The repetitive local sampling and the local distribution theory

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

Previously, the ubiquitous issue regarding severe wasting of computational resources in all forms of molecular simulations due to repetitive local sampling was raised, and the local free energy landscape approach was proposed to address it. This approach is derived from a simple idea of first learning local distributions, and followed by dynamic assembly of which to infer global joint distribution of a target molecular system. When compared with conventional explicit solvent molecular dynamics simulations, a simple and approximate implementation of this theory in protein structural refinement harvested acceleration of about six orders of magnitude without loss of accuracy. While this initial test revealed tremendous benefits for addressing repetitive local sampling, there are some implicit assumptions need to be articulated. Here, I present a more thorough discussion of repetitive local sampling; potential options for learning local distributions; a more general formulation with potential extension to simulation of near equilibrium molecular systems; generalization of repetitive local sampling to repetitive local computation and potential application in accelerating numerical solving of complex equations; the prospect of developing computation driven molecular science; and the connection to mainstream residue pair distance distribution based protein structure prediction/refinement.

This more general development is termed the local distribution theory to release the limitation of strict thermodynamic equilibrium in its potential wide application in both soft condensed molecular systems and complex equations.

Introduction

Molecular simulation has been utilized in a wide variety of disciplines, including but not limited to chemistry, physics, biology and materials science. Its increasing importance is clearly demonstrated by steady growth of relevant publications as shown in Fig. 1. However, atomistic molecular dynamics (MD) simulations, while being effective in revealing underlying atomic mechanisms for many molecular processes, are extremely computationally intensive.[1,2] Historically, scientists have developed two lines of algorithms to accelerate molecular simulations, with one being coarse graining (CG)[3–12] and the other being enhanced sampling (ES).[13–16] Realizing that there are severe wasting of computational resources due to repetitive local sampling (RLS) in all molecular simulations, the local free energy landscape (LFEL) approach was proposed to eliminate such wasting, and its effectiveness was subsequently demonstrated in an approximate implementation in protein structural refinement.[17] The connection among CS, ES and LFEL as various forms of applying "dividing and conquering" and "caching" principle in molecular modeling was summarized.[18] In our initial testing of this new theory, LFEL for amino acid packing in proteins was constructed based on a simple neural network implementation of generalized solvation free energy (GSFE) theory.[19] Further, a computational graph was established through combination of autodifferentiation, coordinate transformation and LFEL cached in trained neural networks. This computational graph was successfully utilized to achieve the only end-to-end and the most efficient protein structural refinement pipeline[17] up to date. Like all present protein structure prediction, design and refinement studies,[20–29] there is an implicit and extremely crude assumption that all high resolution experimental structures were derived under similar environmental (thermodynamic) conditions. Alternatively, differences in thermodynamic and environmental conditions are deemed not important for all high resolution structural data utilized to train models. Such assumptions are apparently not true. Additionally, the LFEL approach as it stands only applies to equilibrium conditions. Here, I explicitly articulate these issues, develop a more general form of the LFEL idea and termed it the local distribution theory (LDT). Meanwhile, more concrete discussions of RLS, more options for fitting local distributions, extension of RLS to repetitive local computation (RLC) and potential applications of LDT in accelerating numerical solving of complex equations are presented.

Repetitive local sampling

In molecular simulations, we have a long history of utilizing RLS in analysis of MD trajectories. For example, when computing pair distribution function $g(r)$ between oxygen atoms of water molecules, instead of counting a specific pair of water molecules or water molecules within a given small space and binning distances of oxygen atom pairs, statistics is usually accumulated by counting all pairs of water molecules in a simulation box to obtain a more smooth curve within a shorter simulation time. Similar tricks are routinely utilized in various analyses of molecular simulation trajectories. The basis of these manipulations is the belief that all molecules of the same chemical identity and composition are indistinguishable, and ensemble average converges to time average for ergodic systems. From a different perspective, all above practice clearly demonstrates that we have been carrying out RLS in essentially all our simulations, except not carefully thinking about its potential utility in saving computational resources in the simulation/sampling stage. This issue was raised previously[17,18] without sufficiently detailed discussions. Some typical examples of RLS in various simulation and/or modeling applications are discussed below.

RLS consumes overwhelming majority of computational resources in regular molecular simulations and exist both within a single simulation task and across different ones. As shown in Fig. 2a, there is a simulation of aqueous solution comprising a few different types of ions and water molecules, with gas-liquid and liquid-solid interfaces under given thermodynamic conditions. After a sufficiently long simulation run, if all snapshots were utilized to analyze distribution of all molecules and ions in a bulk spherical space *A*, one would have obtained a converged LFEL, which is a complex high dimensional distribution that gives correct statistical weight for each thermally reachable structural ensemble (or free energy local minimum) on the one hand, and all possible transition paths connecting these minima with respective statistical significance on the other hand. The exactly same LFEL would have been obtained if another bulk spherical space *B* with the same volume was taken. As a matter of fact, the exactly same LFEL would be obtained for all possible bulk spherical spaces with the same volume. However, for each such separate local space, significant computational resource was consumed to obtain the exactly same result! This is a typical case of RLS in the same simulation task.

While local spaces near various interface certainly have LFELs different from that of bulk, there are regularities that can be learned as well. Such RLS may be effectively described from a slightly different perspective according to the GSFE theory as shown in Fig. 2b. In GSFE theory, each comprising unit of a molecular system is on the one hand a solute unit solvated by its surrounding units, and on the other hand a comprising solvent unit for each of its surrounding units. As all units with the same chemical identity/structure are indistinguishable, so should be LFEL of their local solvent under given thermodynamic conditions if a simulation trajectory is sufficiently long. When our focus is on LFEL surrounding a central unit, different scenarios of interfaces are simply different solvent configurations with corresponding statistical weights and no special treatment is required. More specifically, for a water molecule absorbed on wall of a tube filled with water, its solvent units include both water molecules and molecules belong to the wall surrounding it. To eliminate difficulty of defining interfaces at molecular scales is the very initial motivation for development of the GSFE theory. Additionally, defining local spaces with local coordinates originated from individual molecule is a convenient, efficient and natural choice with two advantages. Firstly, it reduces data requirement and improves accuracy during training/learning of local distributions, and secondly, it facilitates assembly by eliminating the uncertainty of selecting from infinite possible origins for each local spaces during inference for global distribution of the target molecular system.

