https://doi.org/10.1038/s41592-025-02635-0

The need to implement FAIR principles in biomolecular simulations

Rommie E. Amaro, Johan Åqvist, Ivet Bahar, Federica Battistini, Adam Bellaiche, Daniel Beltran, Philip C. Biggin, Massimiliano Bonomi, Gregory R. Bowman, Richard A. Bryce, Giovanni Bussi, Paolo Carloni, David A. Case, Andrea Cavalli, Chia-En A. Chang, Thomas E. Cheatham III, Margaret S. Cheung, Christophe Chipot, Lillian T. Chong, Preeti Choudhary, G. Andres Cisneros, Cecilia Clementi, Rosana Collepardo-Guevara, Peter Coveney, Roberto Covino, T. Daniel Crawford, Matteo Dal Peraro, Bert L. de Groot, Lucie Delemotte, Marco De Vivo, Jonathan W. Essex, Franca Fraternali, Jiali Gao, Josep Ll. Gelpí, Francesco L. Gervasio, Fernando D. González-Nilo, Helmut Grubmüller, Marina G. Guenza, Horacio V. Guzman, Sarah Harris, Teresa Head-Gordon, Rigoberto Hernandez, Adam Hospital, Niu Huang, Xuhui Huang, Gerhard Hummer, Javier Iglesias-Fernández, Jan H. Jensen, Shantenu Jha, Wanting Jiao, William L. Jorgensen, Shina C. L. Kamerlin, Syma Khalid, Charles Laughton, Michael Levitt, Vittorio Limongelli, Erik Lindahl, Kresten Lindorff-Larsen, Sharon Loverde, Magnus Lundborg, Yun L. Luo, F. Javier Luque, Charlotte I. Lynch, Alexander D. MacKerell Jr, Alessandra Magistrato, Siewert J. Marrink, Hugh Martin, J. Andrew McCammon, Kenneth Merz, Vicent Moliner, Adrian J. Mulholland, Sohail Murad, Athi N. Naganathan, Shikha Nangia, Frank Noe, Agnes Noy, Julianna Oláh, Megan L. O'Mara, Mary Jo Ondrechen, Jose N. Onuchic, Alexey Onufriev, Sílvia Osuna, Giulia Palermo, Anna R. Panchenko, Sergio Pantano, Carol Parish, Michele Parrinello, Alberto Perez, Tomas Perez-Acle, Juan R. Perilla, B. Montgomery Pettitt, Adriana Pietropaolo, Jean-Philip Piquemal, Adolfo B. Poma, Matej Praprotnik, Maria J. Ramos, Pengyu Ren, Nathalie Reuter, Adrian Roitberg, Edina Rosta, Carme Rovira, Benoit Roux, Ursula Rothlisberger, Karissa Y. Sanbonmatsu, Tamar Schlick, Alexey K. Shaytan, Carlos Simmerling, Jeremy C. Smith, Yuji Sugita, Katarzyna Świderek, Makoto Taiji, Peng Tao, D. Peter Tieleman, Irina G. Tikhonova, Julian Tirado-Rives, Iñaki Tuñón, Marc W. van der Kamp, David van der Spoel, Sameer Velankar, Gregory A. Voth, Rebecca Wade, Ariel Warshel, Valerie Vaissier Welborn, Stacey D. Wetmore, Travis J. Wheeler, Chung F. Wong, Lee-Wei Yang, Martin Zacharias & Modesto Orozco

In the Big Data era, a change of paradigm in the use of molecular dynamics is required. Trajectories should be stored under FAIR (findable, accessible, interoperable and reusable) requirements to favor its reuse by the community under an open science paradigm.

The communities that embraced data archiving efforts decades ago are now, in the era of data-driven biology, gaining the most from the AI revolution. The structural biology community was a pioneer in this regard, establishing the Protein Data Bank in 1971 and making data accessible using the FAIR principles even before these were articulated[1,2]. The genomics and bioinformatics community has followed the example, establishing many widely used databases[3,4]. By contrast, molecular simulation has been anchored in usage paradigms dating back to the seventies, when molecular dynamics (MD) simulation was first applied to study biomacromolecules[5]. At that time, MD was used by theoretical physicists and chemists in proof-of-concept simulations, but 50 years later, MD has evolved into a cornerstone molecular biology technique that can provide accurate, quantitative analysis and property prediction. MD is now employed by tens of thousands of researchers worldwide, accounting for roughly 15% of global supercomputer usage. Unfortunately, these rich and costly data are not systematically maintained, and when further analyses are required, simulations have to be rerun — an unacceptable situation from scientific, environmental and sustainability standpoints. In this letter, we argue for a collaborative endeavor to archive MD simulation data and describe ongoing efforts to establish cost-effective and sustainable data archiving strategies.

