

# Making high-dimensional molecular distribution functions tractable through Belief Propagation on Factor Graphs

Zachary Smith<sup>1</sup> and Pratyush Tiwary<sup>2</sup>

<sup>1</sup>*Biophysics Program and Institute for Physical Science and Technology, University of Maryland, College Park 20742, USA.*

<sup>2</sup>*Department of Chemistry and Biochemistry and Institute for Physical Science and Technology, University of Maryland, College Park 20742, USA.*

(Dated: 28 June 2021)

Molecular dynamics (MD) simulations provide a wealth of high-dimensional data at all-atom and femtosecond resolution but deciphering mechanistic information from this data is an ongoing challenge in physical chemistry and biophysics. Theoretically speaking, joint probabilities of the equilibrium distribution contain all thermodynamic information, but they prove increasingly difficult to compute and interpret as the dimensionality increases. Here, inspired by tools in probabilistic graphical modeling, we develop a factor graph trained through belief propagation that helps factorize the joint probability into an approximate tractable form that can be easily visualized and used. We validate the study through the analysis of the conformational dynamics of two small peptides with 5 and 9 residues. Our validations include testing the conditional dependency predictions through an intervention scheme inspired by Judea Pearl. Secondly we directly use the belief propagation based approximate probability distribution as a high-dimensional static bias for enhanced sampling, where we achieve spontaneous back-and-forth motion between metastable states that is up to 350 times faster than unbiased MD. We believe this work opens up useful ways to thinking about and dealing with high-dimensional molecular simulations.

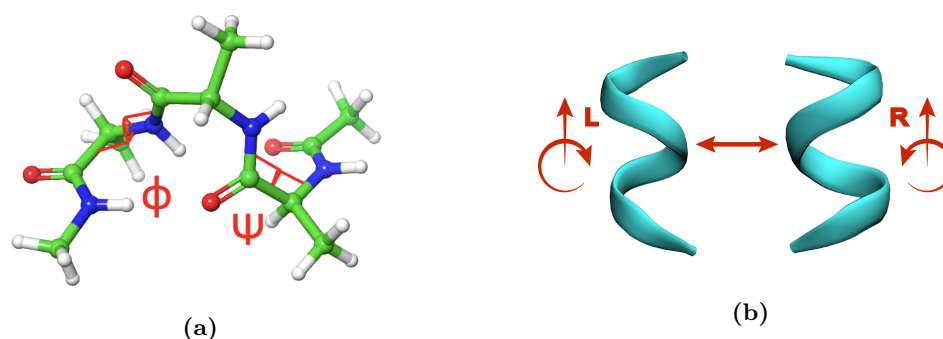
## I INTRODUCTION

Understanding the relationships between the many different degrees of freedom constituting a generic high-dimensional molecular system is a problem of longstanding theoretical and practical interest.<sup>1–3</sup> For a system with  $k \gg 1$  degrees of freedom, in principle one could consider the full  $k$ -body probability distribution  $P$ , which would account for any and all correlations between the  $k$  different components. However, a full  $k$ -body probability distribution is not always necessary as can be intuited with the popularity of the Ising model, where nearest-neighbor interactions can capture arbitrarily long-range communication and correlations, and the full joint probability is not needed to model these relationships. Furthermore, fitting such a  $k$ -dimensional  $P$  to data would involve  $k$ -dimensional histograms which becomes impractical as  $k > 2 \sim 3$ . In fact, learning tractable probabilistic models for high dimensional data is a problem that extends far beyond molecular sciences and pervades numerous aspects of modern day life and data science.<sup>4</sup> Understanding the dependencies in data is thus the first step toward understanding the processes that generated the data, and for further simulating these processes efficiently through advanced sampling methods for instance.<sup>5</sup>