Beyond the illustration in Fig. 2, there are other less obvious forms of RLS. For example, in protein structure prediction, design and refinement with implicit representation of aqueous solution, each residue in a chain has more or less unique surroundings and no direct RLS seems existing. However, in these tasks, each residue experiences many rounds of adjustment or repacking, sampled collisions, favorable and unfavorable configurations from each round is partially or completely discarded and performed on the fly in the next round, engendering significant RLS. Much more computational resource are consumed by RLS across different tasks. Imagine how many times simulations of local packing for water molecules of each popular water force fields have been carried out by thousands of scientists globally! Similarly, packing of amino acids surrounding each of 20 natural amino acids have been carried out numerous times by computational structural bioinformaticians around the world. Such RLS is apparently ubiquitous for simulations of all molecular systems.

Sufficient sampling of complex molecular system has long been our pursuit in our simulation studies. The very fact that we almost always collect statistics from different local spaces and/or utilize indistinguishable property of molecules for better statistics indicates that we rarely achieve sufficient sampling for a given small space or surrounding of a given single molecule. Therefore, it is likely that more accurate global correlations would have been obtained if sufficient statistics was available for all local regions. Since construction of global distributions by assembly of LFEL realizes this very condition, the ability to cache and utilize LFEL properly would not only tremendously reduce amount of computational resources, but also potentially improve accuracy due to effectively more sufficient "local sampling". This is in strong contrast to decades of trade-off in molecular modeling that improved efficiency being always accompanied more or less by reduced accuracy, and increased efficiency being always accompanied by more or less reduction of accuracy! When compared with conventional molecular mechanical force fields[30–33] or knowledge based potentials,[34–36] the ability of accounting for many-body correlations is another advantage of LFEL that is likely to contribute to improved accuracy. It is important to note that many neural network based force fields (NNFF) methodologies have been developed up to date.[37,38] Essentially, development of NNFF and other machine learning based force fields is the mainstream of research bridging artificial intelligence (AI) and molecular simulations with many great successes. NNFF tackles many body correlations and demonstrates improved accuracy while sacrifice some efficiency, and remains in the established framework of "force fields + sampling" without considering RLS.

The local distribution theory

It is well understood that protein folding process and conformational distributions depend upon both its sequence and environmental conditions. However, due to lack to data, in both establishment of traditional knowledge based potentials[34–36] and deep learning studies[21,22] of protein folding, design and structural refinement, it is widely assumed that all experimental structural data may be deemed as obtained under similar conditions, and details of which may be safely ignored in such tasks. Such simplification was similarly utilized in implementing the LFEL approach in protein structure refinement[17] with focus being on coordinates without attending to thermodynamic and solvent conditions. Should detailed modeling of the variation of interested molecular systems under different environmental and/or thermodynamic conditions is desired, inclusion of these variables was essential. Here, previous simplified formulation is extended to deal with such scenarios. Denote environmental and thermodynamic variables (e.g. temperature, pressure, concentrations of relevant molecular species, special restraints) as $\Phi = (\phi_1, \phi_2, \cdots, \phi_k)$, molecular coordinates as $X = (x_1, x_2, \cdots, x_n)$ and local regions of molecular systems as $R = (R_1, R_2, \cdots, R_m)(m \leq n, m = n$ is preferred), the global joint probability density may be expressed by local distributions $P(\Phi, R_i)$ and their correlations as:

$$\begin{aligned} P(\Phi, X) &= P(\Phi, R) \\ &= \frac{P(\Phi, R)}{\prod_{i=1}^{m} P(\Phi, R_i)} \prod_{i=1}^{m} P(\Phi, R_i) \end{aligned} \tag{1}$$

It is important to note that each $R_i (i = 1, 2, \cdots, m)$ represents a dynamic collection of molecular coordinates for the ith specified region and changes with propagating trajectories. When $(m = n)$ or m is close to n, since each local region contains dozens of particles, overlapping among such regions are extensive. Local distributions are essentially LFEL/multi-dimensional potential of mean force (MPMF) for equilibrium systems. The fraction term $\frac{P(\Phi,R)}{\prod_{i=1}^{m} P(\Phi,R_i)}$ includes all complex global correlations among various local regions $R_i (i = 1, 2, \cdots, m)$ and is denoted the global correlation factor (GCF) previously.[18] The product term (hereafter "local term")$\prod_{i=1}^{m} P(\Phi, R_i)$ is simply to treat all local regions as if they were independent. If the GCF was ignored, then overlapping parts of different R_i may have distinct states. In reality, regardless of how many different local regions a molecule x_i participates, it has a unique physical state at any given instant. So all possible configurations with contradicting molecular states for any molecule participating different local regions have probability density zero. Such correction and additional modification of probability density is achieved by the GCF term. However, direct calculation of GCF is intractable for any realistic complex molecular system. Therefore, equation 1 is not directly useful for understanding and predicting behavior of molecular systems. How to approximately and effectively utilize this equation in practice is an open problem, and likely with many potential approximate solutions.

