 doi:10.20944/preprints202012.0081.v3

"Dividing and conquering" and "caching" in molecular modeling

Xiaoyong Cao† and Pu Tian*,†,‡

†*School of Life Sciences, Jilin University, Changchun, China 130012*

‡*School of Artificial Intelligence, Jilin University, Changchun, China 130012*

E-mail: tianpu@jlu.edu.cn

Phone: +86 (0)431 85155287

Abstract

Molecular modeling is widely utilized in subjects including but not limited to physics, chemistry, biology, materials science and engineering. Impressive progress has been made in development of theories, algorithms and software packages. To divide and conquer, and to cache intermediate results have been long standing principles in development of algorithms. Not surprisingly, Most of important methodological advancements in more than half century of molecular modeling are various implementations of these two fundamental principles. In the mainstream classical computational molecular science, tremendous efforts have been invested on two lines of algorithm development. The first is coarse graining, which is to represent multiple basic particles in higher resolution modeling as a single larger and softer particle in lower resolution counterpart, with resulting force fields of partial transferability at the expense of some information loss. The second is enhanced sampling, which realizes "dividing and conquering" and/or "caching" in configurational space with focus either on reaction coordinates and collective variables as in metadynamics and related algorithms, or on the transition matrix

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2021 doi:10.20944/preprints202012.0081.v3

and state discretization as in Markov state models. For this line of algorithms, spatial resolution is maintained but results are not transferable. Deep learning has been utilized to realize more efficient and accurate ways of "dividing and conquering" and "caching" along these two lines of algorithmic research. We proposed and demonstrated the local free energy landscape approach, a new framework for classical computational molecular science. This framework is based on a third class of algorithm that facilitates molecular modeling through partially transferable in resolution "caching" of distributions for local clusters of molecular degrees of freedom. Differences, connections and potential interactions among these three algorithmic directions are discussed, with the hope to stimulate development of more elegant, efficient and reliable formulations and algorithms for "dividing and conquering" and "caching" in complex molecular systems.

Introduction

Impact of molecular modeling in scientific research is clearly embodied by the number of publications. Results of a Web of Science (www.webofknowledge.com) search with various relevant key words is listed in Table 1. However, despite widespread applications, we remain far from accurately predicting and designing molecular systems in general. Further methodological development is highly desired to tap its full potential. Historically, molecular modeling has been approached from a physical or application point of view, and numerous excellent reviews are available in this regard.[1–16] From an algorithmic perspective, "dividing and conquering" (DC) and "caching" intermediate results that need to be computed repetitively are two fundamental principles in development of many important algorithms (e.g. dynamic programming[17]). As a matter of fact, the major focus of modern statistical machine learning is to learn ("caching" relevant information) and then carry out inference on top of which.[18] In this review, we provide a brief discussion of important methodological development in molecular modeling as specific applications of these two principles. The content will be organized as the following. Part II describes fundamental challenges in molecular

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2021 doi:10.20944/preprints202012.0081.v3

modeling; Part III summarizes application of these two fundamental algorithmic principles in two lines of methodological research, coarse graining (CG)[18–27] and enhanced sampling (ES);[28–31] Part IV covers how machine learning, particularly deep learning, facilitate DC and "caching" in CG and ES,[30,32–36] part V introduces local free energy landscape (LFEL) approach, a new framework for computational molecular science based on partially transferable in resolution "caching" of local sampling, and the first implementation of this new framework by generalized solvation free energy (GSFE) theory[37] is briefly discussed; and part VI discusses connections among these three lines of algorithmic development, their specific advantages and prospective explorations. Due to the large body of literature and limited space, we apologize to authors whose excellent work are not cited here.

Table 1: Number of publications from web of science search on Sep, the 8th, 2020

Key words	Number of publications
Molecular dynamics simulation	241,748
Monte Carlo simulation	189,550
QM-MM (quantum mechanical - molecular mechanical) simulation	9907
Dissipative particle dynamics simulation	3693
Langevin dynamics simulation	3893
Molecular modeling	2,072,091
All of the above	2,243,182

Challenges in molecular modeling

Accurate description of molecular interactions

Molecular interactions may be accurately described with high level molecular orbital theories (e.g. coupled cluster theory[38,39]) or sophisticated density functionals combined with large basis sets.[40–44] However, such quantum mechanically detailed computation is prohibitively expensive for any realistic complex molecular systems. Molecular interactions are traditionally represented by explicit functions and pairwise approximations as exemplified by typical

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2021 doi:10.20944/preprints202012.0081.v3

physics based atomistic molecular mechanical (MM) force fields (FF):[45–48]

$$
\begin{aligned}
U(\vec{R}) = & \sum_{bonds} K_b(b-b_0)^2 + \sum_{angles} K_\theta(\theta-\theta_0)^2 \\
& + \sum_{dihedral} K_\chi(1+cos(n\chi-\delta)) \\
& + \sum_{impropers} K_{imp}(\phi-\phi_0)^2 \\
& + \sum_{nonbonded} \left(\epsilon_{ij} \left[\left(\frac{Rmin_{ij}}{r_{ij}} \right)^{12} - \left(\frac{Rmin_{ij}}{r_{ij}} \right)^{6} \right] \right) + \frac{q_i q_j}{\epsilon r_{ij}}
\end{aligned} \tag{1}
$$

or knowledge based potential functions:[49–51] These simple functions, while being amenable to rapid computation and are physically sound grounded near local energy minima (e.g. harmonic behavior of bonding, bending near equilibrium bond lengths and bend angles), are problematic for anharmonic interactions, which are very common in many molecular systems.[52] It is well understood that properly parameterized Lennard-Jones potentials are accurate only near the bottom of its potential well. Frustration are ubiquitous in biomolecular systems and are likely fundamental driving force for conformational fluctuations.[53–55] One may imagine that a molecular system with all its comprising particles at their respective "happy" energy minima positions would likely be a stable "dead" molecule, which may be a good structural support but is likely not able to provide dynamic functional behavior.

Pairwise approximations are usually adopted for its computational convenience, both in terms of dramatically reduced computational cost and tremendously smaller (when compared with possible many body potentials) number of necessary parameters to be fit in FF parameterization. It is widely acknowledged that construction of traditional FF (e.g. equation 1) is a laborious process. Development of polarizable[9,56,57] and more complex FF with larger parameter sets[58] alleviate some shortcomings of earlier counterpart. Expansion based treatments were incorporated to address anharmonicity.[59] However, to tackle limitation of explicit simple functional form and pairwise approximation for better description of molecular interactions remain challenges to be met for molecular modeling community.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2021 doi:10.20944/preprints202012.0081.v3

Additionally, even atomistic simulations are prohibitively expensive for large biomolecular complexes at long time scales (e.g. milliseconds and beyond).[60–62]

Inherent low efficiency in sampling of configurational space

Complexity of molecular systems is rooted in their molecular interactions, which engenders complex and non-linear correlations among molecular degrees of freedom (DOFs). Consequently, effective number of DOFs are greatly reduced. Therefore, complex molecular systems are confined to manifolds[63] of much lower dimensionality with near zero measure in corresponding nominal high dimensional space (NHDS). Consequently, sufficient brute force random sampling in NHDS of interested molecular systems is hopeless.

In stochastic trajectory generation by Monte Carlo (MC) simulations or candidate structural model proposal in protein structure prediction and refinement (or other similar scenarios), new configuration proposal are carried out in NHDS, a lot of effort is inevitably wasted due to sampling outside the actual manifold occupied by the target molecular system. Such wasting may be avoided if we understood all correlations. However, understanding all correlations implicates accurate description of global free energy landscape (FEL) and there is no need to investigate it further! Due to preference of lower energy configurations by typical importance sampling strategies (e.g. Metropolis MC), stochastic trajectories tend to be trapped in local minima of FEL, this is especially true for complex molecular (e.g. biomolecular) systems which have hierarchical rugged FEL with many local minima.[64,65] In trajectory generation by molecular dynamics (MD) simulations, configurational space is explored by laws of classical mechanics and no wasting due to random moves exists. However, molecular systems may well drift away from their true manifolds due to insufficient accuracy of FF. Similar to stochastic trajectory generation, it takes long simulations to map FEL since molecular system tend to staying at any local minimum, achieve equilibrium among many local minima is just as challenging as in the case of stochastic counterpart.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2021 doi:10.20944/preprints202012.0081.v3

DC and "caching" in traditional molecular modeling

To cope with fundamental difficulties in molecular modeling, two distinct lines of methodological development (CG and ES) based on DC and "caching" strategies have been conducted and tremendous progress has been made in understanding of molecular systems. As summarized below:

Coarse graining, a partially transferable "caching" strategy

Atomistic FF parameterization is the most well established coarse graining with a strong theoretical foundation, the Born-Oppenheimer approximation. Theoretically, MMFF are potential of mean force (PMF) obtained by averaging over many electronic DOFs for given atomic configurations. In practice, due to the fact that *ab initio* calculations are expensive and may have significant error when level of theory (and/or basis set) is not sufficient, reference data usually include results from both quantum mechanical (QM) calculations and well-established experimental data.[66–68] The DC strategy is utilized by selecting atomic clusters of various size to facilitate generation of QM reference data. The essential information learned from reference data is then permanently and approximately "cached" in FF parameters through the parameterization process. Due to the separation of time scales between electronic and atomic motion, elimination of electronic DOFs is straight forward but comes with the price of incapability in describing chemical reactions. To harvest benefits of both quantum and atomistic simulations, a well-established DC strategy is to treat a small region involved interested chemical reaction at QM detail and its surrounding with MMFF.[69–74] This series of pioneering work was awarded Nobel prize in 2013, and QM-MM treatment continues to be the mainstream methodology for computational description of chemical reactions.[75,76]

The united atom model (UAM) is the next step in coarse graining,[77] where hydrogen atoms are merged into bonded heavy atoms. This is quite intuitive since hydrogens have

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2021 doi:10.20944/preprints202012.0081.v3

much smaller mass on the one hand, and are difficult to see by experimental detection techniques utilizing electron diffraction (e.g. X-ray crystallography) on the other hand. For both polymeric and biomolecular systems, UAM remains to be expensive for many interested spatial and temporal scales. Therefore, further coarse graining in various forms have been constructed. As a matter of fact, CG is usually used to denote modeling with particles that representing multiple atoms in contrast to atomistic simulations, and the same convention will be adopted in the remaining part of this review unless stated otherwise. Both "Top-down" (that based on reproducing experimental data) and "bottom-up" (that based on reproducing certain properties of atomistic simulations) approaches are utilized.[21,78] For polymeric materials, beads are either utilized to represent monomers or defined on consideration of persistent length,[79] and dissipative particle dynamics (DPD) were proposed to deal with complexities arise from much larger particles.[80] For biomolecular systems, a wide variety of coarse grained models have been developed.[20,21,23,24,81,82] Another important subject of CG methodology development is materials science.[83,84] Earlier definition of CG particles are rather *ad hoc*.[20] More formulations with improved statistical mechanical rigor appeared later on,[22] with radial distribution function based inversion,[79,85–87] entropy divergence[88] and force matching algorithm[89–91] being outstanding examples of systematic development. Present CG is essentially to realize the following mapping as disclosed by equation (4) in ref.[22] :

$$exp\left[-\beta V_{CG}(\mathbf{R}_{CG})\right] \equiv \int d\mathbf{r}\delta(M_R(\mathbf{r}) - \mathbf{R}_{CG}) exp\left[-\beta V(\mathbf{r})\right] \tag{2}$$

with $\mathbf{r}$ and $\mathbf{R}$ being coordinates in higher resolution and CG coordinates, $M_R(\mathbf{r})$ being the map operator from $\mathbf{r}$ to $\mathbf{R}$, V and V_{CG} being potential energy of higher resolution and CG representation respectively. Due to lack of time scale separation (see Fig. 1) for essentially all CG mapping, strict realization of this equation/mapping is not rigorously possible. A naive treatment of CG particles as basic units (with no internal degrees of freedom) would result in wrong thermodynamics.[22] Due to corresponding significant loss of information, it is not

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2021 doi:10.20944/preprints202012.0081.v3

possible to develop a definition of CG and corresponding FF parameterization for comprehensive reproduction of atomistic description of corresponding molecular systems. Different coarse graining have distinct advantages and disadvantages, so choosing proper CG strategy is highly dependent upon specific goal in mind. CG particles are usually isotropic larger and softer particles with pairwise interactions, or simple convex anisotropic object (e.g. soft spheroids) that may be treated analytically.[24,92–94] Such simplifications provide both convenience of computation and certain deficiency for capturing physics of target molecular systems. CG may be carried out iteratively to address increasingly larger spatial scales by "caching" lower resolution CG distributions with ultra CG (UCG) FF.[22,95–99] Pairwise approximation and explicit simple function form remain to be limitations of interaction description for traditional CG FF. When compared with atomistic FF, pairwise approximation deteriorate further due to lack of time scale separation (Fig. 1).

