



**HAL**  
open science

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

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

► **To cite this version:**

Johanna K S Tiemann, Magdalena Szczuka, Lisa Bouarroudj, Mohamed Oussaren, Steven Garcia, et al.. MDverse: Shedding Light on the Dark Matter of Molecular Dynamics Simulations. 2023. hal-04254566v1

**HAL Id: hal-04254566**

**<https://hal.science/hal-04254566v1>**

Preprint submitted on 23 Oct 2023 (v1), last revised 14 Nov 2024 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# 1 MDverse: Shedding Light on the Dark 2 Matter of Molecular Dynamics 3 Simulations

4 Johanna K. S. Tiemann<sup>1\*,†</sup>, Magdalena Szczuka<sup>2</sup>, Lisa Bouarroudj<sup>3</sup>, Mohamed  
5 Oussaren<sup>3</sup>, Steven Garcia<sup>4</sup>, Rebecca J. Howard<sup>5</sup>, Lucie Delemotte<sup>6</sup>, Erik Lindahl<sup>5,6</sup>,  
6 Marc Baaden<sup>7</sup>, Kresten Lindorff-Larsen<sup>1</sup>, Matthieu Chavent<sup>2\*</sup>, Pierre Poulain<sup>3\*</sup>

**\*For correspondence:**

[johanna.tiemann@gmail.com](mailto:johanna.tiemann@gmail.com) (JKST);  
[Matthieu.Chavent@ipbs.fr](mailto:Matthieu.Chavent@ipbs.fr) (MC);  
[pierre.poulain@u-paris.fr](mailto:pierre.poulain@u-paris.fr) (PP)

**Present address:** <sup>†</sup>Novozymes  
A/S, 2800 Kgs. Lyngby, Denmark

7 <sup>1</sup>Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of  
8 Copenhagen, DK-2200 Copenhagen N, Denmark; <sup>2</sup>Institut de Pharmacologie et Biologie  
9 Structurale, CNRS, Université de Toulouse, 205 route de Narbonne, 31400, Toulouse,  
10 France; <sup>3</sup>Université Paris Cité, CNRS, Institut Jacques Monod, F-75013 Paris, France;  
11 <sup>4</sup>Independent researcher, Amsterdam, Netherlands; <sup>5</sup>Department of Biochemistry and  
12 Biophysics, Science for Life Laboratory, Stockholm University, Stockholm, Sweden;  
13 <sup>6</sup>Department of Applied Physics, Science for Life Laboratory, KTH Royal Institute of  
14 Technology, Stockholm, Sweden; <sup>7</sup>Laboratoire de Biochimie Théorique, CNRS,  
15 Université Paris Cité, 13 rue Pierre et Marie Curie, F-75005 Paris, France

16

## 17 Abstract

18 The rise of open science and the absence of a global dedicated data repository for molecular  
19 dynamics (MD) simulations has led to the accumulation of MD files in generalist data repositories,  
20 constituting the *dark matter of MD* - data that is technically accessible, but neither indexed,  
21 curated, or easily searchable. Leveraging an original search strategy, we found and indexed about  
22 250,000 files and 2,000 datasets from Zenodo, Figshare and Open Science Framework. With a  
23 focus on files produced by the Gromacs MD software, we illustrate the potential offered by the  
24 mining of publicly available MD data. We identified systems with specific molecular composition  
25 and were able to characterize essential parameters of MD simulation, such as temperature and  
26 simulation length, and identify model resolution, such as all-atom and coarse-grain. Based on  
27 this analysis, we inferred metadata to propose a search engine prototype to explore collected MD  
28 data. To continue in this direction, we call on the community to pursue the effort of sharing MD  
29 data, and increase populating and standardizing metadata to reuse this valuable matter.

30

## 31 Introduction

32 The volume of data available in biology has increased tremendously (*Marx, 2013; Stephens et al.,*  
33 *2015*), through the emergence of high-throughput experimental technologies, often referred to as -  
34 omics, and the development of efficient computational techniques, associated with high-performance  
35 computing resources. The Open Access (OA) movement to make research results free and avail-  
36 able to anyone (including e.g. the Budapest Open Access Initiative and the Berlin declaration on  
37 Open Access to Knowledge) has led to an explosive growth of research data made available by  
38 scientists (*Wilson et al., 2021*). The FAIR (Findable, Accessible, Interoperable and Reusable) princi-  
39 ples (*Wilkinson et al., 2016*) have emerged to structure the sharing of these data with the goals  
40 of reusing research data and to contribute to the scientific reproducibility. This leads to a world

41 where research data has become widely available and exploitable, and consequently new applica-  
42 tions based on artificial intelligence (AI) emerged. One example is AlphaFold (*Jumper et al., 2021*),  
43 which enables the construction of a structural model of any protein from its sequence. However,  
44 it is important to be aware that the development of AlphaFold was only possible because of the  
45 existence of extremely well annotated and cleaned open databases of protein structures (wwPDB  
46 *Berman et al. (2003)*) and sequences (UniProt *Consortium (2022)*). Similarly, accurate predictions  
47 of NMR chemical shifts and chemical-shift-driven structure determination was only made possible  
48 via a community-driven collection of NMR data in the Biological Magnetic Resonance Data Bank  
49 (*Hoch et al., 2023*). One can easily imagine novel possibilities of AI and deep learning reusing pre-  
50 vious research data in other fields, if that data is curated and made available at a large scale (*Fan*  
51 *and Shi, 2022; Mahmud et al., 2021*).

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

77 Albeit previous attempts in the past (*Tai et al., 2004; Meyer et al., 2010*), there is, as of now,  
78 no central data repository that could host all kinds of MD simulation files. This is not only due to  
79 the huge volume of data and its heterogeneity, but also because interoperability of the many file  
80 formats used adds to the complexity. Thus, faced with the deluge of biosimulation data (*Hospital*  
81 *et al., 2020*), researchers often share their simulation files in multiple generalist data repositories.  
82 This makes it difficult to search and find available data on, for example, a specific protein or a  
83 given set of parameters. We are qualifying this amount of scattered data as *the dark matter of*  
84 *MD*, and we believe it is essential to shed light onto this overlooked but high-potential volume of  
85 data. When unlocked, publicly available MD files will gain more visibility. This will help people to  
86 access and reuse these data more easily and overall, by making MD simulation data more FAIR  
87 (*Wilkinson et al., 2016*), it will also improve the reproducibility of MD simulations (*Elofsson et al.,*  
88 *2019; Porubsky et al., 2020; consortium, 2019*).

89 In this work, we have employed a search strategy to index scattered MD simulation files de-  
90 posited in generalist data repositories. With a focus on the files generated by the Gromacs MD  
91 software, we performed a proof-of-concept large-scale analysis of publicly available MD data. We

92 revealed the high value of these data and highlighted the different categories of the simulated  
93 molecules, as well as the biophysical conditions applied to these systems. Based on these results  
94 and our annotations, we proposed a search engine prototype to easily explore this *dark matter of*  
95 *MD*. Finally, building on this experience, we provide simple guidelines for data sharing to gradually  
96 improve the FAIRness of MD data.

## 97 Results

98 With the rise of open science, researchers increasingly share their data and deposit them into gen-  
99 eralist data repositories, such as Zenodo (<https://zenodo.org>), Figshare (<https://figshare.com>), Open  
100 Science Framework (OSF, <https://osf.io>), and Dryad (<https://datadryad.org/>). In this first attempt to  
101 find out how many files related to MD are deposited in data repositories, we focused our explo-  
102 ration on three major data repositories: Figshare (~3.3 million files, ~112 TB of data, as of January  
103 2023), OSF (~2 million files, as of November 2022)<sup>1</sup>, and Zenodo (~9.9 million files, ~1.3 PB of data,  
104 as of December 2022; *Panero and Benito (2022)*).

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

116 To circumvent this issue, we developed an original and specific search strategy that we called  
117 *Explore and Expand* ( $E_x^2$ ) (see Fig. 1-A and Materials and Methods section) and that relies on a com-  
118 bination of file types and keywords queries. In the *Explore* phase, we searched for files based on  
119 their file types (for instance: .xtc, .gro, etc) with MD-related keywords (for instance: 'molecular dy-  
120 namics', 'Gromacs', 'Martini', etc). Each of these hit files belonged to a dataset, which we further  
121 screened in the *Expand* phase. There, we indexed all files found in a dataset identified in the previ-  
122 ous *Explore* phase with, this time, no restriction to the collected file types (see Fig. 1-A and details  
123 on the data scraping procedure in the Materials and Methods section).

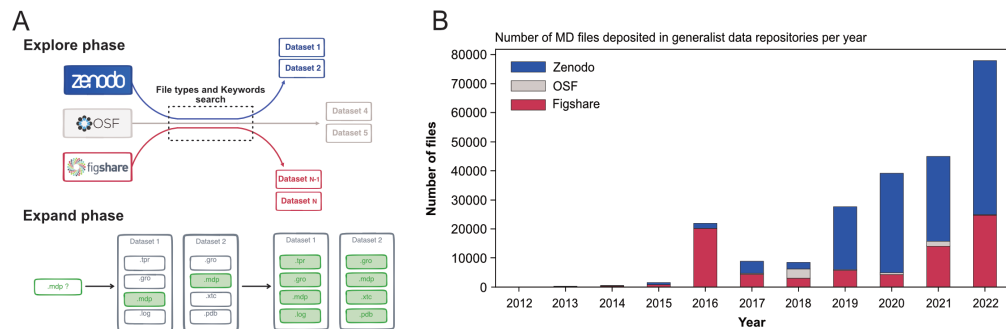
124 Globally, we indexed about 250,000 files and 2,000 datasets that represented 14 TB of data de-  
125 posited between August 2012 and March 2023 (see Table 1). One major difficulty were the numer-  
126 ous files stored in zipped archives, about seven times more than files steadily available in datasets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

135 The first dataset we found related to MD data that has been deposited in August 2012 in  
136 Figshare and corresponds to the work of Fuller et al. (*Fuller et al., 2012*) (see Table 1) but we  
137 may consider the start of more substantial deposition of the MD data to be 2016 with more than  
138 20,000 files deposited, mainly in Figshare (see Fig. 1-B). While the number of files deposited in Zen-  
139 odo was first relatively limited, the last few years (2020-2022) saw a steep increase, passing from

<sup>1</sup>Figures provided by Figshare and OSF user support teams

**Table 1.** Statistics of the MD-related datasets and files found in the data repositories Figshare, OSF, and Zenodo.

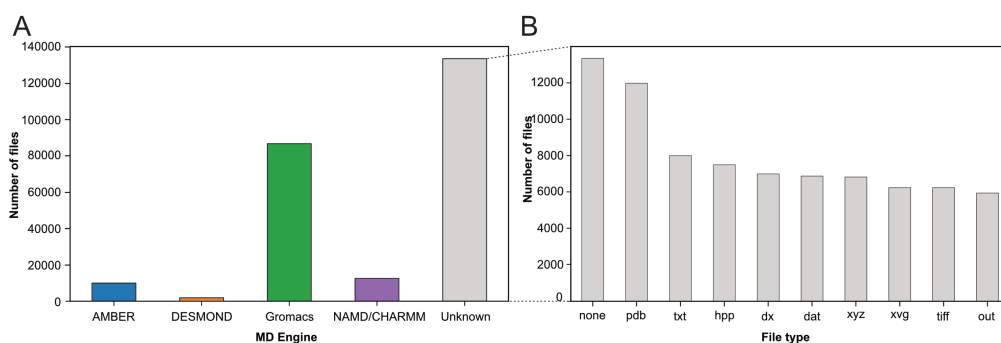
Data repository	datasets	first dataset	latest dataset	files	total size (GB)	zip files	files within zip	total files
Zenodo	1,011	19/11/2014	05/03/2023	20,250	12,851	1,780	141,304	161,554
Figshare	913	20/08/2012	03/03/2023	3,336	736	590	74,720	78,056
OSF	55	24/05/2017	05/02/2023	6,146	495	14	0	6,146
<b>Total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



**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.

140 a few thousands files in 2018 to almost 50,000 files in 2022 (see Fig. 1-B). In 2018, the number of  
 141 MD files deposited in OSF was similar to those in the two other data repositories, but did not take  
 142 off as much as the other data repositories. Zenodo seems to be favored by the MD community  
 143 since 2019, even though Figshare in 2022 also saw a sharp increase in deposited MD files. The  
 144 preference for Zenodo could also be explained by the fact that it is a publicly funded repository de-  
 145 veloped under the European OpenAIRE program and operated by CERN (*European Organization*  
 146 *For Nuclear Research and OpenAIRE, 2013*). Overall, the trend showed a rise of deposited data with  
 147 a steep increase in 2022 (Fig. 1-B). We believe that this trend will continue in future years, which  
 148 will lead to a greater amount of MD data available. It is thus urgent to deploy a strategy to index  
 149 this vast amount of data, and to allow the MD community to easily explore and reuse such gigantic  
 150 resource. The following describes what is already feasible in terms of meta analysis, in particular  
 151 what types of data are deposited in data repositories and the simulation setup parameters used  
 152 by MD experts that have deposited their data.

