An Efficient GaMD Multi-Level Enhanced Sampling Strategy: Application to Polarizable Force Fields Simulations of Large Biological Systems

Frédéric Célerse,[†,‡,¶,△] Théo Jaffrelot Inizan,[†,△] Louis Lagardère,[†,§] Olivier Adjoua,[†] Pierre Monmarché,[†,∥] Yinglong Miao,[⊥] Etienne Derat,[‡] and Jean–Philip Piquemal[*,†,#,@]

†*Sorbonne Université, LCT, UMR 7616 CNRS, Paris, France*

‡*Sorbonne Université, IPCM, UMR 8232 CNRS, Paris, France*

¶*current adress: LCMD EPFL, CH–1015 Lausanne, Switrzerland*

§*Sorbonne Université, IP2CT, FR 2622 CNRS, Paris, France*

∥*Sorbonne Université, LJLL, UMR 7598 CNRS, Paris, France*

⊥ *Center for Computational Biology and Department of Molecular Biosciences, the University of Kansas, KS, USA.*

#*The University of Texas at Austin, Department of Biomedical Engineering, TX, USA*

@*Institut Universitaire de France, Paris, France*

△*Contributed equally to this work*

E-mail: jean-philip.piquemal@sorbonne-universite.fr

Abstract

We introduce a novel multi-level enhanced sampling strategy grounded on Gaussian accelerated Molecular Dynamics (GaMD). First, we propose a GaMD multi-GPUs-accelerated implementation within the Tinker-HP molecular dynamics package. We introduce the new "dual-water" mode and its use with the flexible AMOEBA polarizable force field. By adding harmonic boosts to the water stretching and bonding terms, it accelerates the solvent-solute interactions while enabling speedups thanks to the use of fast multiple–timestep integrators. To further reduce time-to-solution, we couple GaMD to Umbrella Sampling (US). The GaMD—US/dual–water approach is tested on the 1D Potential of Mean Force (PMF) of the solvated CD2–CD58 system (168000 atoms) allowing the AMOEBA PMF to converge within 1 kcal/mol of the experimental value. Finally, Adaptive Sampling (AS) is added enabling AS–GaMD capabilities but also the introduction of the new Adaptive Sampling–US–GaMD (ASUS–GaMD) scheme. The highly parallel ASUS–GaMD setup decreases time to convergence by respectively 10 and 20 times compared to GaMD–US and US. Overall, beside the acceleration of PMF computations, Tinker-HP now allows for the simultaneous use of Adaptive Sampling and GaMD-"dual water" enhanced sampling approaches increasing the applicability of polarizable force fields to large scale simulations of biological systems.

Introduction

Understanding interactions within biomolecules is crucial for many topics such as drug discovery. Some structural modifications, sometime undetected by experiment, can drastically change the nature of the physics ruling interacting complex systems. For this reason, predicting the long timsecale conformational dynamics of proteins is a long standing challenge within the conventional Molecular Dynamics (cMD) community.[1–6] It requires accurate models able to capture the true potential energy hyper-surface and long simulations to both access the large biological processes time-scale and satisfy the ergodicity principle.[7] Accelerating MD has been therefore a central field of research in the last decades.[8–11] Beside these developments, several additional strategies have been pursued overs the years to further speed up the simulations. They include the extensive use of High Performance Computing (HPC) ressources[4,12] and the optimization of GPU–accelerated modeling platforms.[13–15] Alternatively, an intensive algorithmic work has been undertaken, introducing techniques such as multiple–time–step integrator schemes[16,17] or collective variables-driven molecular dynamics methods.[18,19] The latter have been found useful in enhanced sampling and free energy calculation.[20–25] Although such methods are powerful as they can estimate free energies of binding or the stability of secondary and quaternary structures of proteins,[26,27] the free energy estimations can suffer from biases either generated by the initial choice of the collective variable (CV) or by the existence of multiple CV within the mechanism process (e.g dual mechanisms).[28] For these reasons, collective variable–free methods have become increasingly popular.[29] Among them, the recent Gaussian accelerated Molecular Dynamics (GaMD) has shown great promises due to its high sampling acceleration, its user–friendly tunable parameters and its minor additional computational cost.[30] GaMD accelerates conformational sampling by adding a harmonic boost to the potential energy. Coupled with the second order cumulant expansion, GaMD allows us to compute unbiased properties by using an accurate reweighting procedure through cumulant expansion to the second order.

Although new generation many-body polarizable force field (PFFs) are more accurate in describing biomolecular interactions,[31–34] they are computationally more challenging than traditional approaches. Therefore, to overcome these limitations, we provide here a novel general multi–level enhanced sampling strategy, that we apply to the PFF AMOEBA. To do so, we combine together the Tinker–HP massively parallel multi-GPUs platform[15] to a highly scalable GaMD implementation (*level 0*) and to additional enhanced sampling techniques based on recent developments of the field. As a first speedup, we propose an extension of the GaMD formalism with a new GaMD mode enabling the use of flexible water models such as AMOEBA[35,36] and fast multiple–time–step integrators[17] (*level 1*). We then discuss the explicit coupling of such GaMD approach to Umbrella Sampling (US)[37] and Adaptive Sampling (AS)[6] techniques (*level 2*). To demonstrate their applicability to PFF, these physics-based hybrid enhanced sampling strategies are then applied to the Potential of Mean Force (PMF) study of a large biological complex CD2–CD58 interacting via salt bridges with the AMOEBA force field. Finally, we combine all together within the Adaptive Sampling–US–GaMD method (ASUS–GaMD) scheme (*level 3*).

Method: introducing the GaMD "dual water" mode.

GaMD is a potential-biasing method for unconstrained enhanced sampling without the need to set predefined CV. It smooths the potential energy surface by adding a harmonic boost potential as described in the seminal paper[11]. Its general framework makes it suitable for the development of hybrid schemes and variants, such as replica-exchange umbrella sampling GaMD (GaREUS),[38] Ligand GaMD (LiGaMD)[39] and Peptide GaMD (Pep-GaMD).[40]

If the system potential energy is lower than a threshold energy E, a harmonic potential energy boost is applied to smooth the potential energy surface. By denoting $q \in \mathbf{R}^{3N}$ the configurations, when the system potential energy $U(q)$ is lower than a threshold energy E, a boost, which depends on $U(q)$ is added:

$$U'(q) = U(q) + \Delta U^{GaMD}(U(q)) \tag{1}$$

with $\Delta U^{GaMD}(U(q))$ the external harmonic potential boost

$$\Delta U^{GaMD}(U(q)) = \begin{cases} 0 & U(q) \leq E \\ \frac{1}{2}k(E - U(q))^2 & U(q) < E \end{cases} \tag{2}$$

and k the harmonic force constant. The two adjustable GaMD parameters k and E are automatically determined following the original procedure described in ref[30]. The boost intensity can be managed thorough a user–specified uper limit labeled as σ_0 (e.g $10k_BT$) predefined before the simulation. To ensure accurate reweighting with the cumulant expansion the ΔU^{GaMD} standard deviation, $\sigma_{\Delta V}$, should satisfy $\sigma_{\Delta V} < \sigma_0$[30,41,42]. GaMD provides different modes: the boost is either applied on the total potential GaMD–pot, on the dihedral potential GaMD–dih, or on both at the same time GaMD–dual[43,44]. Recently, another mode was introduced:LiGaMD which adds the boost to a ligand non bonded interactions,[39] accelerating the sampling of ligand-protein interactions. It is known that interactions involving water are essential for such systems and that protein stability processes are controlled by water-protein interactions.[6,45,46] To accelerate these interactions, one would like to use the GaMD–dual mode on the non bonded interactions of water molecules. However, such boost requires the evaluation of the complete non bonded energies and, in the context of multi-timestep integrators such as BAOAB–RESPA1[17] where they are split between short and long range, these are only available at outer (large) timestep. This type of integrators enables the use of larger time steps and thus a direct acceleration of molecular dynamics. For example, the BAOAB–RESPA1 is based on a RESPA (Reference System Propagator Algorithm[16]) three-level splitting of forces (bonded, short-range non- bonded and long-range non-bonded) within the Leimkuhler's BAOAB discretization of Langevin dynamics.[47] It allows up to a 7 folds acceleration for polarizable point dipole molecular dynamics.[17] But the fluctuations of the associated bias are such that it has to be evaluated at shorter timesteps, so that the whole procedure is not compatible with multi-timestep integrators such as BAOAB–RESPA1. For similar reasons, the GaMD–dual mode with a bias applied to the complete potential energy is not compatible with even simple RESPA integrators in which the potential energy is split between bonded and non bonded terms. Therefore, GaMD–dual mode becomes rapidly limited by the simulation time. To overcome this issue, we developed a new mode, GaMD–dualwater (denoted GaMD–dualwat), which adds a boost to the protein dihedral potential energy term and the water stretching and bending terms, this time fully compatible with RESPA and RESPA1 like integrators, allowing water molecule to be more flexible and thus favoring their conformational changes.

$$\Delta U^{GaMD-dw}(U(q)) = \Delta U^{dihedral}_{protein}(U(q)) + \Delta U^{stretch}_{water}(U(q)) + \Delta U^{bend}_{water}(U(q)) \tag{3}$$

This mode is enabled by the flexibility of the AMOEBA 03 water model[35] but is not compatible with rigid water models such as TIP3P[48] commonly used with the CHARMM and AMBER force field.[49] This framework allows to further reduce the computational cost gap between PFFs and nPFFs. This new mode, in addition to the other GaMD–dih and GaMD–dual modes, is now available within the Tinker–HP software.[12,15] In the following, we first tested its GPU scalability and performance on the STMV system ($\simeq$1 066 624 atoms), and its sampling efficiency is demonstrated on simulations of the alanine dipeptide and the CD2-CD58 complex. A technical appendix is present at the end of the manuscript and provides the formalism of the method and the associated debiasing equations.