Probability density (free energy in equilibrium) of a specific configuration may be decomposed into three approximately independent contributions. The first is the short range contribution (F_{SR}) that measures the extent of structural stability/compatibility within each local region and is quantified by the local term in equation 1. The second contribution is from mediated interactions (F_{MED} Fig. 3ab) that measures the extent of compatibility among all overlapping local regions, and the third contribution measures direct long range (F_{LR}, Fig. 3b) compatibility within the whole molecular system. Both the second and the third contributions are contained in the GCF term. With the assumption that mediated interactions are independent from long-range interactions, the GCF may be approximately split into F_{MED} and F_{LR} as shown below.

$$\frac{P(\Phi, R)}{\prod_{i=1}^{m} P(\Phi, R_i)} \approx exp(-\sum F_{MED}(\Phi, R)) exp(-\sum F_{LR}(\Phi, R)) \quad (2)$$

The summation is over all mediated and long-range interactions in the given configuration R. In practical computation, separation of F_{SR} and F_{MED} is challenging on the one hand and inefficient on the other hand. In the previous implementation F_{SR} and F_{MED} were merged. Specifically, As shown in Fig 3b, at any given instant, a molecule (particle) in the system experiences free energy driving force additively from local distributions centered on each of its directly interacting neighbors within a preset cutoff. This is in strong contrast to regular MD simulations in which a particle experience direct forces from its directly interacting neighbors. While F_{LR} was not accounted for previously, it may be added in for each particle in each or every few propagation step(s). So in equation 1, local interactions are separated from the GCF, which may be approximately decomposed into mediated and long range interactions. However, local and mediated interactions were computed together in the previous implementation. This choice is somewhat counter intuitive but is feasible and efficient. Since an analytically clean mathematical factorization of the GCF is not available, it is likely that the above approximation is just one of many possible ways to realize practical computation. Distinct molecular systems may have different correlation characteristics and the optimal approximation is likely to be system specific. Nonetheless, the overall idea is quite clear, that is to first train local distributions, which are subsequently to be assembled to compose the global joint distribution (GJD) according to suitable approximation of the equation 1. The core idea of the LDT is to use local distributions to eliminate RLS.

A target molecular system may be propagated similarly as in the case of MD simulations except for the two differences. The first difference is that potential represented in MD is replaced by summation of LDT. The second is that a learning rate α_a, which is implicitly related to temperature is need to be given. The propagation may be carried out in different temperatures other than the one corresponding to the training data. Therefore, tricks such as simulated annealing may be realized just as in regular MD or MC simulations simply by assign a proper temperature specified by a gaussian noise term with variance α_b. In practice, α_a and α_b need not be identical in the following

Langevin equation:

$$X_{t+1} = X_t - \alpha_a \frac{\partial(\sum F_{SR} + \sum F_{MED} + \sum F_{LR})}{\partial X} + \epsilon, \epsilon \sim \mathcal{N}(0, \alpha_b) \tag{3}$$

Challenges and options for fitting local distributions

Training/learning of local terms is by no means trivial. In reality, strictly normalized local distributions is beyond reach and we may approximate them by complex high dimensional unnormalized potential functions. The direct consequence is that resulting free energy unit is arbitrary and is different for different molecular systems. When direct long range interactions are to be added, or comparison of results among different molecular systems are essential, this uncertainty has to be resolved. If long-range interactions with fixed unit may be calculated accurately, then it can serve as a unit-defining quantity among different molecular systems.

Construction of local distributions is essentially a density estimation problem in high dimensional space. Firstly, each local region need to be represented mathematically in a translation, rotation and permutation invariant way for its probability density to be effectively fitted. Such processing of molecular coordinates is termed descriptor function, it has accompanied development of neural network force fields (NNFF),[38,39] and is quite well understood. One possible way of defining a local region is to utilize the position of an given particle as origin for the local coordinates, so $R_i = (x_{i-c}, y_{i-s})$, with x_{i-c} being the origin of the local coordinates defined by a given unit and y_{i-s} being the coordinates of all surrounding molecules within a preset cutoff. It is important to note that the number of molecules may fluctuate and so is dimensionality of y_{i-s}, and padding is a feasible way to address it. So a local distribution is decomposed into local prior $P(y_{i-s})$ and local likelihood $P(x_{i-c}|y_{i-s})$ as shown below:

$$\begin{aligned} P(R_i) &= P(x_{i-c}, y_{i-s}) \\ &= P(x_{i-c}|y_{i-s})P(y_{i-s}) \end{aligned} \tag{4}$$

The likelihood term measures extent of match between the particle at the origin (x_{i-c}) and its surroundings. The prior term represent structural stability of the surrounding under given environmental conditions. In the protein structure refinement implementation,[17] identities of the central amino acids were utilized as labels to train a simple neural network representing likelihood terms and the prior terms were approximated with a simple weight. This strategy is likely to be not very useful for general molecular systems. For example, in a typical molecular system of dilute aqueous solution, the faction of water molecules is the overwhelming majority and training with identity will face extremely unbalanced data, and important differences among minority molecular/ionic species are likely to be lost. To improve fitting of local distributions, accurate description of both likelihood and prior terms are essential.

Like any density estimation application, fitting of local distributions may be carried out directly without decomposing into likelihood and prior terms. As a matter of fact, density estimation problem is of fundamental importance in both statistics and machine learning. Not surprisingly, many neural network architectures have been developed to tackle density estimation in high dimensional space where conventional methods (e.g. kernel density estimators[40]) are not effective. The most widely utilized two types are autoregressive models[41] and normalizing flows.[42,43] The former decompose a target joint density into product of conditional densities, which are modeled by parametric densities (e.g. mixture of gaussians) with trainable parameters. The later utilizing invertible neural network architectures to realize a direct quantitative map from a known density (uniform or gaussian) to the target density space. Establishment of proper correlations among different parametric densities is a highly challenging task for autoregressive models. The invertibility requirement in normalizing flow methodology impose heavy restrictions on neural network architecture and hence its representation power. One outstanding application example of normalizing flow in modeling molecular system is the Boltzmann generator (BG).[44] However, application of BG in complex molecular system remain to be tested. The fundamental difference between BG and LDT is that the former aims to directly model GJD for target molecular systems while the later decompose the problem into fitting and assembly of local distributions. Therefore RLS across different tasks is not addressed by BG. A recent more general approach, Roundtrip,[45] was proposed to overcome weakness of these two density estimation methodology. However, it takes an expensive sampling step to finalize the density estimation. Each available class of methods has their pros and cons, and no theory is available for selection of proper density estimation methodology presently. It might well that better methods will arise in future. For fitting local distributions in specific complex molecular system, many tests are likely necessary to construct a proper neural network model. Different molecular systems may have distinct structural distributions and case by case exploration is probably necessary to achieve high accuracy.