In this work our specific interest is in data coming from Molecular dynamics (MD) simulations, a tool very commonly used to study the dynamics of chemical systems at all-atom spatial resolution and femtosecond temporal resolution. This leads to a deluge of high-dimensional data which can be hard to make sense of. Specifically for biomolecules, here we would like to characterize which residues are fluctuating in a mutually correlated manner, and use this information to better understand the gov-

erning biophysical processes. Furthermore, due to the high temporal resolution of MD, timescales associated with practical problems such as protein folding and drug unbinding are inaccessible even using most modern high power computing resources. Thus we are also interested in using these tractable probability distributions governing the high-dimensional MD to enhance the simulation itself.<sup>6</sup> To do so we propose a framework using methods from probabilistic graphical models<sup>4</sup> to construct a factorization of input degrees of freedom. Among other works in this area, our work is inspired by Ref. 7–8. Given our eventual interest in sampling slow, representative degrees of freedom, we take the liberty to call these degrees of freedom as order parameters (OPs). We chose to use probabilistic graphical models taking inspiration from their successes in a variety of fields ranging from identifying disaster victims<sup>9</sup> to decoding messages<sup>10</sup> where the relationships between unknowns are leveraged to perform an inference task with levels of efficiency that wouldn't be possible otherwise.

The key ideas, detailed in Sec. II, can be summarized as follows. We start with high-dimensional data with complicated and *a priori* unknown dependencies between the various OPs, with the ultimate objective of approximating this intractable high-dimensional probability distribution with a factorized, tractable probability comprising up to triplet terms. For this, we first assess the conditional dependencies between the OPs using Markov random fields (MRF)<sup>11</sup> detailed in section II A. MRFs are graphical models where non-adjacent nodes are conditionally independent. We use graphical lasso,<sup>12</sup> an estimator for the inverse covariance matrix to determine the conditional dependence relationships used in the MRF. In Sec. II B we describe how these conditional depen-



**FIG. 1:** The two model systems simulated in this work capped Ala<sub>3</sub> (a) and Aib<sub>9</sub> peptide (b). Alanine peptide atoms are shown with an example of a  $\phi$  and  $\psi$  angle overlaid in red. Aib<sub>9</sub> peptide is shown as a ribbon to highlight its helical secondary structure.

dependency relationships are converted into a bipartite graph known as factor graph, where the two disjoint sets of vertices denote respectively OPs and the functions connecting them. These functions are then learnt using the belief propagation algorithm (Sec. II C).<sup>8,13</sup> In order to assess the benefits of this factor graph-based approach we apply it to simulations of two well-studied small peptides<sup>14–16</sup> Ala<sub>3</sub> and Aib<sub>9</sub> peptide that display rich conformational dynamics in spite of their small sizes (Fig. 1), and can be described using their  $\phi$  and  $\psi$  dihedral angles. The factor graphs are validated using a range of additional simulations in Sec. III, which include (i) an intervention protocol in the sense of Pearl<sup>17</sup> that directly confirms our predictions of conditional dependency between different degrees of freedom and (ii) a demonstration of how this knowledge can lead to up to three orders of magnitude enhanced sampling along all degrees of freedom, with hysteresis-free back-and-forth movement between different metastable states. We believe this work represents a new use of ideas from the field of probabilistic graphical models<sup>4</sup> in the study of molecular simulations aiding with better interpretations and superior sampling.

## II METHODS

In this Section we describe the theory underlying our key ideas as summarized in the Introduction above, as well as the workflow for its implementation (Fig. 2).

### II.A Markov Random Fields (MRF)

Our starting point is sampling of high-dimensional OPs arising from molecular dynamics (MD) simulations (see supplementary information (SI) for details) but the proposed procedure should be more generally applicable. In order to partition OPs into simpler lower-dimensional factors, we first assess the relationships between the OPs themselves. Consider any 3 random variables A, B, and C. A particularly informative measure of the dependen-

cies between such variables is conditional independence, denoted  $(A \perp\!\!\!\perp B)|C$ .<sup>18</sup> This measure is used in the pairwise Markov property of Markov random fields (MRF), and denotes that the two variables A and B are statistically independent given knowledge of C. Mathematically this can be written as