Results and discussion

Level 0: Efficiency and GPU scalability

The GaMD implementation is such that only a small computational and communication (in parallel) overhead is added compared to cMD. The GaMD–dih and GaMD–dualwat have

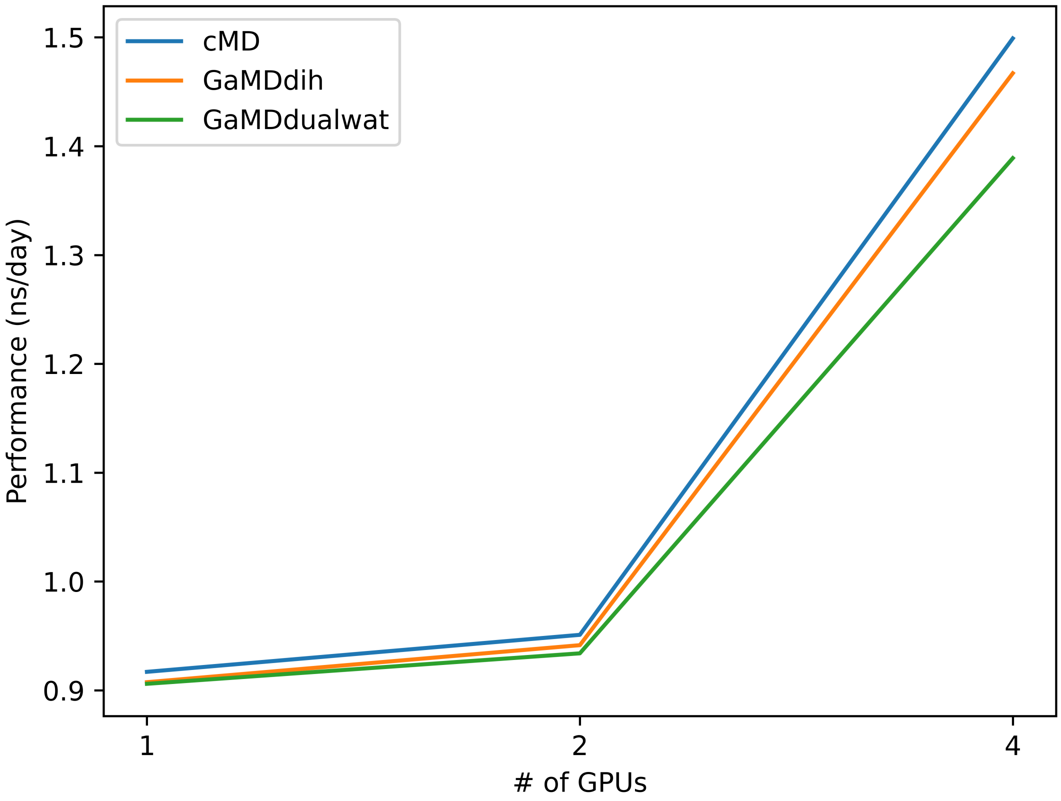

Figure 1: GaMDdih and GaMDdualwat scaling performance on 1/2/4 V100 GPUs (i.e. corresponding to a full node of the Jean Zay machine) on STMV (1 066 624 atoms) with the AMOEBA force field and the BAOAB–RESPA1 10 fs multiple–time–step integrator. The cMD reference, in blue, allows to evaluate the GaMD impacts on the code communications.

been considered on the STMV system (1 066 624 atoms) with the AMOEBA PFF and the 10 fs outer time–step HMR BAOAB–RESPA1 multiple–time–step integrator.[17] V100 GPUs from the national Jean Zay supercalculator have been used for all the benchmark computations. Similar scalability studies have been performed on the Jean Zay multi-CPUs (Figure S1). The AMOEBA GPU simulations were performed on a single node as the multi-node extension of the AMOEBA PFF within the Tinker–HP package is still under development.On 1 and 2 GPUs (Figure 1), the GaMD data communications are negligible, 1%. On 4 GPUs, the communications are increasing and the performance decreases by 7%. Overall, the use of GaMD only slightly alters the performance. This high scalability opens the door to simulate at a high-accuracy large complex biomolecular systems with PFFs.

Level 1: GaMD–dualwat with PFFs

We compared GaMD–dih, GaMD–dual and GaMD–dualwat sampling acceleration on the exploration of the relevant basins of the alanine dipeptide (e.g α_r, α_L and P_{II}). The alanine dipeptide is solvated in a cubic 20 Å water box. We used the many-body AMOEBABIO18 PFF.[50,51] The system was minimized with a RMS of 1 kcal/mol and sampled within the NPT thermodynamic ensemble with the Bussi thermostat[52] and a MonteCarlo barostat[53] at 300 K and 1 atmosphere. We used the Velocity Verlet integrator and a 1 fs time–step.[54] Smooth Particle Mesh Ewald (SPME) algorithm was employed to compute non–covalent interactions[55] with a real space cutoff equal to 7 Å and a Van der Waals cutoff set to 9 Å. For AMOEBA the convergence criteria for multipoles was set to 10^{-5}. After short testing simulations, we found an optimal value of 3 kcal/mol for GaMD–dih and GaMD–dual σ_0, in accordance with ref[11], and 4 kcal/mol for GaMD–dualwat (see Figure S2, Table S1 and S2). We ran 3 independent simulations of 60 ns for each mode. The different sampled basins are also compared to a 1 μs cMD AMOEBA reference.