Energy based models (EBM)[46,47] are good candidates for fitting local distributions, either as a whole or when decomposed into priors and likelihood terms. In EBM, an energy is trained to be associated with a given configuration, thus eliminating the need of a normalization constant, which is a core challenge in fitting local distributions. Present tests of EBMs are mainly in conventional machine learning application scenarios such as computer vision or natural language processing.[48–51] Density distributions for such systems are quite different from complex molecular systems of condensed matter. Since LDT is a new development, significant effort is necessary to search for both proper loss functions, neural network architectures, optimization algorithms and their combinations for EBM to facilitate fitting local distributions in our interested molecular systems.

While neural networks have been black boxes with exceptional fitting capability up to date, and have been utilized with a wide variety of architectures. Efforts are undergoing for building white box neural networks.[52] To realize more physically interpretable and mathematically clean fitting of local distributions is certainly an attractive potential direction to explore.

Connection to conventional AI driven protein structure prediction and refinement

Contact map has played a critical role in development of protein structure prediction.[29] Earlier contact was a simple binary assignment (contact or not) defined by a cutoff distance based mostly on C_β atoms,[29] later on it evolved into residue pair distance distributions (RPDD).[20,24,25,27] Significant effort has been invested in investigating impact of various input information and neural network architectures on RPDD prediction with great progress in understanding. As the only known fully end-to-end protein structure refinement pipeline, GSFE-refinement[17] has a distinct overall pipeline from RPDD based algorithms of protein structure prediction/refinement. With the common goal of describing protein structures, these seemingly very different procedures have to be somehow connected. Fundamentally, all methodology of protein structures reflect the underlying free energy landscape from certain perspective. In GSFE-refinement, the GJD assembled from local distributions (or LFEL) lacks direct long-range correlations beyond spatial range of mediated interactions (Fig. 3) as the method stands now. Certainly, addition of long-range correlations is feasible as discussed above, and is in fact one important task in our future development plan. Sequence information is limited to the target protein in contrast to RPDD based methods, where multiple sequence alignment (MSA) information is usually included as input. In AlphaFold,[20] AlphaFold2 (https://deepmind.com/research/case-studies/alphafold) and many other RPDD based studies,[21,22,24–29] the core information obtained is explicit protein (family) specific RPDD, which are in fact marginalization of the GJD after integrating away all other variables except the distance between the concerning residues. Complex neural networks essentially realize a fitting from input information (protein sequence and MSA) to these marginal distributions without explicit construction of the GJD, approximation of which is the very goal of LDT based methods/models. As shown in Fig. 4, mapping from GJD to RPDD is readily achievable through marginalization. It is important to note that it takes some number of propagation steps (depending upon ruggedness of the underlying FEL) to obtain approximate GJD of sufficient accuracy assuming the underlying local distributions are sufficiently accurate. Marginalization is a deterministic procedure with significant loss of information, specifically correlations among different RPDD. Conversely, with RPDD, one may in principle construct GJD with sufficient sampling and optimization with necessary restraints. However, since correlations among different pair distributions are absent, resulting GJD is highly dependent upon parameters and algorithms utilized in the corresponding reconstruction process. Present mainstream AI-based protein structural prediction/refinement neural networks implicitly cache some projections of local distributions and rules for assembling them into RPDD, each comes with its own loss of information that is hard to retrieve. LDT theory aims to directly construct the most comprehensive GJD directly, thus has the full potential to perform dynamic modeling of relevant molecular processes as long as local distributions were fit for corresponding conditions. However, extending GSFE-refinement for accurately modeling dynamic protein folding is certainly not trivial as data on intermediate states are scarce presently. More importantly, LDT is a general theory applicable to any soft condense matter as long as fitting of corresponding local distributions is accomplished.

Potential extension to near equilibrium scenarios

At molecular scale, temperature, pressure and concentration of comprising molecules have significant fluctuations. In conventional MD simulations, temperature and pressure are usually controlled by various thermostats and barostats[53] with equilibrium assumption. If we have a heterogeneous cell being heated at one side, specifying temperature and pressure within it is a challenge. It might well be that both temperature and pressure are heterogeneous in a live cell (sometimes or always) and we just have no proper way of measuring. To specify temperature and pressure with thermostats and barostats is not a good way since we have no information on heterogeneous temperature in the first place. The probabilistic description of both molecular coordinates and thermodynamic/environmental variables can be of great utility in such scenarios. Assume the target molecular system is near-equilibrium. More specifically, all local distributions in target molecular system are well approximated by local distributions trained from equilibrium data while global molecular system is off equilibrium (e.g. having temperature/pressure gradient). In such scenario, we need thermodynamic variables to be associated with each local distributions. If the number of local regions is defined as the same as number of molecules/particles, we would have a set of relevant variables associated with each particle $\Phi_i = (\phi_{i1}, \phi_{i2}, \cdots, \phi_{ik})$ and denote the environmental conditions as $\mathbf{\Phi} = (\Phi_1, \Phi_2, \cdots, \Phi_n)$ The equation 1 may be expanded as shown below:

$$P(\mathbf{\Phi}, R) = \frac{P(\mathbf{\Phi}, R)}{\prod_{i=1}^{m} P(\Phi_i, R_i)} \prod_{i=1}^{m} P(\Phi_i, R_i) \tag{5}$$

