, even the canonical PDB format is now evolving to allow rendering and analyzing larger files with a gain in performance[28]. This evolution may also encourage the MD community to update its file formats to deal with larger and more heterogeneous data.

A new jointly developed format would need to be modular and flexible enough to not only take into account current but also catch future needs. Here arises a first question: What are the current and future needs of the MD community for such format? While particle coordinates are the current main feature both for input and output standards, other features need to be discussed such as physical/chemical descriptions of the model, experimental data used to create the model, technical details related to the simulation (such as algorithms used, sampling method, and forcefield). Different formats may be used as templates such as MMTF[28], MMCIF[29], JSON (http://www.json.org/), TNF[30]. At this workshop we all agreed that it

would be a great advance if this new standard can follow the FAIR principle[31]: Findable, Accessible, Interoperable, and Reproducible/Reusable. Many details remain to be discussed and the standardization question cannot be solved in one workshop with only a small sample of the MD community but need to be discussed by all main software developers joined with users to ensure usability. To do so another workshop will be held soon in New York to further discuss the question of file format and MD packages interoperability: https://molssi.org/2019/07/29/molssi-workshop-molecular-dynamics-software-interoperability/ .

For further details and discussions interested readers can watch associated videos from the 2018 workshop:

- Introduction of the topic by Mark Abraham (https://youtu.be/2S3qjBIE6Y4)
- Preliminary talk I by Erik Lindahl (https://youtu.be/Hvy8-gyTmj8)
- Preliminary talk II from Alexandre Bonvin (https://youtu.be/48Eb2MLHoYU)
- Breakout discussions presented by Phillip Stansfeld, Mikael Trellet, Daniel Smith and Johanna Tiemann (https://youtu.be/4fnV5EFXDpc)

Streamlining molecular simulations data

The MD simulation is often not a means and an end in itself but instead is run as part of a larger workflow. Such workflows involve joining together the output of many independent programs, such as those used for parameterizing molecules, those for performing molecular dynamics, and those for trajectory analysis. Managing the data movement between different programs in this workflow is challenging for several reasons:

1. The file formats used by different programs in the workflow may be incompatible, thereby preventing certain combinations of tools from being used together.

2. The features and forcefields supported by different programs in the workflow may be incompatible, thereby forcing researchers to choose algorithms and forcefields based on software compatibility rather than for good scientific reasons.
3. Different programs may implement features or forcefields in different ways, thereby meaning that the results of running the workflow will depend on the exact combination of programs (and possibly program versions) used. It is generally not possible to mix-and-match different programs and get the same results.

These challenges have forced researchers to develop workflows using specific software packages and specific forcefields. This creates divisions within the community and makes it difficult to write workflows that function equally well across a number of forcefields and a number of different software packages.

One of the solutions to this problem is the development of programs that convert/handle molecular information between the different file formats such as VMD[32], cpptraj[33], MDAnalysis[34,35], mdtraj[36], LOOS[37,38] and many others for trajectory analysis and TopoGromacs[39], CHARMM-GUI[40], CHAMBER[41], ParmEd (http://parmed.github.io/ParmEd/html/index.html#), InterMol[42] (https://github.com/shirtsgroup/InterMol), and others for topology generation and editing. The aim of these programs is to translate as much information as possible from one molecular file format into another. One recent example is BioSimSpace (https://biosimspace.org/), which provides wrappers that simplify the generation of the command files that are used to control the running of simulations. This allows researchers to write workflows that are independent of the choice of the underlying packages used to perform the simulation. BioSimSpace aims to run all stages of the workflow using the simulation software installed on the researcher's computer that is compatible with the forcefield chosen for the specific calculation.

While translators and program wrappers like ParmEd and BioSimSpace solve some of these problems, they are not a universal solution. They do not solve the issue that different simulation programs use different algorithms (or interpretations of algorithms, for example, different implementations of thermostats or integrators), or that different programs store and represent molecular information in different ways (e.g. SHAKE information for constraining bonds is represented in the molecular topology in GROMACS, while it is a simulation command parameter in NAMD and AMBER). This means MD properties/observables computed with one package will be systematically different by an often small but statistically significant amount from those computed with a different package as shown for free energy calculations[43]. Thus, the version and name of the MD program used to produce a simulation result will affect that result, and must be reported accordingly. Furthermore, MD simulations outputs are mainly trajectories which (1) represent ensemble averages (2) are chaotic in that small differences in initial conditions cause large differences in the subsequent dynamics ('butterfly effect'). This adds another layer of complexity and needs also a consensus on how to further analyze/process these trajectories to provide the final quantities of interest.

