Ten Simple Rules for Sharing Experimental and Clinical Data with the Modeling Community

Matthias König[1*], Jan Grzegorzewski[1], Martin Golebiewski[2], Henning Hermjakob [3,4], Mike Hucka[5], Brett Olivier[6], Sarah M. Keating[7], David Nickerson [8], Falk Schreiber[9,10], Rahuman Sheriff [3], Dagmar Waltemath[11]

1 Institute for Theoretical Biology (ITB), Humboldt University Berlin, Germany
2 Heidelberg Institute for Theoretical Studies (HITS), Heidelberg, Germany
3 European Bioinformatics Institute, European Molecular Biology Laboratory (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge, United Kingdom
4 State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Lifeomics, National Center for Protein Sciences Beijing, 102206 Beijing, China
5 California Institute of Technology, USA
6 Systems Biology Lab, Amsterdam Institute of Molecular and Life Sciences (AIMMS), Vrije Universiteit Amsterdam, Amsterdam, Netherlands
7 University College London, London, United Kingdom
8 Auckland Bioengineering Institute, University of Auckland, New Zealand
9 Department of Computer and Information Science, University of Konstanz, Germany
10 Faculty of Information Technology, Monash University, Australia
11 Medical Informatics Laboratory, University Medicine Greifswald, Greifswald, Germany

Keywords: data sharing, FAIR
* koenigmx@hu.berlin.de

Abstract

Science continues to become more interdisciplinary and to involve increasingly complex data sets. Many projects in the biomedical and health related sciences adhere to the principles of FAIR data sharing, or aim to follow them. Data sharing has been proven to foster collaboration, to lead to better research outcomes, and to help ensure reproducibility of results. Data generated in biomedical and health research are specific in the sense that they are heterogeneous, often big, and highly sensitive in terms of data protection needs and contextuality. Data sharing has to respect these features, but at the same time advances in medical therapy and treatment are time-critical. Modeling and simulation of biomedical processes have become an established tool, and a global community has been developing algorithms, methodologies, and standards for applying biomedical simulation models in clinical research. However, it can be difficult for clinician scientists to follow the specific rules and recommendations for FAIR data sharing within the domain. With this paper, we aim to clarify the standard workflow for sharing experimental and clinical data with the simulation modeling community. By following these recommendations, data sharing will be improved, collaborations will become more effective, and the FAIR publication and subsequent reuse of data will become possible at the level of quality necessary in biomedical and health related sciences.

 doi:10.20944/preprints202108.0303.v1

Author summary

Data sharing improves the quality of scientific reporting, increases scientific outcome, collaboration, publication rates, and visibility, and is a fundamental part of most research projects. In this paper, we outline 10 simple rules for sharing experimental and clinical data with the simulation modeling community.

Introduction

Data provides the evidence for scientific knowledge and science is built on data [1]. Data sharing improves the quality of scientific reporting and increases scientific outcome, collaboration, publication rates, and visibility. Hence, it is beneficial to researchers, society, and funders. Data sharing is part of good scientific practice, enables data reuse, and fosters the scientific discovery process [2]. In addition, when cited properly, researchers get the credit they deserve for the data they generate [3]. Among the essential factors for catalyzing data sharing are (i) clear policies from funders, institutions, journals, and research communities; (ii) credit and incentives for data publication; (iii) explicit funding for data management, data sharing, and data publishing; (iv) practical help with organizing data, finding appropriate repositories and simpler ways of sharing; and (v) training and education in research data management [4]. Within this work we address the last two points.

Computational biology and medicine are fields that are highly dependent on the availability of research data. This requirement is independent of methodology, modeling domain, or complexity of the research object. Data sharing between experimentalists, clinicians and modelers is an essential part of most investigations. Data is needed during model construction (parametrization), in which a subset of the data (training data) is used to calibrate model parameters and model behavior; and during model evaluation (validation), in which a different subset of the data is used to evaluate the model performance by comparing model predictions against test data.

