Nucleic Acids Research, 2023, 1–11
https://doi.org/10.1093/nar/gkad991
Database issue

OXFORD

A new paradigm for molecular dynamics databases: the COVID-19 database, the legacy of a titanic community effort

Daniel Beltrán[1], Adam Hospital[1,*], Josep Lluís Gelpí[2,3] and Modesto Orozco[1,2,*]

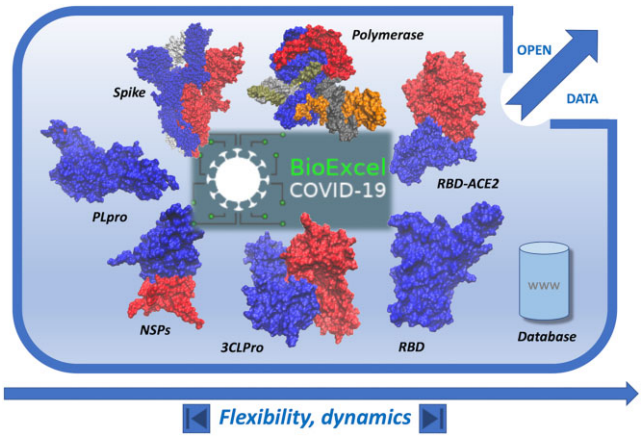
[1]Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology, Barcelona, Spain
[2]Department of Biochemistry and Biomedicine. University of Barcelona, Barcelona, Spain
[3]Barcelona Supercomputing Center (BSC), Barcelona, Spain

[*]To whom correspondence should be addressed. Tel: +34 93 403 7155; Email: adam.hospital@irbbarcelona.org
Correspondence may also be addressed to Modesto Orozco. Email: modesto.orozco@irbbarcelona.org

Abstract

Molecular dynamics (MD) simulations are keeping computers busy around the world, generating a huge amount of data that is typically not open to the scientific community. Pioneering efforts to ensure the safety and reusability of MD data have been based on the use of simple databases providing a limited set of standard analyses on single-short trajectories. Despite their value, these databases do not offer a true solution for the current community of MD users, who want a flexible analysis pipeline and the possibility to address huge non-Markovian ensembles of large systems. Here we present a new paradigm for MD databases, resilient to large systems and long trajectories, and designed to be compatible with modern MD simulations. The data are offered to the community through a web-based graphical user interface (GUI), implemented with state-of-the-art technology, which incorporates system-specific analysis designed by the trajectory providers. A REST API and associated Jupyter Notebooks are integrated into the platform, allowing fully customized meta-analysis by final users. The new technology is illustrated using a collection of trajectories obtained by the community in the context of the effort to fight the COVID-19 pandemic. The server is accessible at https://bioexcel-cv19.bsc.es/#/. It is free and open to all users and there are no login requirements. It is also integrated into the simulations section of the BioExcel-MolSSI *COVID-19 Molecular Structure and Therapeutics Hub*: https://covid.molssi.org/simulations/ and is part of the MDDB effort (https://mddbr.eu).

Graphical abstract



Introduction

Five decades after the first dynamics of folded proteins were obtained (1), molecular dynamics (MD) has become the cornerstone of biomolecular simulations, being heavily used in a wide range of fields, including biophysics, structural biology, biochemistry, enzymology, pharmacology, molecular biology and even virology and cellular biology (2,3). Software and hardware improvements have continuously increased the size of the systems considered and the length of the trajectories collected (4–8). The picosecond-long trajectories of systems containing a few thousand atoms obtained in the 1970s have led to millisecond-long trajectories on systems containing millions of atoms (9–12). These technical improvements, combined with continuous refinements of force fields (FFs) (13–15), have convinced a large community of biologists and chemists of the predictive power of MD simulations. In this

Received: August 14, 2023. **Revised:** October 16, 2023. **Editorial Decision:** October 16, 2023. **Accepted:** October 17, 2023

context, increased interest has led to a dramatic growth in the number of trajectory producers and a 'deluge of data' that is impossible to manage with current standards (16).

For decades, the MD field has worked using a pragmatic approach: trajectories are collected and analyzed by a group, which, after a few months, erases them considering that if the trajectory were to be needed again then it could be re-run at a future date. No data are shared with the rest of the community and some simulations are repeated hundreds of times under identical conditions by different groups (sometimes even by the same one). Computer resources are wasted, data are not subjected to post-processing in search of additional information, learning from trajectories to improve FFs becomes impossible, and meta-analysis seeking to obtain information from a set of systems sharing common characteristics cannot be performed. In summary, and in the context of data management, the MD community should move to a FAIR (Findable, Accessible, Interoperable and Reusable) model that is transversally implemented in all fields of the biosciences.

Early initiatives to store MD data in a reusable format started a decade ago. These were focused on proteins and aimed to cover a representative subset of the PDB to obtain a pan-proteome description of protein dynamics (17–19). A few family-specific databases have since been developed (20–22), and some initiatives have focused on DNA (23,24). All these databases store coordinates—typically without the associated metadata—mainly obtained with a single MD engine by the same group developing the database. They offer the final user a limited set of predefined analyses through a rigid web-graphical interface. Most of these databases were created assuming Markovian ensembles, i.e. single trajectories, and only a few of them allow the generation of meta-trajectories by combining individual trajectories for the same system (23). While these pioneering efforts have shown the community the potential and power of storing MD trajectories, they do not provide a solution for managing the information gathered from last-generation MD simulations. New platforms should be prepared for huge systems and non-Markovian ensembles, should implant system-specific analysis and should face the need for more flexible and tailored analyses from the final user.

Here we present a new paradigm for MD database generation. We demonstrate its power in a new database comprising trajectories obtained by dozens of groups in the context of the effort to fight the COVID-19 pandemic. At the time of submitting this paper, the database contains $\sim 10^4$ (and still growing) trajectories, including classical single trajectories, replica exchange, massive parallel simulations, Markov State Models, biased trajectories, etc. The analyses are tailored for the simulated system, and the user has complete programmatic access to data to perform new on-demand analysis. Developed within the EU MDDB project (https://mddbr.eu/#/), the server is accessible at https://bioexcel-cv19.bsc.es/#/. It is free and open to all users and there are no login requirements. It is also integrated into the simulations section of the BioExcel-MolSSI *COVID-19 Molecular Structure and Therapeutics Hub*: https://covid.molssi.org/simulations/ linked to the EU BioExcel project.

Database design and implementation

Contrary to previous initiatives (17,23,25,26) that distribute the trajectories and associated files in different databases or a combination of files and databases, here we use a single database model based on MongoDB technology (https://www.mongodb.org) to store (in the case of the present BioExcel COVID-19 database) 6 MongoDB collections (Figure 1), namely Projects, Topologies, Analyses, Trajectories, SeqRef and StructRef. Note that to gain efficient data access and retrieval in the document-based structure of MongoDB, the raw trajectory is converted into an in-house binary format (Supplementary Figure S1).