The recordings of this session can be found here:

- Introduction to the topic by John Chodera (https://youtu.be/6xOfN0y_uoQ)
- Preliminary talk I by Philip Stansfeld (https://youtu.be/YPYeujSD-6Y)
- Preliminary talk II by Christopher Woods (https://youtu.be/w1d1xtbGhHc)
- Breakout discussions presented by Christian Blau, Christopher Woods, Jonathan Barnoud and Mark Abraham (https://youtu.be/Z-JfBU3Emug)

Tools for trajectory file sharing

The benefits of sharing data together with the peer-reviewed publication, preprint or as a self-standing research output seem to be many - from receiving additional credit for one's work to

improving reproducibility, reusability or offering potentially new avenues of research[20,44]. Some disciplines, such as protein crystallography or genomics, have open data practices well integrated into their workflow, with metadata being collected throughout the workflow, and those practices are a *de facto* standard in scholarly communication. However, data sharing in the MD community still has not become widely adopted because best practice guidelines or journal recommendations on how to share MD simulations are yet to be established and adopted by the whole community. Making data sharing a standard practice in the field faces both technical and cultural challenges, although these are currently being tackled by some ongoing initiatives and solutions[20,45,46]. Thus, the development of best practices and guidelines for simulation data sharing will be of tremendous value, especially if created with the FAIR principles in mind[31]. To do so, we need to address several important questions regarding *what data* should be shared, *how* and *where*.

Answering to the *what data* question would need longer discussions not limited to a small group of individuals but involving the whole community and especially all the MD packages (another workshop will be held soon to help starting to answer to this question: https://molssi.org/2019/07/29/molssi-workshop-molecular-dynamics-software-interoperability/). The emergence of dedicated tools is now helping to answer to the *how* question. Software such as MDsrv[47], HTMoL[48], Mol* (https://molstar.org), Molmil[49] are now taking advantage of the WebGL API for sharing trajectories through interactive visualization on the web[50].

Other fields of research can help us to answer to the *where* question. Existing databanks, such as wwPDB[51] and Galaxy (https://usegalaxy.org), have been recognized by the scientific community. However, the establishment of an analogous, specialized platform for MD data, poses a great challenge, given the current lack of long-term support for the infrastructure projects of this kind. It is not clear yet who should be responsible for building such platform and how this infrastructure could be funded in a sustainable way, preferably without relying on short-term research grants, to cover the costs of development, maintenance

and data hosting. In the meantime, community-driven, special-purpose platforms like the GPCRmd (http://www.gpcrmd.org), Lipidbook[52] and NMRlipids[45] (http://nmrlipids.blogspot.com), Ligandbook[53], MoDEL[54] and BIGNASim[55] lead the way, providing specialized platforms for deposition and analysis of G protein-coupled receptors (GPCR), lipids, small molecules, proteins, and nucleic acids, respectively. General data sharing resources like Zenodo (https://zenodo.org), FigShare (https://figshare.com), Open Science Framework (https://osf.io) and others, also provide an opportunity for every researcher to deposit their simulation files and trajectories. Nevertheless, those resources may not provide sometimes enough space to sustainably store MD simulations outputs (with file size limits ranging between 5 GB and 50 GB).

To establish an efficient sharing culture, a systematic approach to developing tools and sharing guidelines is necessary, with the participation of the entire community in such activities and efforts. An open and inclusive discussion about best practices in data sharing, identification of short-term solutions based on the currently available frameworks and tools, as well as developing a strategy and requirements for future solutions bespoke to MD community and their needs is necessary. More details about the discussions taking place at the workshop can be found in the following videos:

- Introduction to the topic by Daniel Smith (https://youtu.be/mvesL9Y_9xU)
- Preliminary talk I by Johanna Tiemann (https://youtu.be/VOT6fEc7Iuc)
- Preliminary talk II by Jana Selent (https://youtu.be/TVS75j48mQ8)
- Breakout discussions presented by John Chodera, Karmen Čondić-Jurkić, Samuli Olllila and Lucie Delemotte (https://youtu.be/UIs1isntUPY)

