CJC Chinese Journal of Chemistry 中国化学 · An International Journal

For submission: https://mc.manuscriptcentral.com/cjoc
For published articles: https://onlinelibrary.wiley.com/journal/16147065

Chemistry Authors Up Close

Cite this paper: *Chin. J. Chem.* **2022**, *40*, 160—168. **DOI: 10.1002/cjoc.202100456**

SPONGE: A GPU-Accelerated Molecular Dynamics Package with Enhanced Sampling and AI-Driven Algorithms

Yu-Peng Huang,[‡,a,b,c] Yijie Xia,[‡,a,b,c] Lijiang Yang,[‡,a,b,c,d] Jiachen Wei,[e,f] Yi Isaac Yang,[f] and Yi Qin Gao*[,a,b,c,d,f]

[a] *College of Chemistry and Molecular Engineering, Peking University, Beijing 100871, China*
[b] *Beijing National Laboratory for Molecular Sciences, Peking University, Beijing 100871, China*
[c] *Biomedical Pioneering Innovation Center, Peking University, Beijing 100871, China*
[d] *Beijing Advanced Innovation Center for Genomics, Peking University, Beijing 100871, China*
[e] *State Key Laboratory of Nonlinear Mechanics and Beijing Key Laboratory of Engineered Construction and Mechanobiology, Institute of Mechanics, Chinese Academy of Sciences, Beijing 100190, China*
[f] *Shenzhen Bay Laboratory, Gaoke Innovation Center, Guangqiao Road, Guangming District, Shenzhen, Guangdong 518132, China*

What is the most favorite and original chemistry developed in your research group?
Our research centers at developing methods and theories to unravel molecular mechanisms of chemical and biological systems. By establishing theoretical models, developing enhanced sampling methods combined with machine learning techniques, we are able to conduct comprehensive thermodynamic and dynamic analyses for these complex systems.

How do you get into this specific field? Could you please share some experiences with our readers?
I got into theoretical chemistry as a PhD student. My PhD adviser Prof. Rudolph A. Marcus led me into this field and inspired me by his love of science. Enjoy life, always learn new things and be independent in thinking are something I learnt from my advisers (Professors Dalin Yang, Qihe Zhu, Rudy Marcus, and Martin Karplus) and would love to pass to my students.

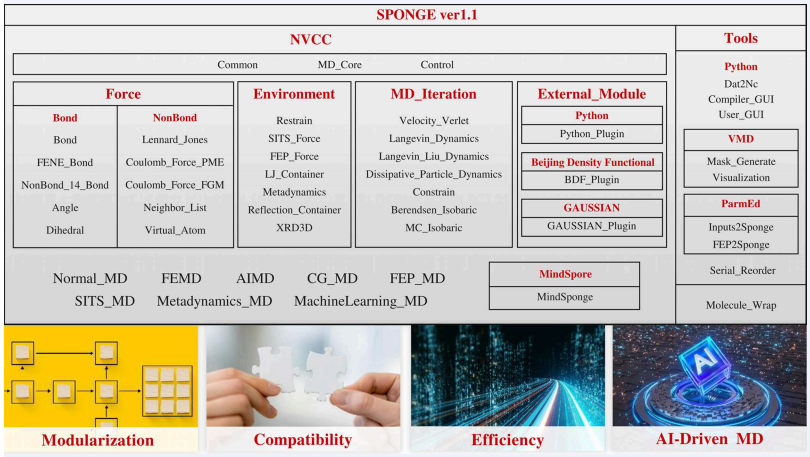
How do you supervise your students?
We learn from each other.

What is the most important personality for scientific research?
Curiosity, passion, and persistence have been of great value to my career.

What are your hobbies? What's your favorite book(s)?
Reading, Ping-Pong, and jogging. I always enjoy reading history.

Who influences you mostly in your life?
Too many, family, academic advisors, friends, students, and colleagues.

Comprehensive Summary



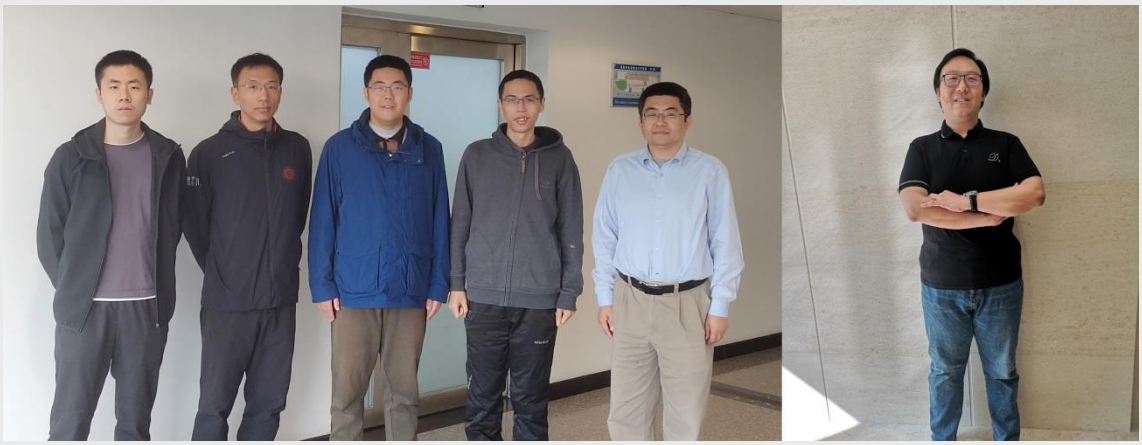
SPONGE (Simulation Package tOward Next GEneration molecular modeling) is a software package for molecular dynamics (MD) simulation of solution and surface molecular systems. In this version of SPONGE, the all-atom potential energy functions used in AMBER MD packages are used by default and other all-atom/coarse-grained potential energy functions are also supported. SPONGE is designed to extend the timescale being approached in MD simulations by utilizing the latest CUDA-enabled graphical processing units (GPU) and adopting highly efficient enhanced sampling algorithms, such as integrated tempering, selective integrated tempering and enhanced sampling of reactive trajectories. It is highly modular and new algorithms and functions can be incorporated con veniently. Particularly, a specialized Python plugin can be easily used to perform the machine learning MD simulation with MindSpore, TensorFlow, PyTorch or other popular machine learning frameworks. Furthermore, a plugin of Finite-Element Method (FEM) is also available to handle metallic surface systems. All these advanced features increase the power of SPONGE for modeling and simulation of complex chemical and biological systems.

Keywords

Molecular dynamics | Molecular modeling | Enhanced sampling | Machine learning | Computational chemistry

*E-mail: gaoyq@pku.edu.cn
‡ These authors contributed equally to this work.