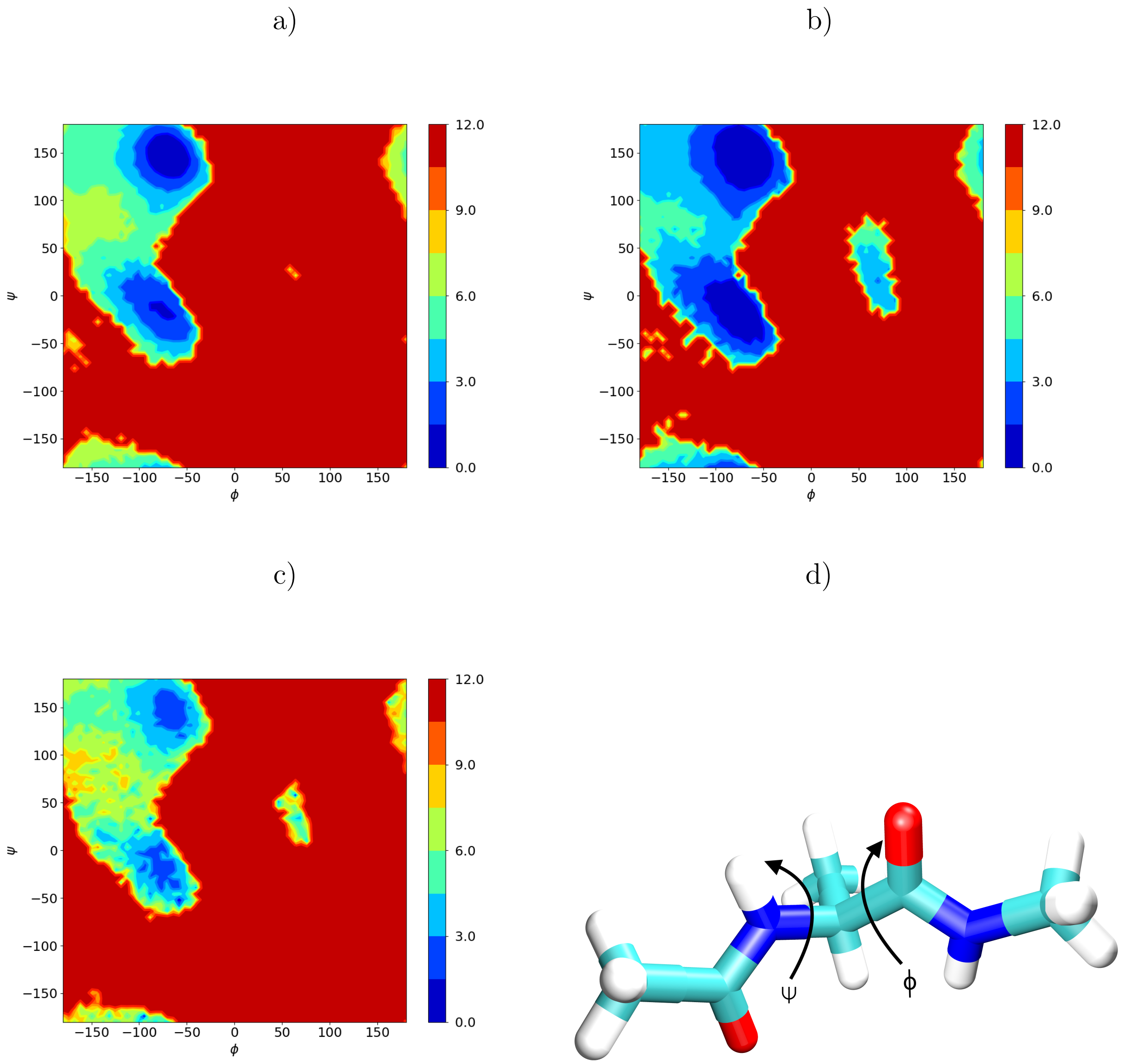

Figure 2: 2D PMF (in kcal/mol) of the alanine dipeptide obtained in AMOEBA for a) GaMD–dih mode (3x60ns) b) GaMD–dual mode (3x60ns) and c) GaMD–dualwat mode (3x60ns) d) Alanine dipeptide representation with the corresponded Φ and Ψ angles.

Reweighted, see Technical Appendix, free energy surfaces obtained from these simulations are depicted on Figure 2 and show that GaMD–dual captures well the α_r (50°,25°), α_L (-75°,-25°) and P$_{II}$ (-75°,150°) basins. These results are consistent with the 1 μs cMD trajectory (see Figure S3 in SI) depicting these three basins. While the GaMD–dih mode captures the α_r basin after 150 ns, the GaMD–dualwat captures it in 100 ns (SI Figure S4). We also observe a sampling acceleration between GaMD–dual and GaMD–dualwat compared to the reference 1 microsecond cMD. To characterize the GaMD boost harmonicity, its distribution anharmonicity γ is calculated as in.[30] γ serves as an indicator of the sampling convergence and reweighting procedure accuracy. Depicted on Figure S5 in SI GaMD–dih as well as GaMD–dual depicts high anharmonicity with respectively 0.252 and 0.016 compared to GaMD–dualwat with 0.0005. Additionally, we see a steep anharmonicity convergence to less than 10^{-3} for GaMD–dualwat while being relatively stable at 2×10^{-1} for GaMD–dih (SI Figure S4). In comparison the anharmonicity is about 0.001 with GaMD–dih and AMBER99SB. PFFs thus increase the statistical noise and stress the importance of using low anharmonicity GaMD modes. In that sense, GaMD–dualwat appears more suitable than GaMD–dual for PFFs simulations with an anharmonicity equal to 0.0005. As stated before, another advantage of GaMD–dualwat is that it can be coupled to multiple–time–step such as BAOAB–RESPA1[17] in contrast to the GaMD–dual mode that remains limited to single timestep integrators. Comparative results of GaMD–dualwat with both integrators can be found in Figure S6 of the SI. Its coupling with multiple–time–step clearly compensates the slightly lower sampling performance compared with GaMD–dual. The sampling enhancement brought by the GaMD–dualwat can be partly related to how it affects the diffusion of water: we report in Table S3 of the SI the self diffusion coefficients of bulk water computed within a same setup (same size of box and same integrator) and observe that it is increased with the GaMD–dualwat mode compared to the simple GaMD–dih one, favoring global conformationnal changes due to water reorganization. While the added sampling efficiency is already significant for the alanine dipeptide, we expect it to be larger on more complex and larger biological systems such as CD2CD58 where water reorganization plays a bigger role. Combined with a highly-parallel GPUs infrastructures and multiple–time–step integrators, the GaMD–dualwat should allow to help reaching very high-resolution conformational space of large molecular systems. In addition to the sampling acceleration it provides, the low associated anharmonicity drastically reduces the statistical noise associated with reweighting.

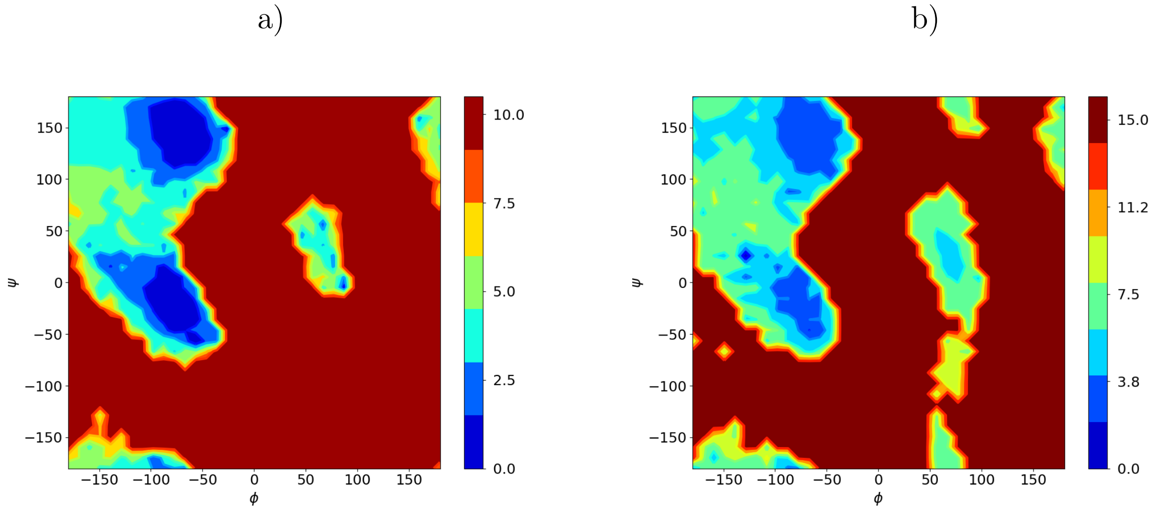

Figure 3: PMFs of a) GaMD dualwat (3x60=180ns)and b) AS–GaMD dualwat (5x25=125ns) simulations for the alanine dipeptide. For the GaMD dualwat simulations we performed 3 independent simulations of 60ns using 1fs timestep with the verlet integrator. The AS–GaMD dualwat simulations were performed using the BAOAB–RESPA1 10fs multi timestep integrator and 5 AS iterations of 5x5ns with a square term.

Level 2: Speeding-up simulations with the parallel AS–GaMD scheme

We further coupled our newly introduced GaMD mode to additional enhanced sampling strategies. Recently, we developed a new adaptive sampling technique (AS) which was shown to allow massive sampling of the SARS–CoV–2 Main Protease conformational space.[56] We coupled these two methodologies together, yielding the AS–GaMD method. The principle is similar to the AS, the only modification being that each cMD at each iteration is now a GaMD simulation. The double bias coming from both AS and GaMD implies that a suitable and careful reweigthing scheme has to be introduced to reconstruct unbiased free energy surface. All mathematical tools for the reweigthing scheme are provided in the Technical

Appendix. We applied this methodology to the same system, the alanine dipeptide using the same GaMD simulation protocol. At each iteration, we projected the structures on the two main dihedral angles space. To push the limit of the AS–GaMD sampling capability we combined a modified version of the AS selection scheme with the BAOAB–RESPA1 multi-timestep integrator. The probability law for selection of new structures was taken as the inverse of the square of the probability density on the reduced space which further amplifies the exploration of undiscovered region.

On Figure 3, we represented the 2D PMF obtained with both AS–GaMD/BAOAB–RESPA1 and GaMD/VERLET simulations. For AS–GaMD/BAOAB–RESPA1 we performed 5 iterations of 5x5ns GaMD–dualwat simulations for a total simulation time of 125ns. As in the previous section, the GaMD/VERLET is composed of 3 independent simulation of 60ns (180ns total). In 30% less simulation time and 5 times less computational time, thanks to the natural AS parallelism, the coupled AS–GaMD/multi-timestep integrator scheme greatly enhance the exploration of the free energy surface. We observed that the α_L region is already captured at the first iteration, i.e with only 25 ns (SI Figure S7). In addition, other states, next to the α_L region, are captured within tens of ns and are still not seen after the whole GaMD simulation. This AS–GaMD/multi-timestep coupling can thus represent an important gain for the sampling of biomolecular systems.

