Received: 2 February 2022 | Revised: 25 April 2022 | Accepted: 28 April 2022

DOI: 10.1002/wcms.1622

FOCUS ARTICLE

WIREs COMPUTATIONAL MOLECULAR SCIENCE WILEY

Pre-exascale HPC approaches for molecular dynamics simulations. Covid-19 research: A use case

Miłosz Wieczór[1,2] | **Vito Genna**[1] | **Juan Aranda**[1] | **Rosa M. Badia**[3] | **Josep Lluís Gelpí**[3,4] | **Vytautas Gapsys**[5] | **Bert L. de Groot**[5] | **Erik Lindahl**[6,7] | **Martí Municoy**[8] | **Adam Hospital**[1] | **Modesto Orozco**[1,4]

[1]Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology, Barcelona, Spain

[2]Department of Physical Chemistry, Gdansk University of Technology, Gdańsk, Poland

[3]Barcelona Supercomputing Center, Barcelona, Spain

[4]Department of Biochemistry and Biomedicine, University of Barcelona, Barcelona, Spain

[5]Max Planck Institute for Multidisciplinary Sciences, Computational Biomolecular Dynamics Group, Goettingen, Germany

[6]Department of Applied Physics, Swedish e-Science Research Center, KTH Royal Institute of Technology, Stockholm, Sweden

[7]Department of Biochemistry and Biophysics, Science for Life Laboratory, Stockholm University, Stockholm, Sweden

[8]Nostrum Biodiscovery, Barcelona, Spain

Correspondence
Adam Hospital and Modesto Orozco, Institute for Research in Biomedicine (IRB Barcelona). The Barcelona Institute of Science and Technology, Barcelona, Spain.
Email: adam.hospital@irbbarcelona.org and modesto.orozco@irbbarcelona.org

Funding information
European Commission (BioExcel-2 project), Grant/Award Number: 823830; Instituto de Salud Carlos III, Grant/Award Number: PT17/0009/0007; Ministerio de Ciencia e Innovación, Grant/Award Numbers: PID2020-116620GB-I00, RTI2018-096704-B-100

Edited by: Peter Schreiner, Editor-in-Chief

Abstract

Exascale computing has been a dream for ages and is close to becoming a reality that will impact how molecular simulations are being performed, as well as the quantity and quality of the information derived for them. We review how the biomolecular simulations field is anticipating these new architectures, making emphasis on recent work from groups in the BioExcel Center of Excellence for High Performance Computing. We exemplified the power of these simulation strategies with the work done by the HPC simulation community to fight Covid-19 pandemics.

This article is categorized under:

Data Science > Computer Algorithms and Programming

Data Science > Databases and Expert Systems

Molecular and Statistical Mechanics > Molecular Dynamics and Monte-Carlo Methods

KEYWORDS

BioExcel, COVID19, exascale, molecular dynamics

1 | NEW COMPUTER ARCHITECTURES

Biomolecular simulations have been linked to computer power since their origin in the 1960s. Thus, access to the Golem computer explained (in part) the leadership of Weizmann's institute in the development of the first force-field

Milosz Wieczor, Vito Genna, and Juan Aranda have contributed equally to this study.

based methods oriented towards the study of biomacromolecules,[1,2] and a large computer placed at CECAM headquarters allowed the first molecular dynamics (MD) simulation of a protein by McCammon, Gellin, and Karplus.[3] Since the 1970s, continuous improvements in computers have allowed molecular simulations, and in particular MD to be established in the structural biology and biophysics worlds.

The increase in computer power seemed unstoppable for decades but, unfortunately, current chip technology has reached a plateau and we cannot expect much more powerful processors to emerge from current computer technology. Thus, while waiting for quantum computers, hardware developers have tried to maintain the increase in processing power by three different strategies: (i) to build specific purpose processors, (ii) to take advantage of secondary processors specialized in basic numerical operations (like the graphical processing units; GPUs) and (iii) to increase the number of cores in the computers.

MD-specific computers such as Anton[4,5] were a major promise for the MD field when first appeared in 2008, but 13 years later, their global impact has been small. On the other hand, GPUs have revolutionized the field allowing small groups access to significant computing power. Yet, no single GPU can perform massive MD simulations, thus requiring the combination of a large number of CPUs (and eventually GPUs) in the simulation. Thus, most massive MD simulations are performed in High-Performance Computing (HPC) environments, especially in gigantic supercomputers created by the aggregation of a massive number of CPUs or CPU/GPUs. Two excellent examples of Tier-0 supercomputers are Japanese Fugaku and American Summit. Rikken supercomputer Fugaku has 7.6 million general-purpose cores A64FX 48C and provides 0.5 ExaFlop of peak power, while American Summit has a hybrid CPU/GPU architecture based on Power9 and NVIDIA Tesla V100 cards, with 2.4 million cores and 0.2 ExaFlop of peak power. Both are amazing pieces of engineering and are approaching to the mythical ExaFlop barrier, but we cannot ignore that Fugaku is only four times faster than the top-of-the line computer five years ago (Chinese Sunway) and has an electrical consumption (close to 30 MW excluding cooling) that raises doubts on the sustainability of a much bigger computer.[6]

In summary, the biosimulation community has largely benefitted from decades of continuous increase in computer power, but we cannot expect much faster processors in the next years. Most likely, increases in aggregated computer power would appear linked to massively parallel architectures. Our community, and in particular the molecular dynamics (MD) field should redefine its objectives and tools having in mind realistic expectations of how new high-performance computers will be.

2 | TAKING ADVANTAGE OF HIGH PERFORMANCE COMPUTING

The quality of an MD simulation depends mainly on three factors: (i) the accuracy of the Hamiltonian representing molecular interactions, (ii) the similarity between the simulated system and the physical reality, and (iii) the exhaustiveness of the sampling of the conformational space. Most MD simulation engines use the same type of classical Hamiltonian defined in the 1960s by Lifson and coworkers,[7] which have been largely refined, reaching a good agreement with experimental measurements when sufficient convergence is achieved for some macromolecular systems.[8–11] It means that in practice, the quality of MD calculations mainly depends on the ability of the simulated model to represent the reality, and on the completeness of the sampling of the relevant configurations. As some of the biologically relevant systems may contain millions of atoms, a tendency in the MD field in the last decade has been to simulate very big systems[12–17] which are expected to be more realistic and better fitting with supercomputer architectures. Unfortunately, the larger the system, the longer the simulation generally should be, and the more complex the force evaluation is. In brief, the reliable treatment of very large systems is limited to a few computational groups with access to massive computer resources, while in everyday practice the simulation system sizes range from 10^4 to 10^5 atoms. The challenge is how to efficiently use supercomputers on biological problems involving such "small" systems.

2.1 | Strategies for parallelization of a single calculation

Molecular dynamics simulation approaches are based on the idea of sampling phase space, and as such the efficiency of the method ultimately comes down to the performance of individual simulations. The total throughput of methods relying, for example, on ensembles is multiplied by the speedup due to parallelization of individual simulations, which in turn is multiplied by speedups from acceleration and algorithmic advances. Several modern MD implementations such as Amber,[18] NAMD,[19] Desmond,[20] and GROMACS[21] have been accelerated either for GPU hardware—using CUDA,

OpenCL or SYCL APIs—or single-instruction multiple-data instruction units available on modern CPUs.[22] However, both hardware and software landscapes have become extremely diverse, and while almost all codes can be compiled on arbitrary systems, it cannot be taken for granted that a code is able to make good use of the hardware available on a particular system. For large projects, it has become an important decision to select a suitable combination of HPC hardware along with optimally accelerated simulation software already at the planning stage. GPU accelerators have been a revolution for MD simulations, just as for many other scientific fields, which is effectively due to new algorithms enabling parallelization of the force evaluation in each timestep over ∼10 000 functional units in each GPU,[23–25] although this is opaque to the user.

On the other end of the spectrum, algorithms to explicitly parallelize simulations by communicating between multiple different nodes have evolved to use advanced load balancing and advanced domain decomposition techniques that limit communication to the neighboring nodes[26] and minimize communication, for example, by allowing an interaction between atoms present on two specific nodes to be evaluated on a third ("neutral-territory" methods).[27,28] While benchmarks occasionally use gigantic systems to showcase codes, biomolecular applications like Covid-19 are typically focused on achieving sampling, for example, for a protein by first reducing the system size as much as possible and the parallelizing the simulation of this given system over as many nodes as possible—so-called "strong scaling", which is considerably more difficult. The algorithms above have enabled truly impressive strong scaling where simulations have been able to make use of 10,000 cores or more,[19,29] resulting in long time scale simulations of individual SARS-CoV-2 proteins,[30] sampling of their binding site dynamics[31], conformational changes,[32] modeling of SARS-CoV-2 proteins whose structure is not yet known,[33] and multiscale simulations, for example, of the SARS-CoV-2 spike protein.[34]

One of the currently most critical bottlenecks is that the particle-mesh Ewald summation[35] algorithms used for long-range electrostatics involve global 3D fast fourier transforms, where the number of communication messages scale as the square of the number of nodes. Fortunately, there are several promising new algorithms in development based, for example, on multipole expansion methods[36,37] or multi-level summation.[38] However, since molecular dynamics is inherently an iterative stepwise method, high-end parallelism is subject to hard challenges of communication bandwidth and latency—it will not help to add more CPU resources to a simulation waiting for communication. These problems are even more severe when compounded with accelerators. First, some GPU codes do not even support multi-node parallelism since the random-access nature of indexing atoms to be communicated is not a great fit for accelerator hardware. Second, it introduces another layer of communication complexity and latency when data must first be sent between GPUs in each node and between CPU and GPU before it can be communicated to other nodes. Third, since the force calculation can be an order of magnitude faster on GPUs, the relative impact of network bandwidth and latency will also become tenfold higher, and thus dominate for much smaller node counts compared with CPU-only simulations. As an example, for GROMACS the current effective limit to strong scaling is somewhere around 40 atoms per CPU core, while it is typically not meaningful to go to <3000 atoms per GPU,[23] and other major codes have qualitatively similar behavior. Notably, this means the relative strong scaling is typically considerably worse for GPU-accelerated simulations, and somewhat paradoxically they might not improve the absolute performance of the simulations in the limit of high-end parallelism, but they make it possible to reach the same performance with order-of-magnitude smaller hardware resources. These freed-up resources can then instead be utilized much more efficiently to parallelize the solution to the scientific problem through combination with ensembles of many accelerated and parallel simulations.

2.2 | Ensemble simulations

In effect, a direct alternative to escape the technical shortcomings of direct parallelism is based on the use of ensembles obtained from independent or loosely coupled simulations. The use of *replicas* is now well established and considered part of the "good practices" in the field,[39,40] especially when one is concerned with simulations of rare events.[41] As an example, in our 2010 study of spontaneous DNA hairpin folding,[42] we were lucky to have found a fast "downhill" folding pathway in our first trajectory (1 μs long). However, to obtain good statistics, we had to collect more than 20 such trajectories, and yet four of them never reached the folded state (after 4 μs), highlighting the variability in the outcomes of individual simulations. In fact, thanks to the chaotic nature of MD simulations, even initially correlated replicas rapidly lose the memory of their common initial state and can thus be analyzed independently to obtain statistical bounds on a given MD-derived quantity. Even more interestingly, they can be also combined to create meta-trajectories that provide ensembles much richer than expected from the length of individual replicas. In this spirit, it was shown that

the multiple replica approach can be used not only to describe equilibrium properties, but also nonequilibrium transitions, even when their characteristic timescales are much larger than those sampled by the individual replicas.[43–45] In this spirit, Pande, Noe, and others[46,47] have demonstrated that short replicas can be combined into a *Markov State Model* (MSM) to trace complex kinetic pathways from the transition probabilities between a discrete number of states in the system, even when no single trajectory samples transitions between all states. In recent years, many generalizations of this scheme have been developed, including history-augmented MSMs that go beyond the simple theory of memoryless Markov processes,[48] or the transition-based reweighting analysis method (TRAM) that combines data from multiple ensembles, including explicitly biased ones.[49]

Indeed, although an unbiased multiple-replica approach is often the most easily interpretable, the use of biasing schemes usually offers better sampling of the configurational space. This is especially important for the exploration of slow conformational transitions that are unlikely to happen within currently accessible simulation times. An example of this approach is *Hamitonian or Temperature replica exchange*[50–53] where either the Hamiltonian or the kinetic energy is perturbed across replicas to improve sampling efficiency by pushing the systems out of metastable states with long residence times. An equally popular intermediate approach, called Solute Tempering (REST), combines these two ideas to perturb the effective temperature of a subsystem of interest.[54,55] In *Free Energy Perturbation (FEP) or Thermodynamic Integration (TI)*, the Hamiltonian is perturbed in discrete steps (FEP, discrete TI) or smoothly (slow growth TI, nonequilibrium TI) from one state to the other, allowing for the calculation of the associated work—which, depending on the Hamiltonian at hand, can be coupled to a conformational transition or an alchemical change.[56–59] In *umbrella sampling* or *metadynamics*,[60–64] additional terms are added to the energy function to favor (in a guided or unguided manner) the exploration of a selected collective variable, often in conjunction with the replica exchange scheme.[51,65] In the *string method*, trajectories are pinned to points forming a path in a low-dimensional subspace, and the displacements tangent and perpendicular to the path are used to iteratively refine this path, until a lowest-energy pathway is found between two initially defined states.[66–68]

Subtler biasing techniques imply the use of information-based biasing as in *Maxwell-Demon Dynamics*, where after a certain time only trajectories approaching a given target are extended or cloned,[69,70] or adaptive approaches (very popular in metadynamics or MSMs), which launch new trajectories from regions of the configurational space that have been undersampled in the original ensemble. As a notable example, the Weighted Ensemble (WE) method provides a rigorous scheme for evolving large batches of short trajectories to obtain kinetic and thermodynamic data without the application of external forces.[71,72] Recently, machine learning approaches hold promise to enhance sampling efficiency by learning a refined biasing scheme from the collected trajectories,[73,74] by re-optimizing the reaction coordinate that is being explored on-the-fly,[75,76] by learning an approximate structural ensemble from which new uncorrelated seeds can be drawn,[77] or by simplifying the structural description of the system without compromising on its physical accuracy.[78] Note that while unbiased replica approaches rely on completely independent trajectories, biased multiple replica approaches usually imply a loose coupling between trajectories, as decisions on a given trajectory should be made periodically based on the results obtained in the others.

The multiple replica approach fits perfectly into massively parallel computers, as each of the simulations is independent (or nearly independent) from the others and as the CPU time required to compute the forces is quite independent of the specific coordinates, a good processor load balance is expected. In fact, unbiased multiple replica approaches can be run in distributed platforms, where volunteers provide access to their personal computers to run independent simulations, which are then integrated to obtain the representation of the system. As the first such distributed setup to achieve Exaflop performance in 2020,[79] the *Folding@home* project[80] is a paradigm for this type of initiatives that has advanced the study of many complex biophysical processes[81–85] and provided excellent results in SARS-Cov2 research (see below).

There are several desirable properties that should characterize a well-scaling ensemble approach fitted to the Exascale era:

- Direct parallelism should ideally be replaced with "trivial" parallelism, meaning that ideally ensembles should be generated by a very large number of short, independent trajectory segments.
- The ideal method should be able to keep the number of parallel runs constant to make resource allocation efficient and predictable.
- Explicit communication between runs should be avoided or minimized, and ideally performed asynchronously as not all computing units perform equally even within a single machine.

- Disk space requirements should be minimized, since 1000s of individual runs can easily overwhelm even purpose-built filesystems.
- Efficient GPU implementations should be considered alongside more trivially paralellized CPU ones to ensure compatibility across all types of HPC resources.
- Finally, popular, out-of-the-box parallelization schemes should be used to ensure interoperability across the increasingly diverse computational cluster architectures.