$$(A \perp\!\!\!\perp B)|C \iff p(A \cap B|C) = p(A|C)p(B|C) \quad (1)$$

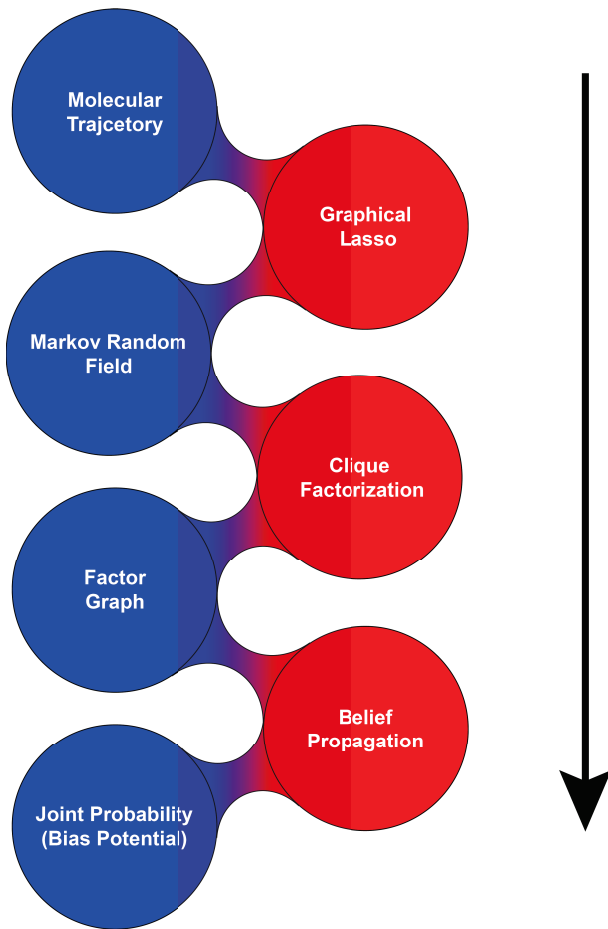
with  $p$  denoting respective probability densities. Here we call variables equivalently as nodes shifting to a graphical modeling perspective. Generalizing Eq. 1 to all possible nodes, the pairwise Markov property is satisfied if two nodes are conditionally independent given all other nodes as shown in Eq. 2, where  $X_i$  and  $X_j$  denote nodes and  $X_{N \setminus \{i,j\}}$  denotes all other nodes.

$$X_i \perp\!\!\!\perp X_j | X_{N \setminus \{i,j\}} \quad (2)$$

A MRF is an undirected graphical model such that all nodes without edges between them have the pairwise Markov property and those with edges do not have it.<sup>4</sup> This can be interpreted as meaning that any correlation between connected nodes can not be “explained away” by a third node causing the correlation. The presence of the pairwise Markov property is estimated for all possible edges by estimating an inverse covariance matrix ( $\Theta$ ). All pairs of variables  $X_i$ ,  $X_j$  that have a zero at the associated position in the inverse covariance matrix ( $\Theta_{i,j}$ ) can be expected to have the pairwise Markov property.<sup>4</sup> To efficiently calculate this inverse covariance matrix, we use the graphical lasso<sup>12</sup> estimator shown in Eq. 3:

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} \left[ \operatorname{tr}(\Sigma\Theta) - \log \det(\Theta) + \lambda \sum_{j \neq k} |\Theta_{jk}| \right] \quad (3)$$

where  $\Theta$  is the inverse covariance,  $\hat{\Theta}$  is the inverse covariance returned by graphical lasso,  $\Sigma$  is the observed covariance,  $\lambda$  is a parameter that controls how strongly off-diagonal terms are penalized,  $\operatorname{tr}$  is the matrix trace, and  $\det$  is the matrix determinant. This  $\lambda$  parameter can be effectively used to control the sparsity of the resulting MRF as is shown later for our examples (Fig. 6). In practice, we tune this parameter to produce the most densely



**FIG. 2:** Schematic of the workflow done in this manuscript as a bipartite graph with objects shown in blue and algorithms shown in red. The workflow proceeds from top to bottom.

connected MRF that contains at most triplet interactions (in other words, the factor graph later described in Sec. IIB should have clique size no larger than three).

## II.B Factor Graphs (FG)

Once we have a MRF, we can use this graph structure to partition the OPs into independent sets or factors that make up the joint distribution. It is important to note that these independent sets are not necessarily disjoint, and the same variable could be a part of different factors as we show for practical examples in Sec. III. The structure of these factors can be obtained from the cliques, or sets of fully connected nodes, of the MRF. Each maximal clique, or clique that is not part of a larger clique, contains the variables in one factor of the joint probability. This relationship is illustrated through a schematic in Fig. 3a. It is important to note that at this point we know the variable sets used for each factor but not their functional form. More technically, this is defined in Eq.