The MongoDB distributed database is deployed in the Barcelona Supercomputing Center (BSC), using an Open Nebula private cloud infrastructure. The current COVID-19 database uses 18 Virtual Machines (VMs), each of them with a local 6 TB disk, configured as a combination of three replication machines (i.e. mirroring data services), each containing six shards (i.e. distributed data servers). Data are split across multiple instances (shards), and each shard is cloned in different machines (replicas). This scheme offers load balancing, thereby distributing a batch of tasks over the set of available resources, and it allows for parallelized data processing, ultimately enhancing the efficiency of the overall processing. The current database has a total capacity of 108 TB (extendable easily to up to 300 TB). The server portal is also implemented on top of a VM deployed in the same private cloud, allowing direct and efficient connection to the database using the Node.js MongoDB driver. Interestingly, the database system can grow easily to incorporate other databases following the same infrastructure model.

Data overview

The BioExcel COVID-19 database is populated with MD trajectories from dozens of simulation groups around the world that used a variety of MD engines. The current database is populated mainly by the following: (i) COVID-19 Molecular Structure and Therapeutics Hub (https://covid.bioexcel.eu/simulations/); (ii) CHARMM-GUI Archive—COVID-19 Proteins Library (https://charmm-gui.org/?doc=archive&lib=covid19); (iii) Amaro Lab COVID-19 Data set (https://amarolab.ucsd.edu/covid19.php); (iv) 'in-house' MD simulations and (v) data directly provided by colleagues. All these databases are monitored continuously to identify new entries to transfer and analyze.

As of August 2023, the database contains > 10K MD simulations, with an accumulated time of 12.7 ms, and covering the following 11 SARS-CoV-2 protein units: *Spike; Angiotensin Converting Enzyme 2 (ACE2); Receptor Binding Domain (RBD); RBD-ACE2 complex; 3C-like protease (3CLpro); Papain-like Protease (PLPro); polymerase; Non-Structural Proteins (NSP); nucleoproteins; membrane proteins;* and *other proteins* (Figure 2). The original simulation systems range from 20 000 to 500 000 atoms and individual trajectories from 20 ns to 100 μs, with the longest accumulated ensemble reaching the 10-millisecond scale (multiple replica simulations). Systems containing millions of atoms are currently being added to the database.

The database incorporates traditional Markovian single-trajectories, as well as non-Markovian ensembles such as enhanced sampling simulations, biased simulations and multiple replica (from 5 to >5000) simulations, some of them compressed into Markov State Models (see Supplementary Table S1).

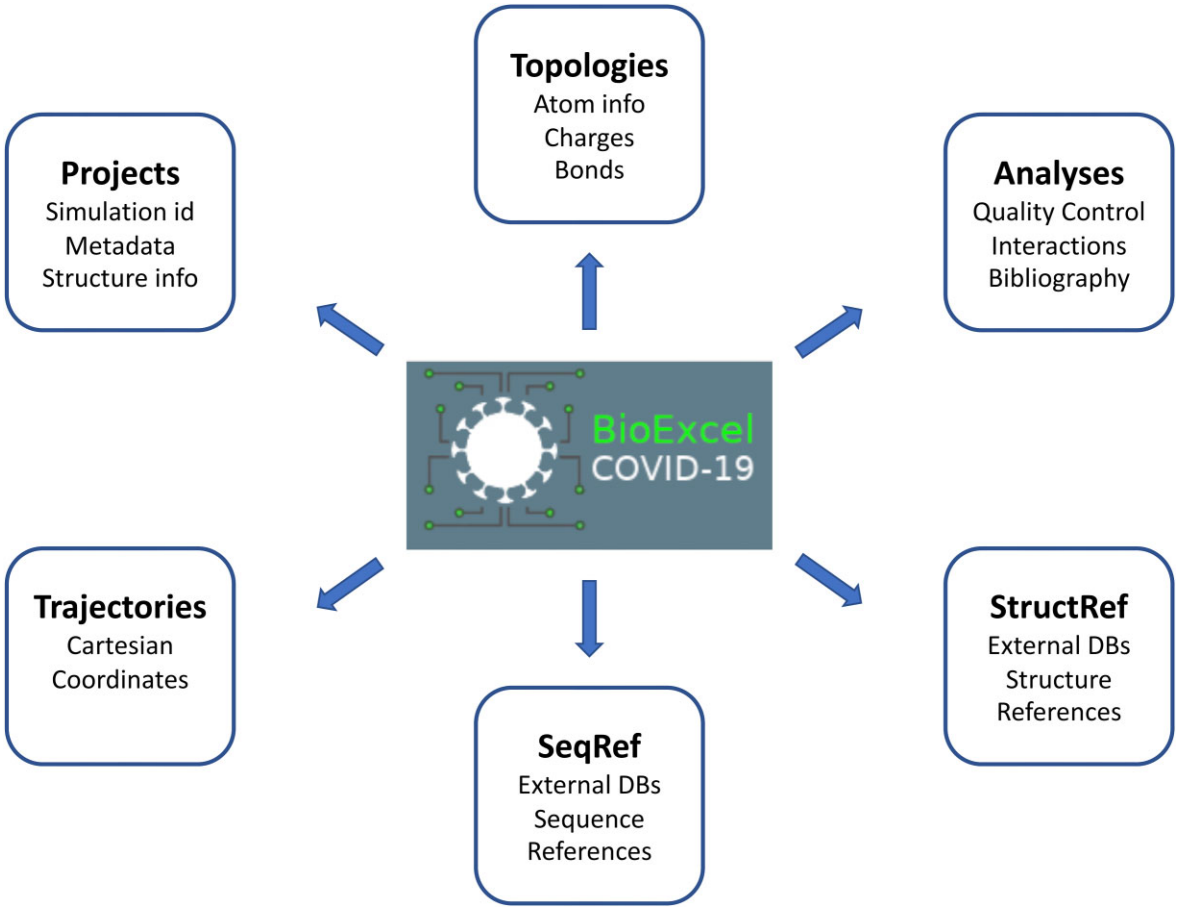


Figure 1. BioExcel COVID-19 database collections. Projects (MD simulation IDs, structure information, simulation metadata); Topologies (bonds, dihedrals, elements, charges); Analyses (general and system specific; see below); SeqRef (references to external biological sequence-based databases: Uniprot, Hmmer, InterProScan); StructRef (references to PDB) and Trajectories (coordinates). Within the database, each simulation is represented by a unique and persistent identifier (e.g. MCV1900002), with a prefix defining the database 'MCV19' and 5 digits for the individual simulations stored.

