Structural Biology and Molecular Biophysics
Computational and Systems Biology

MDverse: Shedding Light on the Dark Matter of Molecular Dynamics Simulations

Reviewed Preprint
v2 • July 1, 2024
Revised by authors

Reviewed Preprint
v1 • September 20, 2023

Johanna K. S. Tiemann, Magdalena Szczuka, Lisa Bouarroudj, Mohamed Oussaren, Steven Garcia, Rebecca J. Howard, Lucie Delemotte, Erik Lindahl, Marc Baaden, Kresten Lindorff-Larsen, Matthieu Chavent, Pierre Poulain

Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, DK-2200 Copenhagen N, Denmark • Institut de Pharmacologie et Biologie Structurale, CNRS, Université de Toulouse, 205 route de Narbonne, 31400, Toulouse, France • Université Paris Cité, CNRS, Institut Jacques Monod, F-75013 Paris, France • Independent researcher, Amsterdam, Netherlands • Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, Stockholm, Sweden • Department of Applied Physics, Science for Life Laboratory, KTH Royal Institute of Technology, Stockholm, Sweden • Laboratoire de Biochimie Théorique, CNRS, Université Paris Cité, 13 rue Pierre et Marie Curie, F-75005 Paris, France

Abstract

The rise of open science and the absence of a global dedicated data repository for molecular dynamics (MD) simulations has led to the accumulation of MD files in generalist data repositories, constituting the *dark matter of MD* — data that is technically accessible, but neither indexed, curated, or easily searchable. Leveraging an original search strategy, we found and indexed about 250,000 files and 2,000 datasets from Zenodo, Figshare and Open Science Framework. With a focus on files produced by the Gromacs MD software, we illustrate the potential offered by the mining of publicly available MD data. We identified systems with specific molecular composition and were able to characterize essential parameters of MD simulation such as temperature and simulation length, and could identify model resolution, such as all-atom and coarse-grain. Based on this analysis, we inferred metadata to propose a search engine prototype to explore the MD data. To continue in this direction, we call on the community to pursue the effort of sharing MD data, and to report and standardize metadata to reuse this valuable matter.

eLife assessment

The study presents a **valuable** tool for searching molecular dynamics simulation data, making such datasets accessible for open science. The authors provide **convincing** evidence that it is possible to identify noteworthy molecular dynamics simulation datasets and that their analysis can produce information of value to the community.

https://doi.org/10.7554/eLife.90061.2.sa3

Introduction

The volume of data available in biology has increased tremendously (Marx, 2013; Stephens et al., 2015), through the emergence of high-throughput experimental technologies, often referred to as - omics, and the development of efficient computational techniques, associated with high-performance computing resources. The Open Access (OA) movement to make research results free and available to anyone (including e.g. the Budapest Open Access Initiative and the Berlin declaration on Open Access to Knowledge) has led to an explosive growth of research data made available by scientists (Wilson et al., 2021). The FAIR (Findable, Accessible, Interoperable and Reusable) principles (Wilkinson et al., 2016) have emerged to structure the sharing of these data with the goals of reusing research data and to contribute to the scientific reproducibility. This leads to a world where research data has become widely available and exploitable, and consequently new applications based on artificial intelligence (AI) emerged. One example is AlphaFold (Jumper et al., 2021), which enables the construction of a structural model of any protein from its sequence. However, it is important to be aware that the development of AlphaFold was only possible because of the existence of extremely well annotated and cleaned open databases of protein structures (wwPDB Berman et al. (2003)) and sequences (UniProt ***Consortium*** **(*2022)*)**. Similarly, accurate predictions of NMR chemical shifts and chemical-shift-driven structure determination was only made possible via a community-driven collection of NMR data in the Biological Magnetic Resonance Data Bank (Hoch et al., 2023). One can easily imagine novel possibilities of AI and deep learning reusing previous research data in other fields, if that data is curated and made available at a large scale (Fan and Shi, 2022; Mahmud et al., 2021).

Molecular Dynamics (MD) is an example of a well-established research field where simulations give valuable insights into dynamic processes, ranging from biological phenomena to material science (Perilla et al., 2015; Hollingsworth and Dror, 2018; Yoo et al., 2020; Alessandri et al., 2021; Krishna et al., 2021). By unraveling motions at details and timescales invisible to the eye, this well- established technique complements numerous experimental approaches (***Bottaro and Lindorff- Larsen, 2018***; Marklund and Benesch, 2019; Fawzi et al., 2021). Nowadays, large amounts of MD data could be generated when modelling large molecular systems (Gupta et al., 2022) or when applying biased sampling methods (Hénin et al., 2022). Most of these simulations are performed to decipher specific molecular phenomena, but typically they are only used for a single publication. We have to confess that many of us used to believe that it was not worth the storage to collect all simulations (in particular since all might not have the same quality), but in hindsight this was wrong. Storage is exceptionally cheap compared to the resources used to generate simulations data, and they represent a potential goldmine of information for researchers wanting to reanalyze them (Antila et al., 2021), in particular when modern machine-learning methods are typically limited by the amount of training data. In the era of open and data-driven science, it is critical to render the data generated by MD simulations not only technically available but also practically usable by the scientific community. In this endeavor, discussions started a few years ago (Abraham et al., 2019; Abriata et al., 2020; Merz et al., 2020) and the MD data sharing trend has been accelerated with the effort of the MD community to release simulation results related to the COVID-19 pandemic (Amaro and Mulholland, 2020; Mulholland and Amaro, 2020) in a centralized database (*https://covid.bioexcel.eu*). Specific databases have also been developed to store sets of simulations related to protein structures (MoDEL: Meyer et al. (2010)), membrane proteins in general (Mem-ProtMD: Stansfeld et al. (2015); Newport et al. (2018)), G-protein coupled receptors in particular (GPCRmd: Rodríguez-Espigares et al. (2020)), or lipids (Lipidbook:Domański et al. (2010), NMRLipids Databank: Kiirikki et al. (2023)).

Albeit previous attempts in the past (Tai et al., 2004; Meyer et al., 2010), there is, as of now, no central data repository that could host all kinds of MD simulation files. This is not only due to the huge volume of data and its heterogeneity, but also because interoperability of the many file formats used adds to the complexity. Thus, faced with the deluge of biosimulation data (Hospital et al., 2020), researchers often share their simulation files in multiple generalist data repositories. This makes it difficult to search and find available data on, for example, a specific protein or a given set of parameters. We are qualifying this amount of scattered data as *the dark matter of MD*, and we believe it is essential to shed light onto this overlooked but high-potential volume of data. When unlocked, publicly available MD files will gain more visibility. This will help people to access and reuse these data more easily and overall, by making MD simulation data more FAIR (Wilkinson et al., 2016), it will also improve the reproducibility of MD simulations (Elofsson et al., 2019; Porubsky et al., 2020; consortium, 2019).

In this work, we have employed a search strategy to index scattered MD simulation files deposited in generalist data repositories. With a focus on the files generated by the Gromacs MD software, we performed a proof-of-concept large-scale analysis of publicly available MD data. We revealed the high value of these data and highlighted the different categories of the simulated molecules, as well as the biophysical conditions applied to these systems. Based on these results and our annotations, we proposed a search engine prototype to easily explore this *dark matter of MD*. Finally, building on this experience, we provide simple guidelines for data sharing to gradually improve the FAIRness of MD data.

Results

With the rise of open science, researchers increasingly share their data and deposit them into generalist data repositories, such as Zenodo (*https://zenodo.org*), Figshare (*https://figshare.com*), Open Science Framework (OSF, *https://osf.io*), and Dryad (*https://datadryad.org/*). In this first attempt to find out how many files related to MD are deposited in data repositories, we focused our exploration on three major data repositories: Figshare (~3.3 million files, ~112 TB of data, as of January 2023), OSF (~2 million files, as of November 2022; Figures provided by Figshare and OSF user support teams), and Zenodo (~9.9 million files, ~1.3 PB of data, as of December 2022; ***Panero and Benito*** (***2022)***).

One immediate strategy to index MD simulation files available in data repositories is to perform a text-based Google-like search. For that, one queries these repositories with keywords such as 'molecular dynamics' or 'Gromacs'. Unfortunately, we experienced many false positives with this search strategy. This could be explained by the strong discrepancy we observed in the quantity and quality of metadata (title, description) accompanying datasets and queried in text-based search. For instance, a description text could be composed of a couple of words to more than 1,200 words. Metadata is provided by the user depositing the data, with no incentive to issue relevant details to support the understanding of the simulation. For the three data repositories studied, no human curation other by that of the providers is performed when submitting data. It is also worth mentioning that title and description are provided as free-text and do not abide to any controlled vocabulary such as a specific MD ontology.

To circumvent this issue, we developed an original and specific search strategy that we called *Explore and Expand* (*Ex*^{2}) (see **Fig. 1-A** and Materials and Methods section) and that relies on a combination of file types and keywords queries. In the *Explore* phase, we searched for files based on their file types (for instance: .xtc, .gro, etc) with MD-related keywords (for instance: 'molecular dynamics', 'Gromacs', 'Martini', etc). Each of these hit files belonged to a dataset, which we further screened in the *Expand* phase. There, we indexed all files found in a dataset identified in the previous *Explore* phase with, this time, no restriction to the collected file types (see **Fig. 1-A** and details on the data scraping procedure in the Materials and Methods section).

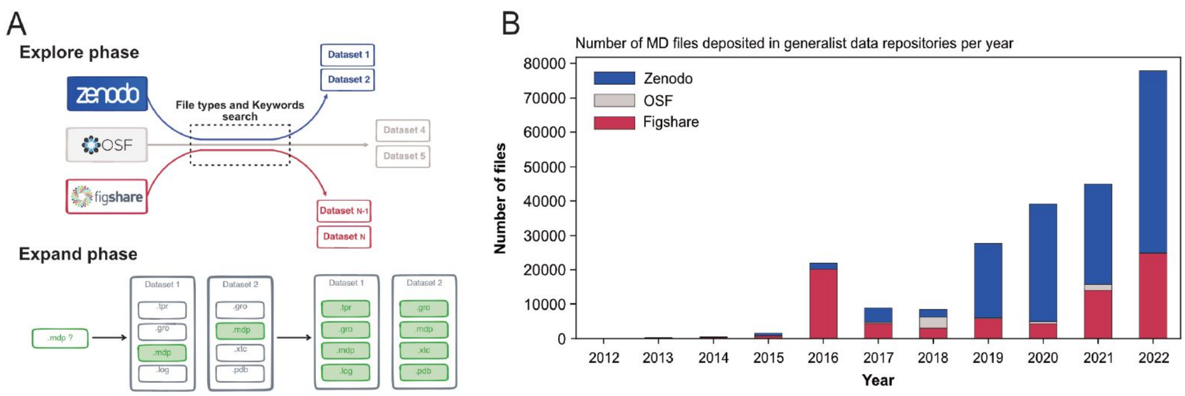


Figure 1.

(A) Explore and Expand (Ex^2) strategy used to index and collect MD-related files. Within the explore phase, we search in the respective data repositories for datasets that contain specific keywords (e.g. "molecular dynamics", "md simulation", "namd", "martini"...) in conjunction with specific file extensions (e.g. "mdp", "psf", "parm7"...), depending on their uniqueness and level of trust to not report false-positives (.i.e not MD related). In the expand phase, the content of the identified datasets is fully cataloged, including files that individually could result in false positives (such as e.g. ".log" files). (B) Number of deposited files in generalist data repositories, identified by our Ex^2 strategy.

Globally, we indexed about 250,000 files and 2,000 datasets that represented 14 TB of data deposited between August 2012 and March 2023 (see **Table 1**). One major difficulty were the numerous files stored in zipped archives, about seven times more than files steadily available in datasets (see **Table 1**). While this choice is very convenient for depositing the files (as one just needs to provide one big zip file to upload to the data repository server), it hinders the analysis of MD files as data repositories only provide a limited preview of the content of the zip archives and completely inhibits, for example, data streaming for remote analysis and visualization. Files within zip files are not indexed and cannot be searched individually. The use of zip archives also hampers the reusability of MD data, since a specific file cannot be downloaded individually. One has to download the entire zip archive (sometimes with a size up to several gigabytes) to extract the one file of interest.

The first dataset we found related to MD data that has been deposited in August 2012 in Figshare and corresponds to the work of Fuller et al. (Fuller et al., 2012) (see **Table 1**) but we may consider the start of more substantial deposition of the MD data to be 2016 with more than 20,000 files deposited, mainly in Figshare (see **Fig. 1-B**). While the number of files deposited in Zenodo was first relatively limited, the last few years (2020-2022) saw a steep increase, passing from a few thousands files in 2018 to almost 50,000 files in 2022 (see **Fig. 1-B**). In 2018, the number of MD files deposited in OSF was similar to those in the two other data repositories, but did not take off as much as the other data repositories. Zenodo seems to be favored by the MD community since 2019, even though Figshare in 2022 also saw a sharp increase in deposited MD files. The preference for Zenodo could also be explained by the fact that it is a publicly funded repository developed under the European OpenAIRE program and operated by CERN (***European Organization For Nuclear Research and OpenAIRE, 2013***). Overall, the trend showed a rise of deposited data with a steep increase in 2022 (**Fig. 1-B**). We believe that this trend will continue in future years, which will lead to a greater amount of MD data available. It is thus urgent to deploy a strategy to index this vast amount of data, and to allow the MD community to easily explore and reuse such gigantic resource. The following describes what is already feasible in terms of meta analysis, in particular what types of data are deposited in data repositories and the simulation setup parameters used by MD experts that have deposited their data.

With our Ex^2 strategy (see **Fig. 1-A**), we assigned the deposited files to the MD packages: AMBER (Ferrer et al., 2012), DESMOND (Bowers et al., 2006), Gromacs (Berendsen et al., 1995; Abraham et al., 2015), and NAMD/CHARMM (Phillips et al., 2020; Brooks et al., 2009), based on their corresponding file types (see Materials and Methods section). In the case of NAMD/CHARMM, file extensions were mostly identical, which prevented us from distinguishing the respective files from these two MD programs. With 87,204 files deposited, the Gromacs program was most represented (see **Fig. 2-A**), followed by NAMD/CHARMM, AMBER, and DESMOND. This statistic is limited as it does not consider more specific databases related to a particular MD program. For example, the DE Shaw Research website contains a large amount of simulation data related to SARS-CoV-2 that has been generated using the ANTON supercomputer (*https://www.deshawresearch.com/downloads/download_trajectory_sarscov2.cgi/*) or other extensively simulated systems of interest to the community. However, this in itself might also serve as a good example, since few automated search strategies will be able to find custom stand-alone web servers as valuable repositories. Here, our goal was not to compare the availability of all data related to each MD program but to give a snapshot of the type of data available at a given time (*i.e*. March 2023) in generalist data repositories. Interestingly, many files (> 133,000) were not directly associated to any MD program (see **Fig. 2-A** label 'Unknown'). We categorized these files based on their extensions (see **Fig. 2-B**). While 10 % of these files were without file extension (**Fig. 2-B**, column *none*), we found numerous files corresponding to structure coordinates such as .pdb (~12,000) and .xyz (~6,800) files. We also got images (.tiff files) and graphics (.xvg files). Finally, we found many text files such as .txt, .dat, and .out which can potentially hold details about how simulations were performed. Focusing further on files related to the Gromacs program,

Data repository	datasets	first dataset	latest dataset	files	total size (GB)	zip files	files within zip	total files
Zenodo	1,011	19/11/2014	05/03/2023	20,250	12,851	1,780	141,304	161,554
Figshare	913	20/08/2012	03/03/2023	3,336	736	590	74,720	78,056
OSF	55	24/05/2017	05/02/2023	6,146	495	14	0	6,146
Total	*1,979*	–	–	*29,732*	*14,082*	*2,384*	*216,024*	*245,756*

Table 1.

