A perspective on the sharing of docking data

Samia Aci-Sèche, Stéphane Bourg, Pascal Bonnet, Joseph Rebehmed, Alexandre de Brevern, Julien Diharce

Title : Actual and global view of sharing molecular docking data

Samia Aci-Sèche, Stéphane Bourg, Pascal Bonnet, Joseph Rebehmed, Alexandre G. de Brevern and Julien Diharce*

Abstract

latter

INTRODUCTION

Drug discovery is nowadays one of the most prevalent fields in scientific research. Hence, the number of drugs proposed to the Food and Drug Administration (FDA) was 51 in 2021, with a 5-year average of 51, whereas it was almost twice less (24) ten years ago[1]. It may be emphasized that the role of the computational approaches is directly linked to the great improvement of the research and the environment around drug discovery during this last decade. Indeed, this period has also been marked by one of the most important technical revolutions in theoretical calculations and numerical simulations, triggered by the arrival of Graphical Process Unit (GPU) hardware for scientific calculations[2,3]. Previously, GPU were notably known to support intense calculation for 3D modelling or visualization.

By consequence, computer-aided drug design (CADD) has known a true expansion during the 2010's. Amongst 70 commercialized drugs using at least one computational approach, half of them has been made during this decade[4]. Indeed, computer techniques allow the proposal a large panel of possible models, from the molecular to the analytic level. Molecular modeling, using techniques such as homology modeling, molecular docking, pharmacophore mapping or molecular dynamics simulations, is used to investigate biological mechanisms associated to the drug, including binding mode, association/dissociation processes, conformational changes, stability of associations, etc[5–7]. In addition, other techniques, such as Machine Learning methods, are used for the prediction of specific properties such as activities, kinetics or ADME-Tox aspects, building models called Quantitative Structure Activities/Properties Relationship, (QSAR/QSPR) models[8,9]. Furthermore, since a decade, we assist to the emergence of Deep Learning approaches, powered by the increase of computational power and by the amount of available data[10].

Amongst those different methodologies, Virtual Screening (VS) is one of the most used in CADD to explore molecular databases and find interesting compounds for a considered target. It is the computational version of the experimental screening, which is by essence very expensive and time-consuming, but remains the experimental method of choice for hit identification and optimization. Virtual screening may be divided in two main approaches: the structure-based and ligand-based approaches. This last, requiring only the knowledge of molecules that bind the biological target of interest, was often used for molecular similarity, pharmacophore query or QSAR/QSPR modeling. But the most widespread method employed in the drug discovery field is a structure-based technique named molecular docking[4,11].
Molecular docking could be of two main natures: small molecule (ligand) to macromolecule or macromolecule to macromolecule, majority represented by ligand-protein and protein-protein docking. Even sharing the main theoretical principles, both are distinct methodologies because of the nature of the entities involved and of the complex interface[12,13]. In this article, we will only focus on the ligand-receptor paradigm of molecular docking.

As depicted earlier, the importance of molecular docking is considerable in drug discovery and has not more to be proved. By consequence, it is necessary to generate suitable docking data in order to have relevant results. However, a crucial question is often forgotten: their free availability to the scientific community. This question is not new and concerns every aspect of the Science and every scientist worldwide. It has been put forward in the article of Wilkinson et al. in 2016[14], by the publication of the FAIR (Findable, Accessible, Interoperable, Reusable) principles. This study is the cornerstone for a new way of thinking in Science, leading to the philosophy of sharing as soon as possible data of studies, with sufficient indications, precisions and clarity for a possible reuse. But in such context, the data to share are not limited to the result of docking campaigns nor to the docking parameters employed. To ensure a proper reuse as the reproducibility of these results, it is also necessary to provide the structural data used to establish the model of docking, before and after preparation, as well as the detailed protocols employed to prepare the structures or to rank and analyze the docked compounds.

The present article will summarize the current state of docking data, starting from the good practices for generating relevant docking results, and then depict the current state of the sharing of those data and what could be the possible improvements and prospects.