When choosing a method, one should always consider the availability of resources, the existence of "natural" reaction coordinates, the expected complexity of the free energy landscape, the natural timescale of the process in question, and the desired levels of detail and accuracy in the final model.

2.3 | Non-equilibrium simulations

Biased replica approaches are typically implemented in the context of equilibrium simulations. For example, in metadynamics, umbrella sampling, or FEP/TI, we expect that the trajectory collected at a given point of the transition (geometrical or alchemical) samples well the neighboring points along the reaction coordinate, and the reversible work associated with this microscopic change will not change by extending the trajectory, which often requires long simulations. Similarly, producing a well-converged MSM requires that the transition probabilities between any two states reach local equilibrium, that is, the trajectory samples an equilibrium situation, which requires long aggregated trajectory times. The need to run a limited number of resource-intensive replicas is undesirable for HPC, as it limits the number of processors that can be allocated to a specific scientific problem.

A potential solution arises from the use of non-equilibrium methods, which allow us to obtain transition free energies by combining a very large number of "irreversible works" derived from short non-equilibrium simulations starting from equilibrium start and end states. Non-equilibrium methods have a solid foundation in Jarzynski equation[86] and its later extension, the Crooks fluctuation theorem.[87] They provide access to equilibrium properties, for example, free energy differences.[88] and are efficient in an HPC environment, as they allow for a distributed use of processors working independently. Within the BioExcel consortium,[89] De Groot and coworkers have pushed the limits of these non-equilibrium techniques in the context of alchemical perturbation, developing a full methodology that goes from defining an optimum pathway for the transition to collecting converged free energy results for the alchemical transition.[90,91] This methodology is becoming very successful in exploring relative binding free energies of related drugs or the impact of mutation in protein interactions[88,92] and has been extensively used in the context of Covid19 research (see below). Note that the technique can also be adapted to biasing schemes related to geometrical rather than Hamiltonian changes, even though initial attempts exposed a number of specific challenges that need to be addressed if a general workflow is to be established.[93,94]

2.4 | Multi-simulation projects and workflows

As already discussed, a typical research project cannot be solved by running a single simulation, but requires combining many of them. For example, to determine the structural role of sequence variants on a short loop of a protein, at least 20n (with n equal to the length of the loop) independent simulations need to be done. The Folding@home strategy is well suited for this type of problems as we are referring to completely independent and asynchronous simulations. There are, however, other cases that require combining different types of simulations that must be arranged using flexible workflows. For example, before running a free energy calculation on the impact of polymorphism on the binding of a large series of drugs to a given target, many calculations need to be done: 3D models and topologies of the drugs need to be generated, force-field parameters have to be determined, and equilibrated structures need to be obtained. In parallel, 3D models of the target protein and its mutants have to be made, optimized, thermalized and equilibrated through MD simulations of both apo and holo forms of the protein. Finally, all these processes must be repeated for hundreds of drugs and sequence variants to reach biologically relevant conclusions, and after some analytics on all the results can be needed. In this context, the cost of running ultra-efficient free energy calculations might have no impact in the global project, as the limiting step can be the manual setup of chains of calculations. Thus, in order to be efficient, manual interventions need to be reduced to a minimum by using suitable applications[95–97] automating all the preparation

steps. Furthermore, the execution of the different programs has to be coordinated by a workflow manager to guarantee no lag time between calculations.

Workflows have been very popular in bioinformatics,[98–102] but their adoption in the biomolecular simulation fields has been slower,[103,104] mainly due to the powerful computational resources needed and the large amounts of data generated. The recent advances in the informatics technology, including faster network bandwidths, optimized and distributed file systems and pre-exascale HPC supercomputers opened the door to biomolecular simulation execution pipelines. The combination of automated workflows with new optimized codes and HPC supercomputers[21,105] allows an efficient usage of the available resources and the implementation of large studies through task parallelization. We anticipate that workflows will be at the core of biosimulations projects in a near future.

Execution of workflows is typically controlled by Workflow Management Systems (WFMS): the software that regulates the execution of a defined sequence of tasks, arranges them in the most efficient way, manages intermediate data, and monitors the whole execution. The list of available WFMS is huge, and is still increasing every year (see the list of workflow managers compiled and regularly updated by the Common Workflow Language (CWL) team[106] at https://github.com/common-workflow-language/common-workflow-language/wiki/Existing-Workflow-systems). Some of these tools also allow for workflow setup by means of Graphical User Interfaces (GUIs).[107–113] Very popular in the bioinformatics field, these drag-and-drop interfaces (KNIME,[107] Pipeline-Pilot,[108] Taverna[109], Galaxy,[110] Unipro UGENE,[111] Kepler[112]) allow for very easy workflow development based on interconnected nodes (building blocks). The chemoinformatics and MD fields have just recently developed GUI-based WFMS, with some available implementations in Galaxy,[113] KNIME,[114–116] Kepler,[117] and Pipeline-Pilot,[108] the most popular being integrated into commercial software. Although most of these WFMS can be used to launch jobs in HPC clusters through integrated remote execution functionalities, they were not designed to run massive HPC jobs.

For this reason, the biomolecular simulation field should rely on a different type of WFMS adapted to work with HPC infrastructure, eventually reaching the Exascale regime. BioExcel CoE joined together some of the most important HPC WFMS around the world in a dedicated workshop (2018), with the goal of identifying specific software gaps and opportunities for improved workflow practices.[118] A representation of the most popular WFMS in the field were present in the workshop: Parsl[119], AdaptiveMD,[120] PyCOMPSs,[121] BioSimSpace,[122] LongBow,[123] FireWorks,[124] Swift,[125] ExTASY,[126,127] and RADICAL.[128,129] Although sharing many of the most important features (automation, data control flow, task management, dependency graph), each WFMS has its own design goal, and adoption of one versus another implies a thorough analysis of the particular research project needs and the WFMS features. Some of the differentiating elements that were identified are the possibility to run adaptive algorithms (ability to change their behavior in runtime), fault tolerance (ability to distinguish between critical and non-critical task failures), flexibility (ability to extend/reduce the number of resources used at runtime), and provenance (ability to store input metadata for reproducibility purposes). Some of these tools have been designed to cover very specific applications, whereas others are able to handle broader, more generic workflows. Finally, the installation and configuration overhead is a critical point that should be carefully considered before choosing a WFMS.

HPC-focused WFMS have demonstrated their power in biomolecular research projects, and Molecular simulation pipelines are now regularly being used to gain insights into the binding free energy and the residence time of ligands in drug development studies.[130] They paved the way for the classical protein-folding problem before AlphaFold2 appeared[131,132] and have been extensively used in recent studies helping the fight against the SARS-CoV-2 virus[34,133–136] (see below). However, despite their obvious impact, HPC workflows are still not widely adopted by the biosimulation community, mainly due to the difficulty of integrating them with the most common simulation tools. One step towards such integration is the BioExcel Building Blocks (BioBB) library: a collection of portable wrappers on top of common biomolecular simulation tools.[137] The library is designed to: i) increase the interoperability between the tools wrapped; ii) facilitate the implementation of biomolecular simulation workflows; and iii) increase the reusability and reproducibility of the generated workflows. The library is being developed following the FAIR principles for research software development best practices. The result is a collection of building block modules, classified according to the functionalities offered (e.g., MD, analysis, chemistry). The BioBB library is compatible with different WFMS, including both GUIs (Galaxy, KNIME) and HPC-focused (PyCOMPSs). Installation of BioBB workflows controlled by PyCOMPSs WFMS can be easily done using BioConda packages.[138]

The community has no doubt that biomolecular simulation workflows will have a clear impact on the future grand challenges in science,[139] and that they will be essential tools for the efficient exploitation of the future Exascale supercomputer generation. However, important aspects linked to the pipeline execution still need further research and discussion to reach a community agreement. Data provenance, historical record of the data;[140,141] workflow metadata,

important information from the input, output and intermediate data; reproducibility, possibility to generate the same results using a particular pipeline;[142,143] reusability, possibility to use a particular pipeline in different resources;[106] and FAIR computational workflows[144] are all crucial points whose treatment is still under debate.

3 | THE PROBLEM OF DATA

As computer power increases, the management of data is becoming the Achilles heel of molecular simulation. As described in a recent review,[145] 1 μs of a medium-sized protein system (10^5 atoms including solvent) generates around 5 Pb of data, and even if only 1‰ of coordinates are stored, a single trajectory would require 5 Tb of disk space. The common practice in the field has been to store trajectories locally and delete them after some period, which means that very valuable (and expensive to obtain) information is lost. FAIR (Findable, Accessible, Interoperable and Reusable) requirements for scientific data[146] force the storage of plain trajectories and associated simulation metadata. This generates a formidable problem that requires at least: (i) efficient metadata to specify the purpose and conditions of the simulation, (ii) efficient ways to store the raw data, (iii) efficient transfer mechanisms, and (iv) flexible analysis portals including virtual research environments (VRE) for remote programmatic access to the data.[147]

The BioExcel CoE and the ELIXIR European Initiative have defined a prototype of EDAM ontologies for metadata of MD simulations.[148] This should be extended and refined to be adopted by already existing[10,149–155] or new simulation databases. Some recent databases such as BigNASim already implement an ontology metadata[154] which helps to classify trajectories, favoring then meta-analysis and the detection of artefactual simulations. Defining what should be stored from a trajectory is a major decision, which is often guided by traditions maintained since times when the throughput of HPC systems was a fraction of the current one. A significant work has been done by communities such as ABC[151,156,157] to define the maximum acceptable frequency for storing data and the amount of solvent that needs to be maintained for reanalysis. In parallel, software developers are creating compression approaches and importing methods used to compress other types of massive data.[158] Again, the community should reach a consensus on the type of compression strategy and the amount of data to be stored, otherwise the FAIR requirement will not be fulfilled.

Despite cleaver approaches to improve transfer efficiency (GridFTP, Aspera, and others; see[145]), current network limitations are severely hampering transferring raw trajectories. Streaming methods that would allow basic analysis in remote before downloading the trajectory are going to be useful to avoid unnecessary downloads[145]. However, it seems that the most sensible approach is to reduce the movement of MD files, with big data centers placed physically close to the computer power concentrating the data, mimicking current situation in other fields such as bioinformatics or genomics. The first MD databases were available around 10–15 years ago and aimed to cover all representative structures in the protein data bank: Dynameomics, MoDEL, or Dynasome[10,149–151] are examples of these families of databases. More recent efforts have been focused on specific families of proteins of special importance, such as GPCRmd that includes trajectories of G-protein coupled receptors, MemProtMD,[156] which contains simulations on 3500 membrane proteins in realistic membrane environments, or TMB-iBIOMES that includes hundreds of nucleosome trajectories.[152] A few databases are focused on small molecules with potential biomedical or biomolecular impact; an example is Cyclo-lib[153] which is focused on cyclodextrins, or BCE,[155] which contains classical trajectories and quantum structural data on small drugs. Finally, BigNASim[154] is, to our knowledge, the only database focused on nucleic acids simulations, and contains hundreds of simulations including the latest ones obtained by the ABC consortium.[159] Recent efforts to create repositories and associated databases developed by the BioExcel CoE to help Covid19 research will be commented below.

Most of these databases are coupled to flexible graphical web interfaces allowing access to key analyses of the trajectories. However, to be really useful, these user interfaces should be flexible and interactive, something made possible with recent database architectures. For example in the BioExcel-CV19 web server, an associated REST API allows the user to extract slices of the trajectory as well as fragments of the molecule, which are then shown by the NGL visualizer. Using these tools, the user can focus their analysis on configurations fulfilling certain criteria, or perform meta-analyses combining parts of different trajectories. The next step towards gaining interactivity for analysis tools coupled to databases is probably the generalization of virtual research environments (VRE), infrastructures that allow the users to deploy and share their software to perform customized analyses on databases. These infrastructures make it possible to perform specific analyses on the computers supporting the databases, eventually only downloading analysis results instead of the original trajectories.

Entering the Exascale era, we must not forget the Exabyte problem that will be created by the future generation of Exaflop computers. Otherwise, massive amounts of resources will be wasted without providing useful information for the community.

4 | COVID-19: A USE CASE FOR HPC-MD SIMULATIONS

SARS-CoV-2, an enveloped and positive-strand RNA coronavirus, was first identified in the city of Wuhan in December 2019 and rapidly spread worldwide due to a high inter-human transmission rate and a relevant percentage of asymptomatic infections.[160–162] As of December 2021, almost 300 million cases and more than 5 million deaths have been registered worldwide, with dramatic implications for the global economy and human well-being. Since its emergence in late 2019, both the SARS-CoV-2 virus and the host response—starting at the main receptor, the angiotensin-converting enzyme 2 (ACE2)—have been extensively studied to understand the disease etiopathogenesis, the structure of the functional viral proteins, and the viral-host interaction pattern to guide a rapid development of both direct-acting antiviral and host-directed therapeutic agents. In this context, the theoreticians' community, assisted by the constantly increasing number of X-ray and high-resolution cryogenic electron microscopy (Cryo-EM) structures available, has promptly reacted and a massive worldwide computational endeavor has been made to unravel the molecular details of the virus' functioning at a multiscale level, from force-field based all-atom simulations to hybrid quantum-mechanics/molecular-mechanics approaches. Here, we will summarize selected studies that have provided key information on the functioning of the virus obtainable uniquely via scientific HPC simulations.

4.1 | RNA-dependent RNA polymerase

The RNA-dependent RNA polymerase (RdRp) is the core protein of the replication-transcription complex, which is the machinery in charge of the transcription and replication of SARS CoV-2's viral genome and subgenomic mRNAs.[163] The RdRp holoenzyme is composed of four subunits: one core nsp12 and cofactors, two nsp8s and one nsp7. In addition to RdRp, other important proteins are involved in the replication complex: a helicase (nsp13), an exonuclease that proofreads the newly synthetized RNA (nsp14), mRNA capping enzymes (nsp14, 16), or a specific endoribonuclease (nsp15).

Despite the great effort invested in vaccinating the global population, the emergence of immune escape variants partially decreased the protection conferred by the vaccine.[164] Moreover, the rapid spread of variants and its possible resistance to current vaccines still make antiviral drugs important means to alleviate the effects of the pandemic and react swiftly to its future outbursts. Thus, due to its central role in the viral life cycle, the RdRp has been an attractive target for antiviral drugs. Among them, Remdesivir has gained much attention currently being the first small-molecule drug approved for the treatment of Covid-19 by FDA in October 2020 (only followed by a Nirmatrelvir/Ritonavir combination in late 2021). Remdesivir is an adenine nucleotide analog that acts by inhibiting RdRp through a delayed chain termination mechanism.[165] Other inhibition strategies have also been deployed towards protein partners involved in the replication machinery or by hacking the virus' replication through lethal mutagenesis.[166]

Although RdRp has been primarily investigated using experimental techniques, theoretical computations have helped understand different aspects of its inner workings. In order to better understand SARS-CoV-2's RdRp, Shaw and co-workers made use of Cryo-EM microscopy, RNA-protein cross-linking and unbiased molecular dynamics simulations to unveil RdRp's backtracking mechanism.[167] In this study, three structures in which the backtracked RNA was accommodated by the nucleotide-triphosphate (NTP) entry channel were resolved, a mechanism also found in cellular RNA polymerases. Starting from the cryo-EM structure that featured two nsp8 cofactors, a nsp7 cofactor, two nsp13 helicase cofactors, an RNA template and a product RNA strand with one mismatched nucleotide at its 3′ end, three independent 5-μs unbiased molecular dynamics simulations were carried out. Although the backtracking process was not captured in the simulations, these trajectories revealed the spontaneous fraying of the 3′ terminal mismatched nucleotide into the NTP entry channel. This positioning would disfavor the RNA translocation and thus the elongation process.