However, despite the efforts of the scientific communities to provide guidelines and tools for open and reproducible science, most data is difficult for modelers to use. One reason is a lack of data accessibility, with researchers rarely making their data available in a manner directly accessible by modelers, despite widespread support from funding agencies, scientific journals, and policy makers [2,5]. A further reason is a lack of data interoperability, with accessible data being difficult to integrate with computational models due to technical challenges with the shared data, for instance poorly annotated data hidden in PDF documents [1]. Many challenges relate to poor data FAIRness, i.e., data not being findable, accessible, interoperable, or reusable [6]. However, simply making data FAIR does not guarantee the data is of high quality nor that it can be used in computational modeling.

It is surprising that despite the importance of data sharing, no guideline or best practice for sharing research data with the modeling community exists. Members of the COMBINE community [7] discussed these problems during their annual meetings [8] and collected best practices. As a result, we provide ten simple rules (Figure 1) on how to share data for computational modeling, addressing issues of data accessibility and quality as well as providing guidelines for the research community and material for education.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 August 2021 doi:10.20944/preprints202108.0303.v1

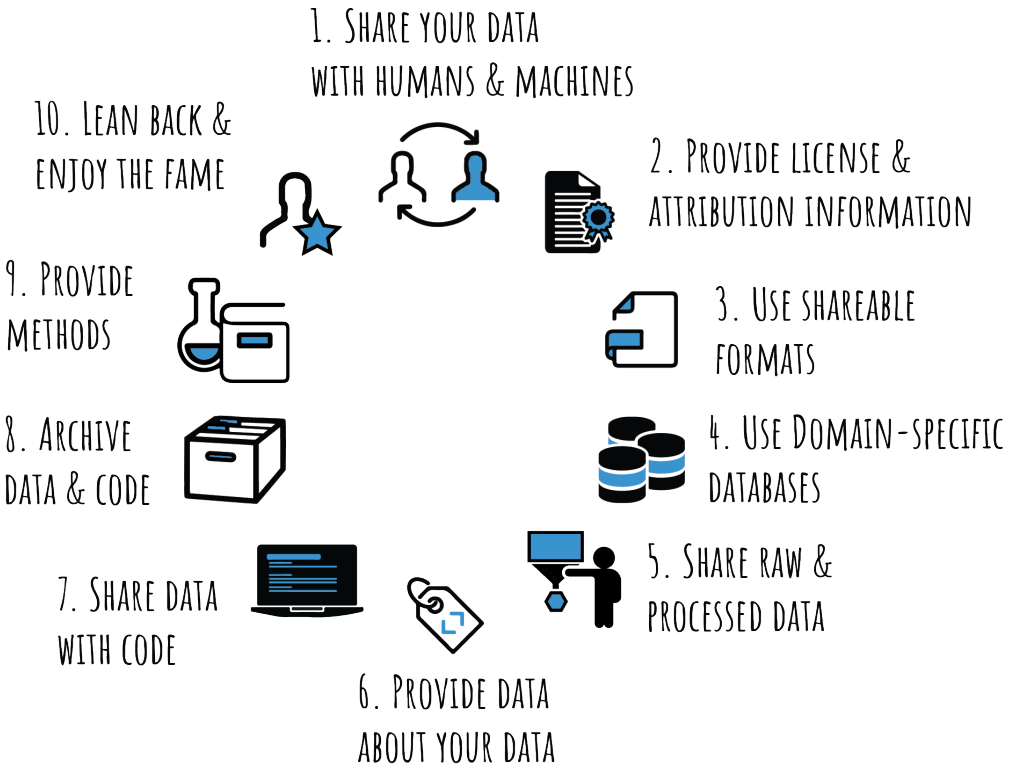

Fig 1. Ten simple rules for sharing data in life science for computational biology and medicine. (1) Share your data in a human and machine accessible manner. (2) Disseminate license and attribution information with data. (3) Store your data in an open standardized format and check that the format is used correctly. (4) Use domain specific formats and databases. (5) Share raw and processed data. Share as much as possible, but not more. (6) Add metadata (data about your data) to make data findable and comprehensible. (7) Share code and workflows for data processing with the data. (8) Archive data and code in a persistent manner accessible via a DOI. (9) Disseminate information on experimental and computational methods and protocols with the data. (10) Lean back and enjoy the additional credit and citations.

