BAR-based Multi-dimensional Nonequilibrium Pulling for Indirect Construction of QM/MM Free Energy Landscapes: Varying the QM Region

Zhaoxi Sun[1*], and Zhirong Liu[1]

[1]*Beijing National Laboratory for Molecular Sciences, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China*

*To whom correspondence should be addressed: z.sun@pku.edu.cn

Abstract

The indirect construction of the free energy landscape at Quantum mechanics (QM)/ molecular mechanics (MM) levels provides a feasible alternative to the direct QM/MM free energy simulations. The main idea under the indirect method is constructing a thermodynamic cycle, exploring the configurational space at a computationally efficient but less accurate low-level Hamiltonian, and performing an alchemical correction to obtain the thermodynamics at an accurate but computationally demanding high-level Hamiltonian. In our previous works, we developed a multi-dimensional nonequilibrium free energy simulation framework to obtain QM/MM free energy landscapes indirectly. Specifically, we considered obtaining semi-empirical QM (SQM) results by combining the MM results and the MM-to-SQM correction and obtaining the QM results by combining the SQM results and the SQM-to-QM correction. In this work, we explore the possibility of changing the region for electronic structure calculations in the multi-scale QM/MM treatment, which could also be considered as a change of the level of theory. More generally, the multi-dimensional nonequilibrium Hamiltonian-variation/perturbation framework could be used to obtain transformations between different Hamiltonians of interest, such as changing the QM theory, the size of the QM region, and the basis set simultaneously.

I. Introduction

Molecular dynamics (MD) simulations provide a feasible route to access the detailed atomistic motions during complex physical, chemical and biological processes. The statistical quantity named free energy difference extracted from MD trajectories depicts the thermodynamic variation along the reaction pathway. The barriers in the free energy profile or the free energy landscape in complex systems are often too high, resulting in a time scale inaccessible in unbiased simulations. As a result, the convergence behavior of the brute-force simulations is poor and the estimates of observables are often biased. Enhanced sampling techniques could be employed to overcome this sampling issue.[1-4] These methods rely on the modifications of the energy landscape coupled with proper post-processing reweighting/perturbation procedure[5-7] to obtain the Boltzmann-weighted statistics in the original unperturbed ensemble. A traditional and representative method is umbrella sampling,[3, 8-10] which adds harmonic biasing potentials to enhance the sampling efficiency in specific regions of phase space. The root of the widely used reweighting methods[11-13] is free energy perturbation (FEP).[14] Its hysteresis problem[5-7] and underestimation of the statistical error[15, 16] could be avoided to some extent by using the bidirectional reweighting scheme named Bennett Acceptance Ratio (BAR).[17, 18] As perturbation-based reweighting methods rely on the overlap between neighboring states to obtain reliable results, if the states of interest show significant differences, the staging technique or the stratification strategy should be employed to increase the overlap in each perturbation step and thus improve the convergence behavior. For instance, in the flipping of a base in nucleotide systems, the base-paired and flipping-out conformational states are significantly different. To obtain a reliable estimate of the free energy profile along the base flipping pathway, a series of intermediate states need to be introduced.[3, 19-21] The more general forms of FEP and BAR are their nonequilibrium extensions, i.e., Jarzynski's Identity (JI)[22] and Crooks' Equation (CE).[23] They generalize the energy difference in FEP and BAR to the overall microscopic nonequilibrium work (NEW) performed on the system during the conformational change.[24-30] The simulation strategy suitable for these estimators is the steered MD (SMD) method, where a time-dependent (often harmonic) biasing potential is added to drive the system from one conformational state to the other. When the target states are distant from each other, similar to the equilibrium case, the nonequilibrium stratification could be employed to improve the convergence of the SMD simulations. The nonequilibrium alternative shows similar accuracy and efficiency in the construction of the potential of mean force (PMF) in various cases.[5-7, 16, 19, 20, 31-38]

As the modifications of the energy function are performed in the Hamiltonian space, another possibility is perturbating in the alchemical space.[36, 39-46] The alchemical method deals with the time-scale problem in a

different way. It introduces thermodynamic cycles and estimates the thermodynamic quantity of interest by combining the results of each leg of the cycle.[38, 47-51] For instance, the free energy of binding of the protein-ligand or host-guest systems is difficult to calculate when the binding/unbinding pathway is complex, but can be easily computed by constructing a thermodynamic cycle and using the alchemical method to calculate the free energy difference of each leg of the cycle.[38, 42, 45, 52-57] Thus, the alchemical method could be viewed as an indirect way to obtain the quantity of interest. Extrapolation of the description of the system could also be performed with the alchemical method.[58-68] The method is computationally efficient when the end states in the alchemical extrapolation are similar in energetics but significantly different in computational costs. For instance, the thermodynamics at ab initio quantum mechanics (QM) levels could be obtained by employing the alchemical method to extrapolate the results at semi-empirical QM (SQM) levels.[15] As molecular mechanics (MM) force fields often show significant differences with various QM and SQM Hamiltonians, performing MM-to-SQM perturbations suffers from convergence problems.[69] As these levels of theory differ significantly in computational cost, introducing intermediate states requires extensive equilibrium sampling under the computationally demanding Hamiltonian, which could degrade the efficiency especially when the intermediate-state sampling is difficult to converge. In this case, the nonequilibrium technique provides a computationally feasible and accurate way to enhance the convergence without equilibrium sampling in the intermediate states.[53, 70-73]

The PMF along the structural collective variable (CV) depicts the variations of thermodynamics in the process, while the PMF along the alchemical pathway is often not physically meaningful. Further, the mechanism of chemical reactions or biological conformational changes could be extracted from the PMF in the physically meaningful space (i.e., the conformational space). Therefore, the free energy landscape in the conformational space could be more informatic.[70, 74] Higher accuracy and efficiency could be obtained by combining the merits of the enhanced sampling techniques in the conformational and alchemical spaces. The enhanced sampling simulation in the conformational space could be used to obtain a converged picture along some physically meaningful CVs under a computationally efficient Hamiltonian, and the perturbation term in the alchemical space could be used to obtain a correction term describing the difference between the computationally efficient Hamiltonian and a more accurate but computationally demanding Hamiltonian. Combining these results, the thermodynamic profile under the accurate but computationally demanding Hamiltonian could be obtained.[67, 75-83] Examples of such simulation scheme are the indirect constructions of free energy landscapes at QM or SQM levels by performing direct free energy simulations at MM or SQM levels and calculating the perturbation term for MM-to-SQM or SQM-to-QM corrections.[15, 67, 69, 75-80, 82, 84-86]

In the multi-scale treatment of complex systems, the whole simulated system is divided into several sub-regions that are described with different models/Hamiltonians.[84, 87-90] Adaptive exchanges between the descriptions/resolutions could sometimes/periodically happen.[91-94] The region that the reaction or conformational change happens is often described with a model of higher accuracy, while the background region could be described with a coarser model. For instance, in the formation/cleavage of chemical bonds, electronic structure calculations are required to describe the QM behavior of bond rearrangements, while the other regions such as solvent molecules far from the reactive region could be treated with MM potentials. As the simulation outcome shows significant dependence on the size of each region, determining the size of each region could be a significant problem in model construction. In practical applications, the region for detailed descriptions is often chosen sufficiently large to converge the simulation outcome on this degree of freedom.

In our previous works on the indirect construction of QM/MM free energy landscapes from multi-dimensional nonequilibrium pulling simulations, we considered obtaining the SQM and QM free energy landscapes by combining MM and SQM free energy landscapes and MM-to-SQM and SQM-to-QM corrections. In the first SQM-from-MM-perturbation case,[69] the description of the system is changed from MM to SQM, where both the level of theory and the size of the QM region are changed. The size of the QM region grows from zero to several residues, and the description of the whole system becomes multi-scale. In the second QM-from-SQM-perturbation case,[15] the QM region is unchanged and only the level of theory is changed. A recent work on equilibrium perturbation-based indirect QM/MM free energy simulation considers the expansion of the QM region,[95] which could alter the thermodynamic profile in a nucleophilic addition reaction. However, due to the significant difference between MM and QM Hamiltonians, statistically significant noises are introduced in the equilibrium perturbation-based extrapolation even with the curve fitting procedure to smooth the PMF, resulting in many unexpected local minima in the free energy profile. In this work, we provide a general nonequilibrium alternative for the indirect scheme. We perform direct nonequilibrium free energy simulations at one QM level and QM region, and calculate the perturbation term with the nonequilibrium alchemical method to correct the free energy profile to another QM level and QM region. The nonequilibrium framework minimizes the statistical noises and the resulting indirect free energy profile is as smooth as the direct result. More generally, as the perturbation is performed on the description of the system (i.e., the Hamiltonian), this perturbation-based indirect scheme could be used to extrapolate on multiple degrees of freedom, e.g., the QM basis set, the QM level, the QM region and so on. For instance, the basis-set extrapolation could be performed between some computationally efficient

ones and some larger ones (e.g., the smaller 3-21G and the larger 6-311G* in the Pople series, or cc-pVDZ and the larger counterparts in the Dunning basis sets for the complete basis set extrapolation), which serves as a nice way to perform the basis-set extrapolation of the thermodynamic profiles. The QM level could also be extrapolated from a low-level one (e.g., PM6) to a high-level one (e.g., MP2). Further, the QM region could be expanded or shrunk to check the convergence of the QM results on the selection/definition of the QM region. In the following parts of the manuscript, we will provide a brief review of the nonequilibrium Hamiltonian-variation scheme and some illustrative examples to prove its applicability in practical systems.

II. Methodology

In the previous case, as the focus of the perturbation variable is the level of theory, we employed the low-level and high-level language to describe the Hamiltonians. In the current case, as we are considering the perturbation in the Hamiltonian space with various degrees of freedom, we write the scheme more generally by using the state-specified Hamiltonian itself. The Hamiltonian H_{k_1,k_2} denotes the Hamiltonian of the system at the state (k_1, k_2). The first dimension describes the conformational change, and the second dimension describes the Hamiltonian/model change. An illustration of the thermodynamic cycle constructed to perform the Hamiltonian perturbation via nonequilibrium transformations is shown in Fig. 1a. Free energy simulations exploring the configurational space are performed under one Hamiltonian, while free energy calculations in the alchemical space provide a correction term accounting for the difference between thermodynamics under different Hamiltonians. Free energy simulations in both spaces are performed with the nonequilibrium stratification framework. The SMD simulations along the configurational CV employ the time-dependent harmonic potential V given below, [96]

$$V(\mathbf{q}) = \frac{k}{2}\left(\xi(\mathbf{q}) - \xi_0(t)\right)^2 \quad (1).$$

Here, k denotes the force constant, $\mathbf{q}$ is the coordinate vector, $\xi_0(t)$ represents the time-dependent protocol for CV variation defining the configurational transformation, and $\xi(\mathbf{q})$ refers to the current value of the CV. To minimize the fluctuations of CV and define a configurational state precisely, a large force constant is often used to achieve the stiff spring limit.[36, 37, 69, 97-101] Care should be taken when the magnitude of the force constant is very large. The integration time step for the equations of motion should be altered accordingly to avoid unstable dynamics and the resulting perturbations of distributions.[15, 24-26] There is no fluctuation for the alchemical CV and its value is varied monotonically according to the pre-defined

schedule.

The conformational change or chemical reaction often involves significant rearrangements of many atoms. Thus, to achieve a good numerical behavior and reduce dissipation, the pulling simulations need to be sufficiently long to relax many involved degrees of freedom. This behavior is not satisfactory for computational and experimental inspections, as a long waiting time is needed before user feedback and protocol diagnosing. Stratification, in this case, is often favored. A long pulling excursion is divided into a series of smaller segments, which enhances the convergence of the nonequilibrium free energy simulation and provides faster results output. We consider the case that the whole reaction or conformational change is divided into K conformational states and thus there are $K-1$ smaller segments. The situation is a bit different for periodic configurational CV, where the number of conformational states is the same as the number of segments.