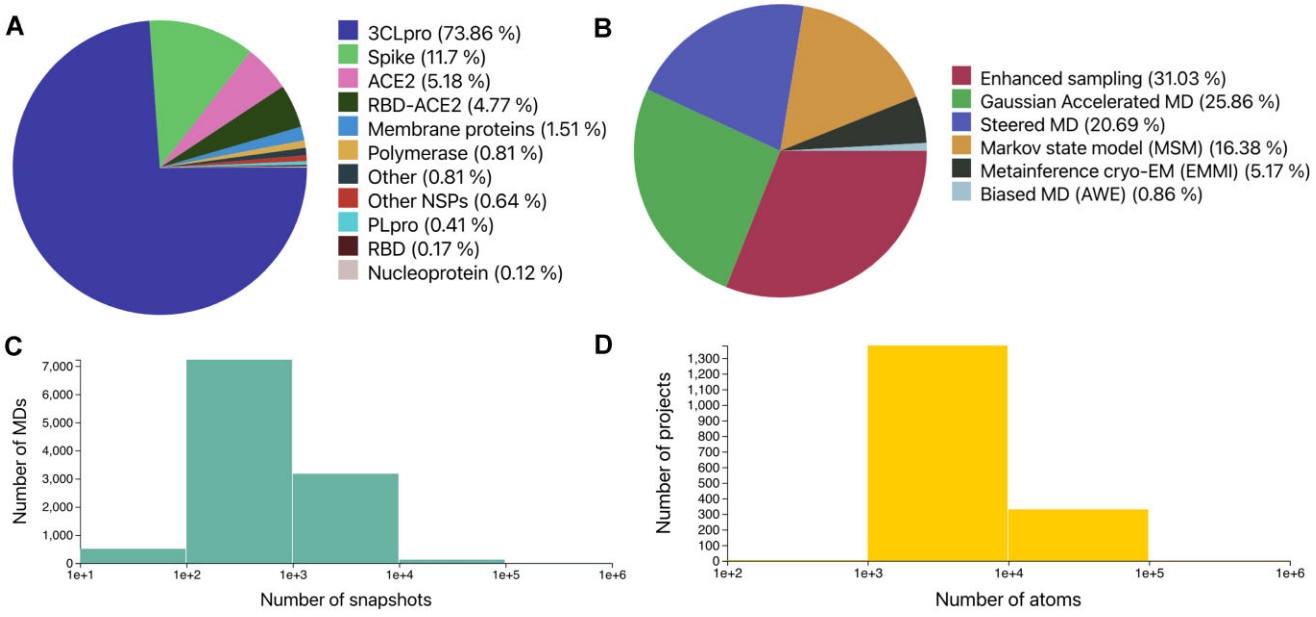


Figure 2. BioExcel COVID-19 MD data statistics (September 2023). (**A**) Simulations divided by SARS-CoV-2 protein units (3CLpro is overrepresented due to the Folding@home COVID-19 Moonshot project https://covid.postera.ai/covid); (**B**) simulations divided by non-classical MD method; (**C**) distribution of simulations by the number of snapshots and (**D**) Distribution of projects by the number of atoms in the system.

Data deposition

The database is open to new simulations and is expected to grow through the addition of new information to be incorporated into the MDDB database infrastructure (https://mddbr.eu/) in the coming months. Deposited data can consist of one or several trajectories (collections) that fulfill a series of requirements:

- Datasets should correspond to MD simulation trajectories of COVID-19-related proteins (e.g. SARS-CoV-2 main protease, Spike, ACE2, etc.), alone or complexed with proteins or small ligands.
- Dehydrated trajectories should be imaged and individual frames superimposed. Trajectory can be split into multiple files if necessary. Accepted formats are explained in the Analysis Pipeline & Database Loader section below.
- Trajectories should be accompanied by a topology file (PDB files are acceptable). The BioExcel COVID-19 internal workflow uses a collection of biomolecular tools for handling trajectory and topology formats. Refer to the tool's documentation for additional information.
- Submission should include a series of mandatory metadata items, but we encourage providing metadata covering the entire form (see BioExcel COVID-19 deposition form in the Appendix of the Supplementary Data).
- The data-providers are responsible for the quality of the deposited trajectory and should fulfill quality requirements shown in (Supplementary Table S2). Additionally, before deposition, the internal analysis pipeline checks the simulation for the following: (i) topology and trajectory coordinate matching; (ii) periodic boundary conditions or imaging problems; (iii) topology problems (e.g. incorrect number of bonds); (iv) root mean square deviations (RMSd, RMSd per residue, pairwise RMSd); (v) radius of gyration (Rgyr); (vi) atomistic fluctuations; (vii) principal component analysis (PCA) and (viii) solvent-accessible surface area (SASA). Future tests will be implemented as accepted by the community. Unusual behaviors (Supplementary Table S2) are communicated to the authors who should confirm if they were expected due to the type of simulation.
- Datasets should ideally (but not mandatorily) be supported by a scientific publication or indexed document (e.g. zenodo). Publications in press or submitted are acceptable, with the possibility to keep data on hold until the publication is available.

We would like to stress here the importance of data and metadata standards. Assessing the quality of the trajectories is not straightforward, especially when working with enhanced sampling or biased simulations. This quality check requires curated information about the simulation, and there is a clear need for a community effort to raise accepted validation metrics.

Analysis pipeline and database loader

In addition to the analysis undertaken in the checking part, the system performs a series of analyses that are system-specific, in most cases proposed by the authors of the simulations (see below). The analyses are performed with a collection of biomolecular tools (Supplementary Table S3), integrated into a reproducible workflow available in GitLab: https://mmb.irbbarcelona.org/gitlab/d.beltran.anadon/MoDEL-workflow. The workflow accepts trajectory and topology formats written by the most popular MD engines (AMBER (27), GROMACS (28), NAMD (29), OpenMM (30) and Desmond (31)). New analyses are constantly being integrated. Note that programmatic access allows tailored analysis by the final user (see below).

The set of analyses is uploaded into the database together with the trajectory data, topology and metadata information (*Project, Topologies, Analyses* and *Trajectories* collections, Figure 1). Furthermore, the loader automatically retrieves data from external biological databases associated with the sequence/structure of interest, thereby extending the amount of accompanying information for each simulation (*SeqRef* and *StructRef* collections, Figure 1).

Programmatic access (REST API)

All the information stored in the database can be accessed programmatically by a REST API: https://bioexcel-cv19.bsc.es/api/rest/docs/, which contains a collection of endpoints divided into the following six categories: *project, references, topology, files, analyses* and *chains*. The endpoints allow a programmatic browse of the database, retrieval of information from the projects, and download of the simulation topology and trajectory. Most of the endpoints take the project accession code (persistent identifier) as input. The *trajectory* endpoint allows the extraction and download of MD trajectories from the database, with the possibility to specify a particular frame range and protein region. As an example, the atomistic coordinates of only the backbone atoms for a representative set of frames, say one every 10 frames, can be directly queried (Supplementary Figure S2). The trajectory is then automatically generated using the power of the noSQL MongoDB. A similar query can be made with the *structure* endpoint to generate a matching structure.

To achieve interoperability between the life sciences and materials disciplines, an implementation of the BioExcel COVID-19 REST API has been developed and deployed, following the Open Databases Integration for Materials Design (OPTIMADE) specifications (https://bioexcel-cv19.bsc.es/optimade/), which are the standard in material science (32).