Another simple and powerful type of CG model for biomolecular systems is Gō model[100,101] and elastic network models (ENM)[102–104] or gaussian network models (GNM)[105,106] with native structure being defined as the equilibrium state, and with quadratic/harmonic interactions between all residues within given cutoff. Only a few parameters (e.g. cutoff distance, spring constant) are needed. Such models "caching" the experimental structures and are proved to be useful in understanding major conformational transitions and slow dynamics of many biomolecular systems.[107,108]

Enhanced sampling, a nontransferable in resolution DC and "caching" strategy

Umbrella sampling (US)[109] is probably the first combination of DC and "caching" strategy for better sampling of molecular system along a given reaction coordinate (RC) (or order parameter) s. DC strategy is first applied by dividing s into windows, information for each window is then partially "cached" by corresponding bias potentials and local statistics. Later on, adaptive US (AUS)[110,111] and weighted histogram analysis method (WHAM)[112] was de-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2021 doi:10.20944/preprints202012.0081.v3

veloped to improve both efficiency and accuracy. MBAR[113,114] was developed to achieve error bound analysis which is lack in WHAM. Further development including adaptive bias force (ABF)[115–117] and metadynamics.[118–120] Details of these methodologies were well explained by excellent reviews.[121–124] The common trick to all of these algorithms (and their variants) is to "cache" visited configurational space with bias potentials/force and local statistics, thus dramatically accelerate sampling of interested rare events. Denote CV as $\mathbf{s}(\mathbf{r})$ ($\mathbf{r}$ being physical coordinates of atoms/particles in the target molecular system), equilibrium distribution and free energy on the CV may be expressed as:[30]

$$p_0(\mathbf{s}) = \int d\mathbf{r}\delta\left[\mathbf{s} - \mathbf{s}(\mathbf{r})\right] p_0(\mathbf{r}) = \langle \delta\left[\mathbf{s} - \mathbf{s}(\mathbf{r})\right] \rangle \tag{3}$$

$$p_0(\mathbf{r}) = \frac{e^{-\beta U(\mathbf{r})}}{\int dr e^{-\beta U(\mathbf{r})}} \tag{4}$$

$$F(\mathbf{s}) = -\frac{1}{\beta} log[p_0(\mathbf{s})] \tag{5}$$

$$F(\mathbf{s}) = -\frac{1}{\beta} log[p(\mathbf{s})] - V(\mathbf{s}) \tag{6}$$

with $p(\mathbf{s})$ being the sampled distribution in simulation with corresponding bias potential $V(\mathbf{s})$ for "caching" of visited configurational space.

The starting point of these "caching" algorithms is specification of reaction coordinates (RC) or collective variables (CVs), which is a very challenging task for complex molecular systems in most cases. Traditionally, principle component analysis (PCA)[125] is the most widely utilized and a robust way for disclosing DOFs associated with the largest variations. To deal with ubiquitous nonlinear correlations, kernels are often used albeit with the difficulty of choosing proper kernels.[126] Additional methodologies, include multidimensional scaling (MDS),[127] isomap,[128] locally linear embedding (LLE),[129] diffusion map[130,131] and sketch map[132] have been developed to map out manifold for high dimensional data. However, each has it own limitations. For example, LLE[129] is sensitive to noise and therefore has difficulty with molecular simulation trajectories which are quite noisy; Isomap[128] requires relatively

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2021 doi:10.20944/preprints202012.0081.v3

homogeneously sampled manifold to be accurate. Both LLE and Isomap do not provide explicit mapping between molecular coordinates and CVs; Diffusion and sketch maps are likely to be more suitable to analyze molecular simulation trajectories. Nonetheless, their successful application for large and complex molecular systems remain to be tested. All of above non-linear mapping algorithm are mainly suitable for manifold on a single scale. When we are interested in finding paths for transitions among known metastable states, transition path sampling (TPS)[133–135] methodology maybe utilized to establish CV.

Apparently, RC and/or CV based ES is a different path for facilitate simulation of complex molecular systems on longer time scales from coarse graining. One apparent plus side is that these algorithms are "in resolution" as no systematic discarding of molecular DOFs occur. With specification of RC and/or CVs, computational resource is directed toward the presumably most interesting dynamics of the target molecular system, and RC and/or CV maybe repetitively refined to obtain mechanistic understanding of interested molecular processes. However, the down side is that "cached" information on local configurational space is not transferable to other similar molecular systems. While rigorous transferability may not be easily established for any CG FF, practical utility of CG FF for molecular systems with similar composition and thermodynamic conditions have been quite common and useful.[24] Therefore, CG FF may be deemed as partially transferable.

An important recent development of DC strategy for enhanced sampling is Markov state models (MSM),[136–139] one great advantage of which is that no RC or CV is needed. Instead, it extracts long-time dynamics from independent short trajectories distributed in configurational space. Many important biomolecular functional processes have been characterized with this great technique.[140–142] The most fundamental assumption is that all states for a target molecular system form an ergodic Markov chain:

$$\pi(t+\tau) = \pi(t)\mathbf{P} \tag{7}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2021 doi:10.20944/preprints202012.0081.v3

with $\pi(t)$ and $\pi(t+\tau)$ being a vector of probabilities for all states at time t and $t+\tau$ respectively. $\mathbf{P}$ is the transition matrix with its element P_{ij} being probability of the molecular system being found in state j after an implied lag time (τ) from the previous state i. Apparently as t goes to infinity for an equilibrium molecular system, a stationary distribution π will arise as defined below:

$$\pi = \pi P \tag{8}$$

The advantage of not needing RC/CV does not come for free but with accompanying difficulties. Firstly, one has to distribute start point of trajectories to statistically important and different part of configurational space, then select proper (usually hierarchical, with each level of hierarchy corresponds to a specific lag time) partition of configurational space into discrete states. This is the key step of DC strategy in MSM. No formal rule is available and experience is important. In many cases some try and error is necessary. Secondly, within each discrete state at a given level of hierarchy, equilibration is assumed to be achieved instantly and this assumption causes systematic discretization error, which fortunately may be controlled with proper partition and sufficiently long lag time.[143] Apparently, metastable states obtained from MSM analysis is molecular system specific and thus not transferable.

Another important class of enhanced sampling is to facilitate sampling with non-Boltzmann distributions and restore property at targeted thermodynamic condition through proper reweight.[144] Most outstanding examples are Tsallis statistics,[145,146] parallel tempering,[147,148] replica exchange molecular dynamics,[149,150] Landau-Wang algorithm[151] and integrated tempering sampling (ITS).[152–154] These algorithms are not direct applications of DC and "caching" strategies and are not discussed further here.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2021 doi:10.20944/preprints202012.0081.v3

Machine learning improves "caching"

Toward *ab initio* accuracy of molecular simulation potentials

Fixed functional form and pairwise approximation of non-bonded interactions are two major factors limiting the accuracy of molecular interaction description in both atomistic and CG FF. Neural network (NN) has capability of approximate arbitrary functions and therefore has the potential to address these two issues. Not surprisingly, significant progress has been made in this regard as summarized by recent excellent reviews.[155–160] Cutoff and attention to local interactions remains the DC strategy for development of machine learning potentials. The major improvement over traditional FF is better "caching" that overcomes pairwise approximation and fixed functional form limitations. NN FF naturally tackle both issues as explicit functions are not necessary since NNs are universal approximators. The significance of many-body potentials[161] and extent of pairwise contributions were analyzed.[162,163] There are also efforts to search for proper simple functional forms, which are expected to be more accurate than present functional forms in traditional FF on the one hand, and alleviate overfitting/generalization difficulty and reduce computational cost of complex NN FF on the other hand,[164,165] especially when training dataset is small. While most machine learning FF are trained by energy data,[155,159] gradient-domain machine learning (GDML) approach[166] directly learns from forces and realizes great savings of data generation.

Just as in the case of traditional FF, transferability and accuracy is always a tradeoff. More transferability implicates less attention is paid to "cache" detailed differences among different molecular systems, hence less accuracy. Exploration in this regard, however, remains not as much as necessary.[167–169] Unlike manual fitting of traditional FF, systematic investigation of tradeoff strategy is potentially feasible for machine learning fitting,[170] and yet to be done for many interesting molecular systems. With expediency of NN training, development of a NN FF hierarchy with increasing transferability/accuracy and decreasing accuracy/transferability is likely to become a pleasing reality in the near future. Rapid fur-

 doi:10.20944/preprints202012.0081.v3

ther development of machine learning potentials, particularly NN potentials, are expected. However, significant challenges for NN potentials remain on better generalization capability, description/treatment of long range interactions,[171,172] wide range of transferability,[173] faster computation[174] and proper characterization of their error bounds. Should further significant progress be made on these issues, it is promising we may have routine molecular simulations with both classical efficiency and *ab initio* accuracy in the near future.

Machine learning and coarse graining

As in the case of constructing atomic level potentials, machine learning has been applied to address two outstanding pending issues in coarse graining, which are definition of CG sites/particles and parameterization of corresponding interactions between/among these sites/particles. Traditional CG FF, suffers from both pairwise approximation and accuracy ceiling of simple fixed functional forms which are easy to fit. By using more complex (but fixed functional form) potentials with a machine learning fitting process, Chan *et. al.*[175] developed ML-BOP CG water model with great success. Deep neural network (DNN) was utilized to facilitate parameterization of CG potentials when given radial distribution functions (RDF) from atomistic simulations.[176] CGnet demonstrated great success with simple model systems (alanine dipeptide).[177] DeePCG model was developed to overcome pair approximation and fixed functional form and demonstrated with water.[178] Using oxygen site to represent water is rather intuitive. However, for more complex biomolecules such as proteins, possibility for selection of CG site explodes. To improve over intuitive or manual try and error definition of CG sites, a number of studies have been carried out[179–182] to provide better and faster options for choosing CG sites. However, no consensus strategy is available up to date and more investigations are desired. The fundamental difficulty is that there is no sufficient time scale separation between explicit CG DOFs and discarded implicit DOFs, regardless of specific selection scheme being utilized. Intuitively, one would expect CG FF parameters to be dependent upon definition of CG sites/particles. In this re-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2021 doi:10.20944/preprints202012.0081.v3

gard, auto-encoders were utilized to construct a generative framework that accomplishes CG representation and parameterization in a unified way.[34] The spirit of generative adversarial networks was utilized to facilitate CG construction and parameterization, particularly with virtual site representation.[183] It was found that description of off-target property by CG exhibit strong correlation with CG resolution, to which on-target property being much less sensitive.[184] Such observation suggests that adjust CG for specific target properties might be a better strategy than searching for a single best CG representation. Despite potentially more severe impact of pairwise approximation for CG FF than in atomistic FF, quantitative analysis in this regard remain to be done.