Statistics of the MD-related datasets and files found in the data repositories Figshare, OSF, and Zenodo.

being currently most represented in the studied data repositories, we demonstrated in the following present possibilities to retrieve numerous information related to deposited MD simulations.

First, we were interested in what file types researchers deposited and thereby find potentially of great value to share. We therefore quantified the types of files generated by Gromacs (**Fig. 3-A**). The most represented file type is the .xtc file (28,559 files, representing 8.6 TB). This compressed (binary) file is used to store the trajectory of an MD simulation and is an important source of information to characterize the evolution of the simulated molecular system as a function of time. It is thus logical to mainly find this type of file shared in data repositories, as it is of great value for reusage and new analyses. Nevertheless, it is not directly readable but needs to be read by a third-party program, such as Gromacs itself, a molecular viewer like VMD (Humphrey et al., 1996) or an analysis library such as MDAnalysis (Michaud-Agrawal et al., 2011; Gowers et al., 2016). In addition, this trajectory file can only be of use in combination with a matching coordinates file, in order to correctly access the dynamics information stored in this file. Thus, as it is, this file is not easily mineable to extract useful information, especially if multiple .xtc and coordinate files are available in one dataset. Interestingly, we found 1,406 .trr files, which contain trajectory but also additional information such as velocities, energy of the system, etc. While this file is especially useful in terms of reusability, the large size (can go up to several 100 GB) limits its deposition in most data repositories. For instance, a file cannot usually exceed 50 GB in Zenodo, 20 GB in Figshare (for free accounts) and 5 GB in OSF. Altogether, Gromacs trajectory files represented about 30,000 files in the three explored generalist repositories (34% of Gromacs files). This is a large number in comparison to existing trajectories stored in known databases dedicated to MD with 1,700 MD trajectories available in MoDEL, 1,737 trajectories (as of November 2022) available in GPCRmd, 5,971 (as of January 2022) trajectories available in MemProtMD and 726 trajectories (as of March 2023) available in the NMRLipids Databank. Although fewer in count, these numbers correspond to manually or semi-automatically curated trajectories of specific systems, mostly proteins and lipids. Thus, ~30,000 MD trajectories available in generalist data repositories may represent a wider spectrum of simulated systems but need to be further analyzed and filtered to separate usable data from less interesting trajectories such as minimization or equilibration runs.

Given the large volume of data represented by .xtc files (see above), we could only scratch the surface of the information stored in these trajectory files by analyzing a subset of 779 .xtc files - one per dataset in which this type of file was found. We were able to get the size of the molecular systems and the number of frames available in these files (**Fig. 3-B**). The system size was up to more than one million atoms for a simulation of the TonB protein (Virtanen et al., 2020). The cumulative distribution of the number of frames showed that half of the files contain more than 10,000 frames. This conformational sampling can be very useful for other research fields besides the MD community that study, for instance, protein flexibility or protein engineering where diverse backbones can be of value. We found an .xtc file containing more than 5 million frames, where the authors probe the picosecond–nanosecond dynamics of T4 lysozyme and guide the MD simulation with NMR relaxation data (Kümmerer et al., 2021). Extending this analysis to all 28,559 .xtc files detected would be of great interest for a more holistic view, but this would require an initial step of careful checking and cleaning to be sure that these files are analyzable. Of note, as .xtc files also contain time stamps, it would be interesting to study the relationship between the time and the number of frames to get useful information about the sampling. Nevertheless, this analysis would be possible only for unbiased MD simulations. So, we would need to decipher if the .xtc file is coming from biased or unbiased simulations, which may not be trivial.

These results bring a first explanation on why there is not a single special-purpose repository for MD trajectory files. Databases dedicated to molecular structures such as the Protein Databank (Berman et al., 2000; Kinjo et al., 2017; Armstrong et al., 2019), or even the recent PDB-dev (Burley et al., 2017), designed for integrative models, cannot accept such large-size files, even less if complete trajectories without reducing the number of frames would be uploaded. This

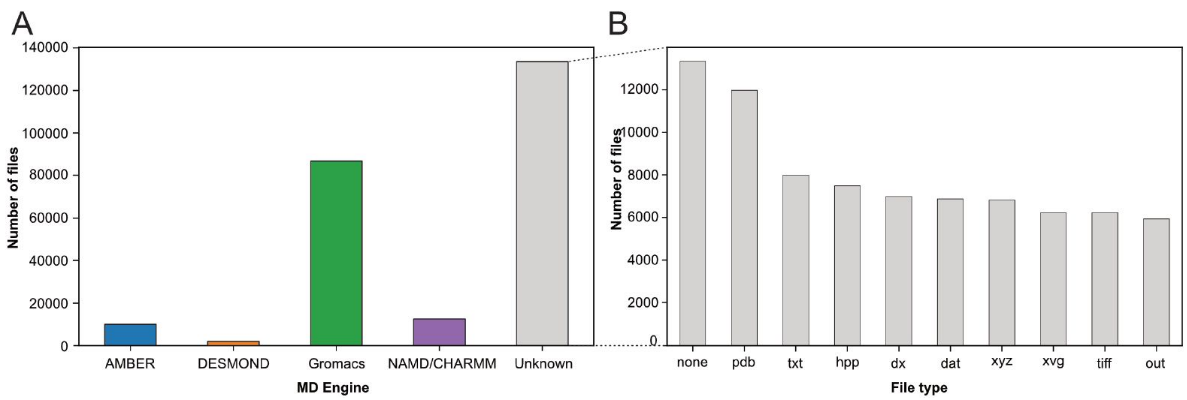


Figure 2.

Categorization of index files based on their file types and assigned MD engine. (A) Distribution of files among MD simulation engines (B) Expansion of (A) MD Engine category "Unknown" into the 10 most observed file types.

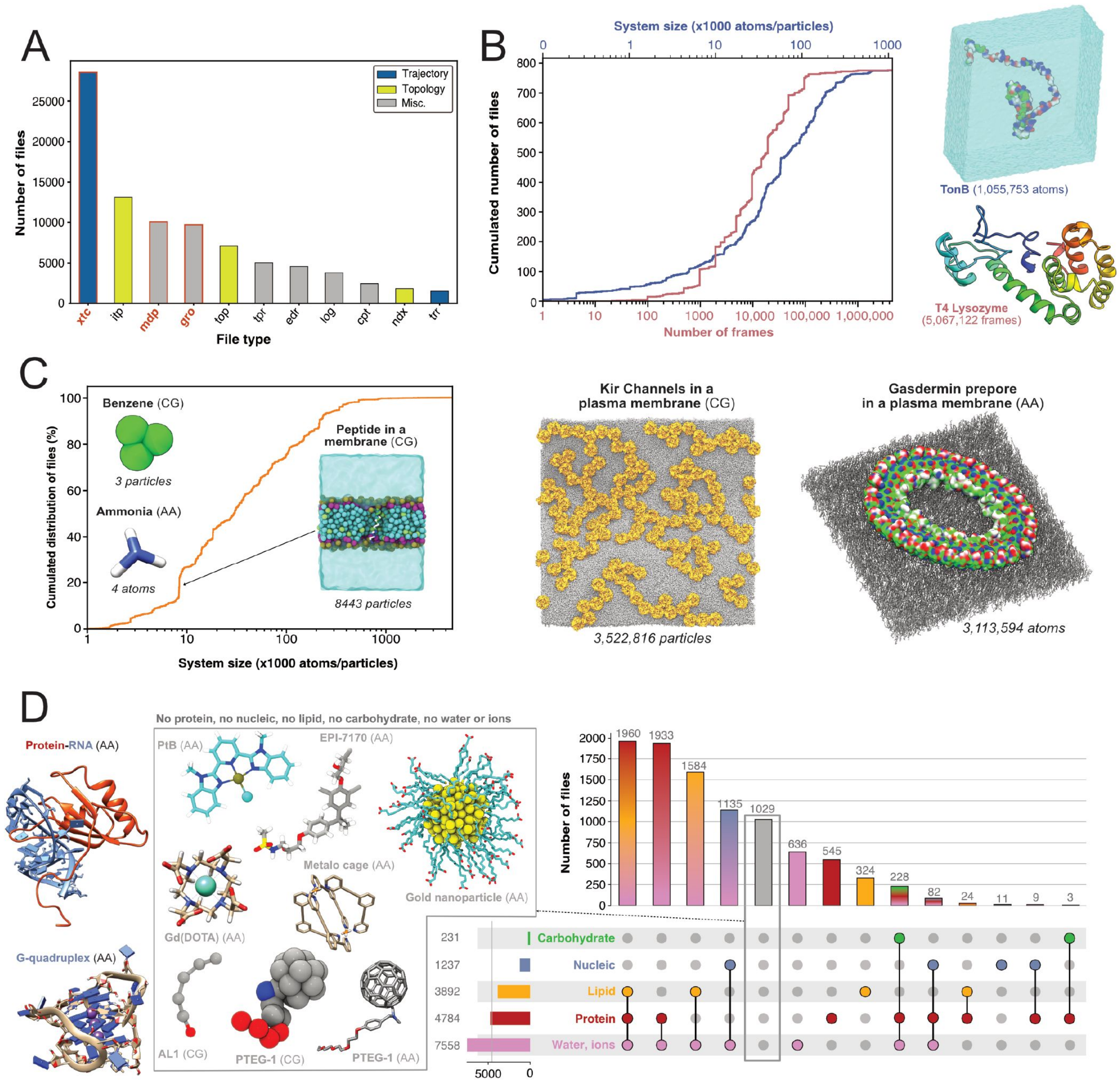


Figure 3.

Content analysis of .xtc and .gro files. (A) Number of Gromacs-related files available in searched data repositories. In red, files used for further analyses. (B) Simple analyze of a subset of .xtc files with the cumulative distribution of the number of frames (in green) and the system size (in orange). (C) Cumulative distribution of the system sizes extracted from .gro files. (D) Upset plot of systems grouped by molecular composition, inferred from the analysis of .gro files. For this figure, 3D structures of representative systems were displayed, including soluble proteins such as TonB and T4 Lysozyme, membrane proteins such as Kir Channels and the Gasdermin prepore, Protein-/RNA and G-quadruplex and other non-protein molecules.

would also require implementing extra steps of data curation and quality control. In addition, the size of the IT infrastructure and the human skills required for data curation represents a significant cost that could probably not be supported by a single institution.

Subsequently, our interest shifted towards exploring which systems are being investigated by MD researchers who deposit their files. We found 9,718 .gro files which are text files that contain the number of particles and the Cartesian coordinates of the system modelled. By parsing the number of particles and the type of residue, we were able to give an overview of all Gromacs systems deposited (**Fig. 3-C,D**). In terms of system size, they ranged from very small - starting with two coarse-grain (CG) particles of graphite (Piskorz et al., 2019), followed by coordinates of a water molecule (3 atoms) (Ivanov et al., 2017), CG model of benzene (3 particles) (Dandekar and Mondal, 2020) and atomistic model of ammonia (4 atoms) (Kelly and Smith, 2020) — to go up to atomistic and coarse-grain systems composed of more than 3 million particles (Duncan et al., 2020; Schaefer and Hummer, 2022) (**Fig. 3-C**). Interestingly, the system sizes in .gro files exceeded those of the analyzed .xtc files (**Fig. 3-B**). Even if we cannot exclude that the limited number of .xtc files analyzed (779 .xtc files selected from 28,559 .xtc files indexed) could explain this discrepancy, an alternate hypothesis is that the size of an .xtc file also depends on the number of frames stored. To reduce the size of .xtc files deposited in data repositories, besides removing some frames, researchers might also remove parts of the system, such as water molecules. As a consequence for reusability, this solvent removal could limit the number of suitable datasets available for researchers interested in re-analysing the simulation with respect to, in this case, water diffusion. While the size of systems extracted from .gro files was homogeneously spread, we observed a clear bump around system sizes of circa 8,500 atoms/particles. This enrichment of data could be explained by the deposition of ∼340 .gro files related to the simulation of a peptide translocation through a membrane (**Fig. 3-C**) (Kabelka et al., 2021). Beyond 1 million particles/atoms, the number of systems is, for the moment, very limited.

We then analyzed residues in .gro files and inferred different types of molecular systems (see **Fig. 3-D**). Two of the most represented systems contained lipid molecules. This may be related to NMRLipids initiative (*http://nmrlipids.blogspot.com*). For several years, this consortium has been ac-tively working on lipid modelling with a strong policy of data sharing and has contributed to share numerous datasets of membrane systems. As illustrated in **Fig 3-C**, a variety of membrane systems, especially membrane proteins, were deposited. This highlights the vitality of this research field, and the will of this community to share their data. We also found numerous systems containing solvated proteins. This type of data, combined with .xtc trajectory files (see above), could be invaluable to describe protein dynamics and potentially train new artificial intelligence models to go beyond the current representation of the static protein structure (Lane, 2023). There was also a good proportion of systems containing nucleic acids alone or in interaction with proteins (1237 systems). At this time, we found only few systems containing carbohydrates that also contained proteins and corresponded to one study to model hyaluronan–CD44 interactions (Vuorio et al., 2017). Maybe a reason for this limited number is that systems containing sugars are often modelled using AMBER force field (Ferrer et al., 2012), in combination with GLYCAM (***Kirschner et al., 2008-03***). A future study on the ∼10,200 AMBER files deposited could retrieve more data related to carbohydrate containing systems. Given the current developments to model glycans (Fadda, 2022), we expect to see more deposited systems with carbohydrates in the coming years.