4 where  $C(G)$  is the set of all maximal cliques on a graph  $G$  and  $f_c(X_c)$  is each factor as a function of the variables contained in the associated clique:

$$p(X) = \prod_{c \in C(G)} f_c(X_c) \quad (4)$$

This clique factorization can be shown more simply in another graph structure called a factor graph (FG). Factor graphs are bipartite graphs with nodes that are either factors or variables, and edges only connecting nodes between these two sets, but not within the sets (i.e. function node connected to variable node is the only allowed edge). This shows which variables are associated with which factor and how the factors are related through mutual connection to the same variable. This representation also simplifies the representation of the joint probability as shown in Eq. 5 because the product is now over factor nodes ( $F$ ) and their associated edges ( $X_a$ ) instead of cliques ( $C(G)$ ) and the nodes that compose them ( $X_c$ ).

$$p(X) = \prod_{a \in F} f_a(X_a) \quad (5)$$

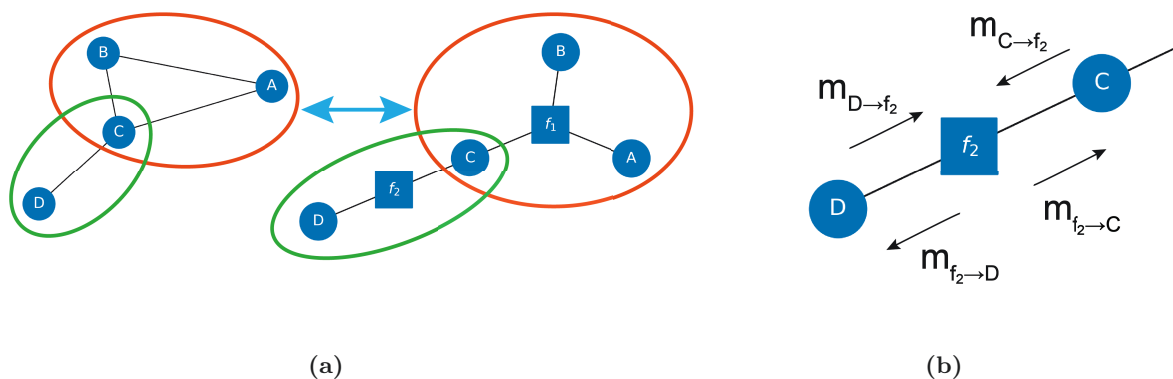
## II.C Belief Propagation

Thus far we have determined the relationship between our set of OPs using graphical lasso and converted this structure into a factor graph using clique factorization. We now know the OP sets we will use for each factor but have not learned the factors themselves or the joint probability they compose. We do so with the belief propagation (also known as sum-product) algorithm.<sup>13,19,20</sup> This method updates an approximate factorized model  $q(X)$  of the joint distribution defined in Eq. 6 and composed of approximate factors  $\tilde{f}_a(X_a)$ :

$$q(X) = \prod_{a \in F} \tilde{f}_a(X_a) \quad (6)$$

The objective of the belief propagation (BP) algorithm, first proposed by Pearl in 1982,<sup>13</sup> is to approximate factors  $\tilde{f}_a(X_a)$  that will be iteratively updated so that their product  $q(X)$  approaches the true joint distribution  $p(X)$  in Eq. 5. To learn such factors, estimates of all other approximate factors are communicated to any single given factor at a time which is then adjusted depending on this “belief” about the other factors. The adjustment itself is carried out so that  $p$  and  $q$  mirror each other as closely as possible given the beliefs.

We now detail the BP equations that are used to calculate marginal probabilities and update messages. Here we index factors with  $a, b$  and variables with  $i, j$ . We denote sets of variables and factors with capital symbols  $X$  and  $F$  while the individual nodes comprising them are indicated with lowercase symbols  $x$  and  $f$  respectively. The sets  $F$  and  $X$  contain all nodes connected to the node with their index, for example  $X_a$  represents all variable nodes  $x$  connected to the factor  $f_a$ .



**FIG. 3:** Depictions of clique factorization and message passing for an arbitrary set of variables where  $A - D$  represent variables and  $f_i$  represent factors. (a) shows an MRF on the left with two cliques circled in orange and green converted into two factors circled in the same colors in the FG on the right. (b) shows all of the incoming and outgoing messages for factor  $f_2$  in the factor graph shown in (a).