Reproducibility of molecular simulations

MD simulations are chaotic and as such, the definition of reproducible results is non-trivial. First, the distinction between repeatability (by the same team and the same computational setup), replicability (by a different team and the same computational setup) and

reproducibility (by a different team, and with a different experimental setup) should be made [14]. Differences in outputs from these three perspectives may indicate different types of errors (bugs in software, human errors, or different choices along the workflow - choice of code, force field, system setup and more). The variability of parameters and dependence of the final results on both software and hardware makes it complicated (but also often unnecessary) to achieve the exact replication/repetition of any given setup, and untangling all the effects would be a difficult task. Focusing on a set of observables that can be calculated and preferably validated against experiments might be a better way of approaching reproducibility in this particular field. Similarly, focusing at observables which, despite the underlying chaoticity of the detailed dynamics, are reproducible without too large variation might be beneficial. Reaching an agreement on which observables we should aim to reproduce and how to properly calculate and report these values is thus desirable. For this, educational efforts are needed: best practice dissemination in terms of calculating statistical properties, for example, are crucial[16]. Coming up with standard benchmarks would also help, where the performance of different software/forcefield combinations for selected tasks could be compared.

In practice, data sharing would help with replicability and reproducibility. Practical challenges come from the size of data sets. However, one can envision sharing at least minimal data sets to improve

- methods reproducibility: provide sufficient details to replicate the study; this is in principle already done in publications, but authors, reviewers, and editors should pay special attention to the question, and sharing directly all input files should be mandatory,
- raw data reproducibility: share a minimum amount of data in the form of MD simulation snapshots, or even better whole trajectories, on existing data sharing repositories - Zenodo, Figshare, OSF, and

- results and inferential reproducibility: share among other analysis code, pipeline/workflow and example used.

Inspiration can be found in other research fields (e.g. Genomics[56] or Proteomics[57]) and existing dedicated initiatives, like MemProtMD[58] (http://memprotmd.bioch.ox.ac.uk), NMRlipids project (www.nmrlipids.blogspot.fi) and GPCRmd (http://www.gpcrmd.org), show that small groups of people focused on a narrow topic can create the necessary structure to share even large datasets in an efficient way. For further details and discussions interested readers can watch associated videos:

- Introduction to the topic by Karmen Čondić-Jurkić (https://youtu.be/lUTQgOXDEP8)
- Preliminary talk I by Helmut Grubmüller (https://youtu.be/cliVmGlrKag)
- Preliminary talk II by Samuli Ollila (https://youtu.be/46s33SonsiU)
- Breakout discussions presented by Mikael Trellet, Alexandre Bonvin, Mark Abraham and Christopher Woods (https://youtu.be/ex0_bqmJwE8)

This article summarizes the discussions started during the workshop held in Stockholm in November 2018. As may be noted by the reader, these discussions have not solved the issues about sharing data that our field is facing. Of course, this has never been the goal of such a small workshop. This workshop was intended to start asking relevant questions. Thus, this document (and the videos associated) can be seen as a road map for future developments. It is now crucial to build a community responsible for transforming these ideas into actions. This community needs to represent a diversity of perspectives by including both MD users and developers, newcomers and more seasoned practitioners, PhD students and postdocs, who are performing MD simulations on a daily basis, and PIs, who may hold the bigger picture views. As a community building effort, we are planning to regularly organize

more specific workshops aiming to address some of the issues raised in this article or to expand the scope of newly recognized problems. Of course, the structure of the workshops limits the number of participants, but care will be taken to ensure the aforementioned diversity of perspectives and roles in the field. In an effort to include as many users as possible in this discussion, the best practices guidelines that will emerge from these workshops will be submitted to the Living Journal of Computational Molecular Science (http://www.livecomsjournal.org/). This journal "*... provides a venue where authors can submit living documents that are updated on an ongoing basis as websites or Wikipedia articles could be, but which still have clear authorship and provide a mechanism for authors to get publication credit for their work*."[59] Hence, researchers interested to help us shape new practices to share data will be able to provide their feedback or directly contribute to the forthcoming document (as per the general idea laid out here: https://livecomsjournal.github.io/about/paper_code/). We hope that our work will act as a first step in a community-driven process of defining best practices for tool development and application in the molecular dynamics field.