153 With our  $Ex^2$  strategy (see Fig. 1-A), we assigned the deposited files to the MD packages: AMBER  
 154 (*Ferrer et al., 2012*), DESMOND (*Bowers et al., 2006*), Gromacs (*Berendsen et al., 1995; Abraham*  
 155 *et al., 2015*), and NAMD/CHARMM (*Phillips et al., 2020; Brooks et al., 2009*), based on their corre-  
 156 sponding file types (see Materials and Methods section). In the case of NAMD/CHARMM, file exten-  
 157 sions were mostly identical, which prevented us from distinguishing the respective files from these  
 158 two MD programs. With 87,204 files deposited, the Gromacs program was most represented (see  
 159 Fig. 2-A), followed by NAMD/CHARMM, AMBER, and DESMOND. This statistic is limited as it does not  
 160 consider more specific databases related to a particular MD program. For example, the DE Shaw  
 161 Research website contains a large amount of simulation data related to SARS-CoV-2 that has been  
 162 generated using the ANTON supercomputer ([https://www.deshawresearch.com/downloads/download\\_](https://www.deshawresearch.com/downloads/download_trajectory_sarscov2.cgi/)  
 163 [trajectory\\_sarscov2.cgi/](https://www.deshawresearch.com/downloads/download_trajectory_sarscov2.cgi/)) or other extensively simulated systems of interest to the community. How-  
 164 ever, this in itself might also serve as a good example, since few automated search strategies will  
 165 be able to find custom stand-alone web servers as valuable repositories. Here, our goal was not



**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.

166 to compare the availability of all data related to each MD program but to give a snapshot of the  
167 type of data available at a given time (*i.e.* March 2023) in generalist data repositories. Interest-  
168 ingly, many files (> 133,000) were not directly associated to any MD program (see Fig. 2-A label  
169 'Unknown'). We categorized these files based on their extensions (see Fig. 2-B). While 10 % of these  
170 files were without file extension (Fig. 2-B, column *none*), we found numerous files corresponding  
171 to structure coordinates such as .pdb (~12,000) and .xyz (~6,800) files. We also got images (.tiff  
172 files) and graphics (.xvg files). Finally, we found many text files such as .txt, .dat, and .out which can  
173 potentially hold details about how simulations were performed. Focusing further on files related  
174 to the Gromacs program, being currently most represented in the studied data repositories, we  
175 demonstrated in the following present possibilities to retrieve numerous information related to  
176 deposited MD simulations.

177 First, we were interested in what file types researchers deposited and thereby find potentially  
178 of great value to share. We therefore quantified the types of files generated by Gromacs (Fig. 3-A).  
179 The most represented file type is the .xtc file (28,559 files, representing 8.6 TB). This compressed  
180 (binary) file is used to store the trajectory of an MD simulation and is an important source of in-  
181 formation to characterize the evolution of the simulated molecular system as a function of time.  
182 It is thus logical to mainly find this type of file shared in data repositories, as it is of great value  
183 for reuse and new analyses. Nevertheless, it is not directly readable but needs to be read by a  
184 third-party program, such as Gromacs itself, a molecular viewer like VMD (*Humphrey et al., 1996*)  
185 or an analysis library such as MDAnalysis (*Michaud-Agrawal et al., 2011; Gowers et al., 2016*). In  
186 addition, this trajectory file can only be of use in combination with a matching coordinates file, in  
187 order to correctly access the dynamics information stored in this file. Thus, as it is, this file is not  
188 easily mineable to extract useful information, especially if multiple .xtc and coordinate files are  
189 available in one dataset. Interestingly, we found 1,406 .trr files, which contain trajectory but also  
190 additional information such as velocities, energy of the system, etc. While this file is especially use-  
191 ful in terms of reusability, the large size (can go up to several 100 GB) limits its deposition in most  
192 data repositories. For instance, a file cannot usually exceed 50 GB in Zenodo, 20 GB in Figshare  
193 (for free accounts) and 5 GB in OSF. Altogether, Gromacs trajectory files represented about 30,000  
194 files in the three explored generalist repositories (34% of Gromacs files). This is a large number  
195 in comparison to existing trajectories stored in known databases dedicated to MD with 1,700 MD  
196 trajectories available in MoDEL, 1,737 trajectories (as of November 2022) available in GPCRmd,  
197 5,971 (as of January 2022) trajectories available in MemProtMD and 726 trajectories (as of March  
198 2023) available in the NMRlipids Databank. Although fewer in count, these numbers correspond to  
199 manually or semi-automatically curated trajectories of specific systems, mostly proteins and lipids.  
200 Thus, ~30,000 MD trajectories available in generalist data repositories may represent a wider spec-  
201 trum of simulated systems but need to be further analyzed and filtered to separate usable data



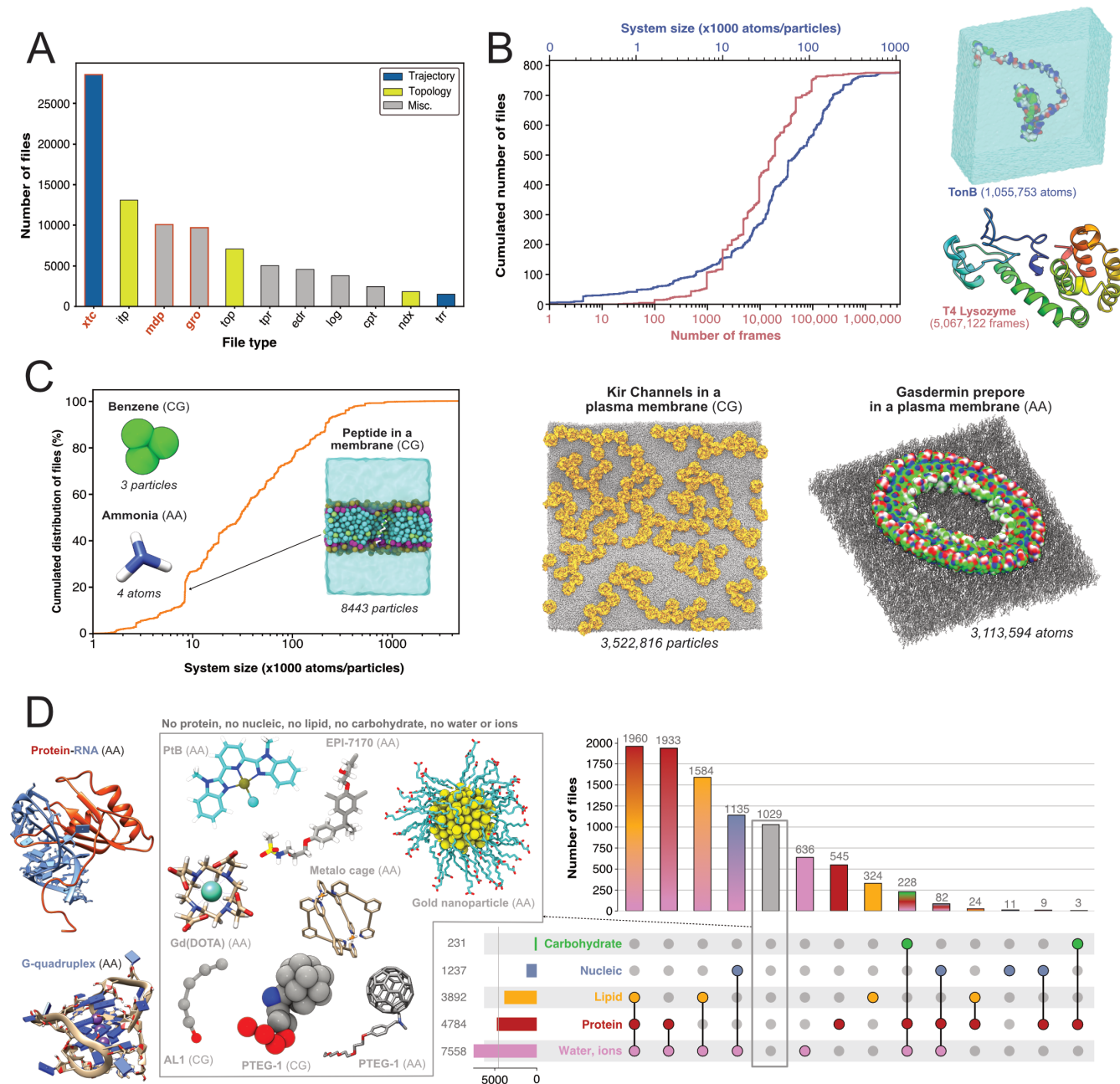
202 from less interesting trajectories such as minimization or equilibration runs.

203 Given the large volume of data represented by .xtc files (see above), we could only scratch the  
204 surface of the information stored in these trajectory files by analyzing a subset of 779 .xtc files -  
205 one per dataset in which this type of file was found. We were able to get the size of the molecular  
206 systems and the number of frames available in these files (Fig. 3-B). Of note, the frames do not  
207 directly translate to the simulation runtime - more information deposited in other files (e.g. .mdp  
208 files) is needed to determine the complete runtime of the simulation. The system size was up  
209 to more than one million atoms for a simulation of the TonB protein (*Virtanen et al., 2020*). The  
210 cumulative distribution of the number of frames showed that half of the files contain more than  
211 10,000 frames. This conformational sampling can be very useful for other research fields besides  
212 the MD community that study, for instance, protein flexibility or protein engineering where diverse  
213 backbones can be of value. We found an .xtc file containing more than 5 million frames, where the  
214 authors probe the picosecond–nanosecond dynamics of T4 lysozyme and guide the MD simulation  
215 with NMR relaxation data (*Kümmerer et al., 2021*). Extending this analysis to all 28,559 .xtc files  
216 detected would be of great interest for a more holistic view, but this would require an initial step  
217 of careful checking and cleaning to be sure that these files are analyzable.

218 These results bring a first explanation on why there is not a single special-purpose repository  
219 for MD trajectory files. Databases dedicated to molecular structures such as the Protein Databank  
220 (*Berman et al., 2000; Kinjo et al., 2017; Armstrong et al., 2019*), or even the recent PDB-dev (*Bur-*  
221 *ley et al., 2017*), designed for integrative models, cannot accept such large-size files, even less if  
222 complete trajectories without reducing the number of frames would be uploaded. This would also  
223 require implementing extra steps of data curation and quality control. In addition, the size of the  
224 IT infrastructure and the human skills required for data curation represents a significant cost that  
225 could probably not be supported by a single institution.