Machine learning in searching for RC/CVs and construction of MSM

To overcome difficulties of earlier nonlinear CV construction algorithms[128–130,132] and to reduce reliance on human experience, auto-encoders, which is well-established for trainable (non-linear) dimensionality reduction, are utilized in a few studies.[185–188] Chen and Ferguson[186] first utilized autoencoders to learn nonlinear CVs that are explicit and differentiable functions of molecular coordinates, thus enabling direct utility in molecular simulations for more effective exploration of configurational space. Further improvement[185] was achieved through circular network nodes and hierarchical network architectures to rank-order CVs. Wehmeyer and Noé[187] developed time-lagged auto-encoder to search for low dimensional embeddings that capture slow dynamics. Ribeiro *et. al.*[188] proposed the reweighted autoencoded variational Bayes to iteratively refine RC and demonstrated in computation of the binding free energy profile for a hydrophobic ligand-substrate system. Building a MSM for any specific molecular system requires tremendous experience and many steps in process are error prone. To overcome these pitfalls, VAMPnet that based on variational approach for Markov process was developed to realize the complete mapping steps from molecular trajectories to Markov states.[189]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2021 doi:10.20944/preprints202012.0081.v3

The local free energy landscape approach

Both CG and ES methodologies facilitate molecular simulation by effectively reducing local sampling. In CG, it is realized through "caching" (integration) of distributions for faster/discarded DOFs with proper CG FF, and thus has the inevitable cost of losing resolution (information), accompanied by the desired attribute of (partial) transferability to various extent. ES reduces lingering time of molecular systems in local minima through "caching" visited local configurational space, which is usually defined by relevant DC strategies, with biasing potentials. When compared with CG, there is no resolution loss. However, "cached" manifold of configurational space is molecular process specific and thus not transferable at all. In molecular modeling community, these two lines of methodologies are developed quite independently. Nevertheless, one might want to ask why not have both advantages in one method, that is to reduce repetitive local sampling without loss of resolution and with "cached" results being partially transferable. The local free energy landscape (LFEL) approach[190] is proposed with this intention in mind. Historically, parameterization of FF by coarse graining has been the only viable framework due to two fundamental constraints. Firstly, in earlier days of molecular modeling, typical computers have memory space of megabytes or less, render it impossible to accommodate millions or more parameters needed to fit complex LFEL; Secondly, while both neural network and autodifferentiation were invented decades ago, the computational molecular science community did not master these techniques for fitting large number of parameters efficiently until recently. With these two constraints removed, possibility for alternative path arise to break monopoly of classical molecular modeling by FF parameterization via coarse graining. Specifically, one may carry out direct fitting of LFEL and all important information on local distributions of molecular DOFs obtained from expensive local sampling may be "cached". This is in strong contrast to coarse graining based parameterization, in which local distributions are substituted by averaging in relevant lower dimensional space projection (e.g. pairwise distances among CG sites). However, it is essential to assemble LFEL and construct FEL of the interested molec-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2021 doi:10.20944/preprints202012.0081.v3

ular system, and this is the core of the LFEL approach. For a molecular system with N DOFs, this LFEL approach may be expressed as:

$$P(r_1, r_2, \cdots, r_N) = P(R_1, R_2, \cdots, R_M)(M \leq N) \tag{9}$$

$$R_i = (r_{i_1}, r_{i_2}, \cdots, r_{i_l}) \tag{10}$$

$$\begin{aligned} & P(R_1, R_2, \cdots, R_M) \\ & = \prod_{i=1}^{M} P(R_i) \frac{P(R_1, R_2, \cdots, R_M)}{\prod_{i=1}^{M} P(R_i)} \qquad (11) \\ & \approx \prod_{i=1}^{M} P(R_i), \text{and sampling all } R\text{s with mediated GCF} \qquad (12) \end{aligned}$$

$$\begin{aligned} G &= -k_B T ln P(r_1, r_2, \cdots, r_N) \\ &= -k_B T ln P(R_1, R_2, \cdots, R_M) \\ &\approx -k_B T \sum_{i}^{M} ln P(R_i), \text{and sampling all } R\text{s with mediated GCF} \end{aligned} \tag{13}$$

a N-DOF molecular system is reorganized into M overlapping regions (Equation 9), each region has some number of DOFs (Equation 10). The key step of LFEL approach is expressed in Equation 11, in which the first product term (addressed as "local term(s)" hereafter) treat M regions as if they were independent, and all correlations among different regions are incorporated by the fraction term, which is termed global correlation fraction (GCF) and is extremely difficult, if ever possible, to be calculated directly. However, GCF is a unnormalized probability distribution, when all molecular DOFs in local terms are (approximately) sampled according to GCF, then we do not need GCF explicitly anymore (Equations 12 and 13). GCF represents two types of global correlations. The first type is mediated correlations among different regions by the fact that they overlap, and relevant molecular DOFs in such

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2021 doi:10.20944/preprints202012.0081.v3

overlapping space shared by different regions should have exact same state for all concerning regions. The second type is direct global correlations among molecular DOFs in different regions caused by genuine long-range interactions (e.g. electrostatic interactions). Satisfying the first type with sampling is trivial, and ensuring all overlapping regions share the exact same state is sufficient (Equations 12 and 13). The second type of global correlations need more involved treatment. These equations are apparently of general utility for any multiple-variable (high-dimensional) problem. In the specific case of a complex molecular system, using one set of coordinates realizes the mediated contribution of GCF. The approximation in Equation 12 is made by ignoring the second type of global correlations. Free energy minimization of a molecular system in thermodynamic equilibrium may be treated as maximization of joint probability (eq. 13). For molecular systems (or biological systems) off equilibrium, the joint distribution remains our focus despite free energy is not well defined anymore. A schematic representation of the LFEL approach in contrast to FF framework is shown in (Fig.2). While we only demonstrated GSFE implementation of LFEL at residue level for protein structural refinement. LFEL approach may be utilized to "cache" local distributions at any spatial scales. Just as there are many methodological developments in the mainstream FF framework, there are certainly many possible ways to develop algorithms in the LFEL approach. We explored a first step toward this direction through a neural network implementation of the generalized solvation free energy (GSFE) theory.[37] In GSFE theory, each comprising unit in a complex molecular system is solvated by its neighboring units. Therefore, each unit is both a solute itself and a comprising solvent unit of its solvent units. Let $(x_i, y_i) = R_i$ denote a region i defined by a solute x_i and its solvent y_i, a molecular system of N units has N overlapping regions. Each local term may be further expanded:

$$\begin{aligned} P(R_i) &= P(x_i, y_i) \\ &= P(x_i|y_i)P(y_i) \end{aligned} \tag{14}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2021 doi:10.20944/preprints202012.0081.v3

Both terms may be learned from either experimental or computational datasets, as long as they are sufficiently representative and reliable. The first term in Equation 14 is the likelihood term when x_i is the given, it quantifies the extent of match between the solute x_i and its solvent y_i. The second term is the local prior term, it quantifies the stability of the solvent environment y_i. Computation of the prior term is more difficult than the likelihood term, but certainly learnable when sufficient data is available. A local maximum likelihood approximation of GSFE (LMLA-GSFE) is to simply ignore local prior terms.

A particular implementation of the LMLA-GSFE for protein structure refinement with residues defined as comprising unit was conducted.[190] In this scheme, GSFE is integrated with autodifferentiation and coordinate transformation to construct a computational graph for free energy optimization. With fully trainable LFEL derived from backbone and C_β atom coordinates of selected experimental protein structures, we achieved superb efficiency and competitive accuracy when compared with state of the art atomistic protein refinement refinement methodologies. With our newly developed pipeline, refinement of typical protein structure decoys (within 300 amino acids) takes a few seconds on a single CPU core, in contrast to a few hours by typical efficient sampling/minimization based algorithms (e.g. FastRelax[191]) and thousands of hours for MD based refinement.[192] In the latest CASP14 refinement contest (predictioncenter.org/casp14/index.cgi), our method ranked the the first for the 13 targets with start GDT-TS score larger than 60. We expect incorporation of complete heavy atom information and local prior terms to further improve this method in the future. GSFE theory in particular and the LFEL approach in general, are certainly extendable to modeling of other soft matter molecular systems.

More on connections among CG, ES and LFEL approach

All of these algorithms have a common goal of accelerating computation of a joint distribution for a given molecular system at some target resolution, albeit from distinct perspectives. The

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2021 doi:10.20944/preprints202012.0081.v3

fundamental underpinning is the fact that molecular correlations among its various DOFs limit a molecular system to a manifold of significantly lower dimension. Both ES and CG in the FF framework and the LFEL approach are distinct strategies to "cache" manifolds from either configurational space (Fig. 3) or physical space perspective (Fig. 4). Commonality and differences of these strategies are summarized in Table 2 and discussed below.

Table 2: Commonality and difference among three types of algorithms

Algorithm	Coarse graining	Enhanced sampling	LFEL approach
Resolution	Lower	In	In
Transferable?	Partial	No	Partial
Dividing space	Physical	Configurational	Physical
Free energy unit	Partially Specified	Specified	Arbitrary

Both MSM and RC/CV based ES are designed to first describe local parts of the approximate manifold in the configurational space formed by all molecular DOFs of the target molecular system. Information for such local configurational space is partially "cached" either as bias potentials or transition counts, which are further processed to map FEL and dynamics of interested molecular processes. Computational process (or educated guess) for establishment of RC/CVs is essentially "caching" results from sampling/guessing local parts of the configurational as approximate relevant manifold(Fig. 3B). Subsequent sampling along RC/CV is hoped to disclose our interested molecular processes (e.g. biomolecular conformational transitions, substrate binding/release in catalysis). Involved molecular DOFs for RC/CVs are not necessarily spatially adjacent on the one hand, and may be different for different molecular processes of the same molecular system. Apparently, RC/CVs are molecular process specific and not transferable, even among different molecular process of the same molecular system. Nonetheless, the methodology for searching CVs may be applied to many different molecular processes/systems.

In contrast, both CG in the FF framework and the LFEL approach are motivated to "cache" relevant information on the complete configurational distribution for local clusters of molecular DOFs. Such local clusters are building blocks for many similar molecular systems

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2021 doi:10.20944/preprints202012.0081.v3

(e.g. AAs in protein molecular systems) and consequently have limited and approximate transferability. In CG, strongly correlated local clusters of molecular DOFs are represented as a single particle, complex many body correlations/interactions of CG particles within selected cutoff distances are represented by simplified CG FF in a lower resolution and longer range correlations/interactions are incorporated either through more coarser CG models or by separate long-range interaction computation. In LMLA-GSFE implementation of LFEL, all complex many body correlations within selected regions (i.e. each solute and its specific solvent) are decomposed into two terms in Equation 14, local likelihoods and local priors in the same resolution, with local priors and direct genuine long-range interactions simply ignored, and LFEL being approximated by local likelihood terms. More and better ways for implementing LFEL are expected in the future.