AUTHOR INFORMATION

Corresponding Author

*Correspondence: matthieu.chavent@ipbs.fr , @Matth_Chavent

Author Contributions

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

Funding Sources

The workshop was supported by BioExcel Centre of Excellence (www.bioexcel.eu).

Acknowledgement

This work was supported by BioExcel Centre of Excellence (www.bioexcel.eu) funded by the European Union contracts H2020-INFRAEDI-02-2018-823830 and H2020-EINFRA-2015-1-675728. MC acknowledges support from CNRS-MITI grants PEPS MPI 2018 and "Modélisation du vivant" 2019. This work was supported by grants from the Gustafsson Foundation and Science for Life Laboratory to LD. HG has been supported by Max Planck Society and the German Research Foundation (DFG), Cluster of excellence Multiscale Imaging and the DFG priority programmes 1648, 755, and 803. OHSO acknowledges financial support from Academy of Finland (315596). DGAS thanks the National Science Foundation for support under Grant No. ACI-1547580. PJS would like to thank Wellcome [208361/Z/17/Z], the BBSRC [BB/P01948X/1, BB/R002517/1, BB/S003339/1 and BB/I019855/1] and MRC [MR/S009213/1] for funding. JKST acknowledges support from the Deutsche Forschungsgemeinschaft (DFG) HI1502/1-2. CW acknowledges support from the EPSRC (EP/N018591/1). We thank Oliver Beckstein and David Mobley for their careful reading and their comments.

Link to the SDMS18 recordings: https://bioexcel.eu/sdms18-recordings/
Discussions from Twitter: can be retrieved/extended by using the hashtag #SDMS18
Several participants from this workshop can be contacted/followed on Twitter: @the_mabraham, @jbarnoud, @amjjbonvin, @Matth_Chavent, @jchodera, @karmecon, @DelemotteLab, @CompBioPhys, @eriklindahl, @NMRlipids, @dga_smith, @pstansfeld, @j0kaso, @chryswoods

REFERENCES

(1) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations Through Multi-Level Parallelism From Laptops to Supercomputers. *SoftwareX* **2015**, *1-2*, 19–25.

(2) Lagardère, L.; Jolly, L.-H.; Lipparini, F.; Aviat, F.; Stamm, B.; Jing, Z. F.; Harger, M.; Torabifard, H.; Cisneros, G. A.; Schnieders, M. J.; Gresh, N.; Maday, Y.; Ren, P. Y.; Ponder, J. W.; Piquemal, J.-P. Tinker-HP: a Massively Parallel Molecular Dynamics Package for Multiscale Simulations of Large Complex Systems with Advanced Point Dipole Polarizable Force Fields. *Chem. Sci.* **2018**, *9*, 956–972.

(3) Jung, J.; Nishima, W.; Daniels, M.; Bascom, G.; Kobayashi, C.; Adedoyin, A.; Wall, M.; Lappala, A.; Phillips, D.; Fischer, W.; Tung, C. S.; Schlick, T.; Sugita, Y.; Sanbonmatsu, K. Y. Scaling Molecular Dynamics Beyond 100,000 Processor Cores for Large-Scale Biophysical Simulations. *J Comput Chem* **2019**.

(4) Stone, J. E.; Hardy, D. J.; Ufimtsev, I. S.; Schulten, K. GPU-Accelerated Molecular Modeling Coming of Age. *J. Mol. Graph. Model.* **2010**, *29*, 116–125.

(5) Doerr, S.; Harvey, M. J.; Noé, F.; De Fabritiis, G. HTMD: High-Throughput Molecular Dynamics for Molecular Discovery. *J. Chem. Theory Comput.* **2016**, *12*, 1845–1852.

(6) Jo, S.; Cheng, X.; Lee, J.; Kim, S.; Park, S. J.; Patel, D. S.; Beaven, A. H.; Lee, K. I.; Rui, H.; Park, S.; Lee, H. S.; Roux, B.; MacKerell, A. D.; Klauda, J. B.; Qi, Y.; Im, W. CHARMM-GUI 10 Years for Biomolecular Modeling and Simulation. *J Comput Chem* **2017**, *38*, 1114–1124.