Numerous experimental and computational studies have focused on understanding Remdesivir's inhibitory mechanism inside RdRp. However, Remdesivir's mode of action is still debated. A theoretical study investigated the possible mechanism of action at the molecular level with a total of 34 μs of simulation.[168] Absolute free energy calculations were

performed to obtain binding free energies of Remdesivir and two other nucleotide analogues inside RdRp, followed by independent short molecular dynamics simulations (100 ns) for each of the systems where Remdesivir had been incorporated in the nascent RNA strand. Their results would support inhibition coupled by a delayed chain termination, or alternatively by destabilizing the protein complex.

In another recent study,[169] equilibrium MD totaling 9 μs of simulation, as well as non-equilibrium free energy calculations, were employed to obtain the free energies of binding of ATP and Remdesivir-TP (RTP) to human and viral RNA polymerases. The molecular mechanism of action of Remdesivir was also addressed. The study showed that while SARS-CoV-2 RdRp displays a binding preference towards RTP versus ATP, human RNA Pol II displays the opposite trend. Moreover, ∼100 ns of hybrid QM/MM MD simulations making use of the string method were performed to decipher the molecular mechanism and reaction free energy profiles of nucleotide activation and incorporation inside the RdRp. These results indicate that RdRp makes use of a self-activated mechanism to achieve nucleotide incorporation, as observed in other polymerases. This highly processive polymerase was found to follow a two-metal ion mechanism of nucleotide incorporation. While Remdesivir is observed to be incorporated more slowly than its natural counterpart, adenine, this difference is too small to account for its biological activity. Finally, free energy calculations unveil a stabilizing effect that occurs during Remdesvir elongation along RdRp's exit channel, which would stall RNA polymerization and explain its antiviral inhibitory effect.

4.2 | Main protease

Proteolytic cleavage of the SARS-CoV-2's virally expressed polyproteins pp1a and pp1ab into its functional proteins is carried out by two viral proteases: the main protease Mpro (nsp5) that cuts 11 sites, and the papain-like protease PLpro (nsp3) which cuts three sites.[163] This a crucial step that enables the formation of the replication machinery in charge of transcription and replication of the viral genome inside the host cell. In addition, viral proteases can act on cellular proteins to help escape the virus from immune response.[170] Thus, many efforts have been directed towards the inhibition of viral proteases as an effective antiviral strategy. Moreover, due to its dissimilarity to human proteases, Mpro and PLpro are promising drug targets. Mpro and PLpro are cysteine proteases composed of three and four domains, respectively. While Mpro is active as a dimer, PLpro acts as a monomer, and as in other cysteine proteases, their active sites are formed by a Cys-His catalytic dyad.

Joint crystallographic and molecular dynamics studies have analyzed the effects and mode of binding/action of substrates or promising inhibitors inside the Mpro/PLpro enzymes.[171] The work by Hummer and co-workers studied SARS-Cov-2 PLpro by means of biochemical, structural and functional analysis. Through multi-μs MD simulations they analyzed the interaction between PLpro and two different protein substrates. In addition, 1-μs long MD simulations were obtained for both SARS-CoV and SARS-CoV-2 PLpros in complex with GRL-0617, a non-covalent inhibitor of SARS-CoV PLpro that was shown to also act against SARS-CoV-2 PLpro. Kovalevsky et al. crystalized Mpro at room temperature and performed 1 μs molecular dynamics simulations to analyze the flexibility of some of the protein's regions.[172] Experimental and theoretical approaches were also combined to study Mpro inhibition by myricetin and its derivatives.[173] Two independent ∼1.5 μs long Gaussian accelerated-MD simulations explained myricetin's dynamic interactions inside Mpro's active site. In addition to drug-design efforts, quantum mechanics (QM) calculations were performed on a small active-site QM-cluster model, including implicit solvation. Reaction pathways and transition states were characterized for myricetin and two other derivatives, explaining the formation of a covalent bond. The calculated activation free energies were obtained through thermal correction. Alternative approaches combining MD-metatrajectories, docking experiments and machine learning strategies have been used to explore binding of millions to billion of chemical compounds as binders of the protease. These massive efforts, which fit into the paradigm of distributed computing, allowed the community to develop chemical moieties that could then undergo lead optimization protocols.[32,174–176]

Additionally, fully theoretical studies have helped understand features not accessible to experiments. Atomistic simulations unveiled the binding preferences of the promising inhibitor ebselen to Mpro through 6 μs of cumulative MD calculations.[177] In another computational study, important insights into structure–function relationships of Mpro were obtained through continuous constant pH MD simulations[178] coupled to replica-exchange enhanced sampling. The study revealed the impact of the protonation of His172, which causes a conformational change of Mpro resulting in its deactivation. Based on their results, the authors provide a set of guidelines to design new inhibitors that could bind to Mpro's residues. Tuñon and co-workers made use of extensive QM/MM-MD simulations to characterize the free energy

profiles of the reaction mechanism of Mpro with a peptidic substrate.[179] Exploration of the reaction free energy landscape was performed at the DFT(B3LYP)/MM level making use of the string and umbrella sampling methods, with a total of 2.4 ns of cumulative sampling. They propose a Mpro reaction mechanism consisting of three steps with four transition states, with deacylation as the rate-limiting step of the overall reaction. The study unveils the most important residues involved in catalysis which can help guide drug-design.

4.3 | Spike: The crucial protein for infection

One of the most widely investigated SARS-CoV-2 protein is the viral Spike protein, a transmembrane homotrimeric class-I fusion glycoprotein which exists in a metastable prefusion architecture and comprises two functional subunits for binding to the host cell receptor ACE2 (S1 subunit) and for fusion of the viral and host cell membranes (S2 subunit).[180] The Spike protein of SARS-CoV is cleaved by a host-cell protease, the transmembrane protease/serine subfamily member 2 (TMPRSS2); an alveolar cell serine protease preferentially expressed on epithelial cells of the respiratory tract.[181–183] TMPRSS2-mediated cleavage of SARS-CoV Spike protein is propaedeutic to host ACE2 binding, membrane fusion and cell-entry mechanism; a highly-concerted and regulated step-wise mechanism. Particular attention has been paid to the highly antigenic S1-based receptor-binding domains (RBD) that by undergoing substantial hinge-like conformational rearrangements, transiently expose (Up conformation) or hide (closed conformation) the interface area required for ACE2 receptor binding and subsequent shedding of S1, refolding of S2 for membrane fusion and virus internalization.[184–186] Many aspects of Spike have been subjected to massive simulations.

Among them, *glycosylation* of both viral and host receptor has been quite a debated topic among theoretical chemists and biochemists. Indeed, it is known that extracellular portion of ACE2 contains at least seven N-glycosylated sites (i.e., N53, N90, N103, N322, N432, N546, N690) and several O-glycosylation sites (e.g., T730),[187,188] with the latter shown to be crucial in the process of protein–protein recognition and binding mode.[185,189,190] To shed light on the role of glycans in mediating virus internalization and spread, Hummer and coworkers have recently performed extensive multi-replica approach MD simulations of the fully-glycosylated human ACE2 receptor bound to the RBD of SARS-CoV-2 Spike and its unbound state.[191] Their results have shown that glycosylation at position N90 in human ACE2 structurally hampers virus binding. This explains, at the atomistic level, the experimental data on the augmented susceptibility to SARS-CoV-2 infection when N90 glycosylation is lost.[191] On the other hand, N322 glycan aids infectivity by hiding a key cryptic epitope known to be the target of CR3022 neutralizing antibody.[192] In their work, Hummer and collaborators have used a classical multi-replica approach of the 11 different glycosylated sites of ACE2. For each of the 11 setups, they carried out three independent MD simulation runs [(1 × 1 μs and 2 × 480 ns) × 11] for a total of ~22 μs trajectory for almost 1 million atoms. The so obtained trajectories were analyzed to dissect protein–protein interactions starting from the Cryo-EM native contacts, which were identified accordingly to the Best et al. approach.[193]

Glycosylation was also intensively analyzed by Amaro and co-workers that have built a full-length model of the glycosylated SARS-CoV-2 S protein, both in the open and closed states and have performed multiple microsecond-long, all-atom molecular dynamics simulations. Their efforts have produces key full-atom structural and dynamical insights on the roles of glycans and their implications on the (i) protein geometry, (ii) the global plasticity of Spike, and (iii) the shielding role against antibody epitope recognition.[194] In general, the numerous glycans found on the whole Spike protein play diverse roles and contribute differently to the global and local conformational rearrangements in a time-dependent manner. Indeed, they observed that the two N165 and N234 glycosylation sites play a major role in shifting the equilibrium ensemble from closed to open RBD conformation. The simulated models pinpoint N165 and N234 positions as a major modulator of the RBD conformational plasticity with N234 glycan that promotes RBD exposure and concurrent stabilization. To provide further support for their findings, they performed additional sets of extensive MD simulations on the N165A/N234A mutant of the SARS-CoV-2 Spike protein, which corroborated the importance of these glycosylated positions in the infective dynamics of the protein.This result was also confirmed by bilayer interferometry experiments. From the technical point of view, this work required an enormous computational effort as the full-length structures were embedded into an equilibrated all-atom membrane bilayer with a composition mimicking the one where the virus buds. Subsequently, explicit water molecules and ions were added, resulting in two final systems of ~1.7 million atoms each for which multiple replicas were simulated collecting more than 15 μs of the overall sampling time. Massively paralel computers were used for this study.

Spike conformational transitions and the functional characterization of SARS-CoV-2 rising variants have been an additional question the scientific community has targeted. As anticipated before, the three-protomeric Spike protein

can adopt different macrostates prevalently determined by the serial combination of each single RBD arrangement. Since RBD exposure (i.e., closed-to-open) is propaedeutic to ACE2 binding and virus internalization, and considering that the vast majority of evolution-selected mutations fall within the RBD domain (contributing to the creation of higher-infective viral strains), dissecting RBD's opening/closing mechanism and energetics at molecular level is of utmost importance. In this regard, S. Gnanakaran and co-workers[195] have characterized the flexibility and conformational transition of the Spike protein in a work where MD simulations using an HPC resource have provided a unique and detailed perspective that other standard biophysical approaches would have not produced. In particular, they have analyzed the effect of D614G mutation on the Spike protein discovering that the mutation heavily affects the interaction pattern of RBD, especially when this adopts the open-state conformation. Briefly, they found three contributions to structural rearrangements associated with the D614G mutation: (i) variation in the inter-protomer contacts, (ii) alteration in the correlations between the single RBDs, and (iii) variation in the specific inter-protomer hydrogen bond pattern. As a consequence, the infection-capable one-up state conformation of RBD is favored when D614G mutation occurs. Based on these results, the authors suggest that these three effects can lead to an increase in infectivity through an increase in the one-up population of the G-form. In this work, mainly unbiased MD simulations were performed, generating five replicas for each of the system analyzed in the study, generating overall ∼20 μs of trajectories for ∼1 million atoms simulated in explicit solvent.

In a similar vein, Ray, Le and Andricioaei studied the long-range allosteric communication within the Spike, using an enforced closed-to-open transition to study dynamical correlations between individual residues.[196] By calculating correlation scores, mutual information-based cross-correlation metrics and connectivity graph networks from a set of over 40 short trajectories, they traced allosteric pathways governing the global motions of the RBD to identify residues that, upon mutation, would most likely disrupt interdomain communications. Their analysis was able to recover the established effect of the D614G substitution on Spike opening, as well as provide a plausible explanation for the biological effects of other observed mutations such as A570D or P681H. However, the authors note that their model does not yet have predictive power, one that would allow to score new mutations by their potential threat to human health.

Nevertheless, the effect of natural mutations of SARS-CoV-2 on Spike structure, conformation, and antigenicity has been the leading-topic recently due to the rapidly growing number of virus variants with increased transmissibility, higher disease severity, resistance to neutralizing antibodies elicited by current vaccines or from previous infection, reduced efficacy of treatments, or failure of diagnostic methods. For this reason, these variants have been named Variants of Concern (VoCs). Gobeil et al.[197] investigated the Spike VoC involved in transmission between minks and humans, as well as the B1.1.7 (alpha variant), B.1.351 (beta variant), and P1 (gamma variant) Spike variants. All variants showed a remarkedly increased ACE2 binding affinity accompanied by a higher propensity of RBDs to stably adopt the functionally-active up states. While adaptation to mink resulted in Spike destabilization, the B.1.1.7 (UK variant) Spike balanced stabilizing and destabilizing mutations. A local destabilizing effect of the RBD E484K mutation was implicated in resistance of the B.1.1.28/P.1 (Brazil Variant) and B.1.351 (South Africa Variant) lineages to neutralizing antibodies. Also, authors have revealed allosteric effects of mutations and mechanistic differences that may govern zoonotic jumps or escaping mechanisms from antibody neutralization. Finally, extensive Adaptive Sampling Molecular Dynamic Simulations (ASMD) were performed by De Fabritis and co-workers[198,199] to explore more under-sampled regions of the conformational space of Spike, hence overcoming energetic barriers separating free energy wells. In ASMD, simulations are launched in sequential batches called epochs utilizing knowledge of the conformational space obtained from all previous epochs. To generate an ensemble of the RBD tip conformations required to start the adaptive sampling routine, authors have performed 100 distinct 50 ns-long simulations in the NVT ensemble with randomized initial velocities for each of the WT and mutant systems. Each iteration consisted of 50 to 100 short independent simulations, conformations from which were selected to further sample the most under-sampled conformational space regions. Time independent component analysis (TICA)[200] was performed to identify slowly changing molecular order parameters and subsequently Markov state models were built. A total of 29 adaptive iterations were performed, yielding total simulation times of 274.8 and 256.8 μs for the WT and Mut systems.

An out-scaling Folding@home distributed computing project[80] involved 1 million people to create a distributed exascale computer with the aim of simulating ∼0.1 s of the entire viral proteome. The brilliant idea of Gregory R. Bowman and collaborators[79] has permitted to observe a dramatic opening of the apo Spike complex, far beyond that seen experimentally, explaining and predicting the existence of "cryptic" epitopes. The project has also highlighted large conformational changes across the proteome, which reveal over 50 "cryptic" pockets that expand targeting options for the design of antivirals, thus paving the way towards dozens of alternative drug discovery projects. Important observations were also made in other aspects of Spike dynamics. The authors, thanks to extensive sampling, successfully

captured this rare event for both glycosylated and unglycosylated protein and found that glycosylation only slightly increases the population of the open state, the effect being smaller than that produced by genetic variation in the protein. The authors also found that opening occurs only for a single RBD at a time and that the scale of spike opening is often substantially larger than has been observed in experimental snapshots in the absence of binding partners. An enormous amount of calculating nodes worldwide were created and authors measured an impressive Folding@home peak performance of 1.01 exaflops. This was achieved at a point when ~280,000 GPUs and 4.8 million CPU cores were performing simulations simultaneously; a performance 5-fold greater than the peak performance of the world's fastest traditional supercomputer at that time, Summit.[6]

All the simulations were performed with the Gromacs package[21] and the immense computational power of peers distributed around the globe has permitted to investigate intensively and with the richest sampling ever obtained 36 distinct proteins associated to SARS-CoV-2 virus. The endeavors have produced (so far) a total outstanding simulated time-scale of ~115 ms (>0,1 s).