The first step of CG is to partition atoms/particles of high resolution representation into highly correlated local clusters that will be represented by corresponding single CG particles, and moderately correlated regions define interaction cutoff for CG particles; The second step is to select a site (usually one of the comprising high resolution particles) to represent the corresponding highly correlated cluster; The third step is to select functional forms to describe molecular interactions among newly defined CG particles, and parameters are optimized by selected loss functions (e.g. differences of average force in force matching[89,90]) based on sampling in the whole configurational space of molecular systems and hopefully to be transferable to some extent. One may imagine that both best clustering and optimal representation sites of clusters may vary with different functional forms used to describe CG particle interactions and in different part of configurational space. Neural network based CG potentials do not have limitation of fixed functional form and pairwise approximations. However, the need to partition molecular systems into transferable clusters and to specify representation site/particle remain. For all different forms of CG, the fundamental essence is to "cache" many body potential of mean force (PMF) in simplified CG FF at a lower resolution. In contrast, LFEL approach is to first using a DC strategy to divide molecular systems

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2021 doi:10.20944/preprints202012.0081.v3

into local regions, then directly "cache" many body PMF (or LFEL) of such local regions in the original resolution. The cached complex local multivariate distributions in NN are subsequently utilized to construct FEL of target molecular system through dynamic puzzle assembly based on sampling with GCF as expressed in Equations 12 and 13. In language of statistical machine learning. Training of LFEL is the learning step, while construction of global FEL is the inference step. The advantage of CG is a simpler resulting physical model, but is inflexible due to fixed clustering and representation on the one hand, and lost resolution/information on the other hand. Properly implemented LFEL while has selected spatial regions comprising many molecular DOFs, composition of such regions are fully dynamic. For example, in GSFE implementation of LFEL, a region is defined by a solute unit and all of its solvent units, and comprising units for the solvent is dynamically updated in each iteration of free energy optimization. Additionally, no loss of resolution is involved for LFEL approach. Hence all difficulties and uncertainties associated with molecular DOF partition, CG site selection and time scale separation, all of which apparently limit transferability of CG FF, disappear. Correspondingly, the extent of transferability of a LFEL model is in principle at least no worse than CG FF. Differences of CG and specific implementation of LFEL by GSFE theory is schematically illustrated in Fig. 4. The superior efficiency of LFEL approach comes with a price. The assembled global FEL has arbitrary unit for two reasons. Firstly, it is extremely difficult to obtain the partition function (normalization constant) for local regions directly during the training/caching stage, therefore we effectively obtain the LFEL up to an unknown constant. Secondly, for two different molecular systems, the number of local regions are usually different and so is the corresponding normalization constant.

These three lines of algorithms may be combined to facilitate molecular modeling. For example, one might first utilize deep learning based near quantum accuracy many body FF to perform atomistic simulations for protein molecular systems, and then extracting local distributions properly with some form of LFEL, which may potentially be utilized to analyze protein molecular systems with near-quantum accuracy and at regular amino-acid based CG

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2021 doi:10.20944/preprints202012.0081.v3

or even much faster speed! Similarly, one may extract and "cache" large body of information from residue level CG simulations with proper LFEL implementation, which may be utilized to achieve ultra CG (UCG) efficiency with residue resolution. Application of CV and MSM based ES algorithm for CG models is straight forward. Combination of LFEL with CV or MSM based ES is more subtle and yet to be investigated.

Conclusions and prospect

The application of "dividing and conquering" and "caching" principle in development of molecular modeling algorithms is briefed. Historically, coarse graining and enhanced sampling have been two independent lines of methodological development in the mainstream FF framework. While they share the common goal of reducing local sampling, the formulations are completely different with distinct (dis)advantages. Coarse graining obtains partial transferable FF but loses resolution, enhanced sampling retains resolution but results are not transferable. The LFEL approach suggests a third strategy to directly approximate global joint distribution by superposition of LFEL, which may be learned from available dataset of either experimental or computational origin. Through integration of coordinate transformation, autodifferentiation and neural network implementation of GSFE, our recent work of protein structure refinement demonstrated that simultaneous realization of transferable in-resolution "caching" of local sampling is not only feasible, but also highly efficient due to replacement of local sampling by differentiation. It is hoped that this review stimulates further development of better "dividing and conquering" strategies for complex molecular systems through more elegant, efficient and accurate ways of "caching" potentially repetitive computations in molecular modeling at various spatial and temporal scales. With diverse molecular systems (e.g. nanomaterials, biomolecular systems), specialization of methodology is essential to take advantage of distinct constraints and characteristics.

Acknowledgement

This work has been supported by the National Key Research and Development Program of China (2017YFB0702500) and by National Natural Science Foundation of China under grant number 31270758.

References

(1) Chodera, J. D.; Mobley, D. L.; Shirts, M. R.; Dixon, R. W.; Branson, K.; Pande, V. S. Alchemical free energy methods for drug discovery: progress and challenges. *Current Opinion in Structural Biology* **2011**, *21*, 150 – 160.

(2) Dror, R. O.; Dirks, R. M.; Grossman, J.; Xu, H.; Shaw, D. E. Biomolecular Simulation: A Computational Microscope for Molecular Biology. *Annual Review of Biophysics* **2012**, *41*, 429–452, PMID: 22577825.

(3) van der Kamp, M. W.; Mulholland, A. J. Combined Quantum Mechanics/Molecular Mechanics (QM/MM) Methods in Computational Enzymology. *BIOCHEMISTRY* **2013**, *52*, 2708–2728.

(4) Canchi, D. R.; Garcia, A. E. In *ANNUAL REVIEW OF PHYSICAL CHEMISTRY, VOL 64*; Johnson, MA and Martinez, TJ,, Ed.; Annual Review of Physical Chemistry; 2013; Vol. 64; pp 273–293.

(5) Hansen, N.; van Gunsteren, W. F. Practical Aspects of Free-Energy Calculations: A Review. *Journal of Chemical Theory and Computation* **2014**, *10*, 2632–2647, PMID: 26586503.

(6) Mobley, D. L.; Gilson, M. K. In *ANNUAL REVIEW OF BIOPHYSICS, VOL 46*; Dill, KA,, Ed.; Annual Review of Biophysics; 2017; Vol. 46; pp 531–558.

(7) Wang, E.; Sun, H.; Wang, J.; Wang, Z.; Liu, H.; Zhang, J. Z. H.; Hou, T. End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design. *CHEMICAL REVIEWS* **2019**, *119*, 9478–9508.

(8) Enkavi, G.; Javanainen, M.; Kulig, W.; Rog, T.; Vattulainen, I. Multiscale Simulations of Biological Membranes: The Challenge To Understand Biological Phenomena in a Living Substance. *CHEMICAL REVIEWS* **2019**, *119*, 5607–5774.

(9) Bedrov, D.; Piquemal, J.-P.; Borodin, O.; MacKerell, A. D., Jr.; Roux, B.; Schroeder, C. Molecular Dynamics Simulations of Ionic Liquids and Electrolytes Using Polarizable Force Fields. *CHEMICAL REVIEWS* **2019**, *119*, 7940–7995.

(10) Marrink, S. J.; Corradi, V.; Souza, P. C. T.; Ingolfsson, H. I.; Tieleman, D. P.; Sansom, M. S. P. Computational Modeling of Realistic Cell Membranes. *CHEMICAL REVIEWS* **2019**, *119*, 6184–6226.

(11) Lee, A. C.-L.; Harris, J. L.; Khanna, K. K.; Hong, J.-H. A Comprehensive Review on Current Advances in Peptide Drug Development and Design. *INTERNATIONAL JOURNAL OF MOLECULAR SCIENCES* **2019**, *20*.

(12) Kang, L.; Liang, F.; Jiang, X.; Lin, Z.; Chen, C. First-Principles Design and Simulations Promote the Development of Nonlinear Optical Crystals. *ACCOUNTS OF CHEMICAL RESEARCH* **2020**, *53*, 209–217.

(13) Nelson, T. R.; White, A. J.; Bjorgaard, J. A.; Sifain, A. E.; Zhang, Y.; Nebgen, B.; Fernandez-Alberti, S.; Mozyrsky, D.; Roitberg, A. E.; Tretiak, S. Non-adiabatic Excited-State Molecular Dynamics: Theory and Applications for Modeling Photophysics in Extended Molecular Materials. *CHEMICAL REVIEWS* **2020**, *120*, 2215–2287.

(14) Shaebani, M. R.; Wysocki, A.; Winkler, R. G.; Gompper, G.; Rieger, H. Computational models for active matter. *NATURE REVIEWS PHYSICS* **2020**, *2*, 181–199.

(15) Spaggiari, G.; Di Pizio, A.; Cozzini, P. Sweet, umami and bitter taste receptors: State of the art of in silico molecular modeling approaches. *TRENDS IN FOOD SCIENCE & TECHNOLOGY* **2020**, *96*, 21–29.

(16) Duncan, A. L.; Song, W.; Sansom, M. S. P. In *ANNUAL REVIEW OF PHARMACOLOGY AND TOXICOLOGY, VOL 60*; Insel, PA,, Ed.; Annual Review of Pharmacology and Toxicology; 2020; Vol. 60; pp 31–50.

(17) Rust, J. *The New Palgrave Dictionary of Economics*; Palgrave Macmillan UK: London, 2016; pp 1–26.

(18) Sambasivan, R.; Das, S.; Sahu, S. K. A Bayesian perspective of statistical machine learning for big data. *COMPUTATIONAL STATISTICS* **2020**, *35*, 893–930.

(19) Rudzinski, J. F.; Noid, W. G. Coarse-graining entropy, forces, and structures. *Journal of Chemical Physics* **2011**, *135*.

(20) Marrink, S. J.; Tieleman, D. P. Perspective on the martini model. *Chemical Society Reviews* **2013**, *42*, 6801–6822.

(21) Noid, W. G. Perspective: Coarse-grained models for biomolecular systems. *Journal of Chemical Physics* **2013**, *139*.

(22) Saunders, M. G.; Voth, G. A. Coarse-graining methods for computational biology. *Annual Review of Biophysics* **2013**, *42*, 73–93.

(23) Ruff, K. M.; Harmon, T. S.; Pappu, R. V. CAMELOT: A machine learning approach for Coarse-grained simulations of aggregation of block-copolymeric protein sequences. *Journal of Chemical Physics* **2015**, *143*, 1–19.

(24) Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse-Grained Protein Models and Their Applications. *Chemical Reviews* **2016**, *116*, 7898–7936.

(25) Hafner, A. E.; Krausser, J.; Saric, A. Minimal coarse-grained models for molecular self-organisation in biology. *CURRENT OPINION IN STRUCTURAL BIOLOGY* **2019**, *58*, 43–52.

(26) Joshi, S. Y.; Deshmukh, S. A. A review of advancements in coarse-grained molecular dynamics simulations. *Molecular Simulation* **2020**, *0*, 1–18.

(27) Gkeka, P. et al. Machine Learning Force Fields and Coarse-Grained Variables in Molecular Dynamics: Application to Materials and Biological Systems. *Journal of Chemical Theory and Computation* **2020**, *16*, 4757–4775, PMID: 32559068.

(28) Bernardi, R. C.; Melo, M. C. R.; Schulten, K. Enhanced sampling techniques in molecular dynamics simulations of biological systems. *BIOCHIMICA ET BIOPHYSICA ACTA-GENERAL SUBJECTS* **2015**, *1850*, 872–877.