Advances in computer technology have made it possible to simulate large, realistic biological systems beyond the millisecond timescale, and we are seeing simulations in the 10^9-particle range, covering entire organelles and even minimal cells, resulting in a "deluge of data"[6] in a field that lacks agreed strategies for data storage. As in the seventies, trajectories obtained after a huge effort are often ignored (or even

deleted) after a hypothesis-driven analysis is presented in a scientific publication. For a field entirely based on sampling, and where the recipe for observations can be described exactly and critically assessed, this is a huge problem. Instead of being able to reanalyze, reuse, and potentially spot undetected artifacts or new features in data, readers are often expected to blindly trust the closed set of statements made by the authors in a paper. The lack of a systematic approach to storing data (and associated provenance and metadata) prevents new studies based on previous trajectories; impedes meta-analyses, extension of trajectories, training of machine learning approaches, optimization of force fields and simulation protocols, generation of new conformations for modeling of reactivity; hampers the use of trajectories to train coarse-grained and mesoscopic models or generative models; and prohibits the integration of MD results into the rich ecosystem of biology databases. Some journals and funding institutions now require the deposition of trajectories. Without a centralized reference repository, this has led to the use of existing generic repositories (for example, Zenodo, Figshare) and the creation of numerous small, independent databases. As a result, we may face vast amounts of dispersed and disconnected data, which are expensive to maintain and often useless for further analysis. It is clear that the community needs to escape from a paradigm that made sense in the seventies but now hinders progress, and move to an open science model.

Establishing an archive for biosimulation data — upon quality assessment — would address these issues, democratize the field, and have a material impact of MD simulations on life science research. The traditional view held by the simulation community that storing and archiving is more expensive than recomputing, which might have been correct in the past, is no longer valid, as demonstrated by the massive Folding@home study on the SARS-CoV-2 main protease[7], or for simulations with many millions of atoms[8]. However, the new science that can be learned from stored trajectories is more important than the cost. For instance, the ABC Consortium[9] was established in 2004 as a community effort generating a multi-gigabyte database of DNA simulations, which had grown to hold 15 terabytes of data by 2019. The original goal of ABC was to study DNA polymorphisms, but the database has become crucial in other fields, such as force field refinement, the study of signal transfer in DNA and the development of coarse-grained models. The current HexABC database contains 400 terabytes of data generated by 14 different groups to explore hexamer dependencies of DNA dynamics. However, its future use, which is difficult to anticipate, might be more important than the current goals of the project. Another example emerged during the COVID-19 pandemic[10,11], when the Molecular Sciences Software Institute (MolSSI), in collaboration with European groups including BioExcel, European Open Science Cloud, European Bioinformatics Institute and Zenodo, created the COVID-19 Molecular Structure and Therapeutics Hub (https://covid.molssi.org). It went live in April 2020, connecting scientists across the global biomolecular simulation community, as well as improving the connection between simulation and experimental and clinical data and their investigators. A further example is MDverse (https://mdverse.github.io/), an effort to make MD trajectories FAIRer by indexing and curating thousands of simulations scattered across the internet. Many other examples are now under development, highlighting the general belief of the community that the traditional paradigm from the seventies should be abandoned and all well-annotated, validated trajectories should be stored and integrated in a general data infrastructure to favor the advance of science and the optimization of computational resources.

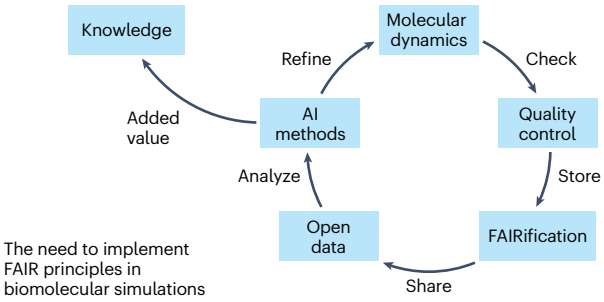

Fig. 1 | Data cycle workflow for implementing FAIR (findable, accessible, interoperable and reusable) principles in biomolecular simulations. The diagram highlights the added value that can be extracted from accessible open data.

