High–Resolution Mining of SARS–CoV–2 Main Protease Conformational Space: Supercomputer–Driven Unsupervised Adaptive Sampling†

Théo Jaffrelot Inizan,a, Frédéric Célerse,a,b, Olivier Adjoua,a Dina El Ahdab,a,c Luc–Henri Jolly,d Chengwen Liu,c Pengyu Ren,c Matthieu Montes,f Nathalie Lagarde,f Louis Lagardère*,a,d Pierre Monmarché*,a,e and Jean-Philip Piquemal*a,e,h

We provide a new unsupervised adaptive sampling strategy capable of producing microsecond-timescale molecular dynamics (MD) simulations using many-body polarizable force fields (PFF) on modern supercomputers. The global exploration problem is decomposed into a set of separate MD trajectories that can be restarted within an iterative/selective process to achieve sufficient phase-space sampling within large biosystems, while accurate statistical properties can be obtained through debiasing. With this pleasingly parallel setup, the Tinker-HP package can be powered by an arbitrary large number of GPUs (Graphics Processing Unit) cards available on pre-exascale supercomputers, reducing to days explorations that would have taken years. We applied the approach to the urgent problem of the modeling of the SARS–CoV–2 Main protease (Mpro) dimer. A 15.14 microsecond high-resolution all-atom simulation (AMOEBA PFF) of its apo state is provided and compared to other available long-timescale non-PFF data. Noticeable differences are found between clustering analysis of the simulations, the AMOEBA adaptive results exhibiting a richer conformational space. Overall, our high-resolution AMOEBA structural analysis captures key experimental observations concerning the stability of the oxyanion hole, a marker of activity through the stability of different stacking and salt bridge interactions. A dissymmetry is found between the enzyme protomers that exhibit different volumes. One of them appears fully inactive while the other is "activable", exhibiting some partial activity features. This activity evaluation can be further traced back to the large flexibility of the C terminal domain, fully captured by AMOEBA but not seen in X-rays due to insufficient electron densities related to the domain high mobility. The C–terminal region of the fully inactive protomer is shown to oscillate between several states, one of them interacting with the other protomer active site, therefore potentially modulating down its activity. Overall, these results reinforce the experimental hypothesis of a full inactivation of the apo state and clearly capture the asymmetric nature of protomers. Additional analysis show that the cavities volumes of the active and distal sites are found to be larger in the most active protomer with AMOEBA. To a larger extend, the PFF finds significantly larger cavities than those obtained with classical, non-polarizable simulations. The consequences on druggability are discussed as additional potential druggable cryptic pockets are found. All data produced within this research are fully accessible to the community for further analysis.

† Sorbonne Université, CNRS, Laboratoire de Chimie Théorique, Paris, France
b Sorbonne Université, CNRS, Institut Parisien de Chimie Moléculaire, Paris, France
c UR EGP, Centre d'Analyses et de Recherche, Faculté des Sciences, Université Saint-

Received Date
Accepted Date
DOI: 0.0000/xxxxxxxxxxx

Cite this DOI: 0.0000/xxxxxxxxxxx
1 Introduction

At the end of December 2019, a novel coronavirus (CoV) inducing severe acute respiratory disease (SARS) and labeled as SARS-CoV-2 developed and rapidly spread across the world. This virus, causing the disease named COVID-19, lead to a global pandemic in 2020 and finally to an urgent global issue. An important effort has been made to gain insights about the action of the virus on the human body. As the genome of the virus has been rapidly determined, a similarity between the SARS-CoV-2 virus and the older SARS-CoV (2003) and Middle East respiratory syndrome coronavirus (MERS-CoV in 2012) was observed. It thus helped researchers to find effective medicines by repurposing approved drugs. Among the different classes of proteins that have been experimentally characterized, the main protease (Mpro) is essential for processing the precursor polyprotein for the replication of the virus. Indeed, proteases are responsible for activating viral proteins for particle assembly. Due to their importance within the replication cycle of the virus, they have been proved to be successful targets for antiviral agents and are used to treat many diseases including HIV and Hepatitis. In the case of SARS-CoV-2, the main protease is called Mpro or 3CLpro. Many efforts are made to refine the crystallographic structure of Mpro as the number of experimental structures available on the Protein Data Bank is increasing. However, while more than one hundred Mpro structures exists and massive efforts to discover successful inhibitor are underway, computational approaches involving virtual screening as well as Molecular Dynamics are involved to help experimenters to in silico optimize their millions of test molecules.

Molecular Dynamics (MD) is a powerful tool which aims at improving the understanding of structural and dynamical features of complex biological systems. It also improves the ability to properly identify promising protein inhibitors. Two main research groups, DE Shaw Research (DESRES) and RIKEN Center for Biosystems Dynamics Research, recently released multi microseconds MD simulations of the Mpro dimer. These MD conformational ensembles both used non-polarizable force fields (n–PFF) such as DES-AMBER and AMBER14. Although they are of great help for the scientific community, conventional MD (cMD) simulation results are however limited by the daunting complexity of Mpro’s conformational space, which requires very large computational resources. In practice, both DESRES and Riken results were obtained on special purpose petascale supercomputers designed for MD (Anton and MD-GRAPE-4A for DESRES and Riken respectively). So, what can be done next? Beside these large scale MD simulations, the question of accuracy still remains open. Indeed, conformational space sampling depends by definition also on the force field used for the simulations. Our group has been involved for many years in the demonstration of the importance of considering explicit many-body effects in classical MD and free energy methods through the use of polarizable force fields (PFFs). Indeed, polarization affects solvation and modify the stability of secondary and quaternary structures of proteins, playing therefore a crucial role in the definition of the conformational space of a protein. Applying such methods to COVID-19 research could provide additional insights for drug modelers and experimental teams. When our project started (end of march 2020) responding to the international High-Performance Computing (HPC) global effort to mitigate the impact of COVID-19 pandemic, performing long timescale MD simulations using new generation PFFs on realistic SARS–CoV–2 proteins encompassing hundred of thousands atoms (or more) such as Mpro was out of reach of generalist supercomputers. Indeed, such simulation would have require years of computation.