Rules

1. Share your data with humans and machines

The first rule of data sharing is to actually make an effort to share your data. Sharing data means making it accessible (online) to both humans and machines via a data repository. Humans should be able to access your data via a web browser and machines via web services and/or persistent links. Use a repository that minimizes hurdles to data access, i.e., if possible avoid resources that require accounts or registration for accessing data or are only accessible to a limited community (e.g., only within an institution). Remember that science is a global endeavor and data should be accessible for researchers worldwide, not only from one country or region. Providing open access to data is not only important for reuse and data mining but to confirm that results presented in a publication are truly based on actual data [9]. Important criteria for choosing a data repository are long-term availability and acceptance within the community. The longer a repository exists, and the more users it has, the better is the support and ecosystem of software tools. Several platforms support scientists in finding

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 August 2021 doi:10.20944/preprints202108.0303.v1

the best repository for their needs, including criteria such as supported data formats (e.g. https://fairsharing.org), archiving services, services for Digital Object Identifiers (DOI), and choice of licenses (e.g. https://www.re3data.org/). Note that there is a clear association between articles that include a data availability statement containing a link to a repository, which have up to 25% higher citation impact on average [3]. So what data should you share? As a rule of thumb, data sharing should be 'as open as possible, as closed as necessary'. For most data, this translates to sharing your data set openly without any restrictions. As a side note, you often share your data with your future self. Hence you should make everything accessible which you will require to reproduce and build on your results.

2. Provide license and attribution information

An often overlooked issue is missing licensing information for a published data set. If you own a specific data set, you have to give explicit rights for access or reuse by others, otherwise all rights are reserved. If your data comes without a license and people are really interested in it, they may request your permission to clarify their rights. However, often people may simply wander off or decide to generate their own data instead of reusing yours. Licenses allow you to explicitly and specifically grant permission to reuse. You can clearly state under which conditions reuse, redistribution, and possibly modifications to your data are allowed. You can do so by referring to predefined standard licenses. You can choose from a variety of Open Data Licenses, i.e. using one of the Creative Commons (CC) licenses. But what license should you choose? We advise you to make your research data accessible under the least restrictive (but compliant) license to allow the widest possible reuse. For example, the Creative Commons license CC-0 hands your data over to the public domain and allows for the broadest reuse, for instance in a context like data aggregation. CC-0 has many benefits for individuals and society with minimal implications for authors [10]. CC-BY provides all the openness but requires attribution, i.e. citation of the primary source, which is especially important for researchers. It is important to note that CC-0 does not mean that you don't request citation, it only means you allow re-use in contexts like data aggregation where attribution might be difficult. One of the most famous examples of open biomedical data is the human genome.

Equally important to the license is the information on how to correctly attribute the data creators. Depending on the data set this may mean to acknowledge others, cite the data set, or to offer co-authorship.

License and attribution information must be distributed with the data. It is advised to provide a human-readable description as well as computer-readable metadata. This includes a copy of or reference to the actual license. License and attribution should clearly be stated in the data description (see also rule 6). Only then can the license information always be transported with the content or data. When sharing data via a database, additional licenses and attributions may apply to the data set.

3. Use shareable formats

To make data useful it should be provided in interoperable and open machine readable formats. Formats should easily be parsable, not require any special software or license, and be supported by a wide range of tools and programming languages. Examples for open formats are JSON, CSV, YAML, XML or HDF5. Interoperable data formats can easily be integrated into modeling workflows. i.e., a CSV format is generally much easier to process than proprietary formats. Data formats should be text-based instead of using binary formats to allow for version control. Version control eases collaborative work as it makes changes on the data visible and trackable. A minimal requirement for

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 August 2021 doi:10.20944/preprints202108.0303.v1

data to be interoperable is that the data is both syntactically parsable and semantically understandable according to the respective standard. For example, you should ensure that there is no issue with the data files, and if possible perform structural checks and content checks. Structural checks ensure that there are no empty rows, no blank headers, etc. Content checks ensure that the values are of the correct types ('string', 'number', 'date', etc.), that their format is valid ('biological database identifiers must match a certain pattern'), and that constraints are respected ('age must be a number greater than 18'). Domain-specific formats often have associated validators, or simple mechanisms for validation (e.g., using XML or JSON schema files). Many communities develop their own domain-specific standards, which should be used whenever possible (see also rule 4).