I) Generation of docking data: basis and reflexes

Generating docking results is not as simple as it seems. Firstly, the choice of a correct protein structure and a set of ligands is not trivial. Moreover, those elements need to be carefully and meticulously prepared.

a) Preparation of the protein for docking

Protein structures can be solved by experimental methods or predicted by means of computational approaches. The Protein Data Bank (PDB) is an on-line database (www.pdb.org) composed by almost two hundred thousand of protein structures, solved by experimental means such as X-Ray crystallography, NMR or cryo-EM techniques. Therefore, mining this database is the first thing to do when one is searching for a structure for docking experiment. But those structure needs to be prepared before the docking. First, structures can sometimes have some defects, especially the non-resolution of some protein portions. Generally, those parts are the most mobile parts in the structure, which are often coupled to non-truly important and functional regions, and are not always needed for the docking calculation. However, the completion of the protein is often mandatory to reach results with real meaning. It can be do with comparative modeling (for example, Modeller[15]), or Deep-Learning approaches which are very numerous now: AlphaFold[16], ColabFold[17], RoseTTAfold[18]… Depending of the missing part (meaning its location, its size or its importance), precaution must be taken when completing the protein structure.

Second, assigning the protonation state of residues is one of the most important things to realize before performing the docking calculation, based on the pH decided by the user. Indeed, charges must be put on every atom of the protein for a good conduction of the docking process. In this matter, some tools are very useful such as the H++[19] or ProPKA[20,21] servers or standalone programs like MCCE. Especially, the buried residues must be treated with care, because the local pH can dramatically vary from the physiological ones, induced by the fact that water molecules cannot reach those chemical functions, hidden by the protein fold. Consequently, the pKa of those amino acids will vary and with sometimes important consequences.

Finally, the structure of the protein deposited in the PDB database is only one rigid conformation, while proteins are dynamic macromolecules. Alternate stable conformations may exist for a binding site and an even limited change of the conformation of the binding cavity can lead to the recognition of new interacting molecules. Thus, considering several conformations of the protein, depending of course of the nature and the flexibility of the site, must be necessary in order to have the more rigorous results with docking methodologies. When other structures are not available, several methodologies such as molecular dynamics (MD) simulations[3] or normal mode analysis could be employed[22].

b) Selection and preparation of the ligand dataset

The second point to consider before the docking calculation itself is the selection and the preparation of the ligand dataset. The number of chemical databases, meaning a library which repertory several compounds, is nowadays quite enormous. Since 2016, a list of 117 databases has been set in the article of Sabe et al.[4], which have been consistently used during this 6-year period. We retrieved amongst them the most known ones, such as ZINC[23], ChEMBL[24], DrugBank[25], PubChem[27] or ChemDiv. Those databases could be commercial or public and list lots of information in addition to the name and structure of the molecule of interest. We may add to this list all the private chemical libraries that may be used for a particular project.

Generally, those databases are provided at 2D SDF format (Structural Data File), SMILES format (Simplified Molecular Input Line Entry Specification, in 1D) or as CVS file. However, some discrepancies could exist in those databases, despite the care of their curator to update and maintain their viability. Those mistakes can be originated from errors during the recording of the compounds, or more complex cases difficult to understand. In addition, sometimes the entry encompasses also other impurities, which couldn't have been separated from the molecule of interest.

Hence, some steps of preparation must be realized before the docking. The first step aims to filter the entries to remove the erroneous compounds and duplicates, and keep only one molecule by entry. Second, a conversion from 1D or 2D format to 3D coordinates must be carried on in order to generate compounds that can be used for docking calculations. The generation of 3D coordinates is done using 4 different steps: tautomer elucidation, hydrogen

atoms addition (in regard to the pH of the environment of the protein target), and definition of the several stereoisomers and the generation of stables conformers, with a particular attention to ring conformations. One must also notice that some steps could be avoided if the initial file possesses those information (especially the tautomer and the asymmetry of carbon atoms). . Several dedicated programs have been proposed in the literature for this purpose as VSprep or GypsumDL.