The challenges that lie ahead for the community are diverse. The technical ones —sustained data storage capacity, bandwidth, and processing capacity for analysis — can be alleviated by a distributed database policy following initiatives such as the EGA infrastructure (European Genome-Phenome Archive; https://ega-archive.org/) and by the commitment of funding institutions and high-performance computing centers, offering storage, bandwidth and processing capabilities. Other key decisions such as quality requirements for storing and maintaining the data, the sparsity of the trajectory, the compression strategy, or whether stored trajectory should be dry or contain also solvent molecules should be taken by the community, keeping in mind that, while storing all the potential information derived by an MD simulation might be impossible, preserving as much data as possible should be a priority.

A centralized management entity should coordinate the federated nodes, defining required metadata (crucial for reproducibility, extension of trajectories, increase of the time density of snapshots, or meta-analysis), setting deposition policies, guaranteeing compliance of FAIR rules and providing a common entry point through web-based and programmatic representational state transfer (REST) API interfaces. The myriads of variants of MD programs, protocols, formats and simulation conditions lead to more complex problems. Recent MD repositories and databases[11] are already prepared to manage not only plain MD trajectories but also Markov state models, ensembles, multiscale simulations (hybrid or combined approaches involving mesoscale, coarse-grained and atomistic methods, as well as quantum mechanics with molecular mechanics), constant pH, replica exchange, and MD trajectories biased with metadynamics or similar methods. NoSQL databases such as MongoDB (with the GridFS file storage and retrieving specification) allow efficient storing and querying of the diversity of outputs provided by MD engines and are already adopted by MD storage initiatives. However, much more work is required for an effective analysis framework that can manage an increasingly large number of MD variants and trajectory formats.

Data should be findable, with each entry registered with a persistent identifier, ideally a DOI, ensuring a proper citation, following the example of the WorkflowHub registry (https://workflowhub.eu/). Furthermore, they should be stored in an interoperable manner, so that they can be read and exploited by current and future data scraping and machine learning algorithms. To this end, the community must reach an agreement to standardize MD data exchange formats with (i) efficient trajectory compression, including simple system specifications

(for example, atom or residue names and connectivity); (ii) key-value trees storing high-level and full simulation settings metadata; and (iii) metadata-based ontology[12], which would allow the user to search databases on the basis of the contents, the nature or even the purpose of the simulations. Standardized provenance should be stored by means of data blocks specifying commands or operations used to generate the trajectory, together with names, stored hash sums of the complete files used for input, and specific software used (with precise versions). This would allow the user to reproduce all the different steps followed to prepare and run the simulation, including modeling of missing residues, physical conditions (for example, pH, salt concentration, temperature and pressure) and force fields, methodology used to obtain parameters involving non-standard molecules (for example, small molecules, membrane systems, ionic coordination), and the equilibration and possibly sampling process. Minimum metadata should include system information, simulation parameters, author(s), data license and copyright, and, importantly, the main purpose of the simulation. The definition of standardized protocols (that is, list of operations) for production run and analysis, including a troubleshooting section, could be added. These, along with a set of metadata-dependent quality control analyses, both general and system specific, are crucial requisites for gaining trust from the community and for defining deposition rules. A data repository following FAIR principles and the associated analysis tools will increase the impact and the reproducibility (complex at the binary level; that is, it is difficult to reproduce exactly the same trajectory owing to numerical errors) of MD in related fields in the life science data ecosystem, from genomics to structural biology and from protein and drug design to molecular biology. MD data would provide unique dynamic information of biological macromolecules fully complementary with the rich information available from the Protein Data Bank. This could be integrated into the life science ecosystem following the approach of the Protein Data Bank in Europe Knowledge Base, designed for the integration and enrichment of 3D structure data and functional annotations[13]. All this information will contribute to knowledge democratization, helping research teams with limited resources and fueling further advances in artificial intelligence (AI) in the scientific domain[14] (Fig. 1).