With near-equilibrium assumption, we may safely learn local distributions from data collected in equilibrium states and relevant environmental conditions. However, propagation of global molecular systems by dynamic assembly of such local distributions is significantly more challenging. Continuity restraints of relevant $\mathbf{\Phi}$ variables is probably necessary, this may be realized through smoothing within certain spatial range. For equilibrium system, propagation of a molecular system under thermal fluctuation may be carried out with Langevin equation (equation 3) with a white noise term associated with a given temperature. However, in near equilibrium scenario, two choices maybe need to be made for propagating the molecular system. The first is utilize either maximum likelihood or bayesian approach to determine control variable at each molecule, with later being significantly more expensive. The second choice is to select a proper smoothing procedure to prevent large variance in control variables during the inference process. Assuming that the joint distribution $P(\Phi_i, R_i)$ has been learned with high accuracy, similar assembly and propagation procedures may be utilized as in the equilibrium case except with $\mathbf{\Phi}$ included and stochastic forces added according to corresponding temperature at each molecule. Large variance of parameters such as temperature and pressure may derail such simple treatment. Significant exploration and development is necessary in these regards. Nonetheless, this opens a potential highly efficient and probabilistic pathway for treatment of near equilibrium massive complex molecular systems (e.g. a cell).

Rapid automatic search for implicit manifold

Due to both local and long range interactions/correlations in condensed molecular systems, the real dimensionality of which is significantly smaller than that corresponds to nominal number of degrees of freedom (DOF). For example, considering 1000 rigid water model molecules in a rigid box, each with 6 DOFs. Its nominal number of DOF for the molecular system is 5997 but its real dimensionality is an unknown but significantly small number dependent upon environmental variables (e.g. temperature, pressure, container material). Local excluded volume interactions, Van der Waals interactions, hydrogen bonding networks, dipolar and multipolar interactions all contribute to correlations and dimensionality reduction in water. Conventional way of understanding underlying manifolds for molecular systems is to perform dimensionality reduction analysis on sufficiently sampled trajectories. However, principal component analysis (PCA) do not treat nonlinear correlations properly, many nonlinear algorithms have their own limitations.[38] More importantly, these dimensionality reduction methodologies are usually utilized as a post processing step for understanding molecular systems after expensive sampling dominated by RLS has been performed. So the goal is to understand manifolds as one of terminal goals, rather than utilizing manifolds to reduce computational cost. Dynamic assembly of local distributions is, however, fundamentally an implicit manifold search process on the one hand, and utilizes manifolds to reduce consumption of computational resources on the other hand. Learned local distributions are essentially implicit local manifolds under relevant conditions. Upon assembly of local distributions in propagation driven by derivatives of local distributions with respect to coordinates, a molecular system either stay on its manifold (free energy valleys) with fluctuations dependent upon temperature or rapidly return to the manifold when being away from it. To state alternatively, construction of GJD by assembly of local distributions according to equation 1 is equivalent to construction of global manifold by stitching together local manifolds embedded in local distributions without any manual intervention.

It is interesting to note that when viewed from the manifold perspective, LDT is effectively a completely automatic, significantly more accurate and more efficient implicit Metadynamics when local distributions were fit accurately and assembled properly. In Metadynamics, one first guess or compute for guiding collective variables (CVs), which is essentially an explicit representation of the manifold for a target molecular system in a given coordinate system. This is a highly challenging task, usually some iterative process is necessary but accuracy of resulting CVs has no guarantee, and no systematic theory is available for explicit searching of CVs. Subsequently explicit biases are accumulated to compute probability density of visited segments along CVs. In a properly implemented LDT, a target molecular system in propagation is automatically and implicitly maintained on its manifold, so the challenge of searching for CVs is met implicitly. Additionally, no bias is necessary and an unnormalized probability density is directly computed for each visited configuration.

Toward computation driven molecular sciences

Recent NNFF has demonstrated significant improvement in accuracy,[38,54–56] albeit with accompanying reduction of efficiency when compared with conventional atomistic MD simulations. With further development of density estimation/fitting, local distributions may be built from near quantum accuracy of NNFF based all atom simulations, and subsequently utilized to compose global distributions via dynamic assembly of local distributions as described by the LDT. Such combination may realize long-desired near-quantum accuracy and superior efficiency beyond conventional coarse grained models. With corresponding dramatic improvement of efficiency brought by LDT, many nanotechnology research may experience a transition from experimental driven to computation driven as spatial and time scales will be accessible by present and computational facility expected in a few years.

For computational molecular biology, lack of data is apparent as exemplified by AI based protein structure prediction, design and refinement studies where solvent and thermodynamic conditions need to be ignored. Deficiency of structural data is even more severe for denatured states of proteins and nucleic acids and other biomolecular systems (e.g. membranes). Presently, modeling of diverse thermodynamic and solvent conditions and denatured states relies heavily on all atom MD simulations, which are limited to micro-second time scales in routine investigations of typical proteins for small research groups, and simulation of large complexes and more extensive biomolecular systems is much more challenging. Development of LDT for efficient and accurate construction of local distributions, when combined with one-time near quantum level MD simulations for general biomolecular systems has the potential of bridging this gap, and realize routine simulations of large molecular complexes on realistic time scales (mini-seconds and longer). Many present experiments dominated molecular biology research (e.g. protein-protein interactions and protein-drug interactions) may experience transition to computation driven with dramatically improved efficiency. This is especially true for proteins and other biomolecules that are marginally stable and hard to express, and store under regular experimental conditions.