(29) Mlynsky, V.; Bussi, G. Exploring RNA structure and dynamics through enhanced sampling simulations. *CURRENT OPINION IN STRUCTURAL BIOLOGY* **2018**, *49*, 63–71.

(30) Yang, Y. I.; Shao, Q.; Zhang, J.; Yang, L.; Gao, Y. Q. Enhanced sampling in molecular dynamics. *Journal of Chemical Physics* **2019**, *151*.

(31) Wang, A.-h.; Zhang, Z.-c.; Li, G.-h. Advances in Enhanced Sampling Molecular Dynamics Simulations for Biomolecules. *CHINESE JOURNAL OF CHEMICAL PHYSICS* **2019**, *32*, 277–286.

(32) Mlynsky, V.; Bussi, G. Exploring RNA structure and dynamics through enhanced sampling simulations. *CURRENT OPINION IN STRUCTURAL BIOLOGY* **2018**, *49*, 63–71.

(33) Okamoto, Y. Protein structure predictions by enhanced conformational sampling methods. *BIOPHYSICS AND PHYSICOBIOLOGY* **2019**, *16*, 344–366.

(34) Wang, W.; Gómez-Bombarelli, R. Coarse-graining auto-encoders for molecular dynamics. *npj Computational Materials* **2019**, *5*.

(35) Lazim, R.; Suh, D.; Choi, S. Advances in Molecular Dynamics Simulations and Enhanced Sampling Methods for the Study of Protein Systems. *International Journal of Molecular Sciences* **2020**, *21*.

(36) Liao, Q. In *COMPUTATIONAL APPROACHES FOR UNDERSTANDING DYNAMICAL SYSTEMS: PROTEIN FOLDING AND ASSEMBLY*; Strodel, B and Barz, B,, Ed.; Progress in Molecular Biology and Translational Science; ACADEMIC PRESS LTD-ELSEVIER SCIENCE LTD: 125 LONDON WALL, LONDON EC2Y 5AS, ENGLAND, 2020; Vol. 170; pp 177–213.

(37) Long, S.; Tian, P. A simple neural network implementation of generalized solvation free energy for assessment of protein structural models. *RSC Advances* **2019**, *9*, 36227–36233.

(38) Čížek, J. On the Correlation Problem in Atomic and Molecular Systems. Calculation of Wavefunction Components in Ursell-Type Expansion Using Quantum-Field Theoretical Methods. *The Journal of Chemical Physics* **1966**, *45*, 4256–4266.

(39) Reyes, A.; Moncada, F.; Charry, J. The any particle molecular orbital approach: A short review of the theory and applications. *INTERNATIONAL JOURNAL OF QUANTUM CHEMISTRY* **2019**, *119*.

(40) Van Houten, J. A Century of Chemical Dynamics Traced through the Nobel Prizes. 1998: Walter Kohn and John Pople. *Journal of Chemical Education* **2002**, *79*, 1297.

(41) Bensberg, M.; Neugebauer, J. Density functional theory based embedding approaches for transition-metal complexes. *Phys. Chem. Chem. Phys.* **2020**, *22*, 26093–26103.

(42) Kraus, P. Basis Set Extrapolations for Density Functional Theory. *Journal of Chemical Theory and Computation* **2020**, *16*, 5712–5722, PMID: 32790303.

(43) Morgante, P.; Peverati, R. The devil in the details: A tutorial review on some undervalued aspects of density functional theory calculations. *INTERNATIONAL JOURNAL OF QUANTUM CHEMISTRY* **2020**, *120*.

(44) Zhang, I. Y.; Xu, X. On the top rung of Jacob's ladder of density functional theory: Toward resolving the dilemma ofSIEandNCE. *WILEY INTERDISCIPLINARY REVIEWS-COMPUTATIONAL MOLECULAR SCIENCE* **2021**, *11*.

(45) Mackerell Jr., A. D. Empirical force fields for biological macromolecules: Overview and issues. *Journal of Computational Chemistry* **2004**, *25*, 1584–1604.

(46) Kumar, A.; Yoluk, O.; MacKerell Jr., A. D. FFParam: Standalone package for CHARMM additive and Drude polarizable force field parametrization of small molecules. *Journal of Computational Chemistry* **2020**, *41*, 958–970.

(47) Oweida, T. J.; Kim, H. S.; Donald, J. M.; Singh, A.; Yingling, Y. G. Assessment of AMBER Force Fields for Simulations of ssDNA. *Journal of Chemical Theory and Computation* **2021**, *17*, 1208–1217, PMID: 33434436.

(48) 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. *Journal of Chemical Information and Modeling* **2021**, *61*, 284–297, PMID: 33307679.

(49) Alford, R. F. et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation* **2017**, *13*, 3031–3048, PMID: 28430426.

(50) Sasse, A.; de Vries, S. J.; Schindler, C. E. M.; de Beauchêne, I. C.; Zacharias, M. Rapid Design of Knowledge-Based Scoring Potentials for Enrichment of Near-Native Geometries in Protein-Protein Docking. *PLOS ONE* **2017**, *12*, 1–19.

(51) Narykov, O.; Bogatov, D.; Korkin, D. DISPOT: a simple knowledge-based protein domain interaction statistical potential. *Bioinformatics* **2019**, *35*, 5374–5378.

(52) Wang, S. Efficiently Calculating Anharmonic Frequencies of Molecular Vibration by Molecular Dynamics Trajectory Analysis. *ACS Omega* **2019**, *4*, 9271–9283.

(53) Ferreiro, D. U.; Komives, E. A.; Wolynes, P. G. Frustration in biomolecules. *Quarterly Reviews of Biophysics* **2014**, *47*, 285?363.

(54) Sulkowska, J. I. On folding of entangled proteins: knots, lassos, links and theta-curves. *CURRENT OPINION IN STRUCTURAL BIOLOGY* **2020**, *60*, 131–141.

(55) Chen, M.; Chen, X.; Schafer, N. P.; Clementi, C.; Komives, E. A.; Ferreiro, D. U.; Wolynes, P. G. Surveying biomolecular frustration at atomic resolution. *NATURE COMMUNICATIONS* **2020**, *11*, 5944.

(56) Warshel, A.; Kato, M.; Pisliakov, A. V. Polarizable force fields: History, test cases, and prospects. *Journal of Chemical Theory and Computation* **2007**, *3*, 2034–2045.

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

(58) Wang, L. P.; Chen, J.; Van Voorhis, T. Systematic parametrization of oolarizable force fields from quantum chemistry data. *Journal of Chemical Theory and Computation* **2013**, *9*, 452–460.

(59) Császár, A. G. Anharmonic molecular force fields. *WIREs Computational Molecular Science* **2012**, *2*, 273–289.

(60) Ulmschneider, J. P.; Ulmschneider, M. B. Molecular Dynamics Simulations Are Redefining Our View of Peptides Interacting with Biological Membranes. *Accounts of Chemical Research* **2018**, *51*, 1106–1116, PMID: 29667836.

(61) Jung, J.; Nishima, W.; Daniels, M.; Bascom, G.; Kobayashi, C.; Adedoyin, A.; Wall, M.; Lappala, A.; Phillips, D.; Fischer, W.; Tung, C.-S.; Schlick, T.; Sugita, Y.; Sanbonmatsu, K. Y. Scaling molecular dynamics beyond 100,000 processor cores for large-scale biophysical simulations. *JOURNAL OF COMPUTATIONAL CHEMISTRY* **2019**, *40*, 1919–1930.

(62) Wolf, S.; Lickert, B.; Bray, S.; Stock, G. Multisecond ligand dissociation dynamics from atomistic simulations. *NATURE COMMUNICATIONS* **2020**, *11*.

(63) Ferguson, A. L.; Panagiotopoulos, A. Z.; Kevrekidis, I. G.; Debenedetti, P. G. Nonlinear dimensionality reduction in molecular simulation: The diffusion map approach. *Chemical Physics Letters* **2011**, *509*, 1 – 11.

(64) *Rugged Free Energy Landscapes*; Springer: Verlag Berlin Heidelberg, 2008.

(65) Westerlund, A. M.; Delemotte, L. InfleCS: Clustering Free Energy Landscapes with Gaussian Mixtures. *Journal of Chemical Theory and Computation* **2019**, *15*, 6752–6759, PMID: 31647864.

(66) Dick, T. J.; Madura, J. D. *Chapter 5 A Review of the TIP4P, TIP4P-Ew, TIP5P, and TIP5P-E Water Models*; Annual Reports in Computational Chemistry; Elsevier, 2005; Vol. 1; pp 59 – 74.

(67) Vanommeslaeghe, K.; Hatcher, E.; Acharya, C.; Kundu, S.; Zhong, S.; Shim, J.; Darian, E.; Guvench, O.; Lopes, P.; Vorobyov, I.; Mackerell Jr., A. D. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of Computational Chemistry* **2010**, *31*, 671–690.

(68) van der Spoel, D. Systematic design of biomolecular force fields. *Current Opinion in Structural Biology* **2021**, *67*, 18–24.

(69) Michael, L.; Warshel, A. Computer simulation of protein folding. *nature* **1975**, *253*, 694–698.

(70) Levitt, M. A simplified representation of protein conformations for rapid simulation of protein folding. *Journal of Molecular Biology* **1976**, *104*, 59 – 107.

(71) Field, M. J.; Bash, P. A.; Karplus, M. A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations. *Journal of Computational Chemistry* **1990**, *11*, 700–733.

(72) Gao, J. *Reviews in Computational Chemistry*; John Wiley Sons, Ltd, 2007; pp 119–185.

(73) Messer, B. M.; Roca, M.; Chu, Z. T.; Vicatos, S.; Kilshtain, A. V.; Warshel, A. Multiscale simulations of protein landscapes: Using coarse-grained models as reference potentials to full explicit models. *Proteins: Structure, Function, and Bioinformatics* **2010**, *78*, 1212–1227.

(74) Mukherjee, S.; Warshel, A. Realistic simulations of the coupling between the protomotive force and the mechanical rotation of the F0-ATPase. *Proceedings of the National Academy of Sciences* **2012**, *109*, 14876–14881.

(75) Jones, L. O.; Mosquera, M. A.; Schatz, G. C.; Ratner, M. A. Embedding Methods for Quantum Chemistry: Applications from Materials to Life Sciences. *JOURNAL OF THE AMERICAN CHEMICAL SOCIETY* **2020**, *142*, 3281–3295.

(76) Nochebuena, J.; Naseem-Khan, S.; Cisneros, G. A. Development and application of quantum mechanics/molecular mechanics methods with advanced polarizable potentials. *WIREs Computational Molecular Science* **2021**, *n/a*, e1515.

(77) Chen, C.; Depa, P.; Sakai, V. G.; Maranas, J. K.; Lynn, J. W.; Peral, I.; Copley, J. R. D. A comparison of united atom, explicit atom, and coarse-grained simulation models for poly(ethylene oxide). *The Journal of Chemical Physics* **2006**, *124*, 234901.

(78) Potter, T. D.; Walker, M.; Wilson, M. R. Self-assembly and mesophase formation in a non-ionic chromonic liquid crystal: insights from bottom-up and top-down coarse-grained simulation models. *Soft Matter* **2020**, *16*, 9488–9498.

(79) Tschöp, W.; Kremer, K.; Batoulis, J.; Bürger, T.; Hahn, O. Simulation of polymer melts. I. Coarse-graining procedure for polycarbonates. *Acta Polymerica* **1998**, *49*, 61–74.