In nonequilibrium pulling simulations, only the nonequilibrium transformations between neighboring states are performed. The reason is that the reliability of the perturbation-based reweighting is highly dependent on the phase space overlap between different states and thus the similarity of the nonequilibrium works during the pulling processes in the forward and backward directions. Therefore, the reweighting is essentially unidirectional or bidirectional in nonequilibrium simulations. In this case, the statistically optimal and non-parametric bidirectional estimator BAR/CE is favored over other estimators. Note that in some cases, using other estimators (e.g., using JI or Gaussian approximations) could save some computational costs. However, these estimators also have their deficiencies such as the underestimation of the standard error (JI) and the systematic bias that is hard to quantify (Gaussian approximation).[15, 16] As the current work aims at achieving the best generality and transferability, we do not consider these case-specific alternations in the estimator selection.

The statistically optimal and asymptotically unbiased bidirectional perturbation estimator CE is described with Eq. (2).[17, 18, 23]

$$\begin{cases} \Delta A_{ij} = \ln \dfrac{\left\langle f(W_{ji}+C)\right\rangle_j}{\left\langle f(W_{ij}-C)\right\rangle_i} + C \\ C = \Delta A_{ij} + \ln(\dfrac{n_j}{n_i}) \end{cases} \qquad (2),$$

where A denotes the dimensionless free energy, ΔA is thus the corresponding difference between free energies of different states, $\langle \ldots \rangle_i$ represents to the canonical average over nonequilibrium realizations

initiated from state i, W_{ij} is the dimensionless work accumulated during the nonequilibrium pulling initiated from state i and ended in state j, n denotes the number of samples in each ensemble, and f is the Fermi weighting function. The bidirectional CE could achieve unbiased estimates with a reasonable sample size for all kinds of distributions, thus achieving the highest transferability and generality. Its dimensionless variance could be expressed as

$$\sigma_{ij}^2 = \frac{\mathrm{Var}\left(f_{ij}\right)}{n_i^2 f_{ij}^2} + \frac{\mathrm{Var}\left(f_{ji}\right)}{n_j^2 f_{ji}^2} \tag{3}.$$

Here, Var represents the absolute variance, and f_{ij} is defined as

$$\begin{aligned} f_{ij} &= \left\langle f(W_{ij} + C_{ij}) \right\rangle_i \\ f_{ji} &= \left\langle f(W_{ji} - C_{ij}) \right\rangle_j \\ n_{ij} f_{ij} &= n_{ji} f_{ji} \end{aligned} \tag{4}.$$

As only the transformations between neighboring states are performed, here $j = i+1$. For the statistically meaningful estimation of the statistical error/variance in Eq. (3), the input of the estimator in Eq. (2) should be statistically independent. Therefore, in the extraction of the initial configurations for nonequilibrium realizations from equilibrium ensembles, we calculate the autocorrelation of the reaction coordinate in each state τ_i and the statistical inefficiency $\phi_i = 1 + 2\tau_i$, and then subsample the whole time series of configurations by the statistical inefficiency to obtain a set of independent configurations. The statistical inefficiency provides an estimate of the computational cost of obtaining an independent sample from equilibrium simulations. Similarly, the computational cost of each sample in nonequilibrium pulling could be estimated as the sum of the statistical inefficiency in the equilibrium state $\phi_{\mathrm{eq},i}$ and the length of pulling simulations $\phi_{\mathrm{NEW},i}$, namely

$$\phi_i = \phi_{\mathrm{NEW},i} + \phi_{\mathrm{eq},i} \tag{5}.$$

As we are considering bidirectional pulling in the current work, the length of pulling for each data point is the sum of the pulling times in the forward (ξ increasing) and backward (ξ decreasing) directions.

The free energy difference and the corresponding statistical error in each segment could be accumulated to obtain the variation of the free energy in the whole process, namely

$$\Delta A_{1k} = \sum_{i=1}^{k-1} \Delta A_{i,i+1} \tag{6},$$

$$\sigma_{1k}^2 = \sum_{i=1}^{k-1} \sigma_{i,i+1}^2 = \sum_{i=1}^{k-1} \left(\frac{\mathrm{Var}\left(f_{i,i+1} \right)}{n_i^2 f_{i,i+1}^2} + \frac{\mathrm{Var}\left(f_{i+1,i} \right)}{n_{i+1}^2 f_{i+1,i}^2} \right) \quad (7).$$

As there are K states, when the subscript $k = K$, Eq. (6) gives the overall free energy difference along the pathway. In the visualization of the simulation results, it is often preferred to set a reference point. Here, we choose the 1st state as the reference point with a free energy of zero,

$$A_1 = 0 \quad (8).$$

Then, the one-dimensional free energy profile along the configurational CV could be obtained by combining Eq. (6) and Eq. (8),

$$A_k = \Delta A_{1k} = \sum_{i=1}^{k-1} \Delta A_{i,i+1} \quad (9).$$

A similar procedure could be employed to construct the free energy profile along the alchemical CV. Combining these two free energy profiles, the two-dimensional free energy surface could be obtained. The subscripts in the above equations should be altered to describe the multi-dimensional case. Specifically, for the configurational CV, there are K_1 states and the state is numbered by k_1, while for the second CV (i.e., the alchemical CV), there are K_2 states and the state is numbered by k_2. The relative free energy of the state $\left(k_1, k_2 \right)$ on the two-dimensional free energy surface could thus be expressed as,

$$A_{k_1 k_2} = \Delta A_{k_1 1,11} + \Delta A_{k_1 k_2, k_1 1} = \sum_{i=1}^{k_1 - 1} \Delta A_{i,1,i+1,1} + \sum_{j=1}^{k_2 - 1} \Delta A_{k_1, j, k_1, j+1} \quad (10).$$

Here, $\Delta A_{k_1 1,11}$ denotes the free energy difference between the reference configurational state 1 and the configurational state k_1 at the 1st alchemical state, and $\Delta A_{k_1 k_2, k_1 1}$ is the correction term accounting for the free energy difference for the k_1 th configurational state at the 1st and k_2 th alchemical Hamiltonians. The statistical error could be similarly obtained with the error propagation procedure, as done in Eq. (7).

Our target is obtaining the free energy landscape at the alchemical state $k_2 = K_2$. The perturbation along the alchemical pathway is relatively small compared with the configurational case, and the free energy profile along the alchemical pathway is artificial and could be varied by defining different mixing functions for intermediate states. Therefore, we do not employ the stratification strategy but use only a single-step bidirectional pulling in the alchemical transformation. In this case, the nonequilibrium technique is especially useful, as equilibrium sampling in the intermediate states is avoided.[53]

In the current work, our illustrative calculation involves changes in both the QM theory and the size of the QM region. Instead of the previous test system ACE-NME (NMA),[15, 69] this time the test case is chosen as a bigger biologically relevant system shown in Fig. 1b. The conformational change in the solvated alanine tripeptide is described with the backbone C-C-N-C dihedral. Multi-scale QM/MM treatment is employed to describe the system. Due to efficiency consideration, our test Hamiltonians includes SQM Hamiltonians of Parametrized Model number 3 (PM3),[102] Parametrized Model number 6 (PM6),[103] and Recife Model 1 (RM1).[104] The multi-scale treatment could include different numbers of residues in the QM region. The small QM region includes the central two residues involved in the definition of the flipping backbone dihedral, and the large QM region includes the central two residues and the 1st residue of the chain (i.e., the ACE cap at the N-terminal of the peptide). In the case of expanding the QM region, $k_2 = 1$ defines the PM3 or PM6 Hamiltonian with a smaller QM region, and $k_2 = 2$ represents the PM6 Hamiltonian with a larger QM region. Namely, the free energy profile along the conformational change is constructed under the PM3 or PM6 Hamiltonian with only the central two residues included in the QM region, and correction terms are calculated between the $(k_1, 1)$ and $(k_1, 2)$ states to obtain the free energy profile under the PM6 Hamiltonian with the three residues included in the QM region. We also considered another case that the QM region is shrunk, which could also be used to check whether the QM results are invariant of the size of the QM region. In this case, $k_2 = 1$ defines the PM3 or PM6 Hamiltonian with a larger QM region, and $k_2 = 2$ represents the PM6 Hamiltonian with a smaller QM region.

In Fig. 1c, we provide a framework altered specifically for the variation of the QM region with fixed QM theory, which could be used when extrapolating the results with different definitions of the QM region. When the QM region 1 is a small QM region and the QM region 2 is defined as a larger QM region, the multi-dimensional free energy simulation is used to expand the size of the QM region. Note that the QM region 2 does not necessarily include all atoms in the QM region 1, and the extrapolation could thus be performed in a more general way. For instance, the QM region 1 could include the 1-3 residues of the alanine tripeptide, while the QM region 2 could be defined as the 2-4 residues of the peptide. The Hamiltonian variation framework could be similarly altered for other types of perturbations, e.g., the basis-set change.

A final part making the multi-dimensional nonequilibrium free energy simulation framework useful for practical applications is employing some curve fitting procedures to minimize the statistical noise introduced

by the correction term. The Savitzky-Golay filter has been used in our previous works and satisfactory results were obtained.[15, 69] Therefore, we recommend using this filter to increase the signal-to-noise ratio and get a smoother PMF, although other curve fitting methods are also usable.

Convergence check is indispensable in free energy simulation. It provides evidence that the simulation results are reliable and reproducible, and thus should be performed in any free energy simulation studies. In nonequilibrium pulling simulations, the ultimate convergence check procedure with the highest reliability is monitoring the sample-size and pulling-speed dependence of various ensemble averages.[42, 45, 105] As our target observable is the relative free energy, we check the pulling-speed and sample-size evolutions of the free energy profiles. As the statistical uncertainty of the free energy is often more biased than the free energy itself, checking the sample-size dependence of the statistical error is also useful.[16, 42, 45, 105] The free energy simulation is considered to be converged when the free energy profile neither changes with further sampling nor changes with slower pulling speeds. According to Eq. (3) or Eq. (7), if the statistical quantity $\frac{\mathrm{Var}\left(f_{i,i+1}\right)}{f_{i,i+1}^2}$ is estimated in an unbiased way, the state-specified variance (or its squared root standard error) $\frac{\mathrm{Var}\left(f_{i,i+1}\right)}{n_i^2 f_{i,i+1}^2}+\frac{\mathrm{Var}\left(f_{i,i-1}\right)}{n_i^2 f_{i,i-1}^2}$ should decrease monotonically with further sampling. Whether this behavior is satisfied provides another hint on the convergence behavior of the free energy simulation. Note that checking the sample-size dependence of the overall variance in the conformational change in Eq. (7) could also be useful, but checking the state-specified statistical quantity for all states by constructing a standard deviation (SD) profile would be more intuitive and useful for diagnosing different behaviors of different configurational states.[15, 36, 42, 69]

III. Computational Details

System preparation. The test system in the current work is the alanine tripeptide with the sequence of ACE-ALA-ALA-NME, which is described with the AMBER14SB[106] force field. The structural observable describing the conformational change is the central periodic backbone dihedral defined by C-C-N-C, which is depicted in Fig. 1b. As the focus of this work is introducing the variation of the QM region into the Hamiltonian-variation framework rather than testing some QM Hamiltonians, we choose the computationally efficient SQM Hamiltonians of PM3[102] and PM6[103] as the QM potential. The system is solvated in TIP3P water molecules[107, 108] and the periodic boundary condition is employed for the simulation box. We tested 2 sizes of QM regions. For the smaller one, the central two residues are included in the QM

region, while the 1st residue is also included in the QM region for the larger one. As the nonequilibrium pulling trajectories are essentially independent and each segment becomes very short (< 1 ps) with the staging technique, each simulation is performed with a single core independently, which avoids the parallelization-related issues degrading the performance of the simulation. In Table S1, we present the single-core timing data for QM/MM simulations with different sizes of the QM region. The simulation speed becomes slower when the QM region is expanded. In the current SQM case, as shown in Table S1, the computational costs of different SQM Hamiltonians with the same definition of the QM region are identical and the increase of the computational cost when the QM region is expanded is modest, the PM3<->PM6 indirect simulation does not lead to computational speedup. However, the test scheme is sufficient to illustrate the applicability of the Hamiltonian-perturbation/variation framework. If the QM theory under consideration includes ab initio QM Hamiltonians, e.g. obtaining the B3LYP results from indirect PM6 free energy simulations, the speedup would be dramatic.[15]

Another illustrative example is the gas-phase alanine tripeptide built with the same force field. The configurational CV is still the backbone C-C-N-C dihedral. Here, we consider the variation of the QM region and/or the change of the QM theory. Two cases are considered for the QM-region variation in the gas-phase simulations. The first one follows the thermodynamic cycle depicted in Fig. S1a, where the smaller QM region includes the 2nd, 3rd and 4th residues of the peptide, and the larger QM region includes the whole alanine tripeptide. The second case is shown in Fig. S1b, where the results with the 1st, 2nd, and 3rd residues included in the QM region are perturbed to those with the whole peptide included in the QM region. The target Hamiltonians in the gas-phase simulations are the RM1 or PM6 Hamiltonian with the whole alanine tripeptide in the QM region. We present the single-core timing information for the gas-phase simulations in Table S1. Due to the exclusion of water molecules in the gas-phase system, the difference between the computational costs with different definitions of the QM region is more dramatic.