Level 3: Pushing the limit of PMF convergence with GaMD–US and ASUS–GaMD

US has been widely used and is mathematically robust but it is still suffer from several issues.[57–59] In addition to the choice of the CVs, it is also difficult to estimate the PMF convergence as it is system dependent. Good indicators to check if convergence is reached are: the overlap between neighboring windows and the evolution of the PMF curve as a function of the simulation time per window. To accelerate the sampling within each window, Oshima et al. recently combined GaMD with replica-exchange and US.[38] Here, we first only applied

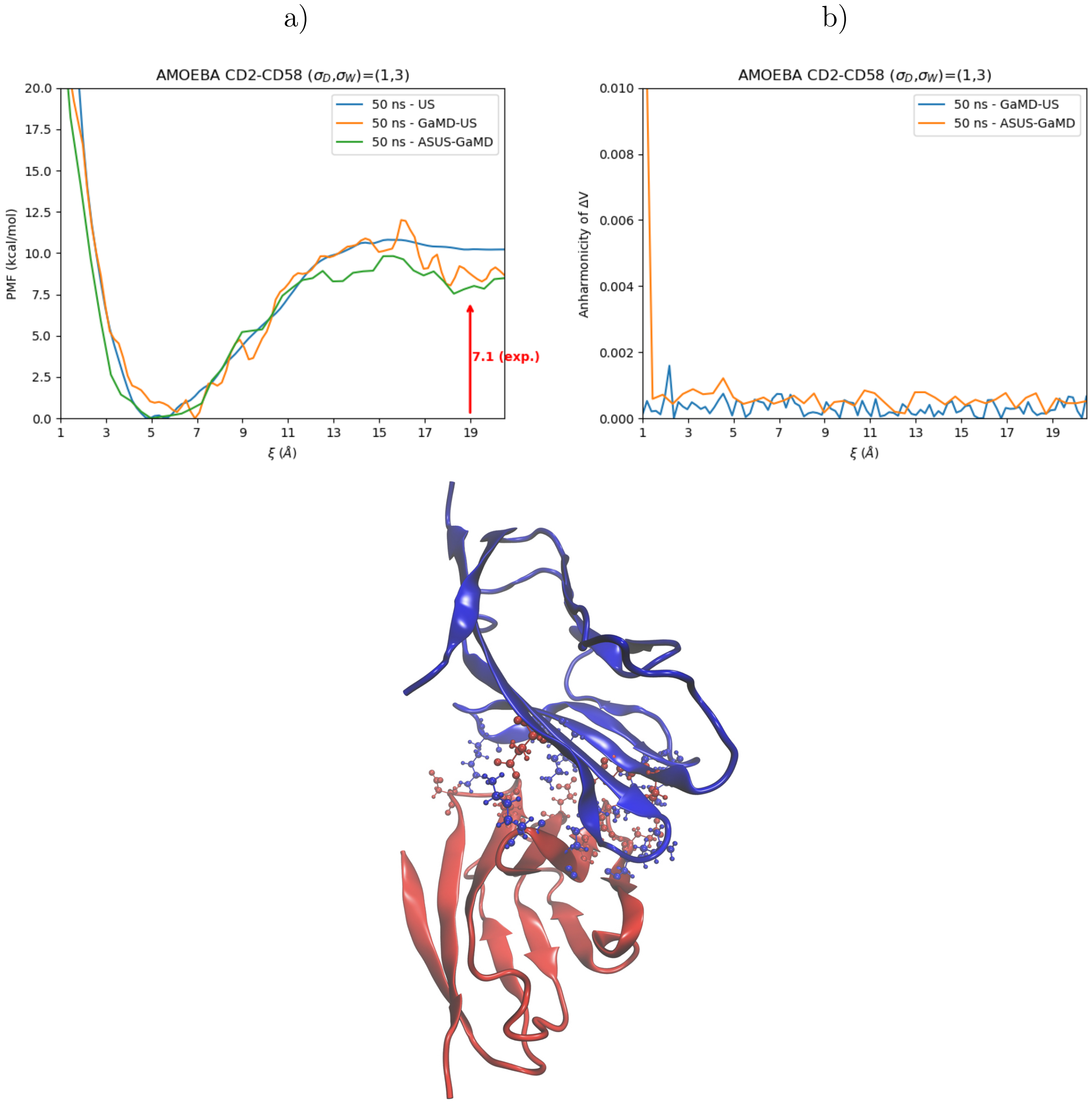

Figure 4: c) PDB 1QA9 CD2CD58 representation with CD2 and CD58 subcomplexes represented respectively in blue and red using the newribbons representation. Residues at the interface considered in the COM distance between the two subcomplexes are represented in blue and red for respective basic and acid residues using the CPK representation. VMD software was employed to generate the structure. PMFs obtained with US, GaMD–US and ASUS–GaMD are depicted in a) and their respective anharmonicity in b).

a GaMD boost in each US window in order to enhance the sampling in the orthogonal space. To demonstrate the PMF convergence acceleration, we studied the dissociation of the salt bridges interface within the CD2CD58 complex. This system, made of several salt bridges and hydrogen bondings interactions, was already studied by some of us.[28] Although it has been shown that PFFs allow a better description of the salt bridges interactions, their computational cost has long hindered the study of such large system. Since the portability of Tinker-HP on multi-GPU and the global acceleration of the PFFs, reaching such system is now easily achievable. To start this study we took the same CD2CD58 complex as in our previous work[28] but we solvated it in a waterbox of 100 × 100 × 100 Å. Counterion were added to neutralize the system. We used the AMOEBABIO18 PFF.[50,51] The system was minimized with a RMS of 1 kcal/mol in the NVT thermodynamic ensemble with the Bussi thermostat.[52] Temperature was set to 300 K while pressure was set to 1 atmosphere. We used the multiple–time–step BAOAB–RESPA1 with a 10 fs timestep with the Hydrogen Mass Repartionning scheme (HMR)[17] and Smooth Particle Mesh Ewald (SPME) algorithm to compute electrostatic and polarization interactions[55] with a real space cutoff of 7 Å and a Van der Waals cutoff of 9 Å. The convergence criteria for polarization was set to 10^{-5}. 39 US windows were generated, ranging from 1 to 20 Å with a width of 0.5 Å between them. CV was chosen as the distance between the center of mass formed by the interfacial residues isolated by Bayas et al. on CD2 and CD58 (Table 1 in ref[60]). A spring constant of 10 kcal/mol.Å^2 was employed to restrain the system along the chosen CV. Each window was run for 5 ns for equilibration and then for 50 ns. Histogram overlap as well as the PMF curve as a function of the simulation time allocated per window were employed to check the convergence of the simulations (SI Figure S9). The final US PMF show a slow decrease of the free energy barrier with the simulation time, suggesting a slow convergence to about 12.5 kcal/mol. Binding affinity was found to be experimentally around 7.1 ± 0.03 kcal/mol, suggesting that our simulations are not converged.[60] In order to improve sampling within each window a new US was performed similar to the previous US protocol but now with an additional GaMD–dualwat potential applied in each window. The GaMD parametrization protocol and reweighting procedure are described in the Technical Appendix and in SI (Figure S8 and Table S4). The optimized GaMD–dualwat parameters σ_0 are equal to 1 and 3 kcal/mol for respectively dihedral and dual water modes. Figure 4 shows the difference between standard US and GaMD–US. The GaMD–US PMF and boost harmonicity converge at 40 ns per window (SI Figure S10 A/, B/ and C/). The predicted free energy barrier is now within the 1 kcal/mol of the experiment. It shows that GaMD–dualwat, even without presence of Replica Exchange, could considerably improve PMF convergence of large systems. It also demonstrates that salt bridges and, more generally protein-protein interactions are well described with PFFs. Further, as demonstrated in Debiec et al.,[61] the improved accuracy of non-PFFs in describing these interactions requires the implicit incorporation of solvent polarization, underscoring the importance of polarization effects in these contexts.

To further push the sampling, we coupled together GaMD, AS and US (ASUS–GaMD). We provide two reweighting schemes that either use modified Multistate Bennett Acceptance Ratio (MBAR) equations or the Rao-Blackwell estimator. The mathematical expressions are general and can be used with any weighted dynamics. Starting from an initial US simulation ($\simeq$ few ns), each window is decomposed in several AS independent trajectories with an additional GaMD–dualwat potential boost (GaMD). Here, we ran 2 iterations of 5×5 ns GaMD–US per window. The PMF evolution can be found in SI (Figure S11) while the resulting PMF is depicted on Figure 4. We observe that ASUS–GaMD reach GaMD–US in one iteration, showing the sampling acceleration impact provided by the AS part within ASUS–GaMD. Note that the rough aspect of the PMF obtained with the methods involving a GaMD bias comes from the debiaising of the boost potential, as can be seen in previous work involving US and GaMD.[38] Although a careful reweighting is needed for the different AS, GaMD and US layers, the overall ASUS–GaMD approach inherits the strong adaptive sampling advantages of being pleasantly parallelizable and considerably accelerates the PMF convergence.