(80) Español, P.; Warren, P. B. Perspective: Dissipative particle dynamics. *Journal of Chemical Physics* **2017**, *146*.

(81) Wu, K.; Xu, S.; Wan, B.; Xiu, P.; Zhou, X. A novel multiscale scheme to accelerate atomistic simulations of bio-macromolecules by adaptively driving coarse-grained coordinates. *JOURNAL OF CHEMICAL PHYSICS* **2020**, *152*.

(82) Perdikari, T. M.; Jovic, N.; Dignon, G. L.; Kim, Y. C.; Fawzi, N. L.; Mittal, J. A coarse-grained model for position-specific effects of post-translational modifications on disordered protein phase separation. *Biophysical Journal* **2021**,

(83) Elliott, J. A. Novel approaches to multiscale modelling in materials science. *INTERNATIONAL MATERIALS REVIEWS* **2011**, *56*, 207–225.

(84) Jankowski, E. et al. Perspective on coarse-graining, cognitive load, and materials simulation. *COMPUTATIONAL MATERIALS SCIENCE* **2020**, *171*.

(85) Lyubartsev, A. P.; Laaksonen, A. Calculation of effective interaction potentials from radial distribution functions: A reverse Monte Carlo approach. *Phys. Rev. E* **1995**, *52*, 3730–3737.

(86) Chiba, S.; Okuno, Y.; Honma, T.; Ikeguchi, M. Force-field parametrization based on radial and energy distribution functions. *Journal of Computational Chemistry* **2019**, *40*, 2577–2585.

(87) Mironenko, A. V.; Voth, G. A. Density Functional Theory-Based Quantum Mechanics/Coarse-Grained Molecular Mechanics: Theory and Implementation. *Journal of Chemical Theory and Computation* **2020**, *16*, 6329–6342, PMID: 32877176.

(88) Rudzinski, J. F.; Noid, W. G. Coarse-graining entropy, forces, and structures. *The Journal of Chemical Physics* **2011**, *135*, 214101.

(89) Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *The Journal of Chemical Physics* **2008**, *128*, 244114.

(90) Noid, W. G.; Liu, P.; Wang, Y.; Chu, J.-W.; Ayton, G. S.; Izvekov, S.; Andersen, H. C.; Voth, G. A. The multiscale coarse-graining method. II. Numerical implementation for coarse-grained molecular models. *The Journal of Chemical Physics* **2008**, *128*, 244115.

(91) Jin, J.; Han, Y.; Pak, A. J.; Voth, G. A. A new one-site coarse-grained model for water: Bottom-up many-body projected water (BUMPer). I. General theory and model. *The Journal of Chemical Physics* **2021**, *154*, 044104.

(92) Shen, H.; Li, Y.; Ren, P.; Zhang, D.; Li, G. Anisotropic Coarse-Grained Model for Proteins Based On Gay–Berne and Electric Multipole Potentials. *Journal of Chemical Theory and Computation* **2014**, *10*, 731–750, PMID: 24659927.

(93) Li, G.; Shen, H.; Zhang, D.; Li, Y.; Wang, H. Coarse-Grained Modeling of Nucleic Acids Using Anisotropic Gay–Berne and Electric Multipole Potentials. *Journal of Chemical Theory and Computation* **2016**, *12*, 676–693, PMID: 26717419.

(94) Tanis, I.; Rousseau, B.; Soulard, L.; Lemarchand, C. A. Assessment of an anisotropic coarse-grained model for cis-1,4-polybutadiene: a bottom-up approach. *Soft Matter* **2021**, *17*, 621–636.

(95) Dama, J. F.; Sinitskiy, A. V.; McCullagh, M.; Weare, J.; Roux, B.; Dinner, A. R.; Voth, G. A. The Theory of Ultra-Coarse-Graining. 1. General Principles. *Journal of Chemical Theory and Computation* **2013**, *9*, 2466–2480, PMID: 26583735.

(96) Davtyan, A.; Dama, J. F.; Sinitskiy, A. V.; Voth, G. A. The Theory of Ultra-Coarse-Graining. 2. Numerical Implementation. *Journal of Chemical Theory and Computation* **2014**, *10*, 5265–5275, PMID: 26583210.

(97) Jin, J.; Voth, G. A. Ultra-Coarse-Grained Models Allow for an Accurate and Transferable Treatment of Interfacial Systems. *JOURNAL OF CHEMICAL THEORY AND COMPUTATION* **2018**, *14*, 2180–2197.

(98) Zhang, Y.; Cao, Z.; Zhang, J. Z.; Xia, F. Double-Well Ultra-Coarse-Grained Model to Describe Protein Conformational Transitions. *Journal of Chemical Theory and Computation* **0**, *0*, null, PMID: 32926616.

(99) Jin, J.; Yu, A.; Voth, G. A. Temperature and Phase Transferable Bottom-up Coarse-Grained Models. *JOURNAL OF CHEMICAL THEORY AND COMPUTATION* **2020**, *16*, 6823–6842.

(100) Ueda, Y.; Taketomi, H.; Gō, N. Studies on protein folding, unfolding, and fluctuations by computer simulation. II. A. Three-dimensional lattice model of lysozyme. *Biopolymers* **1978**, *17*, 1531–1548.

(101) Nymeyer, H.; García, A. E.; Onuchic, J. N. Folding funnels and frustration in off-lattice minimalist protein landscapes. *Proceedings of the National Academy of Sciences* **1998**, *95*, 5921–5928.

(102) Tirion, M. M. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. *Phys. Rev. Lett.* **1996**, *77*, 1905–1908.

(103) AR, A.; SR, D.; RL, J.; MC, D.; Keskin O, B. I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophysical Journal* **2001**, *80*, 505–15.

(104) Togashi, Y.; Flechsig, H. Coarse-Grained Protein Dynamics Studies Using Elastic Network Models. *INTERNATIONAL JOURNAL OF MOLECULAR SCIENCES* **2018**, *19*.

(105) Haliloglu, T.; Bahar, I.; Erman, B. Gaussian Dynamics of Folded Proteins. *Phys. Rev. Lett.* **1997**, *79*, 3090–3093.

(106) Wang, S.; Gong, W.; Deng, X.; Liu, Y.; Li, C. Exploring the dynamics of RNA molecules with multiscale Gaussian network model. *Chemical Physics* **2020**, *538*, 110820.

(107) Yang, L.-W.; Chng, C.-P. Coarse-Grained Models Reveal Functional Dynamics - I. Elastic Network Models – Theories, Comparisons and Perspectives. *Bioinformatics and Biology Insights* **2008**, *2*, BBI.S460, PMID: 19812764.

(108) Chng, C.-P.; Yang, L.-W. Coarse-Grained Models Reveal Functional Dynamics – II. Molecular Dynamics Simulation at the Coarse-Grained Level – Theories and Biological Applications. *Bioinformatics and Biology Insights* **2008**, *2*, BBI.S459, PMID: 19812774.

(109) Torrie, G.; Valleau, J. Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics* **1977**, *23*, 187 – 199.

(110) Mezei, M. Adaptive umbrella sampling: Self-consistent determination of the non-Boltzmann bias. *Journal of Computational Physics* **1987**, *68*, 237 – 248.

(111) Park, S.; Im, W. Theory of adaptive optimization for umbrella sampling. *Journal of Chemical Theory and Computation* **2014**, *10*, 2719–2728.

(112) Kumar, S.; Rosenberg, J. M.; Bouzida, D.; Swendsen, R. H.; Kollman, P. A. THE weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *Journal of Computational Chemistry* **1992**, *13*, 1011–1021.

(113) Shirts, M. R.; Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *Journal of Chemical Physics* **2008**, *129*, 124105.

(114) Paliwal, H.; Shirts, M. R. Using Multistate Reweighting to Rapidly and E ffi ciently Explore Molecular Simulation Parameters Space for Nonbonded Interactions. *Journal of Chemical Theory and Computation* **2013**, *9*, 4700–4717.

(115) Darve, E.; Pohorille, A. Calculating free energies using average force. *The Journal of Chemical Physics* **2001**, *115*, 9169–9183.

(116) Zhao, T.; Fu, H.; Lehievre, T.; Shao, X.; Chipot, C.; Cai, W. The Extended Generalized Adaptive Biasing Force Algorithm for Multidimensional Free-Energy Calculations. *JOURNAL OF CHEMICAL THEORY AND COMPUTATION* **2017**, *13*, 1566–1576.

(117) Miao, M.; Fu, H.; Zhang, H.; Shao, X.; Chipot, C.; Cai, W. Avoiding non-equilibrium effects in adaptive biasing force calculations. *MOLECULAR SIMULATION*

(118) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proceedings of the National Academy of Sciences* **2002**, *99*, 12562–12566.

(119) Raniolo, S.; Limongelli, V. Ligand binding free-energy calculations with funnel metadynamics. *Nature Protocols* **2020**, *15*, 2837–2866.

(120) Kondo, T.; Sasaki, T.; Ruiz-Barragan, S.; Ribas-Ariño, J.; Shiga, M. Refined metadynamics through canonical sampling using time-invariant bias potential: A study of

polyalcohol dehydration in hot acidic solutions. *Journal of Computational Chemistry* **2021**, *42*, 156–165.

(121) Comer, J.; Gumbart, J. C.; Hénin, J.; Lelièvre, T.; Pohorille, A.; Chipot, C. The Adaptive Biasing Force Method: Everything You Always Wanted To Know but Were Afraid To Ask. *The Journal of Physical Chemistry B* **2015**, *119*, 1129–1151, PMID: 25247823.

(122) Fu, H.; Shao, X.; Cai, W.; Chipot, C. Taming Rugged Free Energy Landscapes Using an Average Force. *Accounts of Chemical Research* **2019**, *52*, 3254–3264, PMID: 31680510.

(123) Valsson, O.; Tiwary, P.; Parrinello, M. Enhancing Important Fluctuations: Rare Events and Metadynamics from a Conceptual Viewpoint. *Annual Review of Physical Chemistry* **2016**, *67*, 159–184.

(124) Bussi, G.; Laio, A. Using metadynamics to explore complex free-energy landscapes. *Nature Reviews Physics* **2020**, *2*, 200–212.

(125) *Principal Component Analysis*; Springer: New York Berlin Heidelberg Hong Kong Milan Paris Tokyo, 2002.

(126) Shan, P.; Zhao, Y.; Wang, Q.; Ying, Y.; Peng, S. Principal component analysis or kernel principal component analysis based joint spectral subspace method for calibration transfer. *SPECTROCHIMICA ACTA PART A-MOLECULAR AND BIOMOLECULAR SPECTROSCOPY* **2020**, *227*.

(127) *Multidimensional Scaling*; CHAPMAN & HALL/CRC, 2000.

(128) Tenenbaum, J. B.; Silva, V. d.; Langford, J. C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **2000**, *290*, 2319–2323.

(129) Roweis, S. T.; Saul, L. K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* **2000**, *290*, 2323–2326.

(130) Coifman, R. R.; Lafon, S.; Lee, A. B.; Maggioni, M.; Nadler, B.; Warner, F.; Zucker, S. W. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences* **2005**, *102*, 7426–7431.

(131) Coifman, R. R.; Lafon, S.; Lee, A. B.; Maggioni, M.; Nadler, B.; Warner, F.; Zucker, S. W. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Multiscale methods. *Proceedings of the National Academy of Sciences* **2005**, *102*, 7432–7437.

(132) Ceriotti, M.; Tribello, G. A.; Parrinello, M. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proceedings of the National Academy of Sciences* **2011**, *108*, 13023–13028.