Meta-analyses

The programmatic access above opens the door to meta-analysis that integrates several simulations. For demonstration purposes, a new section has been added to the interface (Meta-analyses), with a collection of Jupyter Notebooks showing the power of the REST API and how it can be used in combination with Python libraries to extract and graphically display information. The notebooks are also available from GitHub: https://github.com/bioexcel/bioexcel_covid19_workflows. The first example analyzes a collection of 78 simulations of the binding of potential drug molecules to the ectodomain of human ACE2 protein from D.E. Shaw´s group. The workflow determines the drug-protein interaction profile for all the ligands, integrating the information to rank the drugs based on the strength of the interaction (Figure 3A) (see https://bioexcel-cv19.bsc.es/#/id/MCV1900103/energies for an example of Gentamicin-ACE2 drug-protein interaction energies). The final plots allow easy identification of the drug candidates with greater interaction

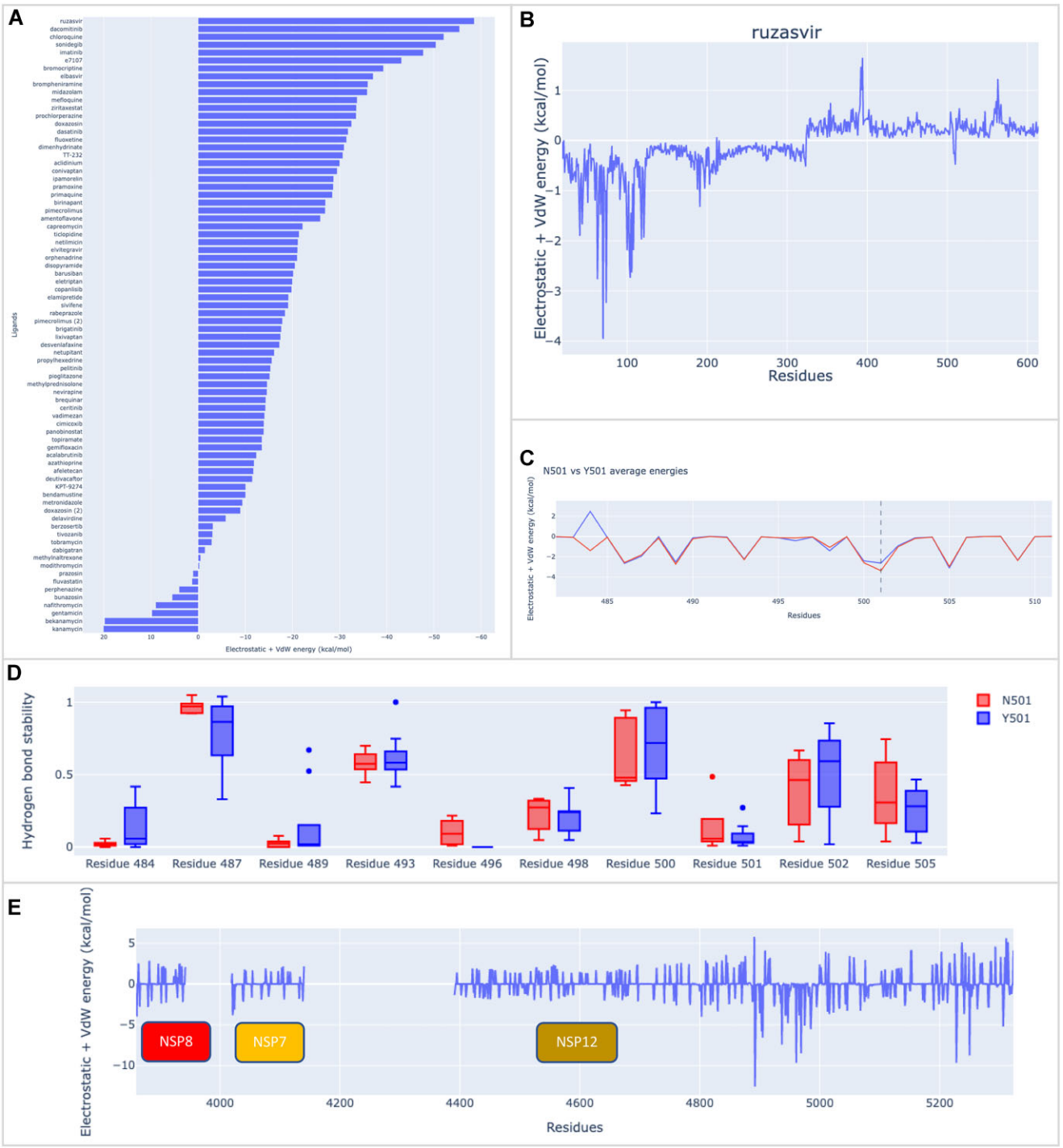


Figure 3. Example of database meta-analyses: (**A**) collection of 78 simulations of the binding of potential drug molecules to the ectodomain of human ACE2 protein ranked by strength of interaction; (**B**) interaction energy profile for the best-ranked drug molecule (Ruzasvir), showing the protein residues with most importance in the interaction; (**C**) RBD-ACE2 protein interaction energy profile comparison (average energies) for two sets of simulations, wild type (N501) and mutated (N501Y); (**D**) stability of the main hydrogen bonds (HBs) involved in the RBD–ACE2 interaction for two sets of RBD-ACE2 simulations, wild type (N501) and mutated (N501Y) and (**E**) RNA-dependent RNA polymerase (RdRp) protein–RNA interaction energy profile, with the NSP8, NSP7 and NSP12 domains highlighted.

(Figure 3A) and exploration of the protein residues that contribute most to the protein-drug binding interaction (Figure 3B). A second example analyses simulations of ACE2 protein interaction with SARS-CoV-2 spike RBD (Receptor Binding Domain) by M. Hongying Chen and coworkers and explores the impact of the N501Y mutation in different variants of the virus on ACE2-Spike binding (33). Although differences in the average interaction energies are mild (Figure 3C), hydrogen bond interactions with the residues surrounding the mutation are altered (Figure 3D). A last example uses a collection of RNA-dependent RNA polymerase (RdRp) simulations to study the interactions of the viral non-structural proteins 12 (NSP12), NSP8 and NSP7 with RNA (Figure 3E). Residues involved in the RNA recognition can be spotted from the interaction energies, which show that while no strong interactions occur with NSP7 and NSP8, strong interactions do take place between RNA and NSP12. The available notebook also exemplifies how to download a number of frames from one of the RdRp simulations and explores the dynamics of these residues through the NGL viewer (34,35).

Data portal

Information stored in the database is accessible through a web-based portal, which offers an easy way to browse and query the data and presents the information in a graphical and interactive way. The interface is divided into sections: *Browse/Search*, *Overview*, *Trajectory*, *Analyses* and *Downloads*.

Browse and search

The whole collection of simulations included in the database are shown in the *Browse* section of the portal (Figure 4A). Accession (persistent id), name of the simulation (short description), unit (Spike, RBD, ACE2, etc.) and available analyses are displayed by entry. The list can be easily filtered using the quick search form at the top of the webpage, which queries the accession number, simulation name, description and author of the trajectory. The table can be also filtered by a particular protein or complex. Each entry displays a list of available analyses, with a direct link to the specific results. The *Search* section allows the user to look for entries using the metadata on the simulation parameters, as well as annotations on the biological role of the protein and COVID-19 related information. This includes searches for variant name, sequence similarity, protein domains, shape of the starting conformation, presence of antibodies/nanobodies and Post Translational Modifications (PTM) (Figure 4B; and Supplementary Table S4). Besides, the possibility to search for organism, gene and protein function is also available, thanks to Uniprot data. Further searches will be implemented based on user's demands. Database collections are indexed to optimize search queries.