Free Energy Simulation. We then consider the sampling in the configurational and alchemical spaces. The dihedral windows are equally spaced from 0° to 360 ° with 2° increments. The force constant applied to achieve the stiff spring limit is 2000 kcal/mol·rad^2. We performed configurational sampling in each dihedral window to obtain independent configurations to initiate nonequilibrium pulling simulations. In each intermediate, minimization, NVT heating to 300 K and 100 ps NPT equilibration are performed under the MM Hamiltonian. We then shift to the multi-scale treatment and equilibrate the system for another 150 ps. After sufficient equilibration, the configurations are extracted with the sampling interval of 0.5 ps, which is sufficiently long to decorrelate the successive samples according to autocorrelation analysis. Bidirectional

nonequilibrium pulling with three pulling speeds including 0.25 ps per 2° segment, 0.5 ps/segment and 0.8 ps/segment are initiated from uncorrelated configurations. The initial configurations used for the nonequilibrium pulling in the alchemical space is obtained from the above initial configurational sampling procedure along the configurational CV. The magnitude for each perturbation of the alchemical control parameter is set to $\Delta\lambda = 0.1$ and the relaxation time between successive perturbations is 1 time step, which is tested to be good enough in the QM-theory-variation case in our previous work.[15] The free energy simulation schemes for the alanine tripeptide in solution and that in vacuo are the same.

Due to the large force constant used to achieve the stiff spring limit, we employ the time step 0.5 fs in the initial configurational sampling and nonequilibrium pulling simulations.[19, 36, 37, 69] Langevin dynamics[109] with the collision frequency of 5 ps^{-1} are implemented for temperature regulation, and in NPT simulations we use isotropic position scaling and the Berendsen barostat to regulate the pressure. For the solvated peptide, the cutoff for non-bonded interactions in the real space is 8 Å and the long-range electrostatics are treated with the PME method.[110] No cutoff is applied for the simulations in vacuo. We use the AMBER[111] suite for MD simulation. All statistical analyses are obtained with homemade codes.

IV. Result and discussion

1. Solvated Peptide.

Convergence behavior of free energy simulations.

To compare the free energy results from direct and indirect free energy simulations, we need to check the convergence behavior of these free energy estimates. Here, we consider the case shown in Fig. 1b as an example. The aim is to obtain the QM/MM results under the PM6/MM Hamiltonian with the 1st, 2nd and 3rd residues included in the QM region. Direct free energy simulations are performed with three Hamiltonians. The first one is the PM3/MM Hamiltonian with the central two residues included in the QM region. The second one uses the PM6/MM Hamiltonian and the central two residues are included in the QM region. The last and targeted Hamiltonian employs the PM6/MM Hamiltonian with the 1-3 residues included in the QM region. In Fig. 2a-c, we presented the sample-size dependence of the free energy profiles along the configurational CV (i.e., the flipping dihedral) with three pulling speeds mentioned in the computational detail section. The pulling speed of 0.5 ps/segment seems sufficiently slow to obtain converged estimates from bidirectional nonequilibrium pulling simulations, which is similar to the previous backbone dihedral case used in our previous works.[15, 69] Thus, we use the results obtained with this pulling speed in the later comparison. As for the sample-size dependence, the initial sample size is 5 and 5 additional samples are

added to the dataset in each iteration. The statistical fluctuations of the free energy profiles after 10 iterations are very small, which indicates that the convergence is reached. Another statistical quantity used for convergence check is the statistical error shown in Fig. 2d-f, where the non-linear and monotonically decreasing behavior is generally observed. As the magnitude of the decrease of the state-specified SD is small when the sample size is large, the time-evolution of a more sample-size sensitive statistical quantity of the time derivative of the overall variance (TDV) is also shown in Fig. S2. As the TDV is scaled by the simulation time, it is more sensitive to the increase of the sample size (i.e., the computational cost). The monotonically decreasing behavior of the state-specified TDV also indicates the convergence of our free energy simulations. For the correction term accounting for the difference between the thermodynamic profiles under different Hamiltonians, we used 40 samples and the convergence behavior is good. Note that in our previous SQM-to-QM (e.g., PM6-to-B3LYP) simulations, 20 samples are already sufficient for well-converged estimates in bidirectional reweighting.

Indirect vs direct.

The free energy results along the configurational CV are combined with those along the alchemical CV. After the smoothing step with the Savitzky-Golay filter, the indirect and direct free energy results are compared. In Fig. 5a, the free energy profile under the target PM6/MM Hamiltonian with 3 residues in the QM region obtained from the direct free energy simulation under that Hamiltonian is compared with the two indirect estimates. The first indirect free energy simulation scheme uses the same QM Hamiltonian of PM6, but only includes the central two residues in the QM region. The alchemical correction term is used to expand the QM region from the central two residues to the 1-3 residues, which could be seen as a test case of the QM-region-variation framework shown in Fig. 1c. The second indirect free energy simulation scheme uses a different QM Hamiltonian of PM3, and includes the central two residues in the QM region. The alchemical correction term accounts for both the change of the QM theory from PM3 to PM6 and the variation of the QM region from two residues to three residues. We can see that the free energy profiles obtained from the direct and indirect schemes agree well. The free energy profiles obtained from the indirect method are as smooth as the direct result, which is better than the equilibrium FEP extrapolation presented in the reference.[95] The current case illustrates that the multi-dimensional nonequilibrium pulling framework could be used in various Hamiltonian-perturbation cases. The QM region could be expanded to check whether the QM region is sufficiently large to ensure the convergence of the free energy results on this degree of freedom. The QM theory could be changed to compare the difference between the descriptions of the same thermodynamic variable at different QM levels.

Another interesting case to check is the variation of the free energy results when the QM region is shrunk. We can indirectly obtain the free energy estimates with a small QM region by reweighting the results with a large QM region. If the large-region and small-region estimates are the same, the small QM region could be sufficiently large to converge the estimates. We still use the Fig. 1b case but with different simulation schemes. The target Hamiltonian is now the PM6/MM Hamiltonian with the central two residues in the QM region. The first indirect scheme uses the PM6/MM Hamiltonian with the 1-3 residues in the QM region, and the alchemical term is used to shrink the QM region from three residues to the central two residues. The second indirect scheme employs the PM3/MM Hamiltonian with the 1-3 residues in the QM region, and the alchemical term is used to account for both the PM3-to-PM6 difference and the variation of the size of the QM region. Again, the direct and indirect estimates are similar, indicating the applicability of the multi-dimensional nonequilibrium Hamiltonian-perturbation/variation framework.

2. Gas-phase Simulations.

The above two test cases show that the proposed framework works well for perturbing the Hamiltonian in the multi-scale treatment for condensed-phase systems. As the last example, we simulate the same system (i.e., the alanine tripeptide) in vacuo, which decouples the influence of the solvent from the thermodynamic behavior of the solute under investigation. The illustrations of the gas-phase simulations for the variations of the QM region and/or the QM level are shown in Fig. S1. The target QM region includes the whole alanine tripeptide, and the simulations are performed under two SQM Hamiltonians (i.e., RM1 and PM6). The sample-size dependences of the direct results including the free energy profile, the uncertainty profile, and the TDV profile under the target Hamiltonians are shown in Fig. S3a-c, respectively. In the first case, we consider the PM6 Hamiltonian with the whole alanine tripeptide in the QM region as the target Hamiltonian. There are two schemes used in the indirect simulations. The first one uses the same QM level (i.e., PM6) and a smaller QM region including the 2^{nd}, 3^{rd}, and 4^{th} residues in the QM region, while the second one is performed at the RM1 level with the 1^{st}, 2^{nd}, and 3^{rd} residues included in the QM region. The first indirect scheme is used to expand the QM region with fixed QM theory, while the second indirect regime expands the QM region in a different way and changes the QM level simultaneously. In this gas-phase case, the direct and indirect estimates again agree well, as shown in Fig. 4a. In the second case, the RM1 Hamiltonian with the whole peptide included in the QM region is the target Hamiltonian, and the indirect free energy simulations follow the same procedure as the first case. In this RM1 case, the accord between the direct and indirect estimates is still very good, as shown in Fig. 4b. Therefore, our nonequilibrium Hamiltonian framework works very well in gas-phase simulations.

Finally, it is worth noting that the above BAR-based perturbation network could be altered to some unidirectional EXP-based one, which would lead to some speedups in some cases. Namely, the nonequilibrium pulling in the alchemical transformation and/or the configurational sampling could be performed unidirectionally, and unidirectional estimators such as the EXP estimator and its Gaussian approximation could be used to estimate the free energy difference. In a following work,[112] we presented such an altered scheme, and uses the selection criterion for initial configurations to further accelerate the computation. The advanced unidirectional pulling framework and the current BAR-based scheme share the same application area but have different advantages. In practical applications, we could choose one according to the computational resources available and the specific features of the simulated system.

IV. Conclusion

Indirect QM/MM free energy simulations provide an alternative to direct free energy simulations at a targeted QM/MM level. The indirect method features the combination of the direct free energy simulation in the configurational space and the free energy correction in the alchemical space. By constructing a thermodynamic cycle, the results at ab initio QM levels such as B3LYP could be obtained indirectly by combining the results at the SQM level and an SQM-to-QM correction. Similarly, the SQM estimates could be obtained by combining the MM results and the MM-to-SQM correction. Further, the indirect scheme could be used to account for the change of the QM region, the basis set and so on. The current work generalizes our previously proposed multi-dimensional nonequilibrium pulling framework for indirect free energy simulation to a general Hamiltonian-variation framework. The nonequilibrium method constructs the thermodynamic profile along the configurational CV with staged bidirectional nonequilibrium pulling and reweighting, and estimates the alchemical correction by performing nonequilibrium pulling in the alchemical space. The multi-dimensional nonequilibrium Hamiltonian-perturbation/variation framework could be used to change the size of the QM region, the level of theory (i.e., the QM Hamiltonian), the basis set and so on. Note that the basis-set extrapolation is quite interesting, as it provides a computationally feasible way to perform the complete basis-set extrapolation of thermodynamic observables, which would be addressed in our following works. Although the single-step perturbation based on equilibrium sampling could also be employed in this case, our nonequilibrium framework provides a fail-safe scheme for absolute convergence.

The nonequilibrium framework has several satisfactory features. The nonequilibrium nature of the perturbation network ensures a good convergence behavior and minor modifications of the theoretical

framework in practical applications. The statistical efficiency and the generality of the method are maximized by using the statistically optimal bidirectional estimator without parametric approximation. Note that the bidirectional estimator also avoids the underestimation of the statistical error to some extent. As the nonequilibrium pulling simulations are independent and the pulling time in each segment is made short by stratification, the efficiency of parallelism is maximized by distributing each nonequilibrium pulling on a single core. This also leads to fast user feedback and enables quick protocol diagnosing, which is very desirable in practical applications.

We chose a larger biologically relevant system for numerical tests in this work. The flipping of the backbone dihedral of alanine tripeptide was simulated under the PM3/MM, PM6/MM or RM1/MM Hamiltonian with different definitions of the QM region. We employed the nonequilibrium Hamiltonian-variation framework to change the QM theory and vary the QM region at the same time. The thermodynamic profile along the flipping pathway obtained from the direct free energy simulation is well reproduced by the indirect method in both expanding and shrinking the QM region. The numerical results of combining the variation of the QM theory with the variation of the QM region are also very good. These numerical validations for the solvated and the gas-phase systems suggest the framework is generally applicable to various physical, chemical and biological systems.

Supporting information

The single-core timing information of the QM/MM simulations of the solvated alanine tripeptide with different sizes of the QM region, the thermodynamic cycle of the indirect nonequilibrium free energy simulations for the alanine tripeptide in vacuo, the time-evolution of the TDV profile for the solvated alanine tripeptide, the sample-size dependences of the free energy profile, the uncertainty profile, and the TDV profile for the gas-phase simulations are provided in the supporting information.