View HTML Article | Supporting Information



Left to Right: Yu-Peng Huang, Lijiang Yang, Jiachen Wei, Yijie Xia, Yi Qin Gao, Yi Isaac Yang

Yu-Peng Huang received his B.S. (2019) from College of Chemistry and Molecular Engineering, Peking University. He is now studying for his Ph.D. in physical chemistry under the supervision of Prof. Yi Qin Gao at Peking University. His research interests include enhance sampling and high performance computing methods in molecular dynamics and simulations of complex chemical and biological systems.

Yijie Xia received his B.S. (2019) from College of Chemistry and Molecular Engineering, Peking University. He is now studying for his Ph.D. in physical chemistry under the supervision of Prof. Yi Qin Gao at Peking University. His research interests include simulations of complex chemical and biological systems.

Lijiang Yang received his B.S. (1997) and Ph.D. (2002) from Nankai University. He joined the University of California, Irvine as a postdoctoral fellow with Prof. Ray Luo in 2003. After his postdoctoral studies (2005—2009) at Texas A&M University, he joined the faculty of Prof. Gao's group at Peking University where now he is a Research Associate Professor. His current research interests include enhanced sampling molecular simulation methods and simulations of complex chemical and biological systems.

Jiachen Wei received his B.S. (2010) from Tongji University and his Ph.D. (2015) from Institute of Mechanics, Chinese Academy of Sciences. He was a visiting postdoctoral researcher at Department of Chemistry, University of Cambridge (2017–2019). He is currently a research associate professor at Institute of Mechanics, Chinese Academy of Sciences. He is also a visiting scientist at Shenzhen Bay Laboratory. His research interests include molecular simulations of soft matter and biological systems, and developing algorithms for coarse-grained simulations.

Yi Isaac Yang received his B.S. (2010) from Shandong University and his Ph.D. (2015) from Peking University. He was a postdoctoral researcher in the group of Prof. Michele Parrinello at ETH Zürich and USI Lugano from 2016 to 2019. He joined the faculty of Prof. Gao's group at Shenzhen Bay Laboratory in 2019, where now he is a research associate professor. His current research interests focus on developing theories, algorithms and software for molecular dynamics simulation and deep learning.

Yi Qin Gao received his B.S. (1993) from Sichuan University, M.S. (1996) from Institute of Chemistry, Chinese Academy of Sciences and Ph. D. (2001) from California Institute of Technology. He performed postdoctoral research at California Institute of Technology with Professor Rudolph A. Marcus (2001—2002) and Harvard University with Professor Martin Karplus (2002—2004). He started his independent career at Texas A&M University as an Assistant Professor (2004—2010). Then he joined College of Chemistry and Molecular Engineering, Peking University where now he is a Cheung Kong Scholar Chair Professor. He has been a Professor of Biomedical Pioneering Innovation Center, Peking University since 2013 and Vice Director of Shenzhen Bay Laboratory since 2020. His current research interests center on enhanced sampling, enzymatic reaction, machine learning assisted molecular simulation technique, chromosome folding and gene regulatory.

Contents

1. Introduction

Molecular dynamics (MD) simulation has been a useful tool in chemistry, physics, biology, materials science and many other fields. It helps in interpreting the experimental data and understanding the relationship between molecular structures, dynamics, and functions. During the last 40 years, various efficient computational algorithms[1-8] and MD programs[9-16] have been developed to study the dynamics of increasingly more complex and larger systems, such as RNA polymerase,[17] membrane proteins in cell membranes,[18] SARS-CoV-2 virus[19-20] and many others. However, as the scope and scale of applications increase, much higher computing capability is required for molecular simulation software. The most direct strategy to reduce the gap between simulations and experiments is to utilize more powerful computational hardware. For example, Anton in D. E. Shaw Research, a specially designed MD platform, can perform millisecond simulations for a single domain protein with system sizes of a few hundred thousand atoms.[21-22] In contrast, use of graphics processing units (GPUs) is probably the most affordable and promising approach for most research groups. From another facet, many advanced computational algorithms extending the simulation time scales have also been developed and widely used. In particular, a number of enhanced sampling methods have been developed in the last few decades to allow fast thermodynamics and/or kinetics calculations. Such methods include but not limited to the widely used umbrella sampling,[23] metadynamics,[24] accelerated MD,[25] replica exchange molecular dynamics (REMD),[26] parallel tempering,[27] simulated tempering,[28] multi-caonical simulation,[29] (especially implemented with the Wang-Landau algorithm[30]) and many others.

In the past 15 years, we have devoted into the development

of efficient molecular simulation methods towards complex chemical and biological systems and designed a series of enhanced sampling methods, which allow fast sampling of the configuration and trajectory spaces and realized fast calculations of thermodynamic and dynamic properties of complex systems. Recently, we have developed a home-grown MD simulation software package named SPONGE (Simulation Package tOward Next GEneration molecular modeling), which implements not only GPU-accelerated conventional MD simulations but also the efficient enhanced sampling methods proposed by our group. The software package is highly modular and additional functions or algorithms, especially the latest deep learning potentials and algorithms can be easily incorporated.

2. Efficient Simulation Method and Software Package for Solution and Surface Molecular Systems

Although there are a number of molecular dynamics simulation software developed, such as AMBER, CHARMM, GROMACS, LAMMPS, NAMD, ACEMD *etc*., we have designed and implemented a new MD software package: SPONGE recently with the expectation of providing an efficient MD algorithm developing framework to incorporate new algorithms and functions more conveniently. The name 'SPONGE' is coming from the full name 'Simulation Package tOward Next GEneration molecular modeling' and it also implies that the goal is to set up a MD platform which can efficiently 'absorb' all kinds of ideas and algorithms like sponge does. SPONGE 1.1 is released under the GNU General Public License v2 (GPLv2) and can be obtained free of charge from: http://www.spongemm.cn.

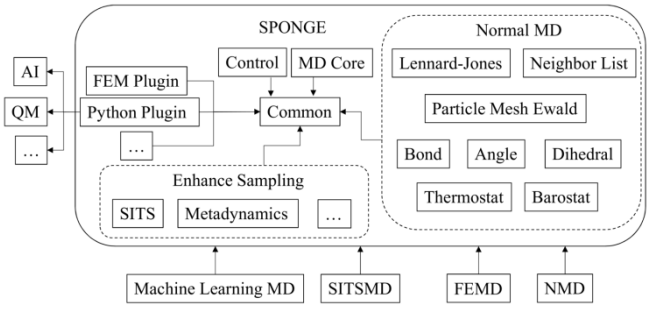


Figure 1 Program structure of SPONGE.

2.1. Classical MD module in SPONGE

The basic MD simulation steps consist of (1) evaluation of potential energy and force, (2) integration of coordinates and momenta, and (3) thermostat or barostat calculations. SPONGE takes the same potential energy function form as AMBER by default, therefore, the similar parameter and topology file formats as AMBER are followed in SPONGE. Currently, the leap-frog and the velocity-Verlet algorithm are available for the integration of coordinates and momenta. The widely used Langevin thermostat and Mont Carlo barostat is used for the temperature and pressure regulation, respectively. Liu's 'middle' thermostat scheme for Langevin dynamics[31] is also supported in constant temperature simulations. SPONGE has an efficient simulation engine implemented using CUDA and C++, so that all calculations of bonded and non-bonded energies and forces are accelerated on CUDA-enabled GPU. Currently, SPONGE can only support simulations on single GPU card. Supporting for running on multiple-GPUs is under development.