Conclusions

Combined with the use of modern GPUs, these sampling techniques allow to drastically reduce time to solution in PFFs evalution of PMFs. Although it is difficult to truly quantify the final speedup (i.e. a PMF convergence remains partially system-dependant), one can see in Figure S12(SI) that if we extrapolate the US convergence, ASUS–GaMD converges 1.4 times faster. Thanks to the native parallelism inherited from AS, the PMF evaluation can be done in one fifth of the simulation time yielding a speedup of 7. If we consider that convergence was already reached with a 25ns per window setup, this factor grows to 14. ASUS–GaMD can thus reduce to days computation that would have taken months. This work also allows to invoke any variant of the combined approaches, offering therefore access to GPU-accelerated GaMD-adaptive sampling (AS–GaMD) simulations that will be helpful to further extend conformational space studies of proteins[6] To conclude, these methodologies will contribute further to allow high-resolution sampling of large biological systems up to millions of atoms using polarizable force field.

Technical Appendix

We denote by $\xi(q)$ the reaction coordinate along which we performed the US simulation and q the configuration. Here, a configuration means the positions $q \in \mathbf{R}^{3N}$ of all the atoms of the system. The imposed US bias potential is

$$U_j^{US}(q) = \mathcal{K}(\xi(q) - \xi_j)^2 \tag{4}$$

with $\mathcal{K}$ the force constant.

We combined the AS, US and GaMD such that each US window $j \in [\![1, ..., M]\!]$, $\xi_1, ..., \xi_M$, is parallelized and accelerated by adaptive sampling replicas and GaMD boost potential:

$$U_j''(q) = U(q) + U^{GaMD}(q) + U_j^{US}(q) \tag{5}$$

We denote by $(q_{j,n})_{n \in 1,N}$ the N configurations generated by the AS replicas of US window j and $(\omega_{j,n})_{n \in 1,N}$ their respective AS weights. These normalized weights are defined as $\omega_{j,n} = \frac{N v_{j,n}}{\sum_{m=1}^{N} v_{j,m}}$ so that $\sum_{n=1}^{N} \omega_{j,n} = N$ with $v_{j,n}$ the unnormalized AS weights. The canonical average of an observable φ is estimated by

$$\langle \varphi \rangle_j'' = \frac{\int \varphi(q) e^{-\beta U_j''(q)} \, \mathrm{d}q}{\int e^{-\beta U_j''(q)} \, \mathrm{d}q} \simeq \frac{\sum_{n=1}^{N} \varphi(q_{j,n}) \omega_{j,n}}{\sum_{n=1}^{N} \omega_{j,n}} = \frac{1}{N} \sum_{n=1}^{N} \varphi(q_{j,n}) \omega_{j,n} \tag{6}$$

In practice, to get a smooth reweighted PMF, the reaction coordinate ξ is discretized in K bins around values $x_1, ..., x_K$. We want to estimate for each $k \in [\![1, ..., K]\!]$ its free energy, up to an additive constant,

$$F(x_k) = -\frac{1}{\beta} \ln \mathbb{P}(\xi(q) \in Bin(x_k)), \tag{7}$$

where q is distributed according to the density probability law $\frac{e^{-\beta U}}{\int e^{-\beta U}}$, i.e.

$$F(x_k) = -\frac{1}{\beta} \ln \frac{\int 1_{\xi(q) \in Bin(x_k)} e^{-\beta U(q)} \, \mathrm{d}q}{\int e^{-\beta U(q)} \, \mathrm{d}q} = -\frac{1}{\beta} \ln \langle \varphi_k \rangle \tag{8}$$

with $\varphi_k = 1_{\xi(q) \in Bin(x_k)}$.

1st step: GaMD with cumulant expansion

We, first, remove the GaMD bias. Here, we want to find a relation between $\langle \varphi \rangle$ and $\langle \varphi \rangle'$ where the prime average represents the canonical average over the potential $U' = U + U^{GaMD}$. Starting from the canonical average, we notice :

$$\langle \varphi \rangle = \frac{\int \varphi(q) e^{-\beta U(q)} \, \mathrm{d}q}{\int e^{-\beta U(q)} \, \mathrm{d}q} = \frac{\int \varphi(q) e^{\beta U^{GaMD}(q)} e^{-\beta U'(q)} \, \mathrm{d}q}{\int e^{\beta U^{GaMD}(q)} e^{-\beta U'(q)} \, \mathrm{d}q} = \frac{\langle \varphi e^{\beta U^{GaMD}} \rangle'}{\langle e^{\beta U^{GaMD}} \rangle'} \tag{9}$$

By applying this with $\varphi = \varphi_k$,

$$F(x_k) = -\frac{1}{\beta} \ln \frac{\langle \varphi_k e^{\beta U^{GaMD}} \rangle'}{\langle e^{\beta U^{GaMD}} \rangle'} = -\frac{1}{\beta} \ln \langle \varphi_k e^{\beta U^{GaMD}} \rangle' + C = F'(x_k) - \frac{1}{\beta} \ln \frac{\langle \varphi_k e^{\beta U^{GaMD}} \rangle'}{\langle \varphi_k \rangle'} + C \tag{10}$$

where C is a constant and $F'(x_k)$ is the free energy $F'(x_k) = -\frac{1}{\beta} \ln \langle \varphi_k \rangle'$. To reduce the estimator variance, we used the cumulant expansion to the second order,

$$\ln \frac{\langle \varphi_k e^{\beta U^{GaMD}} \rangle'}{\langle \varphi_k \rangle'} \simeq \beta \frac{\langle \varphi_k U^{GaMD} \rangle'}{\langle \varphi_k \rangle'} + \frac{\beta^2}{2} \left(\frac{\langle \varphi_k (U^{GaMD})^2 \rangle'}{\langle \varphi_k \rangle'} - \left(\frac{\langle \varphi_k U^{GaMD} \rangle'}{\langle \varphi_k \rangle'} \right)^2 \right) \tag{11}$$

By combining with equation (10), the free energy is rewritten as

$$F(x_k) \simeq -\frac{1}{\beta} \ln \langle \varphi_k \rangle' - \beta \frac{\langle \varphi_k U^{GaMD} \rangle'}{\langle \varphi_k \rangle'} - \frac{\beta^2}{2} \left(\frac{\langle \varphi_k (U^{GaMD})^2 \rangle'}{\langle \varphi_k \rangle'} - \left(\frac{\langle \varphi_k U^{GaMD} \rangle'}{\langle \varphi_k \rangle'} \right)^2 \right) + C \tag{12}$$

2nd step: AS modified MBAR

Finally, we want to express $\langle\varphi\rangle'$ w.r.t the AS weights in each US window $j \in [\![1, ..., M]\!]$. This can be done in two ways either using the MBAR or the Rao-Blackwell estimator.

Modified MBAR

Let's define

$$c'_j = \int e^{-\beta U''_j(q)}\, \mathrm{d}q, \quad F'_j = -\frac{1}{\beta} \ln c'_j \tag{13}$$

The prime comes from the use of the MBAR on the reference energy U' of the previous section. The starting point is to use the MBAR identity (ref[62] equation 5) and notice

$$c'_i \langle e^{\beta U''_j} \alpha_{i,j} \rangle''_i = \int e^{-\beta U''_j(q)} e^{-\beta U''_i(q)} \alpha_{i,j}(q)\, \mathrm{d}q = c'_j \langle e^{\beta U''_i} \alpha_{i,j} \rangle''_j \tag{14}$$

which holds for arbitrary functions $q \longrightarrow \alpha_{ij}(q)$ with $i, j \in [\![1, ..., M]\!]$. Notice that each window generated the same number of configurations N. The MBAR estimator has been proven to be optimal by using

$$\alpha_{i,j}(q) = \frac{1/c'_j}{\sum_{k=1}^{M} e^{-\beta U''_k(q)}/c'_k} \tag{15}$$

and by summing over j

$$c'_i \sum_{j=1}^{M} \left\langle \frac{e^{-\beta U''_j}/c'_j}{\sum_{k=1}^{M} e^{-\beta U''_k}/c'_k} \right\rangle''_i = \sum_{j=1}^{M} c'_j \left\langle \frac{e^{-\beta U''_i}/c'_j}{\sum_{k=1}^{M} e^{-\beta U''_k}/c'_k} \right\rangle''_j \tag{16}$$

We obtain a set of M equations for all $i \in [\![1, ..., M]\!]$

$$c'_i = \sum_{j=1}^{M} \left\langle \frac{e^{-\beta U_i(q)}}{\sum_{k=1}^{M} e^{-\beta U_k(q)}/c'_k} \right\rangle''_j \tag{17}$$

Using equation (6) we obtain the estimators

$$\hat{c}_i' = \frac{1}{N} \sum_{j=1}^{M} \sum_{n=1}^{N} \frac{\omega_{j,n} e^{-\beta U_i(q_{j,n})}}{\sum_{k=1}^{M} e^{-\beta U_k(q_{j,n})}/\hat{c}_k'} \tag{18}$$

and finally with eq (13)

$$F_i' = -\frac{1}{\beta} \ln \left(\frac{1}{N} \sum_{j=1}^{M} \sum_{n=1}^{N} \frac{\omega_{j,n} e^{-\beta U_i(q_{j,n})}}{\sum_{k=1}^{M} e^{\beta(F_k' - U_k(q_{j,n}))}} \right) \tag{19}$$

which must be solve self-consistently.