The MDDB project (https://mddbr.eu/) and similar initiatives aim to establish such a repository, allowing (i) data quality assessment metrics to increase the trust of the community in the deposited data; (ii) common data format, metadata requirements and ontologies to facilitate interoperability; (iii) a minimum set of information needed to store and reproduce the simulations, including data provenance, license and copyright; and (iv) a standard and robust infrastructure to store and share the data, with persistent identifiers and different ways to access them. We believe science will be better served by fully embracing this data-driven view of biomolecular simulation. Furthermore, data-driven initiatives such that supported by this Correspondence would help the interaction with other simulation communities, such as the materials science one, which share some of the problems the biomolecular simulation community is facing.

Rommie E. Amaro[1], Johan Åqvist[2], Ivet Bahar[3,4], Federica Battistini[5], Adam Bellaiche[6], Daniel Beltran[7], Philip C. Biggin[8], Massimiliano Bonomi[9], Gregory R. Bowman[10], Richard A. Bryce[11], Giovanni Bussi[12], Paolo Carloni[13,14], David A. Case[15], Andrea Cavalli[16,17], Chia-En A. Chang[18], Thomas E. Cheatham III[19], Margaret S. Cheung[20,21], Christophe Chipot[22,23,24], Lillian T. Chong[25], Preeti Choudhary[6], G. Andres Cisneros[26,27], Cecilia Clementi[28], Rosana Collepardo-Guevara[29,30,31], Peter Coveney[32,33], Roberto Covino[34,35], T. Daniel Crawford[36,37], Matteo Dal Peraro[38], Bert L. de Groot[39], Lucie Delemotte[40], Marco De Vivo[41], Jonathan W. Essex[42], Franca Fraternali[43], Jiali Gao[44], Josep Ll. Gelpí[5,45], Francesco L. Gervasio[46,47,48,49], Fernando D. González-Nilo[50], Helmut Grubmüller[51], Marina G. Guenza[52], Horacio V. Guzman[53], Sarah Harris[54], Teresa Head-Gordon[55], Rigoberto Hernandez[56], Adam Hospital[7,57] ✉, Niu Huang[58], Xuhui Huang[59], Gerhard Hummer[60,61], Javier Iglesias-Fernández[62], Jan H. Jensen[63], Shantenu Jha[64], Wanting Jiao[65], William L. Jorgensen[66], Shina C. L. Kamerlin[67,68], Syma Khalid[8], Charles Laughton[69], Michael Levitt[70], Vittorio Limongelli[71], Erik Lindahl[40,72], Kresten Lindorff-Larsen[73], Sharon Loverde[74], Magnus Lundborg[40], Yun L. Luo[75], F. Javier Luque[76,77], Charlotte I. Lynch[8], Alexander D. MacKerell Jr[78], Alessandra Magistrato[79], Siewert J. Marrink[80], Hugh Martin[32], J. Andrew McCammon[81,82], Kenneth Merz[83,84], Vicent Moliner[85], Adrian J. Mulholland[86], Sohail Murad[87], Athi N. Naganathan[88], Shikha Nangia[89], Frank Noe[90,91,92,93], Agnes Noy[94], Julianna Oláh[95], Megan L. O'Mara[96], Mary Jo Ondrechen[97], Jose N. Onuchic[92,98,99,100], Alexey Onufriev[101,102,103], Sílvia Osuna[104,105], Giulia Palermo[18,106], Anna R. Panchenko[107,108,109], Sergio Pantano[110,111], Carol Parish[112], Michele Parrinello[113], Alberto Perez[114], Tomas Perez-Acle[115,116], Juan R. Perilla[117], B. Montgomery Pettitt[118], Adriana Pietropaolo[119], Jean-Philip Piquemal[120], Adolfo B. Poma[121], Matej Praprotnik[122,123], Maria J. Ramos[124], Pengyu Ren[125], Nathalie Reuter[126,127], Adrian Roitberg[114], Edina Rosta[128], Carme Rovira[105,129], Benoit Roux[130], Ursula Rothlisberger[131], Karissa Y. Sanbonmatsu[132,133], Tamar Schlick[134,135], Alexey K. Shaytan[136,137], Carlos Simmerling[3,138], Jeremy C. Smith[139,140], Yuji Sugita[141,142,143], Katarzyna Świderek[85], Makoto Taiji[144], Peng Tao[145], D. Peter Tieleman[146], Irina G. Tikhonova[147], Julian Tirado-Rives[66], Iñaki Tuñón[148], Marc W. van der Kamp[149], David van der Spoel[2], Sameer Velankar[6], Gregory A. Voth[150], Rebecca Wade[151], Ariel Warshel[152], Valerie Vaissier Welborn[37,153], Stacey D. Wetmore[154], Travis J. Wheeler[155], Chung F. Wong[156], Lee-Wei Yang[157], Martin Zacharias[158] & Modesto Orozco[5,7] ✉