c) Docking process

Once protein target and chemical libraries prepared, they may be submitted to the docking process. Numerous software could be used for the docking, such as Autodock Vina[28], Glide[29], Gold[30], DOCK[31], rDock[32], PLANTS[33], etc. Numerous papers and reviews have already precisely described and efficiently compared the several docking software available and the power of the different algorithm used[11]. The last step consists in verification and selection of the most interesting ligands for the target. Lots of methods could be applied in order to sort the best molecules: ranking with scoring functions, RMSD comparison with already known ligands, ROC curve in order to see the efficiency of the calculation to separate active form inactive compounds, etc. Bender and collaborators provide a general and practical guide for the treatment of large-scale docking[34]. We have to notice that some software suites as MOE or the one proposed by Schrödinger (Maëstro and the associated stand-alone tools) offer the possibility to prepare target protein and chemical libraries, then performing docking calculation and analyze the results, in the same environment.

II) Statement of docking data sharing

a) Current Statement

Sharing data is nowadays a great stake to consider in modern science. Numerous concerns have been raised for researches that were nor reproductible, nor replicable. Konrad Hinsen in several papers has state that in this last decade. There is confusion between the model, the generated data and the software used to generate them, reinforcing the “push to results” way of thinking. In this paradigm, some researchers do not understand anymore how those software works and how the algorithms are implemented in it[35,36]. As a statement, it seems that the computational techniques tend to become a simple routine tool and not a research area

of development nor progress. One of the solutions proposed by Konrad Hinsen is to further develop the sharing of data of every kind, starting from input files and all the parameters from the software used to the raw results data. That is a statement shared by others group worldwide, that lead to the establishment of the FAIR principles in 2016 to guide the sharing of data in Science[14]. It consists in fact of several good practices for data sharing, from the format, the trackability, the reproducibility and the understandability of generated computational data. Destroying the barrier between the results described in a paper and the raw data, and provides access to the input files could pave the way to a better comprehension by every kind of user to those computational methodologies.

From those principles, many studies and many debates have raised on this subject, the majority focusing for the moment on the treatment of MD data. Thus, workshops[37], online servers for making simulation online[38,39] and also storage and listing service of MD trajectories[40,41] have been proposed in the last 5 years. The most known one is GPCRmd[42], an online database concerning the MD trajectories of all class of GPCR. For each PDB structure associated to a GPCR, this website contains molecular dynamics simulations starting from those structures. All the files associated to the creation of the simulations are available on a special page of the trajectory, in addition to a visual point of view of the trajectory itself, and a list of simple analysis that can be done on the data. Of course, the several data file (input, output, topology, trajectory) are also available for download.

One of the most important tools proposed nowadays is the publication of Simulation Foundry in 2020 by Gudrun Gygli and Juergen Pleiss[43]. This automated workflow on MD trajectory is the first that highlights the implementation of the FAIR principles within the treatment of the data. It allows the generation of the entire set of parameters for the calculation and making the calculation locally. In addition, the workflow comprises analyses of the trajectory generated and also a report in PDF format of the entire protocol and results.

Regarding docking data, unfortunately, few progresses have been made about their sharing to the scientific community. To our knowledge, no solutions, such as the ones already existing for MD data, have been currently proposed for the sharing of docking ones with the community. However, it exists some resources online to help docking users in their processes to generate results. We can for example cite the PDBind database[44], which list several protein-ligand structures with known experimental affinities, but also LIT-PCBA, a curated

dataset for virtual screening and machine learning (ref Rognan) and the DUD-E, a dataset comprising active compounds by targets and associated decoys(ref). With this dataset, one could test his own protocol in order to estimate his robustness. Furthermore, a new breakthrough on the sharing of data is the proposal of entire datasets in ready-to-dock format, following the most classical parameters of physiological environment. This is made for example by ZINC database. Obviously, initial dataset could be download by the users that need to prepare their dataset in another way. Following this lead, other datasets, comprising even billions of molecules are now emerging. The most notable one is VirtualFlow[45] which encompass a workflow for ligand preparation and a ligand dataset comprising 1.4 billion commercially available compounds in ready-to-dock format. The workflow is compatible with lots of docking software and scoring functions, and respects all along the process the FAIR principles. Both ligand dataset for docking calculations by the user, and entire workflow of the VirtualFlow program, are available freely. In addition, through an open-access GNU GPL license, everyone can contribute to improve this workflow. The workflow is available on GitHub (http://www.github.com), a platform of program collaborative development and sharing, with a versioning process very powerful, that can allow following every modification bring on the program. Since the last decade, GitHub has become the standard platform to propose and share online all kind of program, scripts, with an idea of sharing and improvements for and by the concerned community.