Modified Rao-Blackwell estimator

Recently, Ding *et al.*[63,64] derived the MBAR equations using Rao-Blackwell (RB) estimator. The RB theorem characterizes the transformation of a crude estimator into a better estimator that has smaller mean-squared-error w.r.t to the dataset.

We wish to calculate the $i \in [\![1, ..., M]\!]$ relative free energies F_i^* of M thermodynamic states sampled independently, with potential U_i. To compute the relative free energies, the system should be sampled according to Boltzmann distribution. We note $q_{i,n}$ the $n \in [\![1, ..., N_i]\!]$ configurations sampled from state i. To compute the relative free energies of the M thermodynamic states, the configurations $q_{i,n}$ are combined and considered as samples from the generalized ensemble $p_i(q) \propto e^{-\beta(U_i(q)+b_i)}$ where b_i is an unknown biased energy. This biased energy was introduced[63] to adjust the relative weight of state i to be proportional to N_i, leading to

$$F_i = F_i^* + b_i = -\frac{1}{\beta} \ln \frac{N_i}{N} \tag{20}$$

where $N = \sum_{i=1}^{M} N_i$. From this equation we can then use the RB estimator

$$\begin{aligned} F_i &= -\frac{1}{\beta}\ln p_i = -\frac{1}{\beta}\ln \frac{1}{N}\sum_{j=1}^{M}\sum_{n=1}^{N} p_i(q_{j,n}) \\ &= -\frac{1}{\beta}\ln \frac{1}{N}\sum_{j=1}^{M}\sum_{n=1}^{N} \frac{\omega_{j,n} e^{-\beta(U_i(q_{j,n})+b_i)}}{\sum_{k=1}^{M} e^{-\beta(U_k(q_{j,n})+b_i)}} \end{aligned} \tag{21}$$

Combining with equation (20):

$$1 = \frac{1}{N}\sum_{j=1}^{M}\sum_{n=1}^{N} \frac{\omega_{j,n} e^{-\beta(U_i(q_{j,n})+b_i)}}{\sum_{k=1}^{M} e^{-\beta(U_k(q_{j,n})+b_i)}} \tag{22}$$

Thus, the unbiased free energy F_i^* can be calculated using equation (20) after solving (22) for b_i. Equation 22 has major interests: (1) it is more stable, (2) it reduces the number of floating point operations and (3) the problem is reduced to minimizing a convex function. Indeed, if we define

$$g_i(b_1, ..., b_M) = \frac{1}{N}\sum_{j=1}^{M}\sum_{n=1}^{N} \frac{\omega_{j,n} e^{-\beta(U_i(q_{j,n})+b_i)}}{\sum_{k=1}^{M} e^{-\beta(U_k(q_{j,n})+b_i)}} - 1 \tag{23}$$

then solving (22) is equivalent to finding the zeros of $(g_1, \ldots, g_M)$. Moreover, we can remark that the function $g_i = \nabla_{b_i} f$ where the function f is given by

$$f(b_1, ..., b_M) = -\frac{1}{N}\sum_{j=1}^{M}\sum_{n=1}^{N} \omega_{j,n} \ln\left(\sum_{k=1}^{M} e^{-\beta(U_k(q_{j,n})+b_i)}\right) - \sum_{j=1}^{M} b'_j, \tag{24}$$

which means solving (22) is equivalent to finding the critical points of f. Ding *et al.* shown that f is convex so the problem is reduced to minimizing this function which can be done with the L-BFGS method. The reweighting procedure, that use part of the FastMBAR code, takes few minutes on a single GPU. In this work we used the latter procedure thanks to its GPU efficiency.

3rd step: ASUS-GaMD reweighting

With either using the MBAR or the RB estimator procedure, we can extract the, still biased, free energies. The final step is to derive an expression of $\langle\varphi\rangle'$ w.r.t either $\hat{c}'_k$ or F'_k. By setting $c_0 = \int e^{-\beta U'(q)}\,\mathrm{d}q$ and using (6),

$$\begin{aligned} c_0\langle\varphi\rangle' = \int \varphi(q)e^{-\beta U'(q)}\,\mathrm{d}q &= \sum_{i=1}^{M} \frac{\int \varphi(q)e^{-\beta U'(q)}e^{-\beta U''_i(q)}/c'_i\,\mathrm{d}q}{\sum_{j=1}^{M} e^{-\beta U''_j(q)}/c'_j} \\ &= \sum_{i=1}^{M} \left\langle \frac{\varphi}{\sum_{j=1}^{M} e^{-\beta U_j(q)}/c'_j} \right\rangle''_i \\ &\simeq \frac{1}{N}\sum_{i=1}^{M}\sum_{n=1}^{N} \frac{\varphi(q_{i,n})\omega_{i,n}}{\sum_{k=1}^{M} e^{-\beta U_k(q_{j,n})}/\hat{c}'_k} \end{aligned} \tag{25}$$

in other words,

$$c_0\langle\varphi\rangle' \simeq \frac{1}{NM}\sum_{i=1}^{M}\sum_{n=1}^{N} \varphi(q_{i,n})r_{i,n} \tag{26}$$

with $r_{i,n}$ the weight of configuration $q(i,n)$

$$r_{i,n} = \frac{M\omega_{i,n}}{\sum_{k=1}^{M} e^{-\beta U_k(q_{i,n})}/\hat{c}'_k} \tag{27}$$

c_0 is unknown but does not depends of $k \in [\![1,...,K]\!]$ so equation (12) can be rewritten as

$$F(x_k) \simeq -\frac{1}{\beta}\ln\langle c_0\varphi_k\rangle' - \beta\frac{c_0\langle\varphi_k U^{GaMD}\rangle'}{c_0\langle\varphi_k\rangle'} - \frac{\beta^2}{2}\left(\frac{c_0\langle\varphi_k(U^{GaMD})^2\rangle'}{c_0\langle\varphi_k\rangle'} - \left(\frac{c_0\langle\varphi_k U^{GaMD}\rangle'}{c_0\langle\varphi_k\rangle'}\right)^2\right) + C' \tag{28}$$

with

$$\begin{aligned} c_0\langle\varphi_k\rangle' &\simeq \frac{1}{NM}\sum_{i=1}^{M}\sum_{n=1}^{N} r_{i,n}\mathbb{1}_{\xi(q_{i,n})\in Bin(x_k)} \\ c_0\langle\varphi_k U^{GaMD}\rangle' &\simeq \frac{1}{NM}\sum_{i=1}^{M}\sum_{n=1}^{N} U^{GaMD}(q_{i,n})r_{i,n}\mathbb{1}_{\xi(q_{i,n})\in Bin(x_k)} \end{aligned} \tag{29}$$

Acknowledgement

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 thanks 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 Génerale de l'Armement (DGA) Maîtrise NRBC of the French Ministry of Defense. Computations have been performed at GENCI (IDRIS, Orsay, France and TGCC, Bruyères le Chatel) on grant no A0070707671.

Supporting Information Available

- Figure S1: CPUs scalability of GaMD
- Figure S2, Tables S1 and S2: Parametrization of GaMD parameters for the alanine dipeptide
- Figure S3: Free energy surface of alanine dipeptide obtained from a long cMD simulation
- Table S3: The GaMD influence on the water diffusion property
- Figure S4: Evolution of anharmonicity and basins's populations as a function of the simulation time for the alanine dipeptide
- Figure S5: Boost distribution and anharmonicity from AMOEBA GaMD simulations on the alanine dipeptide
- Figure S6: GaMD applied with Velocity Verlet and BAOAB–RESPA1 multi–time–step integrator
- Figure S7: Evolution of the AS–GaMD sampling as a function of the simulation time

- Figure S8 and Table S4: Parametrization of GaMD parameters for the CD2–CD58
- Figure S9: US convergence on CD2–CD58
- Figure S10: GaMD–US convergence on CD2–CD58
- Figure S11: ASUS–GaMD convergence on CD2–CD58
- Figure S12: US, GaMD–US and ASUS–GaMD extrapolated convergence

References

(1) Caves, L. S.; Evanseck, J. D.; Karplus, M. Locally accessible conformations of proteins: multiple molecular dynamics simulations of crambin. *Protein. Sci.* **1998**, *7*, 649–666.