Establishment of a chain of tools from high level first principle calculations to simulation of large complex molecular systems has been long standing wish and efforts for molecular simulation community. Conventionally, coarse-graining has been the only available option and has made great contributions. Development and implementation of LDT in various general molecular systems provides a potential alternative pathway in this regard. However, to realize the potential, significant effort is necessary for development of algorithms in fitting local distributions for a wide variety of molecular systems. Condensed matter in general, and biological systems in particular, are organized in hierarchical structures with distinct correlation patterns over different length and time scales. Such characteristics was well summarized by Anderson[57] decades ago and significant efforts have been invested in multi-scale algorithm development in many subjects.[58–61] As discussed above, local distributions are essentially manifolds of local regions under various composition and environmental conditions. The specific meaning of "local" is dependent upon definition of comprising unit on the one hand , and upon length scales on the other hand. Implementation of LDT on multiple scale, and how should it interact with CG or evolve independently, is a fully open and mysterious field awaits intensive exploration.

Repetitive local computation and extension to solving various equations

View the world from a probabilistic perspective, variables in any systems constitute a configurational space that may be treated probabilistically. Solving integral/differential equations for numerical solutions are ubiquitous in many fields beyond molecular simulations, examples including but not limited to solid and fluid mechanics, meteorology and quantum chemistry. Regardless of time and spatial scales, all of these problems have variables associated with discretized spatial elements, local distributions of relevant variables are calculated numerous times by either different scientist in different projects or one scientist in the same project. With proper density estimation for local distributions and proper assembly strategies, such repetitive local computation (RLC) may be avoided with tremendous potential savings of computational resources. For any set of differential/integral equations in space X (which may be a tensor of various order and dimensions):

$$f(X) = 0 \tag{6}$$

Solving such equations numerically usually involves discretization of space. Each discretized element is associated with some variables and/or their derivatives, and correlations among variables of different elements are specified by the underlying equations (e.g. Navier-Stokes equation for fluid mechanics). Solutions for similar local distributions under typical environmental and boundary conditions of these variables have been obtained many times by many different scientists using similar or different softwares. With a probabilistic point of view, such local distributions are certainly learnable. Just as in the case of molecular simulations, global distributions in relevant space many be inferred from learned local distributions. Certainly, training/learning local distributions for such wide variety of solved equations is not trivial and it is likely that different strategies might be necessary for different equations. However, once relevant local distributions were established, they may be dynamically assembled to approximately construct relevant global distributions with given environmental and boundary conditions much faster than corresponding numerical computation process of solving complex equations. While we have not had a chance to implement this idea for any realistic problem, this speculation is a likely potential.

Conclusions and prospects

RLS in case of molecular simulations, or RLC in general, consumes large amount of computational resources on the one hand and slow down exploration of relevant research fields dramatically on the other hand. The LFEL approach was developed to address RLS previously. However, the formulation and its exemplary implementation in protein structural refinement, while demonstrated tremendous potentials, is limited to a single set of given environmental conditions. Here I propose the local distribution theory to generalize LFEL to address variable environmental conditions and near-equilibrium application scenarios. As a matter of fact, essentially all biological systems are off equilibrium to various extent. Despite the simple theoretical proposal presented here, extending implementation of LDT to near-equilibrium present great challenges and significant exploratory efforts are necessary. It is hoped that discussions and speculations herein stimulate more interest and attract more scientists in further development and application of the local distribution theory.

Acknowledgement

I thank Professor Jingkai Gu and Professor Yaoqi Zhou for comments and encouragement when this theory was conceived.

References

(1) 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.

(2) 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.

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

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

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

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

(7) 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.

(8) 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.

(9) 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.

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

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

(12) 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.

(13) 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.

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

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

(16) 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.

(17) Cao, X.; Tian, P. Molecular free energy optimization on a computational graph. *RSC Adv.* **2021**, *11*, 12929–12937.

(18) Cao, X.; Tian, P. “Dividing and Conquering” and “Caching” in Molecular Modeling. *International Journal of Molecular Sciences* **2021**, *22*.

(19) 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.

(20) Senior, A. W. et al. Improved protein structure prediction using potentials from deep learning. *Nature* **2020**, *577*, 706–710.

(21) Suh, D.; Lee, J. W.; Choi, S.; Lee, Y. Recent Applications of Deep Learning Methods on Evolution- and Contact-Based Protein Structure Prediction. *International Journal of Molecular Sciences* **2021**, *22*.

(22) Pakhrin, S. C.; Shrestha, B.; Adhikari, B.; KC, D. B. Deep Learning-Based Advances in Protein Structure Prediction. *International Journal of Molecular Sciences* **2021**, *22*.

(23) Skolnick, J.; Gao, M. The role of local versus nonlocal physicochemical restraints in determining protein native structure. *Current Opinion in Structural Biology* **2021**, *68*, 1–8, Protein-Carbohydrate Complexes and Glycosylation Sequences and Topology.

(24) Ju, F.; Zhu, J.; Shao, B.; Kong, L.; Liu, T.-Y.; Zheng, W.-M.; Bu, D. CopulaNet: Learning residue co-evolution directly from multiple sequence alignment for protein structure prediction. *Nature Communications* **2021**, *12*, 2535.

(25) Jing, X.; Xu, J. Improved protein model quality assessment by integrating sequential and pairwise features using deep learning. *Bioinformatics* **2020**, *36*, 5361–5367.

(26) Zhao, K.-l.; Liu, J.; Zhou, X.-g.; Su, J.-z.; Zhang, G.-j. Structural bioinformatics MMpred : a distance-assisted multimodal conformation sampling for de novo protein structure prediction. *Bioinformatics* **2021**, 1–8.

(27) Xia, Y.-h.; Peng, C.-x.; Zhou, X.-g.; Zhang, G.-j. Structural bioinformatics A Sequential Niche Multimodal Conformational Sampling Algorithm for Protein Structure Prediction. *Bioinformatics* **2021**, 1–9.

(28) Wu, F.; Xu, J. Deep template-based protein structure prediction. *PLoS computational biology* **2021**, *17*, 1–18.

(29) Zhang, H.; Bei, Z.; Xi, W.; Hao, M.; Ju, Z. Evaluation of residue-residue contact prediction methods : From retrospective to prospective. *PLOS Comput. Biol.* **2021**, *17*, 1–33.

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

(31) 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.