226 Subsequently, our interest shifted towards exploring which systems are being investigated by  
227 MD researchers who deposit their files. We found 9,718 .gro files which are text files that contain  
228 the number of particles and the Cartesian coordinates of the system modelled. By parsing the  
229 number of particles and the type of residue, we were able to give an overview of all Gromacs sys-  
230 tems deposited (Fig. 3-C,D). In terms of system size, they ranged from very small - starting with  
231 two coarse-grain (CG) particles of graphite (*Piskorz et al., 2019*), followed by coordinates of a water  
232 molecule (3 atoms) (*Ivanov et al., 2017*), CG model of benzene (3 particles) (*Dandekar and Mondal,*  
233 *2020*) and atomistic model of ammonia (4 atoms) (*Kelly and Smith, 2020*) — to go up to atomistic  
234 and coarse-grain systems composed of more than 3 million particles (*Duncan et al., 2020; Schae-*  
235 *fer and Hummer, 2022*) (Fig. 3-C). Interestingly, the system sizes in .gro files exceeded those of the  
236 analyzed .xtc files (Fig. 3-B). Even if we cannot exclude that the limited number of .xtc files analyzed  
237 (779 .xtc files selected from 28,559 .xtc files indexed) could explain this discrepancy, an alternate  
238 hypothesis is that the size of an .xtc file also depends on the number of frames stored. To reduce  
239 the size of .xtc files deposited in data repositories, besides removing some frames, researchers  
240 might also remove parts of the system, such as water molecules. As a consequence for reusability,  
241 this solvent removal could limit the number of suitable datasets available for researchers inter-  
242 ested in re-analysing the simulation with respect to, in this case, water diffusion. While the size of  
243 systems extracted from .gro files was homogeneously spread, we observed a clear bump around  
244 system sizes of circa 8,500 atoms/particles. This enrichment of data could be explained by the de-  
245 position of ~340 .gro files related to the simulation of a peptide translocation through a membrane  
246 (Fig. 3-C) (*Kabelka et al., 2021*). Beyond 1 million particles/atoms, the number of systems is, for the  
247 moment, very limited.

248 We then analyzed residues in .gro files and inferred different types of molecular systems (see  
249 Fig. 3-D). Two of the most represented systems contained lipid molecules. This may be related to  
250 NMRlipids initiative (<http://nmrlipids.blogspot.com>). For several years, this consortium has been ac-  
251 tively working on lipid modelling with a strong policy of data sharing and has contributed to share  
252 numerous datasets of membrane systems. As illustrated in Fig 3-C, a variety of membrane sys-



**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.



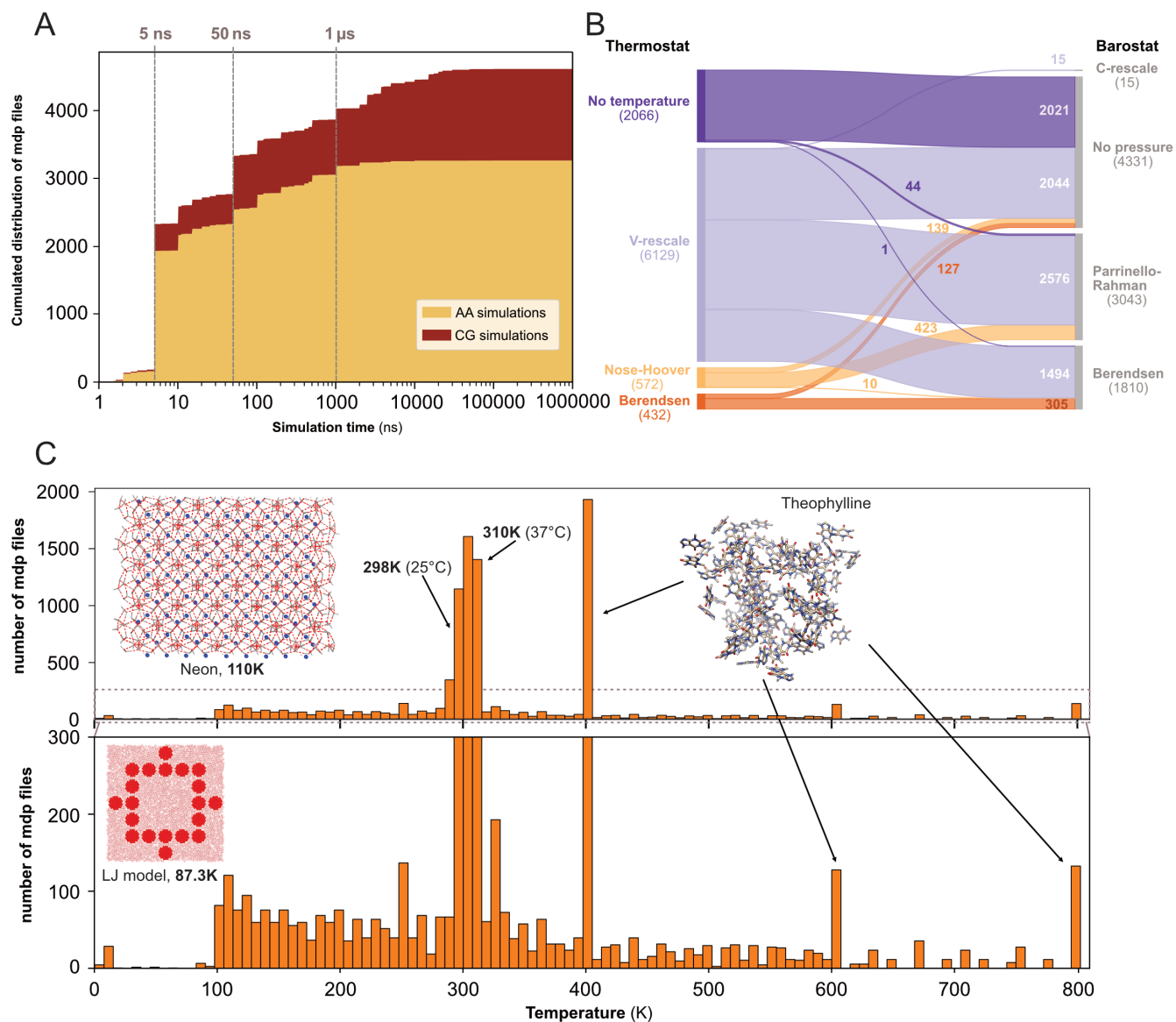
tems, especially membrane proteins, were deposited. This highlights the vitality of this research field, and the will of this community to share their data. We also found numerous systems containing solvated proteins. This type of data, combined with .xtc trajectory files (see above), could be invaluable 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.

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 Mackereel, 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 re-run 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



304 found in .mdp files: the number of steps and the time step. Here, we acknowledge that one can  
305 set up a very long simulation time and stop the simulation before the end or, on contrary, use  
306 a limited time (especially when calculations are performed on HPC resources with wall-time) and  
307 then extend the simulation for a longer duration. Using only the .mdp file, we cannot know if the  
308 simulation reached its term. To do so, comparison with an .xtc file from the same dataset may  
309 help to answer this specific question. However, in this study, we were interested in MD setup  
310 practices, in particular what simulation time researchers would set up their system with - likely in  
311 the mindset to reach that ending time. We restricted this analysis to the 4,623 .mdp files that used  
312 the *md* or *sd* integrator, and that have a simulation time above 1 ns. We found that the majority of  
313 the .mdp files were used for simulations of 50 ns or less (see Fig. 4-A). Further, 697 .mdp files with  
314 simulations times set-up between 50 ns and 1  $\mu$ s and 585 .mdp files with simulation time above  
315 1  $\mu$ s were identified. As analyzing .gro files showed a good proportion of coarse-grained models  
316 (Fig. 3-B,C), we discriminated simulations setups for these two types of models using the time step  
317 as a simple cutoff. We considered that a time step greater than 10 fs (*i.e.* dt=0.01) corresponded  
318 to MD setups for coarse grained models (*Ingólfsson et al., 2014*). Globally, we found that over all  
319 simulations, the setups for atomistic simulations were largely dominant. However, for simulations  
320 with a simulation time above 1  $\mu$ s specifically, coarse-grain simulations represented 86 % of all.

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

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

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

### 349 **Towards a better sharing of MD data**

350 With this work, we have shown that it was possible to not only retrieve MD data from the gen-  
351 eralist data repositories Zenodo, Figshare and OSF, but to shed light onto the *dark matter of MD*  
352 data in terms of learning current scientific practice, extracting valuable topology information, and  
353 analysing how the field is developing. Our objective was not to assess the quality of the data but

354 only to show what kind of data was available. The  $Ex^2$  strategy to find files related to MD sim-  
355 ulations relied on the fact that many MD software output files with specific file extensions. This  
356 strategy could not be applied in research fields where data exhibits non-specific file types. We  
357 experienced this limitation while indexing zip archives related to MD simulations, where we were  
358 able to decide if a zip archive was pertinent for this work only by accessing the list of files contained  
359 in the archive. This valuable feature is provided by data repositories like Zenodo and Figshare, with  
360 some caveats, though.

361 As of March 2023, we managed to index 245,756 files from 1,979 datasets, representing alto-  
362 gether 14 TB of data. This is a fraction of all files stored in data repositories. For instance, as of  
363 December 2022, Zenodo hosted about 9.9 million files for  $\sim 1.3$  PB of data (*Panero and Benito,*  
364 *2022*). All these files are stored on servers available 24/7. This high availability costs human re-  
365 sources, IT infrastructures and energy. Even if MD data represents only 1 % of the total volume  
366 of data stored in Zenodo, we believe it is our responsibility, as a community, to develop a better  
367 sharing and reuse of MD simulation files - and it will neither have to be particularly cumbersome  
368 nor expensive. To this end, we are proposing two solutions. First, improve practices for sharing  
369 and depositing MD data in data repositories. Second, improve the FAIRness of already available  
370 MD data notably by improving the quality of the current metadata.

### 371 **Guidelines for better sharing of MD simulation data**

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

- 375 • Avoid zip or tar archives whose content cannot be properly indexed by data repositories. As  
376 much as possible, deposit original data files directly.
- 377 • Describe the MD dataset with extensive metadata. Provide adequate information along your  
378 dataset, such as:
  - 379 - The scope of the study, e.g. investigate conformation dynamics, benchmark force field,  
380 ...
  - 381 - The method on a basic (e.g. quantum mechanics, all-atom, coarse-grain) or advanced  
382 (accelerated, metadynamics, well-tempered) level.
  - 383 - The MD software: name, version (tag) and whether modifications have been made.
  - 384 - The simulation settings (for each of the steps, including minimization, equilibration and  
385 production): temperature(s), thermostat, barostat, time step, total runtime (simulation  
386 length), force field, additional force field parameters.
  - 387 - The composition of the system, with the precise names of the molecules and their num-  
388 bers, if possible also PDB, UniProt or Ensemble identifiers and whether the default struc-  
389 ture has been modified.
  - 390 - Give information about any post-processing of the uploaded files (e.g. truncation or  
391 stripping of the trajectory), including before and after values of what has been modified  
392 e.g. number of frames or number of atoms of uploaded files
  - 393 - Highlight especially valuable data, e.g. excessively QM-based parameterized molecules,  
394 and their parameter files.

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

- 398 • Link the MD dataset to other associated resources, such as:
  - 399 - The research article (if any) for which these data have been produced. Datasets are  
400 usually mentioned in the research articles, but rarely the other way around, since the  
401 deposition has to be done prior to publication. However, it is eminently possible to

**A**

**MDverse data explorer**

gro files quick search

Enter search term. For instance: Covid, POPC, Gromacs, CHARMM36, 1402417  Add filter

9780 elements found

Show 20 entries

Index	Dataset	ID	Title	Creation date	Authors	Description	File name	Atom number	Protein	Lipid
1	zenodo	1468560	C36 POPC simulation with 17 lipids per leaflet, 300K	2018-10-22	Hanne ...	POPC bilayer with 30 waters per lipid (17+17), at 300K, si...	whole17.gro	7616	false	true
2	zenodo	6526243	Amyloid-beta 16-22 peptide dimer simulation (150nM Na...	2022-05-07	Kav, Ba...	Amyloid-beta 16-22 peptide dimer simulation with the CH...	prod.gro	32020	true	false
3	zenodo	838641	Large DPPC monolayer simulations with Charmm36+OPC ...	2017-08-03	Javanai...	DPPC monolayers simulated at a varying area per lipid in t...	DPPC-31...	232488	false	true
4	zenodo	838641	Large DPPC monolayer simulations with Charmm36+OPC ...	2017-08-03	Javanai...	DPPC monolayers simulated at a varying area per lipid in t...	DPPC-31...	232488	false	true
5	zenodo	838641	Large DPPC monolayer simulations with Charmm36+OPC ...	2017-08-03	Javanai...	DPPC monolayers simulated at a varying area per lipid in t...	DPPC-31...	232488	false	true
6	zenodo	838641	Large DPPC monolayer simulations with Charmm36+OPC ...	2017-08-03	Javanai...	DPPC monolayers simulated at a varying area per lipid in t...	DPPC-31...	232488	false	true
7	zenodo	838641	Large DPPC monolayer simulations with Charmm36+OPC ...	2017-08-03	Javanai...	DPPC monolayers simulated at a varying area per lipid in t...	DPPC-31...	232488	false	true
8	zenodo	838641	Large DPPC monolayer simulations with Charmm36+OPC ...	2017-08-03	Javanai...	DPPC monolayers simulated at a varying area per lipid in t...	DPPC-31...	232488	false	true
9	zenodo	838641	Large DPPC monolayer simulations with Charmm36+OPC ...	2017-08-03	Javanai...	DPPC monolayers simulated at a varying area per lipid in t...	DPPC-31...	232488	false	true
10	zenodo	838641	Large DPPC monolayer simulations with Charmm36+OPC ...	2017-08-03	Javanai...	DPPC monolayers simulated at a varying area per lipid in t...	DPPC-31...	232488	false	true