4. Use domain-specific databases

It is recommended to use domain-specific repositories and data formats whenever possible, as this will greatly simplify integration of your data with other data sets, software tools, and modeling workflows. Examples of such domains are data of genomic sequences, proteins and protein structures, or metabolomic or transcriptomic data (see Table 1). Domain-specific databases are highly relevant for the findability of your data set because they offer an entry point for data search. In addition, these databases are often integrated with other domain-specific tools and workflows. Libraries exist for working with these formats and databases. Compliance of submitted data to the relevant reporting standards promotes consistent and adequate data description, thorough data validation, data discoverability, data reproducibility, data interoperability, and (re)usability. General-purpose repositories such as BioStudies [11], Dryad, Figshare Figshare, Zenodo Zenodo, or Github provide a solution to share the 'unstructured' data that does not fit into specialized repositories [11]. Your specific domain may have its own repository; the time to investigate is well-spent.

Domain	Database	Formats
Metabolomics	MetaboLights [12]	Spectral files
Sequence data	European Nucleotide Archive (ENA) [13]	Feature table
Proteomics	PRotein IDEntification database (PRIDE) [14] Proteome Xchange [15]	mzIdentML mzTab
Gene expression & functional genomics	ArrayExpress [16] Gene Expression Omnibus (GEO) [17]	MAGE-TAB
Protein structures	The Protein Data Bank (PDB) [18] PDBe [19]	PDB file format
Pharmacokinetics	Pharmacokinetics Database (PK-DB) [20]	-
Biological data (unstructured; multi-omics)	BioStudies [11]	-

Table 1. Examples of domain-specific formats and databases. For a good entry point see the ELIXIR (the European Infrastructure for data in the life sciences) recommendations on core data resources and deposition databases as well as the FAIRSHARING collections on standards and databases.

5. Share all raw and processed data

Publish all relevant data as raw data, not just aggregated and highly processed data sets. For example, providing a figure is not the same as sharing the data points. Sharing data for a plot means to provide the underlying raw data and processed data used to compile the figure. We observe that publications in biomedical journals often

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 August 2021 doi:10.20944/preprints202108.0303.v1

contain highly processed data depicted in figures or tables, but lack supplementary material or references to data sets (ideally published in data repositories). Data points in these figures mostly consist of mean or pooled data and error measurements like standard errors or standard deviations. For computational modeling, such pooled and group data are often not (very) useful, particularly if large inter-individual variability exists between the different individual subjects/samples measured and individual data points are not normally distributed. Many algorithms only work with individual data sets, pooled data is the same as no data at all for such applications (e.g. individual-based modeling or parameter fitting). In the context of time course data for ODE based models, to give a specific example, the time courses are very heterogeneous and the mean from individuals can be misleading. To make the data useful for modeling raw and processed data must be shared for individual measurements and subjects, and figures and tables should contain individual data in addition to grouped or pooled data. Often crucial information for modeling and data integration is not relevant for the primary publication and never reported (e.g., body weight, age, sex). In most cases, the data is used in completely new contexts than what the original data creator anticipated. Sharing as much as possible extends the possibilities of subsequent analysis and data integration and makes the data set much more valuable.

However, you should always be careful what raw data to share and not to share. In the context of biomedical research, the protection of patient-derived data is of highest priority, and legal matters must always be obeyed. For example, all sensitive data must be removed from data sets, patient data must be anonymized or at least pseudonymized, and data that would allow re-identification of patients must be removed from data sets. This includes, for instance, genetic information or data about rare diseases.

6. Provide data about your data

To publish FAIR data, it is necessary to clearly state what information is contained in each of the data items, e.g. what has been measured in a specific variable. Metadata (data describing the data) puts the data into context using biological, medical, or computational ontologies and mapping information in the data set to database identifiers. Metadata adds a semantic layer (experimental, biologically, environmentally, etc.) and allows others and your future self to interpret your data. Metadata improves findability as semantic information can additionally be indexed and then used for search and filter functions. One example of a crucial metadata item for computational models is unit information. Units should be defined as SI units, e.g., providing an insulin concentration in pmole/ml is much more helpful than in international units (IU). Adding provenance information and information on the context under which a data set is valid/applicable can be very helpful.