Conflicts of interest

There are no conflicts of interest to declare.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (Grant No. 21633001). Part of the simulation was performed on the high-performance computing platform of the Center for Life Science (Peking University). Dr. Zhaoxi Sun is supported by the PKU-Boya Postdoctoral Fellowship. We

thank anonymous reviewers for valuable comments and critical reading.

Preprint Acknowledgement

Research presented in this article has been posted on a preprint server prior to publication. The corresponding preprint article can be found here: https://doi.org/10.26434/chemrxiv.13634981.

References

1. Echeverria, I.; Amzel, L. M., Helix propensities calculations for amino acids in alanine based peptides using Jarzynski's equality. *Proteins: Structure, Function, and Bioinformatics* **2010**, 78, 1302-1310.

2. Lee, T. S.; Radak, B. K.; Huang, M.; Wong, K. Y.; York, D. M., Roadmaps through free energy landscapes calculated using the multi-dimensional vFEP approach. *J. Chem. Theory Comput.* **2014**, 10, 24-34.

3. Sun, Z.; Wang, X.; Zhang, J. Z. H., Protonation-dependent Base Flipping in The Catalytic Triad of A Small RNA. *Chemical Physics Letters* **2017**, 684, 239-244.

4. Moraca, F.; Amato, J.; Ortuso, F.; Artese, A.; Pagano, B.; Novellino, E.; Alcaro, S.; Parrinello, M.; Limongelli, V., Ligand binding to telomeric G-quadruplex DNA investigated by funnel-metadynamics simulations. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, 114, E2136-E2145.

5. Wood, R. H.; Muhlbauer, W. C. F.; Thompson, P. T., Systematic errors in free energy perturbation calculations due to a finite sample of configuration space: sample-size hysteresis. *Journal of Physical Chemistry* **1991**, 95, 6670-6675.

6. Gore, J.; Ritort, F.; Bustamante, C., Bias and error in estimates of equilibrium free-energy differences from nonequilibrium measurements. *Proc. Natl. Acad. Sci. USA* **2003**, 100, 12564-12569.

7. Zuckerman, D. M.; Woolf, T. B., Theory of a systematic computational error in free energy differences. *Physical Review Letters* **2002**, 89, 180602.

8. Mezei, M., Adaptive Umbrella Sampling: Self-consistent Determination of the Non-Boltzmann Bias. *J. Comput. Phys.* **1987**, 68, 237-248.

9. Hooft, R. W.; van Eijck, B. P.; Kroon, J., An Adaptive Umbrella Sampling Procedure in Conformational Analysis using Molecular Dynamics and Its Application to Glycol. *J. Chem. Phys.* **1992**, 97, 6690-6694.

10. Kästner, J., Umbrella sampling. *Wiley Interdisip. Rev. Comput. Mol. Sci.* **2011**, 1, 932-942.

11. Hub, J. S.; Groot, B. L. D.; Spoel, D. V. D., g_wham—A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates. *J. Chem. Phys.* **2015**, 6, 3713-3720.

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

13. Lee, T. S.; Radak, B. K.; Pabis, A.; York, D. M., A New Maximum Likelihood Approach for Free Energy Profile Construction from Molecular Simulations. *J. Chem. Theory Comput.* **2013**, 9, 153-164.

14. Zwanzig, R. W., High Temperature Equation of State by A Perturbation Method. *J. Chem. Phys.* **1954**, 22, 1420-1426.

15. Sun, Z., BAR-based multi-dimensional nonequilibrium pulling for indirect construction of QM/MM free energy landscapes: from semi-empirical to ab initio. *Phys. Chem. Chem. Phys.* **2019**, 21, 21942-21959

16. Wang, X.; Sun, Z., A Theoretical Interpretation of Variance-based Convergence Citeria in Perturbation-based Theories. *arXiv preprint arXiv:1803.03123* **2018**.

17. Shirts, M. R.; Bair, E.; Hooker, G.; Pande, V. S., Equilibrium Free Energies from Nonequilibrium Measurements using Maximum-likelihood Methods. *Physical review letters* **2003**, 91, 140601.

18. Bennett, C. H., Efficient Estimation of Free Energy Differences from Monte Carlo data. *J. Comput. Phys.* **1976**, 22, 245-268.

19. Sun, Z.; Wang, X.; Zhang, J. Z. H.; He, Q., Sulfur-substitution-induced base flipping in the DNA duplex. *Phys. Chem. Chem. Phys.* **2019**, 21, 14923-14940.

20. Sun, Z.; Zhang, J. Z. H., Thermodynamic Insights of Base Flipping in TNA Duplex: Force Fields, Salt Concentrations, and Free-Energy Simulation Methods. *CCS Chemistry* **2020**, 2, 1026–1039.

21. Lemkul, J. A.; Savelyev, A.; MacKerell Jr, A. D., Induced Polarization Influences The Fundamental Forces in DNA Base Flipping. *J. Phys. Chem. Lett.* **2014**, 5, 2077-2083.

22. Jarzynski, C., A Nonequilibrium Equality for Free Energy Differences. *Physical Review Letters* **1997**, 78, 2690-2693.

23. Mallick, K.; Moshe, M.; Orland, H., Supersymmetry and Nonequilibrium Work Relations. *arXiv preprint arXiv:0711.2059* **2008**.

24. Fass, J.; Sivak, D. A.; Crooks, G. E.; Beauchamp, K. A.; Leimkuhler, B.; Chodera, J. D., Quantifying configuration-sampling error in Langevin simulations of complex molecular systems. *Entropy* **2018**, 20, 318.

25. Sivak, D. A.; Chodera, J. D.; Crooks, G. E., Time Step Rescaling Recovers Continuous-Time Dynamical Properties for

Discrete-Time Langevin Integration of Nonequilibrium Systems. *J. Phys. Chem. B* **2014**, 118, 6466-6474.

26. Sivak, D. A.; Chodera, J. D.; Crooks, G. E., Using Nonequilibrium Fluctuation Theorems to Understand and Correct Errors in Equilibrium and Nonequilibrium Simulations of Discrete Langevin Dynamics. *Phys. Rev. X* **2013**, 3, 011007.

27. Chelli, R.; Procacci, P., A potential of mean force estimator based on nonequilibrium work exponential averages. *Phys. Chem. Chem. Phys.* **2009**, 11, 1152-1158.

28. Nicolini, P.; Frezzato, D.; Chelli, R., Exploiting Configurational Freezing in Nonequilibrium Monte Carlo Simulations. *J. Chem. Theory Comput.* **2011**, 7, 582-593.

29. Chelli, R.; Marsili, S.; Barducci, A.; Procacci, P., Generalization of the Jarzynski and Crooks nonequilibrium work theorems in molecular dynamics simulations. *Phys. Rev. E* **2007**, 75, 050101.

30. Chelli, R.; Marsili, S.; Barducci, A.; Procacci, P., Recovering the Crooks equation for dynamical systems in the isothermal-isobaric ensemble: a strategy based on the equations of motion. *J. Chem. Phys.* **2007**, 126, 044502.

31. Ballard, A. J.; Jarzynski, C., Replica exchange with nonequilibrium switches: enhancing equilibrium sampling by increasing replica overlap. *J. Chem. Phys.* **2012**, 136, 194101.

32. Vaikuntanathan, S.; Jarzynski, C., Escorted free energy simulations: improving convergence by reducing dissipation. *Physical Review Letters* **2008**, 100, 190601.

33. Dickson, A.; Dinner, A. R., Enhanced Sampling of Nonequilibrium Steady States. *Annual Review of Physical Chemistry* **2010**, 61, 441-459.

34. Hudson, P. S.; Woodcock, H. L.; Boresch, S., Use of Nonequilibrium Work Methods to Compute Free Energy Differences Between Molecular Mechanical and Quantum Mechanical Representations of Molecular Systems. *J. Phys. Chem. Lett.* **2015**, 6, 4850-4856.

35. Procacci, P.; Marsili, S., Energy dissipation asymmetry in the non equilibrium folding/unfolding of the single molecule alanine decapeptide. *Chemical Physics* **2010**, 375, 8-15.

36. Wang, X.; Xingzhao, T.; Boming, D.; John Z. H., Z.; Sun, Z., BAR-based Optimum Adaptive Steered MD for Configurational Sampling. *J. Comput. Chem.* **2019**, 40, 1270-1289.

37. Wang, X.; Sun, Z., Determination of Base Flipping Free Energy Landscapes from Nonequilibrium Stratification. *J. Chem. Inf. Model.* **2019**, 59, 2980-2994.

38. Procacci, P., Methodological uncertainties in drug-receptor binding free energy predictions based on classical molecular dynamics. *Curr. Opin. Struct. Biol.* **2021**, 67, 127-134.

39. Swope, W. C., 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.

40. Pham, T. T.; Shirts, M. R., Identifying low variance pathways for free energy calculations of molecular transformations in solution phase. *J. Chem. Phys.* **2011**, 135, 034114.

41. Procacci, P.; Chelli, R., Statistical Mechanics of Ligand–Receptor Noncovalent Association, Revisited: Binding Site and Standard State Volumes in Modern Alchemical Theories. *J. Chem. Theory Comput.* **2017**, 13, 1924-1933.

42. Wang, X.; Tu, X.; Zhang, J. Z. H.; Sun, Z., BAR-based Optimum Adaptive Sampling Regime for Variance Minimization in Alchemical Transformation: The Nonequilibrium Stratification. *Phys. Chem. Chem. Phys.* **2018**, 20, 2009-2021.

43. Shirts, M. R.; Pande, V. S., Solvation free energies of amino acid side chain analogs for common molecular mechanics water models. *J. Chem. Phys.* **2005**, 122, 134508.

44. Hummer, G.; Pratt, L. R.; Garcia, A. E., Hydration free energy of water. *Journal of Physical Chemistry* **1995**, 99, 14188-14194.

45. Sun, Z. X.; Wang, X. H.; Zhang, J. Z. H., BAR-based Optimum Adaptive Sampling Regime for Variance Minimization in Alchemical Transformation. *Phys. Chem. Chem. Phys.* **2017**, 19, 15005-15020.

46. Huai, Z.; Sun, Z., Titration of Adenine in a GA mismatch with Grand Canonical Simulations. *Journal of Computational Biophysics and Chemistry* **2020**, 1-9.

47. Bruckner, S.; Boresch, S., Efficiency of alchemical free energy simulations. II. Improvements for thermodynamic integration. *J. Comput. Chem.* **2011**, 32, 1320–1333.

48. Resat, H.; Mezei, M., Studies on free energy calculations. I. Thermodynamic integration using a polynomial path. *J. Chem. Phys.* **1993**, 99, 6052-6061.