Finally, we found 1,029 .gro files which did not belong to the categories previously described. These files were mostly related to models of small molecules, or molecules used in organic chemistry (Young et al., 2020) and material science (Zheng et al., 2022; Piskorz et al., 2019) (see central panel, **Fig. 3-D**). Several datasets contained lists of small molecules used for calculating free energy of binding (Aldeghi et al., 2015), solubility of molecules (Liu et al., 2016), or osmotic coefficient (Zhu, 2019). Then, we identified models of nanoparticles (Kyrychenko et al., 2012; Pohjolainen et al., 2016), polymers (Sarkar et al., 2020; Karunasena et al., 2021; Gertsen et al., 2020), and drug molecules like EPI-7170, which binds disordered regions of proteins (Zhu et al., 2022). Finally, an interesting case from material sciences was the modelling of the PTEG-1 molecule, an addition of polar triethylene glycol (TEG) onto a fulleropyrrolidine molecule (see central panel, **Fig. 3-D**). This molecule was synthesized to improve semiconductors (Jahani et al., 2014). We found several models related to this peculiar molecule and its derivatives, both atomistic (Qiu et al., 2017; Sami et al., 2022) and coarse grained (Alessandri et al., 2020). With a good indexing of data and appropriate metadata to identify modelled molecules, a simple search, which was previously to this study missing, could easily retrieve different models of the same molecule to compare them or to run multi-scale dynamics simulations. Beyond .gro files, we would like to analyze the ensemble of the ∼ 12,000 .pdb extracted in this study (see **Fig. 2-B**) to better characterize the types of molecular structures deposited.

Another important category of deposited files are those containing information about the topology of the simulated molecules, including file extensions such as .itp and .top. Further, they are often the results of long parametrization processes (Wang et al., 2004; Vanommeslaeghe and MacKerell, 2012; Souza et al., 2021) and therefore of significant value for reusability . Based on our analysis, we indexed almost 20,000 topology files which could spare countless efforts to the MD community if these files could be easily found, annotated and reused. Interestingly, the number of .itp files was elevated (13,058 files) with a total size of 2 GB, while there were less .top files (7,009 files) with a total size of 17 GB. Thus, .itp files seemed to contain much less information than the .top files. Among the remaining file types, .tpr files contain all the information to potentially directly run a simulation. Here, we found 4,987 .tpr files, meaning that it could virtually be possible to rerun almost 5,000 simulations without the burden of setting up the system to simulate. Finally, the 3,730 .log files are also a source of useful information as it is relatively easy to parse this text file to extract details on how MD simulations were run, such as the version of Gromacs, which command line was used to run the simulation, etc.

Our next step was to gain insight into the parameter settings employed by the MD community, which may aid us in identifying preferences in MD setups and potential necessity for further education to avoid suboptimal or outdated configurations. We therefore analyzed 10,055 .mdp files stored in the different data repositories. These text files contain information regarding the input parameters to run the simulations such as the integrator, the number of steps, the different algorithms for barostat and thermostat, etc. (for more details see: *https://manual.gromacs.org/documentation/current/user-guide/mdp-options.html*).

We determined the expected simulation time corresponding to the product of two parameters found in .mdp files: the number of steps and the time step. Here, we acknowledge that one can set up a very long simulation time and stop the simulation before the end or, on contrary, use a limited time (especially when calculations are performed on HPC resources with wall-time) and then extend the simulation for a longer duration. Using only the .mdp file, we cannot know if the simulation reached its term. To do so, comparison with an .xtc file from the same dataset may help to answer this specific question. However, in this study, we were interested in MD setup practices, in particular what simulation time researchers would set up their system with - likely in the mindset to reach that ending time. We restricted this analysis to the 4,623 .mdp files that used the *md* or *sd* integrator, and that have a simulation time above 1 ns. We found that the majority of the .mdp files were used for simulations of 50 ns or less (see **Fig. 4-A**). Further, 697 .mdp files with simulations times set-up between 50 ns and 1 µs and 585 .mdp files with simulation time above 1 µs were identified. As analyzing .gro files showed a good proportion of coarse-grained models (**Fig. 3-B,C**), we discriminated simulations setups for these two types of models using the time step as a simple cutoff. We considered that a time step greater than 10 fs (*i.e.* dt=0.01) corresponded to MD setups for coarse grained models (Ingólfsson et al., 2014). Globally, we found that over all simulations, the setups for atomistic simulations were largely dominant. However, for simulations with a simulation time above 1 µs specifically, coarse-grain simulations represented 86 % of all.

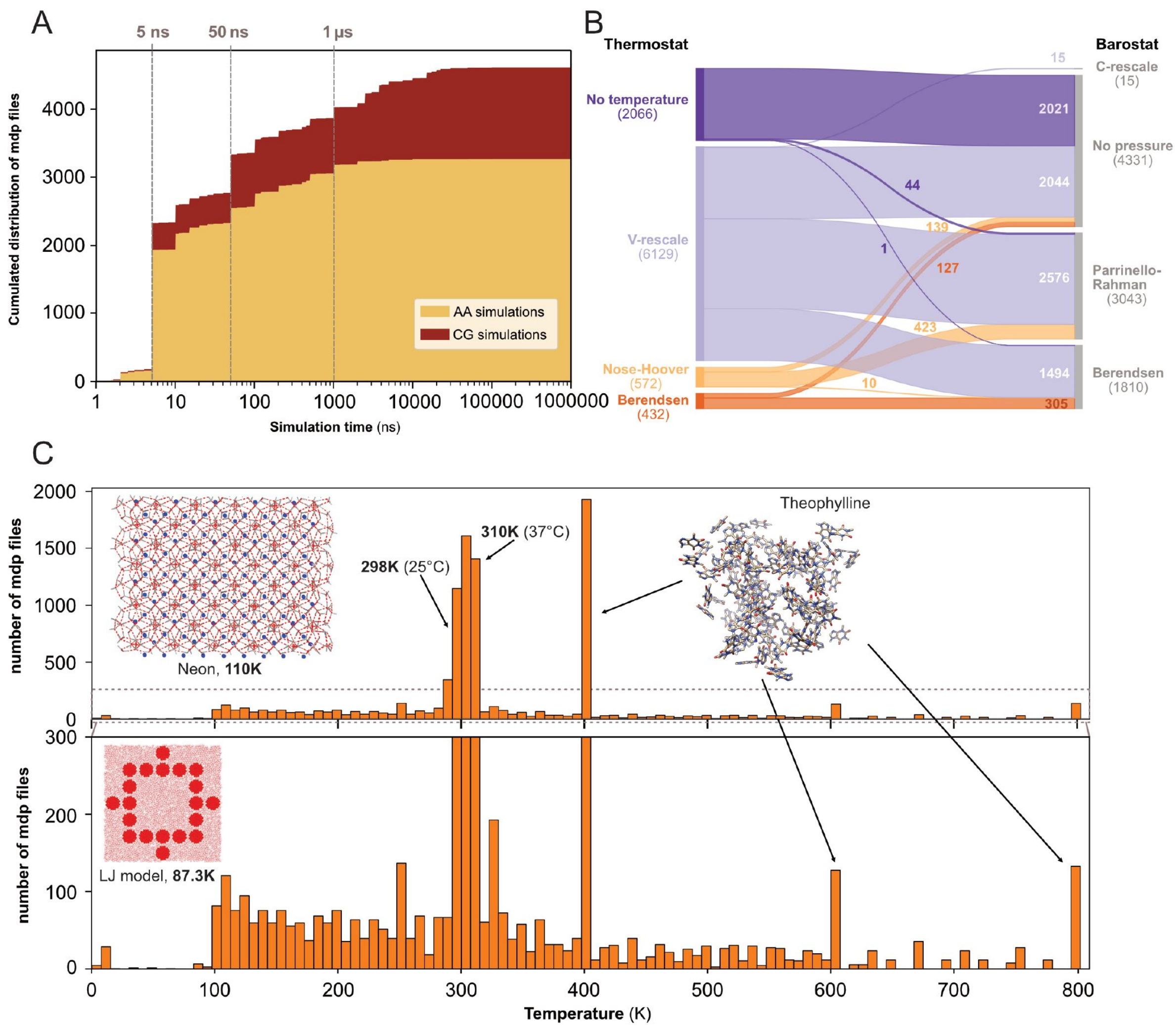


Figure 4.

Content analysis of .mdp files. (A) Cumulative distribution of .mdp files versus the simulation time for all-atom and coarse-grain simulations. (B) Sankey graph of the repartition between different values for thermostat and barostat. (C) Temperature distribution, full scale in upper panel and zoom-in in lower panel.

We then looked into the combinations of thermostat and barostat (see **Fig. 4-B**) from 9,199 .mdp files. The main thermostat used is by far the V-rescale (Bussi et al., 2007) often associated with the Parrinello-Rahman barostat (Parrinello and Rahman, 1981). This thermostat was also used with the Berendsen barostat (Berendsen et al., 1984). In a few cases, we observed the use of the V- rescale thermostat with the very recently developed C-rescale barostat (Bernetti and Bussi, 2020). A total of 2,021 .mdp files presented neither thermostat nor barostat, which means they would not be used in production runs. This could correspond to setups used for energy minimization, or to add ions to the system (with the genion command), or for molecular mechanics with Poisson–Boltzmann and surface area solvation (MM/PBSA) and molecular mechanics with generalised Born and surface area solvation (MM/GBSA) calculations (Genheden and Ryde, 2015).

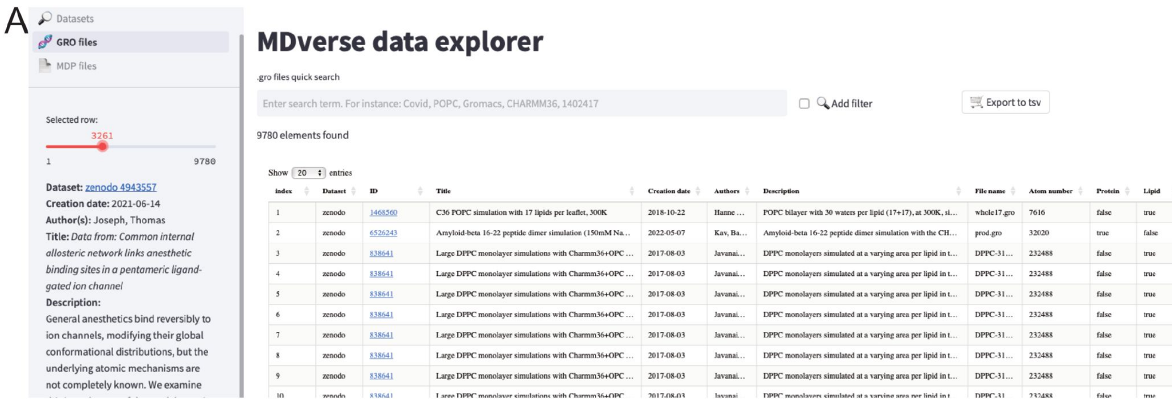
Finally, we analyzed the range of starting temperatures used to perform simulations (see **Fig. 4-C**). We found a clear peak around the temperatures 298 K - 310 K which corresponds to the range between ambient room (298 K - 25 °C) and physiological (310 K - 37 °C) temperatures. Nevertheless, we also observed lower temperatures, which often relate to studies of specific organic systems or simulations of Lennard-Jones models (Jeon et al., 2016). Interestingly, we noticed the appearance of several pikes at 400 K, 600 K, and 800 K, which were not present before the end of the year 2022. These peaks corresponded to the same study related to the stability of hydrated crystals (Dybeck et al., 2023). Overall, this analysis revealed that a wide range of temperatures have been explored, starting mostly from 100 K and going up to 800 K.

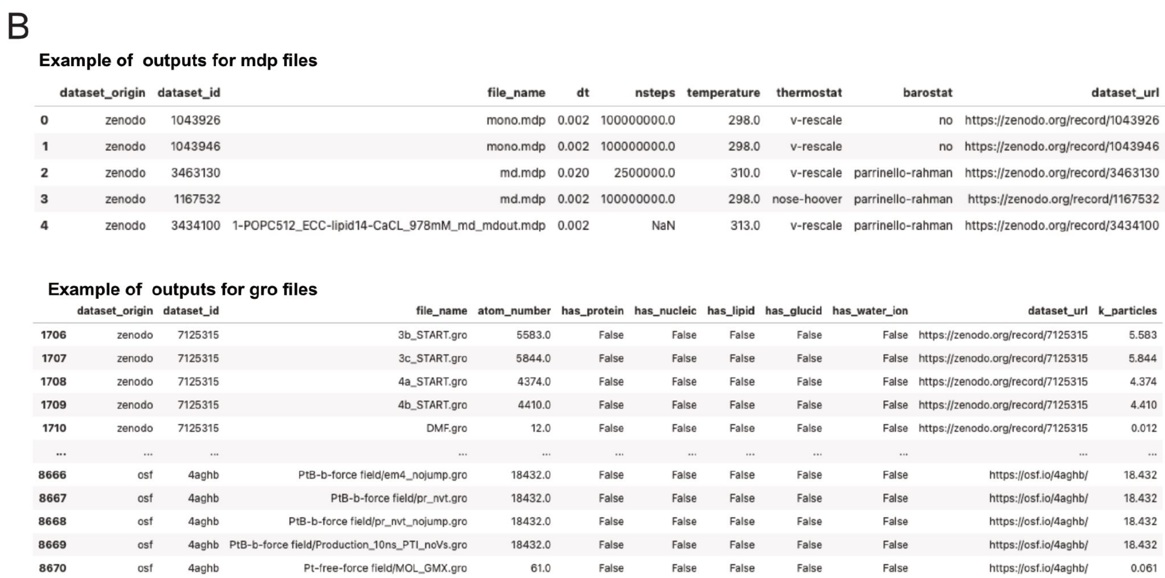
To encourage further analysis of the collected files, we shared our data collection with the community in Zenodo (see Data and code availability section). The data scrapping procedure and data analysis is available on GitHub with a detailed documentation. To let researchers having a quick glance and explore this data collection, we created a prototype web application called *MDverse data explorer* available at *https://mdverse.streamlit.app/* and illustrated in **Fig. 5-A**. With this web application, it is easy to use keywords and filters to access interesting datasets for all MD engines, as well as .gro and .mdp files. Furthermore, when available, a description of the found data is provided and searchable for keywords (**Fig. 5**-A, on the left sidebar). The sets of data found can then be exported as a tab-separated values (.tsv) file for further analysis (**Fig. 5-B**).

Towards a better sharing of MD data

With this work, we have shown that it was possible to not only retrieve MD data from the generalist data repositories Zenodo, Figshare and OSF, but to shed light onto the *dark matter of MD* data in terms of learning current scientific practice, extracting valuable topology information, and analysing how the field is developing. Our objective was not to assess the quality of the data but only to show what kind of data was available. The Ex^2 strategy to find files related to MD simulations relied on the fact that many MD software output files with specific file extensions. This strategy could not be applied in research fields where data exhibits non-specific file types. We experienced this limitation while indexing zip archives related to MD simulations, where we were able to decide if a zip archive was pertinent for this work only by accessing the list of files contained in the archive. This valuable feature is provided by data repositories like Zenodo and Figshare, with some caveats, though.