**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	<a href="https://zenodo.org/record/1043926">https://zenodo.org/record/1043926</a>
1	zenodo	1043946	mono.mdp	0.002	100000000.0	298.0	v-rescale	no	<a href="https://zenodo.org/record/1043946">https://zenodo.org/record/1043946</a>
2	zenodo	3463130	md.mdp	0.020	25000000.0	310.0	v-rescale	parrinello-rahman	<a href="https://zenodo.org/record/3463130">https://zenodo.org/record/3463130</a>
3	zenodo	1167532	md.mdp	0.002	100000000.0	298.0	nose-hoover	parrinello-rahman	<a href="https://zenodo.org/record/1167532">https://zenodo.org/record/1167532</a>
4	zenodo	3434100	1-POPC512_ECC-lipid14-CaCL_978mM_md_moutout.mdp	0.002	NaN	313.0	v-rescale	parrinello-rahman	<a href="https://zenodo.org/record/3434100">https://zenodo.org/record/3434100</a>

**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	<a href="https://zenodo.org/record/7125315">https://zenodo.org/record/7125315</a>	5.583
1707	zenodo	7125315	3c_START.gro	5844.0	False	False	False	False	False	<a href="https://zenodo.org/record/7125315">https://zenodo.org/record/7125315</a>	5.844
1708	zenodo	7125315	4a_START.gro	4374.0	False	False	False	False	False	<a href="https://zenodo.org/record/7125315">https://zenodo.org/record/7125315</a>	4.374
1709	zenodo	7125315	4b_START.gro	4410.0	False	False	False	False	False	<a href="https://zenodo.org/record/7125315">https://zenodo.org/record/7125315</a>	4.410
1710	zenodo	7125315	DMF.gro	12.0	False	False	False	False	False	<a href="https://zenodo.org/record/7125315">https://zenodo.org/record/7125315</a>	0.012
...	...	...	...	...	...	...	...	...	...	...	...
8666	osf	4aghb	PtB-b-force field/em4_nojump.gro	18432.0	False	False	False	False	False	<a href="https://osf.io/4aghb/">https://osf.io/4aghb/</a>	18.432
8667	osf	4aghb	PtB-b-force field/pr_nvt.gro	18432.0	False	False	False	False	False	<a href="https://osf.io/4aghb/">https://osf.io/4aghb/</a>	18.432
8668	osf	4aghb	PtB-b-force field/pr_nvt_nojump.gro	18432.0	False	False	False	False	False	<a href="https://osf.io/4aghb/">https://osf.io/4aghb/</a>	18.432
8669	osf	4aghb	PtB-b-force field/Production_10ns_PTI_noVs.gro	18432.0	False	False	False	False	False	<a href="https://osf.io/4aghb/">https://osf.io/4aghb/</a>	18.432
8670	osf	4aghb	Pt-free-force field/MOL_GMX.gro	61.0	False	False	False	False	False	<a href="https://osf.io/4aghb/">https://osf.io/4aghb/</a>	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.



- 402 submit a revised version, and providing a link to the related research paper in updated  
403 metadata of the MD dataset will ease the reference to the original publication upon data  
404 reuse.
- 405 – The code used to analyze the data, ideally deposited in the repository to guarantee avail-  
406 ability, or in a GitHub or GitLab repository.
  - 407 – Any other datasets that belong to the same study.
- 408 • Provide sufficient files to reproduce simulations and use a clear naming convention to make  
409 explicit links between related files. For instance, for the Gromacs MD engine, trajectory .xtc  
410 files could share the same names as structure .gro files (e.g. proteinA.gro & proteinA.xtc).
  - 411 • Revisit your data deposition after paper acceptance and update information if necessary. Zen-  
412 do and Figshare provide a DOI for every new version of a dataset as well as a 'master' DOI  
413 that always refers to the latest version available.

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

### 421 **Improving metadata of current MD data**

422 While indexing about 2,000 MD datasets, we found that title and description accompanying these  
423 datasets were very heterogeneous in terms of quality and quantity and were difficult for machines  
424 to process automatically. It was sometimes impossible to find even basic information such as the  
425 identity of the molecular system simulated, the temperature or the length of the simulation. With-  
426 out appropriate metadata, sharing data is pointless, and its reuse is doomed to fail (*Musen*, 2022).  
427 It is thus important to close the gap between the availability of MD data and its discoverability and  
428 description through appropriate metadata. We could gradually improve the metadata by following  
429 two strategies. First, since MD engines produce normalized and well-documented files, we could  
430 extract parameters of the simulation by parsing specific files. We already explored this path with  
431 Gromacs, by extracting the molecular size and composition from .gro files and the simulation time  
432 (with some limitations), thermostat and barostat from .mdp files. We could go even further, by  
433 extracting for instance Gromacs version from .log file (if provided) or by identifying the simulated  
434 system from its atomic topology stored in .gro files. This strategy can in principle be applied to  
435 files produced by other MD engines. A second approach that we are currently exploring uses data  
436 mining and named entity recognition (NER) methods (*Perera et al.*, 2020) to automatically identify  
437 the molecular system, the temperature, and the simulation length from existing textual metadata  
438 (dataset title and description), providing they are of sufficient length. Finally, the possibilities af-  
439 forded by large language models supplemented by domain-specific tools (*Bran et al.*, 2023) might  
440 help interpret the heterogenous metadata that is often associated with the simulations.

### 441 **Future works**

442 In the future, it is desirable to go further in terms of analysis and integrate other data repositories,  
443 such as Dryad and Dataverse instances (for example [Recherche Data Gouv](#) in France). The collab-  
444 orative platform for source code GitHub could also be of interest. Albeit dedicated to source code  
445 and not designed to host large-size binary files, GitHub handles small to medium-size text files like  
446 tabular .csv and .tsv data files and has been extensively used to record cases of the Ebola epidemic  
447 in 2014 (*Perkel*, 2016) and the Covid-19 pandemic (<https://github.com/CSSEGISandData/COVID-19>).  
448 Thus, GitHub could probably host small text-based MD simulation files. For Gromacs, we already  
449 found 70,000 parameter .mdp files and 55,000 structure .gro files. Scripts found along these files

450 could also provide valuable insights to understand how a given MD analysis was performed. Fi-  
451 nally, GitHub repositories might also be an entry point to find other datasets by linking to simu-  
452 lation data, such as institutional repositories (see for instance (*Pesce and Lindorff-Larsen, 2023*)).  
453 However, one potential point of concern is that repositories like GitHub or GitLab do not make any  
454 promises about long-term availability of repositories, in particular ones not under active develop-  
455 ment. Archiving of these repositories could be achieved in Zenodo (for data-centric repositories)  
456 or Software Heritage (*Di Cosmo and Zacchioli, 2017*) (for source-code-centric repositories).

457 An obvious next step is the enrichment of metadata with the hope to render open MD data  
458 more findable, accessible and ultimately reusable. Possible strategies have already been detailed  
459 previously in this paper. We could also go further by connecting MD data in the research ecosystem.  
460 For this, two apparent resources need to be linked to MD datasets: their associated research pa-  
461 pers to mine more information and to establish a connection with the scientific context, and their  
462 simulated biomolecular systems, which ultimately could cross-reference MD datasets to reference  
463 databases such as UniProt (*Consortium, 2022*), the PDB (*Berman et al., 2000*) or Lipid Maps (*Sud  
464 et al., 2007*). For already deposited datasets, the enrichment of metadata can only be achieved  
465 via systematic computational approaches, while for future depositions, a clear and uniformly used  
466 ontology and dedicated metadata reference file (as it is used by the PLUMED-NEST: *Bonomi et al.  
467 (2019)*) would facilitate this task.

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

## 471 **Conclusion**

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

## 484 **Methods and Materials**

### 485 **Initial data collection**

486 We searched for MD-related files in the data repositories Zenodo, Figshare and Open Science  
487 Framework (OSF). Queries were designed with a combination of file types and optionally keywords,  
488 depending on how a given file type was solely associated to MD simulations. We therefore built a  
489 list of manually curated and cross-checked file types and keywords ([https://github.com/MDverse/  
490 mdws/blob/main/params/query.yml](https://github.com/MDverse/mdws/blob/main/params/query.yml)). All queries were automated by Python scripts that utilized Ap-  
491 plication Programming Interfaces (APIs) provided by data repositories. Since APIs offered by data  
492 repositories were different, all implementations were performed in dedicated Python(*van Rossum,  
493 1995*) (version 3.9.16) scripts with the NumPy(*Oliphant, 2007*) (version 1.24.2), Pandas(*Wes McKin-  
494 ney, 2010*) (version 1.5.3) and Requests (version 2.28.2) libraries.

495 We made the assumption that files deposited by researchers in data repositories were coherent  
496 and all related to a same research project. Therefore, when an MD-related file was found in a  
497 dataset, all files belonging to this dataset were indexed, regardless of whether their file types were

498 actually identified as MD simulation files. This is the core of the Explore and Expand strategy ( $Ex^2$ )  
499 we applied in this work and illustrated in Fig 1.

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

### 507 **Gromacs files**

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

513 Once downloaded, .mdp files were parsed to extract the following parameters: integrator, time  
514 step, number of steps, temperature, thermostat, and barostat. Values for thermostat and barostat  
515 were normalized according to values provided by the Gromacs documentation. For the simulation  
516 time analysis, we selected .mdp files with the *md* or *sd* integrator and with simulation time above  
517 1 ns to exclude most minimization and equilibrating simulations. For the thermostat and barostat  
518 analysis, only files with non-missing values and with values listed in the Gromacs documentation  
519 were considered.

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

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

### 527 **MDverse data explorer web app**

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

### 530 **System visualization and molecular graphics**

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

#### 536 **TonB**

537 Dataset URL: <https://zenodo.org/record/3756664>

538 Publication (DOI): <https://doi.org/10.1039/D0CP03473H>

#### 539 **T4 Lysozyme**

540 Dataset URL: <https://zenodo.org/record/3989044>

541 Publication (DOI): <https://doi.org/10.1021/acs.jctc.0c01338>

#### 542 **Benzene**

543 Dataset URL: [https://figshare.com/articles/dataset/Capturing\\_Protein\\_Ligand\\_Recognition\\_Pathways\\_](https://figshare.com/articles/dataset/Capturing_Protein_Ligand_Recognition_Pathways_)