b) Discussion and prospects for a better sharing of docking data

What can be done about the sharing of docking data? Of course, normalized protocols, such as the one proposed by VirtualFlow or the one described in the paper of Bender et al.[34], could be proposed any further, for example specialized on some target families of importance (GPCR, kinases protein, etc.), as what was already made for MD data. In addition, the results of large-scale docking campaigns are for now cruelly missing for sharing and availability. One could imagine a specialized repository containing the results of large-scale docking, such as ZINC database or others, against one or several members of protein family. Indeed, we can assume that numerous docking processes against popular targets, with therapeutic potential, are made by researcher around the world, without knowing that others have already realized their virtual screening before. Such sharing using a dedicated infrastructure to list and register the several docking campaign could be of great interest for the entire community of drug

design researchers. The results could be download by the user, and ranked using several criteria, such as different scoring functions, or experimental data if they exist.

However, the principal problems of this kind of architecture that can be easily predict is the issue of the storage and the size of data. Obviously, dataset of millions, even billions of compounds take large space for the storage. Indeed, infrastructure keeping results of docking campaign must take into account several different docking poses for each ligand conformations in a receptor, making the data significantly heavier when the dataset comprises million or billions of compounds. This problematic has been relevantly discussed in the paper of Hospital et al.[46] This interesting review was focused on the MD data, which are obviously more affected by the size issue due to the inherent nature of the data. They highlight two important things for the storage of data: the need of standard format for the files and also compression processes of the files in order to gain space. They also underlined the need to a long-term sustainability plan in order to storage with safety on specific repositories the data of MD data.

Obviously, even if the scale is very different compared to the case of MD calculations, those issues about the size of data and their long-term sustainability remain relevant for docking data and must be considered in order to improve the clarity and the application of FAIR principles on those kinds of data.
It is interesting even to see that Data in Brief shares some papers on MD results and fewer on drug design researches, and most of these last come from experiments (70 in 2021 and 2022), and not from *in silico* studies (0)[47,48].

III) Concluding remarks

CADD is now an important field in life science research, thanks to the technological progress made during this last decade. In particular, virtual screening using docking protocols remains the main methodology for drug discovery processes. Our goal here was double: i) to sensibilize users of docking methodologies to the main points in order to create relevant docking data, from preparation of inputs to verification of the results and ii) what is done now to the sharing of data regarding docking calculations and what can be done in the future.

As we have described, there is actually a lot of online resources for assistance and help to the users, starting from the large number of chemiotheques, to the several software and workflow that can be used for making docking campaign, even some respecting the principles of FAIR data for the sharing and a good understanding. However, there are only scarce elements about the sharing itself of docking results. We believe that, even with the issues that can from this sharing and storage, the entire CADD scientific community could benefit from the creation of such online storage of docking campaign results and freely shared to everyone. However, questions of data size and sustainability in long-term are two major bottlenecks to creation of such infrastructure.

Here are some good practices guidelines to follow for the sharing of such data:

(i) precise and clear description of the structure used (with or without optimization, but clearly noted), with date and place. For example, structure of the PDB supplemented by Modeller version 10.1 by taking another local support, with optimization of the loops (and therefore the Python code used).

(ii) If a molecular dynamics approach or other method are used to have conformers, the set of parameters used for the simulation and also for the structure clustering. It is advisable to have at least the different configurations and if possible the trajectories (with the parameter files).

(iii) a precise description of the origin of the libraries, the protocol and the software used with their versions for the 3D conversion, the refinements used to generate the entire dataset.

(iv) The same for docking with the different scoring approaches used.

The difficulty here is identical to that of MD simulations results, the place that raw data can take.