As of March 2023, we managed to index 245,756 files from 1,979 datasets, representing altogether 14 TB of data. This is a fraction of all files stored in data repositories. For instance, as of December 2022, Zenodo hosted about 9.9 million files for ~1.3 PB of data (Panero and Benito, 2022). All these files are stored on servers available 24/7. This high availability costs human resources, IT infrastructures and energy. Even if MD data represents only 1 % of the total volume of data stored in Zenodo, we believe it is our responsibility, as a community, to develop a better sharing and





B

Example of outputs for mdp files

	dataset_origin	dataset_id	file_name	dt	nsteps	temperature	thermostat	barostat	dataset_url
0	zenodo	1043926	mono.mdp	0.002	100000000.0	298.0	v-rescale	no	https://zenodo.org/record/1043926
1	zenodo	1043946	mono.mdp	0.002	100000000.0	298.0	v-rescale	no	https://zenodo.org/record/1043946
2	zenodo	3463130	md.mdp	0.020	2500000.0	310.0	v-rescale	parrinello-rahman	https://zenodo.org/record/3463130
3	zenodo	1167532	md.mdp	0.002	100000000.0	298.0	nose-hoover	parrinello-rahman	https://zenodo.org/record/1167532
4	zenodo	3434100	1-POPC512_ECC-lipid14-CaCL_978mM_md_mdout.mdp	0.002	NaN	313.0	v-rescale	parrinello-rahman	https://zenodo.org/record/3434100

Example of outputs for gro files

	dataset_origin	dataset_id	file_name	atom_number	has_protein	has_nucleic	has_lipid	has_glucid	has_water_ion	dataset_url	k_particles
1706	zenodo	7125315	3b_START.gro	5583.0	False	False	False	False	False	https://zenodo.org/record/7125315	5.583
1707	zenodo	7125315	3c_START.gro	5844.0	False	False	False	False	False	https://zenodo.org/record/7125315	5.844
1708	zenodo	7125315	4a_START.gro	4374.0	False	False	False	False	False	https://zenodo.org/record/7125315	4.374
1709	zenodo	7125315	4b_START.gro	4410.0	False	False	False	False	False	https://zenodo.org/record/7125315	4.410
1710	zenodo	7125315	DMF.gro	12.0	False	False	False	False	False	https://zenodo.org/record/7125315	0.012
...	...	...	...	...	...	...	...	...	...	...	...
8666	osf	4aghb	PtB-b-force field/em4_nojump.gro	18432.0	False	False	False	False	False	https://osf.io/4aghb/	18.432
8667	osf	4aghb	PtB-b-force field/pr_nvt.gro	18432.0	False	False	False	False	False	https://osf.io/4aghb/	18.432
8668	osf	4aghb	PtB-b-force field/pr_nvt_nojump.gro	18432.0	False	False	False	False	False	https://osf.io/4aghb/	18.432
8669	osf	4aghb	PtB-b-force field/Production_10ns_PTI_noVs.gro	18432.0	False	False	False	False	False	https://osf.io/4aghb/	18.432
8670	osf	4aghb	Pt-free-force field/MOL_GMX.gro	61.0	False	False	False	False	False	https://osf.io/4aghb/	0.061

Figure 5.

Snapshots of the MDverse data explorer, a prototype search engine to explore collected files and datasets. (A) General view of the web application. (B) Focus on the .mdp and .gro files sets of data exported as .tsv files. The web application also includes links to their original repository.

reuse of MD simulation files - and it will neither have to be particularly cumbersome nor expensive. To this end, we are proposing two solutions. First, improve practices for sharing and depositing MD data in data repositories. Second, improve the FAIRness of already available MD data notably by improving the quality of the current metadata.

Guidelines for better sharing of MD simulation data

Without a community-approved methodology for depositing MD simulation files in data repositories, and based on the current experience we described here, we propose a few simple guidelines when sharing MD data to make them more FAIR (Findable, Accessible, Interoperable and Reusable):

- Avoid zip or tar archives whose content cannot be properly indexed by data repositories. As much as possible, deposit original data files directly.
- Describe the MD dataset with extensive metadata. Provide adequate information along your dataset, such as:
    - - The scope of the study, e.g. investigate conformation dynamics, benchmark force field, …
    - - The method on a basic (e.g. quantum mechanics, all-atom, coarse-grain) or advanced (accelerated, metadynamics, well-tempered) level.
    - - The MD software: name, version (tag) and whether modifications have been made.
    - - The simulation settings (for each of the steps, including minimization, equilibration and production): temperature(s), thermostat, barostat, time step, total runtime (simulation length), force field, additional force field parameters.
    - - The composition of the system, with the precise names of the molecules and their numbers, if possible also PDB, UniProt or Ensemble identifiers and whether the default structure has been modified.
    - - Give information about any post-processing of the uploaded files (e.g. truncation or stripping of the trajectory), including before and after values of what has been modified e.g. number of frames or number of atoms of uploaded files
    - - Highlight especially valuable data, e.g. excessively QM-based parameterized molecules, and their parameter files.

Store this metadata in the description of the dataset. An adaptation of the Minimum Information About a Simulation Experiment (MIASE) guidelines (Waltemath et al., 2011) in the context of MD simulations would be useful to define required metadata.

- Link the MD dataset to other associated resources, such as:
    - - The research article (if any) for which these data have been produced. Datasets are usually mentioned in the research articles, but rarely the other way around, since the deposition has to be done prior to publication. However, it is eminently possible to submit a revised version, and providing a link to the related research paper in updated metadata of the MD dataset will ease the reference to the original publication upon data reuse.
    - - The code used to analyze the data, ideally deposited in the repository to guarantee availability, or in a GitHub or GitLab repository.
    - - Any other datasets that belong to the same study.

- Provide sufficient files to reproduce simulations and use a clear naming convention to make explicit links between related files. For instance, for the Gromacs MD engine, trajectory .xtc files could share the same names as structure .gro files (e.g. proteinA.gro & proteinA.xtc).

- Revisit your data deposition after paper acceptance and update information if necessary. Zenodo and Figshare provide a DOI for every new version of a dataset as well as a 'master' DOI that always refers to the latest version available.

These guidelines are complementary to the reliability and reproducibility checklist for molecular dynamics simulations (Commun Biol, 2023). Eventually, they could be implemented in machine actionable Data Management Plan (maDMP) (Miksa et al., 2019). So far, MD metadata is formalized as free text. We advocate for the creation of a standardized and controlled vocabulary to describe artifacts and properties of MD simulations. Normalized metadata will, in turn, enable scientific knowledge graphs (Auer, 2018; Färber and Lamprecht, 2021) that could link MD data, research articles and MD software in a rich network of research outputs.

Converging on a set of metadata and format requires a large consensus of different stakeholders, from users, to MD program developers, and journal editors. It would be especially useful to organize specific workshops with representatives of all these communities to collectively tackle this specific issue.

Improving metadata of current MD data

While indexing about 2,000 MD datasets, we found that title and description accompanying these datasets were very heterogeneous in terms of quality and quantity and were difficult for machines to process automatically. It was sometimes impossible to find even basic information such as the identity of the molecular system simulated, the temperature or the length of the simulation. Without appropriate metadata, sharing data is pointless, and its reuse is doomed to fail (Musen, 2022). It is thus important to close the gap between the availability of MD data and its discoverability and description through appropriate metadata. We could gradually improve the metadata by following two strategies. First, since MD engines produce normalized and well-documented files, we could extract parameters of the simulation by parsing specific files. We already explored this path with Gromacs, by extracting the molecular size and composition from .gro files and the simulation time (with some limitations), thermostat and barostat from .mdp files. We could go even further, by extracting for instance Gromacs version from .log file (if provided) or by identifying the simulated system from its atomic topology stored in .gro files. This strategy can in principle be applied to files produced by other MD engines. A second approach that we are currently exploring uses data mining and named entity recognition (NER) methods (Perera et al., 2020) to automatically identify the molecular system, the temperature, and the simulation length from existing textual metadata (dataset title and description), providing they are of sufficient length. Finally, the possibilities afforded by large language models supplemented by domain-specific tools (Bran et al., 2023) might help interpret the heterogenous metadata that is often associated with the simulations.

Future works

In the future, it is desirable to go further in terms of analysis and integrate other data repositories, such as Dryad and Dataverse instances (for example Recherche Data Gouv in France). The collaborative platform for source code GitHub could also be of interest. Albeit dedicated to source code and not designed to host large-size binary files, GitHub handles small to medium-size text files like tabular .csv and .tsv data files and has been extensively used to record cases of the Ebola epidemic in 2014 (Perkel, 2016) and the Covid-19 pandemic (*https://github.com/CSSEGISandData/COVID-19*). Thus, GitHub could probably host small text-based MD simulation files. For Gromacs, we already found 70,000 parameter .mdp files and 55,000 structure .gro files. Scripts found along these files could also provide valuable insights to understand how a given MD analysis was performed. Finally, GitHub repositories might also be an entry point to find other datasets by linking to simulation data, such as institutional repositories (see for instance (Pesce and Lindorff-Larsen, 2023)). However, one potential point of concern is that repositories like GitHub or GitLab do not make any promises about long-term availability of repositories, in particular ones not under active development. Archiving of these repositories could be achieved in Zenodo (for data-centric repositories) or Software Heritage (Di Cosmo and Zacchiroli, 2017) (for source-code-centric repositories).

An obvious next step is the enrichment of metadata with the hope to render open MD data more findable, accessible and ultimately reusable. Possible strategies have already been detailed previously in this paper. We could also go further by connecting MD data in the research ecosystem. For this, two apparent resources need to be linked to MD datasets: their associated research papers to mine more information and to establish a connection with the scientific context, and their simulated biomolecular systems, which ultimately could cross-reference MD datasets to reference databases such as UniProt (Consortium, 2022), the PDB (Berman et al., 2000) or Lipid Maps (Sud et al., 2007). For already deposited datasets, the enrichment of metadata can only be achieved via systematic computational approaches, while for future depositions, a clear and uniformly used ontology and dedicated metadata reference file (as it is used by the PLUMED-NEST: Bonomi et al. (2019)) would facilitate this task.

Eventually, front-end solutions such as the MDverse data explorer tool can evolve to being more user-friendly by interfacing the structures and dynamics with interactive 3D molecular viewers (Tiemann et al., 2017; Kampfrath et al., 2022; Martinez and Baaden, 2021).

Conclusion

In this work, we showed that sharing data generated from MD simulations is now a common practice. From Zenodo, Figshare and OSF alone, we indexed about 250,000 files from 2,000 datasets, and we showed that this trend is increasing. This data brings incentive and opportunities at different levels. First, for researchers who cannot access high-performance computing (HPC) facilities, or do not want to rerun a costly simulation to save time and energy, simulations of many systems are already available. These simulations could be useful to reanalyze existing trajectories, to extend simulations with already equilibrated systems or to compare simulations of a dedicated molecular system modelled with different settings. Second, building annotated and highly curated datasets for artificial intelligence will be invaluable to develop dynamic generative deep-learning models. Then, improving metadata along available data will foster their reuse and will mechanically increase the reproducibility of MD simulations. At last, we see here the occasion to push for good practices in the setup and production of MD simulations.

Methods and Materials

Initial data collection

We searched for MD-related files in the data repositories Zenodo, Figshare and Open Science Framework (OSF). Queries were designed with a combination of file types and optionally keywords, depending on how a given file type was solely associated to MD simulations. We therefore built a list of manually curated and cross-checked file types and keywords (*https://github.com/MDverse/mdws/blob/main/params/query.yml*). All queries were automated by Python scripts that utilized Application Programming Interfaces (APIs) provided by data repositories. Since APIs offered by data repositories were different, all implementations were performed in dedicated Python(van Rossum, 1995) (version 3.9.16) scripts with the NumPy(Oliphant, 2007) (version 1.24.2), Pandas(Wes McKin-ney, 2010) (version 1.5.3) and Requests (version 2.28.2) libraries.

We made the assumption that files deposited by researchers in data repositories were coherent and all related to a same research project. Therefore, when an MD-related file was found in a dataset, all files belonging to this dataset were indexed, regardless of whether their file types were actually identified as MD simulation files. This is the core of the Explore and Expand strategy (Ex^2) we applied in this work and illustrated in **Fig 1**. By default, the last version of the datasets was collected.

When a zip file was found in a dataset, its content was extracted from a preview provided by Zenodo and Figshare. This preview was not provided through APIs, but as HTML code, which we parsed using the Beautiful Soup library (version 4.11.2). Note that the zip file preview for Zenodo was limited to the first 1,000 files. To avoid false-positive files collected from zip archives, a final cleaning step was performed to remove all datasets that did not share at least one file type with the file type list mentioned above. In the case of OSF, there was no preview for zip files, so their content has not been retrieved.

Gromacs files

After the initial data collection, Gromacs .mdp and .gro files were downloaded with the Pooch library (version 1.6.0). When a .mdp or .gro file was found to be in a zip archive, the latter was downloaded and the targeted .mdp or .gro file was selectively extracted from the archive. The same procedure was applied for a subset of .xtc files that consisted of about one .xtc file per Gromacs datasets.

Once downloaded, .mdp files were parsed to extract the following parameters: integrator, time step, number of steps, temperature, thermostat, and barostat. Values for thermostat and barostat were normalized according to values provided by the Gromacs documentation. For the simulation time analysis, we selected .mdp files with the *md* or *sd* integrator and with simulation time above 1 ns to exclude most minimization and equilibrating simulations. For the thermostat and barostat analysis, only files with non-missing values and with values listed in the Gromacs documentation were considered.

The .gro files were parsed with the MDAnalysis library (Michaud-Agrawal et al., 2011) to extract the number of particles of the system. Values found in the residue name column were also extracted and compared to a list of residues we manually associated to the following categories: protein, lipid, nucleic acid, glucid and water or ions (*https://github.com/MDverse/mdws/blob/main/params/residue_names.yml*).

The .xtc files were analyzed using the gmxcheck command (*https://manual.gromacs.org/current/onlinehelp/gmx-check.html*) to extract the number of particles and the number of frames.

MDverse data explorer web app

The MDverse data explorer web application was built in Python with the Streamlit library. Data was downloaded from Zenodo (see the Data and code availability section).

System visualization and molecular graphics

Molecular graphics were performed with VMD (Humphrey et al., 1996) and Chimera (Pettersen et al., 2004). For all visualizations, .gro files containing molecular structure were used. In the case of the two structures in **Fig. 3-B**, .xtc files were manually assigned to their corresponding .gro (for the TonB protein) or .tpr (for the T4 Lysozyme) files based on their names in their datasets.