- 544 [in\\_Coarse-Grained\\_Simulation/12517490/1](#)  
545 Publication (DOI): <https://doi.org/10.1021/acs.jpcltt.0c01683>
- 546 **Ammonia**  
547 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](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)  
548 [Using\\_Molecular\\_Dynamics\\_with\\_Explicit\\_Polarization\\_and\\_Induced\\_Polarity\\_Decoupling\\_An\\_On\\_the\\_Fly\\_Polarization\\_Approach/11702442](#)  
549 [the\\_Fly\\_Polarization\\_Approach/11702442](#)  
550 Publication (DOI): <https://doi.org/10.1021/acs.jctc.9b01139>
- 551 **Peptide with membrane**  
552 Dataset URL: <https://zenodo.org/record/4371296>  
553 Publication (DOI): <https://doi.org/10.1021/acs.jcim.0c01312>
- 554 **Kir channels**  
555 Dataset URL: <https://zenodo.org/record/3634884>  
556 Publication (DOI): <https://doi.org/10.1073/pnas.1918387117>
- 557 **Gasdermin**  
558 Dataset URL: <https://zenodo.org/record/6797842>  
559 Publication (DOI): <https://doi.org/10.7554/eLife.81432>
- 560 **Protein-RNA**  
561 Dataset URL: <https://zenodo.org/record/1308045>  
562 Publication (DOI): <https://doi.org/10.1371/journal.pcbi.1006642>
- 563 **G-quadruplex**  
564 Dataset URL: <https://zenodo.org/record/5594466>  
565 Publication (DOI): <https://doi.org/10.1021/jacs.1c11248>
- 566 **Ptb**  
567 Dataset URL: <https://osf.io/4aghb/>  
568 Publication (DOI): <https://doi.org/10.1073/pnas.2116543119>
- 569 **EPI-7170**  
570 Dataset URL: <https://zenodo.org/record/7120845>  
571 Publication (DOI): <https://doi.org/10.1038/s41467-022-34077-z>
- 572 **Gold nanoparticle**  
573 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\\_/2481241](https://acs.figshare.com/articles/dataset/Fluorescence_Probing_of_Thiol_Functionalized_Gold_Nanoparticles_Is_Alkylthiol_Coating_of_a_Nanoparticle_as_Hydrophobic_as_Expected_/2481241)  
574 [Gold\\_Nanoparticles\\_Is\\_Alkylthiol\\_Coating\\_of\\_a\\_Nanoparticle\\_as\\_Hydrophobic\\_as\\_Expected\\_/2481241](#)  
575 Publication (DOI): <https://doi.org/10.1021/jp3060813>
- 576 **Gd(DOTA)**  
577 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](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)  
578 [Molecular\\_Dynamics\\_Simulations\\_Toward\\_a\\_Rational\\_Optimization\\_of\\_MRI\\_Contrast\\_Agents/20334621](#)  
579 Publication (DOI): <https://doi.org/10.1021/acs.inorgchem.2c01597>
- 580 **Metallo cage**  
581 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](https://acs.figshare.com/articles/dataset/Rationalizing_the_Activity_of_an_Artificial_Diels-Alderase_Establishing_Efficient_and_Accurate_Protocols_for_Calculating_Supramolecular_Catalysis/11569452)  
582 [Diels-Alderase\\_Establishing\\_Efficient\\_and\\_Accurate\\_Protocols\\_for\\_Calculating\\_Supramolecular\\_Catalysis/11569452](#)  
583 [11569452](#)  
584 Publication (DOI): <https://doi.org/10.1021/jacs.9b10302>
- 585 **AL1**  
586 Dataset URL: [https://acs.figshare.com/articles/dataset/Nucleation\\_Mechanisms\\_of\\_Self-Assembled\\_Physisorbed\\_Monolayers\\_on\\_Graphite/8846045](https://acs.figshare.com/articles/dataset/Nucleation_Mechanisms_of_Self-Assembled_Physisorbed_Monolayers_on_Graphite/8846045)  
587 [Physisorbed\\_Monolayers\\_on\\_Graphite/8846045](#)  
588 Publication (DOI): <https://doi.org/10.1021/acs.jpcc.9b01234>

### 589 **PTEG-1 (all-atom)**

590 Dataset URL: [https://figshare.com/articles/dataset/PTEG-1\\_PP\\_and\\_N-DMBI\\_atomistic\\_force\\_fields/5458144](https://figshare.com/articles/dataset/PTEG-1_PP_and_N-DMBI_atomistic_force_fields/5458144)

592 Publication (DOI): <https://doi.org/10.1039/C7TA06609K>

### 593 **PTEG-1 (coarse-grain)**

594 Dataset URL: [https://figshare.com/articles/dataset/Neat\\_and\\_P3HT-Based\\_Blend\\_Morphologies\\_for\\_PCBM\\_and\\_PTEG-1/12338633](https://figshare.com/articles/dataset/Neat_and_P3HT-Based_Blend_Morphologies_for_PCBM_and_PTEG-1/12338633)

596 Publication (DOI): <https://doi.org/10.1002/adfm.202004799>

### 597 **Theophylline**

598 Dataset URL: [https://figshare.com/articles/dataset/A\\_Comparison\\_of\\_Methods\\_for\\_Computing\\_Relative\\_Anhydrous\\_Hydrate\\_Stability\\_with\\_Molecular\\_Simulation/21644393](https://figshare.com/articles/dataset/A_Comparison_of_Methods_for_Computing_Relative_Anhydrous_Hydrate_Stability_with_Molecular_Simulation/21644393)

600 Publication (DOI): <https://doi.org/10.1021/acs.cgd.2c00832>

## 601 **Data and code availability**

602 Data files produced from the data collection and processing are shared in Parquet format in the  
603 Zenodo repository: <https://zenodo.org/record/7856806>

604 Python scripts to search and index MD files, and to download and parse .mdp and .gro files are  
605 open-source, freely available on GitHub (<https://github.com/MDverse/mdws>) and archived in Soft-  
606 ware Heritage ([swh:1:dir:4d30b00345a732dc9f79d3c8bfae38b35b8f2c4](https://softwareheritage.org/1/1/dir:4d30b00345a732dc9f79d3c8bfae38b35b8f2c4)). A detailed documenta-  
607 tion is provided along the scripts to easily reproduce the data collection and processing.

608 Jupyter notebooks used to analyze results and create the figures of this paper are open-source,  
609 freely available on GitHub (<https://github.com/MDverse/mdda>) and archived in Software Heritage  
610 ([swh:1:dir:1f8497f72134cef0a9724c955bb03c751f52cccd](https://softwareheritage.org/1/1/dir:1f8497f72134cef0a9724c955bb03c751f52cccd)).

611 The code of the MDverse data explorer web application is open-source, freely available on  
612 GitHub (<https://github.com/MDverse/mdde>) and archived in Software Heritage ([swh:1:dir:1fc8b8eaabf4a9087e](https://softwareheritage.org/1/1/dir:1fc8b8eaabf4a9087e)).

## 613 **Acknowledgments**

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

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

## 624 **Author contributions**

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

## 631 **References**

632 **Abraham M**, Apostolov R, Barnoud J, Bauer P, Blau C, Bonvin AMJJ, Chavent M, Chodera J, Čondić Jurkić K, Dele-  
633 motte L, Grubmüller H, Howard RJ, Jordan EJ, Lindahl E, Ollila OHS, Selent J, Smith DGA, Stansfeld PJ, Tiemann



- 634 JKS, Trellet M, et al. Sharing Data from Molecular Simulations. *Journal of Chemical Information and Modeling*.  
635 2019 Oct; 59(10):4093–4099. <https://doi.org/10.1021/acs.jcim.9b00665>, doi: 10.1021/acs.jcim.9b00665.
- 636 **Abraham MJ**, Murtola T, Schulz R, Páll S, Smith JC, Hess B, Lindahl E. GROMACS: High performance molecular  
637 simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*. 2015; 1-2:19 – 25.  
638 doi: 10.1016/j.softx.2015.06.001.
- 639 **Abriata LA**, Lepore R, Dal Peraro M. About the need to make computational models of biological macro-  
640 molecules available and discoverable. *Bioinformatics*. 2020 May; 36(9):2952–2954. [https://doi.org/10.1093/](https://doi.org/10.1093/bioinformatics/btaa086)  
641 [bioinformatics/btaa086](https://doi.org/10.1093/bioinformatics/btaa086), doi: 10.1093/bioinformatics/btaa086.
- 642 **Aldeghi M**, Heifetz A, Bodkin MJ, Knapp S, Biggin PC. Accurate calculation of the absolute free energy of binding  
643 for drug molecules. *Chemical Science*. 2015; 7(1):207–218. doi: 10.1039/c5sc02678d.
- 644 **Alessandri R**, Grünewald F, Marrink SJ. The Martini Model in Materials Science. *Advanced Materials*. 2021;  
645 33(24):2008635. doi: 10.1002/adma.202008635.
- 646 **Alessandri R**, Sami S, Barnoud J, Vries AH, Marrink SJ, Havenith RWA. Resolving Donor–Acceptor Interfaces  
647 and Charge Carrier Energy Levels of Organic Semiconductors with Polar Side Chains. *Advanced Functional*  
648 *Materials*. 2020; 30(46):2004799. doi: 10.1002/adfm.202004799.
- 649 **Amaro RE**, Mulholland AJ. A Community Letter Regarding Sharing Biomolecular Simulation Data for COVID-19.  
650 *Journal of Chemical Information and Modeling*. 2020 Jun; 60(6):2653–2656. [https://doi.org/10.1021/acs.jcim.](https://doi.org/10.1021/acs.jcim.0c00319)  
651 [0c00319](https://doi.org/10.1021/acs.jcim.0c00319), doi: 10.1021/acs.jcim.0c00319, publisher: American Chemical Society.
- 652 **Antila HS**, M Ferreira T, Ollila OHS, Miettinen MS. Using Open Data to Rapidly Benchmark Biomolecular Simu-  
653 lations: Phospholipid Conformational Dynamics. *Journal of Chemical Information and Modeling*. 2021 Feb;  
654 61(2):938–949. <https://doi.org/10.1021/acs.jcim.0c01299>, doi: 10.1021/acs.jcim.0c01299, publisher: American  
655 Chemical Society.
- 656 **Armstrong DR**, Berrisford JM, Conroy MJ, Gutmanas A, Anyango S, Choudhary P, Clark AR, Dana JM, Deshpande  
657 M, Dunlop R, Gane P, Gáborová R, Gupta D, Haslam P, Koča J, Mak L, Mir S, Mukhopadhyay A, Nadzirin N, Nair  
658 S, et al. PDBe: improved findability of macromolecular structure data in the PDB. *Nucleic Acids Research*.  
659 2019; 48(D1):D335–D343. doi: 10.1093/nar/gkz990.
- 660 **Auer S**, Towards an Open Research Knowledge Graph. Zenodo; 2018. <https://doi.org/10.5281/zenodo.1157185>,  
661 doi: 10.5281/zenodo.1157185.
- 662 **Berendsen HJC**, Postma JPM, Gunsteren WFV, DiNola A, Haak JR. Molecular dynamics with coupling to an  
663 external bath. *The Journal of Chemical Physics*. 1984 04; 81(8):3684 – 3690. [https://aip.scitation.org/doi/10.](https://aip.scitation.org/doi/10.1063/1.448118)  
664 [1063/1.448118](https://aip.scitation.org/doi/10.1063/1.448118), doi: 10.1063/1.448118.
- 665 **Berendsen HJC**, van der Spoel D, van Drunen R. GROMACS: A Message-Passing Parallel Molecular Dynam-  
666 ics Implementation. *Computer Physics Communications*. 1995 Sep; 91(1-3):43–56. doi: 10.1016/0010-  
667 4655(95)00042-E.
- 668 **Berman H**, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. *Nature structural biology*.  
669 2003; 10(12):980. doi: 10.1038/nsb1203-980.
- 670 **Berman HM**, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data  
671 Bank. *Nucleic Acids Research*. 2000; 28(1):235–242. doi: 10.1093/nar/28.1.235.
- 672 **Bernetti M**, Bussi G. Pressure control using stochastic cell rescaling. *The Journal of Chemical Physics*. 2020;  
673 153(11):114107. doi: 10.1063/5.0020514.
- 674 **Bonomi M**, Bussi G, Camilloni C, Tribello GA, Banáš P, Barducci A, Bernetti M, Bolhuis PG, Bottaro S, Branduardi  
675 D, Capelli R, Carloni P, Ceriotti M, Cesari A, Chen H, Chen W, Colizzi F, De S, De La Pierre M, Donadio D, et al.  
676 Promoting transparency and reproducibility in enhanced molecular simulations. *Nature Methods*. 2019 Aug;  
677 16(8):670–673. <https://www.nature.com/articles/s41592-019-0506-8>, doi: 10.1038/s41592-019-0506-8.
- 678 **Bottaro S**, Lindorff-Larsen K. Biophysical experiments and biomolecular simulations: A perfect match? *Sci-*  
679 *ence*. 2018 07; 361(6400):355 – 360. <http://science.sciencemag.org/content/361/6400/355>, doi: 10.1126/sci-  
680 [ence.aat4010](http://science.sciencemag.org/content/361/6400/355).
- 681 **Bowers KJ**, Chow DE, Xu H, Dror RO, Eastwood MP, Gregersen BA, Klepeis JL, Kolossvary I, Moraes MA, Sacerdoti  
682 FD, Salmon JK, Shan Y, Shaw DE. Scalable Algorithms for Molecular Dynamics Simulations on Commodity  
683 Clusters. *ACM/IEEE SC 2006 Conference (SC'06)*. 2006; p. 43–43. doi: 10.1109/sc.2006.54.