Overview

This section of the website presents information about the molecule of interest (Figure 4C). When available, the starting PDB codes are used to retrieve data from the PDB database (36) and link to external databases such as Uniprot (37), PDBe (38,39) and 3DBionotes (40). Simulation metadata include a short description of the simulation, title, method, authors, program, license and citation.

Trajectory

This section shows the simulation using a powerful web component based on the NGL viewer (Figure 4D). The panel is directly connected to the database and it uses streaming technologies to display the trajectory snapshots as they are being transferred. For the purpose of visualization, a video is shown considering a reduced trajectory, with the possibility to select the number of frames. Regions of interest (specific domains, epitopes, residues that are mutated along virus evolution…) can be highlighted. The NGL viewer is highly customizable, with a hidden panel allowing modification of the default representations, the addition of new ones, or the adjustment of the trajectory playback settings and visualization properties (Figure 4E). A selection of the most important simulation parameters, essential for reproducibility purposes, are included just below the visualization panel (Figure 4F). Finally, families, domains and sites revealed by an InterProScan (41) analysis for each of the chain sequences are also presented (Figure 4G). The regions are mapped onto the chain sequences and linked to the NGL viewer to easily identify them on the structure.

Analyses

The database contains a series of system-specific precomputed analyses (performed following feedback from the authors of the simulations; see *Help* pages of the server). Analyses are presented in fully interactive plots, allowing zoom, information on hover and with connection to the NGL viewer, which opens specific frames or highlights specific residues. Analyses can be grouped into three categories (see Analysis Pipeline & Database Loader section): i) general geometrical analysis; ii) interactions; and iii) system-specific analyses. The first ones include global (RMSd, Rgyr, pairwise RMSd, PCA) and local (RMSd per residue, SASA) descriptors (Figure 5A–C). Interaction analyses include Poisson Boltzmann – classical molecular interaction potentials estimate of interaction energies (42), as well as residue contacts and hydrogen bonds (Figure 5D, E). Finally, system-specific analyses show information considered essential by the authors of the simulations and they are specifically coupled to the molecule of interest. Examples of these analyses are Spike mutations or epitopes (Figure 5F, G), or the Markov State Model (MSM) centroids, population and transitions between them. Epitopes have been mined from experimental structures in the PDB, defining the SARS-CoV-2 interface residues as those being at less than 5 Ångstroms from the antibodies. MSM visualization is based on the states and transitions that the authors of the trajectory defined when possible, and is adjusted to a reduced number of states otherwise (e.g. using RMSd between macrostates as edges) (Figure 6).

Downloads

This section gives direct access to the topology and trajectory files. The PDB is used as topology format, while XTC (GROMACS (28)) is the format used to keep the trajectory. Before incorporation into the database, the trajectory is processed (PBC removed, solvent stripped), maintaining the original number of frames from the uploaded (original) simulation. All data used on the website are available to download in *json* format, through a '*Data in this page*' link at the top right corner of the screen, and are programmatically accessible through the REST API (see above).

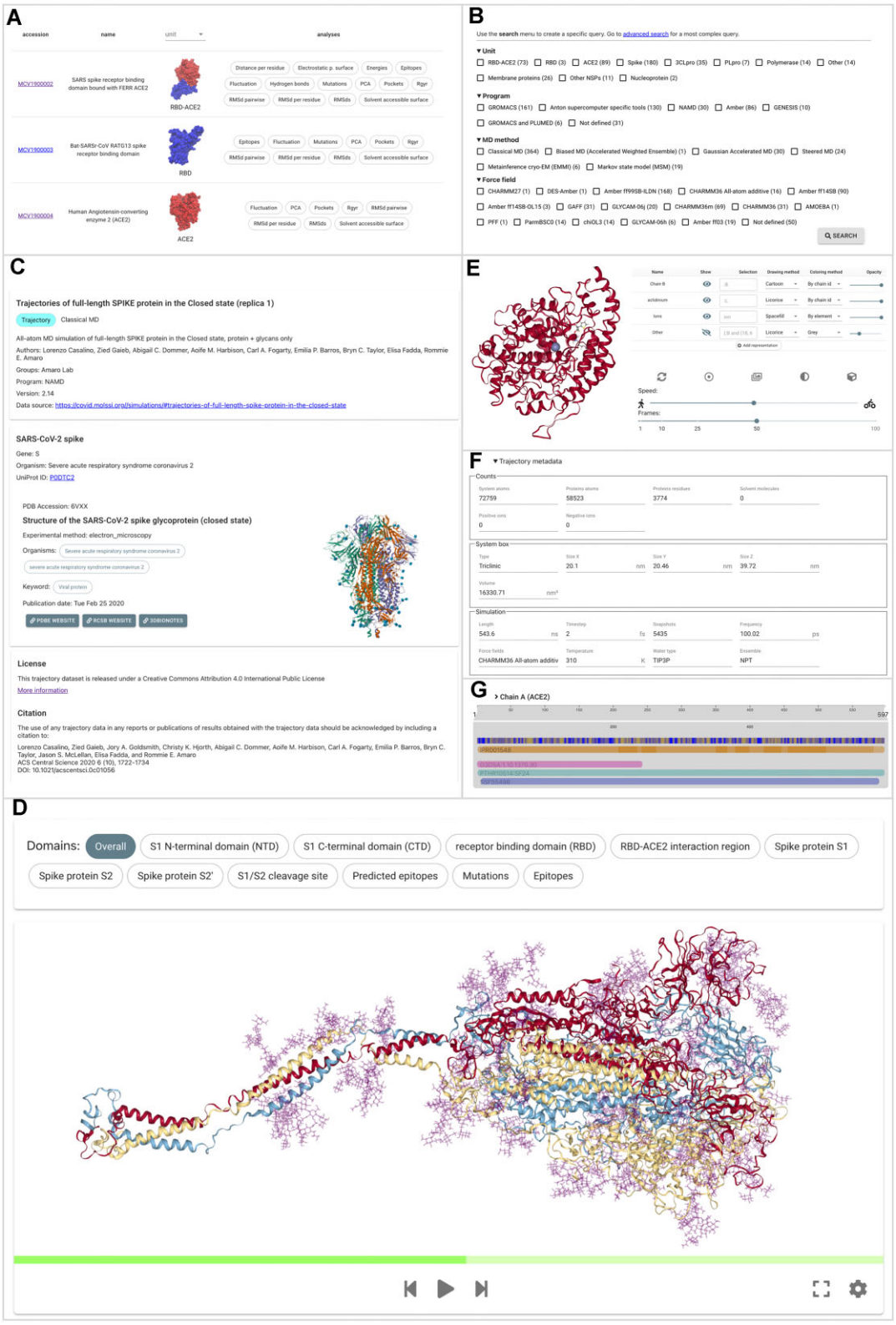


Figure 4. Data portal sections. (**A**) Browse; (**B**) Search; (**C**) Overview, including simulation and molecule metadata, license, citation and links to external databases; (**D**) Trajectory NGL visualization, with NGL customization (**E**), simulation metadata (**F**) and domain highlights (**G**).