The classical MD module in SPONGE can be used to perform classical MD simulations on chemical and biological systems such as ionic aqueous solution, ionic crystal surface, and protein aqueous solution, *etc*. To illustrate the reliability and efficiency of SPONGE classical MD module, we simulated a system which consists of 7000 SPC/E water molecules in a cubic periodic box using both SPONGE and AMBER 16. The simulations were carried in NPT ensemble, NVT ensemble and NVE ensemble for 100 ns, respectively. The temperature of the whole system was maintained at 300 K and the pressure was regulated to 1 bar. Hardware/software configurations and simulation conditions applied in the testing of GPU-accelerated MD packages: SPONGE and AMBER 16 are summarized in Table 1.

Table 1 Hardware/software configurations and simulation conditions for the testing of GPU-accelerated MD packages

Item	Specification
Operating System	Ubuntu 16.04
CPU	AMD Ryzen 7 2700X@3.7 GHz
GPU	NVIDIA GeForce RTX 2080 SUPER
Compiler	NVCC
CUDA	CUDA Toolkit 9.0
SPONGE	SPONGE 1.1 (Single Precision, SP)[a]
AMBER	AMBER 16 (Hybrid Precision, SPDP)
Thermostat	Langevin Dynamics, Temperature 300 K Collision Frequency 3.0 ps^{-1}
Nonbonded	Cutoff 10.0 Å, Skin[b] 2.0 Å

[a] SPONGE 1.1 can only support single precision mode currently. The next version of SPONGE will support the hybrid precision mode. [b] Cutoff is the truncation distance for evaluating nonbonded interactions and Skin is the distance buffer in addition to Cutoff for building atom neighbor list.

The bulk water density calculated in NPT ensemble by the two MD programs agrees with each other very well (1.031 g/cm^3 for AMBER 16 and 1.039 g/cm^3 for SPONGE). As for NVE ensemble, the total energy deviation in SPONGE 1.1 is ~5% during 100 ns simulation and the corresponding value of AMBER 16 is ~1%. Other thermodynamic and kinetic characteristics of the system obtained in AMBER 16 and SPONGE 1.1 are also in good agreement, as shown in Figure 2.

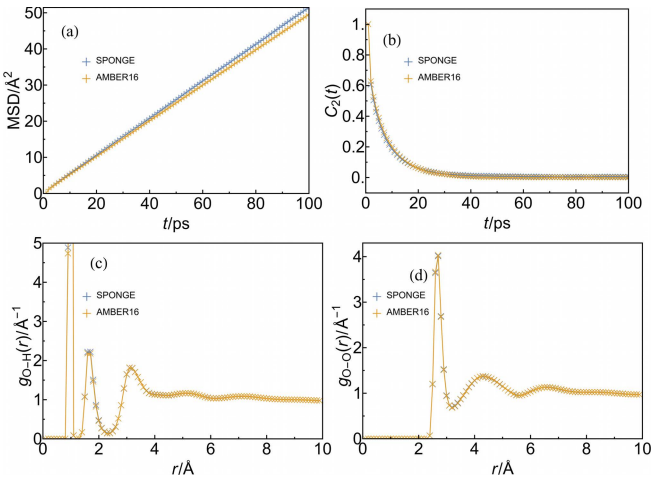


Figure 2 Comparison of thermodynamic and kinetic characters. (a) Mean-square displacement (MSD) of oxygen atoms. (b) Rotation relaxation of the dipole of water molecules. (c) Radial distribution function (RDF) of O—H. (d) RDF of O—O.

Next, we applied SPONGE to the systems which are well studied using MD simulations and tried to reproduce the results obtained by the previous simulations, for example, the interfacial transport of sodium ions by Peng and coworkers.[32] The simula-

tion system was composed of a hydrated Na^+, hydration water molecules (molecule number is from 1 to 5) and a NaCl (001) surface. AMBER 14 was used to perform MD simulations in this study. The same system and simulation setup were followed in SPONGE simulations. As shown in Figure 3, the similar diffusion rates of Na^+ with respect to the hydration number are obtained in SPONGE and Ref. [32].

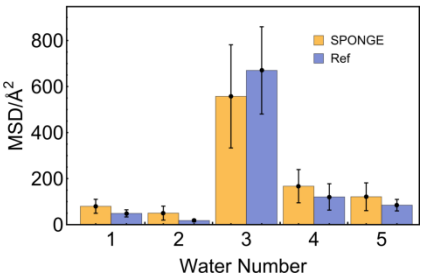


Figure 3 MSD of the hydrated Na^+ obtained by SPONGE and Ref. [32].

To show the reliability of SPONGE in biological systems, we then applied both SPONGE 1.1 and AMBER 16 to simulate a system with the alanine dipeptide (ACE-ALA-NME) solvated in a cubic periodic water box. AMBER FF96 all atom protein force field was used to modeled the peptide and 655 water molecules were described using TIP3P water model. The simulation was performed for 100 ns in NVT ensemble. As shown in Table 2, the differences of the different potential energy terms obtained by AMBER 16 and SPONGE are smaller than the fluctuations. In addition, the comparison of single point energies is shown as Figure S1 in SI.

Table 2 Comparison of different energy terms obtained by AMBER 16 and SPONGE 1.1

		E_{bond}	E_{angle}	$E_{dihedral}$	E_{elec}	E_{vdw}	E_{1-4}
$\langle E \rangle^a$	AMBER 16	1209.8	8.834	4.522	−8564.4	1281.0	49.16
	SPONGE 1.1	1210.7	8.847	4.536	−8572.4[b]	1285.1[c]	49.12
σ_E^a	AMBER 16	35.4	2.366	1.429	83.3	45.1	2.69
	SPONGE 1.1	35.4	2.342	1.418	83.8	45.2	2.65

[a] The average and the standard deviation of energy is in unit of kcal/mol. [b] PME algorithm is applied to calculate the electrostatic interactions. [c] Heaviside step function is used as the switch function in Lennard-Jones potential calculations.

Then the benchmark performance tests of SPONGE on the biological systems with 100—500 thousand atoms (Figure 4) were executed on a NVIDIA GeForce RTX 2080 Super GPU. The performances show that it is practical to use SPONGE to simulate large-scale biological systems (Table 3).

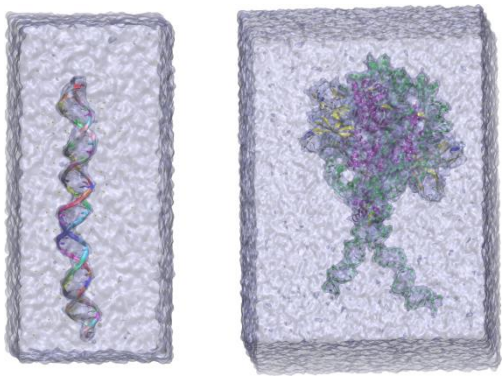


Figure 4 The two biological systems for performance tests.