(7) Marrink, S. J.; Corradi, V.; Souza, P. C. T.; Ingólfsson, H. I.; Tieleman, D. P.; Sansom, M. S. P. Computational Modeling of Realistic Cell Membranes. *Chem. Rev.* **2019**, acs.chemrev.8b00460.

(8) Bottaro, S.; Lindorff-Larsen, K. Biophysical Experiments and Biomolecular Simulations: a Perfect Match? *Science* **2018**, *361*, 355–360.

(9) Im, W.; Liang, J.; Olson, A.; Zhou, H.-X.; Vajda, S.; Vakser, I. A. Challenges in Structural Approaches to Cell Modeling. *J. Mol. Biol.* **2016**, *428*, 2943–2964.

(10) Chavent, M.; Duncan, A. L.; Sansom, M. S. Molecular Dynamics Simulations of Membrane Proteins and Their Interactions: From Nanoscale to Mesoscale. *Curr. Opin. Struct. Biol.* **2016**, *40*, 8–16.

(11) Elofsson, A.; Hess, B.; Lindahl, E.; Onufriev, A.; Van Der Spoel, D.; Wallqvist, A. Ten Simple Rules on How to Create Open Access and Reproducible Molecular Simulations of Biological Systems. *PLoS Comput Biol* **2019**, *15*, e1006649.

(12) Berman, H. M.; Kleywegt, G. J.; Nakamura, H.; Markley, J. L. The Protein Data Bank at 40: Reflecting on the Past to Prepare for the Future. *Structure (London, England : 1993)* **2012**, *20*, 391–396.

(13) Plesser, H. E. Reproducibility vs. Replicability: a Brief History of a Confused Terminology. *Front Neuroinform* **2017**, *11*, 76.

(14) Hinsen, K. ActivePapers: a Platform for Publishing and Archiving Computer-Aided Research. *F1000Res* **2014**, *3*, 289.

(15) Barba, L. A. Terminologies for Reproducible Research. *CoRR* **2018**.

(16) Grossfield, A.; Patrone, P. N.; Roe, D. R.; Schultz, A. J.; Siderius, D. W.; Zuckerman, D. M. Best Practices for Quantification of Uncertainty and Sampling Quality in Molecular Simulations [Article v1.0]. *Living Journal of Computational Molecular Science* **2018**, *1*.

(17) Walters, W. P. Modeling, Informatics, and the Quest for Reproducibility. *J Chem Inf Model* **2013**, *53*, 1529–1530.

(18) Graham, S. C.; Nagar, B.; Privé, G. G.; Deane, J. E. Molecular Models Should Not Be Published Without the Corresponding Atomic Coordinates. *Proc Natl Acad Sci USA* **2019**, *116*, 11099–11100.

(19) Romero, R.; Yuen, T.; New, M. I.; Zaidi, M.; Haider, S. Reply to Graham Et Al.: in Silico Atomistic Coordinates and Molecular Dynamics Simulation Trajectories of the

Glucocerebrosidase-Saposin C Complex. *Proc Natl Acad Sci USA* **2019**, *116*, 11101–11102.

(20) Data Sharing and the Future of Science. *Nat Commun* **2018**, *9*.

(21) Introducing eLife's First Computationally Reproducible Article. **2019**.

(22) 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 (London, England : 1993)* **2017**, *25*, 1317–1318.

(23) Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kalé, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J Comput Chem* **2005**, *26*, 1781–1802.

(24) Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: the Biomolecular Simulation Program. *J Comput Chem* **2009**, *30*, 1545–1614.

(25) Salomon Ferrer, R.; Case, D. A.; Walker, R. C. An Overview of the Amber Biomolecular Simulation Package. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2012**, *3*, 198–210.

(26) Rackers, J. A.; Wang, Z.; Lu, C.; Laury, M. L.; Lagardère, L.; Schnieders, M. J.; Piquemal, J.-P.; Ren, P.; Ponder, J. W. Tinker 8: Software Tools for Molecular Design. *J. Chem. Theory Comput.* **2018**, *14*, 5273–5289.

(27) Eastman, P.; Friedrichs, M. S.; Chodera, J. D.; Radmer, R. J.; Bruns, C. M.; Ku, J. P.; Beauchamp, K. A.; Lane, T. J.; Wang, L.-P.; Shukla, D.; Tye, T.; Houston, M.; Stich, T.; Klein, C.; Shirts, M. R.; Pande, V. S. OpenMM 4: a Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *J. Chem. Theory Comput.* **2013**, *9*, 461–469.