49. Resat, H.; Mezei, M., Studies on free energy calculations. II. A theoretical approach to molecular solvation. *J. Chem. Phys.* **1994**, 101, 6126-6140.
50. Paliwal, H.; Shirts, M. R., A Benchmark Test Set for Alchemical Free Energy Transformations and Its Use to Quantify Error in Common Free Energy Methods. *J. Chem. Theory Comput.* **2011**, 7, 4115-34.
51. Fenwick, M. K.; Escobedo, F. A., On the use of Bennett's acceptance ratio method in multi-canonical-type simulations. *J. Chem. Phys.* **2004**, 120, 3066-74.
52. Sun, Z.; Wang, X.; Zhang, J. Z., Determination of Binding Affinities of 3-Hydroxy-3-Methylglutaryl Coenzyme A Reductase Inhibitors from Free Energy calculation. *Chemical Physics Letters* **2019**, 723, 1-10.
53. Huai, Z.; Yang, H.; Li, X.; Sun, Z., SAMPL7 TrimerTrip host-guest binding affinities from extensive alchemical and end-point free energy calculations. *Journal of Computer-Aided Molecular Design* **2021**, 35, 117-129.
54. Sun, Z.; Wang, X.; Zhang, J. Z., Theoretical understanding of the thermodynamics and interactions in transcriptional regulator TtgR-ligand binding. *Phys. Chem. Chem. Phys.* **2020**, 22, 1511-1524.
55. Sun, Z.; Wang, X.; Zhao, Q.; Zhu, T., Understanding Aldose Reductase-Inhibitors interactions with free energy simulation. *Journal of Molecular Graphics and Modelling* **2019**, 91, 10-21.
56. Wang, X.; Sun, Z., Understanding PIM-1 kinase inhibitor interactions with free energy simulation. *Phys. Chem. Chem. Phys.* **2019**, 21, 7544-7558.
57. Huai, Z.; Shen, Z.; Sun, Z., Binding Thermodynamics and Interaction Patterns of Inhibitor-Major Urinary Protein-I Binding from Extensive Free-Energy Calculations: Benchmarking AMBER Force Fields. *J. Chem. Inf. Model.* **2021**, 61, 284-297.
58. Heimdal, J.; Rydberg, P.; Ryde, U., Protonation of the proximal histidine ligand in heme peroxidases. *J. Phys. Chem. B* **2008**, 112, 2501-10.
59. Mikulskis, P.; Cioloboc, D.; Andrejić, M.; Khare, S.; Brorsson, J.; Genheden, S.; Mata, R. A.; Söderhjelm, P.; Ryde, U., Free-energy perturbation and quantum mechanical study of SAMPL4 octa-acid host-guest binding energies. *Journal of computer-aided molecular design* **2014**, 28, 375-400.
60. Fox, S. J.; Pittock, C.; Tautermann, C. S.; Fox, T.; Christ, C.; Malcolm, N. O.; Essex, J. W.; Skylaris, C. K., Free energies of binding from large-scale first-principles quantum mechanical calculations: application to ligand hydration energies. *J. Phys. Chem. B* **2013**, 117, 9478-85.
61. Genheden, S.; Ryde, U.; Söderhjelm, P., Binding affinities by alchemical perturbation using QM/MM with a large QM system and polarizable MM model. *J. Comput. Chem.* **2015**, 2114-2124.
62. Genheden, S.; Martinez, A. I. C.; Criddle, M. P.; Essex, J. W., Extensive all-atom Monte Carlo sampling and QM/MM corrections in the SAMPL4 hydration free energy challenge. *Journal of Computer-Aided Molecular Design* **2014**, 28, 187-200.
63. Fox, S. J.; Pittock, C.; Tautermann, C. S.; Fox, T.; Christ, C.; Malcolm, N. O. J.; Essex, J. W.; Skylaris, C. K., Free Energies of Binding from Large-Scale First-Principles Quantum Mechanical Calculations: Application to Ligand Hydration Energies. *J. Phys. Chem. B* **2013**, 117, 9478-85.
64. Woods, C. J.; Manby, F. R.; Mulholland, A. J., An efficient method for the calculation of quantum mechanics/molecular mechanics free energies. *J. Chem. Phys.* **2008**, 128, 152-159.
65. Caveayland, C.; Skylaris, C. K.; Essex, J. W., Direct Validation of the Single Step Classical to Quantum Free Energy Perturbation. *J. Phys. Chem. B* **2014**, 119, 1017-25.
66. Rod, T. H.; Ryde, U., Quantum mechanical free energy barrier for an enzymatic reaction. *Phys. Rev. Lett.* **2005**, 94.
67. Lameira, J. S.; Kupchencko, I.; Warshel, A., Enhancing Paradynamics for QM/MM Sampling of Enzymatic Reactions. *J. Phys. Chem. B* **2016**, 120, 2155.
68. Klimovich, P. V.; Shirts, M. R.; Mobley, D. L., Guidelines for the Analysis of Free Energy Calculations. *Journal of Computer-Aided Molecular Design* **2015**, 29, 397-411.
69. Wang, X.; He, Q.; Sun, Z., BAR-Based Multi-Dimensional Nonequilibrium Pulling for Indirect Construction of a QM/MM Free Energy Landscape. *Phys. Chem. Chem. Phys.* **2019**, 21, 6672-6688
70. Sun, Z., SAMPL7 TrimerTrip Host-Guest Binding Poses and Binding Affinities from Spherical-Coordinates-Biased Simulations. *Journal of Computer-Aided Molecular Design* **2021**, 35, 105-115.
71. Procacci, P., Precision and computational efficiency of nonequilibrium alchemical methods for computing free energies of

solvation. II. Unidirectional estimates. *J. Chem. Phys.* **2019**, 151, 144115.

72. Procacci, P.; Guarnieri, G., SAMPL7 blind predictions using nonequilibrium alchemical approaches. *Journal of Computer-Aided Molecular Design* **2021**.

73. Nerattini, F.; Chelli, R.; Procacci, P., II. Dissociation free energies in drug–receptor systems via nonequilibrium alchemical simulations: application to the FK506-related immunophilin ligands. *Phys. Chem. Chem. Phys.* **2016**, 18, 15005-15018.

74. Sun, Z.; He, Q.; Li, X.; Zhu, Z., SAMPL6 host–guest binding affinities and binding poses from spherical-coordinates-biased simulations. *Journal of Computer-Aided Molecular Design* **2020**, 34, 589-600.

75. Gao, J.; Luque, F. J.; Orozco, M., Induced dipole moment and atomic charges based on average electrostatic potentials in aqueous solution. *J. Chem. Phys.* **1993**, 98, 2975-2982.

76. Luzhkov, V.; Warshel, A., Microscopic models for quantum mechanical calculations of chemical processes in solutions: LD/AMPAC and SCAAS/AMPAC calculations of solvation energies. *J. Comput. Chem.* **1992**, 13, 199–213.

77. Wesolowski, T.; Warshel, A., Ab Initio Free Energy Perturbation Calculations of Solvation Free Energy Using the Frozen Density Functional Approach. *Journal of Physical Chemistry* **1994**, 98, 5183-5187.

78. Gao, J.; Xia, X., A priori evaluation of aqueous polarization effects through Monte Carlo QM-MM simulations. *Science* **1992**, 258, 631-5.

79. Zheng, Y. J.; Merz, K. M., Mechanism of the human carbonic anhydrase II-catalyzed hydration of carbon dioxide. *Journal of the American Chemical Society* **1992**, 114, 10498-10507.

80. Plotnikov, N. V.; Warshel, A., Exploring, refining, and validating the paradynamics QM/MM sampling. *J. Phys. Chem. B* **2012**, 116, 10342-10356.

81. Bentzien, J.; Muller, R. P.; Florián, J.; Warshel, A., Hybrid ab initio quantum mechanics/molecular mechanics calculations of free energy surfaces for enzymatic reactions: the nucleophilic attack in subtilisin. *J. Phys. Chem. B* **1998**, 102, 2293-2301.

82. Plotnikov, N.; Kamerlin, S. C. L.; Warshel, A., ParaDynamics: An Effective and Reliable Model for Ab Initio QM/MM Free Energy Calculations and Related Tasks. *J. Phys. Chem. B* **2011**, 115, 7950-62.

83. Polyak, I.; Benighaus, T.; Boulanger, E.; Thiel, W., Quantum mechanics/molecular mechanics dual Hamiltonian free energy perturbation. *J. Chem. Phys.* **2013**, 139, 578.

84. Sun, Z.; Zhu, T.; Wang, X.; Mei, Y.; Zhang, J. Z., Optimization of convergence criteria for fragmentation methods. *Chemical Physics Letters* **2017**, 687, 163-170.

85. Liu, W.; Sakane, S.; And, R. H. W.; Doren, D. J., The Hydration Free Energy of Aqueous Na+ and Cl- at High Temperatures Predicted by ab Initio/Classical Free Energy Perturbation: 973 K with 0.535 g/cm3 and 573 K with 0.725 g/cm3. *J. Phys. Chem. A* **2002**, 106, 1409-1418.

86. Olsson, M. A.; Söderhjelm, P.; Ryde, U., Converging ligand-binding free energies obtained with free-energy perturbations at the quantum mechanical level. *J. Comput. Chem.* **2016**, 37, 1589-1600.

87. Raghavachari, K.; Saha, A., Accurate composite and fragment-based quantum chemical models for large molecules. *Chemical reviews* **2015**, 115, 5643-5677.

88. Collins, M. A.; Bettens, R. P., Energy-based molecular fragmentation methods. *Chemical reviews* **2015**, 115, 5607-5642.

89. Sahu, N.; Gadre, S. R., Molecular tailoring approach: a route for ab initio treatment of large clusters. *Accounts of chemical research* **2014**, 47, 2739-2747.

90. Kitaura, K.; Ikeo, E.; Asada, T.; Nakano, T.; Uebayasi, M., Fragment molecular orbital method: an approximate computational method for large molecules. *Chemical Physics Letters* **1999**, 313, 701-706.

91. Heidari, M.; Cortes-Huerto, R.; Kremer, K.; Potestio, R., Concurrent coupling of realistic and ideal models of liquids and solids in Hamiltonian adaptive resolution simulations. *EUR PHYS J E* **2018**, 41, 64.

92. Kreis, K.; Tuckerman, M. E.; Donadio, D.; Kremer, K.; Potestio, R., From Classical to Quantum and Back: A Hamiltonian Scheme for Adaptive Multiresolution Classical/Path-Integral Simulations. *J. Chem. Theory Comput.* **2016**, 12, 3030-3039.

93. Jackson, N. E.; Webb, M. A.; de Pablo, J. J., Layered nested Markov chain Monte Carlo. *J. Chem. Phys.* **2018**, 149, 072326.

94. Chen, Y.; Kale, S.; Weare, J.; Dinner, A. R.; Roux, B., Multiple Time-Step Dual-Hamiltonian Hybrid Molecular Dynamics – Monte Carlo Canonical Propagation Algorithm. *J. Chem. Theory Comput.* **2016**, 12, 1449-1458.

95. Jia-Ning, W.; Wei, L.; Pengfei, L.; Yan, M.; Wenxin, H.; Jun, Z.; Xiaoliang, P.; Yihan, S.; Ye, M., *Accelerated Computation of Free*

Energy Profile at Ab Initio Quantum Mechanical/Molecular Mechanics Accuracy via a Semiempirical Reference-Potential. 4. Adaptive QM/MM. 2020.

96. Hummer, G.; Szabo, A., From the Cover: Free energy reconstruction from nonequilibrium single-molecule pulling experiments. *Proceedings of the National Academy of Science* **2001**, 98, 3658-3661.

97. Hummer, G.; Szabo, A., Free Energy Reconstruction from Nonequilibrium Single-molecule Pulling Experiments. *Proc. Natl. Acad. Sci. USA* **2001**, 98, 3658-3661.

98. Hummer, G.; Szabo, A., Free Energy Surfaces from Single-molecule Force Spectroscopy. *Cheminform* **2005**, 36, 504-513.

99. Paramore, S.; Ayton, G. S.; Voth, G. A., Extending the Fluctuation Theorem to Describe Reaction Coordinates. *J. Chem. Phys.* **2007**, 126, 051102.

100. Balsera; Stepaniants; Izrailev; Oono; Schulten, Reconstructing potential energy functions from simulated force-induced unbinding processes. *Biophysical Journal* **1997**, 73, 1281.

101. Marsili, S.; Procacci, P., Free energy reconstruction in bidirectional force spectroscopy experiments: The effect of the device stiffness. *J. Phys. Chem. B* **2010**, 114, 2509-2516.

102. Stewart, J. J., Optimization of parameters for semiempirical methods II. Applications. *J. Comput. Chem.* **1989**, 10, 221-264.

103. Stewart, J. J., Optimization of parameters for semiempirical methods V: modification of NDDO approximations and application to 70 elements. *Journal of Molecular modeling* **2007**, 13, 1173-1213.

104. Rocha, G. B.; Freire, R. O.; Simas, A. M.; Stewart, J. J. P., RM1: A reparameterization of AM1 for H, C, N, O, P, S, F, Cl, Br, and I. *J. Comput. Chem.* **2006**, 27, 1101-1111.

105. Sun, Z.; Wang, X.; Song, J., Extensive Assessment of Various Computational Methods for Aspartate's pKa Shift. *J. Chem. Inf. Model.* **2017**, 57, 1621-1639.

106. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, 11, 3696-3713.

107. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, 79, 926-935.

108. Price, D. J.; Brooks III, C. L., A Modified TIP3P Water Potential for Simulation with Ewald Summation. *J. Chem. Phys.* **2004**, 121, 10096-10103.