4.4 | BioExcel workflows

The combination of BioExcel Building Blocks and the PyCOMPSs workflow manager illustrates how workflows can help reaching the High Throughput (HT) regime in HPC infrastructures, running hundreds of calculations using hundreds of cores, thus exploiting thousands of cores in one single job. The possibility of running these huge executions will help the efficient usage of large supercomputers such as the ones that are to become available in the coming years. Workflows built using these tools have been used in some of the COVID-19 related research presented in this review. One example is a pipeline for the calculation of binding free energy differences upon protein residue mutations (Figure 1). The relative changes in binding free energies (ΔΔGs) are computed using a so-called alchemical fast-growth thermodynamic integration (TI) method that does not require a quasi-equilibrium simulation during the alchemical transition.[201] The workflow starts from two independent equilibrium simulations: WT and mutant. These simulations need to sufficiently sample the end state ensembles, as the free energy accuracy will depend on the sampling

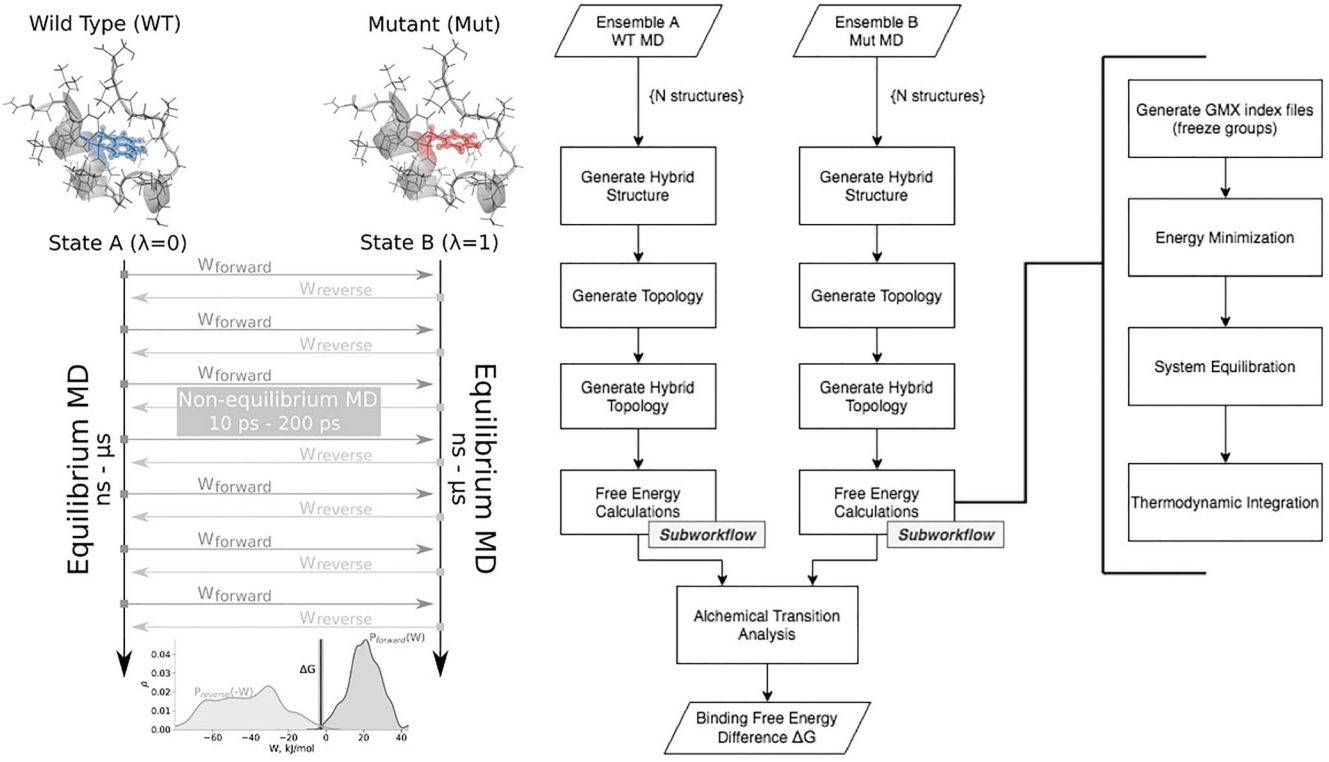


FIGURE 1 Left: Schematic representation of the nonequilibrium alchemical free energy calculations (taken from Aldeghi, 2019). Right: Alchemical free energy calculations workflow implemented with BioExcel building blocks.

convergence. From the generated trajectories snapshots are selected to start fast (picoseconds-long) transitions driving the system in the forward (WT to mutant) and reverse (mutant to WT) directions. The work values required to perform these transitions are collected and the Crooks Fluctuation Theorem[87] is used to calculate the free energy difference between the two states.

The whole pipeline was implemented using the BioExcel Building Blocks library, wrapping GROMACS and pmx biomolecular tools.[90,91] pmx is used to generate hybrid structures and topologies for the mutated residues and GROMACS to perform molecular dynamics simulations. The PyCOMPSs workflow manager is used to automatically distribute and parallelize the high number of transitions between the wild type and mutant states of the protein. As all these integration steps are completely independent calculations, they can in principle all be run at the same time. The flexibility added by PyCOMPSs allows a variable number of HPC cores depending on their availability (see use example in Figure 2). Here, a particular alchemical free energy calculation was performed on a 218,179-atom system, running a total number of short 1000 TI runs using 32 nodes (1536 cores) of the MareNostrum supercomputer in a single job. This particular execution took 4 h, using 100% of the CPUs available during most of this time. The number of nodes/cores to be used in the pipeline is completely flexible as PyCOMPSs is taking care of distributing the independent simulations across the number of available cores (see examples of pre-exascale workflows built within the BioExcel CoE including the one described in this section in the BioExcel GitHub repository: https://github.com/bioexcel/biobb_hpc_workflows).

These workflows were heavily used to understand the effect of mutations in both virus and host on spike recognition by ACE2, finding, for example, that most human ACE2 polymorphisms have a negligible effect on the binding affinity of the RBD, but also that the presence of some rare polymorphisms would protect a small fraction of the population from virus infection (see Figure 3).

Similarly, Gapsys & de Groot have explored an effect of ACE2 mutations on the ACE2-RBD binding creating a two-level screening strategy (Figure 4). Firstly, the residues at the ACE2-RBD interface were exhaustively scanned with Rosetta Flex ddG[202] protocol. This approach provides a good compromise for accuracy and computational efficiency[203] allowing to predict changes in binding affinity for a large number of mutations. In the second step, a selection of mutations based on the predictions from the previous step was subjected to the pmx/Gromacs based alchemical free energy calculation protocol. While these calculations are computationally demanding, they also offer high prediction accuracy.[204] In addition to estimating changes in the ACE2-RBD binding affinity, relative changes in ACE2 stability upon amino acid mutation were evaluated as well. All in all, this multi-step strategy allows identifying the residue positions in ACE2 and their mutations which would have the largest impact on the binding with the viral RBD. This strategy can be easily adjusted to probe for RBD mutations and their effect on the binding affinity, thus identifying viral variants of higher potency. Furthermore, by replacing ACE2, as the RBD binding partner, to an antibody or peptide, such mutation scan can be used to design novel therapeutics by enhancing their binding affinity to the viral RBD.

4.5 | BioExcel COVID-19 hub: Shaping the future of open data

The entire scientific community reacted to the COVID-19 pandemic by redirecting efforts to study the SARS-CoV-2 infection and its mechanisms of action. The biomolecular simulation field contributed since day one with an incredible number of MD simulations, which in turn produced an enormous amount of data distributed around the different groups working on them. BioExcel CoE, in collaboration with the Molecular Sciences Software Institute (MolSSI), developed the COVID-19 Molecular Structure and Therapeutics Hub (https://covid.bioexcel.eu/), a website presented as a community-driven data repository and curation service for molecular structures, models, therapeutics, and simulations related to COVID-19 computational research. The repository was designed from scratch to share data (including MD data) from the scientific community, making them completely open to better tackle the COVID-19 global emergency. The Hub has become a reference repository with huge amounts of useful information gathered in one single portal. Renowned groups in the field (e.g., D.E. Shaw, Riken, Folding@home) have contributed to the repository, summing up to milliseconds of trajectory data. Trajectories stored include different structures involved in the process of virus infection and virus life cycle, such as the ones previously presented in this review (Spike, Protease, RNA Polymerase).

Today the Hub is an essential repository for the field, a central point where to find and download useful data for research. The BioExcel-CV19 database and associated web server (Figure 5) expand the power of the Hub, including interactive graphical representations of the trajectories and analyses performed on them. The main objective of the BioExcel-CV19 project is to generate a tool for scientists interested in the COVID-19 research to interactively and graphically check key structural and dynamic features stemming from MDs. As these features vary depending on the

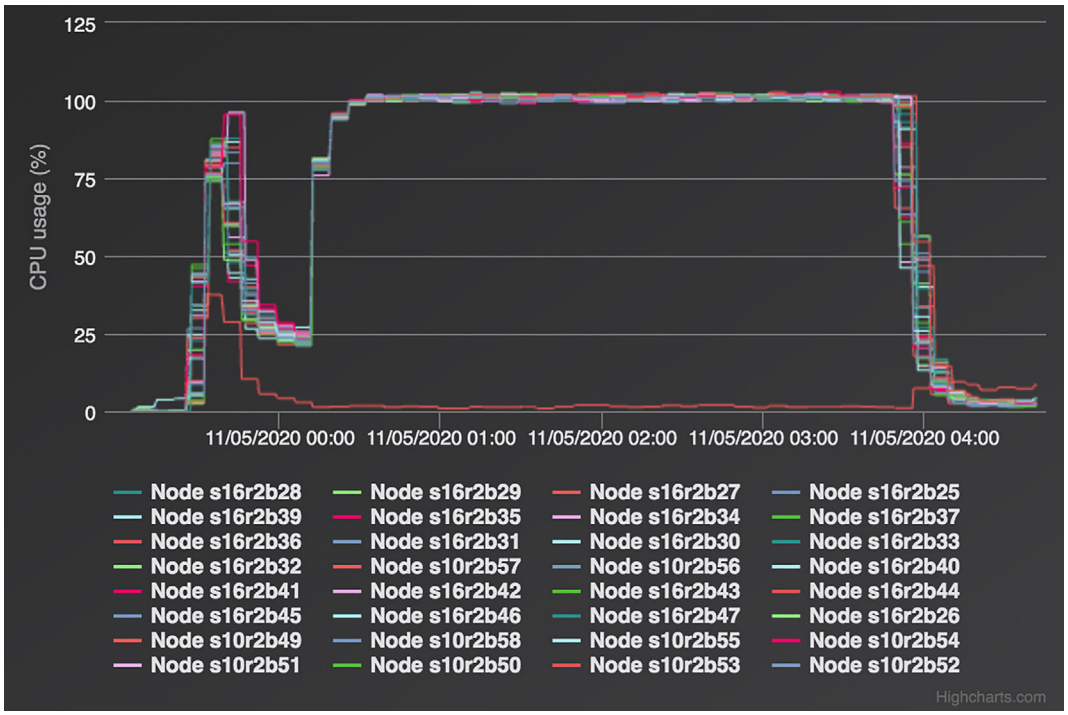
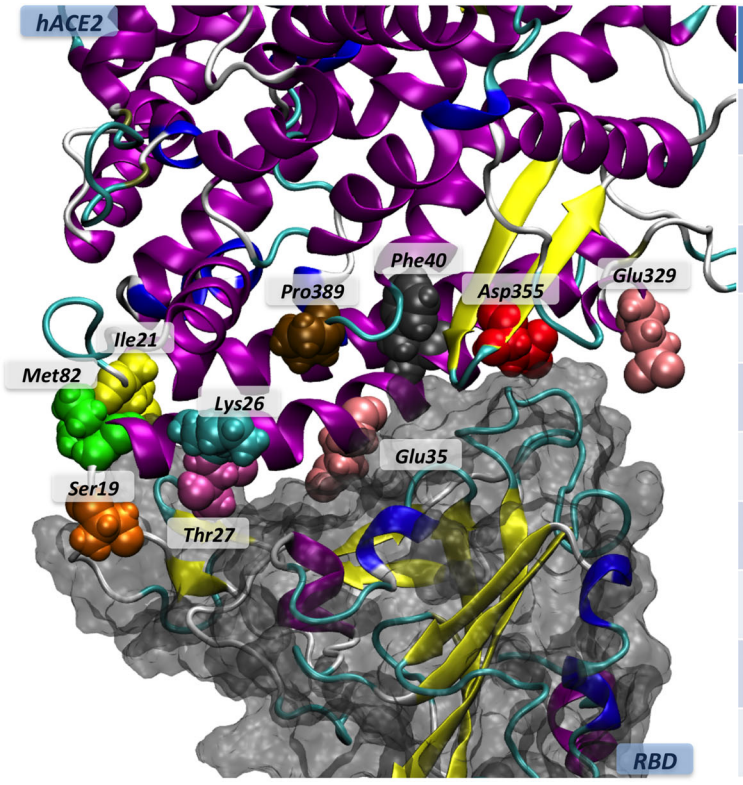


FIGURE 2 High parallelization reached thanks to the PyCOMPSs workflow manager. Example of a single job using 32 nodes (1536 cores) of the Marenostrum supercomputer at the BSC. Each line shows the CPU usage in % of a single MareNostrum node.



Mutation	Frequency	SARS-CoV-2 RBD hACE2 ΔΔG
Ser19Pro	0.000313	0.86
Ile21Val	0.000011	-0.65
Lys26Arg	0.003883	-1.12
Thr27Ala	0.000011	0.1
Glu35Lys	0.000016	4.25
Phe40Leu	0.000061	1.02
Met82Ile	0.000024	1.44
Glu329Gly	0.000034	0.36
Asp355Asn	0.000012	3.42
Pro389His	0.000038	1.36

FIGURE 3 Impact of human polymorphisms in RBD-hACE2 binding free energy, computed with BioExcel workflows including GROMACS and pmx. Human ACE2 protein mutations with higher frequency in the population are shown. Positive $\Delta\Delta G$ values indicate that a mutation reduces binding affinity. The $\Delta\Delta G$ values are in kcal/mol.