Nevertheless, efforts must be done on the application of FAIR principals for docking data, in order to favorize a better connection between the molecular model of protein-ligand interactions and biological activities and by extension, the experimental methodologies. The availability of such data freely could also be used for the development of deep learning approaches applied to docking and analysis of protein-ligand interactions, such as recent papers just begin to realize[49].

References:

[1] A. Mullard, 2021 FDA approvals, Nat Rev Drug Discov. 21 (2022) 83–88. https://doi.org/10.1038/d41573-022-00001-9.

[2] B. Kohnke, C. Kutzner, H. Grubmüller, A GPU-Accelerated Fast Multipole Method for GROMACS: Performance and Accuracy, J. Chem. Theory Comput. 16 (2020) 6938–6949. https://doi.org/10.1021/acs.jctc.0c00744.
[3] T.-S. Lee, D.S. Cerutti, D. Mermelstein, C. Lin, S. LeGrand, T.J. Giese, A. Roitberg, D.A. Case, R.C. Walker, D.M. York, GPU-Accelerated Molecular Dynamics and Free Energy Methods in Amber18: Performance Enhancements and New Features, J. Chem. Inf. Model. 58 (2018) 2043–2050. https://doi.org/10.1021/acs.jcim.8b00462.
[4] V.T. Sabe, T. Ntombela, L.A. Jhamba, G.E.M. Maguire, T. Govender, T. Naicker, H.G. Kruger, Current trends in computer aided drug design and a highlight of drugs discovered via computational techniques: A review, European Journal of Medicinal Chemistry. 224 (2021) 113705. https://doi.org/10.1016/j.ejmech.2021.113705.
[5] J. Diharce, E. Bignon, S. Fiorucci, S. Antonczak, Exploring Dihydroflavonol-4-Reductase Reactivity and Selectivity by QM/MM-MD Simulations, ChemBioChem. 23 (2022) e202100553. https://doi.org/10.1002/cbic.202100553.
[6] H. Fu, H. Zhang, H. Chen, X. Shao, C. Chipot, W. Cai, Zooming across the Free-Energy Landscape: Shaving Barriers, and Flooding Valleys, J. Phys. Chem. Lett. 9 (2018) 4738–4745. https://doi.org/10.1021/acs.jpclett.8b01994.
[7] N. Yasuo, M. Sekijima, Improved Method of Structure-Based Virtual Screening via Interaction-Energy-Based Learning, J. Chem. Inf. Model. 59 (2019) 1050–1061. https://doi.org/10.1021/acs.jcim.8b00673.
[8] T. Huang, G. Sun, L. Zhao, N. Zhang, R. Zhong, Y. Peng, Quantitative Structure-Activity Relationship (QSAR) Studies on the Toxic Effects of Nitroaromatic Compounds (NACs): A Systematic Review, International Journal of Molecular Sciences. 22 (2021). https://doi.org/10.3390/ijms22168557.
[9] K. Heikamp, J. Bajorath, Support vector machines for drug discovery, Expert Opinion on Drug Discovery. 9 (2014) 93–104. https://doi.org/10.1517/17460441.2014.866943.
[10] J. Jiménez-Luna, F. Grisoni, G. Schneider, Drug discovery with explainable artificial intelligence, Nature Machine Intelligence. 2 (2020) 573–584. https://doi.org/10.1038/s42256-020-00236-4.
[11] X.-Y. Meng, H.-X. Zhang, M. Mezei, M. Cui, Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery, CAD. 7 (2011) 146–157. https://doi.org/10.2174/157340911795677602.
[12] L.L. Conte, C. Chothia, J. Janin, The atomic structure of protein-protein recognition sites11Edited by A. R. Fersht, Journal of Molecular Biology. 285 (1999) 2177–2198. https://doi.org/10.1006/jmbi.1998.2439.
[13] R.P. Bahadur, M. Zacharias, The interface of protein-protein complexes: Analysis of contacts and prediction of interactions, Cellular and Molecular Life Sciences. 65 (2008) 1059–1072. https://doi.org/10.1007/s00018-007-7451-x.
[14] M.D. Wilkinson, M. Dumontier, Ij.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data. 3 (2016) 160018. https://doi.org/10.1038/sdata.2016.18.
[15] B. Webb, A. Sali, Comparative Protein Structure Modeling Using MODELLER, Current Protocols in Bioinformatics. 54 (2016) 5.6.1-5.6.37. https://doi.org/10.1002/cpbi.3.