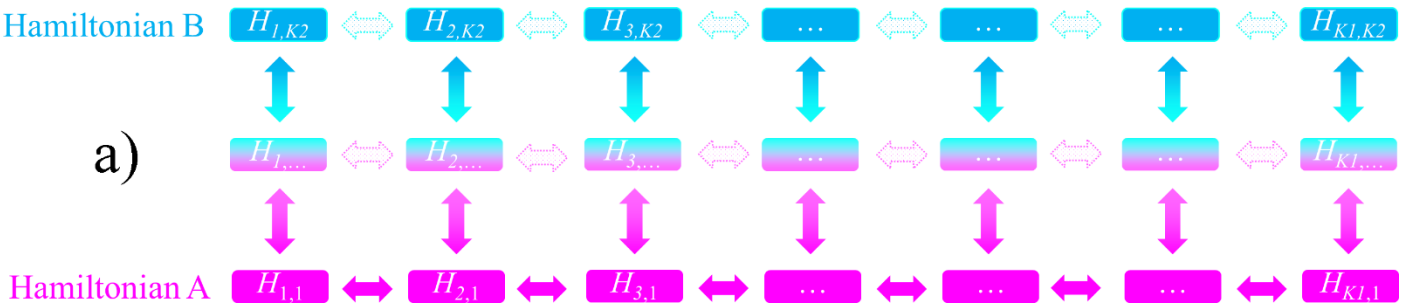
109. Pastor, R. W.; Brooks, B. R.; Szabo, A., An analysis of the accuracy of Langevin and molecular dynamics algorithms. *Mol. Phys.* **1988**, 65, 1409-1419.

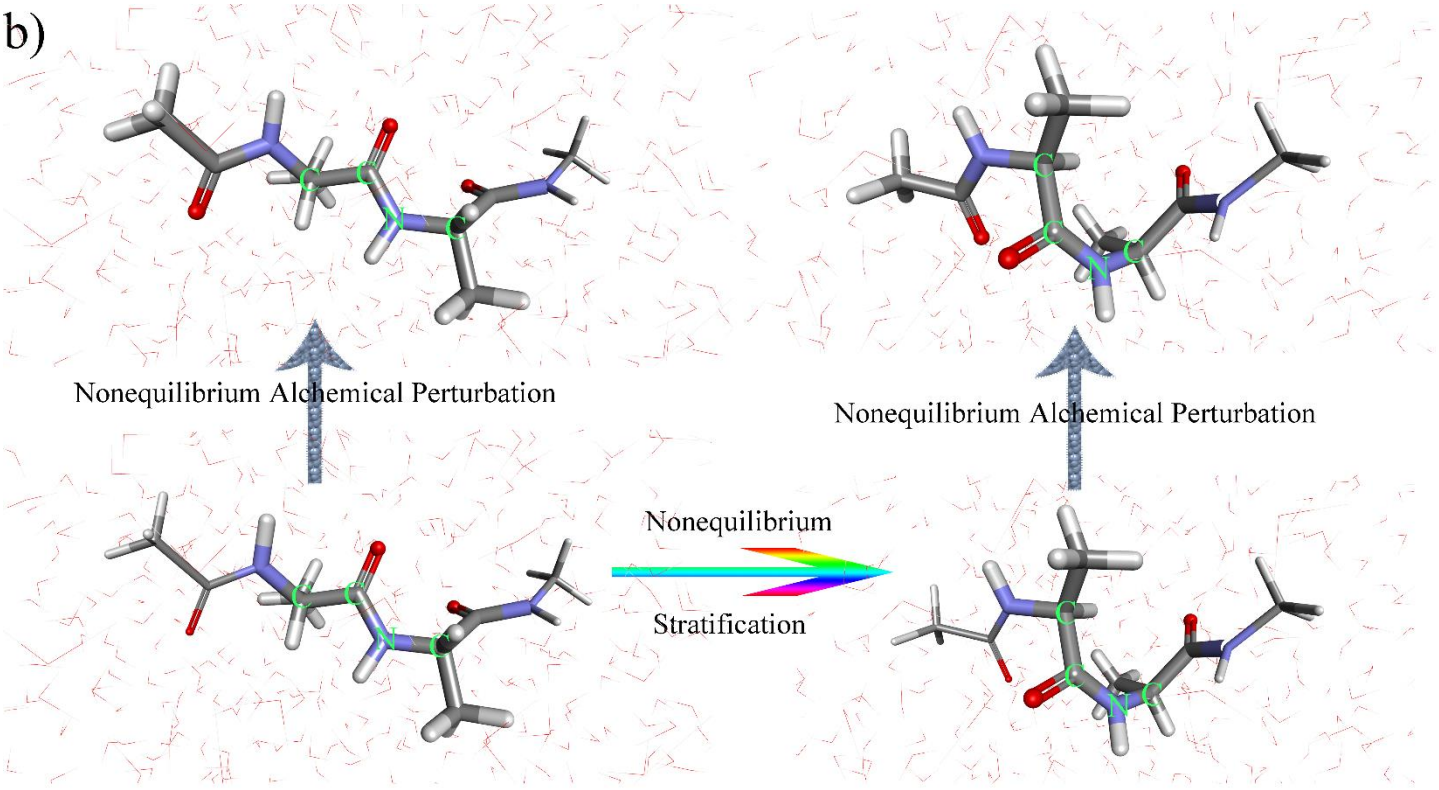
110. York, D. M.; Darden, T. A.; Pedersen, L. G., The effect of long‐range electrostatic interactions in simulations of macromolecular crystals: A comparison of the Ewald and truncated list methods. *J. Chem. Phys.* **1993**, 99, 8345-8348.

111. Case, D. A.; Cheatham, T. E.; Tom, D.; Holger, G.; Luo, R.; Merz, K. M.; Alexey, O.; Carlos, S.; Bing, W.; Woods, R. J., The Amber Biomolecular Simulation Programs. *J. Comput. Chem.* **2005**, 26, 1668-1688.

112. Zhaoxi, S.; Qiaole, H., *Seeding the Multi-dimensional Nonequilibrium Pulling for Hamiltonian Variation: Indirect QM/MM Free Energy Simulations*. 2021.

Fig. 1. a) An illustration of the Hamiltonian perturbation framework via multi-dimensional nonequilibrium free energy calculations. Hamiltonian perturbation is performed between neighboring states with nonequilibrium transformations. H_{k_1,k_2} denotes the Hamiltonian of the system at the k_1 th configurational state and k_2 th alchemical state. The target free energy landscape is at the Hamiltonian state $k_2 = K_2 = 2$. The indirect scheme performs direct free energy calculation at the Hamiltonian state $k_2 = 1$ to explore the configurational space, and adds a correction term $H_{k_1,1}$ <-> $H_{k_1,2}$ to perturb the thermodynamic profile to obtain the result at the target Hamiltonian state. Only the transformations described with solid arrows are performed due to efficiency considerations in nonequilibrium free energy simulations. b) A practical view of the thermodynamic cycle describing the dihedral flipping process with different sizes of the QM region. The reaction coordinate is the backbone dihedral in alanine tripeptide (C-C-N-C). The solute atoms drawn with thicker sticks are described with the QM Hamiltonian, while the other solute atoms (thinner sticks) and the line-type solvent atoms are described with MM potentials. Here, the nonequilibrium pulling in the configurational space is performed with the central two residues included in the QM region, and the nonequilibrium pulling in the alchemical space is used to perturb the results to the Hamiltonian with the 1-3 residues included in the QM region. c) The QM-region-variation variant of Fig. 1a. The QM region 1 is defined as $k_2 = 1$ and the QM region 2 has $k_2 = K_2 = 2$. The goal of the indirect scheme is obtaining the free energy landscape with the QM region 2. The direct free energy simulation with the QM region 1 is the transformation from $H_{1,1}$ to $H_{k_1,1}$, and the indirect scheme adds a correction term $H_{k_1,1}$ <-> $H_{k_1,2}$ in each state along the configurational CV, i.e. $k_1 = 1,\ldots,K_1$.




b)
Nonequilibrium Alchemical Perturbation
Nonequilibrium Alchemical Perturbation
Nonequilibrium
Stratification

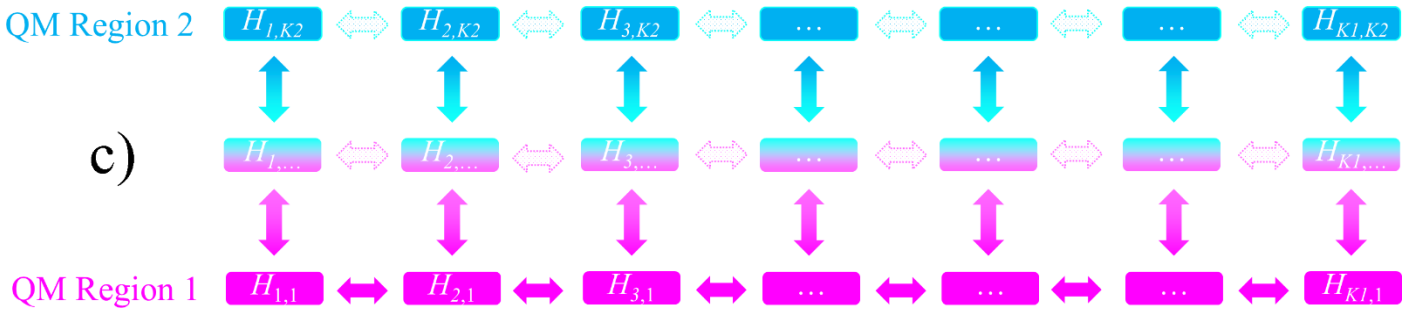
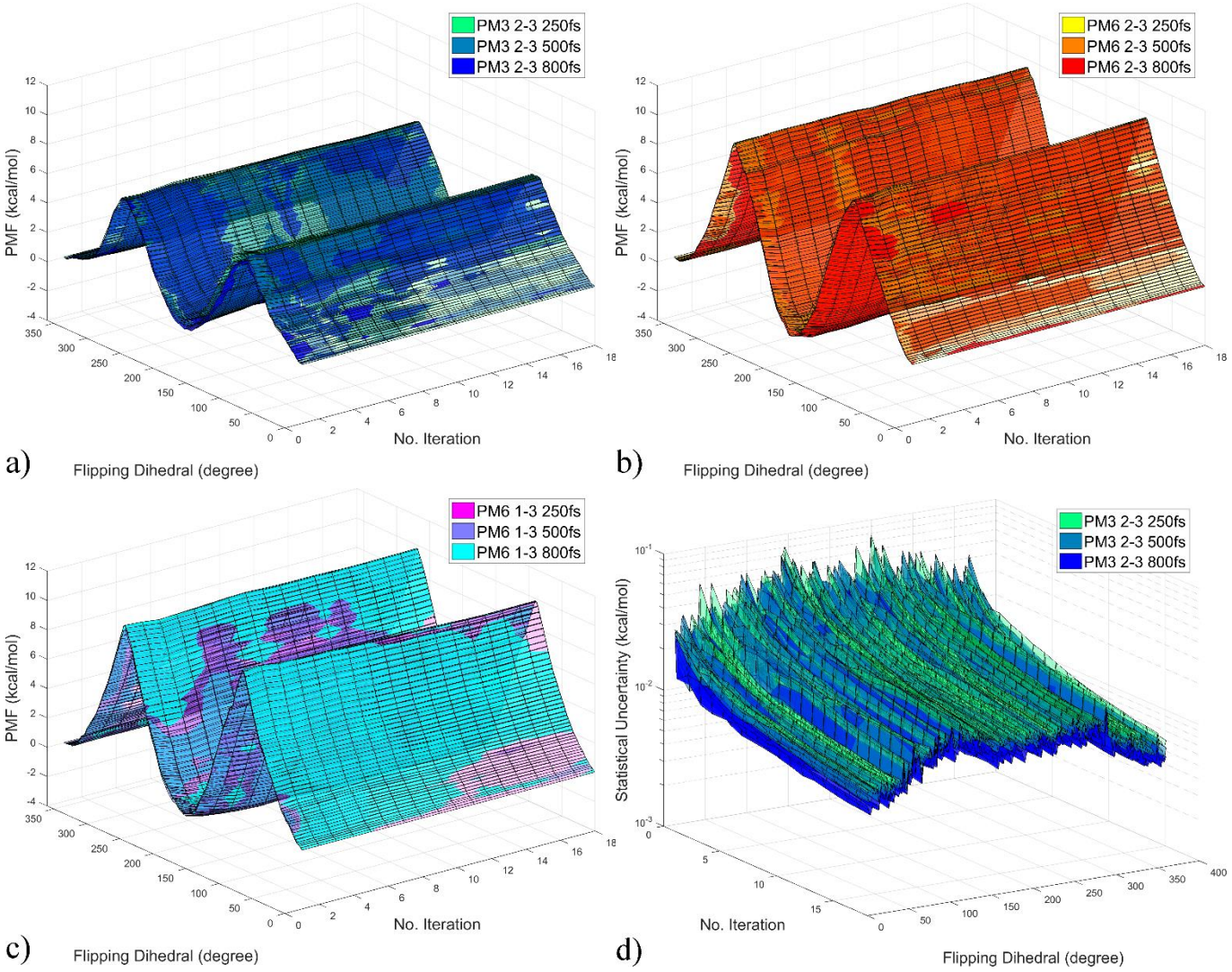
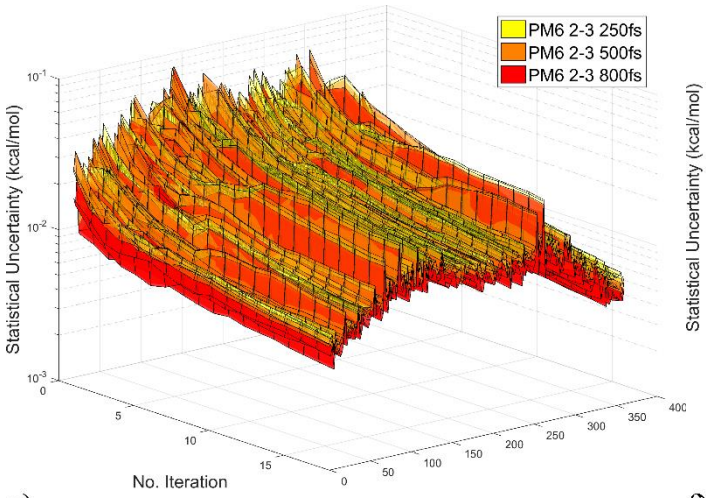
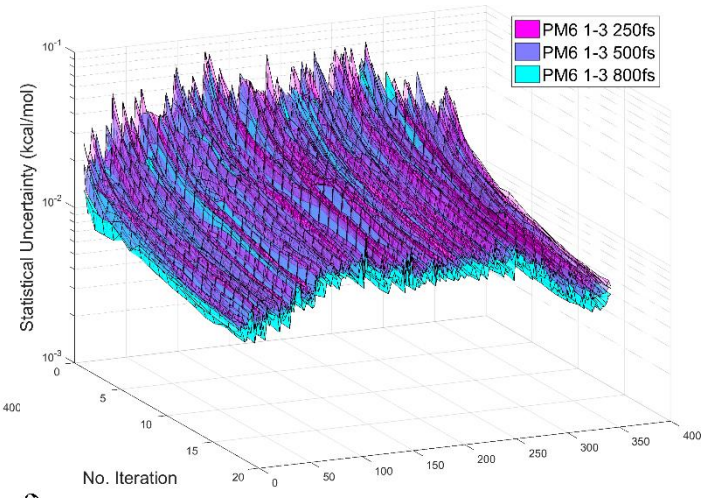