- 684 **Bran AM**, Cox S, White AD, Schwaller P, ChemCrow: Augmenting large-language models with chemistry tools;  
685 2023.
- 686 **Brooks BR**, Brooks CL, Mackerell AD, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S,  
687 Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodosek M, Im W, Kuczera K, et al. CHARMM:  
688 the biomolecular simulation program. *Journal of computational chemistry*. 2009 07; 30(10):1545 – 1614. doi:  
689 [10.1002/jcc.21287](https://doi.org/10.1002/jcc.21287).
- 690 **Burley SK**, Kurisu G, Markley JL, Nakamura H, Velankar S, Berman HM, Sali A, Schwede T, Trewella J. PDB-Dev:  
691 a Prototype System for Depositing Integrative/Hybrid Structural Models. *Structure (London, England : 1993)*.  
692 2017 09; 25(9):1317 – 1318. doi: [10.1016/j.str.2017.08.001](https://doi.org/10.1016/j.str.2017.08.001).
- 693 **Bussi G**, Donadio D, Parrinello M. Canonical sampling through velocity rescaling. *The Journal of*  
694 *Chemical Physics*. 2007; 126(1):014101. [http://jcp.aip.org/resource/1/jcpsa6/v126/i1/p014101\\_s1](http://jcp.aip.org/resource/1/jcpsa6/v126/i1/p014101_s1), doi:  
695 [10.1063/1.2408420](https://doi.org/10.1063/1.2408420).
- 696 **Commun Biol e**. Reliability and reproducibility checklist for molecular dynamics simulations. *Communications*  
697 *Biology*. 2023 Mar; 6(1). <https://doi.org/10.1038/s42003-023-04653-0>, doi: 10.1038/s42003-023-04653-0.
- 698 **consortium TP**. Promoting transparency and reproducibility in enhanced molecular simulations. *Nat Methods*.  
699 2019 08; 16(8):670 – 673. <https://www.nature.com/articles/s41592-019-0506-8>, doi: 10.1038/s41592-019-0506-  
700 8.
- 701 **Consortium TU**. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*. 2022 11;  
702 51(D1):D523–D531. <https://doi.org/10.1093/nar/gkac1052>, doi: 10.1093/nar/gkac1052.
- 703 **Dandekar BR**, Mondal J. Capturing Protein–Ligand Recognition Pathways in Coarse-Grained Simulation. *The*  
704 *Journal of Physical Chemistry Letters*. 2020; 11(13):5302–5311. doi: [10.1021/acs.jpcllett.0c01683](https://doi.org/10.1021/acs.jpcllett.0c01683).
- 705 **Di Cosmo R**, Zacchiroli S. Software Heritage: Why and How to Preserve Software Source Code. In: *Proceedings*  
706 *of the 14th International Conference on Digital Preservation, iPRES 2017 Japan*; 2017. .
- 707 **Domański J**, Stansfeld PJ, Sansom MSP, Beckstein O. Lipidbook: a public repository for force-field param-  
708 eters used in membrane simulations. *The Journal of membrane biology*. 2010; 236(3):255 – 258. doi:  
709 [10.1007/s00232-010-9296-8](https://doi.org/10.1007/s00232-010-9296-8).
- 710 **Duncan AL**, Corey RA, Sansom MSP. Defining how multiple lipid species interact with inward rectifier potassium  
711 (Kir2) channels. *Proc Natl Acad Sci USA*. 2020 04; 117(14):7803 – 7813. doi: [10.1073/pnas.1918387117](https://doi.org/10.1073/pnas.1918387117).
- 712 **Dybeck EC**, Thiel A, Schnieders MJ, Pickard FC, Wood GPF, Krzyzaniak JF, Hancock BC. A Comparison of Methods  
713 for Computing Relative Anhydrous–Hydrate Stability with Molecular Simulation. *Crystal Growth & Design*.  
714 2023; 23(1):142–167. doi: [10.1021/acs.cgd.2c00832](https://doi.org/10.1021/acs.cgd.2c00832).
- 715 **Elofsson A**, Hess B, Lindahl E, Onufriev A, Spoel DVD, Wallqvist A. Ten simple rules on how to create open  
716 access and reproducible molecular simulations of biological systems. *PLOS Computational Biology*. 2019;  
717 15(1):e1006649. doi: [10.1371/journal.pcbi.1006649](https://doi.org/10.1371/journal.pcbi.1006649).
- 718 **European Organization For Nuclear Research**, OpenAIRE, Zenodo. CERN; 2013. <https://www.zenodo.org/>,  
719 doi: 10.25495/7GXX-RD71.
- 720 **Fadda E**. Molecular simulations of complex carbohydrates and glycoconjugates. *Current Opinion in Chemical*  
721 *Biology*. 2022; 69:102175. doi: [10.1016/j.cbpa.2022.102175](https://doi.org/10.1016/j.cbpa.2022.102175).
- 722 **Fan FJ**, Shi Y. Effects of data quality and quantity on deep learning for protein-ligand binding affinity predic-  
723 tion. *Bioorganic & Medicinal Chemistry*. 2022; 72:117003. [https://www.sciencedirect.com/science/article/pii/](https://www.sciencedirect.com/science/article/pii/S0968089622003960)  
724 [S0968089622003960](https://doi.org/10.1016/j.bmc.2022.117003), doi: <https://doi.org/10.1016/j.bmc.2022.117003>.
- 725 **Fawzi NL**, Parekh SH, Mittal J. Biophysical studies of phase separation integrating experimental and computa-  
726 tional methods. *Current Opinion in Structural Biology*. 2021 Oct; 70:78–86. [https://www.sciencedirect.com/](https://www.sciencedirect.com/science/article/pii/S0959440X21000580)  
727 [science/article/pii/S0959440X21000580](https://doi.org/10.1016/j.sbi.2021.04.004), doi: [10.1016/j.sbi.2021.04.004](https://doi.org/10.1016/j.sbi.2021.04.004).
- 728 **Ferrer RS**, Case DA, Walker RC. An overview of the Amber biomolecular simulation package. *Wiley Interdisci-*  
729 *plinary Reviews: Computational Molecular Science*. 2012; 3(2):198 – 210. doi: [10.1002/wcms.1121](https://doi.org/10.1002/wcms.1121).
- 730 **Fuller JC**, Jackson RM, Edwards TA, Wilson AJ, Shirts MR. Modeling of Arylamide Helix Mimetics in the p53  
731 Peptide Binding Site of hDM2 Suggests Parallel and Anti-Parallel Conformations Are Both Stable. *PLOS ONE*.  
732 2012 08; 7(8):1–17. <https://doi.org/10.1371/journal.pone.0043253>, doi: [10.1371/journal.pone.0043253](https://doi.org/10.1371/journal.pone.0043253).

- 733 **Färber M**, Lamprecht D. The data set knowledge graph: Creating a linked open data source for data  
734 sets. *Quantitative Science Studies*. 2021 12; 2(4):1324–1355. [https://doi.org/10.1162/qss\\_a\\_00161](https://doi.org/10.1162/qss_a_00161), doi:  
735 10.1162/qss\_a\_00161.
- 736 **Genheden S**, Ryde U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert  
737 Opinion on Drug Discovery*. 2015; 10(5):449–461. <https://doi.org/10.1517/17460441.2015.1032936>, doi:  
738 10.1517/17460441.2015.1032936, pMID: 25835573.
- 739 **Gertsen AS**, Sørensen MK, Andreasen JW. Nanostructure of organic semiconductor thin films: Molecular dy-  
740 namics modeling with solvent evaporation. *Physical Review Materials*. 2020; 4(7):075405. doi: 10.1103/phys-  
741 revmaterials.4.075405.
- 742 **Gowers R**, Linke M, Barnoud J, Reddy T, Melo M, Seyler S, Domański J, Dotson D, Buchoux S, Kenney I, Beck-  
743 stein O. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. In: *Pro-  
744 ceedings of the Python in Science Conference SciPy*; 2016. <https://doi.org/10.25080/majora-629e541a-00e>, doi:  
745 10.25080/majora-629e541a-00e.
- 746 **Gupta C**, Sarkar D, Tieleman DP, Singharoy A. The ugly, bad, and good stories of large-scale biomolecular  
747 simulations. *Current Opinion in Structural Biology*. 2022 Apr; 73:102338. [https://www.sciencedirect.com/  
748 science/article/pii/S0959440X22000112](https://www.sciencedirect.com/science/article/pii/S0959440X22000112), doi: 10.1016/j.sbi.2022.102338.
- 749 **Hoch JC**, Baskaran K, Burr H, Chin J, Eghbalnia HR, Fujiwara T, Gryk MR, Iwata T, Kojima C, Kurisu G, et al.  
750 Biological Magnetic Resonance Data Bank. *Nucleic Acids Research*. 2023; 51(D1):D368–D376.
- 751 **Hollingsworth SA**, Dror RO. Molecular Dynamics Simulation for All. *Neuron*. 2018 Sep; 99(6):1129–1143. [https://  
752 www.sciencedirect.com/science/article/pii/S0896627318306846](https://www.sciencedirect.com/science/article/pii/S0896627318306846), doi: 10.1016/j.neuron.2018.08.011.
- 753 **Hospital A**, Battistini F, Soliva R, Gelpí JL, Orozco M. Surviving the deluge of biosimulation data. *WIREs Compu-  
754 tational Molecular Science*. 2020; 10(3):e1449. <https://onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1449>, doi:  
755 10.1002/wcms.1449.
- 756 **Humphrey W**, Dalke A, Schulten K. VMD: visual molecular dynamics. *J Mol Graph*. 1996; 14(1):33 – 8, 27–8.
- 757 **Hénin J**, Lelièvre T, Shirts MR, Valsson O, Delemotte L. Enhanced Sampling Methods for Molecular Dynamics  
758 Simulations [Article v1.0]. *Living Journal of Computational Molecular Science*. 2022; 4(1). doi: 10.33011/live-  
759 coms.4.1.1583.
- 760 **Ingólfsson HI**, Lopez CA, Uusitalo JJ, Jong DHd, Gopal SM, Periole X, Marrink SJ. The power of coarse graining  
761 in biomolecular simulations. *Wiley Interdisciplinary Reviews: Computational Molecular Science*. 2014 05;  
762 4(3):225 – 248. <https://onlinelibrary.wiley.com/doi/full/10.1002/wcms.1169>, doi: 10.1002/wcms.1169.
- 763 **Ivanov P**, Mu J, Leay L, Chang SY, Sharrad CA, Masters AJ, Schroeder SLM. Organic and Third Phase in  
764 HNO<sub>3</sub>/TBP/n-Dodecane System: No Reverse Micelles. *Solvent Extraction and Ion Exchange*. 2017; 35(4):251–  
765 265. doi: 10.1080/07366299.2017.1336048.
- 766 **Jahani F**, Torabi S, Chiechi RC, Koster LJA, Hummelen JC. Fullerene derivatives with increased dielectric con-  
767 stants. *Chemical Communications*. 2014; 50(73):10645–10647. doi: 10.1039/c4cc04366a.
- 768 **Jeon JH**, Javanainen M, Martinez-Seara H, Metzler R. Protein Crowding in Lipid Bilayers Gives Rise to  
769 Non-Gaussian Anomalous Lateral Diffusion of Phospholipids and Proteins. *Physical Review X*. 2016; doi:  
770 10.1103/physrevx.6.021006.
- 771 **Jumper J**, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Židek A,  
772 Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J,  
773 Back T, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021 Aug; 596(7873):583–  
774 589. <https://www.nature.com/articles/s41586-021-03819-2>, doi: 10.1038/s41586-021-03819-2, number: 7873  
775 Publisher: Nature Publishing Group.
- 776 **Kabelka I**, Brožek R, Vácha R. Selecting Collective Variables and Free-Energy Methods for Peptide Translo-  
777 cation across Membranes. *Journal of Chemical Information and Modeling*. 2021; 61(2):819–830. doi:  
778 10.1021/acs.jcim.0c01312.
- 779 **Kampfrath M**, Staritzbichler R, Hernández GP, Rose AS, Tiemann JKS, Scheuermann G, Wiegrefe D, Hildebrand  
780 PW. MDsrv: visual sharing and analysis of molecular dynamics simulations. *Nucleic Acids Research*. 2022;  
781 50(W1):W483–W489. doi: 10.1093/nar/gkac398.