7. Share data with code

Data is often highly processed, and only the results are shown in figures and tables. However, to enable reproducing analysis results, one must be able to apply an identical analysis pipeline. Software tools, libraries, and workflows change over time, and this often leads to changes in results, due to different parameters, new algorithms, or simple implementation errors. Your shared data should therefore also contain code and workflows used in data processing, or at least clearly state the used software, its version, and methods. An example is RNAseq data with raw data being the raw counts, whereas the processed and analyzed data is often something like differential gene expression between conditions. For reproducibility of the analysis, the workflow for processing the data should be provided as code. The ideal case is if the complete code which generated the figures from the raw data is provided. This allows us to easily update the analysis

 doi:10.20944/preprints202108.0303.v1

pipeline and reuse the pipeline if additional data sets are generated (which is often the case for validation of computational models).

8. Archive data and code

An important aspect of findable and accessible data for use in computational models is to provide a standard identification mechanism to make data locatable. Archive your code and data in a separate 3rd-party archiving site such as Zenodo, as well as any long-term access repositories that are provided by your institution (e.g., data.caltech.edu for Caltech). Note that archiving is not equivalent to making your code and data available in code-sharing sites such as GitHub. Get unique and stable identifiers for the data or data set. Even once shared data is often lost due to either resource decay and link decay. We highly recommend using a repository with resolvable identifiers REF identifiers paper, such as DOIs which are accessible and resolvable long term. In case of a dedicated database, these can be database identifiers which should be uniquely resolvable (e.g., identifiers.org information). The journal Scientific Data maintains a good list of archives you can look at.

9. Provide methods

To evaluate the usefulness of data for computational modeling it is often necessary to understand the experimental and computational methods with their setup, as well as the procedures and underlying protocols used to generate the raw and processed data. Minimal information guidelines for the respective fields exist which describe which information should be provided as metadata with the data. The FAIRsharing resource makes a wide range of minimum information guidelines available for researchers via the MIBBI FAIRsharing collection (https://fairsharing.org/collection/MIBBI). If possible link your data set to a method section in a publication or other online descriptions of the protocols (see for instance https://protocols.io). Importantly, any information on the experimental setup and protocol are better than no information.

10. Lean back and enjoy the fame

So you shared your data with license and attribution information, people will be able to find it via metadata in their favorite repositories, and you made it easy to integrate data and processing into other people's computational modeling workflows because you provided easy to parse computer-readable formats and code. What's next? Lean back and enjoy your fame, you made an important contribution to scientific research and computational models using your data could answer important questions in biology and medicine. Thanks for your efforts. Your data matters.

Summary

Publishing the data behind biomedical and clinical studies is good scientific practice, and it encourages scientific discourse. As a result, the data can be transparently checked and further reused, and scientific results can obtain a higher level of curation and trust. In this paper, we outline the recommended workflow for sharing data between biomedical and clinician scientists and the biomodeling and simulation community. We focus on mathematical models describing biomedical systems, such as disease progression, organ level models, or biochemical processes leading to disorders. Typical data that needs to be shared are genomics, proteomics, metabolomics, but also patient-specific measurements. For any of these data types, it will be important to

understand that the data should remain understandable for both humans and machines. Formal representations and semantic annotations using domain-specific standards are key factors. Data can only be reused when it is equipped with a license and deposited in a findable repository. When talking about 'data', this includes not only the raw and processed data from measurements, but also software code, scripts, documentation, and all relevant metadata such as provenance information. We would like to encourage the community to adhere to these recommendations, which fully respect the FAIR principles for data stewardship.

Acknowledgments

MK and JG are supported by the Federal Ministry of Education and Research (BMBF, Germany) within the research network Systems Medicine of the Liver (LiSyM, grant number 031L0054). MK is supported by the DFG within the Research Unit Programme FOR 5151 "QuaLiPerF (Quantifying Liver Perfusion–Function Relationship in Complex Resection—A Systems Medicine Approach)" by grant 436883643.