QM Region 2
c)
QM Region 1
$H_{1,K2}$
$H_{2,K2}$
$H_{3,K2}$
$H_{K1,K2}$
$H_{1,1}$
$H_{2,1}$
$H_{3,1}$
$H_{K1,1}$

Fig. 2. In QM/MM nonequilibrium pulling simulations of the solvated alanine tripeptide with different pulling speeds and different sizes of QM regions, the dependences of a-c) free energy profiles and d-f) SD profiles on the sample size. The initial sample size is 5 and further 5 samples are added to the dataset in each iteration. In the legend, the time represents the pulling time for each 2° segment, and the #-# flag denotes the residues included in the QM region. For instance, "PM6 2-3 500 fs" indicates the simulation setup of the PM6 Hamiltonian with the 2nd and 3rd residues included in the QM region and 500 fs pulling time in each direction for each 2° segment along the conformational CV.




PM6 2-3 250fs
PM6 2-3 500fs
PM6 2-3 800fs
Statistical Uncertainty (kcal/mol)
No. Iteration
Flipping Dihedral (degree)

e)


PM6 1-3 250fs
PM6 1-3 500fs
PM6 1-3 800fs
Statistical Uncertainty (kcal/mol)
No. Iteration
Flipping Dihedral (degree)

f)

Fig. 3. For the solvated alanine tripeptide, the comparison between the direct and indirect results under the PM6/MM Hamiltonians a) with the 1st, 2nd and 3rd residues included in the QM region and b) with the central 2nd and 3rd residues included in the QM region. The error bars are obtained with the standard error propagation procedure. The #-# flag in the legend denotes the residues included in the QM region. For instance, "2-3" leads to the simulation setup that the 2nd and 3rd residues are included in the QM region, while the 1st and 4th residues of the solute (i.e., the ACE and NME caps) and the water molecules are represented by MM Hamiltonians.

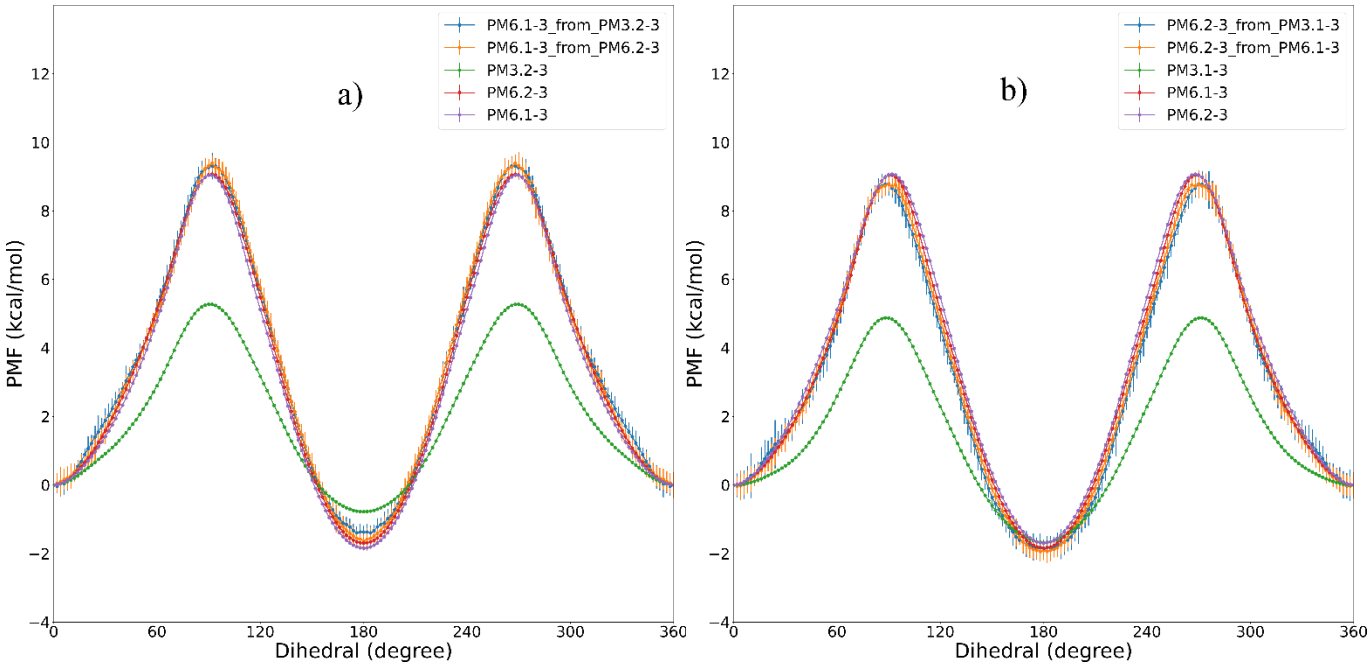
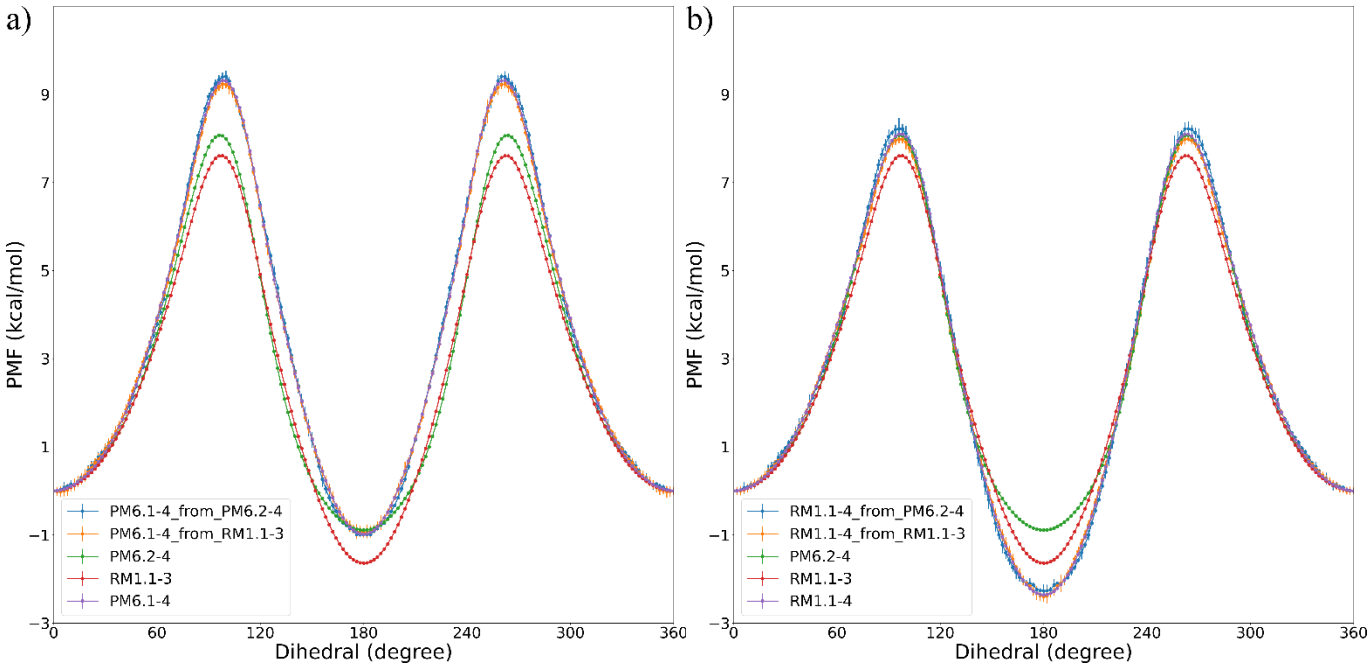


Fig. 4. In the gas-phase simulations, the comparison between the direct and indirect results with the whole alanine tripeptide included in the QM region at a) the PM6 level and b) the RM1 level. The error bars are obtained with the standard error propagation procedure. The #-# flag in the legend denotes the residues included in the QM region. For instance, "1-4" leads to the simulation setup that the whole alanine tripeptide is included in the QM region.



Supporting Information: BAR-based Multi-dimensional Nonequilibrium Pulling for Indirect Construction of QM/MM Free Energy Landscape: Varying the QM Region

Zhaoxi Sun[1*], and Zhirong Liu[1]

[1]*Beijing National Laboratory for Molecular Sciences, College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China*

*To whom correspondence should be addressed: z.sun@pku.edu.cn

Table S1. The single-core timing information (in ns/day) of QM/MM simulations of the solvated alanine tripeptide and the gas-phase one with different definitions of the QM region. As the nonequilibrium pulling simulations are independent and each of them is short in time, we perform each of them with a single core, thus avoiding the degradation of computational efficiency due to parallelization-related issues (e.g., communications overhead). We used Intel(R) Xeon(R) E5-2690 v4 @ 2.60GHz for the timing test.