- 782 **Karunasena C**, Li S, Heifner MC, Ryno SM, Risko C. Reconsidering the Roles of Noncovalent In-  
783 tramolecular “Locks” in  $\pi$ -Conjugated Molecules. *Chemistry of Materials*. 2021; 33(23):9139–9151. doi:  
784 [10.1021/acs.chemmater.1c02335](https://doi.org/10.1021/acs.chemmater.1c02335).
- 785 **Kelly BD**, Smith WR. Alchemical Hydration Free-Energy Calculations Using Molecular Dynamics with Explicit  
786 Polarization and Induced Polarity Decoupling: An On-the-Fly Polarization Approach. *Journal of Chemical*  
787 *Theory and Computation*. 2020; 16(2):1146–1161. doi: [10.1021/acs.jctc.9b01139](https://doi.org/10.1021/acs.jctc.9b01139).
- 788 **Kiirikki A**, Antila H, Bort L, Buslaev P, Fernando F, Ferreira TM, Fuchs P, Garcia-Fandino R, Gushchin I, Kav B, Kula  
789 P, Kurki M, Kuzmin A, Madsen J, Miettinen M, Nencini R, Piggot T, Pineiro A, Samantray S, Suarez-Leston F, et al.  
790 NMRlipids Databank makes data-driven analysis of biomembrane properties accessible for all. *ChemRxiv*.  
791 2023 Feb; <https://doi.org/10.26434/chemrxiv-2023-jrjpw>, doi: 10.26434/chemrxiv-2023-jrjpw.
- 792 **Kinjo AR**, Bekker GJ, Suzuki H, Tsuchiya Y, Kawabata T, Ikegawa Y, Nakamura H. Protein Data Bank Japan (PDBj):  
793 updated user interfaces, resource description framework, analysis tools for large structures. *Nucleic Acids*  
794 *Research*. 2017; 45(D1):D282–D288. doi: 10.1093/nar/gkw962.
- 795 **Kirschner KN**, Yongye AB, Tschampel SM, González-Outeiriño J, Daniels CR, Foley BL, Woods RJ. GLYCAM06:  
796 a generalizable biomolecular force field. *Carbohydrates*. *Journal of computational chemistry*. 2008-03;  
797 29(4):622 – 655. doi: [10.1002/jcc.20820](https://doi.org/10.1002/jcc.20820).
- 798 **Krishna S**, Sreedhar I, Patel CM. Molecular dynamics simulation of polyamide-based materials – A review.  
799 *Computational Materials Science*. 2021 Dec; 200:110853. <https://www.sciencedirect.com/science/article/pii/S0927025621005711>, doi: [10.1016/j.commatsci.2021.110853](https://doi.org/10.1016/j.commatsci.2021.110853).
- 800  
801 **Kyrychenko A**, Karpushina GV, Svechkarev D, Kolodezny D, Bogatyrenko SI, Kryshtal AP, Doroshenko AO.  
802 Fluorescence Probing of Thiol-Functionalized Gold Nanoparticles: Is Alkylthiol Coating of a Nanoparticle  
803 as Hydrophobic as Expected? *The Journal of Physical Chemistry C*. 2012; 116(39):21059–21068. doi:  
804 [10.1021/jp3060813](https://doi.org/10.1021/jp3060813).
- 805 **Kümmerer F**, Orioli S, Harding-Larsen D, Hoffmann F, Gavrilov Y, Teilum K, Lindorff-Larsen K. Fitting Side-  
806 Chain NMR Relaxation Data Using Molecular Simulations. *Journal of Chemical Theory and Computation*.  
807 2021; 17(8):5262–5275. doi: [10.1021/acs.jctc.0c01338](https://doi.org/10.1021/acs.jctc.0c01338).
- 808 **Lane TJ**. Protein structure prediction has reached the single-structure frontier. *Nature Methods*. 2023; p. 1–4.  
809 doi: [10.1038/s41592-022-01760-4](https://doi.org/10.1038/s41592-022-01760-4).
- 810 **Liu S**, Cao S, Hoang K, Young KL, Paluch AS, Mobley DL. Using MD Simulations To Calculate How Sol-  
811 vents Modulate Solubility. *Journal of Chemical Theory and Computation*. 2016; 12(4):1930–1941. doi:  
812 [10.1021/acs.jctc.5b00934](https://doi.org/10.1021/acs.jctc.5b00934).
- 813 **Mahmud M**, Kaiser MS, McGinnity TM, Hussain A. Deep Learning in Mining Biological Data. *Cognitive Compu-*  
814 *tation*. 2021; 13(1):1–33. <https://doi.org/10.1007/s12559-020-09773-x>, doi: 10.1007/s12559-020-09773-x.
- 815 **Marklund EG**, Benesch JL. Weighing-up protein dynamics: the combination of native mass spectrometry and  
816 molecular dynamics simulations. *Current Opinion in Structural Biology*. 2019 Feb; 54:50–58. [https://www.](https://www.sciencedirect.com/science/article/pii/S0959440X18300630)  
817 [sciencedirect.com/science/article/pii/S0959440X18300630](https://www.sciencedirect.com/science/article/pii/S0959440X18300630), doi: [10.1016/j.sbi.2018.12.011](https://doi.org/10.1016/j.sbi.2018.12.011).
- 818 **Martinez X**, Baaden M. UnityMol prototype for FAIR sharing of molecular-visualization experiences: from  
819 pictures in the cloud to collaborative virtual reality exploration in immersive 3D environments. *Acta Crystal-*  
820 *lographica Section D*. 2021; 77(6):746–754. doi: [10.1107/s2059798321002941](https://doi.org/10.1107/s2059798321002941).
- 821 **Marx V**. Biology: The Big Challenges of Big Data. *Nature*. 2013; 498(7453):255–260. doi: [10.1038/498255a](https://doi.org/10.1038/498255a).
- 822 **Wes McKinney**. Data Structures for Statistical Computing in Python. In: Stéfan van der Walt, Jarrod Millman,  
823 editors. *Proceedings of the 9th Python in Science Conference*; 2010. p. 56 – 61. doi: [10.25080/Majora-92bf1922-](https://doi.org/10.25080/Majora-92bf1922-00a)  
824 [00a](https://doi.org/10.25080/Majora-92bf1922-00a).
- 825 **Merz KMJ**, Amaro R, Cournia Z, Rarey M, Soares T, Tropsha A, Wahab HA, Wang R. Editorial: Method and  
826 Data Sharing and Reproducibility of Scientific Results. *Journal of Chemical Information and Modeling*. 2020  
827 Dec; 60(12):5868–5869. <https://doi.org/10.1021/acs.jcim.0c01389>, doi: [10.1021/acs.jcim.0c01389](https://doi.org/10.1021/acs.jcim.0c01389), publisher:  
828 American Chemical Society.
- 829 **Meyer T**, D’Abramo M, Hospital A, Rueda M, Ferrer-Costa C, Pérez A, Carrillo O, Camps J, Fenollosa C, Repchevsky  
830 D, Gelpí JL, Orozco M. MoDEL (Molecular Dynamics Extended Library): A Database of Atomistic Molecular  
831 Dynamics Trajectories. *Structure*. 2010 Nov; 18(11):1399–1409. [https://www.cell.com/structure/abstract/](https://www.cell.com/structure/abstract/S0969-2126(10)00353-9)  
832 [S0969-2126\(10\)00353-9](https://doi.org/10.1016/j.str.2010.07.013), doi: [10.1016/j.str.2010.07.013](https://doi.org/10.1016/j.str.2010.07.013), publisher: Elsevier.

- 833 **Michaud-Agrawal N**, Denning EJ, Woolf TB, Beckstein O. MDAnalysis: A toolkit for the analysis of molec-  
834 ular dynamics simulations. *Journal of computational chemistry*. 2011 04; 32(10):2319 – 2327. doi:  
835 [10.1002/jcc.21787](https://doi.org/10.1002/jcc.21787).
- 836 **Miksa T**, Simms S, Mietchen D, Jones S. Ten principles for machine-actionable data management plans. *PLOS*  
837 *Computational Biology*. 2019 03; 15(3):1–15. <https://doi.org/10.1371/journal.pcbi.1006750>, doi: [10.1371/jour-](https://doi.org/10.1371/journal.pcbi.1006750)  
838 [nal.pcbi.1006750](https://doi.org/10.1371/journal.pcbi.1006750).
- 839 **Mulholland AJ**, Amaro RE. COVID19 - Computational Chemists Meet the Moment. *Journal of Chemical*  
840 *Information and Modeling*. 2020 Dec; 60(12):5724–5726. <https://doi.org/10.1021/acs.jcim.0c01395>, doi:  
841 [10.1021/acs.jcim.0c01395](https://doi.org/10.1021/acs.jcim.0c01395).
- 842 **Musen MA**. Without Appropriate Metadata, Data-Sharing Mandates Are Pointless. *Nature*. 2022 Sep;  
843 609(7926):222–222. doi: [10.1038/d41586-022-02820-7](https://doi.org/10.1038/d41586-022-02820-7).
- 844 **Newport TD**, Sansom MSP, Stansfeld PJ. The MemProtMD database: a resource for membrane-embedded  
845 protein structures and their lipid interactions. *Nucleic Acids Research*. 2018; 47(Database issue):gky1047–.  
846 doi: [10.1093/nar/gky1047](https://doi.org/10.1093/nar/gky1047).
- 847 **Oliphant TE**. Python for Scientific Computing. *Computing in Science & Engineering*. 2007; 9(3):10–20. doi:  
848 [10.1109/MCSE.2007.58](https://doi.org/10.1109/MCSE.2007.58).
- 849 **Panero P**, Benito J, OpenAIRE Webinar: Zenodo - open digital repository. Zenodo; 2022. [https://doi.org/10.](https://doi.org/10.5281/zenodo.7417839)  
850 [5281/zenodo.7417839](https://doi.org/10.5281/zenodo.7417839), doi: [10.5281/zenodo.7417839](https://doi.org/10.5281/zenodo.7417839).
- 851 **Parrinello M**, Rahman A. Polymorphic transitions in single crystals: A new molecular dynamics method.  
852 *Journal of Applied Physics*. 1981; 52(12):7182. [http://jap.aip.org/resource/1/japiau/v52/i12/p7182\\_s1](http://jap.aip.org/resource/1/japiau/v52/i12/p7182_s1), doi:  
853 [10.1063/1.328693](https://doi.org/10.1063/1.328693).
- 854 **Perera N**, Dehmer M, Emmert-Streib F. Named Entity Recognition and Relation Detection for Biomedical Infor-  
855 mation Extraction. *Frontiers in Cell and Developmental Biology*. 2020; 8:673. doi: [10.3389/fcell.2020.00673](https://doi.org/10.3389/fcell.2020.00673).
- 856 **Perilla JR**, Goh BC, Cassidy CK, Liu B, Bernardi RC, Rudack T, Yu H, Wu Z, Schulten K. Molecular dynamics  
857 simulations of large macromolecular complexes. *Current opinion in structural biology*. 2015 Apr; 31:64–74.  
858 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4476923/>, doi: [10.1016/j.sbi.2015.03.007](https://doi.org/10.1016/j.sbi.2015.03.007).
- 859 **Perkel J**. Democratic Databases: Science on GitHub. *Nature*. 2016 Oct; 538(7623):127–128. doi:  
860 [10.1038/538127a](https://doi.org/10.1038/538127a).
- 861 **Pesce F**, Lindorff-Larsen K. Combining experiments and simulations to examine the temperature-dependent  
862 behaviour of a disordered protein. *bioRxiv*. 2023; [https://www.biorxiv.org/content/early/2023/03/05/2023.03.](https://www.biorxiv.org/content/early/2023/03/05/2023.03.04.531094)  
863 [04.531094](https://www.biorxiv.org/content/early/2023/03/05/2023.03.04.531094), doi: [10.1101/2023.03.04.531094](https://doi.org/10.1101/2023.03.04.531094).
- 864 **Pettersen EF**, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. UCSF Chimera—a visual-  
865 ization system for exploratory research and analysis. *Journal of computational chemistry*. 2004; 25(13):1605  
866 – 1612. [http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=15264254&retmode=ref&](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=15264254&retmode=ref&cmd=prlinks)  
867 [cmd=prlinks](http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=15264254&retmode=ref&cmd=prlinks), doi: [10.1002/jcc.20084](https://doi.org/10.1002/jcc.20084), pettersen, Eric F Goddard, Thomas D Huang, Conrad C Couch, Gregory  
868 S Greenblatt, Daniel M Meng, Elaine C Ferrin, Thomas E P41-RR01081/RR/NCRR NIH HHS/United States Re-  
869 search Support, U.S. Gov't, P.H.S. *United States Journal of computational chemistry J Comput Chem*. 2004  
870 Oct;25(13):1605-12.
- 871 **Phillips JC**, Hardy DJ, Maia JDC, Stone JE, Ribeiro JV, Bernardi RC, Buch R, Fiorin G, Hémin J, Jiang W, McGreevy R,  
872 Melo MCR, Radak BK, Skeel RD, Singharoy A, Wang Y, Roux B, Aksimentiev A, Luthey-Schulten Z, Kalé LV, et al.  
873 Scalable molecular dynamics on CPU and GPU architectures with NAMD. *The Journal of Chemical Physics*.  
874 2020; 153(4):044130. doi: [10.1063/5.0014475](https://doi.org/10.1063/5.0014475).
- 875 **Piskorz TK**, Gobbo C, Marrink SJ, Feyter SD, Vries AHd, Esch JHv. Nucleation Mechanisms of Self-Assembled  
876 Physisorbed Monolayers on Graphite. *The Journal of Physical Chemistry C*. 2019; 123(28):17510–17520. doi:  
877 [10.1021/acs.jpcc.9b01234](https://doi.org/10.1021/acs.jpcc.9b01234).
- 878 **Pohjolainen E**, Chen X, Malola S, Groenhof G, Häkkinen H. A Unified AMBER-Compatible Molecular Mechanics  
879 Force Field for Thiolate-Protected Gold Nanoclusters. *Journal of Chemical Theory and Computation*. 2016;  
880 12(3):1342–1350. doi: [10.1021/acs.jctc.5b01053](https://doi.org/10.1021/acs.jctc.5b01053).
- 881 **Porubsky VL**, Goldberg AP, Rampadarath AK, Nickerson DP, Karr JR, Sauro HM. Best Practices for Making  
882 Reproducible Biochemical Models. *Cell Systems*. 2020; 11(2):109–120. doi: [10.1016/j.cels.2020.06.012](https://doi.org/10.1016/j.cels.2020.06.012).