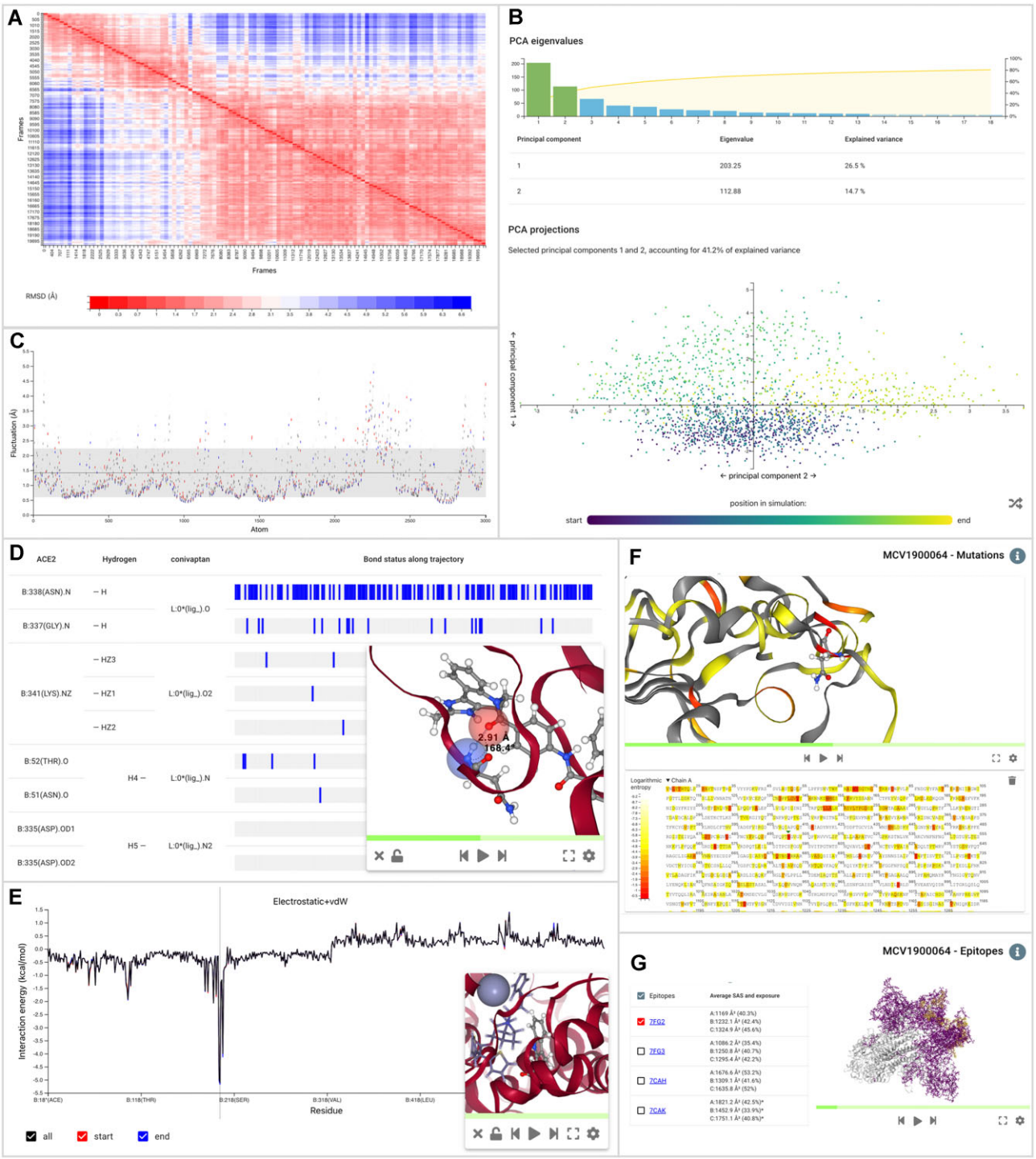


Figure 5. Analysis sections of the server, including: (i) general geometric analyses: pairwise RMSd (**A**); principal component analysis (**B**) and atomistic fluctuation (**C**); (ii) interaction analyses: hydrogen bonds (**D**) and interaction energies (**E**) (electrostatic + VdW) and (iii) system-specific analyses: positions of the most preeminent mutations appearing in the SARS-CoV-2 RBD protein (**F**) and list, position and SASA of known epitopes in the same protein (**G**). All the analyses are interactive and connected to the NGL viewer, opening specific frames or highlighting specific residues depending on the particular analysis (D, E).

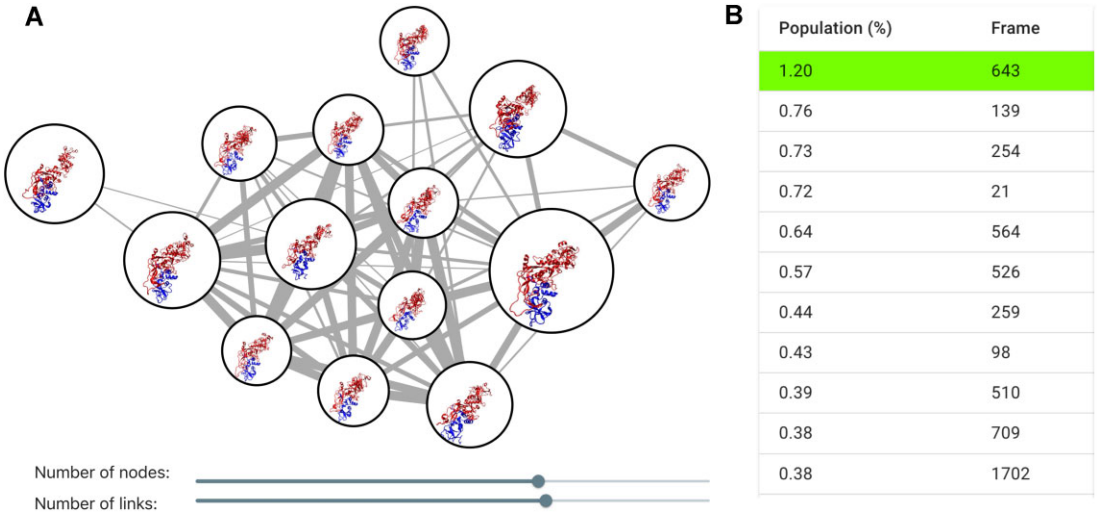


Figure 6. Example of specific analysis for a Markov State Model (MSM) entry in the database. (**A**) Schematic representation of the MSM, including the centroids (graph nodes) and possible transitions (edges); (**B**) population of the MSM centroids, with the associated reference frame in the MSM trajectory. Both the graph and table are interactive and linked to the corresponding structure visualization with NGL viewer.