Origin of the structures displayed in this work:

TonB

Dataset URL: *https://zenodo.org/record/3756664*

Publication (DOI): *https://doi.org/10.1039/D0CP03473H*

T4 Lyzozyme

Dataset URL: *https://zenodo.org/record/3989044*

Publication (DOI): *https://doi.org/10.1021/acs.jctc.0c01338*

Benzene

Dataset URL: *https://figshare.com/articles/dataset/Capturing_Protein_Ligand_Recognition_Pathways_in_Coarse-Grained_Simulation/12517490/1*

Publication (DOI): *https://doi.org/10.1021/acs.jpclett.0c01683*

Ammonia

Dataset URL: *https://figshare.com/articles/dataset/Alchemical_Hydration_Free-Energy_Calculations_Using_Molecular_Dynamics_with_Explicit_Polarization_and_Induced_Polarity_Decoupling_An_On_the_Fly_Polarization_Approach/11702442*

Publication (DOI): *https://doi.org/10.1021/acs.jctc.9b01139*

Peptide with membrane

Dataset URL: *https://zenodo.org/record/4371296*

Publication (DOI): *https://doi.org/10.1021/acs.jcim.0c01312*

Kir channels

Dataset URL: *https://zenodo.org/record/3634884*

Publication (DOI): *https://doi.org/10.1073/pnas.1918387117*

Gasdermin

Dataset URL: *https://zenodo.org/record/6797842*

Publication (DOI): *https://doi.org/10.7554/eLife.81432*

Protein-RNA

Dataset URL: *https://zenodo.org/record/1308045*

Publication (DOI): *https://doi.org/10.1371/journal.pcbi.1006642*

G-quadruplex

Dataset URL: *https://zenodo.org/record/5594466*

Publication (DOI): *https://doi.org/10.1021/jacs.1c11248*

Ptb

Dataset URL: *https://osf.io/4aghb/*

Publication (DOI): *https://doi.org/10.1073/pnas.2116543119*

Epi 7170

Dataset URL: *https://zenodo.org/record/7120845*

Publication (DOI): *https://doi.org/10.1038/s41467-022-34077-z*

Gold nanoparticle

Dataset URL: Dataset URL: *https://acs.figshare.com/articles/dataset/Fluorescence_Probing_of_Thiol_Functionalized_Gold_Nanoparticles_Is_Alkylthiol_Coating_of_a_Nanoparticle_as_Hydrophobic_as_Expected_/2481241Publication* (DOI): *https://doi.org/10.1021/jp3060813*

Gd(DOTA)

Dataset URL: *https://acs.figshare.com/articles/dataset/Modeling_Gd_sup_3_sup_Complexes_for_Molecular_Dynamics_Simulations_Toward_a_Rational_Optimization_of_MRI_Contrast_Agents/20334621*
Publication (DOI): https://doi.org/10.1021/acs.inorgchem.2c01597.

Metalo cage

Dataset URL: *https://acs.figshare.com/articles/dataset/Rationalizing_the_Activity_of_an_Artificial_Diels-Alderase_Establishing_Efficient_and_Accurate_Protocols_for_Calculating_Supramolecular_Catalysis/11569452*

Publication (DOI): *https://doi.org/10.1021/jacs.9b10302*

Al1

Dataset URL: *https://acs.figshare.com/articles/dataset/Nucleation_Mechanisms_of_Self-Assembled_Physisorbed_Monolayers_on_Graphite/8846045*

Publication (DOI): *https://doi.org/10.1021/acs.jpcc.9b01234*

PTEG-1 (all-atom)

Dataset URL: *https://figshare.com/articles/dataset/PTEG-1_PP_and_N-DMBI_atomistic_force_fields/5458144*

Publication (DOI): *https://doi.org/10.1039/C7TA06609K*

PTEG-1 (coarse-grain)

Dataset URL: *https://figshare.com/articles/dataset/Neat_and_P3HT-Based_Blend_Morphologies_for_PCBM_and_PTEG-1/12338633*

Publication (DOI): *https://doi.org/10.1002/adfm.202004799*

Theophylline

Dataset URL: *https://figshare.com/articles/dataset/A_Comparison_of_Methods_for_Computing_Relative_Anhydrous_Hydrate_Stability_with_Molecular_Simulation/21644393*

Publication (DOI): *https://doi.org/10.1021/acs.cgd.2c00832*

Data and code availability

Data files produced from the data collection and processing are shared in Parquet format in the Zenodo repository: *https://zenodo.org/record/7856806*. They are freely available under the Creative Commons Attribution 4.0 International license (CC-BY).

Python scripts to search and index MD files, and to download and parse .mdp and .gro files are open-source (under the AGPL-3.0 license), freely available on GitHub (*https://github.com/MDverse/mdws*) and archived in Software Heritage (swh:1:dir:4d30b00345a732dcf9f79d3c8bfae38b35b8f2c4). A detailed documentation is provided along the scripts to easily reproduce the data collection and processing.

Jupyter notebooks used to analyze results and create the figures of this paper are open-source (under the BSD 3-Clause license), freely available on GitHub (*https://github.com/MDverse/mdda*) and archived in Software Heritage (swh:1:dir:1f8497f72134cef0a9724c955bb03c751f52cccd).

The code of the MDverse data explorer web application is open-source (under the BSD 3-Clause license), freely available on GitHub (*https://github.com/MDverse/mdde*) and archived in Software Heritage (swh:1:dir:1fc8b8eaabf4a9087e6d5b0ec5ed97031482bcbf).

Acknowledgements

We thank Lauri Mesilaakso, Bryan White, Jorge Hernansanz Biel, Zihwei Li and Kirill Baranov for their participation in the Copenhagen BioHackathon 2020, whose results showcased the need for a more advanced search strategy. We acknowledge Massimiliano Bonomi, Giovanni Bussi, Patrick Fuchs and Elise Lehoux for helpful discussions and suggestions. We also thank the Zenodo, Figshare and OSF support teams for providing figures on the content of their respective data repository and for their help in using APIs.

This work was supported by Institut français du Danemark (Blåtand program, 2021), the Data Intelligence Institute of Paris (diiP, IdEx Université Paris Cité, ANR-18-IDEX-0001, 2023). JKST and KL-L acknowledge funding by the Novo Nordisk Foundation [NNF18OC0033950 to KL-L], and workshops funded by The BioExcel Center-of-Excellence (grant agreements 823830, 101093290).

Author contributions

The original idea was conceived by EL together with JKST, MC, RH and LD. JKST, MC and PP supervised the project. JKST, PP, MC and SG conceived the search strategy. PP, JKST and LB implemented the search strategy. PP performed the analysis and interpreted the results with MS, JKST, MC and KL-L. PP and MO generated the MDverse web interface. JKST, PP and MC discussed all designs and results. MC and PP designed the figures. JKST, MB, MC and PP wrote the manuscript with input from all authors.

References

Abraham M *et al.* (2019) **Sharing Data from Molecular Simulations** *Journal of Chemical Information and Modeling* **59**:4093–4099 https://doi.org/10.1021/acs.jcim.9b00665

Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, Lindahl E. (2015) **GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers** *SoftwareX* **1**:19–25 https://doi.org/10.1016/j.softx.2015.06.001

Abriata LA, Lepore R, Dal Peraro M. (2020) **About the need to make computational models of biological macromolecules available and discoverable** *Bioinformatics* **36**:2952–2954 https://doi.org/10.1093/bioinformatics/btaa086

Aldeghi M, Heifetz A, Bodkin MJ, Knapp S, Biggin PC (2015) **Accurate calculation of the absolute free energy of binding for drug molecules** *Chemical Science* **7**:207–218 https://doi.org/10.1039/c5sc02678d

Alessandri R, Grünewald F, Marrink SJ (2021) **The Martini Model in Materials Science** *Advanced Materials* **33** https://doi.org/10.1002/adma.202008635

Alessandri R, Sami S, Barnoud J, Vries AH, Marrink SJ, Havenith RWA (2020) **Resolving Donor–Acceptor Interfaces and Charge Carrier Energy Levels of Organic Semiconductors with Polar Side Chains** *Advanced Functional Materials* **30** https://doi.org/10.1002/adfm.202004799

Amaro RE, Mulholland AJ (2020) **A Community Letter Regarding Sharing Biomolecular Simulation Data for COVID-19** *Journal of Chemical Information and Modeling* :2653–2656 https://doi.org/10.1021/acs.jcim.0c00319

Antila HS, M Ferreira T, Ollila OHS, Miettinen MS (2021) **Using Open Data to Rapidly Benchmark Biomolecular Simulations: Phospholipid Conformational Dynamics** *Journal of Chemical Information and Modeling* :938–949 https://doi.org/10.1021/acs.jcim.0c01299

Armstrong DR *et al.* (2019) **PDBe: improved findability of macromolecular structure data in the PDB** *Nucleic Acids Research* **48**:D335–D343 https://doi.org/10.1093/nar/gkz990

Auer S (2018) **Towards an Open Research Knowledge Graph** https://doi.org/10.5281/zenodo.1157185

Berendsen HJC, Postma JPM, Gunsteren WFV, DiNola A, Haak JR (1984) **Molecular dynamics with coupling to an external bath** *The Journal of Chemical Physics* **81**:3684–3690 https://doi.org/10.1063/1.448118

Berendsen HJC, van der Spoel D, van Drunen R. (1995) **GROMACS: A Message-Passing Parallel Molecular Dynamics Implementation** *Computer Physics Communications* **91**:43–56 https://doi.org/10.1016/0010-4655(95)00042-E

Berman H, Henrick K, Nakamura H. (2003) **Announcing the worldwide Protein Data Bank** *Nature structural biology* **10** https://doi.org/10.1038/nsb1203-980

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) **The Protein Data Bank** *Nucleic Acids Research* **28**:235–242 https://doi.org/10.1093/nar/28.1.235

Bernetti M, Bussi G. (2020) **Pressure control using stochastic cell rescaling** *The Journal of Chemical Physics* **153** https://doi.org/10.1063/5.0020514

Bonomi M *et al.* (2019) **Promoting transparency and reproducibility in enhanced molecular simulations** *Nature Methods* **16**:670–673 https://doi.org/10.1038/s41592-019-0506-8

Bottaro S, Lindorff-Larsen K. (2018) **Biophysical experiments and biomolecular simulations: A perfect match?** *Science* **361**:355–360 https://doi.org/10.1126/science.aat4010

Bowers KJ *et al.* (2006) **Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters** *ACM/IEEE SC 2006 Conference (SC'06)* :43–43 https://doi.org/10.1109/sc.2006.54

Bran AM, Cox S, White AD, Schwaller P (2023) **ChemCrow: Augmenting large-language models with chemistry tools**

Brooks BR *et al.* (2009) **CHARMM: the biomolecular simulation program** *Journal of computational chemistry* **30**:1545–1614 https://doi.org/10.1002/jcc.21287

Burley SK, Kurisu G, Markley JL, Nakamura H, Velankar S, Berman HM, Sali A, Schwede T, Trewhella J. (2017) **PDB-Dev: a Prototype System for Depositing Integrative/Hybrid Structural Models** *Structure (London, England : 1993)* **25**:1317–1318 https://doi.org/10.1016/j.str.2017.08.001

Bussi G, Donadio D, Parrinello M. (2007) **Canonical sampling through velocity rescaling** *The Journal of Chemical Physics* **126** https://doi.org/10.1063/1.2408420

Biol Commun (2023) **Reliability and reproducibility checklist for molecular dynamics simulations** *Communications Biology* **6** https://doi.org/10.1038/s42003-023-04653-0

consortium TP (2019) **Promoting transparency and reproducibility in enhanced molecular simulations** *Nat Methods* **16**:670–673 https://doi.org/10.1038/s41592-019-0506-8

Consortium TU (2022) **UniProt: the Universal Protein Knowledgebase in 2023** *Nucleic Acids Research* **51**:D523–D531 https://doi.org/10.1093/nar/gkac1052

Dandekar BR, Mondal J. (2020) **Capturing Protein–Ligand Recognition Pathways in Coarse-Grained Simulation** *The Journal of Physical Chemistry Letters* **11**:5302–5311 https://doi.org/10.1021/acs.jpclett.0c01683

Di Cosmo R, Zacchiroli S. (2017) **Software Heritage: Why and How to Preserve Software Source Code** *In: Proceedings of the 14th International Conference on Digital Preservation*

Domański J, Stansfeld PJ, Sansom MSP, Beckstein O. (2010) **Lipidbook: a public repository for force-field parameters used in membrane simulations** *The Journal of membrane biology* **236**:255–258 https://doi.org/10.1007/s00232-010-9296-8

Duncan AL, Corey RA, Sansom MSP (2020) **Defining how multiple lipid species interact with inward rectifier potassium (Kir2) channels** *Proc Natl Acad Sci USA* **117**:7803–7813 https://doi.org/10.1073/pnas.1918387117

Dybeck EC, Thiel A, Schnieders MJ, Pickard FC, Wood GPF, Krzyzaniak JF, Hancock BC (2023) **A Comparison of Methods for Computing Relative Anhydrous–Hydrate Stability with Molecular Simulation** *Crystal Growth & Design* **23**:142–167 https://doi.org/10.1021/acs.cgd.2c00832

Elofsson A, Hess B, Lindahl E, Onufriev A, Spoel DVD, Wallqvist A. (2019) **Ten simple rules on how to create open access and reproducible molecular simulations of biological systems** *PLOS Computational Biology* **15** https://doi.org/10.1371/journal.pcbi.1006649

European Organization For Nuclear Research, OpenAIRE, Zenodo (2013) **European Organization For Nuclear Research, OpenAIRE, Zenodo. CERN; 2013. https://www.zenodo.org/, doi: 10.25495/7GXK-RD71.** https://doi.org/10.25495/7GXK-RD71

Fadda E. (2022) **Molecular simulations of complex carbohydrates and glycoconjugates** *Current Opinion in Chemical Biology* **69** https://doi.org/10.1016/j.cbpa.2022.102175

Fan FJ, Shi Y. (2022) **Effects of data quality and quantity on deep learning for protein-ligand binding affinity prediction** *Bioorganic & Medicinal Chemistry* **72** https://doi.org/10.1016/j.bmc.2022.117003

Fawzi NL, Parekh SH, Mittal J. (2021) **Biophysical studies of phase separation integrating experimental and computational methods** *Current Opinion in Structural Biology* **70**:78–86 https://doi.org/10.1016/j.sbi.2021.04.004

Ferrer RS, Case DA, Walker RC (2012) **An overview of the Amber biomolecular simulation package** *Wiley Interdisciplinary Reviews: Computational Molecular Science* **3**:198–210 https://doi.org/10.1002/wcms.1121

Fuller JC, Jackson RM, Edwards TA, Wilson AJ, Shirts MR (2012) **Modeling of Arylamide Helix Mimetics in the p53 Peptide Binding Site of hDM2 Suggests Parallel and Anti-Parallel Conformations Are Both Stable** *PLOS ONE* **7**:1–17 https://doi.org/10.1371/journal.pone.0043253