[1]Department of Molecular Biology, University of California San Diego, La Jolla, CA, USA. [2]Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden. [3]Laufer Center for Physical and Quantitative Biology, Stony Brook University, Stony Brook, NY, USA. [4]Department of Biochemistry and Cell Biology, Renaissance School of Medicine, Stony Brook University, Stony Brook, NY, USA. [5]Department of Biochemistry and Molecular Biomedicine, University of Barcelona, Barcelona, Spain. [6]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, UK. [7]Institute for Research in Biomedicine (IRB Barcelona), Barcelona, Spain. [8]Structural Bioinformatics and Computational Biochemistry, Department of Biochemistry, University of Oxford, Oxford, UK. [9]Institut Pasteur, Université Paris Cité, CNRS UMR 3528, Computational Structural Biology Unit, Paris, France. [10]Department of Biochemistry and Biophysics, University of Pennsylvania, Philadelphia, PA, USA. [11]Division of Pharmacy and Optometry, University of Manchester, Manchester, UK. [12]Scuola Internazionale Superiore di Studi Avanzati-SISSA, Trieste, Italy. [13]Computational Biomedicine, Institute of Advanced Simulations IAS-5/Institute for Neuroscience and Medicine INM-9,

Forschungszentrum Jülich GmbH, Jülich, Germany. [14]Department of Physics and Universitätsklinikum, RWTH Aachen University, Aachen, Germany. [15]Department of Chemistry & Chemical Biology, Rutgers University, Piscataway, NJ, USA. [16]Istituto Italiano di Tecnologia, Drug Discovery and Development, Bologna, Italy. [17]Centre Européen de Calcul Atomique et Moléculaire (CECAM), Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. [18]Department of Chemistry, University of California, Riverside, CA, USA. [19]Department of Medicinal Chemistry, College of Pharmacy, University of Utah, Salt Lake City, UT, USA. [20]Department of Physics, University of Washington, Seattle, Washington, USA. [21]Pacific Northwest National Laboratory, Richland, Washington, USA. [22]LIA CNRS-UIUC, UMR 7019, Université de Lorraine, Vandœuvre-lès-Nancy, France. [23]Department of Physics, University of Illinois at Urbana-Champaign, Urbana, IL, USA. [24]Department of Biochemistry and Molecular Biology, The University of Chicago, Chicago, IL, USA. [25]Department of Chemistry, University of Pittsburgh, Pittsburgh, PA, USA. [26]Department of Chemistry and Biochemistry, University of Texas at Dallas, Richardson, TX, USA. [27]Department of Physics, University of Texas at Dallas, Richardson, TX, USA. [28]Theoretical and Computational Biophysics, Department of Physics, Freie Universität Berlin, Berlin, Germany. [29]Maxwell Centre, Cavendish Laboratory, Department of Physics, University of Cambridge, Cambridge, UK. [30]Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge, UK. [31]Department of Genetics, University of Cambridge, Cambridge, UK. [32]Centre for Computational Science, University College London, London, UK. [33]Institute for Informatics, University of Amsterdam, Amsterdam, Netherlands. [34]Institute of Computer Science, Goethe University Frankfurt, Frankfurt am Main, Germany. [35]Frankfurt Institute for Advanced Studies, Frankfurt am Main, Germany. [36]Molecular Sciences Software Institute, Blacksburg, VA, USA. [37]Department of Chemistry, Virginia Tech, Blacksburg, VA, USA. [38]Institute of Bioengineering, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. [39]Computational Biomolecular Dynamics Group, Department of Theoretical and Computational Biophysics, Max Planck Institute for Multidisciplinary Sciences, Göttingen, Germany. [40]Department of Applied Physics, Science for Life Laboratory, KTH Royal Institute of Technology, Solna, Sweden. [41]Laboratory of Molecular Modelling & Drug Discovery, Istituto Italiano di Tecnologia, Genoa, Italy. [42]School of Chemistry and Chemical Engineering, University of Southampton, Southampton, UK. [43]Institute of Structural and Molecular Biology, University College London, London, UK. [44]Department of Chemistry, University of Minnesota, Minneapolis, MN, USA. [45]Barcelona Supercomputing Center (BSC), Barcelona, Spain. [46]Pharmaceutical Sciences, University of Geneva, Geneva, Switzerland. [47]Institute of Pharmaceutical Sciences of Western Switzerland, Geneva, Switzerland. [48]Chemistry Department, University College London, London, UK. [49]Swiss Bioinformatics