[16] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S.A.A. Kohl, A.J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A.W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold, Nature. 596 (2021) 583–589. https://doi.org/10.1038/s41586-021-03819-2.

[17] M. Mirdita, K. Schütze, Y. Moriwaki, L. Heo, S. Ovchinnikov, M. Steinegger, ColabFold: making protein folding accessible to all, Nature Methods. 19 (2022) 679–682. https://doi.org/10.1038/s41592-022-01488-1.

[18] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G.R. Lee, J. Wang, Q. Cong, L.N. Kinch, R.D. Schaeffer, C. Millán, H. Park, C. Adams, C.R. Glassman, A. DeGiovanni, J.H. Pereira, A.V. Rodrigues, A.A. van Dijk, A.C. Ebrecht, D.J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M.K. Rathinaswamy, U. Dalwadi, C.K. Yip, J.E. Burke, K.C. Garcia, N.V. Grishin, P.D. Adams, R.J. Read, D. Baker, Accurate prediction of protein structures and interactions using a three-track neural network, Science. 373 (2021) 871–876. https://doi.org/10.1126/science.abj8754.

[19] R. Anandakrishnan, B. Aguilar, A.V. Onufriev, H++ 3.0: automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations, Nucleic Acids Research. 40 (2012) W537–W541. https://doi.org/10.1093/nar/gks375.

[20] M.H.M. Olsson, C.R. Søndergaard, M. Rostkowski, J.H. Jensen, PROPKA3: Consistent Treatment of Internal and Surface Residues in Empirical pKa Predictions, J. Chem. Theory Comput. 7 (2011) 525–537. https://doi.org/10.1021/ct100578z.

[21] C.R. Søndergaard, M.H.M. Olsson, M. Rostkowski, J.H. Jensen, Improved Treatment of Ligands and Coupling Effects in Empirical Calculation and Rationalization of pKa Values, J. Chem. Theory Comput. 7 (2011) 2284–2295. https://doi.org/10.1021/ct200133y.

[22] J.A. Bauer, J. Pavlović, V. Bauerová-Hlinková, Normal Mode Analysis as a Routine Part of a Structural Investigation, Molecules. 24 (2019). https://doi.org/10.3390/molecules24183293.

[23] J.J. Irwin, K.G. Tang, J. Young, C. Dandarchuluun, B.R. Wong, M. Khurelbaatar, Y.S. Moroz, J. Mayfield, R.A. Sayle, ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery, J. Chem. Inf. Model. 60 (2020) 6065–6073. https://doi.org/10.1021/acs.jcim.0c00675.

[24] D. Mendez, A. Gaulton, A.P. Bento, J. Chambers, M. De Veij, E. Félix, M.P. Magariños, J.F. Mosquera, P. Mutowo, M. Nowotka, M. Gordillo-Marañón, F. Hunter, L. Junco, G. Mugumbate, M. Rodriguez-Lopez, F. Atkinson, N. Bosc, C.J. Radoux, A. Segura-Cabrera, A. Hersey, A.R. Leach, ChEMBL: towards direct deposition of bioassay data, Nucleic Acids Research. 47 (2019) D930–D940. https://doi.org/10.1093/nar/gky1075.

[25] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, M. Wilson, DrugBank 5.0: a major update to the DrugBank database for 2018, Nucleic Acids Research. 46 (2018) D1074–D1082. https://doi.org/10.1093/nar/gkx1037.

[26] N. Huang, B.K. Shoichet, J.J. Irwin, Benchmarking Sets for Molecular Docking, J. Med. Chem. 49 (2006) 6789–6801. https://doi.org/10.1021/jm0608356.

[27] S. Kim, Exploring Chemical Information in PubChem, Current Protocols. 1 (2021) e217. https://doi.org/10.1002/cpz1.217.