Färber M, Lamprecht D. (2021) **The data set knowledge graph: Creating a linked open data source for data sets** *Quantitative Science Studies* **2**:1324–1355 https://doi.org/10.1162/qss_a_00161

Genheden S, Ryde U. (2015) **The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities** *Expert Opinion on Drug Discovery* **10**:449–461 https://doi.org/10.1517/17460441.2015.1032936

Gertsen AS, Sørensen MK, Andreasen JW (2020) **Nanostructure of organic semiconductor thin films: Molecular dynamics modeling with solvent evaporation** *Physical Review Materials* **4** https://doi.org/10.1103/physrevmaterials.4.075405

Gowers R *et al.* (2016) **MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations** https://doi.org/10.25080/majora-629e541a-00e

Gupta C, Sarkar D, Tieleman DP, Singharoy A. (2022) **The ugly, bad, and good stories of large-scale biomolecular simulations** *Current Opinion in Structural Biology* **73** https://doi.org/10.1016/j.sbi.2022.102338

Hoch JC *et al.* (2023) **Biological Magnetic Resonance Data Bank** *Nucleic Acids Research* **51**:D368–D376

Hollingsworth SA, Dror RO (2018) **Molecular Dynamics Simulation for All** *Neuron* **99**:1129–1143 https://doi.org/10.1016/j.neuron.2018.08.011

Hospital A, Battistini F, Soliva R, Gelpí JL, Orozco M. (2020) **Surviving the deluge of biosimulation data** *WIREs Computational Molecular Science* **10** https://doi.org/10.1002/wcms.1449

Humphrey W, Dalke A, Schulten K. (1996) **VMD: visual molecular dynamics** *J Mol Graph* **14**:33–8

Hénin J, Lelièvre T, Shirts MR, Valsson O, Delemotte L. (2022) **Enhanced Sampling Methods for Molecular Dynamics Simulations [Article v1.0]** *Living Journal of Computational Molecular Science* **4** https://doi.org/10.33011/livecoms.4.1.1583

Ingólfsson HI, Lopez CA, Uusitalo JJ, Jong DHd, Gopal SM, Periole X, Marrink SJ (2014) **The power of coarse graining in biomolecular simulations** *Wiley Interdisciplinary Reviews: Computational Molecular Science* **4**:225–248 https://doi.org/10.1002/wcms.1169

Ivanov P, Mu J, Leay L, Chang SY, Sharrad CA, Masters AJ, Schroeder SLM (2017) **Organic and Third Phase in HNO3/TBP/n-Dodecane System: No Reverse Micelles** *Solvent Extraction and Ion Exchange* **35**:251–265 https://doi.org/10.1080/07366299.2017.1336048

Jahani F, Torabi S, Chiechi RC, Koster LJA, Hummelen JC (2014) **Fullerene derivatives with increased dielectric constants** *Chemical Communications* **50**:10645–10647 https://doi.org/10.1039/c4cc04366a

Jeon JH, Javanainen M, Martinez-Seara H, Metzler R. (2016) **Protein Crowding in Lipid Bilayers Gives Rise to Non-Gaussian Anomalous Lateral Diffusion of Phospholipids and Proteins** *Physical Review X* https://doi.org/10.1103/physrevx.6.021006

Jumper J *et al.* (2021) **Highly accurate protein structure prediction with AlphaFold** *Nature* :583–589 https://doi.org/10.1038/s41586-021-03819-2

Kabelka I, Brožek R, Vácha R. (2021) **Selecting Collective Variables and Free-Energy Methods for Peptide Translocation across Membranes** *Journal of Chemical Information and Modeling* **61**:819–830 https://doi.org/10.1021/acs.jcim.0c01312

Kampfrath M, Staritzbichler R, Hernández GP, Rose AS, Tiemann JKS, Scheuermann G, Wiegreffe D, Hildebrand PW (2022) **MDsrv: visual sharing and analysis of molecular dynamics simulations** *Nucleic Acids Research* **50**:W483–W489 https://doi.org/10.1093/nar/gkac398

Karunasena C, Li S, Heifner MC, Ryno SM, Risko C. (2021) **Reconsidering the Roles of Noncovalent Intramolecular "Locks" in π-Conjugated Molecules** *Chemistry of Materials* **33**:9139–9151 https://doi.org/10.1021/acs.chemmater.1c02335

Kelly BD, Smith WR (2020) **Alchemical Hydration Free-Energy Calculations Using Molecular Dynamics with Explicit Polarization and Induced Polarity Decoupling: An On–the–Fly Polarization Approach** *Journal of Chemical Theory and Computation* **16**:1146–1161 https://doi.org/10.1021/acs.jctc.9b01139

Kiirikki A *et al.* (2023) **NMRlipids Databank makes data-driven analysis of biomembrane properties accessible for all** https://doi.org/10.26434/chemrxiv-2023-jrpwm

Kinjo AR, Bekker GJ, Suzuki H, Tsuchiya Y, Kawabata T, Ikegawa Y, Nakamura H. (2017) **Protein Data Bank Japan (PDBj): updated user interfaces, resource description framework, analysis tools for large structures** *Nucleic Acids Research* **45**:D282–D288 https://doi.org/10.1093/nar/gkw962

Kirschner KN, Yongye AB, Tschampel SM, González-Outeiriño J, Daniels CR, Foley BL, Woods RJ **GLYCAM06: a generalizable biomolecular force field** *Carbohydrates. Journal of computational chemistry* **29**:622–655 https://doi.org/10.1002/jcc.20820

Krishna S, Sreedhar I, Patel CM (2021) **Molecular dynamics simulation of polyamide-based materials – A review** *Computational Materials Science* **200** https://doi.org/10.1016/j.commatsci.2021.110853

Kyrychenko A, Karpushina GV, Svechkarev D, Kolodezny D, Bogatyrenko SI, Kryshtal AP, Doroshenko AO (2012) **Fluorescence Probing of Thiol-Functionalized Gold Nanoparticles: Is Alkylthiol Coating of a Nanoparticle as Hydrophobic as Expected?** *The Journal of Physical Chemistry C* **116**:21059–21068 https://doi.org/10.1021/jp3060813

Kümmerer F, Orioli S, Harding-Larsen D, Hoffmann F, Gavrilov Y, Teilum K, Lindorff-Larsen K. (2021) **Fitting Side-Chain NMR Relaxation Data Using Molecular Simulations** *Journal of Chemical Theory and Computation* **17**:5262–5275 https://doi.org/10.1021/acs.jctc.0c01338

Lane TJ (2023) **Protein structure prediction has reached the single-structure frontier** *Nature Methods* :1–4 https://doi.org/10.1038/s41592-022-01760-4

Liu S, Cao S, Hoang K, Young KL, Paluch AS, Mobley DL (2016) **Using MD Simulations To Calculate How Solvents Modulate Solubility** *Journal of Chemical Theory and Computation* **12**:1930–1941 https://doi.org/10.1021/acs.jctc.5b00934

Mahmud M, Kaiser MS, McGinnity TM, Hussain A. (2021) **Deep Learning in Mining Biological Data** *Cognitive Computation* **13**:1–33 https://doi.org/10.1007/s12559-020-09773-x

Marklund EG, Benesch JL (2019) **Weighing-up protein dynamics: the combination of native mass spectrometry and molecular dynamics simulations** *Current Opinion in Structural Biology* **54**:50–58 https://doi.org/10.1016/j.sbi.2018.12.011

Martinez X, Baaden M. (2021) **UnityMol prototype for FAIR sharing of molecular-visualization experiences: from pictures in the cloud to collaborative virtual reality exploration in immersive 3D environments** *Acta Crystallographica Section D* **77**:746–754 https://doi.org/10.1107/s2059798321002941

Marx V. (2013) **Biology: The Big Challenges of Big Data** *Nature* **498**:255–260 https://doi.org/10.1038/498255a

McKinney Wes, Walt Stéfan van der, Millman Jarrod (2010) **Data Structures for Statistical Computing in Python** *Proceedings of the 9th Python in Science Conference* :56–61 https://doi.org/10.25080/Majora-92bf1922-00a

Merz KMJ, Amaro R, Cournia Z, Rarey M, Soares T, Tropsha A, Wahab HA, Wang R. (2020) **Editorial: Method and Data Sharing and Reproducibility of Scientific Results** *Journal of Chemical Information and Modeling* :5868–5869 https://doi.org/10.1021/acs.jcim.0c01389

Meyer T *et al.* (2010) **MoDEL (Molecular Dynamics Extended Library): A Database of Atomistic Molecular Dynamics Trajectories** *Structure* :1399–1409 https://doi.org/10.1016/j.str.2010.07.013

Michaud-Agrawal N, Denning EJ, Woolf TB, Beckstein O. (2011) **MDAnalysis: A toolkit for the analysis of molecular dynamics simulations** *Journal of computational chemistry* **32**:2319–2327 https://doi.org/10.1002/jcc.21787

Miksa T, Simms S, Mietchen D, Jones S. (2019) **Ten principles for machine-actionable data management plans** *PLOS Computational Biology* **15**:1–15 https://doi.org/10.1371/journal.pcbi.1006750

Mulholland AJ, Amaro RE (2020) **COVID19 - Computational Chemists Meet the Moment** *Journal of Chemical Information and Modeling* **60**:5724–5726 https://doi.org/10.1021/acs.jcim.0c01395

Musen MA (2022) **Without Appropriate Metadata, Data-Sharing Mandates Are Pointless** *Nature* **609**:222–222 https://doi.org/10.1038/d41586-022-02820-7

Newport TD, Sansom MSP, Stansfeld PJ (2018) **The MemProtMD database: a resource for membrane-embedded protein structures and their lipid interactions** *Nucleic Acids Research* **47** https://doi.org/10.1093/nar/gky1047

Oliphant TE (2007) **Python for Scientific Computing** *Computing in Science & Engineering* **9**:10–20 https://doi.org/10.1109/MCSE.2007.58

Panero P, Benito J (2022) **OpenAIRE Webinar: Zenodo - open digital repository** https://doi.org/10.5281/zenodo.7417839

Parrinello M, Rahman A. (1981) **Polymorphic transitions in single crystals: A new molecular dynamics method** *Journal of Applied Physics* **52** https://doi.org/10.1063/1.328693

Perera N, Dehmer M, Emmert-Streib F. (2020) **Named Entity Recognition and Relation Detection for Biomedical Information Extraction** *Frontiers in Cell and Developmental Biology* **8** https://doi.org/10.3389/fcell.2020.00673

Perilla JR, Goh BC, Cassidy CK, Liu B, Bernardi RC, Rudack T, Yu H, Wu Z, Schulten K. (2015) **Molecular dynamics simulations of large macromolecular complexes** *Current opinion in structural biology* **31**:64–74 https://doi.org/10.1016/j.sbi.2015.03.007

Perkel J. (2016) **Democratic Databases: Science on GitHub** *Nature* **538**:127–128 https://doi.org/10.1038/538127a

Pesce F, Lindorff-Larsen K. (2023) **Combining experiments and simulations to examine the temperature-dependent behaviour of a disordered protein** https://doi.org/10.1101/2023.03.04.531094

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) **UCSF Chimera–a visualization system for exploratory research and analysis** *Journal of computational chemistry* **25**:1605–1612 https://doi.org/10.1002/jcc.20084

Phillips JC *et al.* (2020) **Scalable molecular dynamics on CPU and GPU architectures with NAMD** *The Journal of Chemical Physics* **153** https://doi.org/10.1063/5.0014475

Piskorz TK, Gobbo C, Marrink SJ, Feyter SD, AHd Vries, JHv Esch (2019) **Nucleation Mechanisms of Self-Assembled Physisorbed Monolayers on Graphite** *The Journal of Physical Chemistry C* **123**:17510–17520 https://doi.org/10.1021/acs.jpcc.9b01234

Pohjolainen E, Chen X, Malola S, Groenhof G, Häkkinen H. (2016) **A Unified AMBER-Compatible Molecular Mechanics Force Field for Thiolate-Protected Gold Nanoclusters** *Journal of Chemical Theory and Computation* **12**:1342–1350 https://doi.org/10.1021/acs.jctc.5b01053

Porubsky VL, Goldberg AP, Rampadarath AK, Nickerson DP, Karr JR, Sauro HM (2020) **Best Practices for Making Reproducible Biochemical Models** *Cell Systems* **11**:109–120 https://doi.org/10.1016/j.cels.2020.06.012

Qiu L, Liu J, Alessandri R, Qiu X, Koopmans M, Havenith RWA, Marrink SJ, Chiechi RC, Koster LJA, Hummelen JC (2017) **Enhancing doping efficiency by improving host-dopant miscibility for fullerene-based n-type thermoelectrics** *Journal of Materials Chemistry A* **5**:21234–21241 https://doi.org/10.1039/c7ta06609k

Rodríguez-Espigares I *et al.* (2020) **GPCRmd uncovers the dynamics of the 3D-GPCRome** *Nature Methods* **17**:777–787 https://doi.org/10.1038/s41592-020-0884-y

Sami S, Alessandri R, Wijaya JBW, Grünewald F, AHd Vries, Marrink SJ, Broer R, Havenith RWA (2022) **Strategies for Enhancing the Dielectric Constant of Organic Materials** *The Journal of Physical Chemistry C* **126**:19462–19469 https://doi.org/10.1021/acs.jpcc.2c05682

Sarkar A *et al.* (2020) **Self-Sorted, Random, and Block Supramolecular Copolymers via Sequence Controlled, Multicomponent Self-Assembly** *Journal of the American Chemical Society* **142**:7606–7617 https://doi.org/10.1021/jacs.0c01822

Schaefer SL, Hummer G. (2022) **Sublytic gasdermin-D pores captured in atomistic molecular simulations** *eLife* **11** https://doi.org/10.7554/elife.81432

Souza PCT *et al.* (2021) **Martini 3: a general purpose force field for coarse-grained molecular dynamics** *Nature Methods* :1–7 https://doi.org/10.1038/s41592-021-01098-3

Stansfeld PJ, Goose JE, Caffrey M, Carpenter EP, Parker JL, Newstead S, Sansom MSP (2015) **MemProtMD: Automated Insertion of Membrane Protein Structures into Explicit Lipid Membranes** *Structure (London, England : 1993)* **23**:1350–1361 https://doi.org/10.1016/j.str.2015.05.006

Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE (2015) **Big Data: Astronomical or Genomical?** *PLOS Biology* **13** https://doi.org/10.1371/journal.pbio.1002195

Sud M *et al.* (2007) **LMSD: LIPID MAPS structure database** *Nucleic Acids Research* **35**:D527–D532 https://doi.org/10.1093/nar/gkl838

Tai K, Murdock S, Wu B, Ng MH, Johnston S, Fangohr H, Cox SJ, Jeffreys P, Essex JW, Sansom MSP (2004) **BioSimGrid: towards a worldwide repository for biomolecular simulations** *Organic & Biomolecular Chemistry* :3219–3221 https://doi.org/10.1039/B411352G