Table 3 Performance tests of SPONGE on large biological systems under NVT conditions.

System	Atom number	Cutoff/Å	Time step/fs	Performance (ns/d)
DNA linker between two nucleosomes	108 373	8.0	4.0	140
DNA wrapped around histone	553 708	8.0	4.0	20

It needs to be mentioned that the performance of SPONGE 1.1 (in terms of speed) is currently lower than that of the latest version of AMBER and GROMACS (Tables S8 and S9 in SI). However, we believe that as a newly developed program, SPONGE has great potential in performance improvement. We are consistently working to improve the efficiency of this simulation package. Furthermore, it is enhanced sampling, machine learning based algorithms and other unique functions that distinguish SPONGE from traditional molecular simulation packages. These unique features of SPONGE are the focus of the following sections.

2.2. Enhanced sampling module in SPONGE

Besides the basic MD simulation algorithms, SPONGE now supports several enhanced sampling algorithms. As stated in Introduction section, in order to accelerate thermodynamics and/or kinetics calculations, many enhanced sampling methods have been proposed. Usually, most of the enhanced sampling methods can be classified into two categories: collective variable based and collective variable free methods. The enhanced sampling methods belong to the former category usually introduce bias potentials along predefined reaction coordinates or collective variables to the Hamiltonian of the systems to accelerate thermodynamics calculations. The most popular and widely used enhanced sampling methods such as umbrella sampling and metadynamics fall into this category. Unfortunately, the proper reaction coordinates are not easily identified for many systems, such that, generalized-ensemble methods that do not need predefined collective variables are also desired. In most collective variable free methods, the canonical probability distribution is altered to a distribution that induces a broader sampling of the potential energy. As a result, an efficient sampling of the configurational space will be obtained accordingly. Some examples of generalized-ensemble methods are parallel tempering, simulated tempering, replica exchange molecular dynamics (REMD), simulated tempering, multi-canonical simulation, among others.

At present the enhanced sampling methods implemented in SPONGE are integrated tempering sampling (ITS) and its variants proposed by our group. In the following, the basic idea of enhanced sampling methods ITS, selective integrated tempering sampling (SITS) and enhanced sampling of reactive trajectories are briefly introduced at first, then the implementation and testing of SITS enhanced sampling module of SPONGE is presented.

2.2.1. Integrated tempering sampling method. In 2008, we proposed the integrated tempering sampling (ITS)[33-34] method, which combines some of the advantages of both biasing potential and generalized-ensemble methods. It possesses the most important feature of generalized-ensemble methods: no predefined collective variable is needed in ITS. Secondly, ITS has almost the similar computational cost as biasing potential methods. Instead of running parallel simulations at many different temperatures in REMD, ITS generates an energy distribution covering a broad range of energies in a single simulation. Therefore, the ITS method avoids multiple parallel calculations and exchange operations between parallel trajectories, and thus requires fewer computational resources. In ITS an effective potential U_{eff} is generated at simulation temperature β_0 based on a sum-over-temperature non-Boltzmann distribution $f'(U) = \sum_i^N n_k f_k(U)$, which can be considered as a linear combination of a series of Boltzmann dis-

tribution $f_k(U) = \exp(-\beta_k U)$ at a range of different temperatures $\{\beta_k\}$ with coefficients $\{n_k\}$. This operation allows efficient sampling in a desired energy range without requiring a pre-defined collective variable:

$$U_{\text{eff}} = -\log \sum_k^N n_k e^{-\beta_k U} / \beta_0 \quad (1)$$

where U is the potential energy of the system under study, $\beta_0 = 1/k_B T_0$ (k_B is the Boltzmann constant and T_0 is the temperature of the system), β_k denotes a series of temperatures that cover both low and high temperatures around T_0, and n_k denotes weighting factors obtained through an iterative procedure. The biased force F_{eff} used in the simulations is:

$$F_{\text{eff}} = -\partial U_{\text{eff}} / \partial r = F \cdot \sum_k^N n_k \beta_k e^{-\beta_k U} / \beta_0 \sum_k^N n_k e^{-\beta_k U} \quad (2)$$

where F is the force in the original system. In ITS simulations, a converged calculation yields a biased distribution function in the configuration space, $p_{\text{ITS}}(r) \propto e^{-\beta_0 U_{\text{eff}}(r)}$. The desired distribution $p_x(r)$ at any object temperature $\beta_x \in \{\beta_k\}$ is easily recovered as:

$$p_x(r) = p_{\text{ITS}}(r) e^{-[\beta_x U(r) - \beta_0 U_{\text{eff}}(U)]} \quad (3)$$

In contrast to the REMD method, ITS not only requires fewer computational resources, but also circumvents the problem of re-equilibration for the kinetic energy arising from the exchange events in REMD. Recently, You *et al.* showed ITS is highly efficient and found that ITS is equivalent to simulated tempering with the attempt switching frequency going to infinity.[35]

To characterize the sampling efficiency in obtaining the thermodynamics properties, the convergences of thermodynamic calculations were investigated for accelerated molecular dynamics (AMD), REMD and ITS (Table 4). The model system is an ALA-PRO peptide that has a high free-energy barrier of ~20 kcal/mol for its transition between *trans* and *cis* conformations. It is clear that the potential of mean force (PMF) obtained from the AMD simulations converges very slowly owing to the under-sampling of low-energy states. REMD and ITS both obtain better convergence of thermodynamics calculations. In particular, ITS simulations give the best convergence with the least computational time.

Table 4 Convergence and computational costs for ALA-PRO peptide

Method	Trajectory length/ns	Wall clock time/h	CPU time/h	RMSD[a]/ (kcal·mol^{-1})
AMD	800	12.4	99	1.22
REMD	2400	25	600	0.48
ITS	800	12.5	100	0.21

[a] RMSD: the root-mean-square deviation of free energies that is used to characterize the convergence of the methods. $\text{RMSD}(F_1, F_2) = \sqrt{E(F_1 - F_2)^2}$.