In this paper, we propose to overcome these limitations by introducing a new unsupervised adaptive sampling method strategy based on statistical models and principal component analysis (PCA). It is then deployed on a generalist supercomputer. Since the global exploration problem is decomposed into a set of separate MD trajectories, the process can be restarted using an iterative selection method, and the various computations can take place on a large number of Graphic Processor Units (GPUs) that are now available in generalist supercomputers. Such a strategy enables the Tinker-HP package, which recently proposed a GPU-accelerated implementation, to perform multi-microseconds MD simulations within a few days, where years would have been required with single-card or CPUs conventional MD simulations. We additionally provide the capability to unbiased our simulations, which enables to fully exploit the total amount of MD trajectories to compute statistical properties that can therefore benefit from the long simulations. After describing our sampling strategy, we will detail our conformational space exploration results that notably expand over those obtained by other groups. We will unveil critical structural behavior, not captured with n–PFF, such as differences in clustering results, active site volumes and interactions between the C-terminal chain with the active site.
2 Unsupervised adaptive sampling strategy for exploration: exploiting pre-exascale machine and GPUs

Adaptive sampling has been used for many years and has proven to be a powerful exploration tool to study protein folding and dynamics, ligand binding and a variety of rare events phenomena.\textsuperscript{24,27} For this family of approaches, multiple iterations of independent molecular dynamics simulations are performed, basing the initial conditions at each iteration on the results of previous iteration steps. We propose here a new unsupervised (i.e. fully automated) adaptive sampling strategy dedicated to our specific use of PFF within large supercomputer systems allowing for the simultaneous use in production of hundreds (thousands if available) of GPUs cards. This characteristic is important as it allows to benefit from the full potential of pre-exascale supercomputers, and will naturally transfer on future exascale machines. The results that will be presented here benefit from a GPU acceleration thanks to the newly developed Tinker-HP GPUs code\textsuperscript{23} that was first used here for production to handle COVID-19 simulations. However the procedure is completely general and can be applied to any homogeneous or heterogeneous computational platforms compatible with Tinker-HP\textsuperscript{23,28} or any MD software. Therefore, in view of the particular distribution of available numerical resources, the simulations are organized by iterations as follows. At the beginning of each iteration, some initial structures are selected among the configurations sampled in the past iterations, from which independent MD simulations are run, generating new configurations. The selection of the initial structures at each iteration follows an adaptive procedure designed to enhance the exploration of a low-dimensional space of slow variables. More precisely, denote by $M_0$ the number of configurations available at the beginning of iteration $k \geq 0$, and $(q_{ij})_{1 \leq i \leq M_0}$ these configurations. Here, a configuration means the positions $q \in R^{3N}$ of all the atoms of the system. In particular, at the very beginning of the algorithm, we suppose that we start with $M_0 \geq 1$ configurations, obtained from an initial conventional MD simulation (which is in practice non-polarizable), or previous available works. At the beginning of iteration $k$, first, the protein is aligned in all configurations, using the backbone atoms of the 6LU7 crystal structure from the Protein Data Bank. A principal component analysis (PCA)\textsuperscript{29} is then performed, using the scikit-learn\textsuperscript{30} and mdtraj\textsuperscript{31} packages, on the protein atoms $(q_{ij})_{1 \leq i \leq M_0}$, from which the $n = 4$ principal modes are considered. This choice was made upon a global analysis of the first 20 PCA modes of the first AMOeba 0.14\textmu s that showed that $n > 4$ modes had variance contribution below 4\% (Figures 1 SI). This has also been corroborated by an analysis of Riken and DE Shaw trajectories, for which respectively 3 and 4 PCA modes are above 4\% (Figures 2 SI). We denote by $\xi_k : R^{3N} \rightarrow R^n$ the orthogonal projection on these $n$ principal modes and we write $x_i = \xi_k(q_{ij})$. At the beginning of iteration $k$, this represents the current guess of slow variables of the system, and in order to enhance the sampling, we would like to explore all the values of these slow variables. In other words, ideally, we would like the values of $x$ sampled to be uniformly distributed over some compact set of $R^n$. The selection procedure is designed to push the exploration in the direction of this ideal target. The density $\rho_k$ of the collective variables is approximated thanks to a Gaussian kernel, i.e. for $x \in R^n$

$$\rho_k(x) = \frac{1}{(2\pi \sigma^2)^{n/2}M_k} \sum_{i=1}^{M_k} \exp \left( -\frac{|x-x_i|^2}{2\sigma^2} \right),$$

for some $\sigma > 0$. In practice we used D.W. Scott method, implemented in Scipy\textsuperscript{32} to estimate a suitable bandwidth $\sigma$. Denote by $q_k$ the number of MD trajectories that are going to be run during iteration $k$. In order to select the initial structures $(q_{ij})_{1 \leq i \leq M_0}$ of these simulations, the indexes $I_1, \ldots, I_{q_k}$ are generated as independent random variables in $\{1, \ldots, M_0\}$ distributed according to

$$P(I = i) = \frac{\rho_k^{-1}(x_i)}{\sum_{j=1}^{M_0} \rho_k^{-1}(x_j)}.$$  

In other words, among all the structures currently available, $q_i$ is selected to be the initial structure of a new simulation with a probability inversely proportional to its density (in the low-dimensional space given by the first four PCA components). The effect of this selection can intuitively be illustrated as follows: if two domains of similar size (in the sense of the Lesbesgue measure on $R^n$) have been visited, with one that concentrates most of the past trajectories while the other contains only a few points, then approximately half of the new initial structures will be selected in each domain; on the contrary, a uniform selection among the past configurations would have put much more weight on the dense domain.