Tiemann JKS, Guixà-González R, Hildebrand PW, Rose AS (2017) **MDsrv: viewing and sharing molecular dynamics simulations on the web** *Nat Methods* **14**:1123–1124 https://doi.org/10.1038/nmeth.4497 van Rossum G. (1995) **Python Tutorial**

Vanommeslaeghe K, MacKerell AD (2012) **Automation of the CHARMM General Force Field (CGenFF) I: Bond Perception and Atom Typing** *Journal of Chemical Information and Modeling* **52**:3144–3154 https://doi.org/10.1021/ci300363c

Virtanen SI, Kiirikki AM, Mikula KM, Iwaï H, Ollila OHS (2020) **Heterogeneous dynamics in partially disordered proteins** *Physical Chemistry Chemical Physics* **22**:21185–21196 https://doi.org/10.1039/d0cp03473h

Vuorio J, Vattulainen I, Martinez-Seara H. (2017) **Atomistic fingerprint of hyaluronan–CD44 binding** *PLoS Computational Biology* **13** https://doi.org/10.1371/journal.pcbi.1005663

Waltemath D *et al.* (2011) **Minimum Information About a Simulation Experiment (MIASE)** *PLOS Computational Biology* **7**:1–4 https://doi.org/10.1371/journal.pcbi.1001122

Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) **Development and testing of a general amber force field** *Journal of computational chemistry* **25**:1157–1174 https://doi.org/10.1002/jcc.20035

Wilkinson MD *et al.* (2016) **The FAIR Guiding Principles for scientific data management and stewardship** *Scientific data* **3** https://doi.org/10.1038/sdata.2016.18

Wilson SL, Way GP, Bittremieux W, Armache JP, Haendel MA, Hoffman MM (2021) **Sharing biological data: why, when, and how** *FEBS Letters* **595**:847–863 https://doi.org/10.1002/1873-3468.14067

Yoo J, Winogradoff D, Aksimentiev A. (2020) **Molecular dynamics simulations of DNA-DNA and DNA-protein interactions** *Current Opinion in Structural Biology* **64**:88–96 https://doi.org/10.1016/j.sbi.2020.06.007

Young TA, Martí-Centelles V, Wang J, Lusby PJ, Duarte F. (2020) **Rationalizing the Activity of an "Artificial Diels-Alderase": Establishing Efficient and Accurate Protocols for Calculating Supramolecular Catalysis** *Journal of the American Chemical Society* **142**:1300–1310 https://doi.org/10.1021/jacs.9b10302

Zheng X *et al.* (2022) **Elucidation of the key role of Pt…Pt interactions in the directional self-assembly of platinum(II) complexes** *Proceedings of the National Academy of Sciences* **119** https://doi.org/10.1073/pnas.2116543119

Zhu J, Salvatella X, Robustelli P. (2022) **Small molecules targeting the disordered transactivation domain of the androgen receptor induce the formation of collapsed helical states** *Nature Communications* **13** https://doi.org/10.1038/s41467-022-34077-z

Zhu S. (2019) **Validation of the Generalized Force Fields GAFF, CGenFF, OPLS-AA, and PRODRGFF by Testing Against Experimental Osmotic Coefficient Data for Small Drug-Like Molecules** *Journal of Chemical Information and Modeling* **59**:4239–4247 https://doi.org/10.1021/acs.jcim.9b00552

Editors

Reviewing Editor
Shozeb Haider
University College London, London, United Kingdom

Senior Editor
Qiang Cui
Boston University, Boston, United States of America

Reviewer #1 (Public Review):

Summary:

Tiemann et al. have undertaken an original study on the availability of molecular dynamics (MD) simulation datasets across the Internet. There is a widespread belief that extensive, well-curated MD datasets would enable the development of novel classes of AI models for structural biology. However, currently, there is no standard for sharing MD datasets. As generating MD datasets is energy-intensive, it is also important to facilitate the reuse of MD datasets to minimize energy consumption. Developing a universally accepted standard for depositing and curating MD datasets is a huge undertaking. The study by Tiemann et al. will be very valuable in informing policy developments toward this goal.

Strengths:

The study presents an original approach to addressing a growing concern in the field. It is clear that adopting a more collaborative approach could significantly enhance the impact of MD simulations in modern molecular sciences.

The timing of the work is appropriate, given the current interest in developing AI models for describing biomolecular dynamics.

Weaknesses:

The study primarily focuses on one major MD engine (GROMACS), although this limitation is not significant considering the proof-of-concept nature of the study.

https://doi.org/10.7554/eLife.90061.2.sa2

Reviewer #2 (Public Review):

Summary:

Molecular dynamics (MD) data is deposited in public, non-specialist repositories. This work starts from the premise that these data are a valuable resource as they could be used by other researchers to extract additional insights from these simulations; it could also potentially be used as training data for ML/AI approaches. The problem is that mining these data is difficult because they are not easy to find and work with. The primary goal of the authors was to discover and index these difficult-to-find MD datasets, which they call the "dark matter of the MD universe" (in contrast to data sets held in specialist databases).

The authors developed a search strategy that avoided the use of ill-defined metadata but instead relied on the knowledge of the restricted set of file formats used in MD simulations as a true marker for the data they were looking for. Detection of MD data marked a data set as relevant with a follow-up indexing strategy of all associated content. This "explore-and-expand" strategy allowed the authors for the first time to provide a realistic census of the MD data in non-specialist repositories.

As a proof of principle, they analyzed a subset of the data (primarily related to simulations with the popular Gromacs MD package) to summarize the types of simulated systems (primarily biomolecular systems) and commonly used simulation settings.

Based on their experience they propose best practices for metadata provision to make MD data FAIR (findable, accessible, interoperable, reusable).

A prototype search engine that works on the indexed datasets is made publicly available. All data and code are made freely available as open source/open data.

Strengths:

- The novel search strategy is based on relevant data to identify full datasets instead of relying on metadata and thus is likely to have many true positives and few false positives.

- The paper provides a first glimpse at the potential hidden treasures of MD simulations and force field parametrizations of molecules.

- Analysis of parameter settings of MD simulations from how researchers *actually* run simulations can provide valuable feedback to MD code developers for how to document/educate users. This approach is much better than analyzing what authors write in the Methods sections.

- The authors make a prototype search engine available.

- The guidelines for FAIR MD data are based on experience gained from trying to make sense of the data.

Weaknesses:

- So far the work is a proof-of-concept that focuses on MD data produced by Gromacs (which was prevalent under all indexed and identified packages).

As discussed in the manuscript, some types of biomolecules are likely underrepresented because different communities have different preferences for force fields/MD codes (for example: carbohydrates with AMBER/GLYCAM using AMBER MD instead of Gromacs).

- Materials sciences seem to be severely under-represented - commonly used codes in this area such as LAMMPS are not even detected, and only very few examples could be identified. As it is, the paper primarily provides an insight into the *biomolecular* MD simulation world.

The authors succeed in providing a first realistic view on what MD data is available in public repositories. In particular, their explore-expand approach has the potential to be customized for all kinds of specialist simulation data, whereby specific artifacts are used as fiducial markers instead of metadata. The more detailed analysis is limited to Gromacs simulations and primarily biomolecular simulations (even though MD is also widely used in other fields such as the materials sciences). This restricted view may simply be correlated with the user community of Gromacs and hopefully, follow-up studies from this work will shed more light on this shortcoming.

The study quantified the number of trajectories currently held in structured databases as ~10k vs ~30k in generalist repositories. To go beyond the proof-of-principle analysis it would be interesting to analyze the data in specialist repositories in the same way as the one in the generalist ones, especially as there are now efforts underway to create a database for MD simulations (Grant 'Molecular dynamics simulation for biology and chemistry research' to establish MDDB' DOI 10.3030/101094651). One should note that structured databases do not invalidate the approach pioneered in this work; if anything they are orthogonal to each other and both will likely play an important role in growing the usefulness of MD simulations in the future.

https://doi.org/10.7554/eLife.90061.2.sa1

Reviewer #3 (Public Review):

Molecular dynamics (MD) simulations nowadays are an essential element of structural biology investigations, complementing experiments and aiding their interpretation by revealing transient processes or details (such as the effects of glycosylation on the SARS-CoV-2 spike protein, for example (Casalino et al. ACS Cent. Sci. 2020; 6, 10, 1722-1734 https://doi.org/10.1021/acscentsci.0c01056) that cannot be observed directly. MD simulations can allow for the calculation of thermodynamic, kinetic, and other properties and the prediction of biological or chemical activity. MD simulations can now serve as "computational assays" (Huggins et al. WIREs Comput Mol Sci. 2019; 9:e1393. https://doi.org/10.1002/wcms.1393). Conceptually, MD simulations have played a crucial role in developing the understanding that the dynamics and conformational behaviour of biological macromolecules are essential to their function, and are shaped by evolution. Atomistic simulations range up to the billion atom scale with exascale resources (e.g. simulations of SARS-CoV-2 in a respiratory aerosol. Dommer et al. The International Journal of High Performance Computing Applications. 2023; 37:28-44. doi:10.1177/10943420221128233), while coarse-grained models allow simulations on even larger length- and timescales. Simulations with combined quantum mechanics/molecular mechanics (QM/MM) methods can investigate biochemical reactivity, and overcome limitations of empirical forcefields (Cui et al. J. Phys. Chem. B 2021; 125, 689 https://doi.org/10.1021/acs.jpcb.0c09898).

MD simulations generate large amounts of data (e.g. structures along the MD trajectory) and increasingly, e.g. because of funder mandates for open science, these data are deposited in publicly accessible repositories. There is real potential to learn from these data en masse, not only to understand biomolecular dynamics but also to explore methodological issues. Deposition of data is haphazard and lags far behind experimental structural biology, however, and it is also hard to answer the apparently simple question of "what is out there?". This is the question that Tiemann et al explore in this nice and important work, focusing on simulations run with the widely used GROMACS package. They develop a search strategy and identify almost 2,000 datasets from Zenodo, Figshare and Open Science Framework. This provides a very useful resource. For these datasets, they analyse features of the simulations (e.g. atomistic or coarse-grained), which provides a useful snapshot of current simulation approaches. The analysis is presented clearly and discussed insightfully. They also present a search engine to explore MD data, the MDverse data explorer, which promises to be a very useful tool.

As the authors state: "Eventually, front-end solutions such as the MDverse data explorer tool can evolve being more user-friendly by interfacing the structures and dynamics with interactive 3D molecular viewers". This will make MD simulations accessible to non-specialists and researchers in other areas. I would envisage that this will also include approaches using interactive virtual reality for an immersive exploration of structure and dynamics, and virtual collaboration (e.g. O'Connor et al., Sci. Adv.4, eaat2731 (2018). DOI:10.1126/sciadv.aat2731)

The need to share data effectively, and to compare simulations and test models, was illustrated clearly in the COVID-19 pandemic, which also demonstrated a willingness and commitment to data sharing across the international community (e.g. Amaro and Mulholland, J. Chem. Inf. Model. 2020, 60, 6, 2653-2656 https://doi.org/10.1021/acs.jcim

.0c00319; Computing in Science & Engineering 2020, 22, 30-36 doi: 10.1109/MCSE.2020.3024155). There are important lessons to learn here, for simulations to be reproducible and reliable, for rapid testing, for exploiting data with machine learning, and for linking to data from other approaches. Tiemann et al. discuss how to develop these links, providing good perspectives and suggestions.

I agree completely with the statement of the authors that "Even if MD data represents only 1 % of the total volume of data stored in Zenodo, we believe it is our responsibility, as a community, to develop a better sharing and reuse of MD simulation files - and it will neither have to be particularly cumbersome nor expensive. To this end, we are proposing two solutions. First, improve practices for sharing and depositing MD data in data repositories. Second, improve the FAIRness of already available MD data notably by improving the quality of the current metadata."

This nicely states the challenge to the biomolecular simulation community. There is a clear need for standards for MD data and associated metadata. This will also help with the development of standards of best practice in simulations. The authors provide useful and detailed recommendations for MD metadata. These recommendations should contribute to discussions on the development of standards by researchers, funders, and publishers. Community organizations (such as CCP-BioSim and HECBioSim in the UK, BioExcel, CECAM, MolSSI, learned societies etc) have an important part to play in these developments, which are vital for the future of biomolecular simulation.

https://doi.org/10.7554/eLife.90061.2.sa0

Author response:

The following is the authors' response to the original reviews.

> ***eLife assessment***
>
> *The study presents a valuable tool for searching molecular dynamics simulation data, making such data sets accessible for open science. The authors provide convincing evidence that it is possible to identify useful molecular dynamics simulation data sets and their analysis can produce valuable information.*
>
> ***Public Reviews***
>
> ***Reviewer #1 (Public Review):***
>
> *Summary:*
>
> *Tiemann et al. have undertaken an original study on the availability of molecular dynamics (MD) simulation datasets across the Internet. There is a widespread belief that extensive, well-curated MD datasets would enable the development of novel classes of AI models for structural biology. However, currently, there is no standard for sharing MD datasets. As generating MD datasets is energy-intensive, it is also important to facilitate the reuse of MD datasets to minimize energy consumption. Developing a universally accepted standard for depositing and curating MD datasets is a huge undertaking. The study by Tiemann et al. will be very valuable in informing policy developments toward this goal.*
>
> *Strengths:*
>
> *The study presents an original approach to addressing a growing concern in the field. It is clear that adopting a more collaborative approach could significantly enhance the*

impact of MD simulations in modern molecular sciences.

The timing of the work is appropriate, given the current interest in developing AI models for describing biomolecular dynamics.

Weaknesses:

The study primarily focuses on one major MD engine (GROMACS), although this limitation is not significant considering the proof-of-concept nature of the study.

We thank the reviewer for his/her comments. Moving forward, our plan includes expanding this research to encompass other MD engines used in biomolecular simulations and materials sciences, such as NAMD, Charmm, Amber, LAMMPS, etc. However, this requires parsing associated files to supplement the sparse metadata generally available for the related datasets

Reviewer #2 (Public Review):

Summary:

Molecular dynamics (MD) data is deposited in public, non-specialist repositories. This work starts from the premise that these data are a valuable resource as they could be used by other researchers to extract additional insights from these simulations; it could also potentially be used as training data for ML/AI approaches. The problem is that mining these data is difficult because they are not easy to find and work with. The primary goal of the authors was to discover and index these difficult-to-find MD datasets, which they call the "dark matter of the MD universe" (in contrast to data sets held in specialist databases).

The authors developed a search strategy that avoided the use of ill-defined metadata but instead relied on the knowledge of the restricted set of file formats used in MD simulations as a true marker for the data they were looking for. Detection of MD data marked a data set as relevant with a follow-up indexing strategy of all associated content. This "explore-and-expand" strategy allowed the authors for the first time to provide a realistic census of the MD data in non-specialist repositories.