2.2.2. Selective integrated tempering sampling method. Since the enhanced sampling of ITS is performed in the energy space, it is convenient to divide the system into subspaces and to enhance the sampling for a preselected subsystem. Therefore, the selective integrated tempering sampling or SITS[36] was proposed, which is especially practical in the simulations of the systems including a large amount of explicit solvent. For example, in explicit solvent simulations of protein folding, the protein atoms can be targeted for enhanced sampling and a large amount of protein conformations can be sampled while the solvent is kept at the near-room-temperature conformations. At first, the system is divided into different components. For example, in a protein solution, the protein is considered as one component (the central group, labeled as P) and the water as the other (the bath, labeled as W). The potential energy of this system U is then written as:

$$U = E_P + E_W + E_{PW} \quad (4)$$

where E_P, E_W, and E_{PW} are, respectively, the internal energy of the central group (protein), the internal energy of all water molecules and the energy of interaction between the protein and water. The differentiated sampling of such a system is conveniently achieved by introducing an effective potential in the form:

$$U_{\text{eff}} = E_W - \log \sum_k^N n_k e^{-\beta_k (E_P + E_{PW})} / \beta_0 \quad (5)$$

As shown in Eq. 5, the enhanced sampling is applied selectively to the degrees of freedom that are involved in the region of interest, whereas the rest of the system such as the solvent is kept as close as possible to its equilibrium.

SITS has been applied to study protein folding and DNA base flipping in aqueous solution.[37] SITS is also naturally introduced to QM/MM calculations,[38-41] in which only a small part of the simulation system is treated using quantum mechanics and the rest by classical molecular mechanics. For such calculations, since normally the interested events, such as chemical reaction, occur in the quantum region, it is desirable to explore the molecular configurations of the quantum mechanically treated subsystem. One can therefore make use of the SITS scheme to enhance sampling over the QM region whereas keep the MM part less perturbed by introducing the following effective potential:

$$U_{\text{eff}} = E_{\text{MM}} - \log \sum_k^N n_k e^{-\beta_k (E_{\text{QM}} + E_{\text{QM/MM}})} / \beta_0 \quad (6)$$

where E_{MM} is the self-energy of the MM region (*e.g.*, the solvent), E_{QM} is the self-energy of the QM region (*e.g.*, the reacting molecule), and $E_{\text{QM/MM}}$ interacting energy between the QM and MM regions. In a series of studies, we applied SITS-QM/MM to the aliphatic Claisen rearrangement reaction, in which the solute, or the reactant, alone was embedded in QM-frames, and solvent molecules were treated with MM. Since the high-performance sampling of QM-treated parts, the reactant reached a conformational equilibrium between compact and extended conformation in simulations. And the compact conformation, which was the proper configuration for the subsequent electron rearrangement, was found to be polarized with respect to the extended conformation according to Mulliken population analysis of QM/MM-calculated data. SITS allows the chemical transition process overcoming high barriers to be realized in silico without any intrinsic reaction coordinates (IRC). SITS herein exhibits excellent capacity and adaptability for sampling chemical events.

2.2.3. Enhanced sampling of reactive trajectories. Both ITS and SITS are highly effective in conformation searching and thermodynamics calculations. However, dynamics as well as kinetics information of the original system is lost owing to the use of effective Hamiltonians in ITS/SITS. Inspired by transition path sampling (TPS),[42] we introduced a combined approach that takes advantages of both ITS/SITS-MD and TPS (shooting) methods (enhanced sampling of reactive trajectories, ESoRT[38-40] for short). In this method, we first carry out ITS/SITS-MD simulations to identify the active phase space of the reaction of interest and then path sampling (shooting) is performed on the original potential-energy surface starting from the phase space points identified by the ITS/SITS-MD simulation. Since the statistical weight of each trajectory can be calculated using Eq. 3, the rate constant of the original system can be obtained directly as the ratio of the reactive and total trajectories:

$$\text{Rate} = \sum_{\text{suc traj}} e^{f(v)/k_B T} / t \sum_{\text{all traj}} e^{f(v)/k_B T} \quad (7)$$

Although ITS/SITS-MD provides a biased sampling of the phase space, the phase space points which are more likely to be 'reactive' will be chosen with a high probability for the forward/backward trajectory shooting. Therefore, ESoRT has four major advantages compared to the traditional TPS: (1) the initial trajectories are automatically generated by efficient ITS/SITS-MD simulation; (2) a thorough sampling of the phase space avoids the

entrapment of trajectories in a particular pathway(s), allowing the search of multiple pathways separated by high barriers; (3) ESoRT further reduces the computational cost by reducing the sampling over the unsuccessful transition paths; and (4) a direct calculation of both reactive and non-reactive trajectories that avoids the predetermined reaction coordinate and the calculation of the reactive flux.

ESoRT sampling method was successfully applied to the QM/MM simulation of Claisen rearrangement mentioned above. The prerequisite for the enhanced sampling of reactive trajectories is to generate trajectories that connect the reactant and product without or merely with a few predefined reaction coordinates. Successfully reacted trajectories from QM(DFTB)/MM simulations were collected and registered as 'reactive trajectories', with atomic coordinates, velocities, charges and the effective weighting factors recorded. Then, along each reactive trajectory, a series of configurations, which contain both reactant-like structures and configurations near transition states but not product-like ones, were selected as the initial structures for the following transition trajectory shooting. Finally, after a number of short (~2 ps) NVE simulations, the rate constant was retrieved as the ratio between the probability of reacted and that of non-reacted trajectories, each reweighted to reflect their probability distribution in the unbiased ensemble. The results calculated were in good agreement with the experimental results. The reaction pathway information can also be obtained by projecting the calculated visiting probabilities to the post-selected coordinate. By providing dynamic observations of transition events, this method puts the investigation of condensed phase reactions under molecular-detailed perspectives.

2.2.4. Test of SPONGE on enhanced sampling – SITS. SITS enhanced sampling method has been implemented as an independent module in SPONGE. As shown in section 2.2.1. and 2.2.2., the algorithm of SITS is easily to be implemented. However, the partition of the energy and force in readily available MD program causes difficulties. Since SPONGE is specifically designed to be easily controlled, it is much easier for SITS module of SPONGE accessing and updating all the internal data. The same alanine dipeptide (ACE-ALA-NME) test case mentioned above is used to illustrate the efficiency and the accuracy of the SITS module. As shown in Figure 5, the potential of mean force in ψ calculated using SPONGE reproduced the results of Ref. [36].

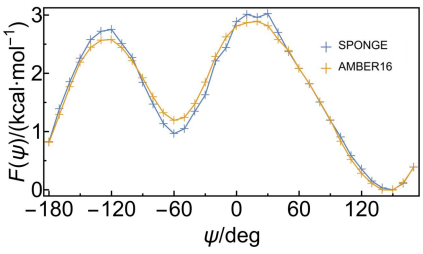


Figure 5 Comparison of the potential of mean force in ψ using SPONGE and Ref. [36].

2.3. Other unique functions in SPONGE

Besides ITS and SITS enhanced sampling methods, metadynamics and the combination of metadynamics and ITS/SITS: metaSITS is also implemented in SPONGE.

In addition, SPONGE provides a specific FEM module to calculate electrostatic interactions using finite element method (Figure 6). As shown in Figure 6, FEM module first initializes finite element grids, after which the electric potential distribution is obtained by solving Poisson's equation under a given boundary condition. The principle of FEM is presented in Figure 7 and the mathematical details of this method will be illustrated elsewhere. FEM is particularly useful to systems with strong polarizable effect, for example, the systems containing metal surfaces. As shown in Figure 8, the electric field induced by ions and metals would strongly influence the dynamics of charged particles. For this type of systems, SPONGE suggests the using of FEM to calculate the molecular electrostatic interaction instead of the traditional particle mesh Ewald method used in the most other MD programs.