(32) 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.

(33) 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.

(34) 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.

(35) 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.

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

(37) 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.

(38) 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.

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

(40) Parzen, E. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics* **1962**, *33*, 1065 – 1076.

(41) Uria, B.; Côté, M.-A.; Gregor, K.; Murray, I.; Larochelle, H. Neural Autoregressive Distribution Estimation. *Journal of Machine Learning Research* **2016**, *17*, 1–37.

(42) Kobyzev, I.; Brubaker, M. A. Normalizing Flows: Introduction and Ideas. *ArXiv* **2019**, 1–35.

(43) Papamakarios, G.; Nalisnick, E. Normalizing Flows for Probabilistic Modeling and Inference. *ArXiv* **2019**, 1–60.

(44) Noé, F.; Olsson, S.; Kohler, J.; Wu, H. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* **2019**, *365*, eaaw1147.

(45) Liu, Q.; Xu, J.; Jiang, R.; Hung, W. Density estimation using deep generative neural networks. *Proceedings of the National Academy of Sciences* **2021**, *118*.

(46) Cun, Y.; Huang, F. Loss functions for discriminative training of energy-based models. AISTATS 2005 - Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics. 2005; pp 206–213, 10th International Workshop on Artificial Intelligence and Statistics, AISTATS 2005 ; Conference date: 06-01-2005 Through 08-01-2005.

(47) BakIr, G.; Hofmann, T.; Schölkopf, B.; Smola, A. J.; Taskar, B.; Vishwanathan, S. *Predicting Structured Data*; The MIT Press, 2007.

(48) Zhao, J.; Mathieu, M.; Lecun, Y.; Artificial, F. Energy-based generative adversarial networks. *ArXiv* **2017**, 1–17.

(49) Liu, M.; Yan, K.; Oztekin, B.; Ji, S. GraphEBM: Molecular Graph Generation with Energy-Based Models. *ArXiv* **2020**,

(50) Mordatch, I. Compositional Visual Generation with Energy Based Models. *ArXiv* **2020**,

(51) Grathwohl, W.; Duvenaud, D. Your classifier is secretly an energy based model and you should treat it like one. *ArXiv* **2020**, 1–23.

(52) Chan, K. H. R.; Yu, Y.; You, C.; Qi, H.; Wright, J.; Ma, Y. ReduNet: A White-box Deep Network from the Principle of Maximizing Rate Reduction. 2021.

(53) Frenkel, D.; Smit, B. In *Understanding Molecular Simulation (Second Edition)*, second edition ed.; Frenkel, D., Smit, B., Eds.; Academic Press: San Diego, 2002; pp 139–163.

(54) Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W. Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Phys. Rev. Lett.* **2018**, *120*, 143001.

(55) Bedolla, E.; Padierna, L. C.; Castañeda-Priego, R. Machine learning for condensed matter physics. *Journal of Physics: Condensed Matter* **2020**, *33*, 053001.

(56) Lu, D.; Wang, H.; Chen, M.; Lin, L.; Car, R.; E, W.; Jia, W.; Zhang, L. 86 PFLOPS Deep Potential Molecular Dynamics simulation of 100 million atoms with ab initio accuracy. *Computer Physics Communications* **2021**, *259*, 107624.

(57) Anderson, P. More is different. *Science* **1972**, *177*, 393–396.

(58) Franco, A. A.; Rucci, A.; Brandell, D.; Frayret, C.; Gaberscek, M.; Jankowski, P.; Johansson, P. Boosting Rechargeable Batteries R&D by Multiscale Modeling: Myth or Reality? *Chemical Reviews* **2019**, *119*, 4569–4627, PMID: 30859816.

(59) Henning, P.; Peterseim, D. Oversampling for the Multiscale Finite Element Method. *Multiscale Modeling & Simulation* **2013**, *11*, 1149–1175.

(60) Abdulle, A.; Weinan, E.; Engquist, B.; Vanden-Eijnden, E. The heterogeneous multiscale method. *Acta Numerica* **2012**, *21*, 1?87.

(61) Ramstead, M. J. D.; Kirchhoff, M. D.; Constant, A.; Friston, K. J. Multiscale integration: beyond internalism and externalism. *Synthese* **2021**, *198*, 41–70.

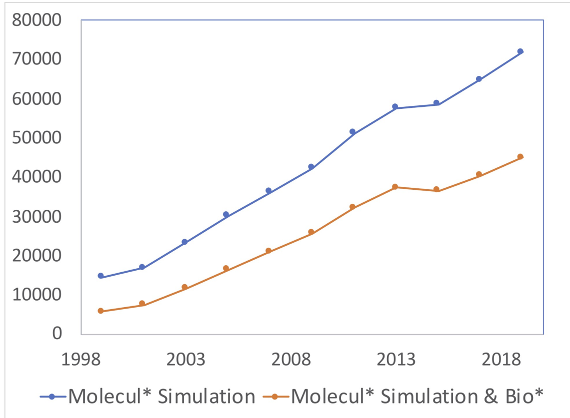

Figure 1: The number of publications retrieved from web of science on Jun. 1st 2021 with subject word "molecul* simulation" and "molecul* simulation & bio" respectively. The corresponding time frame is every two years starting from 1999. The first data point is the number of papers published in year 1999 and 2000.