[28] J. Eberhardt, D. Santos-Martins, A.F. Tillack, S. Forli, AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings, J. Chem. Inf. Model. 61 (2021) 3891–3898. https://doi.org/10.1021/acs.jcim.1c00203.
[29] R.A. Friesner, J.L. Banks, R.B. Murphy, T.A. Halgren, J.J. Klicic, D.T. Mainz, M.P. Repasky, E.H. Knoll, M. Shelley, J.K. Perry, D.E. Shaw, P. Francis, P.S. Shenkin, Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy, J. Med. Chem. 47 (2004) 1739–1749. https://doi.org/10.1021/jm0306430.
[30] P.A. Greenidge, R.A. Lewis, P. Ertl, Boosting Pose Ranking Performance via Rescoring with MM-GBSA, Chemical Biology & Drug Design. 88 (2016) 317–328. https://doi.org/10.1111/cbdd.12763.
[31] M.M. Mysinger, B.K. Shoichet, Rapid Context-Dependent Ligand Desolvation in Molecular Docking, J. Chem. Inf. Model. 50 (2010) 1561–1573. https://doi.org/10.1021/ci100214a.
[32] S. Ruiz-Carmona, D. Alvarez-Garcia, N. Foloppe, A.B. Garmendia-Doval, S. Juhos, P. Schmidtke, X. Barril, R.E. Hubbard, S.D. Morley, rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids, PLOS Computational Biology. 10 (2014) e1003571. https://doi.org/10.1371/journal.pcbi.1003571.
[33] O. Korb, T. Stützle, T.E. Exner, Empirical Scoring Functions for Advanced Protein−Ligand Docking with PLANTS, J. Chem. Inf. Model. 49 (2009) 84–96. https://doi.org/10.1021/ci800298z.
[34] B.J. Bender, S. Gahbauer, A. Luttens, J. Lyu, C.M. Webb, R.M. Stein, E.A. Fink, T.E. Balius, J. Carlsson, J.J. Irwin, B.K. Shoichet, A practical guide to large-scale docking, Nat Protoc. 16 (2021) 4799–4832. https://doi.org/10.1038/s41596-021-00597-z.
[35] K. Hinsen, Verifiability in computer-aided research: the role of digital scientific notations at the human-computer interface, PeerJ Computer Science. 4 (2018) e158. https://doi.org/10.7717/peerj-cs.158.
[36] K. Hinsen, Computational science: shifting the focus from tools to models [version 2; peer review: 2 approved], F1000Research. 3 (2014). https://doi.org/10.12688/f1000research.3978.2.
[37] M. Abraham, R. Apostolov, J. Barnoud, P. Bauer, C. Blau, A.M.J.J. Bonvin, M. Chavent, J. Chodera, K. Čondić-Jurkić, L. Delemotte, H. Grubmüller, R.J. Howard, E.J. Jordan, E. Lindahl, O.H.S. Ollila, J. Selent, D.G.A. Smith, P.J. Stansfeld, J.K.S. Tiemann, M. Trellet, C. Woods, A. Zhmurov, Sharing Data from Molecular Simulations, J. Chem. Inf. Model. 59 (2019) 4093–4099. https://doi.org/10.1021/acs.jcim.9b00665.
[38] G. Bayarri, A. Hospital, M. Orozco, 3dRS, a Web-Based Tool to Share Interactive Representations of 3D Biomolecular Structures and Molecular Dynamics Trajectories, Frontiers in Molecular Biosciences. 8 (2021). https://www.frontiersin.org/articles/10.3389/fmolb.2021.726232.
[39] M. Kampfrath, R. Staritzbichler, G.P. Hernández, A.S. Rose, J.K.S. Tiemann, G. Scheuermann, D. Wiegreffe, P.W. Hildebrand, MDsrv: visual sharing and analysis of molecular dynamics simulations, Nucleic Acids Research. 50 (2022) W483–W489. https://doi.org/10.1093/nar/gkac398.
[40] T.D. Newport, M.S.P. Sansom, P.J. Stansfeld, The MemProtMD database: a resource for membrane-embedded protein structures and their lipid interactions, Nucleic Acids Research. 47 (2019) D390–D397. https://doi.org/10.1093/nar/gky1047.
[41] P.W. Hildebrand, A.S. Rose, J.K.S. Tiemann, Bringing Molecular Dynamics Simulation Data into View, Trends in Biochemical Sciences. 44 (2019) 902–913. https://doi.org/10.1016/j.tibs.2019.06.004.