From the initial structures $(q_{ij})_{1 \leq i \leq M_0}$, $s_k$ independent MD simulations are sampled, and the state of each simulation is recorded every 0.1 ns (the initial structure is not recorded, since it has already been recorded in one of the past iterations). Here, independent means that the initial velocities (sampled according to the equilibrium Gaussian density) and the white noises of the Langevin thermostats are independent (and, of course, independent from previous iterations, so that a trajectory starting at some configuration $q_i$ will be different from the trajectory that initially produced this $q_i$). At the end of this $k^{th}$ iteration step $(q_{ij})_{M_0 < j \leq M_{k+1}}$, have been added, and iteration $k + 1$ starts. The procedure penalizes areas that have already been extensively visited, and is in a way reminiscent of the metadynamics\textsuperscript{33} method except that the statistical biasing is done through a selection step between each iteration rather than a biasing force updated along the trajectory. By comparison with metadynamics, this unsupervised selection step has the advantage to overcome...
the critical choice of an initial collective variable at the beginning of the simulation reinforcing automation of sampling scheme. This strategy belongs to the family of counts based adaptive sampling algorithms, where one only exploits the number of passages in the different states (micro or macro) visited in the previous iterations to choose which state to restart trajectories from. These are known to be efficient for pure exploration purpose (as it is the case here), even though more refined algorithms exist when some information is available as to where the sampling should be guided. However, contrary to what is usually done in the context of Markov State Models (MSMs), the states are not defined by applying a clustering algorithm to the already explored structures, but are the projection on the $n$ principal components generated by PCA (here, $n = 4$ as we discussed) of all the previous data. This has the advantage to provide an unsupervised sampling strategy that does not rely on a particular clustering algorithm (and therefore of its associated parameters) and to treat every point of this 4 dimension representation differently.

At the end of the simulation, $M_K$ configurations have been sampled with $K$ the total number of iterations. For a large $K$, the distribution of these configurations does not converge to the canonical distribution because of the statistical bias induced by the selection. To compute thermodynamical quantities, this bias should be taken into account. In that case, we interpret the previous selection as an importance sampling scheme. Thus, we have to compute a score $\omega_i > 0$ for each $i \in \{1, \ldots, M_K\}$ so that the canonical average of an observable $\varphi$ is estimated by

$$\langle \varphi \rangle \simeq \frac{\sum_{i=1}^{M_K} \omega_i \varphi(q_i)}{\sum_{i=1}^{M_K} \omega_i}.$$ 

The score $\omega_i$ is the ratio between the probabilities to obtain $q_i$ in the biased simulation and in an unbiased simulation (where, between each iteration, the next initial conditions are uniformly chosen among all currently available configurations, i.e. all with probability $1/M_K$). As a consequence, it is computed as follows: for all $i \leq M_0$, $\omega_i = 1$. Suppose by induction that $\omega_i$ has been computed for all $i \leq M_{k-1}$ for some $k$. Let $(i_1, \ldots, i_{\alpha})$ be the indexes that have been randomly selected for the initial conditions at the beginning of iteration $k$. For each $h \in \{i_1, \ldots, i_{\alpha}\}$, compute

$$\alpha_h = \frac{1}{M_k} P(I = h) - \frac{P_k(x_h)}{M_k} \sum_{j=1}^{M_k} \rho_k^{-1}(x_j).$$

Then, the score of all the configurations that are generated during iteration $k$ from the initial condition $q_h$ is $\alpha_h \omega_h$. That way, $\omega_i$ is computed for all $i \leq M_k$.

This latest point is important since it means the total simulation time can be used to compute average statistical properties that are unbiased and therefore exploitable. For example, it is possible to compare them to those obtained upon performing conventional MD runs.

Finally, it should be noticed that, instead of the PCA, this adaptive sampling strategy may be used with any other collective variables and/or dimensionality reduction algorithm. Overall the procedure is fully unsupervised, fast and can be used within Tinker-HP in a fully automated way.

3 Large scale unsupervised adaptive simulation using polarizable force fields (PFF) and GPUs

3.1 Simulation protocol

The presented all atom simulation was performed using the newly developed GPUs module within the Tinker-HP software, which is part of the Tinker 8 platform. This newly developed module is able efficiently exploit mixed precision offering a strong acceleration of simulations using GPUs. The 98 694 atoms initial structure of the fully solvated MPF pro dimer was extracted from the protein Data Bank (PDB: 6LU7) and the AMOEBA PFF was used to describe all atoms (protein and water). Periodic boundary conditions using a cubic box of side length of 100 Å were used. Langevin molecular dynamics simulations were performed using the BAOAB–RESPA1 integrator using a 10 fs outer time step, a preconditioned conjugate gradient polarization solver (with a $10^{-5}$ convergence threshold), hydrogen–mass repartitioning (HMR) and random initial velocities. Periodic boundary conditions were employed using the Smooth Particle Mesh Ewald (SPME) method with a grid of dimension $128 \times 128 \times 128$ Å. The Ewald–cutoff was taken to 7 Å and the van der Waals cutoff to 9 Å. We started the simulation by running 14 trajectories of 10 ns of representative structures extracted from the Riken trajectory using the PCA algorithm as a guiding thread (Figure 1 SI). A first adaptive sampling selection is then conducted on those initial structures. As explained in the previous method section, we chose the first four PCA components, which was found to be a relevant compromise between computational cost and exploration performance. At each iteration, the adaptive sampling procedure is used on the newly computed PCA components in order to select 100 structures. Then, 100 independent molecular simulations of 10ns were performed in the NVT ensemble at 300K on single NVIDIA V100 GPU cards. Each trajectory belonging to the same adaptive sampling iteration were run simultaneously on the HPE Jean Zay Supercomputer (IDRIS, GENCI, France). A single adaptive sampling iteration taking less than 18 hours to complete, it allows a production rate of 15.14μs in two weeks. Overall, the presented simulation ran over 10 working days in line with computer center resources availability. Additional conventional MD were performed on similar GPUs cards with the same simulation protocol on the Irene Joliot Curie Machine (TGCC, GENCI, France).