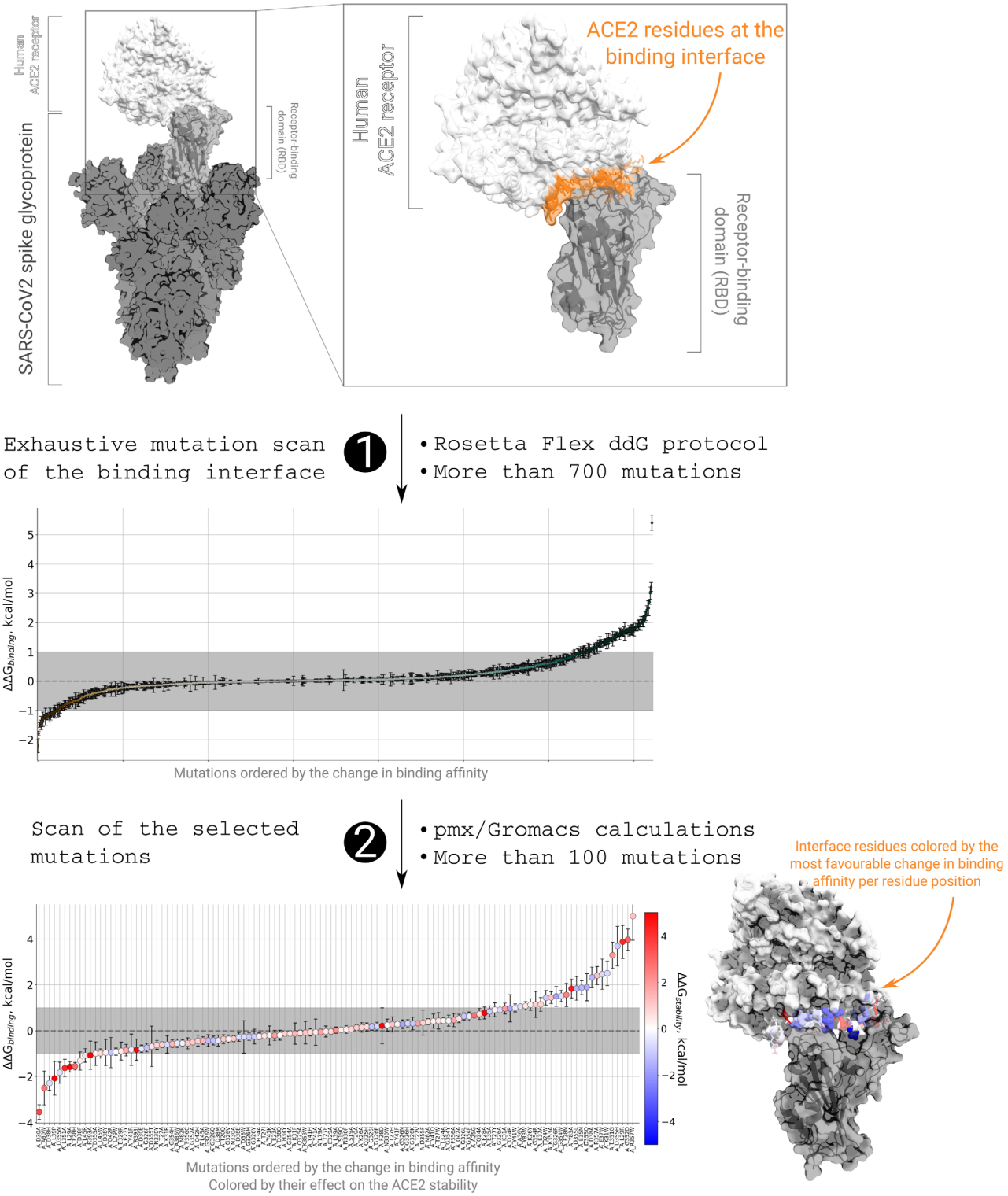


FIGURE 4 A multi-step strategy to predict the effect of ACE2 mutations on the ACE-RBD binding affinity. In the first step, the ACE2 residues at the protein binding interface are scanned by means of the computationally efficient Rosetta flex ddG protocol. In the second step, a selection of the mutations is probed by the computationally more demanding MD based free energy calculations using pmx and Gromacs software packages. The calculations allow evaluating the effect of mutations on the protein binding affinity, as well as on the stability of the ACE2 protein. This strategy can be further adapted to predict the effects of virus mutations, design high affinity antibodies or peptide therapeutics.

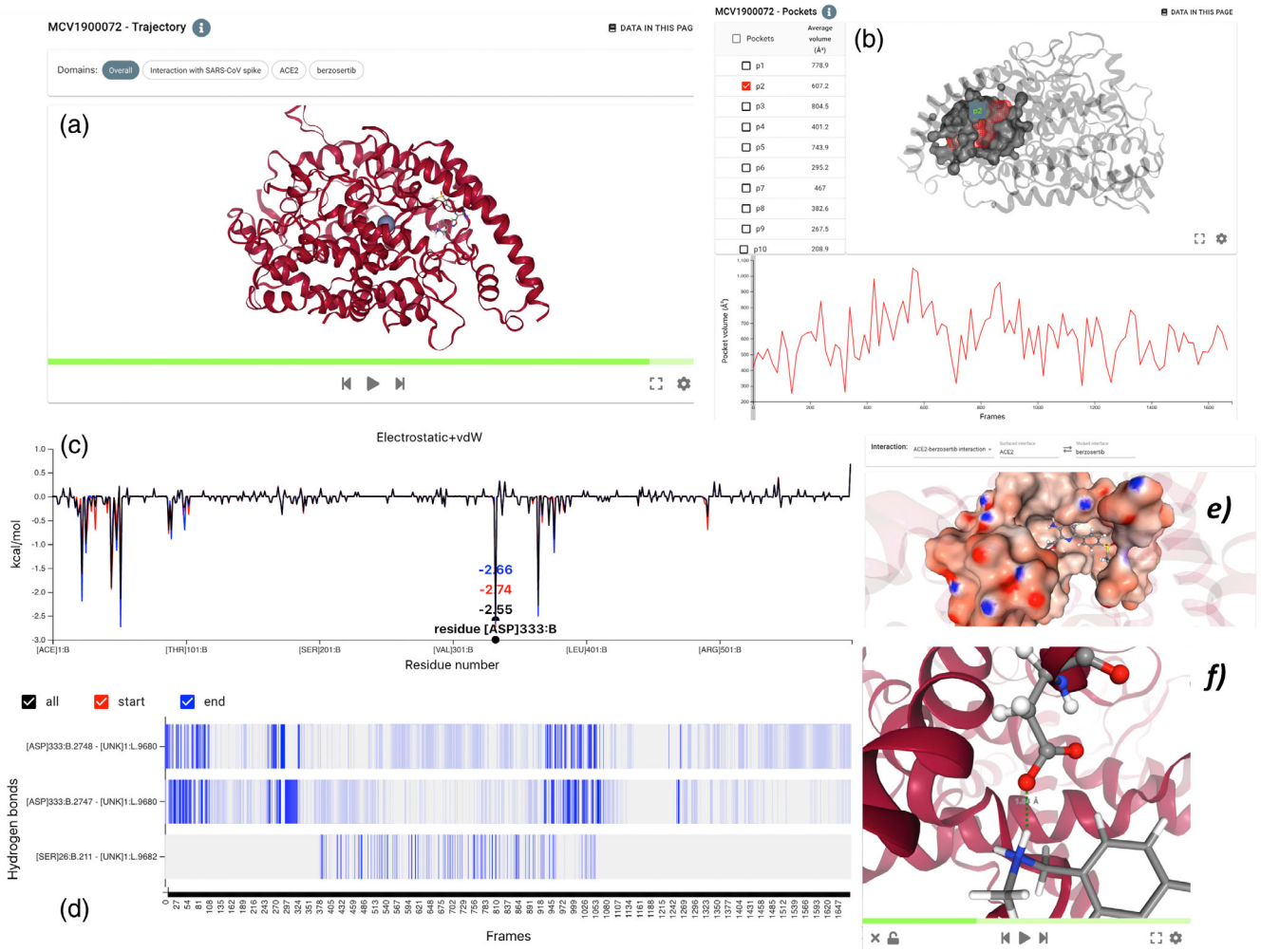


FIGURE 5 Screenshots from the BioExcel-CV19 web server. (a) Trajectory representation (*berzosertib*, an FDA approved drug molecule binding to the ectodomain of human ACE2); (b) analysis of the ligand pocket flexibility; (c) Analysis of electrostatic and Van der Waals interactions between the ligand and the ACE2 receptor; (d) analysis of hydrogen bonding between the ligand and the ACE2 receptor; (e) electrostatic potential surface in the ligand-binding pocket; (f) insight on the main interaction between the ligand and the ACE2 receptor involving an aspartic acid. URL: https://bioexcel-cv19.bsc.es/#/browse/MCV1900072

structure analyzed, specific analyses were performed, uploaded to the database, and represented in the web portal. These analyses and key features were collected by direct interaction with the authors of the simulations. As an example, trajectories corresponding to the viral RBD-hACE2 complex include interface observables (e.g., residue distances, hydrogen bonds), allowing an easy analysis of their behavior along the simulation (see Figure 5).

All the analyses integrated in the web portal are completely interactive. Whenever possible, a direct link from the analysis to the 3D representation is offered, using NGL viewer tool. The fast extraction of a particular snapshot from the whole trajectory is possible thanks to the NoSQL Mongo database backend powering the server. The whole set of trajectories (atomistic 3D coordinates for every atom and every frame of the simulation) and analyses are stored in this distributed database and efficiently retrieved on the fly from any web portal request. The entire pipeline is automated and new COVID-19 related trajectories are currently being processed. The set of analyses will be continuously extended, according to the suggestions of the authors of MD simulations.

5 | THE ROAD AHEAD

MD and, in general, biomolecular simulation tools are no longer marginal techniques used by a small set of groups to confirm already known facts. Rather, the methods are used by a huge community, helping to understand the

mechanisms of life, making predictions and guiding experiments. The Covid-19 pandemic has highlighted the incredible power of MD simulations, either alone or combined with experimental techniques, to reveal atomistic details of biologically relevant mechanisms. These research projects have also shown how large the computational requirements of state-of-the-art MD simulations are. It is not only a question of access to massive computer platforms, but also of strategies for how to use them in an efficient manner. While the next-generation supercomputers will be designed by hardware specialists, and future biological questions will come out of experimental labs, the bio-simulation community has to provide a robust interface between computer science and biology. Exascale supercomputing will be soon a reality, while Exascale distributed platforms are already here. If we learn from the past, computer scientists will soon dream of the YottaFlop computers.[205] The computational science community should escape the futile discussion between capacity and capability of computers and our community should strive to make the best of whatever new computer technology emerges, and to fully exploit the data derived from simulations.

AUTHOR CONTRIBUTIONS

Miłosz Wieczór: Writing – original draft (equal); writing – review and editing (equal). **Vito Genna:** Writing – original draft (equal); writing – review and editing (equal). **Juan Aranda:** Writing – original draft (equal); writing – review and editing (equal). **Rosa M. Badia:** Writing – review and editing (supporting). **Josep Lluís Gelpí:** Writing – review and editing (supporting). **Vytautas Gapsys:** Writing – original draft (supporting); writing – review and editing (supporting). **Bert de Groot:** Conceptualization (supporting); writing – original draft (supporting); writing – review and editing (supporting). **Erik Lindahl:** Writing – original draft (supporting); writing – review and editing (supporting). **Martí Municoy:** Formal analysis (supporting). **Adam Hospital:** Conceptualization (lead); writing – original draft (lead); writing – review and editing (lead). **Modesto Orozco:** Conceptualization (lead); writing – original draft (lead); writing – review and editing (lead).

ACKNOWLEDGMENTS

We are indebted to BioExcel partners for help and discussion and all the colleagues involved in Covid19 research for offering their research to the community. This study has been supported by the BioExcel-2. Center of Excellence for Computational Biomolecular Research" (823830), the Spanish Ministry of Science (RTI2018-096704-B-100, PID2020-116620GB-I00) and the Instituto de Salud Carlos III—Instituto Nacional de Bioinformática (ISCIII PT 17/0009/0007 co-funded by the Fondo Europeo de Desarrollo Regional). Funding was also provided by the MINECO Severo Ochoa Award of Excellence from the Government of Spain (awarded to IRB Barcelona). Modesto Orozco is an ICREA (Institució Catalana de Recerca i Estudis Avancats) academia researcher and Juan Aranda is a Juan de la Cierva Fellow. Funding was also provided by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 894489. The IRB is supported by MINECO Severo Ochoa Award of Excellence from the Government of Spain. [Correction added on 25 June 2022, after first online publication: The funding provided by European Union's Horizon 2020 research and innovation programme has been added in the Acknowledgments section in this version.]

DATA AVAILABILITY STATEMENT

Data collected from other people simulations is available as indicated in the manuscript

ORCID

Rosa M. Badia https://orcid.org/0000-0003-2941-5499
Vytautas Gapsys https://orcid.org/0000-0002-6761-7780
Martí Municoy https://orcid.org/0000-0003-4399-153X
Modesto Orozco https://orcid.org/0000-0002-8608-3278

RELATED WIREs ARTICLE

Surviving the Deluge of Biosimulation Data

REFERENCES

1. Lifson S, Warshel A. Consistent force field for calculations of conformations, vibrational spectra, and enthalpies of cycloalkane and n-alkane molecules. J Chem Phys. 1968;49(11):5104–7. https://doi.org/10.1063/1.1670007

2. Levitt M, Lifson S. Refinement of protein conformations using a macromolecular energy minimization procedure. J Mol Biol. 1969;46(2):269–79. https://doi.org/10.1016/0022-2836(69)90421-5
3. McCammon JA, Gelin BR, Karplus M. Dynamics of folded proteins. Nature. 1977;267(5612):585–90. https://doi.org/10.1038/267585a0
4. Shaw DE, Deneroff MM, Dror RO, Kuskin JS, Larson RH, Salmon JK, et al. Anton, a special-purpose machine for molecular dynamics simulation. Commun ACM. 2008;51(7):91–7. https://doi.org/10.1145/1364782.1364802
5. Shaw DE, Grossman JP, J.A. Bank, Batson B, Butts JA, Chao JC, et al. Anton 2: raising the Bar for performance and programmability in a special-purpose molecular dynamics supercomputer. International conference for high performance computing, networking, storage and analysis, SC. New York City, USA: IEEE; 2014. p. 41–53. https://doi.org/10.1109/SC.2014.9
6. TOP500 News. TOP500: The List. 2021. https://www.top500.org/
7. Bixon M, Lifson S. Potential functions and conformations in cycloalkanes. Tetrahedron. 1967;23(2):769–84. https://doi.org/10.1016/0040-4020(67)85023-3
8. Palazzesi F, Prakash MK, Bonomi M, Barducci A. Accuracy of current all-atom force-fields in modeling protein disordered states. J. Chem. Theory Comput. 2015;11(1):2–7. https://doi.org/10.1021/ct500718s
9. Dans PD, Ivani I, Hospital A, Portella G, González C, Orozco M. How accurate are accurate force-fields for B-DNA? Nucleic Acids Res. 2017;45(7):4217–30. https://doi.org/10.1093/nar/gkw1355
10. Rueda M, Ferrer-Costa C, Meyer T, Pérez A, Camps J, Hospital A, et al. A consensus view of protein dynamics. Proc Natl Acad Sci U S A. 2007;104(3):796–801. https://doi.org/10.1073/pnas.0605534104
11. Pérez A, Lankas F, Luque FJ, Orozco M. Towards a molecular dynamics consensus view of B-DNA flexibility. Nucleic Acids Res. 2008;36(7):2379–94. https://doi.org/10.1093/nar/gkn082
12. Jung J, Kobayashi C, Kasahara K, Tan C, Kuroda A, Minami K, et al. New parallel computing algorithm of molecular dynamics for extremely huge scale biological systems. J Comput Chem. 2021;42(4):231–41. https://doi.org/10.1002/jcc.26450
13. Palermo G, Bonvin AMJJ, Dal Peraro M, Amaro RE, Tozzini V. Editorial: multiscale modeling from macromolecules to cell: opportunities and challenges of biomolecular simulations. Front Mol Biosci. 2020;7:194. https://doi.org/10.3389/fmolb.2020.00194
14. Zhao G, Perilla JR, Yufenyuy EL, Meng X, Chen B, Ning J, et al. Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. Nature. 2013;497(7451):643–6. https://doi.org/10.1038/nature12162
15. Perilla JR, Goh BC, Cassidy CK, Liu B, Bernardi RC, Rudack T, et al. Molecular dynamics simulations of large macromolecular complexes. Curr Opin Struct Biol. 2015;31:64–74. https://doi.org/10.1016/j.sbi.2015.03.007
16. Durrant JD, Kochanek SE, Casalino L, Ieong PU, Dommer AC, Amaro RE. Mesoscale all-atom influenza virus simulations suggest new substrate binding mechanism. ACS Cent. Sci. 2020;6(2):189–96. https://doi.org/10.1021/acscentsci.9b01071
17. Yu A, Pak AJ, He P, Monje-Galvan V, Casalino L, Gaieb Z, et al. A multiscale coarse-grained model of the SARS-CoV-2 virion. Biophys J. 2021;120(6):1097–104. https://doi.org/10.1016/j.bpj.2020.10.048
18. Salomon-Ferrer R, Case DA, Walker RC. An overview of the Amber biomolecular simulation package. Wiley Interdiscip Rev Comput Mol Sci. 2013;3(2):198–210. https://doi.org/10.1002/wcms.1121
19. Phillips JC, Hardy DJ, Maia JDC, Stone JE, Ribeiro JV, Bernardi RC, et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. J Chem Phys. 2020;153(4):044130. https://doi.org/10.1063/5.0014475
20. Bowers KJ, Chow DE, Xu H, Dror RO, Eastwood MP, Gregersen BA, et al. Scalable algorithms for molecular dynamics simulations on commodity clusters. IEEE Xplore. 2007;43. https://doi.org/10.1109/sc.2006.54
21. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, et al. Gromacs: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX. 2015;1–2:19–25. https://doi.org/10.1016/j.softx.2015.06.001
22. Páll S, Hess B. A flexible algorithm for calculating pair interactions on SIMD architectures. Comput Phys Commun. 2013;184(12):2641–50. https://doi.org/10.1016/j.cpc.2013.06.003
23. Páll S, Zhmurov A, Bauer P, Abraham M, Lundborg M, Gray A, et al. Heterogeneous parallelization and acceleration of molecular dynamics simulations in GROMACS. J Chem Phys. 2020;153(13):134110. https://doi.org/10.1063/5.0018516
24. Götz AW, Williamson MJ, Xu D, Poole D, Le Grand S, Walker RC. Routine microsecond molecular dynamics simulations with AMBER on GPUs. 1. Generalized born. J. Chem. Theory Comput. 2012;8(5):1542–55. https://doi.org/10.1021/ct200909j
25. Stone JE, Hardy DJ, Ufimtsev IS, Schulten K. GPU-accelerated molecular modeling coming of age. J Mol Graph Model. 2010;29(2):116–25. https://doi.org/10.1016/j.jmgm.2010.06.010
26. Plimpton S. Fast parallel algorithms for short-range molecular dynamics. J Comput Phys. 1995;117(1):1–19. https://doi.org/10.1006/jcph.1995.1039
27. Bowers KJ, Dror RO, Shaw DE. Overview of neutral territory methods for the parallel evaluation of pairwise particle interactions. J Phys. 2005;16(1):300–4. https://doi.org/10.1088/1742-6596/16/1/041
28. Bowers KJ, Dror RO, Shaw DE. Zonal methods for the parallel execution of range-limited N-body simulations. J Comput Phys. 2007;221(1):303–29. https://doi.org/10.1016/j.jcp.2006.06.014
29. Kutzner C, Apostolov R, Hess B, Grubmüller H. Scaling of the GROMACS 4.6 molecular dynamics code on SuperMUC. Adv Parallel Comput. 2014;25:722–7. https://doi.org/10.3233/978-1-61499-381-0-722
30. Jaffrelot Inizan T, Célerse F, Adjoua O, el Ahdab D, Jolly LH, Liu C, et al. High-resolution mining of the SARS-CoV-2 main protease conformational space: supercomputer-driven unsupervised adaptive sampling. Chem Sci. 2021;12(13):4889–907. https://doi.org/10.1039/d1sc00145k