Institute, Geneva, Switzerland. [50]Center for Bioinformatics and Integrative Biology (CBIB), Facultad Ciencias de la Vida, Universidad Andrés Bello, Santiago, Chile. [51]Department of Theoretical and Computational Biophysics, Max-Planck-Institute for Multidisciplinary Sciences, Göttingen, Germany. [52]Department of Chemistry and Biochemistry, University of Oregon, Eugene, OR, USA. [53]Institut de Ciència de Materials de Barcelona, CSIC, Barcelona, Spain. [54]School of Physics and Astronomy, University of Sheffield, Sheffield, UK. [55]Pitzer Theory Center and Departments of Chemistry, Bioengineering, Chemical and Biomolecular Engineering, University of California, Berkeley, Berkeley, CA, USA. [56]Department of Chemistry, Johns Hopkins University, Baltimore, MD, USA. [57]Spanish National Institute of Bioinformatics (INB)/ELIXIR-ES, Barcelona, Spain. [58]National Institute of Biological Sciences, Beijing, China. [59]Department of Chemistry, Data Science Institute, University of Wisconsin-Madison, Madison, WI, USA. [60]Department of Theoretical Biophysics, Max Planck Institute of Biophysics, Frankfurt am Main, Germany. [61]Institute for Biophysics, Goethe University Frankfurt, Frankfurt am Main, Germany. [62]NBD | Nostrum Biodiscovery, Barcelona, Spain. [63]Department of Chemistry, University of Copenhagen, Copenhagen, Denmark. [64]Department of Electrical and Computer Engineering, Rutgers University, New Brunswick, NJ, USA. [65]Ferrier Research Institute, Victoria University of Wellington, Wellington, New Zealand. [66]Department of Chemistry, Yale University, New Haven, CT, USA. [67]Department of Chemistry, Lund University, Lund, Sweden. [68]School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, GA, USA. [69]School of Pharmacy and Centre for Biomolecular Sciences, University of Nottingham, Nottingham, UK. [70]Department of Structural Biology, Stanford University School of Medicine, Stanford, CA, USA. [71]Euler Institute, Faculty of Biomedical Sciences, Università della Svizzera italiana (USI), Lugano, Switzerland. [72]Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, Solna, Sweden. [73]Structural Biology and NMR Laboratory & the Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, Copenhagen, Denmark. [74]Department of Chemistry, College of Staten Island, The City University of New York, New York, NY, USA. [75]College of Pharmacy, Western University of Health Sciences, Pomona, CA, USA. [76]Institute of Biomedicine, University of Barcelona, Santa Coloma de Gramanet, Spain. [77]Institute of Theoretical and Computational Chemistry, University of Barcelona, Santa Coloma de Gramanet, Spain. [78]Department of Pharmaceutical Sciences, School of Pharmacy, University of Maryland, Baltimore, MD, USA. [79]National Research Council-Institute of Material Foundry at Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy. [80]Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Groningen, the Netherlands. [81]Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA, USA. [82]Department of Pharmacology, University of California, San Diego, La Jolla, CA, USA. [83]Lerner Research Institute Cleveland Clinic, Cleveland, OH, USA. [84]Department of Chemistry, Michigan State University, East Lansing, MI, USA. [85]BioComp Group, Institute of Advanced Materials (INAM), Universitat Jaume I, Castellón, Spain. [86]Centre for Computational Chemistry, School of Chemistry, University of Bristol, Bristol, UK. [87]Department of Chemical and Biological Engineering, Illinois Institute of Technology, Chicago, IL, USA. [88]Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai, India. [89]Department of Biomedical and Chemical Engineering, Syracuse University, Syracuse, NY, USA. [90]Department of Physics, Freie Universität Berlin, Berlin, Germany. [91]Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin, Germany. [92]Department of Chemistry, Rice University, Houston, TX, USA. [93]Microsoft Research AI4Science, Berlin, Germany. [94]School of Physics, Engineering and Technology, University of York, York, UK. [95]Department of Inorganic and Analytical Chemistry, Budapest University of Technology and Economics, Budapest, Hungary. [96]Australian Institute for Bioengineering and Nanotechnology, The University of Queensland, Brisbane, Queensland, Australia. [97]Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA, USA.