The complete 15.14μs trajectory with and without water are free of access though the Swiss National Supercomputing Center
(CSCS) and have been deposited on the BioExcel/Molssi COVID-19 community portal. A movie depicting the progress of the exploration can be found in SI.

3.2 Performance of the Adaptive Sampling Exploration: comparisons with other available simulations

As we mentioned in the method section, we use the PCA as an intermediate quantity to orient the consecutive sampling iteration. However, it is also a good quantity to quickly assess the performance of the adaptive sampling scheme for the exploration of the conformational space. Indeed, the analysis of MD trajectories with PCA is a well-known strategy known in the community as the "essential dynamics". PCA being a dimensionality reduction algorithm that evaluates directions maximizing the variance of the dataset, it is thus a reveal of a system conformational diversity. Therefore, it can be seen as a way to assess the amount of sampling and can also detect explicit "essential motions" otherwise not discernible using predefined collective variables. Thus, it is interesting to compare the amount of sampling on the space of these reduced variables. This is why we projected the Riken, the DESRES and the first 2µs Tinker-HP data set on the first two PCA components of the first 2µs of the Tinker-HP data set (Figure 1a, b). One can see that, in this space, the Tinker-HP adaptive scheme already captured the Riken and DESRES major main PCA features. It also appears that the Riken trajectory sampled a portion of conformational space close to the Tinker-HP data set while the DESRES trajectory seems to explore only the area the most sampled by Tinker-HP. The same procedure was applied for the PCA components and associated data of the entire Tinker-HP data set (Figure 1c, d) and it is striking that a much larger portion of conformational space has been sampled by our adaptive scheme. Additionally, we also projected the same data sets on the first two principal components of the Riken trajectory which gives the same justification of the larger sampling obtained by our method (Figure 7 SI).

As a preliminary conclusion, we can say that our adaptive sampling scheme allowed us to generate a multi-microsecond polarizable MD simulation that sampled a vast area of the free energy landscape. In addition, we analyzed the Root-Mean Square Deviation (RMSD) on protein backbones versus the radius of gyration (Figure 3 SI) for the AMOEBA 15.14 µs. It revealed large conformational changes. Variations for the radius of gyration are about 3 Å, while it varies of 1 Å for non-polarizable conventional MD. In addition, important changes are also observed in different important areas of the protease such as the dimerization site. The RMSD of the protein backbone versus the RMSD of the chain Å dimerization site (Figure 4 SI) depicts large fluctuations between 6–7 Å. DESRES and Riken trajectories exhibited only 2 Å, which is in the order of the size of the observed PCA features. Finally, to provide another view of these simulations differences, we performed dynamic cross-correlation map (DCCM) analysis for three datasets (RIKEN, DESRES and Tinker-HP). DCCMs were generated based on the Cα atom of each residue by using the functionality provided in MD-TASK package. As shown in Figure 5 SI, DCCM obtained from Tinker-HP trajectory shows more positive/negative values than those obtained from non-PFF trajectories (RIKEN and DESRES), indicating a stronger correlated/anti-correlated atom motion in Tinker-HP. It is worth mentioning that strong anti-correlation motions are observed between the α-helical region of each monomer of Mpro (residue number around 220-280 and 470-570) in Tinker-HP trajectory. By contrast, the corresponding regions have much weaker (anti-)correlation in both DESRES and RIKEN trajectories. The anti-correlation observed in PFF trajectory may need further confirmation from experiment, given that these α-helical regions are responsible for the dimerization process of Mpro. Overall, these first observations of the differences between the non-polarizable and the polarizable simulations motivate a further analysis of the different simulations.

3.3 Unsupervised Clustering and Extraction of unbiased relative free energy between representative domains

First, if the PCA analysis reveals useful information, a proper clustering of the produced ensembles is a more precise and quantitative framework to discus differences between simulations and possible new features captured by the AMOEBA force field. Therefore, we applied to all trajectories the density-based spatial clustering of applications with noise (DBSCAN) method. DBSCAN is an unsupervised machine learning algorithm that groups together data in clusters according to their density. It has the particularity to label points as noise if they are not in a dense region and are then not assigned to any cluster. DBSCAN is particularly well suited in our case as it is especially designed to target arbitrary shape clusters. To evaluate the density, DBSCAN uses two parameters, ε the distance for which two points are considered to be neighbors and MinPts the minimum number of points to define a cluster. ε has been chosen using the nearest neighbor graph procedure, i.e. by plotting the distance to the nearest n-neighbor for each point, ordered from the largest to the smallest value, and evaluating ε for which the graph starts an elbow. For a given ε we then scanned different values of MinPts until relatively large clusters covering a wide range of the space are found. In practice we evaluated the distance to the 4th nearest neighbor on the 4-Dimensions composed of the first four 15.14µs principal PCA components (Figure 6 SI). For DESRES and Riken, after aligning the structures to their respective PDB, we projected them on this 4D space.