31. Gossen J, Albani S, Hanke A, Joseph BP, Bergh C, Kuzikov M, et al. A blueprint for high affinity SARS-CoV-2 Mpro inhibitors from activity-based compound library screening guided by analysis of protein dynamics. ACS Pharmacol Transl Sci. 2021;4(3):1079–95. https://doi.org/10.1021/acsptsci.0c00215
32. Grottesi A, Bešker N, Emerson A, Manelfi C, Beccari AR, Frigerio F, et al. Computational studies of SARS-CoV-2 3CLpro: insights from MD simulations. Int J Mol Sci. 2020;21(15):5346. https://doi.org/10.3390/IJMS21155346
33. Mehregan A, Pérez-Conesa S, Zhuang Y, Elbahnsi A, Pasini D, Lindahl E, et al. Probing effects of the SARS-CoV-2 E protein on membrane curvature and intracellular calcium. bioRxiv. 2021;2021.05.28.446179. https://doi.org/10.1101/2021.05.28.446179
34. Casalino L, Dommer AC, Gaieb Z, Barros EP, Sztain T, Ahn SH, et al. AI-driven multiscale simulations illuminate mechanisms of SARS-CoV-2 spike dynamics. Int. J. High Perform. Comput. Appl. 2021;35(5):432–51. https://doi.org/10.1177/10943420211006452
35. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh Ewald method. J Chem Phys. 1995;103 (19):8577–93. https://doi.org/10.1063/1.470117
36. Shamshirgar DS, Yokota R, Tornberg AK, Hess B. Regularizing the fast multipole method for use in molecular simulation. J Chem Phys. 2019;151(23):234113. https://doi.org/10.1063/1.5122859
37. Kohnke B, Ullmann TR, Beckmann A, Kabadshow I, Haensel D, Morgenstern L, et al. Gromex: a scalable and versatile fast multipole method for biomolecular simulation. Lecture notes in computational science and engineering. Volume 136. Cham: Springer; 2020. p. 517–43.
38. Hardy DJ, Wu Z, Phillips JC, Stone JE, Skeel RD, Schulten K. Multilevel summation method for electrostatic force evaluation. J. Chem. Theory Comput. 2015;11(2):766–79. https://doi.org/10.1021/ct5009075
39. Knapp B, Ospina L, Deane CM. Avoiding false positive conclusions in molecular simulation: the importance of replicas. J. Chem. Theory Comput. 2018;14(12):6127–38. https://doi.org/10.1021/acs.jctc.8b00391
40. Gapsys V, de Groot BL. On the importance of statistics in molecular simulations for thermodynamics, kinetics and simulation box size. Elife. 2020;9:1–21. https://doi.org/10.7554/ELIFE.57589
41. Jiang W, Phillips JC, Huang L, Fajer M, Meng Y, Gumbart JC, et al. Generalized scalable multiple copy algorithms for molecular dynamics simulations in NAMD. Comput Phys Commun. 2014;185(3):908–16. https://doi.org/10.1016/j.cpc.2013.12.014
42. Portella G, Orozco M. Multiple routes to characterize the folding of a small DNA hairpin. Angew Chem Int Ed. 2010;49(42):7673–6. https://doi.org/10.1002/anie.201003816
43. Chodera JD, Swope WC, Pitera JW, Dill KA. Long-time protein folding dynamics from short-time molecular dynamics simulations. Curr Opin Struct Biol. 2006;5(4):1214–26. https://doi.org/10.1137/06065146X
44. Pande VS, Beauchamp K, Bowman GR. Everything you wanted to know about Markov state models but were afraid to ask. Methods. 2010;52(1):99–105. https://doi.org/10.1016/j.ymeth.2010.06.002
45. Sirur A, de Sancho D, Best RB. Markov state models of protein misfolding. J Chem Phys. 2016;144(7):75101. https://doi.org/10.1063/1.4941579
46. Bowman GR, Pande VS, Noé F. An introduction to Markov state models and their application to long timescale molecular simulation. Vol 797. Dordrecht, NL: Springer; 2014. p. 148. https://books.google.com/books/about/An_Introduction_to_Markov_State_Models_a.html?id=ggnGBAAAQBAJ
47. Nüske F, Wu H, Prinz JH, Wehmeyer C, Clementi C, Noé F. Markov state models from short non-equilibrium simulations: analysis and correction of estimation bias. J Chem Phys. 2017;146(9):094104. https://doi.org/10.1063/1.4976518
48. Suárez E, Adelman JL, Zuckerman DM. Accurate estimation of protein folding and unfolding times: beyond Markov state models. J. Chem. Theory Comput. 2016;12(8):3473–81. https://doi.org/10.1021/acs.jctc.6b00339
49. Wua H, Paul F, Wehmeyer C, Noé F. Multiensemble Markov models of molecular thermodynamics and kinetics. Proc Natl Acad Sci U S A. 2016;113(23):E3221–30. https://doi.org/10.1073/pnas.1525092113
50. Bussi G. Hamiltonian replica exchange in GROMACS: a flexible implementation. Mol Phys. 2014;112(3–4):379–84. https://doi.org/10.1080/00268976.2013.824126
51. Sugita Y, Kitao A, Okamoto Y. Multidimensional replica-exchange method for free-energy calculations. J Chem Phys. 2000;113(15): 6042–51. https://doi.org/10.1063/1.1308516
52. Hukushima K, Nemoto K. Exchange Monte Carlo method and application to spin glass simulations. J Physical Soc Japan. 1996;65(6): 1604–8. https://doi.org/10.1143/JPSJ.65.1604
53. Sugita Y, Okamoto Y. Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett. 1999;314(1–2):141–51. https://doi.org/10.1016/S0009-2614(99)01123-9
54. Liu P, Kim B, Friesner RA, Berne BJ. Replica exchange with solute tempering: a method for sampling biological systems in explicit water. Proc Natl Acad Sci U S A. 2005;102(39):13749–54. https://doi.org/10.1073/pnas.0506346102
55. Wang L, Friesner RA, Berne BJ. Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (REST2). J Phys Chem B. 2011;115(30):9431–8. https://doi.org/10.1021/jp204407d
56. Zwanzig RW. High-temperature equation of state by a perturbation method. I. Nonpolar gases. J Chem Phys. 1954;22(8):1420–6. https://doi.org/10.1063/1.1740409
57. Jorgensen WL, Thomas LL. Perspective on free-energy perturbation calculations for chemical equilibria. J. Chem. Theory Comput. 2008;4(6):869–76. https://doi.org/10.1021/ct800011m
58. Bash PA, Singh UC, Langridge R, Kollman PA. Free energy calculations by computer simulation. Science (80-.). 1987;236(4801):564–8. https://doi.org/10.1126/science.3576184

59. Straatsma TP, McCammon JA. Multiconfiguration thermodynamic integration. J Chem Phys. 1991;95(2):1175–88. https://doi.org/10.1063/1.461148
60. Torrie GM, Valleau JP. Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling. J Comput Phys. 1977;23(2):187–99. https://doi.org/10.1016/0021-9991(77)90121-8
61. Laio A, Parrinello M. Escaping free-energy minima. Proc Natl Acad Sci U S A. 2002;99(20):12562–6. https://doi.org/10.1073/pnas.202427399
62. Kumar S, Rosenberg JM, Bouzida D, Swendsen RH, Kollman PA. THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. J Comput Chem. 1992;13(8):1011–21. https://doi.org/10.1002/jcc.540130812
63. Grubmüller H. Predicting slow structural transitions in macromolecular systems: conformational flooding. Phys Rev E. 1995;52(3):2893–906. https://doi.org/10.1103/PhysRevE.52.2893
64. Barducci A, Bussi G, Parrinello M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. Phys Rev Lett. 2008;100(2):020603. https://doi.org/10.1103/PhysRevLett.100.020603
65. Piana S, Laio A. A bias-exchange approach to protein folding. J Phys Chem B. 2007;111(17):4553–9. https://doi.org/10.1021/jp0678731
66. Maragliano L, Fischer A, Vanden-Eijnden E, Ciccotti G. String method in collective variables: minimum free energy paths and isocommittor surfaces. J Chem Phys. 2006;125(2):024106. https://doi.org/10.1063/1.2212942
67. Pan AC, Sezer D, Roux B. Finding transition pathways using the string method with swarms of trajectories. J Phys Chem B. 2008;112(11):3432–40. https://doi.org/10.1021/jp0777059
68. Maragliano L, Roux B, Vanden-Eijnden E. Comparison between mean forces and swarms-of-trajectories string methods. J. Chem. Theory Comput. 2014;10(2):524–33. https://doi.org/10.1021/ct400606c
69. Rueda M, Cubero E, Laughton CA, Orozco M. Exploring the counterion atmosphere around DNA: what can be learned from molecular dynamics simulations? Biophys J. 2004;87(2):800–11. https://doi.org/10.1529/biophysj.104.040451
70. Sfriso P, Emperador A, Orellana L, Hospital A, Gelpí JL, Orozco M. Finding conformational transition pathways from discrete molecular dynamics simulations. J. Chem. Theory Comput. 2012;8(11):4707–18. https://doi.org/10.1021/ct300494q
71. Zuckerman DM, Chong LT. Weighted ensemble simulation: review of methodology, applications, and software. Annu Rev Biophys. 2017;46(1):43–57. https://doi.org/10.1146/annurev-biophys-070816-033834
72. Zwier MC, Adelman JL, Kaus JW, Pratt AJ, Wong KF, Rego NB, et al. WESTPA: an interoperable, highly scalable software package for weighted ensemble simulation and analysis. J Chem Theory Comput. 2015;11(2):800–9. https://doi.org/10.1021/ct5010615
73. Zimmerman MI, Bowman GR. FAST conformational searches by balancing exploration/exploitation trade-offs. J. Chem. Theory Comput. 2015;11(12):5747–57. https://doi.org/10.1021/acs.jctc.5b00737
74. Zimmerman MI, Porter JR, Sun X, Silva RR, Bowman GR. Choice of adaptive sampling strategy impacts state discovery, transition probabilities, and the apparent mechanism of conformational changes. J. Chem. Theory Comput. 2018;14(11):5459–75. https://doi.org/10.1021/acs.jctc.8b00500
75. Brandt S, Sittel F, Ernst M, Stock G. Machine learning of biomolecular reaction coordinates. J Phys Chem Lett. 2018;9(9):2144–50. https://doi.org/10.1021/acs.jpclett.8b00759
76. Jung H, Covino R, Hummer G. Artificial intelligence assists discovery of reaction coordinates and mechanisms from molecular dynamics simulations. arXiv. 2019;1901.04595. https://arxiv.org/abs/1901.04595v1
77. Noé F, Olsson S, Köhler J, Wu H. Boltzmann generators: sampling equilibrium states of many-body systems with deep learning. Science (80-). 2019;365(6457):eaaw1147. https://doi.org/10.1126/science.aaw1147
78. Wang J, Olsson S, Wehmeyer C, Pérez A, Charron NE, Fabritiis G, et al. Machine learning of coarse-grained molecular dynamics force fields. ACS Cent Sci. 2019;5(5):755–67. https://doi.org/10.1021/acscentsci.8b00913
79. Zimmerman MI, Porter JR, Ward MD, Singh S, Vithani N, Meller A, et al. SARS-CoV-2 simulations go exascale to predict dramatic spike opening and cryptic pockets across the proteome. Nat Chem. 2021;13(7):651–9. https://doi.org/10.1038/s41557-021-00707-0
80. Beberg AL, Ensign DL, Jayachandran G, Khaliq S, Pande VS. Folding@home: lessons from eight years of volunteer distributed computing. New York City, USA: IEEE Xplore; 2009. https://doi.org/10.1109/IPDPS.2009.5160922
81. Sultan MM, Denny RA, Unwalla R, Lovering F, Pande VS. Millisecond dynamics of BTK reveal kinome-wide conformational plasticity within the apo kinase domain. Sci Rep. 2017;7(1):15604. https://doi.org/10.1038/s41598-017-10697-0
82. Sun X, Singh S, Blumer KJ, Bowman GR. Simulation of spontaneous G protein activation reveals a new intermediate driving GDP unbinding. Elife. 2018;7:e38465. https://doi.org/10.7554/eLife.38465
83. Meng Y, Shukla D, Pande VS, Roux B. Transition path theory analysis of c-Src kinase activation. Proc Natl Acad Sci U S A. 2016;113(33):9193–8. https://doi.org/10.1073/pnas.1602790113
84. Shukla D, Peck A, Pande VS. Conformational heterogeneity of the calmodulin binding interface. Nat Commun. 2016;7:10910. https://doi.org/10.1038/ncomms10910
85. Ryckbosch SM, Wender PA, Pande VS. Molecular dynamics simulations reveal ligand-controlled positioning of a peripheral protein complex in membranes. Nat Commun. 2017;8(1):6. https://doi.org/10.1038/s41467-016-0015-8
86. Jarzynski C. Nonequilibrium equality for free energy differences. Phys Rev Lett. 1997;78(14):2690–3. https://doi.org/10.1103/PhysRevLett.78.2690
87. Crooks GE. Entropy production fluctuation theorem and the nonequilibrium work relation for free energy differences. Phys Rev E - Stat. 1999;60(3):2721–6. https://doi.org/10.1103/PhysRevE.60.2721