(2) Schlitter, J.; Engels, M.; Krüger, P. Targeted molecular dynamics: a new approach for searching pathways of conformational transitions. *J. Mol. Graph.* **1994**, *12*, 84–89.

(3) Markwick, P. R.; McCammon, J. A. Studying functional dynamics in bio-molecules using accelerated molecular dynamics. *Phys. Chem. Chem. Phys.* **2011**, *13*, 20053–20065.

(4) Shaw, D. E.; Dror, R. O.; Salmon, J. K.; Grossman, J.; Mackenzie, K. M.; Bank, J. A.; Young, C.; Deneroff, M. M.; Batson, B.; Bowers, K. J., et al. Millisecond-scale molecular dynamics simulations on Anton. Proceedings of the conference on high performance computing networking, storage and analysis. 2009; pp 1–11.

(5) Lindorff-Larsen, K.; Maragakis, P.; Piana, S.; Shaw, D. E. Picosecond to Millisecond Structural Dynamics in Human Ubiquitin. *The Journal of Physical Chemistry B* **2016**, *120*, 8313–8320, PMID: 27082121.

(6) Jaffrelot Inizan, T.; Célerse, F.; Adjoua, O.; El Ahdab, D.; Jolly, L.-H.; Liu, C.; Ren, P.; Montes, M.; Lagarde, N.; Lagardère, L.; Monmarché, P.; Piquemal, J.-P. High-resolution mining of the SARS-CoV-2 main protease conformational space: supercomputer-driven unsupervised adaptive sampling. *Chem. Sci.* **2021**, *12*, 4889–4907.

(7) Cho, K.; Joannopoulos, J. Ergodicity and dynamical properties of constant-temperature molecular dynamics. *Phys. Rev. A.* **1992**, *45*, 7089.

(8) Hamelberg, D.; Mongan, J.; McCammon, J. A. Accelerated molecular dynamics: a

promising and efficient simulation method for biomolecules. *J. Chem. Phys.* **2004**, *120*, 11919–11929.

(9) Voter, A. F. Hyperdynamics: Accelerated molecular dynamics of infrequent events. *Phys. Rev. Lett.* **1997**, *78*, 3908.

(10) Pierce, L. C.; Salomon-Ferrer, R.; Augusto F. de Oliveira, C.; McCammon, J. A.; Walker, R. C. Routine access to millisecond time scale events with accelerated molecular dynamics. *J. Chem. Theory. Comput.* **2012**, *8*, 2997–3002.

(11) Miao, Y.; Feixas, F.; Eun, C.; McCammon, J. A. Accelerated molecular dynamics simulations of protein folding. *J. Comput. Chem.* **2015**, *36*, 1536–1549.

(12) Lagardère, L.; Jolly, L.-H.; Lipparini, F.; Aviat, F.; Stamm, B.; Jing, Z. F.; Harger, M.; Torabifard, H.; Cisneros, G. A.; Schnieders, M. J., et al. Tinker-HP: a massively parallel molecular dynamics package for multiscale simulations of large complex systems with advanced point dipole polarizable force fields. *Chem. Sci.* **2018**, *9*, 956–972.

(13) Götz, A. W.; Williamson, M. J.; Xu, D.; Poole, D.; Le Grand, S.; Walker, R. C. Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. *Journal of Chemical Theory and Computation* **2012**, *8*, 1542–1555, PMID: 22582031.

(14) Páll, S.; Zhmurov, A.; Bauer, P.; Abraham, M.; Lundborg, M.; Gray, A.; Hess, B.; Lindahl, E. Heterogeneous parallelization and acceleration of molecular dynamics simulations in GROMACS. *The Journal of Chemical Physics* **2020**, *153*, 134110.

(15) Adjoua, O.; Lagardère, L.; Jolly, L.-H.; Durocher, A.; Very, T.; Dupays, I.; Wang, Z.; Inizan, T. J.; Célerse, F.; Ren, P., et al. Tinker-HP: Accelerating Molecular Dynamics Simulations of Large Complex Systems with Advanced Point Dipole Polarizable Force Fields Using GPUs and Multi-GPU Systems. *J. Chem. Theory. Comput.* **2021**, *17*, 2034–2053.

(16) Tuckerman, M. E.; Berne, B. J.; Rossi, A. Molecular dynamics algorithm for multiple time scales: Systems with disparate masses. *J. Chem. Phys.* **1991**, *94*, 1465–1469.

(17) Lagardère, L.; Aviat, F.; Piquemal, J.-P. Pushing the limits of multiple-time-step strategies for polarizable point dipole molecular dynamics. *J. Phys. Chem. Lett.* **2019**, *10*, 2593–2599.

(18) Fiorin, G.; Klein, M. L.; Hénin, J. Using collective variables to drive molecular dynamics simulations. *Mol. Phys.* **2013**, *111*, 3345–3362.

(19) Tribello, G. A.; Bonomi, M.; Branduardi, D.; Camilloni, C.; Bussi, G. PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* **2014**, *185*, 604–613.

(20) Brotzakis, Z. F.; Limongelli, V.; Parrinello, M. Accelerating the calculation of protein–ligand binding free energy and residence times using dynamically optimized collective variables. *J. Chem. Theory. Comput.* **2018**, *15*, 743–750.

(21) Kabelka, I.; Brozek, R.; Vácha, R. Selecting Collective Variables and Free-Energy Methods for Peptide Translocation across Membranes. *J. Chem. Inf. Mod.* **2021**, *61*, 819–830.

(22) Hovan, L.; Comitani, F.; Gervasio, F. L. Defining an optimal metric for the path collective variables. *J. Chem. Theory. Comput.* **2018**, *15*, 25–32.

(23) Bonati, L.; Rizzi, V.; Parrinello, M. Data-driven collective variables for enhanced sampling. *J. Phys. Chem. Lett.* **2020**, *11*, 2998–3004.

(24) Hashemian, B.; Millán, D.; Arroyo, M. Modeling and enhanced sampling of molecular systems with smooth and nonlinear data-driven collective variables. *J. Chem. Phys.* **2013**, *139*, 12B601_1.

(25) Chen, P.-Y.; Tuckerman, M. E. Molecular dynamics based enhanced sampling of collective variables with very large time steps. *J. Chem. Phys.* **2018**, *148*, 024106.

(26) Henin, J.; Fiorin, G.; Chipot, C.; Klein, M. L. Exploring multidimensional free energy landscapes using time-dependent biases on collective variables. *J. Chem. Theory. Comput.* **2010**, *6*, 35–47.

(27) Gardner, J. M.; Abrams, C. F. Energetics of flap opening in HIV-1 protease: string method calculations. *J. Phys. Chem. B.* **2019**, *123*, 9584–9591.

(28) Célerse, F.; Lagardère, L.; Derat, E.; Piquemal, J.-P. Massively parallel implementation of Steered Molecular Dynamics in Tinker-HP: comparisons of polarizable and non-polarizable simulations of realistic systems. *J. Chem. Theory. Comput.* **2019**, *15*, 3694–3709.

(29) Rodriguez, A.; d'Errico, M.; Facco, E.; Laio, A. Computing the free energy without collective variables. *J. Chem. Theory. Comput.* **2018**, *14*, 1206–1215.

(30) Miao, Y.; Feher, V. A.; McCammon, J. A. Gaussian accelerated molecular dynamics: Unconstrained enhanced sampling and free energy calculation. *J. Chem. Theory. Comput.* **2015**, *11*, 3584–3595.

(31) Melcr, J.; Piquemal, J.-P. Accurate Biomolecular Simulations Account for Electronic Polarization. *Frontiers in Molecular Biosciences* **2019**, *6*, 143.

(32) Shi, Y.; Ren, P.; Schnieders, M.; Piquemal, J.-P. *Reviews in Computational Chemistry Volume 28*; John Wiley Sons, Ltd, 2015; Chapter 2, pp 51–86.

(33) Jing, Z.; Liu, C.; Cheng, S. Y.; Qi, R.; Walker, B. D.; Piquemal, J.-P.; Ren, P. Polarizable Force Fields for Biomolecular Simulations: Recent Advances and Applications. *Annual Review of Biophysics* **2019**, *48*, 371–394, PMID: 30916997.

(34) Gresh, N.; Cisneros, G. A.; Darden, T. A.; Piquemal, J.-P. Anisotropic, Polarizable Molecular Mechanics Studies of Inter- and Intramolecular Interactions and Ligand-

Macromolecule Complexes. A Bottom-Up Strategy. *J. Chem. Theory Comput.* **2007**, *3*, 1960–1986.

(35) Ren, P.; Ponder, J. W. Polarizable atomic multipole water model for molecular mechanics simulation. *J. Phys. Chem. B.* **2003**, *107*, 5933–5947.