[42] I. Rodríguez-Espigares, M. Torrens-Fontanals, J.K.S. Tiemann, D. Aranda-García, J.M. Ramírez-Anguita, T.M. Stepniewski, N. Worp, A. Varela-Rial, A. Morales-Pastor, B. Medel-Lacruz, G. Pándy-Szekeres, E. Mayol, T. Giorgino, J. Carlsson, X. Deupi, S. Filipek, M. Filizola, J.C. Gómez-Tamayo, A. Gonzalez, H. Gutiérrez-de-Terán, M. Jiménez-Rosés, W. Jespers, J. Kapla, G. Khelashvili, P. Kolb, D. Latek, M. Marti-Solano, P. Matricon, M.-T. Matsoukas, P. Miszta, M. Olivella, L. Perez-Benito, D. Provasi, S. Ríos, I. R. Torrecillas, J. Sallander, A. Sztyler, S. Vasile, H. Weinstein, U. Zachariae, P.W. Hildebrand, G. De Fabritiis, F. Sanz, D.E. Gloriam, A. Cordomi, R. Guixà-González, J. Selent, GPCRmd uncovers the dynamics of the 3D-GPCRome, Nature Methods. 17 (2020) 777–787. https://doi.org/10.1038/s41592-020-0884-y.

[43] G. Gygli, J. Pleiss, Simulation Foundry: Automated and F.A.I.R. Molecular Modeling, J. Chem. Inf. Model. 60 (2020) 1922–1927. https://doi.org/10.1021/acs.jcim.0c00018.

[44] Z. Liu, M. Su, L. Han, J. Liu, Q. Yang, Y. Li, R. Wang, Forging the Basis for Developing Protein–Ligand Interaction Scoring Functions, Acc. Chem. Res. 50 (2017) 302–309. https://doi.org/10.1021/acs.accounts.6b00491.

[45] C. Gorgulla, A. Boeszoermenyi, Z.-F. Wang, P.D. Fischer, P.W. Coote, K.M. Padmanabha Das, Y.S. Malets, D.S. Radchenko, Y.S. Moroz, D.A. Scott, K. Fackeldey, M. Hoffmann, I. Iavniuk, G. Wagner, H. Arthanari, An open-source drug discovery platform enables ultra-large virtual screens, Nature. 580 (2020) 663–668. https://doi.org/10.1038/s41586-020-2117-z.

[46] A. Hospital, F. Battistini, R. Soliva, J.L. Gelpí, M. Orozco, Surviving the deluge of biosimulation data, WIREs Comput Mol Sci. 10 (2020). https://doi.org/10.1002/wcms.1449.

[47] S. Tanwar, P. Auberger, G. Gillet, M. DiPaola, K. Tsaioun, B.O. Villoutreix, A new ChEMBL dataset for the similarity-based target fishing engine FastTargetPred: Annotation of an exhaustive list of linear tetrapeptides, Data in Brief. 42 (2022) 108159. https://doi.org/10.1016/j.dib.2022.108159.

[48] M. Tarpley, H. Oladapo, T.B. Caligan, R.U. Onyenwoke, K.P. Williams, Data supporting a pilot high-throughput screen of a drug library for identification of DYRK1A inhibitors and high-content imaging analysis of identified harmine analogs, Data in Brief. 37 (2021) 107189. https://doi.org/10.1016/j.dib.2021.107189.

[49] F. Gentile, J.C. Yaacoub, J. Gleave, M. Fernandez, A.-T. Ton, F. Ban, A. Stern, A. Cherkasov, Artificial intelligence–enabled virtual screening of ultra-large chemical libraries with deep docking, Nature Protocols. 17 (2022) 672–697. https://doi.org/10.1038/s41596-021-00659-2.