References

1. Molloy JC. The Open Knowledge Foundation: open data means better science. PLoS biology. 2011;9:e1001195. doi:10.1371/journal.pbio.1001195.

2. Fecher B, Friesike S, Hebing M. What drives academic data sharing? PloS one. 2015;10:e0118053. doi:10.1371/journal.pone.0118053.

3. Colavizza G, Hrynaszkiewicz I, Staden I, Whitaker K, McGillivray B. The citation advantage of linking publications to research data. PloS one. 2020;15:e0230416. doi:10.1371/journal.pone.0230416.

4. Lucraft M, Baynes G, Allin K, Hrynaszkiewicz I, Khodiyar V. Five Essential Factors for Data Sharing; 2019. Available from: `https://figshare.com/articles/journal_contribution/Five_Essential_Factors_for_Data_Sharing/7807949/1`.

5. Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, Ioannidis JPA. Public availability of published research data in high-impact journals. PloS one. 2011;6:e24357. doi:10.1371/journal.pone.0024357.

6. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data. 2016;3:160018. doi:10.1038/sdata.2016.18.

7. Hucka M, Nickerson DP, Bader GD, Bergmann FT, Cooper J, Demir E, et al. Promoting Coordinated Development of Community-Based Information Standards for Modeling in Biology: The COMBINE Initiative. Frontiers in bioengineering and biotechnology. 2015;3:19. doi:10.3389/fbioe.2015.00019.

8. Waltemath D, Golebiewski M, Blinov ML, Gleeson P, Hermjakob H, Hucka M, et al. The first 10 years of the international coordination network for standards in systems and synthetic biology (COMBINE). Journal of integrative bioinformatics. 2020;17. doi:10.1515/jib-2020-0005.

9. Miyakawa T. No raw data, no science: another possible source of the reproducibility crisis; 2020.

10. Hrynaszkiewicz I, Cockerill MJ. Open by default: a proposed copyright license and waiver agreement for open access research and data in peer-reviewed journals. BMC research notes. 2012;5:494. doi:10.1186/1756-0500-5-494.

11. Sarkans U, Gostev M, Athar A, Behrangi E, Melnichuk O, Ali A, et al. The BioStudies database-one stop shop for all data supporting a life sciences study. Nucleic acids research. 2018;46:D1266–D1270. doi:10.1093/nar/gkx965.

12. Haug K, Salek RM, Conesa P, Hastings J, de Matos P, Rijnbeek M, et al. MetaboLights–an open-access general-purpose repository for metabolomics studies and associated meta-data. Nucleic acids research. 2013;41:D781–D786. doi:10.1093/nar/gks1004.

13. Toribio AL, Alako B, Amid C, Cerdeño-Tarrága A, Clarke L, Cleland I, et al. European Nucleotide Archive in 2016. Nucleic acids research. 2017;45:D32–D36. doi:10.1093/nar/gkw1106.

14. Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic acids research. 2019;47:D442–D450. doi:10.1093/nar/gky1106.

15. Deutsch EW, Csordas A, Sun Z, Jarnuczak A, Perez-Riverol Y, Ternent T, et al. The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. Nucleic acids research. 2017;45:D1100–D1106. doi:10.1093/nar/gkw936.

16. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, et al. ArrayExpress update–simplifying data submissions. Nucleic acids research. 2015;43:D1113–D1116. doi:10.1093/nar/gku1057.

17. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. NCBI GEO: archive for functional genomics data sets–10 years on. Nucleic acids research. 2011;39:D1005–D1010. doi:10.1093/nar/gkq1184.

18. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic acids research. 2000;28:235–242. doi:10.1093/nar/28.1.235.

19. Armstrong DR, Berrisford JM, Conroy MJ, Gutmanas A, Anyango S, Choudhary P, et al. PDBe: improved findability of macromolecular structure data in the PDB. Nucleic acids research. 2020;48:D335–D343. doi:10.1093/nar/gkz990.

20. Grzegorzewski J, Brandhorst J, Green K, Eleftheriadou D, Duport Y, Barthorscht F, et al. PK-DB: pharmacokinetics database for individualized and stratified computational modeling. Nucleic acids research. 2020;doi:10.1093/nar/gkaa990.