The central idea in our workflow is to mix kernel density estimation<sup>21,22</sup> with traditional belief propagation. Specifically, we use kernel density estimation to model the marginal associated with each factor  $p(X_a)$ . We initialize all messages  $m$  to arrays of ones equivalent to passing no information about their factors. After this initialization step, our iteration begins as follows:

1. Select a factor to update, this is done at random for our simple case but could use more sophisticated methods.<sup>23</sup>
2. Apply Eq. 7 updating  $\tilde{f}_a(X_a)$  such that  $q(X_a) = p(X_a)$  from the kernel density estimator.

$$q(X_a) = \tilde{f}_a(X_a) \prod_{x_i \in X_a} m_{x_i \rightarrow f_a}(x_i) \quad (7)$$

3. Apply Eq. 8 to update the factor's outgoing messages.

$$\begin{aligned} m_{f_a \rightarrow x_i}(x_i) &= \sum_{X_a \setminus x_i} \tilde{f}_a(X_a) \prod_{x_j \in X_a \setminus x_i} m_{x_j \rightarrow f_a}(x_j) \\ &= \sum_{X_a \setminus x_i} \frac{q(X_a)}{m_{x_i \rightarrow f_a}(x_i)} \end{aligned} \quad (8)$$

4. Apply Eq. 9 to update the messages outgoing from each variable associated with the factor.

$$m_{x_i \rightarrow f_a}(x_i) = \prod_{f_b \in F_x \setminus f_a} m_{f_b \rightarrow x_i}(x_i) \quad (9)$$

5. Repeat steps 1 – 4 until factors converge.

This procedure yields a factorized model for the joint probability that when marginalized will reproduce the

marginal probabilities from kernel density estimation. This particular implementation of belief propagation was designed to be easily extended to use time evolving density estimators similar to that of metadynamics.<sup>6,24,25</sup>

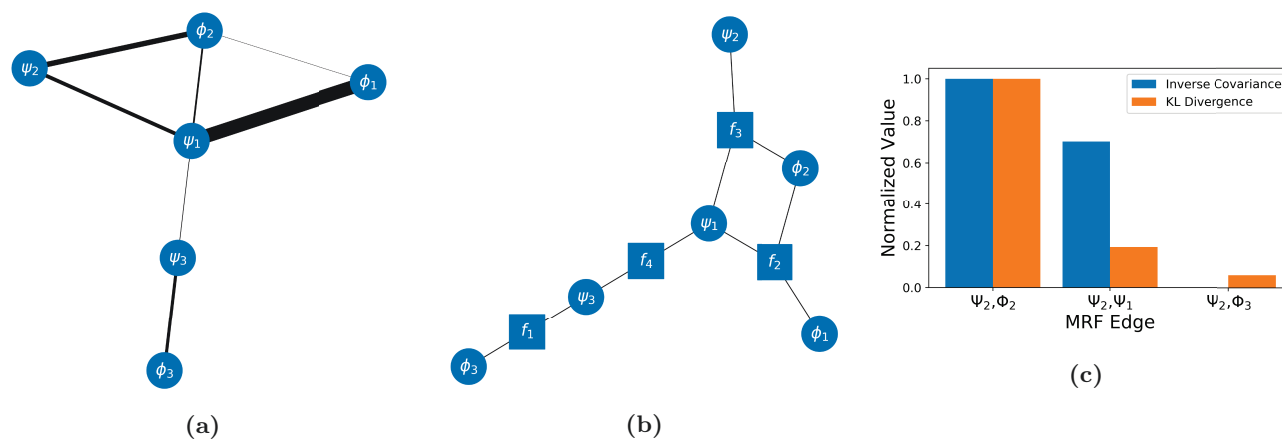
### III RESULTS

In this section, we report detailed results of applying the methods described so far in this manuscript to two test biomolecules, namely capped  $\text{Ala}_3$  and  $\text{Aib}_9$ .<sup>14,16,26,27</sup> These respectively comprise 5 and 9 residues in explicit water, thus containing many complicated intertwined degrees of freedom. While the methods could easily be extended to consider solvent degrees of freedom as well,<sup>28</sup> here we focus on the protein torsional degrees of freedom as the input OPs, numbering 6 and 18 for  $\text{Ala}_3$  and  $\text{Aib}_9$  respectively. For both systems we learn Markov random fields using graphical lasso. In addition, for  $\text{Ala}_3$  we learn a factor graph and use belief propagation to estimate the factors themselves. Thus we reduce the 6- and 18- dimensional probability distributions respectively to a combinations of pair and triplet interactions. The learned approximate lower-dimensional interactions are then analyzed and validated for both the systems in different ways, including intervention,<sup>17</sup> enhanced sampling,<sup>5,29</sup> and network analysis.