(133) Dellago, C.; Bolhuis, P. G.; Geissler, P. L. *Advances in Chemical Physics*; John Wiley Sons, Ltd, 2003; Chapter 1, pp 1–78.

(134) Rogal, J.; Bolhuis, P. G. Multiple state transition path sampling. *The Journal of Chemical Physics* **2008**, *129*, 224107.

(135) Buijsman, P.; Bolhuis, P. G. Transition path sampling for non-equilibrium dynamics without predefined reaction coordinates. *The Journal of Chemical Physics* **2020**, *152*, 044108.

(136) Yao, Y.; Cui, R. Z.; Bowman, G. R.; Silva, D.-A.; Sun, J.; Huang, X. Hierarchical Nyström methods for constructing Markov state models for conformational dynamics. *The Journal of Chemical Physics* **2013**, *138*, 174106.

(137) Chodera, J. D.; Noé, F. Markov state models of biomolecular conformational dynamics. *Current Opinion in Structural Biology* **2014**, *25*, 135–144.

(138) Husic, B. E.; Pande, V. S. Markov State Models: From an Art to a Science. *Journal of the American Chemical Society* **2018**, *140*, 2386–2396, PMID: 29323881.

(139) Nagel, D.; Weber, A.; Stock, G. MSMPathfinder: Identification of Pathways in Markov State Models. *Journal of Chemical Theory and Computation* **2020**, *16*, 7874–7882, PMID: 33141565.

(140) Sadiq, S. K.; Noé, F.; De Fabritiis, G. Kinetic characterization of the critical step in HIV-1 protease maturation. *Proceedings of the National Academy of Sciences* **2012**, *109*, 20449–20454.

(141) Kohlhoff, K. J.; Shukla, D.; Lawrenz, M.; Bowman, G. R.; Konerding, D. E.; Belov, D.; Altman, R. B.; Pande, V. S. Cloud-based simulations on Google Exacycle reveal ligand modulation of GPCR activation pathways. *Nature Chemistry* **2014**, *6*, 15–21.

(142) Yang, M.; Tang, Y.; Weng, J.; Liu, Z.; Wang, W. The Role of Calcium in Regulating the Conformational Dynamics of d-Galactose/d-Glucose-Binding Protein Revealed by Markov State Model Analysis. *Journal of Chemical Information and Modeling* **2021**, *61*, 891–900, PMID: 33445873.

(143) Wu, H.; Noé, F. Variational Approach for Learning Markov Processes from Time Series Data. *Journal of Nonlinear Science* **2020**, *30*, 23–66.

(144) Zuckerman, D. M.; Chong, L. T. Weighted Ensemble Simulation: Review of Methodology, Applications, and Software. *Annual Review of Biophysics* **2017**, *46*, 43–57, PMID: 28301772.

(145) Tsallis, C. Some comments on Boltzmann-Gibbs statistical mechanics. *Chaos, Solitons & Fractals* **1995**, *6*, 539 – 559, Complex Systems in Computational Physics.

(146) Plastino, A. Why Tsallis' statistics? *Physica A: Statistical Mechanics and its Applications* **2004**, *344*, 608 – 613, Proceedings of the International Workshop on 'Trends and perspectives in extensive and non-extensive statistical mechanics', in honor of the 60th birthday of Constantino Tsallis.

(147) Swendsen, R. H.; Wang, J.-S. Replica Monte Carlo Simulation of Spin-Glasses. *Phys. Rev. Lett.* **1986**, *57*, 2607–2609.

(148) Appadurai, R.; Nagesh, J.; Srivastava, A. High resolution ensemble description of metamorphic and intrinsically disordered proteins using an efficient hybrid parallel tempering scheme. *Nature communications* **2021**, *12*, 958.

(149) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters* **1999**, *314*, 141 – 151.

(150) Peng, C.; Wang, J.; Shi, Y.; Xu, Z.; Zhu, W. Increasing the Sampling Efficiency of Protein Conformational Change by Combining a Modified Replica Exchange Molecular Dynamics and Normal Mode Analysis. *Journal of Chemical Theory and Computation* **2021**, *17*, 13–28, PMID: 33351613.

(151) Wang, F.; Landau, D. P. Efficient, Multiple-Range Random Walk Algorithm to Calculate the Density of States. *Phys. Rev. Lett.* **2001**, *86*, 2050–2053.

(152) Gao, Y. Q. An integrate-over-temperature approach for enhanced sampling. *The Journal of Chemical Physics* **2008**, *128*, 064105.

(153) Yang, L.; Liu, C.-W.; Shao, Q.; Zhang, J.; Gao, Y. Q. From Thermodynamics to Kinetics: Enhanced Sampling of Rare Events. *Accounts of Chemical Research* **2015**, *48*, 947–955, PMID: 25781363.

(154) Yang, Y. I.; Niu, H.; Parrinello, M. Combining Metadynamics and Integrated Temper-

ing Sampling. *JOURNAL OF PHYSICAL CHEMISTRY LETTERS* **2018**, *9*, 6426–6430.

(155) Behler, J. Perspective: Machine learning potentials for atomistic simulations. *Journal of Chemical Physics* **2016**, *145*.

(156) Ceriotti, M. Unsupervised machine learning in atomistic simulations, between predictions and understanding. *Journal of Chemical Physics* **2019**, *150*.

(157) Lunghi, A.; Sanvito, S. A unified picture of the covalent bond within quantum-accurate force fields: From organic molecules to metallic complexes' reactivity. *SCIENCE ADVANCES* **2019**, *5*.

(158) Mueller, T.; Hernandez, A.; Wang, C. Machine learning for interatomic potential models. *Journal of Chemical Physics* **2020**, *152*.

(159) Noé, F.; Tkatchenko, A.; Müller, K.-R.; Clementi, C. Machine Learning for Molecular Simulation. *Annual Review of Physical Chemistry* **2020**, *71*, 361–390, PMID: 32092281.

(160) Gkeka, P. et al. Machine Learning Force Fields and Coarse-Grained Variables in Molecular Dynamics: Application to Materials and Biological Systems. *Journal of Chemical Theory and Computation* **2020**, *16*, 4757–4775, PMID: 32559068.

(161) Takahashi, A.; Seko, A.; Tanaka, I. Linearized machine-learning interatomic potentials for non-magnetic elemental metals: Limitation of pairwise descriptors and trend of predictive power. *The Journal of Chemical Physics* **2018**, *148*, 234106.

(162) Deringer, V. L.; Csányi, G. Machine learning based interatomic potential for amorphous carbon. *Phys. Rev. B* **2017**, *95*, 094203.

(163) Bartók, A. P.; Kermode, J.; Bernstein, N.; Csányi, G. Machine Learning a General-Purpose Interatomic Potential for Silicon. *Phys. Rev. X* **2018**, *8*, 041048.

(164) Slepoy, A.; Peters, M.; AP, T. Searching for globally optimal functional forms for interatomic potentials using genetic programming with parallel tempering. *J Comput Chem* **2007**, *28*, 2465–71.

(165) Qu, C.; Bowman, J. M. A fragmented, permutationally invariant polynomial approach for potential energy surfaces of large molecules: Application to N-methyl acetamide. *The Journal of Chemical Physics* **2019**, *150*, 141101.

(166) Chmiela, S.; Sauceda, H. E.; Müller, K.-R. Towards exact molecular dynamics simulations with machine-learned force fields. *Nature Communications* **2018**, *9*, 3887.

(167) Artrith, N.; Behler, J. High-dimensional neural network potentials for metal surfaces: A prototype study for copper. *Phys. Rev. B* **2012**, *85*, 045439.

(168) Podryabinkin, E. V.; Tikhonov, E. V.; Shapeev, A. V.; Oganov, A. R. Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning. *Phys. Rev. B* **2019**, *99*, 064114.

(169) Jinnouchi, R.; Karsai, F.; Kresse, G. On-the-fly machine learning force field generation: Application to melting points. *Phys. Rev. B* **2019**, *100*, 014105.

(170) Huan, T. D.; Batra, R.; Chapman, J.; Kim, C.; Chandrasekaran, A.; Ramprasad, R. Iterative-Learning Strategy for the Development of Application-Specific Atomistic Force Fields. *The Journal of Physical Chemistry C* **2019**, *123*, 20715–20722.

(171) Ghasemi, S. A.; Hofstetter, A.; Saha, S.; Goedecker, S. Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network. *Phys. Rev. B* **2015**, *92*, 045131.

(172) Ko, T. W.; Finkler, J. A.; Goedecker, S.; Behler, J. A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer. *Nature Communications* **2021**, *12*, 398.

(173) Harris, W. H. Machine Learning Transferable Physics-Based Force Fields using Graph Convolutional Neural Networks. 2020.

(174) Zhang, Y.; Hu, C.; Jiang, B. Embedded Atom Neural Network Potentials: Efficient and Accurate Machine Learning with a Physically Inspired Representation. *The Journal of Physical Chemistry Letters* **2019**, *10*, 4962–4967, PMID: 31397157.

(175) Chan, H.; Cherukara, M. J.; Narayanan, B.; Loeffler, T. D.; Benmore, C.; Gray, S. K.; Sankaranarayanan, S. K. Machine learning coarse grained models for water. *Nature Communications* **2019**, *10*, 1–14.

(176) Moradzadeh, A.; Aluru, N. R. Transfer-Learning-Based Coarse-Graining Method for Simple Fluids: Toward Deep Inverse Liquid-State Theory. *Journal of Physical Chemistry Letters* **2019**, *10*, 1242–1250.

(177) Wang, J.; Olsson, S.; Wehmeyer, C.; Pérez, A.; Charron, N. E.; De Fabritiis, G.; Noé, F.; Clementi, C. Machine Learning of Coarse-Grained Molecular Dynamics Force Fields. *ACS Central Science* **2019**, *5*, 755–767.

(178) Zhang, L.; Han, J.; Wang, H.; Car, R.; Weinan, W. E. DeePCG: Constructing coarse-grained models via deep neural networks. *Journal of Chemical Physics* **2018**, *149*.

(179) Chakraborty, M.; Xu, C.; White, A. D. Encoding and selecting coarse-grain mapping operators with hierarchical graphs. *Journal of Chemical Physics* **2018**, *149*.

(180) Webb, M. A.; Delannoy, J. Y.; De Pablo, J. J. Graph-Based Approach to Systematic Molecular Coarse-Graining. *Journal of Chemical Theory and Computation* **2019**,

(181) Giulini, M.; Menichetti, R.; Shell, M. S.; Potestio, R. An information theory-based approach for optimal model reduction of biomolecules. **2020**, 1–22.

(182) Li, Z.; Wellawatte, G. P.; Chakraborty, M.; Gandhi, H. A.; Xu, C.; White, A. D. Graph

neural network based coarse-grained mapping prediction. *Chemical Science* **2020**, *11*, 9524–9531.

(183) Durumeric, A. E.; Voth, G. A. Adversarial-residual-coarse-graining: Applying machine learning theory to systematic molecular coarse-graining. *Journal of Chemical Physics* **2019**, *151*.

(184) Khot, A.; Shiring, S. B.; Savoie, B. M. Evidence of information limitations in coarse-grained models. *Journal of Chemical Physics* **2019**, *151*.

(185) Chen, W.; Tan, A. R.; Ferguson, A. L. Collective variable discovery and enhanced sampling using autoencoders: Innovations in network architecture and error function design. *Journal of Chemical Physics* **2018**, *149*.

(186) Chen, W.; Ferguson, A. L. Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration. *Journal of Computational Chemistry* **2018**, *39*, 2079–2102.