88. Baumann HM, Gapsys V, De Groot BL, Mobley DL. Challenges encountered applying equilibrium and nonequilibrium binding free energy calculations. J Phys Chem B. 2021;125(17):4241–61. https://doi.org/10.1021/acs.jpcb.0c10263
89. BioExcel, 2021. Centre of Excellence for Computation Biomolecular Research. https://bioexcel.eu/
90. Gapsys V, Michielssens S, Seeliger D, de Groot BL. pmx: automated protein structure and topology generation for alchemical perturbations. J Comput Chem. 2015;36(5):348–54. https://doi.org/10.1002/jcc.23804
91. Gapsys V, De Groot BL. Pmx webserver: a user friendly Interface for Alchemistry. J Chem Inf Model. 2017;57(2):109–14. https://doi.org/10.1021/acs.jcim.6b00498
92. Gapsys V, Pérez-Benito L, Aldeghi M, Seeliger D, Vlijmen H, Tresadern G, et al. Large scale relative protein ligand binding affinities using non-equilibrium alchemy. Chem Sci. 2020;11(4):1140–52. https://doi.org/10.1039/c9sc03754c
93. Wolf S, Stock G. Targeted molecular dynamics calculations of free energy profiles using a nonequilibrium friction correction. J. Chem. Theory Comput. 2018;14(12):6175–82. https://doi.org/10.1021/acs.jctc.8b00835
94. Pohjolainen E, Malola S, Groenhof G, Häkkinen H. Exploring strategies for labeling viruses with gold nanoclusters through non-equilibrium molecular dynamics simulations. Bioconjug Chem. 2017;28(9):2327–39. https://doi.org/10.1021/acs.bioconjchem.7b00367
95. Jo S, Kim T, Iyer VG, Im W. CHARMM-GUI: a web-based graphical user interface for CHARMM. J Comput Chem. 2008;29(11):1859–65. https://doi.org/10.1002/JCC.20945
96. Hospital A, Andrio P, Fenollosa C, Cicin-Sain D, Orozco M, Gelpí JL. MDWeb and MDMoby: an integrated web-based platform for molecular dynamics simulations. Bioinformatics. 2012;28(9):1278–9. https://doi.org/10.1093/BIOINFORMATICS/BTS139
97. Hospital A, Faustino I, Collepardo-Guevara R, González C, Gelpí JL, Orozco M. NAFlex: a web server for the study of nucleic acid flexibility. Nucleic Acids Res. 2013;41(W1):W47–55. https://doi.org/10.1093/NAR/GKT378
98. Spjuth O, Bongcam-Rudloff E, Hernández GC, Forer L, Giovacchini M, Guimera RV, et al. Experiences with workflows for automating data-intensive bioinformatics. Biol Direct. 2015;10(1):1–12. https://doi.org/10.1186/s13062-015-0071-8
99. Fjukstad B, Bongo LA. A review of scalable bioinformatics pipelines. Data Sci Eng. 2017;2(3):245–51. https://doi.org/10.1007/s41019-017-0047-z
100. Leipzig J. A review of bioinformatic pipeline frameworks. Brief Bioinform. 2017;18(3):530–6. https://doi.org/10.1093/bib/bbw020
101. Atkinson M, Gesing S, Montagnat J, Taylor I. Scientific workflows: past, present and future. Fut Generat Comput Syst. 2017;75:216–27. https://doi.org/10.1016/j.future.2017.05.041
102. Reiter T, Brooks PT, Irber L, Joslin SEK, Reid CM, Scott C, et al. Streamlining data-intensive biology with workflow systems. GigaScience. 2021;10(1):1–19. https://doi.org/10.1093/gigascience/giaa140
103. Cheatham TE, Roe DR. The impact of heterogeneous computing on workflows for biomolecular simulation and analysis. Comput Sci Eng. 2015;17(2):30–9. https://doi.org/10.1109/MCSE.2015.7
104. Shade A, Teal TK. Computing workflows for biologists: a roadmap. PLoS Biol. 2015;13(11):e1002303. https://doi.org/10.1371/journal.pbio.1002303
105. Kutzner C, Páll S, Fechner M, Esztermann A, de Groot BL, Grubmüller H. More bang for your buck: improved use of GPU nodes for GROMACS 2018. J Comput Chem. 2019;40(27):2418–31. https://doi.org/10.1002/jcc.26011
106. M. R. Crusoe, M. Crusoe, S. Abeln, A. Iosup, P. Amstutz, J. Chilton, *et al.*, Methods included: standardizing computational reuse and portability with the common workflow language, arXiv. 2021. https://doi.org/10.48550/arXiv.2105.07028
107. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, et al. KNIME: the Konstanz information miner. Studies in classification, data analysis, and knowledge organization. Berlin, Heidelberg: Springer; 2008. p. 319–26. https://doi.org/10.1007/978-3-540-78246-9_38
108. Warr WA. Scientific workflow systems: pipeline pilot and KNIME. J Comput Aided Mol Des. 2012;26(7):801–4. https://doi.org/10.1007/s10822-012-9577-7
109. Oinn T, Addis M, Ferris J, Marvin D, Senger M, Greenwood M, et al. Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics. 2004;20(17):3045–54. https://doi.org/10.1093/bioinformatics/bth361
110. Afgan E, Baker D, Beek M, Blankenberg D, Bouvier D, Čech M, et al. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res. 2016;44(W1):W3–W10. https://doi.org/10.1093/nar/gkw343
111. Okonechnikov K, Golosova O, Fursov M, UGENE team. Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics. 2012;28(8):1166–7. https://doi.org/10.1093/bioinformatics/bts091
112. Ludäscher B, Altintas I, Berkley C, Higgins D, Jaeger E, Jones M, et al. Scientific workflow management and the Kepler system: research articles. Concurr. Comput. Pract. Exp. 2006;18(10):1039–65. https://doi.org/10.1002/cpe.v18:10
113. Bray SA, Lucas X, Kumar A, Grüning BA. The ChemicalToolbox: reproducible, user-friendly cheminformatics analysis on the galaxy platform. J Chem. 2020;12(1):1–7. https://doi.org/10.1186/s13321-020-00442-7
114. Beisken S, Meinl T, Wiswedel B, de Figueiredo LF, Berthold M, Steinbeck C. KNIME-CDK: workflow-driven cheminformatics. BMC Bioinformatics. 2013;14(1):1–4. https://doi.org/10.1186/1471-2105-14-257
115. V. A. Eyrich, 2021. The Schrödinger KNIME Extensions. https://www.schrodinger.com/products/knime-extensions.
116. Gally JM, Bourg S, Do QT, Aci-Sèche S, Bonnet P. VSPrep: a general KNIME workflow for the preparation of molecules for virtual screening. Mol Inform. 2017;36(10):1700023. https://doi.org/10.1002/minf.201700023
117. Purawat S, Ieong PU, Malmstrom RD, Chan GJ, Yeung AK, Walker RC, et al. A Kepler workflow tool for reproducible AMBER GPU molecular dynamics. Biophys J. 2017;112(12):2469–74. https://doi.org/10.1016/j.bpj.2017.04.055
118. Naden L, Ellis S, Jha S. MolSSI and BioExcel workflow workshop 2018 report. arXiv. 2019. https://arxiv.org/abs/1905.11863v1

119. Babuji Y, Woodard A, Li Z, Katz DS, Clifford B, Kumar R, et al. Parsl: pervasive parallel programming in Python. HPDC 2019: Proceedings of the 28th International Symposium on High-Performance Parallel and Distributed Computing. NYC: ACM (Association for Computing Machinery); 2019. p. 25–36. https://doi.org/10.1145/3307681.3325400
120. Ossyra J, Sedova A, Tharrington A, Noé F, Clementi C, Smith JC. Porting adaptive ensemble molecular dynamics workflows to the summit supercomputer. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). Vol 11887. Cham, CH: Springer; 2019. p. 397–417. https://doi.org/10.1007/978-3-030-34356-9_30
121. Tejedor E, Becerra Y, Alomar G, Queralt A, Badia RM, Torres J, et al. PyCOMPSs: parallel computational workflows in python. Int J High Perform Comput Appl. 2017;31(1):66–82. https://doi.org/10.1177/1094342015594678
122. Hedges L, Mey A, Laughton C, Gervasio F, Mulholland A, Woods C, et al. BioSimSpace: an interoperable python framework for biomolecular simulation. J Open Source Softw. 2019;4(43):1831. https://doi.org/10.21105/joss.01831
123. Gebbie-Rayet J, Shannon G, Loeffler HH, Laughton CA. Longbow: a lightweight remote job submission tool. J Open Res Softw. 2016;4(1):1. https://doi.org/10.5334/jors.95
124. Jain A, Ong SP, Chen W, Medasani B, Qu X, Kocher M, et al. FireWorks: a dynamic workflow system designed for high-throughput applications. Concurr Comput Pract Exp. 2015;27(17):5037–59. https://doi.org/10.1002/CPE.3505
125. Zhao Y, Li Y, Raicu I, Lu S, Lin C, Zhang Y, et al. A service framework for scientific workflow management in the cloud. IEEE Trans Serv Comput. 2015;8(6):930–44. https://doi.org/10.1109/TSC.2014.2341235
126. Balasubramanian V, Bethune I, Shkurti A, Breitmoser E, Hruska E, Clementi C, et al. ExTASY: scalable and flexible coupling of MD simulations and advanced sampling techniques. Proceedings of the 2016 IEEE 12th International Conference on e-Science, e-Science; 2017. NYC: IEEE; 2016. p. 361–70. https://doi.org/10.1109/eScience.2016.7870921
127. Hruska E, Balasubramanian V, Lee H, Jha S, Clementi C. Extensible and scalable adaptive sampling on supercomputers. J. Chem. Theory Comput. 2020;16(12):7915–25. https://doi.org/10.1021/acs.jctc.0c00991
128. Turilli M, Balasubramanian V, Merzky A, Paraskevakos I, Jha S. Middleware building blocks for workflow systems. Comput Sci Eng. 2019;21(4):62–75. https://doi.org/10.1109/MCSE.2019.2920048
129. Merzky A, Turilli M, Maldonado M, Santcroos M, Jha S. Using pilot systems to execute many task workloads on supercomputers. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol 11332. Cham, CH: Springer; 2019. p. 61–82. https://doi.org/10.1007/978-3-030-10632-4_4
130. Dakka J, Turilli M, Wright DW, Zasada SJ, Balasubramanian V, Wan S, et al. High-throughput binding affinity calculations at extreme scales. BMC Bioinform. 2018;19(18):33–45. https://doi.org/10.1186/s12859-018-2506-6
131. Lee H, Turilli M, Jha S, Bhowmik D, Ma H, Ramanathan A. DeepDriveMD: deep-learning driven adaptive molecular simulations for protein folding. Proceedings of DLS 2019: deep learning on supercomputers—held in conjunction with SC 2019: the international conference for high performance computing, networking, storage and analysis. NYC: IEEE; 2019. p. 12–9. https://doi.org/10.1109/DLS49591.2019.00007
132. Ma H, Bhowmik D, Lee H, Turilli M, Young M, Jha S, et al. Deep generative model driven protein folding simulations. Adv Parallel Comput. 2020;36:45–55. https://doi.org/10.3233/APC200023
133. Vermaas JV, Sedova A, Baker MB, Boehm S, Rogers DM, Larkin J, et al. Supercomputing pipelines search for therapeutics against covid-19. Comput. Sci. Eng. 2021;23(1):7–16. https://doi.org/10.1109/MCSE.2020.3036540
134. Bhati AP, Wan S, Alfè D, Clyde AR, Bode M, Tan L, et al. Pandemic drugs at pandemic speed: accelerating COVID-19 drug discovery with hybrid machine learning- and physics-based simulations on high performance computers. arXiv. 2021;20:1–19.
135. Wan S, Bhati A, Wade A, Alfè D, Coveney P. Thermodynamic and structural insights into the repurposing of drugs that bind to SARS-CoV-2 main protease. Biol Med Chem. 2021;1–10. https://doi.org/10.33774/CHEMRXIV-2021-03NRL-V2
136. Al Saadi A, Alfe D, Babuji Y, Bhati A, Blaiszik B, Brettin T, et al. IMPECCABLE: integrated modeling PipelinE for COVID cure by assessing better LEads. arXiv. 2020;20. https://doi.org/10.48550/arXiv.2010.06574
137. Andrio P, Hospital A, Conejero J, Jordá L, Pino M, Codo L, et al. BioExcel building blocks, a software library for interoperable biomolecular simulation workflows. Sci Data. 2019;6(1):1–8. https://doi.org/10.1038/s41597-019-0177-4
138. Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods. 2018;15(7):475–6. https://doi.org/10.1038/s41592-018-0046-7
139. Krylov A, Windus TL, Barnes T, Marin-Rimoldi E, Nash JA, Pritchard B, et al. Perspective: computational chemistry software and its advancement as illustrated through three grand challenge cases for molecular science. J Chem Phys. 2018;149(18):180901. https://doi.org/10.1063/1.5052551
140. Ferdous R, Roy B, Roy CK, Schneider KA. Workflow provenance for big data: from modelling to reporting. Cham: Springer; 2020. p. 185–200.
141. Khan FZ, Soiland-Reyes S, Sinnott RO, Lonie A, Goble C, Crusoe MR. Sharing interoperable workflow provenance: a review of best practices and their practical application in CWLProv. Gigascience. 2019;8(11):1–27. https://doi.org/10.1093/gigascience/giz095
142. Meng H, Thain D. Facilitating the reproducibility of scientific workflows with execution environment specifications. Procedia Comput Sci. 2017;108:705–14. https://doi.org/10.1016/j.procs.2017.05.116
143. Cohen-Boulakia S, Belhajjame K, Collin O, Chopard J, Froidevaux C, Gaignard A, et al. Scientific workflows for computational reproducibility in the life sciences: status, challenges and opportunities. Future Gener Comput Syst. 2017;75:284–98. https://doi.org/10.1016/j.future.2017.01.012