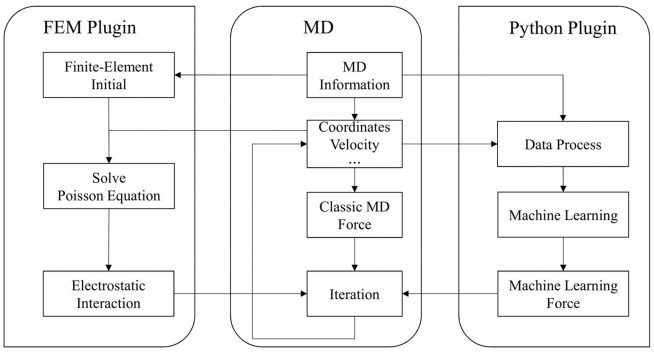


Figure 6 Flowchart of MD with FEM and Python plugins.

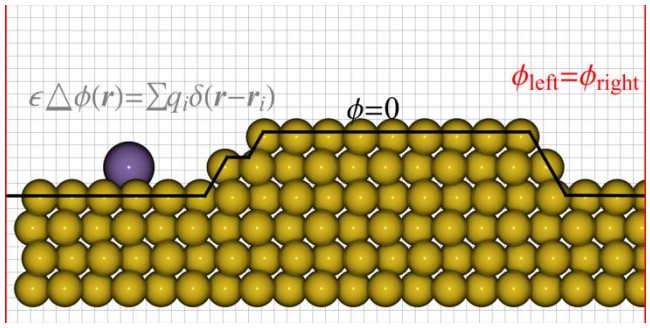


Figure 7 Illustration of FEM in SPONGE. Above the metal surface, the potential Φ satisfies Poisson's Equation. Closing to the metal surface, the equipotential surface (black line) gives the boundary condition. The red line gives the periodic boundary condition.

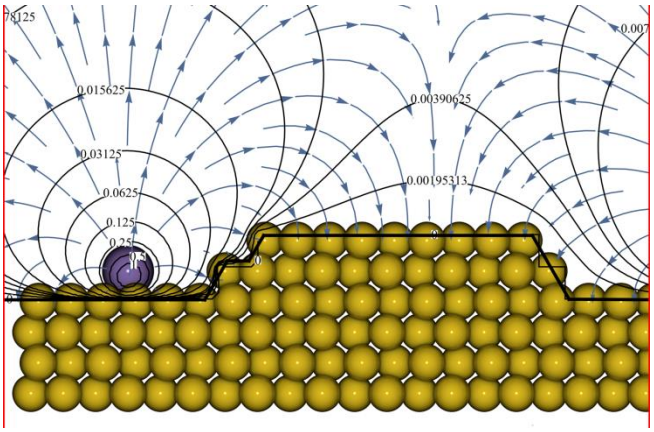


Figure 8 The numerical solution of Poisson's Equation above the metal surface. The electric field lines show that the electric field induced by ions and metals would strongly influence the dynamics of charged particles.

SPONGE also provides a Dissipative Particle Dynamics (DPD) module for coarse-grained simulations of macromolecules, soft matter, complex fluids, and so forth.[43-44] This module is designed to address micro-to-mesoscale problems with preserved hydrody-

namics. Moreover, the DPD module without the conservative portion of the force field can also be implemented as a thermostat. As an example, we performed simulations of the mono-component DPD fluids in a cubic box of side length 8.52 nm. Both SPONGE and LAMMPS were used to measure the diffusional properties of the fluids. As shown in SI, almost the same exponential decay of the translational diffusion coefficient *Dt* with dissipative constant *γ* are obtained both in SPONGE and LAMMPS. Details about the DPD simulation example can be found in SI.

Python programming interface is supported by SPONGE (Figure 6). Deep learning, with the artificial intelligent Go game program "AlphaGo" being their representative, showed great potential in many areas, including computational chemistry. Since most of the current machine learning frameworks are based on python, the support of python programming interface enables SPONGE the capability of incorporating deep learning techniques to develop high precision force fields and efficient enhanced sampling algorithms. As an example, a low-cost fluctuating charge model was implemented in SPONGE for the MD simulations of cellulose with the degree of polymerization of n = 6 in ionic liquid $[C_1MIM]^+ \cdot Cl^-$. In classical molecular dynamics simulations, charge distributions of molecules are usually fixed. Although this assumption is reasonable for systems with weak polarizability, it may fail in the highly polarizable systems. The fluctuating charge model is a feasible way to include polarizability of the system, which requires charge refitting based on quantum chemical calculations every certain steps in simulations, such that, huge computational cost is requested. By taking use of SITS module, python programming interface of SPONGE and the neural network algorithms implemented in MindSpore, one can largely reduce the computational cost of the fluctuating charge model. Firstly, SITS was used to obtain sufficient typical molecular conformations, then quantum chemical calculations were performed for these typical conformations to yield varied charge distributions. These pre-obtained conformations and charge distributions then were used to build up the training and testing set for the supervised learning models provided by MindSpore. When the training of the neural network model was successfully accomplished, the updating of molecular charges could be immediately achieved as the output of the neural network through using the newly reached conformations as the input. In this way, the computational costs of traditional fluctuating charge models were largely reduced. We are also trying to transfer the other deep learning based molecular simulation algorithms into SPONGE, such as Information Distilling of Metastability (IDM, to perform clustering in the meantime of reducing the dimensionality for molecular systems),[45] variational adversarial density estimation (VADE, to approximate the free energy surface by parametric models and without supervision)[46] and targeted adversarial learning optimized sampling (TALOS, to re-formulate the enhanced sampling problem as a distribution learning problem).[47]

3. Conclusions

As the continuous increase of the demanding on both system size and simulation time, a number of efficient and reliable sampling methods have been developed in the recent two decades for studies of complex chemical and biological systems. A series of efficient algorithms based on the integrated tempering sampling approach developed in our group can enhance sampling in both phase and trajectory space, which combines the advantages of both biasing potential and generalized-ensemble method. The successful applications of the ITS/SITS to the chemical and biological systems have shown its efficiency in configuration searching and thermodynamics calculations. Furthermore, ITS/SITS can be combined with the transition path sampling technique (ESoRT) to enhance the sampling of reactive trajectories and rate constant calculations. All these methods are expected to have broad applications for a large variety of complex systems.