As a proof of principle, they analyzed a subset of the data (primarily related to simulations with the popular Gromacs MD package) to summarize the types of simulated systems (primarily biomolecular systems) and commonly used simulation settings.

Based on their experience they propose best practices for metadata provision to make MD data FAIR (findable, accessible, interoperable, reusable).

A prototype search engine that works on the indexed datasets is made publicly available. All data and code are made freely available as open source/open data.

Strengths:

The novel search strategy is based on relevant data to identify full datasets instead of relying on metadata and thus is likely to have many true positives and few false positives.

The paper provides a first glimpse at the potential hidden treasures of MD simulations and force field parametrizations of molecules.

*Analysis of parameter settings of MD simulations from how researchers *actually* run simulations can provide valuable feedback to MD code developers for how to document/educate users. This approach is much better than analyzing what authors write in the Methods sections.*

> *The authors make a prototype search engine available.*
>
> *The guidelines for FAIR MD data are based on experience gained from trying to make sense of the data.*
>
> *Weaknesses:*
>
> *So far the work is a proof-of-concept that focuses on MD data produced by Gromacs (which was prevalent under all indexed and identified packages).*
>
> *As discussed in the manuscript, some types of biomolecules are likely underrepresented because different communities have different preferences for force fields/MD codes (for example: carbohydrates with AMBER/GLYCAM using AMBER MD instead of Gromacs).*
>
> *Materials sciences seem to be severely under-represented --- commonly used codes in this area such as LAMMPS are not even detected, and only very few examples could be identified. As it is, the paper primarily provides an insight into the *biomolecular* MD simulation world.*
>
> *The authors succeed in providing a first realistic view on what MD data is available in public repositories. In particular, their explore-expand approach has the potential to be customized for all kinds of specialist simulation data, whereby specific artifacts are used as fiducial markers instead of metadata. The more detailed analysis is limited to Gromacs simulations and primarily biomolecular simulations (even though MD is also widely used in other fields such as the materials sciences). This restricted view may simply be correlated with the user community of Gromacs and hopefully, follow-up studies from this work will shed more light on this shortcoming.*
>
> *The study quantified the number of trajectories currently held in structured databases as ~10k vs ~30k in generalist repositories. To go beyond the proof-of-principle analysis it would be interesting to analyze the data in specialist repositories in the same way as the one in the generalist ones, especially as there are now efforts underway to create a database for MD simulations (Grant 'Molecular dynamics simulation for biology and chemistry research' to establish MDDB' DOI 10.3030/101094651). One should note that structured databases do not invalidate the approach pioneered in this work; if anything they are orthogonal to each other and both will likely play an important role in growing the usefulness of MD simulations in the future.*

We thank the reviewer for his/her comments. As mentioned to Reviewer 1, we intend to extend this work to other MD engines in the near future to go beyond Gromacs and even biomolecular simulations. Furthermore, as the value of accessing and indexing specialized MD databases such as MDDB, MemprotMD, GPCRmd, NMRLipids, ATLAS, and others has been mentioned by the reviewer, it is indeed one of our next steps to continue to expand the MDverse catalog of MD data. This indexing may also extend the visibility and widespreaded adoptability of these specific databases.

> ***Reviewer #3 (Public Review):***
>
> *Molecular dynamics (MD) simulations nowadays are an essential element of structural biology investigations, complementing experiments and aiding their interpretation by revealing transient processes or details (such as the effects of glycosylation on the SARS-CoV-2 spike protein, for example (Casalino et al. ACS Cent. Sci. 2020; 6, 10, 1722-1734 https://doi.org/10.1021/acscentsci.0c01056) that cannot be observed directly. MD simulations can allow for the calculation of thermodynamic, kinetic, and other properties and the prediction of biological or chemical activity. MD simulations can now serve as "computational assays" (Huggins et al. WIREs Comput Mol Sci. 2019; 9:e1393.*

https://doi.org/10.1002/wcms.1393). Conceptually, MD simulations have played a crucial role in developing the understanding that the dynamics and conformational behaviour of biological macromolecules are essential to their function, and are shaped by evolution. Atomistic simulations range up to the billion atom scale with exascale resources (e.g. simulations of SARS-CoV-2 in a respiratory aerosol. Dommer et al. The International Journal of High Performance Computing Applications. 2023; 37:28-44. doi:10.1177/10943420221128233), while coarse-grained models allow simulations on even larger length- and timescales. Simulations with combined quantum mechanics/molecular mechanics (QM/MM) methods can investigate biochemical reactivity, and overcome limitations of empirical forcefields (Cui et al. J. Phys. Chem. B 2021; 125, 689 https://doi.org/10.1021/acs.jpcb.0c09898).

MD simulations generate large amounts of data (e.g. structures along the MD trajectory) and increasingly, e.g. because of funder mandates for open science, these data are deposited in publicly accessible repositories. There is real potential to learn from these data en masse, not only to understand biomolecular dynamics but also to explore methodological issues. Deposition of data is haphazard and lags far behind experimental structural biology, however, and it is also hard to answer the apparently simple question of "what is out there?". This is the question that Tiemann et al explore in this nice and important work, focusing on simulations run with the widely used GROMACS package. They develop a search strategy and identify almost 2,000 datasets from Zenodo, Figshare and Open Science Framework. This provides a very useful resource. For these datasets, they analyse features of the simulations (e.g. atomistic or coarse-grained), which provides a useful snapshot of current simulation approaches. The analysis is presented clearly and discussed insightfully. They also present a search engine to explore MD data, the MDverse data explorer, which promises to be a very useful tool.

As the authors state: "Eventually, front-end solutions such as the MDverse data explorer tool can evolve being more user-friendly by interfacing the structures and dynamics with interactive 3D molecular viewers". This will make MD simulations accessible to non-specialists and researchers in other areas. I would envisage that this will also include approaches using interactive virtual reality for an immersive exploration of structure and dynamics, and virtual collaboration (e.g. O'Connor et al., Sci. Adv.4, eaat2731 (2018). DOI:10.1126/sciadv.aat2731)

The need to share data effectively, and to compare simulations and test models, was illustrated clearly in the COVID-19 pandemic, which also demonstrated a willingness and commitment to data sharing across the international community (e.g. Amaro and Mulholland, J. Chem. Inf. Model. 2020, 60, 6, 2653-2656 https://doi.org/10.1021/acs.jcim.0c00319; Computing in Science & Engineering 2020, 22, 30-36 doi: 10.1109/MCSE.2020.3024155). There are important lessons to learn here, for simulations to be reproducible and reliable, for rapid testing, for exploiting data with machine learning, and for linking to data from other approaches. Tiemann et al. discuss how to develop these links, providing good perspectives and suggestions.

I agree completely with the statement of the authors that "Even if MD data represents only 1 % of the total volume of data stored in Zenodo, we believe it is our responsibility, as a community, to develop a better sharing and reuse of MD simulation files - and it will neither have to be particularly cumbersome nor expensive. To this end, we are proposing two solutions. First, improve practices for sharing and depositing MD data in data repositories. Second, improve the FAIRness of already available MD data notably by improving the quality of the current metadata."

This nicely states the challenge to the biomolecular simulation community. There is a clear need for standards for MD data and associated metadata. This will also help with

> *the development of standards of best practice in simulations. The authors provide useful and detailed recommendations for MD metadata. These recommendations should contribute to discussions on the development of standards by researchers, funders, and publishers. Community organizations (such as CCP-BioSim and HECBioSim in the UK, BioExcel, CECAM, MolSSI, learned societies etc) have an important part to play in these developments, which are vital for the future of biomolecular simulation.*

We thank the reviewer for his/her comments. Beyond the points mentioned to Reviewers 1 and 2, as the reviewer suggested, it would be of great interest to combine innovative and immersive approaches to visualize and possibly interact with the data collected. This is indeed more and more amenable thanks to technologies such as WebGL and programs such as Mol*, or even - as also pointed out by the reviewer - through virtual reality, for example with the mentioned Narupa framework or with the UnityMol software. For a comprehensive review on MD trajectory visualization and associated challenges, we refer to our recent review article https://doi.org/10.3389/fbinf.2024.1356659.

> ***Recommendations for the authors:***
>
> ***Reviewer #1 (Recommendations For The Authors):***
>
> *Some minor text editing would improve the readability of the manuscript.*
>
> *It would be very useful if the authors could share their perspectives on the best and most efficient approach to sharing datasets and code associated with a publication. My concern lies in the fact that Github, which is currently the dominant platform for sharing code, is not well-suited for hosting large MD datasets. As a result, researchers often need to adopt a workflow where code is shared on Github and datasets are stored elsewhere (e.g., Zenodo). While this is feasible, it adds extra work. Ideally, a transparent process could be developed to seamlessly share code and datasets linked to a study through a unified interface.*

We thank the reviewer for this excellent suggestion. To our knowledge, there is yet no easy framework to jointly store and share code and data, linked to their scientific publication. Of course, code can be submitted to “generic” databases along with the data, but at the current state, those do not provide such useful features like collaborative work & track recording as done to the extent of GitHub.

Although GitHub is indeed a suitable platform to deposit code, we strongly advise researchers to archive their code in Software Heritage. In addition to preserving source code, Software Heritage provides a unique identifier called SWHID that unambiguously makes reference to a specific version of the source code.

So far, it is the responsibility of the scientific publication authors to link datasets and source codes (whether in GitHub or Software Heritage) in their paper, but also to make the reverse link from the data and code sharing platforms to the paper after publication.

As mentioned by the reviewer, a unified interface that could ease this process would significantly contribute to FAIR-ness in MD.

> ***Reviewer #2 (Recommendations For The Authors):***
>
> *L180: I am not aware that TRR files contain energy terms as stated here, my understanding was that EDR files primarily served that purpose.*
>
> *“...available in one dataset. Interestingly, we found 1,406 .trr files, Which contain trajectory but also additional information such as velocities, energy of the system, etc’*

> *While the file is especially useful in terms of reusability, the large size (can go up to several 100GB) limits its deposition in most...”*

Indeed, our formulation was ambiguous. The EDR files contain the detailed information on energies, whereas TRR files contain numerous values from the trajectory such as coordinates, velocities, forces and to some extent also energies

(https://manual.gromacs.org/current/reference-manual/file-formats.html#trr)

> *L207: The text states that the total time was not available from XTC files, only the number of frames. However, XTC files record time stamps in addition to frame numbers. As long as these times are in the Gromacs standard of picoseconds, the simulation time ought to be available from XTCs.*
>
> *“...systems and the number of frames available in the files (Fig. 3-B). Of note, the frames do not directly translate to the simulation runtime - more information deposited in other files (e.g. .mdp files) is needed to determine the complete runtime of the simulation. The system was up...”.*

Thank you for the useful comment, we removed this sentence. We now mention that studying the simulation time would be of interest in the future, especially when we will perform an exhaustive analysis of XTC files.

“Of note, as .xtc files also contain time stamps, it would be interesting to study the relationship between the time and the number of frames to get useful information about the sampling. Nevertheless, this analysis would be possible only for unbiased MD simulations. So, we would need to decipher if the .xtc file is coming from biased or unbiased simulations, which may not be trivial.”

> *Analysis of MDP files: Were these standard equilibrium MD or can you distinguish biased MD or free energy calculations?*

Currently we do not distinguish between biased and unbiased MD, but in the future we may attempt to do so, e.g. by correlating it with standard equilibration force-fields/parameters, timesteps or similar. Nevertheless, a true distinction will remain challenging.

> *L336: typo: pikes -> spikes (or peaks?)*
>
> *“...simulations of Lennard-Jones models (Jeon et al., 2016). Interestingly, we noticed the appearance of several pikes at 400K, 600K and 800K, which were not present before the end of the year 2022. These peaks correspond to the same study related to the stability of hydrated crystals (Dybeck et al., 2023)’ Overall, thhis analysis revealed that a wide range of temperatures have been explored,...”*

Thank you. We have corrected this typo.

> *Make clear how multiple versions of data sets are handled, e.g., if v1, v2, and v3 of a dataset are provided in Zenodo then which one is counted or are all counted?*

We collected the latest version only of datasets, as exposed by default by the Zenodo API. To reflect this, we added the following sentence to the Methods and Materials section, Initial data collection sub-section:

“By default, the last version of the datasets was collected.”

> *L248 Analysis of GRO files seems fairly narrow because PDB files are very often used for exactly the same purpose, even in the context of Gromacs simulations, not the least because it is familiar to structural biologists that may be interested in representative MD snapshots. Despite all the shortcomings of abusing the PDB format for MD, it is an attempt at increased interoperability. Perhaps the authors can make sure that readers understand that choosing GRO for analysis may give a somewhat skewed picture, even within Gromacs simulations.*

Thanks for this comment. We collected about 12,000 PDB files that could indeed be output from Gromacs simulations and easily be shared due to the universality of this format, but that could as well come from different sources (like other MD packages or the PDB database itself). We purposely decided to limit our study to files strictly associated with the Gromacs package, like MDP and XTC file types. However, we will extend our survey to all other structure-like formats and especially the PDB file type. We reflected this purpose in the following sentence (after line 281)

"Beyond .gro files, we would like to analyze the ensemble of the ~12,000 .pdb files extracted in this study (see Figure 2-B) to better characterize the types of molecular structures deposited."

> *A simple template metadata file would be welcome (e.g., served from a GitHub/GitLab repository so that it can be improved with community input).*

Thank you for this suggestion that we fundamentally agree with. However, the generation of such a file is a major task, and we believe that the creation of a metadata file template requires far-reaching considerations, therefore is beyond the scope of this paper and should not be decided by a small group of researchers. Indeed, this topic requires a large consensus of different stakeholders, from users, to MD program developers, and journal editors. It would be especially useful to organize dedicated workshops with representatives of all these communities to tackle this specific issue, as mentioned by Reviewer3 in his/her public review. As a basis for this discussion, we humbly proposed at the end of this manuscript a few non-constraining guidelines based on our experience retrieving the data.

To emphasize this statement, we added the following sentence at the end of the "Guidelines for better sharing of MD simulation data" section (line 420):

"Converging on a set of metadata and format requires a large consensus of different stakeholders from users, to MD program developers, and journal editors. It would be especially useful to organize specific workshops with representatives of all these communities to collectively tackle this specific issue."

> *In "Data and code availability" it would be good to specify licenses in addition to stating "open source". Thank you for pointing out that GitLab/GitHub are not archives and that everyone should be strongly encouraged to submit data to stable archival repositories.*

We added the corresponding licenses for code and data in the "Data and code availability" section.

> ***Reviewer #3 (Recommendations For The Authors)***
>
> *The paper is well written, with very few typographical or other minor errors.*
>
> *Minor points:*

> *Line 468-9 "can evolve being more user-friendly" should be "can evolve to being more user-friendly", I think.*

Thank you, we have changed the wording accordingly.

https://doi.org/10.7554/eLife.90061.2.sa4