- 883 **Qiu L**, Liu J, Alessandri R, Qiu X, Koopmans M, Havenith RWA, Marrink SJ, Chiechi RC, Koster LJA, Hummelen JC.  
884 Enhancing doping efficiency by improving host-dopant miscibility for fullerene-based n-type thermoelectrics.  
885 *Journal of Materials Chemistry A*. 2017; 5(40):21234–21241. doi: 10.1039/c7ta066609k.
- 886 **Rodríguez-Espigares I**, Torrens-Fontanals M, Tiemann JKS, Aranda-García D, Ramírez-Anguita JM, Stepniewski  
887 TM, Worp N, Varela-Rial A, Morales-Pastor A, Medel-Lacruz B, Pándy-Szekeres G, Mayol E, Giorgino T, Carlsson  
888 J, Deupi X, Filipek S, Filizola M, Gómez-Tamayo JC, Gonzalez A, Gutiérrez-de Terán H, et al. GPCRmd uncovers  
889 the dynamics of the 3D-GPCRome. *Nature Methods*. 2020 Aug; 17(8):777–787. [https://www.nature.com/  
890 articles/s41592-020-0884-y](https://www.nature.com/articles/s41592-020-0884-y), doi: 10.1038/s41592-020-0884-y.
- 891 **Sami S**, Alessandri R, Wijaya JBW, Grünewald F, Vries AHd, Marrink SJ, Broer R, Havenith RWA. Strategies  
892 for Enhancing the Dielectric Constant of Organic Materials. *The Journal of Physical Chemistry C*. 2022;  
893 126(45):19462–19469. doi: 10.1021/acs.jpcc.2c05682.
- 894 **Sarkar A**, Sasmal R, Empereur-mot C, Bochicchio D, Kompella SVK, Sharma K, Dhiman S, Sundaram B, Agasti SS,  
895 Pavan GM, George SJ. Self-Sorted, Random, and Block Supramolecular Copolymers via Sequence Controlled,  
896 Multicomponent Self-Assembly. *Journal of the American Chemical Society*. 2020; 142(16):7606–7617. doi:  
897 10.1021/jacs.0c01822.
- 898 **Schaefer SL**, Hummer G. Sublytic gasdermin-D pores captured in atomistic molecular simulations. *eLife*. 2022;  
899 11:e81432. doi: 10.7554/elife.81432.
- 900 **Souza PCT**, Alessandri R, Barnoud J, Thallmair S, Faustino I, Grünewald F, Patmanidis I, Abdizadeh H, Bruininks  
901 BMH, Wassenaar TA, Kroon PC, Melcr J, Nieto V, Corradi V, Khan HM, Domański J, Javanainen M, Martinez-  
902 Seara H, Reuter N, Best RB, et al. Martini 3: a general purpose force field for coarse-grained molecular  
903 dynamics. *Nature Methods*. 2021; p. 1–7. doi: 10.1038/s41592-021-01098-3.
- 904 **Stansfeld PJ**, Goose JE, Caffrey M, Carpenter EP, Parker JL, Newstead S, Sansom MSP. MemProtMD: Automated  
905 Insertion of Membrane Protein Structures into Explicit Lipid Membranes. *Structure (London, England : 1993)*.  
906 2015 07; 23(7):1350 – 1361. doi: 10.1016/j.str.2015.05.006.
- 907 **Stephens ZD**, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, Iyer R, Schatz MC, Sinha S, Robinson GE. Big Data:  
908 Astronomical or Genomical? *PLOS Biology*. 2015; 13(7):e1002195. doi: 10.1371/journal.pbio.1002195.
- 909 **Sud M**, Fahy E, Cotter D, Brown A, Dennis EA, Glass CK, Merrill AH, Murphy RC, Raetz CRH, Russell DW, Subrama-  
910 niam S. LMSD: LIPID MAPS structure database. *Nucleic Acids Research*. 2007 Jan; 35(Database):D527–D532.  
911 <https://doi.org/10.1093/nar/gkl838>, doi: 10.1093/nar/gkl838.
- 912 **Tai K**, Murdock S, Wu B, Ng MH, Johnston S, Fangohr H, Cox SJ, Jeffreys P, Essex JW, Sansom MSP. BioSimGrid:  
913 towards a worldwide repository for biomolecular simulations. *Organic & Biomolecular Chemistry*. 2004 Nov;  
914 2(22):3219–3221. <https://pubs.rsc.org/en/content/articlelanding/2004/ob/b411352g>, doi: 10.1039/B411352G,  
915 publisher: The Royal Society of Chemistry.
- 916 **Tiemann JKS**, Guixà-González R, Hildebrand PW, Rose AS. MDSrv: viewing and sharing molecular dynamics  
917 simulations on the web. *Nat Methods*. 2017 12; 14(12):1123 – 1124. [https://www.nature.com/articles/nmeth.  
918 4497](https://www.nature.com/articles/nmeth.4497), doi: 10.1038/nmeth.4497.
- 919 **van Rossum G**. Python Tutorial. Amsterdam: Centrum voor Wiskunde en Informatica (CWI); 1995.
- 920 **Vanommeslaeghe K**, MacKerell AD. Automation of the CHARMM General Force Field (CGenFF) I: Bond Per-  
921 ception and Atom Typing. *Journal of Chemical Information and Modeling*. 2012; 52(12):3144–3154. doi:  
922 10.1021/ci300363c.
- 923 **Virtanen SI**, Kiirikki AM, Mikula KM, Iwai H, Ollila OHS. Heterogeneous dynamics in partially disordered proteins.  
924 *Physical Chemistry Chemical Physics*. 2020; 22(37):21185–21196. doi: 10.1039/d0cp03473h.
- 925 **Vuorio J**, Vattulainen I, Martinez-Seara H. Atomistic fingerprint of hyaluronan–CD44 binding. *PLoS Computa-  
926 tional Biology*. 2017; 13(7):e1005663. doi: 10.1371/journal.pcbi.1005663.
- 927 **Waltemath D**, Adams R, Beard DA, Bergmann FT, Bhalla US, Britten R, Chelliah V, Cooling MT, Cooper J, Crampin  
928 EJ, Garry A, Hoops S, Hucka M, Hunter P, Klipp E, Laibe C, Miller AK, Moraru I, Nickerson D, Nielsen P, et al.  
929 Minimum Information About a Simulation Experiment (MIASE). *PLOS Computational Biology*. 2011 04; 7(4):1–  
930 4. <https://doi.org/10.1371/journal.pcbi.1001122>, doi: 10.1371/journal.pcbi.1001122.
- 931 **Wang J**, Wolf RM, Caldwell JW, Kollman PA, Case DA. Development and testing of a general amber force field.  
932 *Journal of computational chemistry*. 2004 07; 25(9):1157 – 1174. doi: 10.1002/jcc.20035.

- 933 **Wilkinson MD**, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, Santos  
934 LBdS, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers  
935 R, Gonzalez-Beltran A, et al. The FAIR Guiding Principles for scientific data management and stewardship.  
936 *Scientific data*. 2016 03; 3:160018. doi: [10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18).
- 937 **Wilson SL**, Way GP, Bittremieux W, Armache JP, Haendel MA, Hoffman MM. Sharing biological data: why, when,  
938 and how. *FEBS Letters*. 2021; 595(7):847–863. [https://febs.onlinelibrary.wiley.com/doi/abs/10.1002/1873-3468.](https://febs.onlinelibrary.wiley.com/doi/abs/10.1002/1873-3468.14067)  
939 [14067](https://doi.org/10.1002/1873-3468.14067), doi: <https://doi.org/10.1002/1873-3468.14067>.
- 940 **Yoo J**, Winogradoff D, Aksimentiev A. Molecular dynamics simulations of DNA-DNA and DNA-protein interac-  
941 tions. *Current Opinion in Structural Biology*. 2020 Oct; 64:88–96. doi: [10.1016/j.sbi.2020.06.007](https://doi.org/10.1016/j.sbi.2020.06.007).
- 942 **Young TA**, Martí-Centelles V, Wang J, Lusby PJ, Duarte F. Rationalizing the Activity of an “Artificial Diels-Alderase”:  
943 Establishing Efficient and Accurate Protocols for Calculating Supramolecular Catalysis. *Journal of the Ameri-*  
944 *can Chemical Society*. 2020; 142(3):1300–1310. doi: [10.1021/jacs.9b10302](https://doi.org/10.1021/jacs.9b10302).
- 945 **Zheng X**, Chan MHY, Chan AKW, Cao S, Ng M, Sheong FK, Li C, Goonetilleke EC, Lam WWY, Lau TC, Huang  
946 X, Yam VWW. Elucidation of the key role of Pt···Pt interactions in the directional self-assembly of plat-  
947 inum(II) complexes. *Proceedings of the National Academy of Sciences*. 2022; 119(12):e2116543119. doi:  
948 [10.1073/pnas.2116543119](https://doi.org/10.1073/pnas.2116543119).
- 949 **Zhu J**, Salvatella X, Robustelli P. Small molecules targeting the disordered transactivation domain of the andro-  
950 gen receptor induce the formation of collapsed helical states. *Nature Communications*. 2022; 13(1):6390.  
951 doi: [10.1038/s41467-022-34077-z](https://doi.org/10.1038/s41467-022-34077-z).
- 952 **Zhu S**. Validation of the Generalized Force Fields GAFF, CGenFF, OPLS-AA, and PRODRGFF by Testing Against  
953 Experimental Osmotic Coefficient Data for Small Drug-Like Molecules. *Journal of Chemical Information and*  
954 *Modeling*. 2019; 59(10):4239–4247. doi: [10.1021/acs.jcim.9b00552](https://doi.org/10.1021/acs.jcim.9b00552).