Our choice of using the AMOEBA 15.14µs PCA components as the starting point of the clustering is driven by the conformational diversity brought by the coupling of the polarizable force field and the adaptive sampling scheme. For visualization, clusters are then projected on the first two principal components (Figure 2). To
Figure 1 Riken and DESRES datasets superposed to the 6LU7 protein backbone and projected on the first two PCA components fitted to, respectively, the 2\(\mu\)s (a), (b)) and 15.14\(\mu\)s (c), (d)) of the simulation.
Figure 2 DBSCAN clustering of a) DESRES (100 µs) b) Riken (10 µs) datasets and c) the Tinker–HP 15 µs simulation.
evaluate the quality of the clustering we used three scoring methods for unknown labeled data\textsuperscript{44}: Silhouette coefficient, Calinski-Harabasz and Davies-Bouldin indices. These indices confirmed our parameter optimization procedure and the high quality of the clustering. Our new adaptive sampling scheme has the main advantage that it offers access to true statistical properties such as free energies. To understand the clusters stability, the free energies for each cluster are computed (Figure 3\textsuperscript{c}, \textsuperscript{d}) through the evaluation of the probability distribution over the total number of structures. Notice that, as all the structures are not part of a cluster, the clusters probabilities do not sum to one. The unbiased probability distribution (Figure 3\textsuperscript{a}, \textsuperscript{b}) is estimated with the de-biasing procedure explained in the previous section. The de-biasing step preserves the trend between clusters but increases the probabilities. It means the five clusters were disadvantaged by the adaptive sampling. For example, the biased simulation assessed 8\% of probability of presence for the cluster 1, which should have contained, in an unbiased simulation, 20\% of the configurations. Besides, the cluster 1 is indeed the most explored region by both DESRES and Riken. Hence, the algorithm managed to disadvantage this part of the conformational space which is what we could have expected as it favored intermediate transition areas to the detriment of dense regions in order to discover new regions. The effect of the polarizability on structural properties such as volumes and RMSF is further depicted in the next Section. Overall, our approach demonstrated our capability to reach high-resolution conformational space exploration using a PFF. We identified 5 different clusters using AMOEBA (see Figure 2). While some of these clusters were already identified in previous MD using n–PFF (AMBER/DES-AMBER) (see Figure 1), we have detected new domains and their unbiased relative free energy shows that none of them could be neglected. It is now possible to use these clusters to unravel some statistical properties that could be useful to guide further ensemble docking simulations and/or to help to interpret experimental results.

4 Correlation with experimental data and new features

Figure 4 Representation of the $\pi-\pi$ stacking between His163 and Phe140 (green points) and several distances of interest which are responsible of the stability of the active site (black dashed lines).

To ensure the validity of our AMOEBA simulations, we compared computed properties with available experimental data. Since the beginning of the covid–19 pandemic various X–Ray structures have been released (PDB: 6Y84, 6LU7, 1UJ1, 6Y2G, ...). They provided important insights on specific interactions between residues as well as structural information about the active site. To be consistent with Riken simulations we used as reference the same PDB, 6LU7. Note that DESRES used another PDB, 6Y84, that we used as a reference in the computation of its properties. Recently, Zhou et al. published an experimental study of the apo structure (PDB 1UJ1\textsuperscript{46}). They found several features allowing for the characterization of the presence of the oxianion hole structure which is a key structural element of the activity of each protomer. In particular, they propose to monitor a salt bridge between Glu166 and His172 and a $\pi-\pi$ stacking between Phe140 and His163. These features were initially also observed for SARS–CoV–1\textsuperscript{47}. The oxianion hole is responsible of the stabilization of the substrate in the active site and is of a crucial importance for the enzyme's kinetic. Indeed, when the stacking and the salt bridge are broken, a rearrangement occurs leading to the collapse of the oxianion hole. In this case, Glu166 potentially interacts with His163 instead of His172 as suggested recently\textsuperscript{46}. In practice, the absence of a well-structured oxianion hole leads to the inhibition of the enzyme's activity. Experimentally, it is known that the M\textsuperscript{pro} monomeric form is inactive while the active form is
Figure 5: Representation of several distances of interest with Glu166 and the catalytic dial as well as the π–π stacking between Phe140 and His163. Each calculation has been performed on both chains of the system for each simulation. The π–π stacking has been evaluated following the approach of Branduardi et al. [45].
a homodimer containing two protomers. In the holo-state, the first protomer is active while the second one found inactive exactly like in the previous SARS-CoV-1 virus.

From then we investigated these two features: the salt bridge that Glu166 could forms with either His172 or His163, and the stability of the π–π stacking between Phe140 and His163. To study the stacking interaction, we use a stacking-index developed by Branduardi and Parinello who described it as a product of 2 fermi functions, one considering the radial dependence, the other the angular dependence of the interaction. The model provides an index ranging from 0 for a non-stacked interaction to 0.6 for a perfect one. The Glu166 interactions and π–π stacking were thus calculated for both chains of all Riken, DESRES and Tinker-HP structures and then classified into histograms. Finally, each histograms have been unbiased and extrapolated using univariate kernel density estimator. Final results are given on Figure 5.

For the salt bridge formed by Glu 166, we observed in the case of Tinker–HP an asymmetry between both protomers. In one protomer the Glu166 does not interact with His172 exhibiting a well-defined absence of activity of the protomer. The situation appears more complex in the other protomer where we observe an oscillation between two states presenting either a formed salt bridge or and absence of salt bridge leading to only some partial activity features. These two states demonstrate that the oxionion hole is either absent or partially disorganized in both chains in the apo structure, which is consistent with experimental data on the apo state where it was postulated that the structure was exhibiting a non activity of both its protomer. It also explains the asymmetry observed in the holo state where only one protomer is found to be active, a similar feature that was previously observed in SARS–CoV-1. In fact, our simulations show a fully inactive protomer (no salt bridge) coupled to an "activable" second protomer that exhibit some activation features (two states) when compared to its inactive counterpart. Similar interpretations could be deduced from the DESRES and Riken simulations despite a lower definition of the salt bridge interactions, a well-known difficult feature to capture with n–PFF and that is usually better described when including polarization. This disordering is also highlighted by another asymmetry between protomers while considering the interaction or absence of interaction of Glu166 with His163. We note the stronger asymmetry for DESRES while in the case of Riken and Tinker–HP we could again observe a mixture between interacting/non–interacting states. In that connection, a better conservation of the catalytic dial is observed in the Riken and Tinker–HP simulations with a smaller Cys145—His41 distance compared to DESRES.

Finally, a last marker is studied to confirm our observation: the π–π stacking between Phe140 and His163. Results are depicted on Figure 5. Tinker–HP does not capture this stacking in one protomer while again two mix-states (stacked and unstacked) are observed in the other protomer. The same observations can be done for DESRES and Riken although the states are less well defined in connection with the well-known difficult of capturing π–π stacking with n–PFF. Despite these differences observed in the 3 simulations (PFF handle more easily salt bridges and π–π stacking), all these features, combined together, converge to the fact that the active site is not well enough organized to exhibit active protomers and thus the apo state simulated by Riken, DESRES and Tinker–HP should not be active, corroborating experiment. However, simulations describe an asymmetric situation while one protomer is fully inactive, the other appears "activable" with some remaining activity features.