(36) Ponder, J. W.; Wu, C.; Ren, P.; Pande, V. S.; Chodera, J. D.; Schnieders, M. J.; Haque, I.; Mobley, D. L.; Lambrecht, D. S.; DiStasio Jr, R. A., et al. Current status of the AMOEBA polarizable force field. *J. Phys. Chem. B.* **2010**, *114*, 2549–2564.

(37) Torrie, G. M.; Valleau, J. P. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *J. Comp. Phys.* **1977**, *23*, 187–199.

(38) Oshima, H.; Re, S.; Sugita, Y. Replica-exchange umbrella sampling combined with gaussian accelerated molecular dynamics for free-energy calculation of biomolecules. *J. Chem. Theory. Comput.* **2019**, *15*, 5199–5208.

(39) Miao, Y.; Bhattarai, A.; Wang, J. Ligand Gaussian accelerated molecular dynamics (LiGaMD): Characterization of ligand binding thermodynamics and kinetics. *J. Chem. Theory. Comput.* **2020**, *16*, 5526–5547.

(40) Wang, J.; Miao, Y. Peptide Gaussian accelerated molecular dynamics (Pep-GaMD): Enhanced sampling and free energy and kinetics calculations of peptide binding. *J. Chem. Phys.* **2020**, *153*, 154109.

(41) Miao, Y.; McCammon, J. A. *Annu. Rep. Comput. Chem.*; Elsevier, 2017; Vol. 13; pp 231–278.

(42) Wang, J.; Arantes, P. R.; Bhattarai, A.; Hsu, R. V.; Pawnikar, S.; Huang, Y.-m. M.; Palermo, G.; Miao, Y. Gaussian accelerated molecular dynamics: Principles and applications. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2021**, e1521.

(43) Wang, Y.-T.; Chan, Y.-H. Understanding the molecular basis of agonist/antagonist mechanism of human mu opioid receptor through gaussian accelerated molecular dynamics method. *Sci. Rep.* **2017**, *7*, 1–11.

(44) Palermo, G. Structure and dynamics of the CRISPR–Cas9 catalytic complex. *J. Chem. Inf. Model.* **2019**, *59*, 2394–2406.

(45) de Courcy, B.; Piquemal, J.-P.; Garbay, C.; Gresh, N. Polarizable water molecules in ligand- macromolecule recognition. Impact on the relative affinities of competing pyrrolopyrimidine inhibitors for FAK kinase. *J. Am. Chem. Soc.* **2010**, *132*, 3312–3320.

(46) El Ahdab, D.; Lagardère, L.; Inizan, T. J.; Célerse, F.; Liu, C.; Adjoua, O.; Jolly, L.-H.; Gresh, N.; Hobaika, Z.; Ren, P.; Maroun, R. G.; Piquemal, J.-P. Interfacial Water Many-Body Effects Drive Structural Dynamics and Allosteric Interactions in SARS-CoV-2 Main Protease Dimerization Interface. *The Journal of Physical Chemistry Letters* **2021**, *12*, 6218–6226, PMID: 34196568.

(47) Leimkuhler, B.; Matthews, C. Robust and efficient configurational molecular sampling via Langevin dynamics. *The Journal of chemical physics* **2013**, *138*, 05B601_1.

(48) Mark, P.; Nilsson, L. Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K. *The Journal of Physical Chemistry A* **2001**, *105*, 9954–9960.

(49) MacKerell Jr, A. D.; Bashford, D.; Bellott, M.; Dunbrack Jr, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S., et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B.* **1998**, *102*, 3586–3616.

(50) Shi, Y.; Xia, Z.; Zhang, J.; Best, R.; Wu, C.; Ponder, J. W.; Ren, P. Polarizable atomic multipole-based AMOEBA force field for proteins. *J. Chem. Theory. Comput.* **2013**, *9*, 4046–4063.

(51) Zhang, C.; Lu, C.; Jing, Z.; Wu, C.; Piquemal, J.-P.; Ponder, J. W.; Ren, P. AMOEBA polarizable atomic multipole force field for nucleic acids. *J. Chem. Theory. Comput.* **2018**, *14*, 2084–2108.

(52) Bussi, G.; Donadio, D.; Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

(53) Chow, K.-H.; Ferguson, D. M. Isothermal-isobaric molecular dynamics simulations with Monte Carlo volume sampling. *Comput. Phys. Commun.* **1995**, *91*, 283–289.

(54) Swope, W. C.; Andersen, H. C.; Berens, P. H.; Wilson, K. R. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.* **1982**, *76*, 637–649.

(55) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. A smooth particle mesh Ewald method. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(56) Inizan, T. J.; Célerse, F.; Adjoua, O.; El Ahdab, D.; Jolly, L.-H.; Liu, C.; Ren, P.; Montes, M.; Lagarde, N.; Lagardère, L., et al. High-resolution mining of the SARS-CoV-2 main protease conformational space: supercomputer-driven unsupervised adaptive sampling. *Chem. Sci.* **2021**, *12*, 4889–4907.

(57) You, W.; Tang, Z.; Chang, C.-e. A. Potential mean force from umbrella sampling simulations: What can we learn and what is missed? *J. Chem. Theory. Comput.* **2019**, *15*, 2433–2443.

(58) Baştuğ, T.; Chen, P.-C.; Patra, S. M.; Kuyucak, S. Potential of mean force calculations of ligand binding to ion channels from Jarzynski's equality and umbrella sampling. *J. Chem. Phys.* **2008**, *128*, 04B614.

(59) Mills, M.; Andricioaei, I. An experimentally guided umbrella sampling protocol for biomolecules. *J. Chem. Phys.* **2008**, *129*, 114101.

(60) Bayas, M. V.; Kearney, A.; Avramovic, A.; Van Der Merwe, P. A.; Leckband, D. E. Impact of salt bridges on the equilibrium binding and adhesion of human CD2 and CD58. *J. Biol. Chem.* **2007**, *282*, 5589–5596.

(61) Debiec, K. T.; Gronenborn, A. M.; Chong, L. T. Evaluating the strength of salt bridges: a comparison of current biomolecular force fields. *J. Phys. Chem. B.* **2014**, *118*, 6561–6569.

(62) Shirts, M. R.; Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **2008**, *129*, 124105.

(63) Ding, X.; Vilseck, J. Z.; Hayes, R. L.; Brooks, C. L. Gibbs Sampler-Based -Dynamics and Rao–Blackwell Estimator for Alchemical Free Energy Calculation. *Journal of Chemical Theory and Computation* **2017**, *13*, 2501–2510, PMID: 28510433.

(64) Ding, X.; Vilseck, J. Z.; Brooks, C. L. Fast Solver for Large Scale Multistate Bennett Acceptance Ratio Equations. *J. Chem. Theory. Comput.* **2019**, *15*, 799–802.

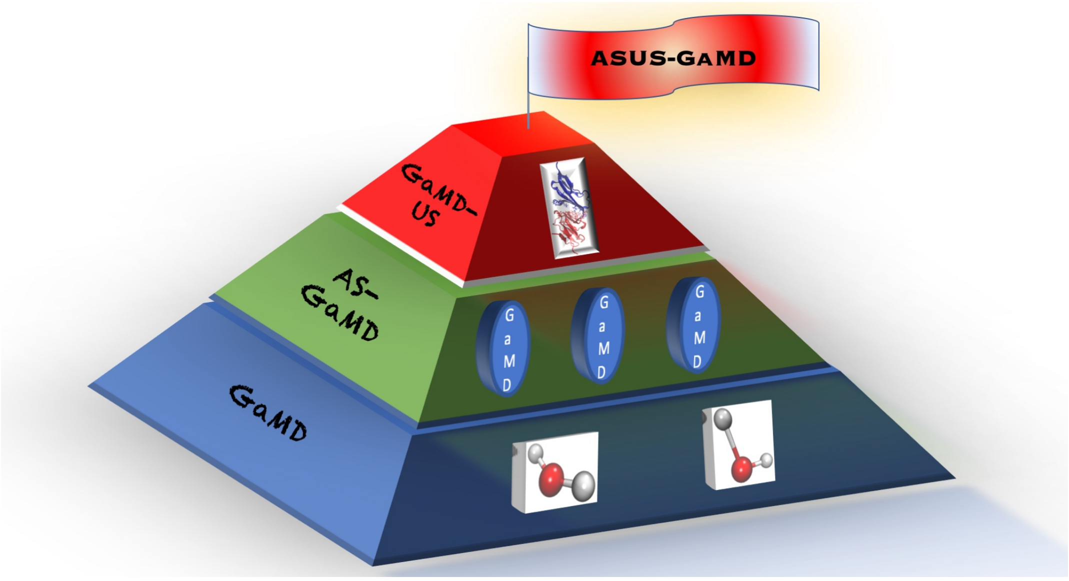

Figure 5: TOC