[98]Center for Theoretical Biological Physics, Rice University, Houston, TX, USA. [99]Department of Physics and Astronomy, Rice University, Houston, TX, USA. [100]Department of BioSciences, Rice University, Houston, TX, USA. [101]Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA. [102]Department of Physics, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA. [103]Center for Soft Matter and Biological Physics, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA. [104]Institut de Química Computacional i Catàlisi (IQCC) and Departament de Química, Universitat de Girona, Girona, Spain. [105]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain. [106]Department of Bioengineering, University of California Riverside, Riverside, CA, USA. [107]Department of Pathology and Molecular Medicine, School of Medicine, Queen's University, Kingston, Ontario, Canada. [108]Department of Biology and Molecular Sciences, Queen's University, Kingston, Ontario, Canada. [109]School of Computing, Queen's University, Kingston, Ontario, Canada. [110]Biomolecular Simulations Group, Institut Pasteur de Montevideo, Montevideo, Uruguay. [111]Bioinformatics Area, DETEMA, Faculty of Chemistry, Udelar, Montevideo, Uruguay. [112]Department of Chemistry, Gottwald Center for the Sciences, University of Richmond, Richmond, VA, USA. [113]Atomistic Simulations, Italian Institute of Technology, Genova, Italy. [114]Department of Chemistry and Quantum Theory Project, University of Florida, Gainesville, FL, USA. [115]Computational Biology Lab, Fundación Ciencia & Vida, Santiago, Chile. [116]Facultad de Ingeniería, Universidad San Sebastián, Santiago, Chile. [117]Department of Chemistry and Biochemistry, University of Delaware, Newark, DE, USA. [118]University of Texas Medical Branch, Galveston, TX, USA. [119]Dipartimento di Scienze della Salute, Università di Catanzaro, Catanzaro, Italy. [120]Laboratory of Theoretical Chemistry, Department of Chemistry, Sorbonne University, Paris, France. [121]Biosystems and Soft Matter Division, Institute of Fundamental Technological Research, Polish Academy of Sciences, Warsaw, Poland. [122]Laboratory for Molecular Modeling, National Institute of Chemistry, Ljubljana, Slovenia. [123]Department of Physics, Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia. [124]Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, Porto, Portugal. [125]Department of Biomedical Engineering, The University of Texas at Austin, Austin, TX, USA. [126]Department of Chemistry, University of Bergen, Bergen, Norway. [127]Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway. [128]Department of Physics and Astronomy, University College London, London, UK. [129]Departament de Química Inorgànica i Orgànica (Secció de Química Orgànica) and Institut de Química Teòrica i Computacional (IQTCUB), Universitat de Barcelona, Barcelona, Spain. [130]Department of Chemistry, University of Chicago, Chicago, IL, USA. [131]Laboratory of Computational Chemistry and Biochemistry, Institute of Chemical Sciences and Engineering, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland. [132]Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM, USA. [133]New Mexico Consortium, Los Alamos, NM, USA. [134]Department of Chemistry and Courant Institute of Mathematical Sciences, New York University, New York, NY, USA. [135]Simons Center for Computational Physical Chemistry, New York University, New York, NY, USA. [136]Department of Biology, Lomonosov Moscow State University, Moscow, Russia. [137]International Laboratory of Bioinformatics, AI and Digital Sciences Institute, Faculty of Computer Science, HSE University, Moscow, Russia. [138]Department of Chemistry, Stony Brook University, Stony Brook, NY, USA. [139]Biosciences Division and Center for Molecular Biophysics, Oak Ridge National Laboratory, Oak Ridge, TN, USA. [140]Department of Biochemistry and Cellular and Molecular Biology, University of Tennessee, Knoxville, TN, USA. [141]Computational Biophysics Research Team, RIKEN Center for Computational Science, Kobe, Japan. [142]Theoretical Molecular Science Laboratory, RIKEN Cluster for Pioneering Research, Saitama, Japan. [143]Laboratory for Biomolecular Function Simulation, RIKEN Center for Biosystems Dynamics Research, Kobe, Japan. [144]Laboratory for Computational Molecular Design, RIKEN Center for Biosystems Dynamics Research, Kobe, Japan. [145]Department of Chemistry, O'Donnell Data Science and Research Computing Institute, Center for Drug Discovery, Design, and Delivery (CD4), Southern Methodist University, Dallas, TX, USA. [146]Department of Biological Sciences and Centre for Molecular Simulation, University of Calgary, Calgary, Alberta, Canada. [147]School of Pharmacy, Queen's University Belfast, Belfast, UK. [148]Departamento de Química Física, Universidad de Valencia, Burjassot, Spain. [149]School of Biochemistry, University of Bristol, Bristol, UK. [150]Department of Chemistry, Chicago Center for Theoretical Chemistry, Institute for Biophysical Dynamics, and James Franck Institute, The University of Chicago, Chicago, IL, USA. [151]Molecular and Cellular Modeling Group, Heidelberg Institute for Theoretical Studies (HITS), Heidelberg, Germany. [152]Department of Chemistry, University of Southern California, Los Angeles, CA, USA. [153]Macromolecules Innovation Institute, Virginia Tech, Blacksburg, VA, USA. [154]Department of Chemistry and Biochemistry, University of Lethbridge, Lethbridge, Alberta, Canada. [155]R. Ken Coit College of Pharmacy, University of Arizona, Tucson, Arizona, USA. [156]Department of Chemistry and Biochemistry, University of Missouri-St. Louis, St. Louis, MO, USA. [157]Institute of Bioinformatics and Structural Biology, National Tsing Hua University, Hsinchu, Taiwan. [158]Physics Department, Technical University of Munich, Garching, Germany.
✉e-mail: adam.hospital@irbbarcelona.org; modesto.orozco@irbbarcelona.org