(28) Bradley, A. R.; Rose, A. S.; Pavelka, A.; Valasatava, Y.; Duarte, J. M.; Prlić, A.; Rose, P. W. MMTF—an Efficient File Format for the Transmission, Visualization, and Analysis of Macromolecular Structures. *PLoS Comput Biol* **2017**, *13*, e1005575.

(29) Bourne, P. E.; Berman, H. M.; McMahon, B.; Watenpaugh, K. D.; Westbrook, J. D.; Fitzgerald, P. Macromolecular Crystallographic Information File. *Meth. Enzymol.* **1997**, *277*, 571–590.

(30) Lundborg, M.; Apostolov, R.; Spangberg, D.; Gardenas, A.; Van Der Spoel, D.; Lindahl, E. An Efficient and Extensible Format, Library, and API for Binary Trajectory Data From Molecular Simulations. *J Comput Chem* **2014**, *35*, 260–269.

(31) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; 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.; 't Hoen, P. A. C.; 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.; van Schaik, R.; Sansone, S.-A.; Schultes, E.; Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; 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*, 160018.

(32) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J Mol Graph* **1996**, *14*, 33–38, 27–28.

(33) Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J. Chem. Theory Comput.* **2013**, *9*, 3084–3095.

(34) Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O. MDAnalysis: a Toolkit for the Analysis of Molecular Dynamics Simulations. *J Comput Chem* **2011**,

32, 2319–2327.
(35) Gowers, R.; Linke, M.; Barnoud, J.; Reddy, T.; Melo, M.; Seyler, S.; Domański, J.; Dotson, D.; Buchoux, S.; Kenney, I.; Beckstein, O. MDAnalysis: a Python Package for the Rapid Analysis of Molecular Dynamics Simulations; SciPy, 2016; pp 98–105.
(36) McGibbon, R. T.; Beauchamp, K. A.; Harrigan, M. P.; Klein, C.; Swails, J. M.; Hernández, C. X.; Schwantes, C. R.; Wang, L.-P.; Lane, T. J.; Pande, V. S. MDTraj: a Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109*, 1528–1532.
(37) Romo, T. D.; Grossfield, A. LOOS: an Extensible Platform for the Structural Analysis of Simulations. *Conf Proc IEEE Eng Med Biol Soc* **2009**, *2009*, 2332–2335.
(38) Romo, T. D.; Leioatts, N.; Grossfield, A. Lightweight Object Oriented Structure Analysis: Tools for Building Tools to Analyze Molecular Dynamics Simulations. *J Comput Chem* **2014**, *35*, 2305–2318.
(39) Vermaas, J. V.; Hardy, D. J.; Stone, J. E.; Tajkhorshid, E.; Kohlmeyer, A. TopoGromacs: Automated Topology Conversion From CHARMM to GROMACS Within VMD. *J Chem Inf Model* **2016**, *56*, 1112–1116.
(40) Lee, J.; Cheng, X.; Swails, J. M.; Yeom, M. S.; Eastman, P. K.; Lemkul, J. A.; Wei, S.; Buckner, J.; Jeong, J. C.; Qi, Y.; Jo, S.; Pande, V. S.; Case, D. A.; Brooks, C. L., III; MacKerell, A. D., Jr.; Klauda, J. B.; Im, W. CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *J. Chem. Theory Comput.* **2015**, *12*, 405–413.
(41) Crowley, M. F.; Williamson, M. J.; Walker, R. C. CHAMBER: Comprehensive Support for CHARMM Force Fields Within the AMBER Software. *International Journal of Quantum Chemistry* **2009**, *109*, 3767–3772.
(42) 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.* **2016**, *31*, 147–161.
(43) Loeffler, H. H.; Bosisio, S.; Duarte Ramos Matos, G.; Suh, D.; Roux, B.; Mobley, D. L.; Michel, J. Reproducibility of Free Energy Calculations Across Different Molecular Simulation Software Packages. *J. Chem. Theory Comput.* **2018**, *14* (11), 5567–5582.
(44) Woelfle, M.; Olliaro, P.; Todd, M. H. Open Science Is a Research Accelerator. *Nature Chem* **2011**, *3*, 745–748.
(45) Botan, A.; Favela-Rosales, F.; Fuchs, P. F. J.; Javanainen, M.; Kanduč, M.; Kulig, W.; Lamberg, A.; Loison, C.; Lyubartsev, A.; Miettinen, M. S.; Monticelli, L.; Määttä, J.; Ollila, O. H. S.; Retegan, M.; Róg, T.; Santuz, H.; Tynkkynen, J. Toward Atomistic Resolution Structure of Phosphatidylcholine Headgroup and Glycerol Backbone at Different Ambient Conditions. *J Phys Chem B* **2015**, *119*, 15075–15088.
(46) The PLUMED consortium. Promoting Transparency and Reproducibility in Enhanced Molecular Simulations. *Nat. Methods* **2019**, *16*, 670–673.
(47) Tiemann, J. K. S.; Guixà-González, R.; Hildebrand, P. W.; Rose, A. S. MDsrv: Viewing and Sharing Molecular Dynamics Simulations on the Web. *Nat. Methods* **2017**, *14*, 1123–1124.
(48) Carrillo-Tripp, M.; Alvarez-Rivera, L.; Lara-Ramírez, O. I.; Becerra-Toledo, F. J.; Vega-Ramírez, A.; Quijas-Valades, E.; González-Zavala, E.; González-Vázquez, J. C.; García-Vieyra, J.; Santoyo-Rivera, N. B.; Chapa-Vergara, S. V.; Meneses-Viveros, A. HTMoL: Full-Stack Solution for Remote Access, Visualization, and Analysis of Molecular Dynamics Trajectory Data. *J. Comput. Aided Mol. Des.* **2018**, *32*, 869–876.
(49) Bekker, G.-J.; Nakamura, H.; Kinjo, A. R. Molmil: a Molecular Viewer for the PDB and Beyond. *J Cheminform* **2016**, *8*, 42.
(50) Hildebrand, P. W.; Rose, A. S.; Tiemann, J. K. S. Bringing Molecular Dynamics Simulation Data Into View. *Trends Biochem. Sci.* **2019**.
(51) Berman, H.; Henrick, K.; Nakamura, H. Announcing the Worldwide Protein Data Bank. *Nat. Struct. Biol.* **2003**, *10*, 980.
(52) Domański, J.; Stansfeld, P. J.; Sansom, M. S. P.; Beckstein, O. Lipidbook: a Public Repository for Force-Field Parameters Used in Membrane Simulations. *J. Membr.*