One way to measure some potential global differences between the different simulations is to measure the active site volume in each cluster and to depict the observed trend similarly to the π–π stacking previously. Beside the main active site cavity, the main protease exhibits 2 other cavities: the distal site and the dimerization site. Represented on Figure 5, these cavities are considered as potential targets for drug inhibitors. An accurate description of each of these cavities is essential to the estimation of efficient inhibitors. For each cluster of each dataset, we thus estimated those 3 cavity volumes. Volumes were calculated for each isolated clusters using the POVME 3.0 software. For each cavities, a 1.0 Å grid spacing has been chosen. The residues 7–198, 198–306 and all residues within 3.5 Å from the other monomer have been selected for the active, distal and dimerization sites with respectively a 12 Å, 10 Å and 10 Å. 1 000 structures were randomly chosen per cluster for the analysis. When a cluster had less than 1 000 structures, we chose all the structures. Detailed information are given in SI on the size of each cluster as well as their relative size (see Table 1 in SI). Similarly to the π–π stacking and the Glu166 distances, we used the univariate kernel density estimator on the volumes. The final volumes are depicted on Figure 6. Additionally, each cluster has a normal distribution supporting the quality of DBSCAN clusters. Represented by black arrows, different trends appear. For the 3 cavities, we observed a similarity between the single DESSCAN clusters, the clusters 1 and 2 from Riken and Tinker–HP’s clusters 1 and 2. Overall, while Tinker–HP clusters 1 and 2 are in good agreement with Riken and DESRES clusters, our clusters 3, 4 and 5 appear to be different. These observations highlight that differences indeed occur between clusters and between different datasets, going in the same direction of the previous analysis of the π–π stacking between residues Phe140 and His163 in chains A and B. For Tinker–HP, we observed a contraction for the three cavities in the cluster 3 while in cluster 4 and especially cluster 5, we observed a strong difference with a non–negligible increase of the cavity volumes. Cavities from clusters 4/5 depict stronger volumes fluctuations when using the polarizable AMOeba FF. While cavity volumes coming from AMBER/DES–AMBER simulations and from clusters 1 and 2 from AMOeba simulations are in agreement, the AMOeba results clearly captures an additional feature not captured by the DES-
Figure 6 Representation of the 3 cavities considered in this study: the dimerization site, active site and distal site. For each cavities trends coming from each clusters are depicted and superposed in three different graphs. Each curve has been unbiased according to the unbiasing approach described in this work. Cavity volumes are the sum of volumes found in both protomers.
AMBER and AMBER simulations. This information could be important for designing potential new inhibitor. Consequently, since strong differences between methods are observed in the volume evaluations of the different clusters, it is interesting to estimate the global protomer volumes if one wants to try to capture further the discussed asymmetry. Protomer volumes can be found on Figure 7. The protomer 1 (eg predicted to be non-active) depicts a strong gaussian behavior while protomer 2 (eg shown predicted to be oscillating between an active and a non-active state) is characterized by a spread gaussian with more important associated volume compared to protomer 1. This increase of volume is therefore concomitant with the previous asymmetry related to the various discussed structural markers. It is worth noting that this asymmetry is also found out for the DESRES simulation but to a lesser extent compared to the AMOEBA Tinker–HP simulations. Concerning the Riken dataset, this feature is not found out as both protomers depict a similar gaussian trend with very similar values.

Finally, it is also possible to study local fluctuations in the structural dynamics of the Mpro dimer system to uncover other types of differences between datasets. We calculated the fluctuation of residues in each cluster on the same 1000 previously randomly chosen structures per cluster using the Root Mean Square Fluctuation (RMSF). These were calculated on the 5 clusters from Tinker-HP (AMOEBA), the 3 clusters from Riken (AMBER) and the single cluster from DESRES (DES-AMBER). Results are depicted on Figure 8. The most interesting fluctuation as well as the main differences between clusters originates from a different spatial rearrangement of the C-terminal region of the protein (eg residues 300 to 306 on chains A and B of the dimer). In fact, this region is highly dynamical, which is in accordance with experimental X-ray observations where the electron density of the C-terminal domain was insufficient for backbone tracing, suggesting the flexibility of this region. Visual enlargements of this region is provided in subgraphs of Figure 8 for chains A and B that do not differ significantly. Cluster 1 from the DESRES simulation depicts the same fluctuation as the cluster 1 from the Riken simulation. This behaviour of the C-terminal region in these two clusters is characterized by a π–π interaction between Phe305 and His41, eventually blocking the access for any ligand to the active site. When the C terminal region does not interact with His41, it adopts an unfolded configuration which shows a high flexibility of these terminal amino acid. Structural representations can be found in Figure 9. As this event is observed on the active site of only one chain and not both of them, it could be another marker of the previously mentioned protomer inactivation. We also observed such fluctuations in the cluster 1 and 2 extracted from our Tinker–HP/AMOEBA simulations. However, in cluster 1, while the Phe305–His41 π–π interaction is indeed observed, we measure a lower fluctuation of the chain A for cluster 1. It corresponds to a weaker interaction between Phe305 and His41 as configurations where the C-terminal branch is less structured are preferred. A similar feature is observed for the cluster 2 of Riken, but with an inversion of fluctuation peaks between A and B. Overall, the clusters 1 and 2 obtained from the Tinker–HP and Riken simulations appear relatively similar in the PCA space. They correspond to clusters where the C terminal region can oscillate between two states: one with a π–π stacking interaction between Phe305 and His41, and another with a less structured C-terminal branch with higher flexibility. Clusters 4 and 5 from our Tinker–HP simulations and to a lesser extent Riken’s cluster 3 correspond to another configuration of the C-terminal region. Representative pictures are provided in Figure 9 for each cluster C-terminal conformations. In these clusters, the C-terminal region appears more preserved/organized exhibiting as it is localized further from the active site. To summarize the discussion concerning this specific feature, the high C–terminal flexibility observed in the X-ray experiments can be traced back to a modulated access to the active site linked to an absence of the π–π stacking between Phe305 and His41. In other words, the C–terminal region of the fully inactive protomer is shown to oscillate between several states and one of them directly interact with the other protomer active site. Such interaction tends to block the active site access and therefore modulating down the activity of the potentially most active site. This high flexibility is captured by both the Riken and Tinker-HP, exemplifying the importance of the local conformational sampling and comforting the experimental analysis of a full inactivation of the apo state.