Partly to efficiently incorporate these enhanced sampling simulations, we have designed and developed a home-grown MD software platform SPONGE. The comprehensive tests on classical and ITS/SITS enhanced sampling simulations have shown that SPONGE is a reliable and robust simulation program. In our opinion, a promising MD program should meet two requirements: 1. as many as possible commonly used force fields should be supported, 2. as many as possible efficient MD algorithms should be easily incorporated. Therefore, SPONGE is designed to be modularization in structure and to be ready for users incorporating new algorithms and functions. In addition, SPONGE supports not only classical MD simulations (AMBER, OPLS-AA, CHARMM27/CMAP and CHARMM36 force fields are currently supported), enhanced sampling simulations, but also deep learning based force fields/algorithms, which is implemented through a flexible python programming interface. Furthermore, we are now integrating SPONGE into the machine learning framework: MindSpore. In this framework, all force computations including bond, angle, dihedral, Lennard-Jones and Coulomb interactions in MD calculations are expressed as MindSpore arrays and operations. Therefore, a pioneering branch of SPONGE called MindSPONGE is invoked, which is designed to be an end-to-end differentiable MD engine. Since it is based on MindSpore which has a developing neural network ecosystem, MindSPONGE enables researchers to incorporate machine learning models into their workflows seamlessly. In this sense, SPONGE is indeed a simulation package toward next generation molecular modeling and is expected to be robust, self-learning, and widely applicable.

Supporting Information

The supporting information for this article is available on the WWW under https://doi.org/10.1002/cjoc.202100456.

Acknowledgement

The authors thank the National Key R&D Program of China (2017YFA0204702), the National Natural Science Foundation of China (21821004, 21873007 and 21927901) and CAAI-Huawei MindSpore Open Fund for financial support. This research was supported, in part, through the computational resources provided by the SZBL supercomputing center.

References

[1] Berendsen, H. J. C.; Postma, J. P. M.; Vangunsteren, W. F.; Dinola, A.; Haak, J. R. Molecular-Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

[2] Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald - an N.Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

[3] Huang, X. H.; Bowman, G. R.; Bacallado, S.; Pande, V. S. Rapid equilibrium sampling initiated from nonequilibrium data. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 19765–19769.

[4] Luo, R.; David, L.; Gilson, M. K. Accelerated Poisson-Boltzmann calculations for static and dynamic systems. *J. Comput. Chem.* **2002**, *23*, 1244–1253.

[5] Onufriev, A. V.; Case, D. A. Generalized Born Implicit Solvent Models for Biomolecules. *Annu. Rev. Biophys.* **2019**, *48*, 275–296.

[6] Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C. Numerical-Integration of Cartesian Equations of Motion of a System with Constraints - Molecular-Dynamics of N-Alkanes. *J. Comput. Phys.* **1977**, *23*, 327–341.

[7] Wu, X. W.; Brooks, B. R. Self-guided Langevin dynamics simulation method. *Chem. Phys. Lett.* **2003**, *381*, 512–518.

[8] Plimpton, S.; Hendrickson, B. Parallel Molecular-Dynamics Algorithms for Simulation of Molecular-Systems. *ACS Sym. Ser.* **1995**, *592*, 114–132.

[9] Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187–217.

[10] Berendsen, H. J. C.; Vanderspoel, D.; Vandrunen, R. Gromacs - a Message-Passing Parallel Molecular-Dynamics Implementation. *Comput. Phys. Commun.* **1995**, *91*, 43–56.

[11] Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E.; Debolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. Amber, a Package of Computer-Programs for Applying Molecular Mechanics, Normal-Mode Analysis, Molecular-Dynamics and Free-Energy Calculations to Simulate the Structural and Energetic Properties of Molecules. *Comput. Phys. Commun.* **1995**, *91*, 1–41.

[12] Plimpton, S. Fast Parallel Algorithms for Short-Range Molecular-Dynamics. *J. Comput. Phys.* **1995**, *117*, 1–19.

[13] Rackers, J. A.; Wang, Z.; Lu, C.; Laury, M. L.; Lagardere, L.; Schnieders, M. J.; Piquemal, J. P.; Ren, P. Y.; Ponder, J. W. Tinker 8: Software Tools for Molecular Design. *J. Chem. Theory Comput.* **2018**, *14*, 5273–5289.

[14] Harvey, M. J.; Giupponi, G.; De Fabritiis, G. ACEMD: Accelerating Biomolecular Dynamics in the Microsecond Time Scale. *J. Chem. Theory Comput.* **2009**, *5*, 1632–1639.

[15] Eastman, P.; Friedrichs, M. S.; Chodera, J. D.; Radmer, R. J.; Bruns, C. M.; Ku, J. P.; Beauchamp, K. A.; Lane, T. J.; Wang, L. P.; Shukla, D.; Tye, T.; Houston, M.; Stich, T.; Klein, C.; Shirts, M. R.; Pande, V. S. OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. *J. Chem. Theory Comput.* **2013**, *9*, 461–469.

[16] Kobayashi, C.; Jung, J.; Matsunaga, Y.; Mori, T.; Ando, T.; Tamura, K.; Kamiya, M.; Sugita, Y. GENESIS 1.1: A hybrid-parallel molecular dynamics simulator with enhanced sampling algorithms on multiple computational platforms. *J. Comput. Chem.* **2017**, *38*, 2193–2206.

[17] Zhang, L.; Pardo-Avila, F.; Unarta, I. C.; Cheung, P. P. H.; Wang, G.; Wang, D.; Huang, X. H. Elucidation of the Dynamics of Transcription Elongation by RNA Polymerase II using Kinetic Network Models. *Acc. Chem. Res.* **2016**, *49*, 687–694.

[18] Im, W.; Roux, B. Ion permeation and selectivity of OmpF porin: A theoretical study based on molecular dynamics, brownian dynamics, and continuum electrodiffusion theory. *J. Mol. Biol.* **2002**, *322*, 851–869.

[19] Choi, Y. K.; Cao, Y.; Frank, M.; Woo, H.; Park, S. J.; Yeom, M. S.; Croll, T. I.; Seok, C.; Im, W. Structure, Dynamics, Receptor Binding, and Antibody Binding of the Fully Glycosylated Full-Length SARS-CoV-2 Spike Protein in a Viral Membrane. *J. Chem. Theory Comput.* **2021**, *17*, 2479–2487.

[20] Woo, H.; Park, S. J.; Choi, Y. K.; Park, T.; Tanveer, M.; Cao, Y. W.; Kern, N. R.; Lee, J.; Yeom, M. S.; Croll, T. I.; Seok, C.; Im, W. Developing a Fully Glycosylated Full-Length SARS-CoV-2 Spike Protein Model in a Viral Membrane. *J. Phys. Chem. B* **2020**, *124*, 7128–7137.

[21] Shaw, D. E.; Deneroff, M. M.; Dror, R. O.; Kuskin, J. S.; Larson, R. H.; Salmon, J. K.; Young, C.; Batson, B.; Bowers, K. J.; Chao, J. C.; Eastwood, M. P.; Gagliardo, J.; Grossman, J. P.; Ho, C. R.; Ierardi, D. J.; Kolossvary, I.; Klepeis, J. L.; Layman, T.; McLeavey, C.; Moraes, M. A.; Mueller, R.; Priest, E. C.; Shan, Y. B.; Spengler, J.; Theobald, M.; Towles, B.; Wang, S. C. Anton, a Special-Purpose Machine for Molecular Dynamics Simulation. *Conf. Proc. Int. Symp. C* **2007**, 1–12.