Published online: 02 April 2025

References

1. wwPDB Consortium. *Nucleic Acids Res.* **47**, D520–D528 (2019). (D1).
2. Wilkinson, M. D. et al. *Sci. Data* **3**, 160018 (2016).
3. Thakur, M. et al. *Nucleic Acids Res.* **51**, D9–D17 (2023). (D1).
4. Rigden, D. J. & Fernández, X. M. *Nucleic Acids Res.* **50**, D1–D10 (2022). (D1).
5. McCammon, J. A., Gelin, B. R. & Karplus, M. *Nature* **267**, 585–590 (1977).
6. Hospital, A. et al. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **10**, e1449 (2020).
7. von Delft, F. et al. *Nature* **594**, 330–332 (2021).
8. Dommer, A. et al. *Int. J. High Perform. Comput. Appl.* **37**, 28–44 (2023).
9. da Rosa, G. et al. *Biophys. Rev.* **13**, 995–1005 (2021).
10. Amaro, R. E. & Mulholland, A. J. *J. Chem. Inf. Model.* **60**, 2653–2656 (2020).
11. Beltrán, D. et al. *Nucleic Acids Res.* **52**, D393–D403 (2024). (D1).
12. Hospital, A. et al. *Nucleic Acids Res.* **44**, D272–D278 (2016). (D1).
13. PDBe-KB Consortium. *Nucleic Acids Res.* **48**, D344–D353 (2019). (D1).
14. Dessimoz, C. & Thomas, P. D. *Sci. Data* **11**, 268 (2024).

Acknowledgements

The authors thank the whole MD community for useful inputs and discussions. The MDDB project is supported by European Union's Horizon Europe programme under grant agreement 101094651 awarded to M.O., E.L., S.V., J.L.G., J.I., A.C. and P.C.B.

Author contributions

M.O. conceived the idea. M.O., A.H. and E.L. wrote the draft of the paper, which was corrected first by other members of the MDDB project (S.V., P.C.B., J.L.G., J.I.-F., A.C.) and later by all the authors.

Competing interests

The authors declare no competing interests.