#### III.A Capped $\text{Ala}_3$

We performed 1  $\mu\text{s}$  long unbiased MD simulation (details in SI) recording three  $\phi$  and three  $\psi$  dihedral angles every 200 fs. The trajectory of these six OPs was used to construct a MRF using graphical lasso as described in Sec. II A and this MRF was converted to a FG using





**FIG. 4:** Graphical models of Ala<sub>3</sub> calculated using a 1  $\mu$ s long molecular dynamics trajectory and comparison of the intervention results with the inverse covariance calculated with graphical lasso. (a) shows the MRF from graphical lasso with edge thickness showing the underlying inverse covariance, i.e. pairs of variables with thinner edges between them are more conditionally independent. (b) shows the FG calculated using clique factorization on (a). (c) shows the result of the intervention test that validates (a), by comparing the inverse covariance to the KL divergence between the observed probability during the intervention test and the independent probability model. We see that graphical lasso will favor drawing edges that clearly do not have the pairwise Markov property.

clique factorization as described in Sec. II B. The MRF and the FG so learned are shown in Fig. 4.

### III.A.1 Validating the MRF through intervention

In order to validate this MRF, we perform a series of enhanced sampling simulations in the spirit of the intervention approach pioneered by Pearl.<sup>17</sup> For each of these simulations we chose two OPs and restrained all other dihedrals in the MRF using a quadratic bias potential. This intervention procedure prevents any large fluctuations in the restrained OPs allowing us to directly assess the conditional independence hypothesis between the unrestrained OPs. Specifically, if a given pair of OPs are indeed conditionally independent given other OPs, then in the presence of such a restraint, the joint probability for this pair of OPs should be the same as the product of their marginal probabilities. Thus, in light of the MRF shown in Fig. 4, under respective constraints, we expect the joint probability to be closest to the product of constituent marginals for the pair  $\Psi_2, \Phi_2$ , followed by  $\Psi_2, \Psi_1$  and finally for  $\Psi_2, \Phi_3$ . In order to assess this hypothesis, we use the Kullback–Leibler (KL) divergence<sup>30</sup> to measure how close the sampled distribution is to the product of marginals. This represents an assessment of how well Eq. 1 models the sampled distribution. In Fig. 4 we show that indeed this is the case by comparing the KL divergence to the magnitude of the inverse covariance from graphical lasso. We see that the order of the inverse covariance and KL divergence match showing that graphical lasso did draw edges that lack conditional independence. The simulations were run for 1  $\mu$ s using all simulation details (except the restraint) identical to the