(187) Wehmeyer, C.; Noé, F. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *Journal of Chemical Physics* **2018**, *148*.

(188) Ribeiro, J. M. L.; Bravo, P.; Wang, Y.; Tiwary, P. Reweighted autoencoded variational Bayes for enhanced sampling (RAVE). *Journal of Chemical Physics* **2018**, *149*.

(189) Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. VAMPnets for deep learning of molecular kinetics. *Nature Communications* **2018**, *9*, 1–11.

(190) Cao, X.; Tian, P. Molecular free energy optimization on a computational graph. 2020.

(191) Khatib, F.; Cooper, S.; Tyka, M. D.; Xu, K.; Makedon, I.; Popović, Z.; Baker, D.; Players, F. Algorithm discovery by protein folding game players. *Proceedings of the National Academy of Sciences* **2011**, *108*, 18949–18953.

 doi:10.20944/preprints202012.0081.v3

(192) Feig, M. Computational protein structure refinement: Almost there, yet still so far to go. *Wiley Interdiplinary Reviews: Computational Molecular science* **2017**, *7*, e1307.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2021 doi:10.20944/preprints202012.0081.v3

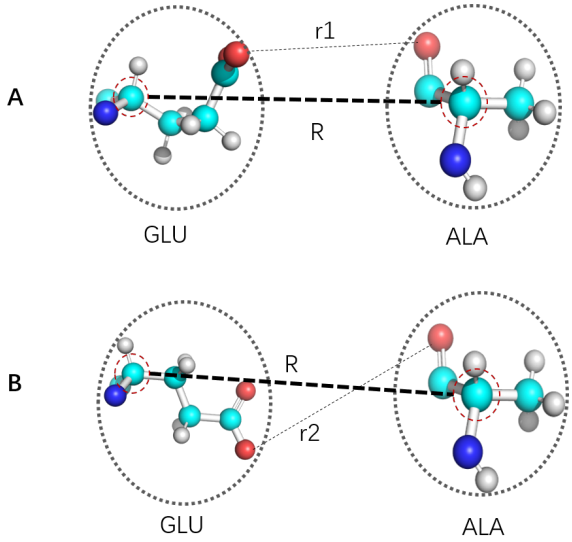

Figure 1: Schematic illustration of time scale separation issue in CG. A) and B) show two situations with C_α distances between two amino acids GLU and ALA being R, but with GLU have different conformations. If C_α atoms were defined as CG site, then these two relative conformation with distinct interactions would be treated as the same. In A) and B), CG site distance in both A) and B) are R, but many other pairs of atoms have distinct distances as exemplified by r_1 and r_2. Such treatment would only be true if for any small amount of displacement of C_α, side chains accomplished many rotations and thus may be accurately represented by averaging (i.e. with good time scale separation). This issue is apparently not limited to the specific definition of C_α being CG site, but rather general for essentially all CG development.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2021 doi:10.20944/preprints202012.0081.v3

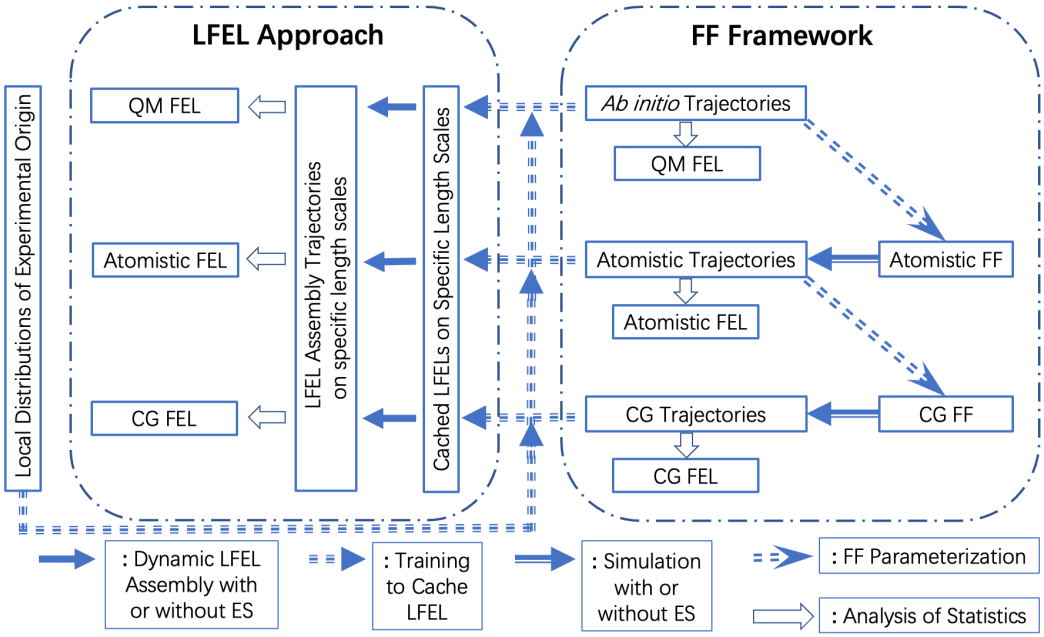

Figure 2: Schematic illustration of the LFEL approach in contrast to present mainstream FF framework. FF parameterization is the foundation for present classical computational molecular science. Training of neural network for "caching" LFEL is the foundation for LFEL approach, the source data can be either of experimental or computational origin. In FF framework, simulation (with or without ES) is driven by FF, in LFEL approach, propagation of molecular systems to minimize free energy (or maximize joint probability) is driven by compromise among many LFELs. Expensive repetitive local sampling in FF framework is substituted by differentiation w.r.t. LFELs.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2021 doi:10.20944/preprints202012.0081.v3

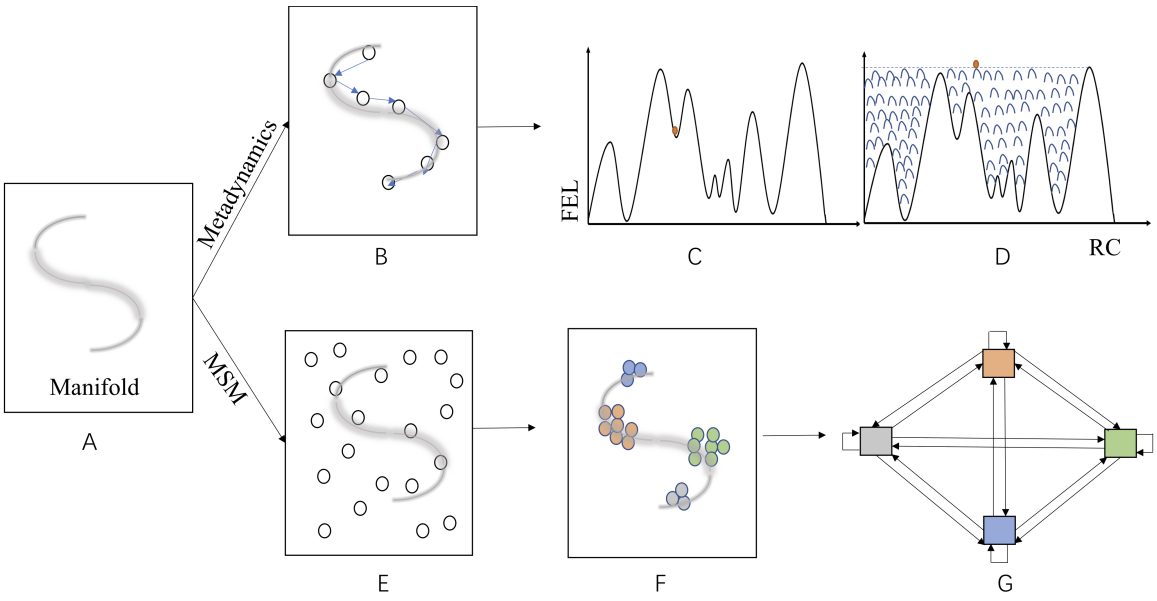

Figure 3: Schematic illustration of essential features for enhanced sampling by Metadynamics and MSM. A) The "S" shape grey line represents the unknown manifold in the configurational space (represented by the square) of a molecular system. B) Small circles connected by blue arrows represent computed (guessed) RC/CVs for the molecular system, which is utilized to conduct Metadynamics simulations. C) The FEL of the molecular system along the computed/selected RC/CV in B). D) "Caching" of the LFEL by bias potentials (gaussians represented by blue bell shaped lines) accumulated in the course of Metadynamics simulations. E) Distribution of the molecular system to the whole configurational space at the start of a MSM simulation, small circles represent initial start points for short MSM trajectories. F) Sampling results of short MSM trajectories fall mainly near the manifold, distinct "states" are represented by different colors. G) Establishment of transition matrix by transition counts between "states" obtained from short trajectories.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2021 doi:10.20944/preprints202012.0081.v3

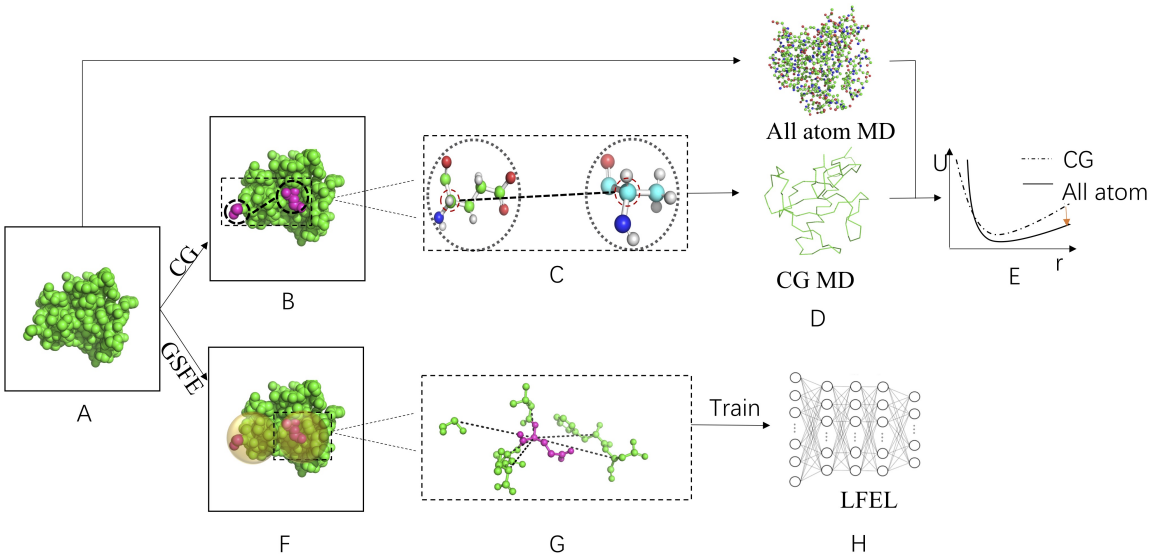

Figure 4: Schematic illustration of difference between CG and GSFE implementation of LFEL using protein as an example. A) Target molecular systems in physical space. Due to the goal of constructing partially transferable models and/or force fields, usually many different but similar molecular systems are considered. B) Selection of local atom/particle clusters to be represented as one particle in CG model. C) Selection of CG sites. D) Comparison between atomistic (or higher resolution) simulation results and CG (lower resolution) results. E) Adjust of CG FF parameter according to comparison from D). F)Definition of solvent region for each solute unit. G) Feature extraction for each solute. H) "Caching" of LFEL with neural network by training with prepared data sets.