To conclude, we cannot discuss here all the analysis that could be extracted from the data. In particular, we will not fully discuss here the solvation aspects of these simulations comparisons that require to go through terabits of solvated protein data. As this study is started, we can nevertheless highlight some preliminary results. Indeed, the presence of the water molecule discussed in the Apo crystal structure is observed with AMOEBA. Such water molecule interacts with the nitrogen of His41 and His163 and is clearly present in our Tinker-HP simulations (see Figure 9, from SI). This is not surprising as usually structural water molecules are known to be better described using PFF due to their high dipole moment values. To date, such agreement with experimental results is not found in other testset as the water molecules are found further away from the cyclic Nitrogen of the histidines. However, we will dedicate a full analysis paper (D. El Ahdab, in preparation) deciphering the complete simulation data (DESRES/Riken/Tinker-HP) to unravel the specific and various complex cases of the Mpro solvatation.

5 Comparative Druggability Analysis

In order to check if all the previous features could affect the druggability of the Mpro dimer system, we decided to search if new cryptic pockets are detected in each clusters. By taking into account the same sets as for the cavity volumes analysis, cryptic
Figure 7 Graphical representation of the distal + active sites for protomer 1 (on left) and protomer 2 (on right) for the DESRES, Riken and Tinker–HP simulations.

Figure 8 Representation of the RMSF for each cluster of each simulations (Tinker–HP, Riken and DESRES). Zoom for both chains (A and B) are represented in subgraph as it corresponds to the C–terminal part and is the most important fluctuations (residues 300 to 306 for chain A and B).

pockets were searched using the DoGSite Scorer software, an automated tool for pocket detection and pocket descriptors calculation. DoGSite Scorer detected 18 pockets located on the chain A or at the interface of chains A and B of the SARS–CoV–2 protease 6LU7 crystal structure. Among these pockets, 6 are already described in the literature: the pockets 'P_1_1', 'P_3' and 'P_15' corresponding to the dimerization site; the 'P_2' pocket corresponding to the active site and the 'P_6' and 'P_11' pockets located in the distal region. These 18 pockets were used as a reference and all pockets detected on the DESRES, Riken and Tinker-HP simulations.
<table>
<thead>
<tr>
<th></th>
<th>DESRES</th>
<th>Riken</th>
<th>Tinker–HP</th>
<th>Experimentally observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glu166–His172 long distance</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Glu166–His163 short distance</td>
<td>√</td>
<td>×</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>His41–Cys145 short distance</td>
<td>×</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Disordered Phe140–His163 π–π stacking</td>
<td>√</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>C terminal flexibility</td>
<td>×</td>
<td>√</td>
<td>√</td>
<td>√</td>
</tr>
<tr>
<td>Theory predicted enhanced volume cavities</td>
<td>×</td>
<td>×</td>
<td>√</td>
<td>N.A.</td>
</tr>
<tr>
<td>Theory predicted volume asymmetry between protomers</td>
<td>√</td>
<td>×</td>
<td>√</td>
<td>N.A.</td>
</tr>
</tbody>
</table>

**Table 1** Summary of the Mpro structural dynamics features predicted (√) or not (×) in our analysis on the DESRES (DES-AMBER), Riken (AMBER) and Tinker–HP (AMOEBA) simulations as compared to available experimental data extracted from references [46–48] if available (N.A. = Non available).

**Figure 9** Representation of the 3 possible states of the C terminal end. The C terminal end presented in blue depicts most of the states in clusters 1 and 2 with a His41 and Phe305 stacking. The C terminal end presented in lime depicts most of the states in cluster 3 and when presented in purple it depicts most of the states in clusters 4 and 5.
selected structures were assigned to these reference pockets by comparing the list of residues of the different pockets and selecting the reference pocket with the maximum number of common residues. When the maximum number of common residues was inferior to 5, and the ratio between the maximum number of common residues and the number of residues in the predicted pocket was inferior to 0.25, the pocket was not assigned to any reference pocket and was defined as a new cryptic pocket. New cryptic pockets were named after the first structure in which they were detected and added to the set of reference pockets. For example, the ‘R_c1_s2_P21’ mentioned in the Figure 11, is the pocket P_14 detected by DoGSite Scorer in the structure 1 (s1) of the cluster 1 (c1) of the Riken (R) simulations. The results of pockets assignment and new cryptic pockets identification are presented in the Figure 11. We observed that the reference pockets previously highlighted as ‘active site’, ‘dimerization site’ and ‘distal site’, except ‘P_6’, are particularly conserved and detected in a large majority of analyzed structures. However, a consequent number of other pockets were also detected: 1. in a few structures such as ‘R_c1_s2_P21’, ‘R_c1_s18_P14’ or ‘T_c4_s19_P3’ or 2. in many structures, such as ‘R_c1_s2_P20’, ‘R_c1_s2_P25’ or ‘R_c1_s4_P7’. Interestingly, 3 pockets were only retrieved in the clusters 4 and 5 of the Tinker-HP simulations: ‘T_c4_s2_P8’, ‘T_c4_s5_P5’ and ‘T_c4_s6_P9’. The last one, ‘T_c4_s6_P9’ is of particular interest since its volume is equal to 199 Å³ and its druggability score, DrugScore, reaches 0.62. We repeated the pocket detection and analysis procedure on 100 randomly selected structures (20 for each of the 5 clusters identified within the Tinker-HP simulations (see Figure 10, from SI). We observed that the 3 previously identified pockets ‘T_c4_s2_P8’, ‘T_c4_s5_P5’ and ‘T_c4_s6_P9’ were also detected on the structures randomly selected in the cluster 4 and 5 of the Tinker HP simulations but also partially in the cluster 3. We then evaluated if all the pockets assigned to the ‘T_c4_s6_P9’ pocket displayed similar properties. We observed that the mean volume of these pockets was 215 Å³ but few structures were presenting extreme values far superior to this mean volume (Figure 11 from SI). Similarly, the DrugScore mean value was 0.37 but with large variations among the structures and the clusters (see Figure 12, from SI). As a comparison, we also computed the DrugScore values distribution for each newly identified pocket, i.e. pockets that were not detected in the 6LU7 structure (Figure 13 from SI). One pocket, ‘R_c1_s2_P21’ displays peculiar properties with a mean druggability value of 0.6 and a mean volume value of 150 Å³ which seems to indicate that this pocket may only accommodate very small compounds. The discovery of
6 Conclusion and Perspectives