Biol. **2010**, *236*, 255–258.

(53) Domański, J.; Beckstein, O.; Iorga, B. I. Ligandbook — an Online Repository for Small and Drug-Like Molecule Force Field Parameters. *Bioinformatics* **2017**, btx037.

(54) Meyer, T.; D'Abramo, M.; Hospital, A.; Rueda, M.; Ferrer-Costa, C.; Pérez, 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 (London, England : 1993)* **2010**, *18*, 1399–1409.

(55) Hospital, A.; Andrio, P.; Cugnasco, C.; Codo, L.; Becerra, Y.; Dans, P. D.; Battistini, F.; Torres, J.; Goñi, R.; Orozco, M.; Gelpí, J. L. BIGNASim: a NoSQL Database Structure and Analysis Portal for Nucleic Acids Simulation Data. *Nucleic Acids Res.* **2016**, *44*, D272–D278.

(56) Kaye, J.; Heeney, C.; Hawkins, N.; de Vries, J.; Boddington, P. Data Sharing in Genomics — Re-Shaping Scientific Practice. *Nature Reviews Genetics 2009 10:5* **2009**, *10*, 331–335.

(57) Martens, L.; Vizcaíno, J. A. A Golden Age for Working with Public Proteomics Data. *Trends Biochem. Sci.* **2017**, *42*, 333–341.

(58) Stansfeld, P. J.; Goose, J. E.; Caffrey, M.; Carpenter, E. P.; Parker, J. L.; Newstead, S.; Sansom, M. S. P. MemProtMD: Automated Insertion of Membrane Protein Structures Into Explicit Lipid Membranes. *Structure (London, England : 1993)* **2015**, *23*, 1350–1361.

(59) Mobley, D. L.; Shirts, M. R.; Zuckerman, D. M. Why We Need the Living Journal of Computational Molecular Science. *Living Journal of Computational Molecular Science* **2017**, *1*, 2031.