Conclusions

As MD simulations increase in complexity and accuracy, the need for storing such simulations becomes clearer. Such storage is a requirement to not only guarantee the reproducibility of the results and the robustness of conclusions derived from the simulations but also to facilitate further analysis and meta-analysis by the community. Tools to store trajectories and make them accessible to the MD community are required and should be adopted by the community in the same way as the structural biology community adopted the PDB. A nice example of a community effort was done in the context of the fight against the COVID-19 pandemic. The processing, curating and sharing of these data to facilitate proper analysis are as necessary as they are challenging. Several criteria to perform these steps have been discussed and applied to set up a new database as proof of concept. A new tool for researchers to easily access stored data, both programmatically and through visual interface, has also been developed. Although it is clear that maintaining a centralized infrastructure for MD simulations is not realistic, we believe this project will be the perfect prototype for new federated/distributed infrastructures, where compatible DB deployments will be connected by a central server, avoiding unnecessary data transfer. We expect BioExcel COVID-19 database to lay the groundwork for future MD databases.

Data availability

The server is accessible at https://bioexcel-cv19.bsc.es/#/. It is free and open to all users and there are no login requirements. It is also integrated in the simulations section of the BioExcel-MolSSI *COVID-19 Molecular Structure and Therapeutics Hub*: https://covid.molssi.org/simulations/.

The analyses workflow is available at: https://mmb.irbbarcelona.org/gitlab/d.beltran.anadon/MoDEL-workflow.

The meta-analyses implemented in Jupyter Notebooks are available at: https://github.com/bioexcel/bioexcel_covid19_workflows.

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

We are indebted to all the simulation authors who have kindly shared their data with us. Complete authors list can be found in Supplementary Table S4. We are also indebted to Aurélien Luciani, for his work in the DB design and original implementation.

This work was supported by the Center of Excellence for HPC H2020 European Commission; 'BioExcel Centre of Excellence for Computational Biomolecular Research' (BioExcel-2 [823830]; BioExcel-3 [European Union: 101093290; Ministerio de Ciencia e Innovación: PCI2022-134976-2]); Spanish Ministry of Science [RTI2018-096704-B-100, PID2021-122478NB-I00]; Instituto de Salud Carlos III–Instituto Nacional de Bioinformatica, Fondo Europeo de Desarrollo Regional [ISCIII PT 17/0009/0007]; European Regional Development Fund, ERFD Operative Program for Catalunya, the Catalan Government AGAUR [SGR2021 00863]; European Union 'MDDB: Molecular Dynamics Data Bank. The European Repository for Biosimulation Data' [101094651] and the EU Human Brain Flagship program. We acknowledge the use of Fenix Infrastructure resources, which are partially funded from the European Union's Horizon 2020 research and innovation programme through the ICEI project under the grant agreement No. 800858. IRB Barcelona is the recipient of a Severo Ochoa Award of Excellence from the Spanish *Ministerio de Asuntos Económicos y Transformación Digital* (MINECO). Modesto Orozco is an ICREA Academy scholar. Data for the project is kept on BSC disks thanks to the Red Española de Supercomputación (RES) DATA [DATA-2020-1-0037] project.

Funding

European Union 'MDDB: Molecular Dynamics Data Bank. The European Repository for Biosimulation Data'

[101094651]; BioExcel Centre of Excellence for Computational Biomolecular Research (BioExcel-2 and BioExcel-3) [823830, 101093290]. Funding for open access charge: European and national (Spanish) projects.

Conflict of interest statement

None declared.