Solvated:

QM theory/region / Terms	PM6 2-3	PM6 1-3	PM3 2-3
	H-ALA-ALA-H	ACE-ALA-ALA-H	H-ALA-ALA-H
Number of atoms including the link atoms	22	27	22
speed(ns/day)	0.71	0.58	0.71
speedup(QM/PM6 2-3)	1.00	1.22	1.00

Vacuo:

QM theory/region / Terms	RM1 2-4	RM1 1-3	RM1 1-4
	H-ALA-ALA-NME	ACE-ALA-ALA-H	ACE-ALA-ALA-NME
Number of atoms including the link atoms	27	27	32
speed(ns/day)	5.11	4.26	2.90
speedup(QM/RM1 2-4)	1.00	1.20	1.76

Fig. S1. An illustration of the thermodynamic cycle to obtain the free energy estimates with the whole alanine tripeptide included in the QM region. The reaction coordinate C-C-N-C describing the dihedral flipping is explicitly marked. The simulation is performed in vacuo. The atoms drawn with thicker sticks are described with the QM Hamiltonian, while the other (thinner sticks) are described with MM potentials. a) The nonequilibrium pulling in the configurational space is performed with the 2nd, 3rd and 4th residues included in the QM region, and the nonequilibrium pulling in the alchemical space is used to perturb the results to the Hamiltonian with the whole alanine tripeptide included in the QM region. b) The nonequilibrium pulling in the configurational space is performed with the 1st, 2nd, and 3rd residues included in the QM region, and the nonequilibrium pulling in the alchemical space is used to perturb the results to the Hamiltonian with the whole alanine tripeptide included in the QM region.

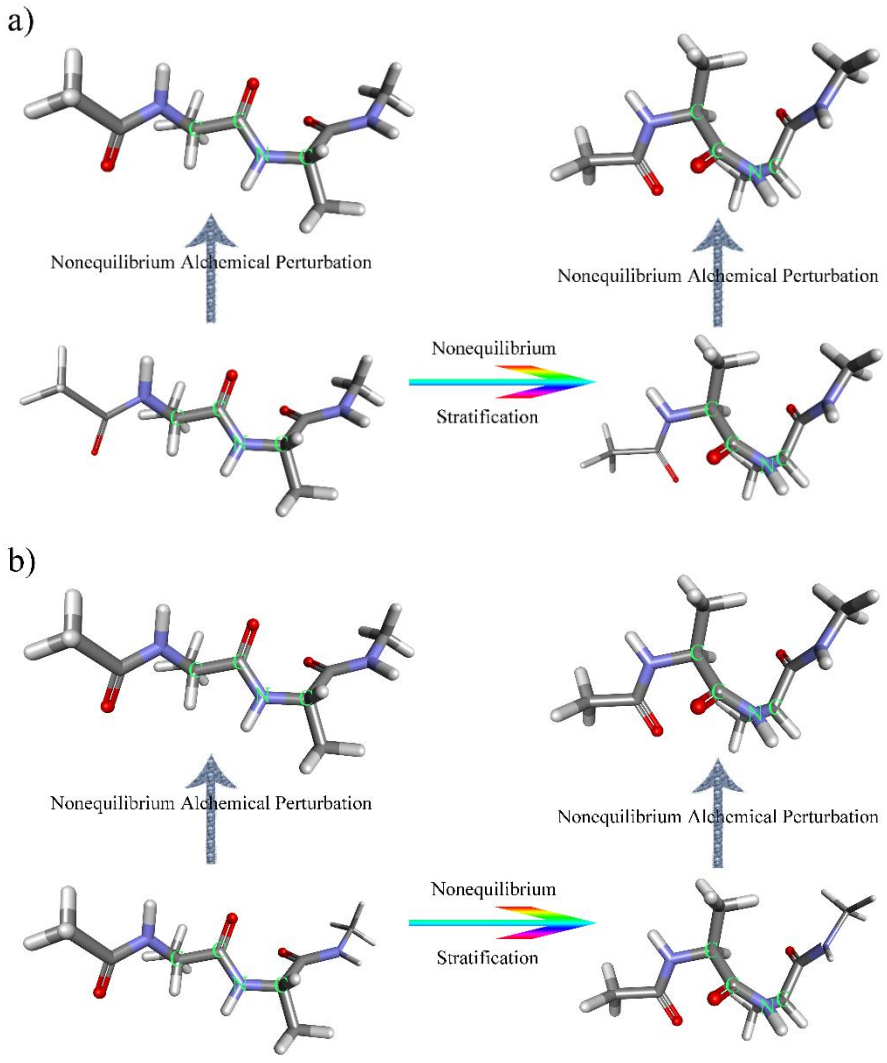
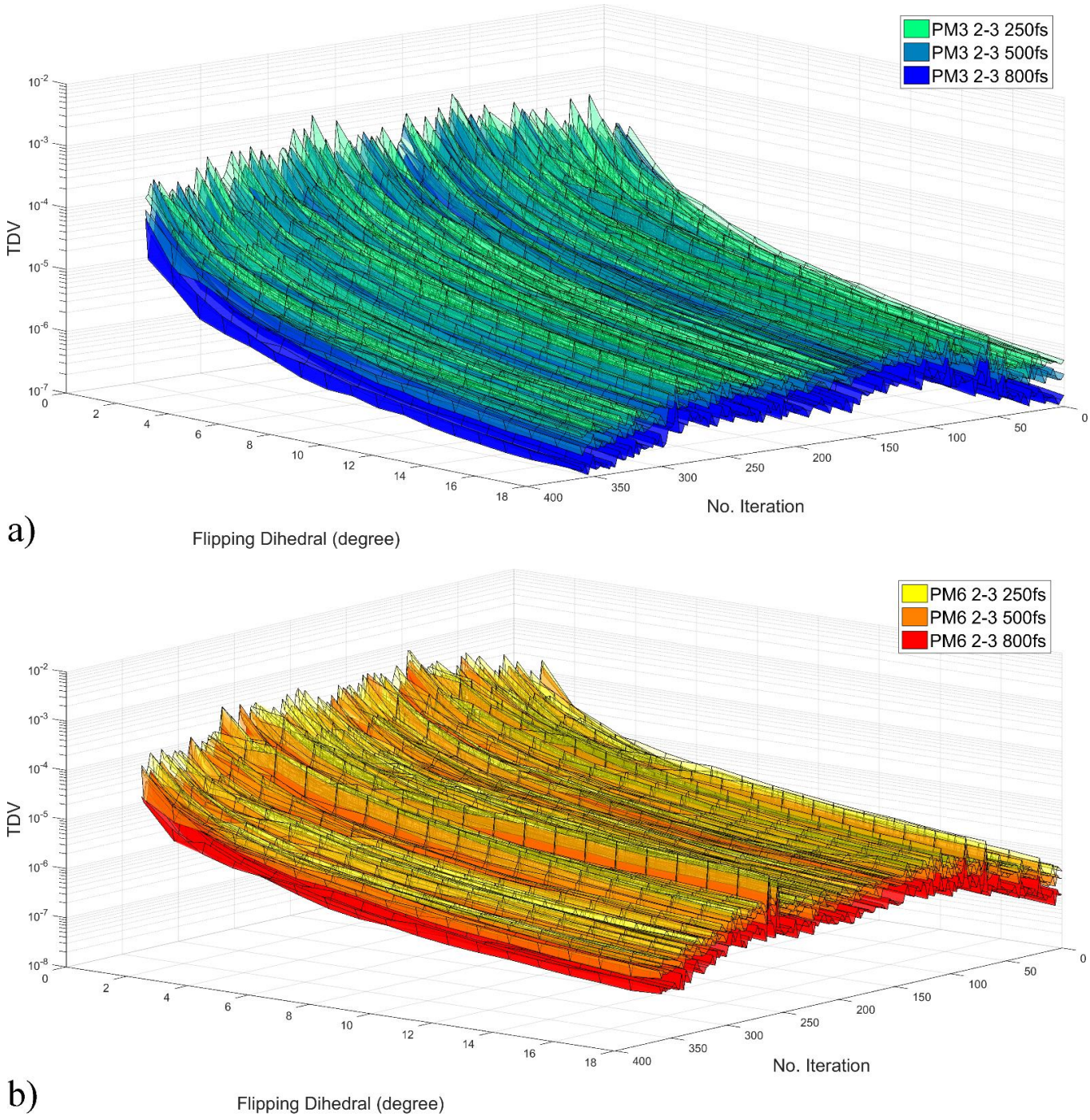
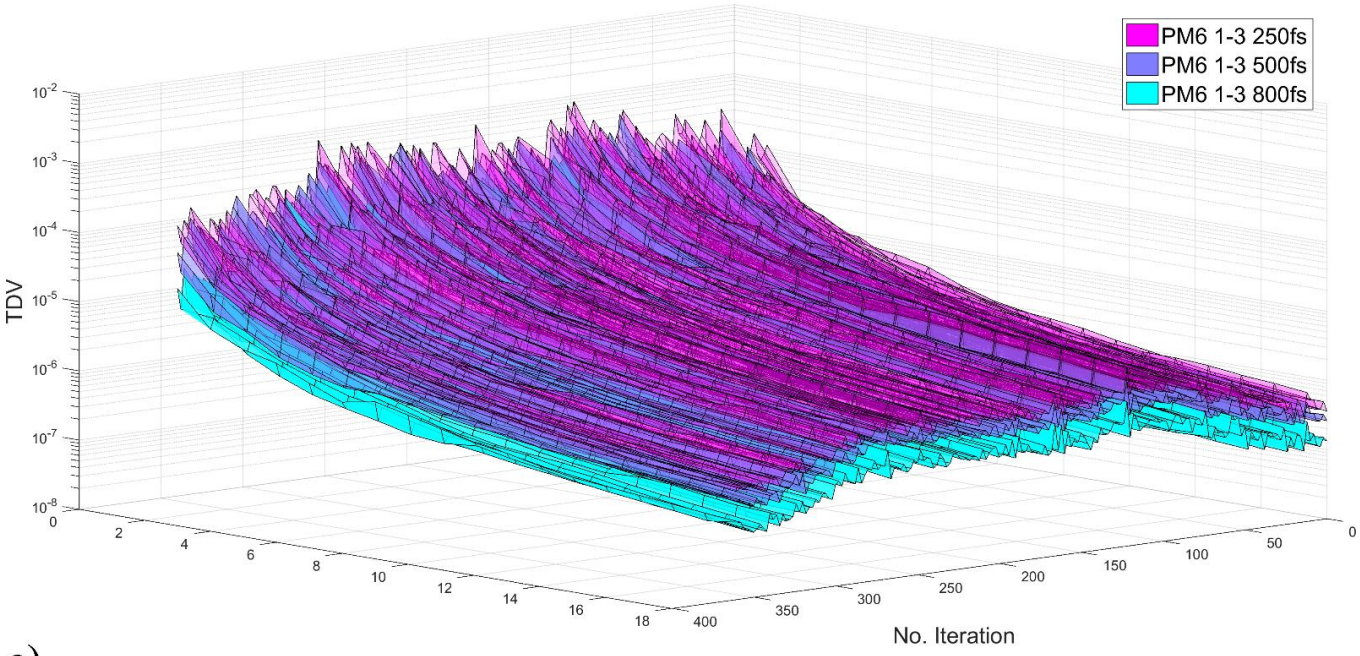


Fig. S2. In QM/MM nonequilibrium pulling simulations of the solvated alanine tripeptide with different pulling speeds and different sizes of QM regions, the dependences of the time derivative of overall variance (TDV) on the sample size. The initial sample size is 5 and further 5 samples are added to the dataset in each iteration. In the legend, the time represents the pulling time for each 2° segment, and the #-# flag denotes the residues included in the QM region. For instance, "PM6 2-3 500 fs" indicates the simulation setup of the PM6 Hamiltonian with the 2nd and 3rd residues included in the QM region and 500 fs pulling time in each direction for each 2° segment along the configurational CV.




PM6 1-3 250fs
PM6 1-3 500fs
PM6 1-3 800fs
TDV
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
0
2
4
6
8
10
12
14
16
18
Flipping Dihedral (degree)
400
350
300
250
200
150
100
50
0
No. Iteration

c)

Fig. S3. In the gas-phase simulations with the whole alanine tripeptide included in the QM region at the RM1 and PM6 levels, the dependences of a) the free energy profile, b) the statistical uncertainty, and c) the TDV on the sample size. The initial sample size is 5 and further 5 samples are added to the dataset in each iteration. In the legend, the time represents the pulling time for each 2° segment, and the #-# flag denotes the residues included in the QM region. The pulling speed of 0.5 ps/segment is observed to be sufficiently slow to ensure convergence.