In this work, that was designed in order to respond to the urgent need of COVID-19 research, we demonstrated that it is now possible to perform large scale/long timescale MD simulations with PFF such as AMOEBA that allow for the explicit inclusion of physical many-body effects. Due to inherent complexity of the SARS-CoV-2 proteins, performing such higher-resolution simulations is important as they could provide additional information about the virus constituents structural dynamics to the COVID-19 experimental and computational research communities. To do so, we proposed a fully unsupervised adaptive sampling strategy that can be used on any type of computational resources. This fully automated framework allows for production simulations that benefit from the advances in supercomputing and from our recent Tinker-HP massively parallel software enhancements, that can now efficiently handle GPU-accelerated large petascale computers using lower precision arithmetic and MPI. In order to extract new information from this type of simulation, we also provided the necessary steps to debias the obtained data to collect useful and accurate structural dynamics features.

Results were compared to state-of-the-art large scale available data. The new generation PFF data were shown to capture most of structural dynamics features discussed in the experimental literature, confirming that the M_proof structural dynamics features discussed in the experimental literature to identify small compounds able to modulate the SARS-CoV-2 protease activity. All the herein discussed pockets are represented within the 6LU7 structure in Figure 10.

The new structural information provided here could help such studies. These data could also be important for the longer-term design of new inhibitors or to understand chemical reactivity at an atomic level via hybrid QM/MM simulations that could also use PFF to model the environment. Finally, thanks to the presented divide and conquer strategy, our AMOEBA adaptive MD simulations were shown to be simultaneously computationally competitive and more in line with available experimental data on the apo state of the SARS-CoV-2 Main protease. Using 100 GPUs cards, we show that an acceptable and competitive time to solution could be achieved as our "microsecond" results were obtained in only 10 business days (i.e. 7 full production days as each iteration took 17 hours on single V100 cards) on the academic (and multi-purpose) Jean Zay supercomputer. It is worth noting that if each simulation had run on one node (4 cards vs 1, using the Tinker-HP multi-GPU features), the time to solution would have been reduced by two. In practice, a similar exploration to the available community data was roughly achieved in only 2.5 days (Figure 1 Figure 7 SI). It is important to note that Tinker-HP can also produce order of magnitude faster simulation using n-PFF and GPUs. Since such simulation also capture many experimental details, our dual-level strategy is confirmed. Indeed, an optimal setup consists in producing first long adaptive non-polarizable simulation that can be further refined with polarizable potentials within additional adaptive iterations. That way, our approach could also use Folding@home COVID-19 community results as input (or any available data shared on the Bioexcel/Molssi repository) in order to reinject a maximum of potentially new/useful information into COVID-19 research. Indeed, it is important to recall the importance of proposing accurate (and as much as possible converged) simulations of the COVID-19 targets. Presently, the developed theoretical approach applied here to "small" M_proof protein is currently "mining" other SARS-CoV-2 proteins, including the million atoms Spike-ACE2 protein complex. Results will be shared as they are finalized. As a final perspective, we can mention that the present strategy is platform agnostic and not limited to supercomputers. Therefore, it can also be used at a smaller scale on "cheap" laboratory GPUs clusters which can benefit from the computational power of low arithmetic to obtain local supercomputing capabilities. On the other side of the spectrum, with the coming of the Exascale era and the HPC-Artificial Intelligence (AI) convergence, the "big iron" supercomputer systems, and their cloud-computing counterparts, will considerably extend the high accuracy conformational mining capabilities leading to a universe of possibilities for the in silico modeling of complex biological systems.
Conflicts of interest
There are no conflicts to declare.

Acknowledgements

This work has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (grant agreement No 810367), project EMC2 (JPP). FC acknowledges funding from the French state funds managed by the CalSimLab LABEX and the ANR within the Investissements d’Avenir program (reference ANR11-IDEX-0004-02) and support from the Direction Générale de l’Armement (DGA) Maîtrise NRBC of the French Ministry of Defense. DEA acknowledges funding from the Lebanese National Council for Scientific Research, CNRS-L. Adaptive sampling computations have been performed at GENCI thanks to a COVID19 emergency allocation on the Jean Zay machine (IDRIS, Orsay, France) on grant no A0070707671. Additional molecular dynamics simulations have been performed on the Irene Joliot Curie machine thanks to a PRACE COVID-19 emergency grant (project COVID-HP). The authors thank the Swiss National Supercomputing Center (CSCS) for hosting our data trough the FENIX infrastructure. JPP acknowledges a special COVID-19 funding from Sorbonne Université. PR is grateful for support by the Robert A. Welch Foundation (F-1691) and National Institutes of Health (R01GM106137 and R01GM114237).

References