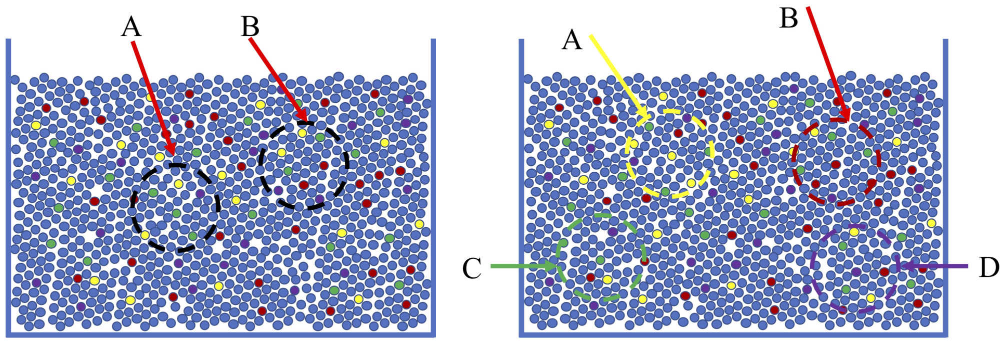

Figure 2: Schematic illustration of RLS. Left: the spatial perspective. A) and B) are two different spherical bulk spaces. We expect the same local distributions after sufficiently long simulations of the whole molecular system. In such cases, spherical and partial spherical spaces near or on interfaces have different local distributions from that of the bulk, special treatment of such spherical spaces engenders significant difficulty. Right: indistinguishable particle and GSFE perspective. All particles of the same species are indistinguishable, so should be local distributions of local regions defined by spherical spaces with such a particle as the origin. This removes the need for special treatment of all interfacial issues as different interfaces may be simply defined as more cases of particle packing surrounding a given particle with well defined statistical weight under given thermodynamic and environmental conditions. A), B), C) and D) are examples of surrounding local regions of different particle species.

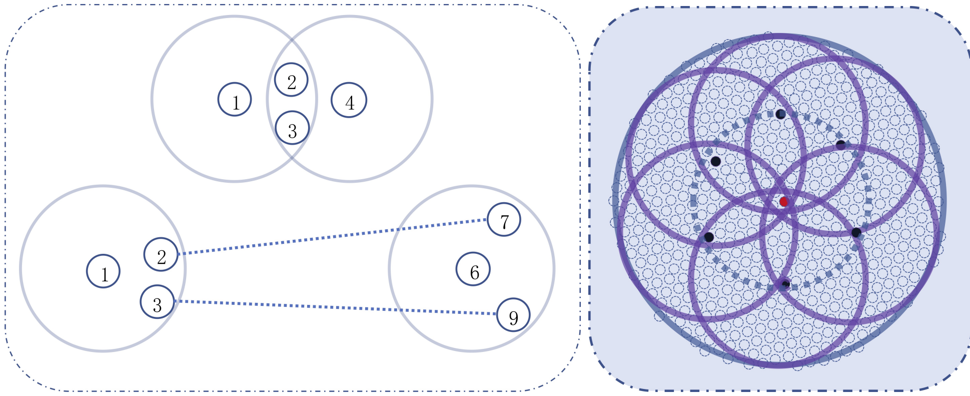

Figure 3: Schematic representation of the short range, mediated interactions and long range interactions as implemented in ref. Left: particles (1,2,3), (2,3,4) and (6,7,9) are directly interacting with short range interactions. (1,4) are interacting through mediation by (2,3), (2,7) and (3,9) have direct long range interactions. Right: here the focus is the central red particle, which define a region with boundary being shown as dotted partially transparent blue line. Each of all other particles within this region defines a local distribution, six of the most further of such regions are represented as purple circles. The central red particle experience forces from all of local distributions surrounding each of its neighbors. In this way, short range and mediated interactions are effectively accounted for simultaneously. In summary, for the central red particle, it experiences short range interactions from particles within the dotted partial transparent blue circle, mediated interactions from particles between the dotted blue circle and large solid blue circle, long range interactions from the region outside the large blue circle.

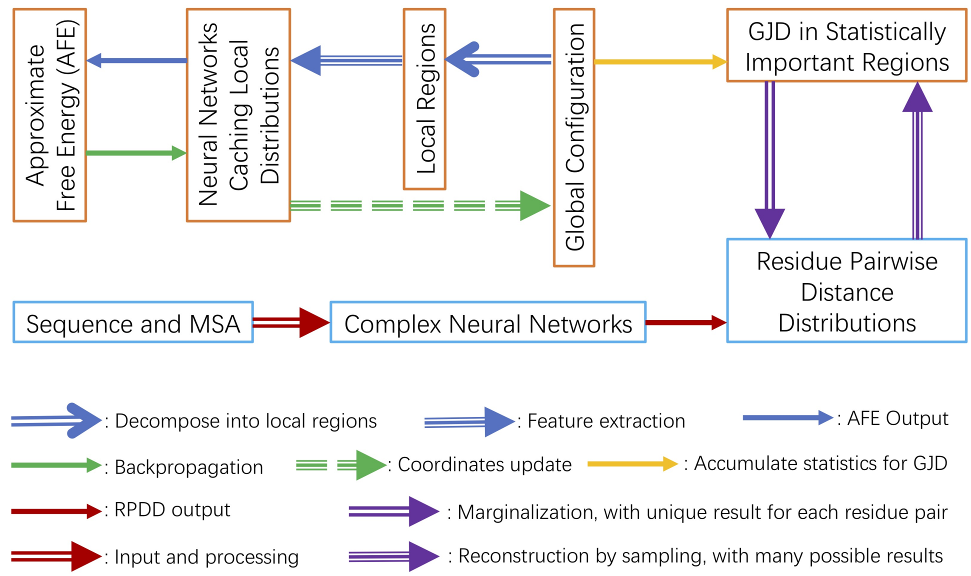

Figure 4: Schematic comparison between LDT based end-to-end protein structure modeling (top orange boxes) and mainstream RPDD based protein structure prediction and refinement schemes (bottom blue boxes). It is important to note that LDT based modeling aims to generate the GJD, which is the most comprehensive information for any complex molecular system and is generally applicable. The marginalization from the GJD to pairwise residue distance distributions is an irreversible process with deterministic results and significant information loss on correlations among different pairwise distances. The converse process is a highly expensive process with expensive sampling and optimization involved, due to complexity of correlations among different distances, resulting global distribution is highly dependent both on initialization and the optimization procedures.