[22] Shaw, D. E.; Grossman, J. P.; Bank, J. A.; Batson, B.; Butts, J. A.; Chao, J. C.; Deneroff, M. M.; Dror, R. O.; Even, A.; Fenton, C. H.; Forte, A.; Gagliardo, J.; Gill, G.; Greskamp, B.; Ho, C. R.; Ierardi, D. J.; Iserovich, L.; Kuskin, J. S.; Larson, R. H.; Layman, T.; Lee, L. S.; Lerer, A. K.; Li, C.; Killebrew, D.; Mackenzie, K. M.; Mok, S. Y. H.; Moraes, M. A.; Mueller, R.; Nociolo, L. J.; Peticolas, J. L.; Quan, T.; Ramot, D.; Salmon, J. K.; Scarpazza, D. P.; Ben Schafer, U.; Siddique, N.; Snyder, C. W.; Spengler, J.; Tang, P. T. P.; Theobald, M.; Toma, H.; Towles, B.; Vitale, B.; Wang, S. C.; Young, C. Anton 2: Raising the bar for performance and programmability in a special-purpose molecular dynamics supercomputer. *Int. Conf. High Perfor.* **2014**, 41–53.

[23] Torrie, G. M.; Valleau, J. P. Non-Physical Sampling Distributions in Monte-Carlo Free-Energy Estimation - Umbrella Sampling. *J. Comput. Phys.* **1977**, *23*, 187–199.

[24] Bussi, G.; Laio, A.; Parrinello, M. Equilibrium free energies from nonequilibrium metadynamics. *Phys. Rev. Lett.* **2006**, *96*, 090601.

[25] 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.

[26] Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

[27] Hansmann, U. H. E. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.* **1997**, *281*, 140–150.

[28] Marinari, E.; Parisi, G. Simulated Tempering - a New Monte-Carlo Scheme. *Europhys. Lett.* **1992**, *19*, 451–458.

[29] Berg, B. A.; Neuhaus, T. Multicanonical Algorithms for 1st Order Phase-Transitions. *Phys. Lett. B* **1991**, *267*, 249–253.

[30] Wang, F. G.; Landau, D. P. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.* **2001**, *86*, 2050–2053.

[31] Zhang, Z. J.; Liu, X. Z. J.; Chen, Z. F.; Zheng, H. F.; Yan, K.; Liu, J. A unified thermostat scheme for efficient configurational sampling for classical/quantum canonical ensembles via molecular dynamics. *J. Chem. Phys.* **2017**, *147*, 034109.

[32] Peng, J. B.; Cao, D. Y.; He, Z. L.; Guo, J.; Hapala, P.; Ma, R. Z.; Cheng, B. W.; Chen, J.; Xie, W. J.; Li, X. Z.; Jelinek, P.; Xu, L. M.; Gao, Y. Q.; Wang, E. G.; Jiang, Y. The effect of hydration number on the interfacial transport of sodium ions. *Nature* **2018**, *557*, 701–705.

[33] Gao, Y. Q. An integrate-over-temperature approach for enhanced sampling. *J. Chem. Phys.* **2008**, *128*, 064105.

[34] Yang, L.; Liu, C.-W.; Shao, Q.; Zhang, J.; Gao, Y. Q. From Thermodynamics to Kinetics: Enhanced Sampling of Rare Events. *Acc. Chem. Res.* **2015**, *48*, 947–955.

[35] You, Z.; Li, L.; Lu, J.; Ge, H. Integrated tempering enhanced sampling method as the infinite switching limit of simulated tempering. *J. Chem. Phys.* **2018**, *149*, 084114.

[36] Yang, L. J.; Gao, Y. Q. A selective integrated tempering method. *J. Chem. Phys.* **2009**, *131*, 214109.

[37] Yin, Y.; Yang, L.; Zheng, G.; Gu, C.; Yi, C.; He, C.; Gao, Y. Q.; Zhao, X. S. Dynamics of spontaneous flipping of a mismatched base in DNA duplex. *Proc. Nat. Acad. Sci. U. S. A.* **2014**, *111*, 8043–8048.

[38] Zhang, J.; Yang, Y. I.; Yang, L.; Gao, Y. Q. Conformational Preadjustment in Aqueous Claisen Rearrangement Revealed by SITS-QM/MM MD Simulations. *J. Phys. Chem. B* **2015**, *119*, 5518–5530.

[39] Zhang, J.; Yang, Y. I.; Yang, L.; Gao, Y. Q. Dynamics and Kinetics Study of "In-Water" Chemical Reactions by Enhanced Sampling of Reactive Trajectories. *J. Phys. Chem. B* **2015**, *119*, 14505–14514.

[40] Zhang, J.; Zhang, Z.; Yang, Y. I.; Liu, S. R.; Yang, L. J.; Gao, Y. Q. Rich Dynamics Underlying Solution Reactions Revealed by Sampling and Data Mining of Reactive Trajectories. *ACS Central Sci.* **2017**, *3*, 407–414.

[41] Lei, Y. K.; Zhang, J.; Zhang, Z.; Gao, Y. Q. Dynamic Electric Field Complicates Chemical Reactions in Solutions. *J. Phys. Chem. Lett.* **2019**, *10*, 2991–2997.

[42] Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.* **2002**, *53*, 291–318.

[43] Groot, R. D.; Warren, P. B. Dissipative particle dynamics: Bridging the

gap between atomistic and mesoscopic simulation. *J. Chem. Phys.* **1997**, *107*, 4423–4435.
[44] Wei, J. C.; Tian, H.; Zhou, R.; Shao, Y. F.; Song, F.; Gao, Y. Q. Topological Constraints with Optimal Length Promote the Formation of Chromosomal Territories at Weakened Degree of Phase Separation. *J. Phys. Chem. B* **2021**, *125*, 9092–9101.
[45] Zhang, J.; Lei, Y. K.; Che, X.; Zhang, Z.; Yang, Y. I.; Gao, Y. Q. Deep Representation Learning for Complex Free-Energy Landscapes. *J. Phys. Chem. Lett.* **2019**, *10*, 5571–5576.
[46] Zhang, J.; Lei, Y. K.; Yang, Y. I.; Gao, Y. Q. Deep learning for variational multiscale molecular modeling. *J. Chem. Phys.* **2020**, *153*, 174115.
[47] Zhang, J.; Yang, Y. I.; Noe, F. Targeted Adversarial Learning Optimized Sampling. *J. Phys. Chem. Lett.* **2019**, *10*, 5791–5797.

Manuscript received: July 13, 2021
Manuscript revised: September 7, 2021
Manuscript accepted: September 17, 2021
Accepted manuscript online: September 22, 2021