References

1. McCammon,J.A., Gelin,B.R. and Karplus,M. (1977) Dynamics of folded proteins. *Nature*, **267**, 585–590.
2. Dror,R.O., Dirks,R.M., Grossman,J.P., Xu,H. and Shaw,D.E. (2012) Biomolecular simulation: a computational microscope for molecular biology. *Annu. Rev. Biophys.*, **41**, 429–452.
3. Huggins,D.J., Biggin,P.C., Dämgen,M.A., Essex,J.W., Harris,S.A., Henchman,R.H., Khalid,S., Kuzmanic,A., Laughton,C.A., Michel,J., *et al.* (2019) Biomolecular simulations: from dynamics and mechanisms to computational assays of biological activity. *WIREs Comput. Mol. Sci.*, **9**, e1393.
4. Wieczór,M., Genna,V., Aranda,J., Badia,R.M., Gelpí,J.L., Gapsys,V., de Groot,B.L., Lindahl,E., Municoy,M., Hospital,A., *et al.* (2023) Pre-exascale HPC approaches for molecular dynamics simulations. Covid-19 research: a use case. *WIREs Comput. Mol. Sci.*, **13**, e1622.
5. Páll,S., Zhmurov,A., Bauer,P., Abraham,M., Lundborg,M., Gray,A., Hess,B. and Lindahl,E. (2020) Heterogeneous parallelization and acceleration of molecular dynamics simulations in GROMACS. *J. Chem. Phys.*, **153**, 134110.
6. Götz,A.W., Williamson,M.J., Xu,D., Poole,D., Le Grand,S. and Walker,R.C. (2012) Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized born. *J. Chem. Theory Comput.*, **8**, 1542–1555.
7. Salomon-Ferrer,R., Götz,A.W., Poole,D., Le Grand,S. and Walker,R.C. (2013) Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald. *J. Chem. Theory Comput.*, **9**, 3878–3888.
8. Mao,R., Zhang,H., Bie,L., Liu,L.N. and Gao,J. (2023) Million-atom molecular dynamics simulations reveal the interfacial interactions and assembly of plant PSII-LHCII supercomplex. *RSC Adv.*, **13**, 6699–6712.
9. Stevens,J.A., Grünewald,F., van Tilburg,P.A.M., König,M., Gilbert,B.R., Brier,T.A., Thornburg,Z.R., Luthey-Schulten,Z. and Marrink,S.J. (2023) Molecular dynamics simulation of an entire cell. *Front. Chem.*, **11**, 1106495.
10. Casalino,L., Seitz,C., Lederhofer,J., Tsybovsky,Y., Wilson,I.A., Kanekiyo,M. and Amaro,R.E. (2022) Breathing and tilting: mesoscale simulations illuminate influenza glycoprotein vulnerabilities. *ACS Cent. Sci.*, **8**, 1646–1663.
11. Dommer,A., Casalino,L., Kearns,F., Rosenfeld,M., Wauer,N., Ahn,S.H., Russo,J., Oliveira,S., Morris,C., Bogetti,A., *et al.* (2023) #COVIDisAirborne: aI-enabled multiscale computational microscopy of delta SARS-CoV-2 in a respiratory aerosol. *Int. J. High Perform. Comput. Appl.*, **37**, 28–44.
12. Coshic,K. and Aksimentiev,A. (2023) The structure and dynamics of a fully packaged RNA virus. *Biophys. J.*, **122**, 443a–444a.
13. Lindorff-Larsen,K., Piana,S., Palmo,K., Maragakis,P., Klepeis,J.L., Dror,R.O. and Shaw,D.E. (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins*, **78**, 1950–1958.
14. Ivani,I., Dans,P.D., Noy,A., Pérez,A., Faustino,I., Hospital,A., Walther,J., Andrio,P., Goñi,R., Balaceanu,A., *et al.* (2016) Parmbsc1: a refined force field for DNA simulations. *Nat. Methods*, **13**, 55–58.
15. Tian,C., Kasavajhala,K., Belfon,K.A.A., Raguette,L., Huang,H., Migues,A.N., Bickel,J., Wang,Y., Pincay,J., Wu,Q., *et al.* (2020) ff19SB: amino-acid-specific protein backbone parameters trained against quantum mechanics energy surfaces in solution. *J. Chem. Theory Comput.*, **16**, 528–552.
16. Hospital,A., Battistini,F., Soliva,R., Gelpí,J.L. and Orozco,M. (2020) Surviving the deluge of biosimulation data. *WIREs Comput.Mol. Sci.*, **10**, e1449.
17. Meyer,T., D'Abramo,M., Hospital,A., Rueda,M., Ferrer-Costa,C., Pérez,A., Carrillo,O., Camps,J., Fenollosa,C., Repchevsky,D., *et al.* (2010) MoDEL (Molecular Dynamics Extended Library): a database of atomistic molecular dynamics trajectories. *Structure*, **18**, 1399–1409.
18. Rueda,M., Ferrer-Costa,C., Meyer,T., Pérez,A., Camps,J., Hospital,A., Gelpí,J.L. and Orozco,M. (2007) A consensus view of protein dynamics. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 796–801.
19. van der Kamp,M.W., Schaeffer,R.D., Jonsson,A.L., Scouras,A.D., Simms,A.M., Toofanny,R.D., Benson,N.C., Anderson,P.C., Merkley,E.D., Rysavy,S., *et al.* (2010) Dynameomics: a comprehensive database of protein dynamics. *Structure*, **18**, 423–435.
20. Rodríguez-Espigares,I., Torrens-Fontanals,M., Tiemann,J.K.S., Aranda-García,D., Ramírez-Anguita,J.M., Stepniewski,T.M., Worp,N., Varela-Rial,A., Morales-Pastor,A., Medel-Lacruz,B., *et al.* (2020) GPCRmd uncovers the dynamics of the 3D-GPCRome. *Nat. Methods*, **17**, 777–787.
21. Zivanovic,S., Bayarri,G., Colizzi,F., Moreno,D., Gelpí,J.L., Soliva,R., Hospital,A. and Orozco,M. (2020) Bioactive conformational Ensemble server and database. A public framework to speed up. *J. Chem. Theory Comput.*, **16**, 6586–6597.
22. Torrens-Fontanals,M., Peralta-García,A., Talarico,C., Guixà-González,R., Giorgino,T. and Selent,J. (2022) SCoV2-MD: a database for the dynamics of the SARS-CoV-2 proteome and variant impact predictions. *Nucleic Acids Res.*, **50**, D858–D866.
23. Hospital,A., Andrio,P., Cugnasco,C., Codo,L., Becerra,Y., Dans,P.D., Battistini,F., Torres,J., Goñi,R., Orozco,M., *et al.* (2016) BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data. *Nucleic Acids Res.*, **44**, D272–D278.
24. Sun,R., Li,Z. and Bishop,T.C.T.M.B. (2019) Library of nucleosome simulations. *J. Chem. Inf. Model.*, **59**, 4289–4299.
25. Thibault,J.C., Facelli,J.C. and Cheatham,T.E. (2013) iBIOMES: managing and sharing biomolecular simulation data in a distributed environment. *J. Chem. Inf. Model.*, **53**, 726–736.
26. Thibault,J.C., Cheatham,T.E. and Facelli,J.C. (2014) iBIOMES Lite: summarizing biomolecular simulation data in limited settings. *J. Chem. Inf. Model.*, **54**, 1810–1819.
27. Case,D.A., Cheatham,T.E., Darden,T., Gohlke,H., Luo,R., Merz,K.M., Onufriev,A., Simmerling,C., Wang,B. and Woods,R.J. (2005) The Amber biomolecular simulation programs. *J. Comput. Chem.*, **26**, 1668–1688.
28. Abraham,M.J., Murtola,T., Schulz,R., Páll,S., Smith,J.C., Hess,B. and Lindahl,E. (2015) GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, **1–2**, 19–25.
29. Phillips,J.C., Braun,R., Wang,W., Gumbart,J., Tajkhorshid,E., Villa,E., Chipot,C., Skeel,R.D., Kalé,L. and Schulten,K. (2005) Scalable molecular dynamics with NAMD. *J. Comput. Chem.*, **26**, 1781–1802.
30. Eastman,P., Swails,J., Chodera,J.D., McGibbon,R.T., Zhao,Y., Beauchamp,K.A., Wang,L.P., Simmonett,A.C., Harrigan,M.P., Stern,C.D., *et al.* (2017) OpenMM 7: rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.*, **13**, e1005659.
31. Schrödinger (2021) *Schrödinger Release 2022-1: Desmond Molecular Dynamics System, D.E.Shaw Research, New York, NY, 2021*. Maestro-Desmond Interoperability Tools. Schrödinger, New York, NY.
32. Andersen,C.W., Armiento,R., Blokhin,E., Conduit,G.J., Dwaraknath,S., Evans,M.L., Fekete,Á., Gopakumar,A., Gražulis,S., Merkys,A., *et al.* (2021) OPTIMADE, an API for exchanging materials data. *Sci. Data*, **8**, 217.

33. Cheng,M.H., Krieger,J.M., Banerjee,A., Xiang,Y., Kaynak,B., Shi,Y., Arditi,M. and Bahar,I. (2022) Impact of new variants on SARS-CoV-2 infectivity and neutralization: a molecular assessment of the alterations in the spike-host protein interactions. *iScience*, **25**, 103939.
34. Nguyen,H., Case,D.A. and Rose,A.S. (2018) NGLview-interactive molecular graphics for Jupyter notebooks. *Bioinformatics*, **34**, 1241–1242.
35. Rose,A.S., Bradley,A.R., Valasatava,Y., Duarte,J.M., Prlić,A. and Rose,P.W. (2018) NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics*, **34**, 3755–3758.
36. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
37. The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
38. Consortium, P.D.-K. (2019) PDBe-KB: a community-driven resource for structural and functional annotations. *Nucleic Acids Res.*, **48**, D344–D353.
39. Armstrong,D.R., Berrisford,J.M., Conroy,M.J., Gutmanas,A., Anyango,S., Choudhary,P., Clark,A.R., Dana,J.M., Deshpande,M., Dunlop,R., *et al.* (2020) PDBe: improved findability of macromolecular structure data in the PDB. *Nucleic Acids Res.*, **48**, D335–D343.
40. Macias,J.R., Sanchez-Garcia,R., Conesa,P., Ramirez-Aportela,E., Martinez Gonzalez,M., Wert-Carvajal,C., Parra-Perez,A.M., Segura Mora,J., Horrell,S., Thorn,A., *et al.* (2021) 3DBionotes COVID-19 edition. *Bioinformatics*, **22**, 4258–4260.
41. Quevillon,E., Silventoinen,V., Pillai,S., Harte,N., Mulder,N., Apweiler,R. and Lopez,R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.*, **33**, W116–W120.
42. Gelpí,J.L., Kalko,S.G., Barril,X., Cirera,J., de La Cruz,X., Luque,F.J. and Orozco,M. (2001) Classical molecular interaction potentials: improved setup procedure in molecular dynamics simulations of proteins. *Proteins*, **45**, 428–437.

Received: August 14, 2023. **Revised:** October 16, 2023. **Editorial Decision:** October 16, 2023. **Accepted:** October 17, 2023