144. Goble C, Cohen-Boulakia S, Soiland-Reyes S, Garijo D, Gil Y, Crusoe MR, et al. FAIR computational workflows. Data Intell. 2020;2(1–2):108–21. https://doi.org/10.1162/dint_a_00033
145. Hospital A, Battistini F, Soliva R, Gelpí JL, Orozco M. Surviving the deluge of biosimulation data. WIREs Comput Mol Sci. 2020;10(3):e1449. https://doi.org/10.1002/wcms.1449
146. Stall S, Yarmey L, Cutcher-Gershenfeld J, Hanson B, Lehnert K, Nosek B, et al. Make scientific data FAIR. Nature. 2019;570(7759):27–9. https://doi.org/10.1038/d41586-019-01720-7
147. Elofsson A, Hess B, Lindahl E, Onufriev A, van der Spoel D, Wallqvist A. Ten simple rules on how to create open access and reproducible molecular simulations of biological systems. PLoS Comput Biol. 2019;15(1):e1006649. https://doi.org/10.1371/journal.pcbi.1006649
148. Ison J, Kalas M, Jonassen I, Bolser D, Uludag M, McWilliam H, et al. EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. Bioinformatics. 2013;29(10):1325–32. https://doi.org/10.1093/bioinformatics/btt113
149. Meyer T, D'Abramo M, Hospital A, Rueda M, Ferrer-Costa C, Pérez A, et al. MoDEL (molecular dynamics extended library): a database of atomistic molecular dynamics trajectories. Structure. 2010;18(11):1399–409. https://doi.org/10.1016/j.str.2010.07.013
150. van der Kamp MW, Schaeffer RD, Jonsson AL, Scouras AD, Simms AM, Toofanny RD, et al. Dynameomics: a comprehensive database of protein dynamics. Structure. 2010;18(4):423–35. https://doi.org/10.1016/j.str.2010.01.012
151. Hensen U, Meyer T, Haas J, Rex R, Vriend G, Grubmüller H. Exploring protein dynamics space: the dynasome as the missing link between protein structure and function. PLoS One. 2012;7(5):e33931. https://doi.org/10.1371/journal.pone.0033931
152. Sun R, Li Z, Bishop TC. TMB-iBIOMES: an iBIOMES-lite database of nucleosome trajectories and meta-analysis. ChemRxiv. 2019. https://doi.org/10.26434/chemrxiv.7793939
153. Mixcoha E, Rosende R, Garcia-Fandino R, Piñeiro Á. Cyclo-lib: a database of computational molecular dynamics simulations of cyclodextrins. Bioinformatics. 2016;32(21):3371–3. https://doi.org/10.1093/bioinformatics/btw289
154. Hospital A, Andrio P, Cugnasco C, Codo L, Becerra Y, Dans PD, et al. BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data. Nucleic Acids Res. 2016;44(D1):D272–8. https://doi.org/10.1093/nar/gkv1301
155. Zivanovic S, Bayarri G, Colizzi F, Moreno D, Gelpí JL, Soliva R, et al. Bioactive conformational ensemble server and database. A public framework to speed up in Silico drug discovery. J. Chem. Theory Comput. 2020;16(10):6586–97. https://doi.org/10.1021/acs.jctc.0c00305
156. Newport TD, Sansom MSP, Stansfeld PJ. The MemProtMD database: a resource for membrane-embedded protein structures and their lipid interactions. Nucleic Acids Res. 2019;47(D1):D390–7. https://doi.org/10.1093/NAR/GKY1047
157. Pasi M, Maddocks JH, Beveridge D, Bishop TC, Case DA, Cheatham T III, et al. μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. Nucleic Acids Res. 2014;42(19):12272–83. https://doi.org/10.1093/nar/gku855
158. Lundborg M, Apostolov R, Spångberg D, Gärdenäs A, Van Der Spoel D, Lindahl E. An efficient and extensible format, library, and API for binary trajectory data from molecular simulations. J Comput Chem. 2014;35(3):260–9. https://doi.org/10.1002/JCC.23495
159. Dans PD, Balaceanu A, Pasi M, Patelli AS, Petkevičiūtė D, Walther J, et al. The static and dynamic structural heterogeneities of B-DNA: extending Calladine-Dickerson rules. Nucleic Acids Res. 2019;47(21):11090–102. https://doi.org/10.1093/nar/gkz905
160. Tabata S, Imai K, Kawano S, Ikeda M, Kodama T, Miyoshi K, et al. Clinical characteristics of COVID-19 in 104 people with SARS-CoV-2 infection on the diamond princess cruise ship: a retrospective analysis. Lancet Infect Dis. 2020;20(9):1043–50. https://doi.org/10.1016/S1473-3099(20)30482-5
161. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579(7798):270–3. https://doi.org/10.1038/s41586-020-2012-7
162. Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. Addendum: a pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;588(7836):E6–6. https://doi.org/10.1038/s41586-020-2951-z
163. V'kovski P, Kratzel A, Steiner S, Stalder H, Thiel V. Coronavirus biology and replication: implications for SARS-CoV-2. Nat Rev Microbiol. 2021;19(3):155–70. https://doi.org/10.1038/s41579-020-00468-6
164. Hu J, Peng P, Cao X, Wu K, Chen J, Wang K, et al. Increased immune escape of the new SARS-CoV-2 variant of concern omicron. Cell Mol Immunol. 2022;19(2):293–5. https://doi.org/10.1038/s41423-021-00836-z
165. Gordon CJ, Tchesnokov EP, Woolner E, Perry JK, Feng JY, Porter DP, et al. Remdesivir is a direct-acting antiviral that inhibits RNA-dependent RNA polymerase from severe acute respiratory syndrome coronavirus 2 with high potency. J Biol Chem. 2020;295(20):6785–97. https://doi.org/10.1074/jbc.RA120.013679
166. Gordon CJ, Tchesnokov EP, Schinazi RF, Götte M. Molnupiravir promotes SARS-CoV-2 mutagenesis via the RNA template. J Biol Chem. 2021;297(1):100770. https://doi.org/10.1016/j.jbc.2021.100770
167. Malone B, Chen J, Wang Q, Llewellyn E, Choi YJ, Olinares PDB, et al. Structural basis for backtracking by the SARS-CoV-2 replication-transcription complex. Proc Natl Acad Sci U S A. 2021;118(19):e2102516118. https://doi.org/10.1073/pnas.2102516118
168. Byléhn F, Menéndez CA, Perez-Lemus GR, Alvarado W, De Pablo JJ. Modeling the binding mechanism of remdesivir, favilavir, and ribavirin to SARS-CoV-2 RNA-dependent RNA polymerase. ACS Cent. Sci. 2021;7(1):164–74. https://doi.org/10.1021/acscentsci.0c01242
169. Aranda J, Wieczor M, Brun-Heath I, Terrazas M, Orozco M. Mechanism of reaction of RNA-dependent RNA polymerase from SARS-CoV-2. Chem Catal. 2022. [In press] https://doi.org/10.1016/j.checat.2022.03.019
170. Shin D, Mukherjee R, Grewe D, Bojkova D, Baek K, Bhattacharya A, et al. Papain-like protease regulates SARS-CoV-2 viral spread and innate immunity. Nature. 2020;587(7835):657–62. https://doi.org/10.1038/s41586-020-2601-5

171. Sacco MD, Ma C, Lagarias P, Gao A, Townsend JA, Meng X, et al. Structure and inhibition of the SARS-CoV-2 main protease reveal strategy for developing dual inhibitors against Mpro and cathepsin L. Sci Adv. 2020;6(50):eabe0751. https://doi.org/10.1126/sciadv.abe0751
172. Kneller DW, Phillips G, O'Neill HM, Jedrzejczak R, Stols L, Langan P, et al. Structural plasticity of SARS-CoV-2 3CL Mpro active site cavity revealed by room temperature X-ray crystallography. Nat Commun. 2020;11(1):1–6. https://doi.org/10.1038/s41467-020-16954-7
173. Su H, Yao S, Zhao W, Zhang Y, Liu J, Shao Q, et al. Identification of pyrogallol as a warhead in design of covalent inhibitors for the SARS-CoV-2 3CL protease. Nat Commun. 2021;12(1):1–12. https://doi.org/10.1038/s41467-021-23751-3
174. Weng YL, Naik SR, Dingelstad N, Lugo MR, Kalyaanamoorthy S, Ganesan A. Molecular dynamics and in silico mutagenesis on the reversible inhibitor-bound SARS-CoV-2 main protease complexes reveal the role of lateral pocket in enhancing the ligand affinity. Sci Rep. 2021;11(1):1–22. https://doi.org/10.1038/s41598-021-86471-0
175. Ton AT, Gentile F, Hsing M, Ban F, Cherkasov A. Rapid identification of potential inhibitors of SARS-CoV-2 main protease by deep docking of 1.3 billion compounds. Mol Inform. 2020;39(8):2000028. https://doi.org/10.1002/MINF.202000028
176. Mekni N, Coronnello C, Langer T, de Rosa M, Perricone U. Support vector machine as a supervised learning for the prioritization of novel potential SARS-CoV-2 Main protease inhibitors. Int J Mol Sci. 2021;22(14):7714. https://doi.org/10.3390/IJMS22147714
177. Menéndez CA, Byléhn F, Perez-Lemus GR, Alvarado W, de Pablo JJ. Molecular characterization of ebselen binding activity to SARS-CoV-2 main protease. Sci Adv. 2020;6(37):345–56. https://doi.org/10.1126/sciadv.abd0345
178. Verma N, Henderson JA, Shen J. Proton-coupled conformational activation of SARS coronavirus main proteases and opportunity for designing small-molecule broad-spectrum targeted covalent inhibitors. J Am Chem Soc. 2020;142(52):21883–90. https://doi.org/10.1021/JACS.0C10770
179. Ramos-Guzmán CA, Ruiz-Pernía JJ, Tuñón I. Unraveling the SARS-CoV-2 main protease mechanism using multiscale methods. ACS Catal. 2020;10(21):12544–54. https://doi.org/10.1021/acscatal.0c03420
180. Gobeil SMC, Janowska K, McDowell S, Mansouri K, Parks R, Stalls V, et al. Effect of natural mutations of SARS-CoV-2 on spike structure, conformation, and antigenicity. Science (80-). 2021;373(6555):eabi6226. https://doi.org/10.1126/SCIENCE.ABI6226
181. Matsuyama S, Nagata N, Shirato K, Kawase M, Takeda M, Taguchi F. Efficient activation of the severe acute respiratory syndrome coronavirus spike protein by the transmembrane protease TMPRSS2. J Virol. 2010;84(24):12658–64. https://doi.org/10.1128/jvi.01542-10
182. Glowacka I, Bertram S, Müller MA, Allen P, Soilleux E, Pfefferle S, et al. Evidence that TMPRSS2 activates the severe acute respiratory syndrome coronavirus spike protein for membrane fusion and reduces viral control by the humoral immune response. J Virol. 2011;85(9):4122–34. https://doi.org/10.1128/jvi.02232-10
183. Sternberg A, Naujokat C. Structural features of coronavirus SARS-CoV-2 spike protein: targets for vaccination. Life Sci. 2020;257:118056. https://doi.org/10.1016/j.lfs.2020.118056
184. Hwang SS, Lim J, Yu Z, Kong P, Sefik E, Xu H, et al. MRNA destabilization by BTG1 and BTG2 maintains T cell quiescence. Science (80–.). 2020;367(6483):1255–60. https://doi.org/10.1126/science.abb2507
185. Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell. 2020;181(2):281–292.e6. https://doi.org/10.1016/j.cell.2020.02.058
186. Duan L, Zheng Q, Zhang H, Niu Y, Lou Y, Wang H. The SARS-CoV-2 spike glycoprotein biosynthesis, structure, function, and antigenicity: implications for the design of spike-based vaccine immunogens. Front Immunol. 2020;11:576622. https://doi.org/10.3389/fimmu.2020.576622
187. Shajahan A, Archer-Hartmann S, Supekar NT, Gleinich AS, Heiss C, Azadi P. Comprehensive characterization of N- and O- glycosylation of SARS-CoV-2 human receptor angiotensin converting enzyme 2. Glycobiology. 2021;31(4):410–24. https://doi.org/10.1093/GLYCOB/CWAA101
188. Yan R, Zhang Y, Li Y, Xia L, Guo Y, Zhou Q. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science (80–.). 2020;367(6485):1444–8. https://doi.org/10.1126/science.abb2762
189. Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S, et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature. 2020;581(7807):215–20. https://doi.org/10.1038/s41586-020-2180-5
190. Barros EP, Casalino L, Gaieb Z, Dommer AC, Wang Y, Fallon L, et al. The flexibility of ACE2 in the context of SARS-CoV-2 infection. Biophys J. 2021;120(6):1072–84. https://doi.org/10.1016/J.BPJ.2020.10.036
191. Mehdipour AR, Hummer G. Dual nature of human ACE2 glycosylation in binding to SARS-CoV-2 spike. Proc Natl Acad Sci U S A. 2021;118(19):e2100425118. https://doi.org/10.1073/pnas.2100425118
192. Yuan M, Wu NC, Zhu X, Lee CCD, So RTY, Lv H, et al. A highly conserved cryptic epitope in the receptor binding domains of SARS-CoV-2 and SARS-CoV. Science (80–.). 2020;368(6491):630–3. https://doi.org/10.1126/science.abb7269
193. Best RB, Hummer G, Eaton WA. Native contacts determine protein folding mechanisms in atomistic simulations. Proc Natl Acad Sci. 2013;110(44):17874–9. https://doi.org/10.1073/PNAS.1311599110
194. Casalino L, Gaieb Z, Goldsmith JA, Hjorth CK, Dommer AC, Harbison AM, et al. Beyond shielding: the roles of Glycans in the SARS-CoV-2 spike protein. ACS Cent. Sci. 2020;6(10):1722–34. https://doi.org/10.1021/ACSCENTSCI.0C01056
195. Mansbach RA, Chakraborty S, Nguyen K, Montefiori DC, Korber B, Gnanakaran S. The SARS-CoV-2 spike variant D614G favors an open conformational state. Sci Adv. 2021;7(16):3671–87. https://doi.org/10.1126/sciadv.abf3671
196. Ray D, Le L, Andricioaei I. Distant residues modulate conformational opening in SARS-CoV-2 spike protein. Proc Natl Acad Sci USA. 2021;118(43):e2100943118. https://doi.org/10.1073/pnas.2100943118

197. Gobeil SM-C, Janowska K, McDowell S, Mansouri K, Parks R, Manne K, et al. D614G mutation alters SARS-CoV-2 spike conformation and enhances protease cleavage at the S1/S2 junction. Cell Rep. 2021;34(2):108630. https://doi.org/10.1016/j.celrep.2020.108630
198. Doerr S, de Fabritiis G. On-the-fly learning and sampling of ligand binding by high-throughput molecular simulations. J. Chem. Theory Comput. 2014;10(5):2064–9. https://doi.org/10.1021/ct400919u
199. Doerr S, Harvey MJ, Noé F, de Fabritiis G. HTMD: high-throughput molecular dynamics for molecular discovery. J. Chem. Theory Comput. 2016;12(4):1845–52. https://doi.org/10.1021/acs.jctc.6b00049
200. Pérez-Hernández G, Paul F, Giorgino T, de Fabritiis G, Noé F. Identification of slow molecular order parameters for Markov model construction. J Chem Phys. 2013;139(1):015102. https://doi.org/10.1063/1.4811489
201. Aldeghi M, de Groot BL, Gapsys V. Accurate calculation of free energy changes upon amino acid mutation. Methods Mol Biol. 2019; 1851:19–47.
202. Barlow KA, Ó Conchúir S, Thompson S, Suresh P, Lucas JE, Heinonen M, et al. Flex ddG: Rosetta ensemble-based estimation of changes in protein-protein binding affinity upon mutation. J Phys Chem B. 2018;122(21):5389–99. https://doi.org/10.1021/acs.jpcb.7b11367
203. Aldeghi M, Gapsys V, de Groot BL. Accurate estimation of ligand binding affinity changes upon protein mutation. ACS Cent. Sci. 2018; 4(12):1708–18. https://doi.org/10.1021/acscentsci.8b00717
204. Gapsys V, Michielssens S, Seeliger D, de Groot BL. Accurate and rigorous prediction of the changes in protein free energies in a large-scale mutation scan. Angew. Chem Int. Ed. 2016;55(26):7364–8. https://doi.org/10.1002/anie.201510054
205. Goñi R, Orozco M. The Yottaflop frontier of atomistic molecular dynamics simulations. Theoretical and quantum chemistry at the Dawn of the 21st century. Oakville, CA: Apple Academic Press; 2018. p. 597–616.

How to cite this article: Wieczór M, Genna V, Aranda J, Badia RM, Gelpí JL, Gapsys V, et al. Pre-exascale HPC approaches for molecular dynamics simulations. Covid-19 research: A use case. WIREs Comput Mol Sci. 2022. e1622. https://doi.org/10.1002/wcms.1622