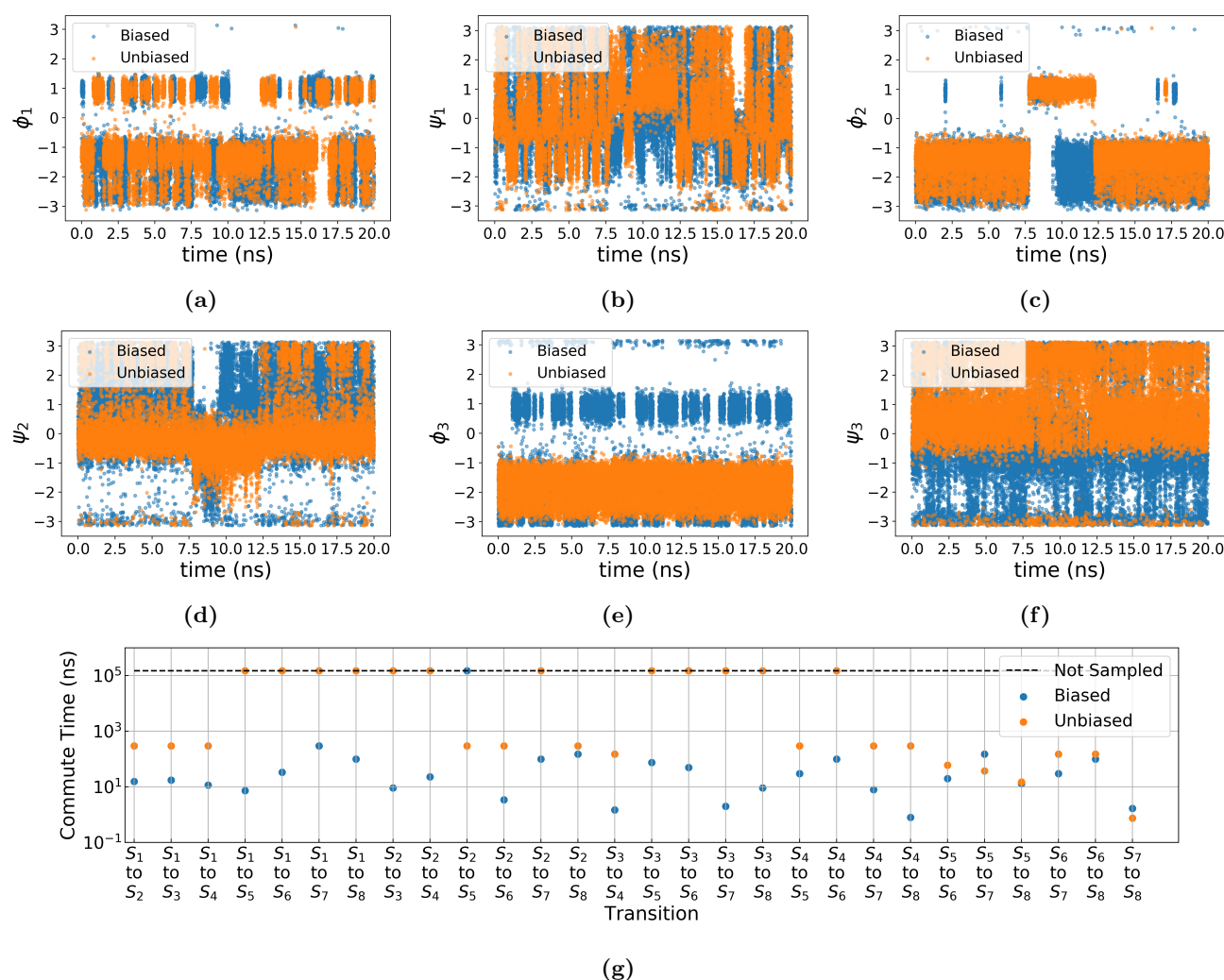
initial Ala<sub>3</sub> simulations, detailed in SI.

### III.A.2 Validating the FG through enhanced sampling

With confidence in the MRF through the intervention procedure demonstrated in Fig. 4, these MRFs were then converted to factor graphs with up to triplet interactions. Given such a joint probability distribution expressed as product of terms up to triplets, we can directly use it to build a very high-dimensional bias potential that enhances the sampling simultaneously in all degrees of freedom, as defined below:

$$V(X) = k_B T \log(p(X)) \approx k_B T \log(q(X)) \quad (10)$$

The central idea here is that the exact high-dimensional probability distribution  $p(X)$ , if we knew it, would have led to the ideal bias for importance sampling which perfectly enhanced all fluctuations in the molecular system.<sup>6</sup> The quality of the approximation  $q(X)$  in Eq. 6 can thus be ascertained by using it to construct an approximate bias as in Eq. 10. When added as a static bias to the system’s potential energy, a good approximation  $p \approx q$  should then lead to spontaneous and enhanced back-and-forth movement between different metastable states. We convert the joint probability  $q$  to a bias potential using Eq. 10 with the caveat that we do not bias any regions of configurational space with a sampled probability of less than 0.001 in order to avoid adding bias to transition states. This threshold can be made higher or lower without affecting the nature of this discussion. The output of this calculation is a set of static external bias files to be used with PLUMED.<sup>31–33</sup> The factor graph bias was used to conduct a 150 ns simulation using the same parameters used in the previous unbiased simula-

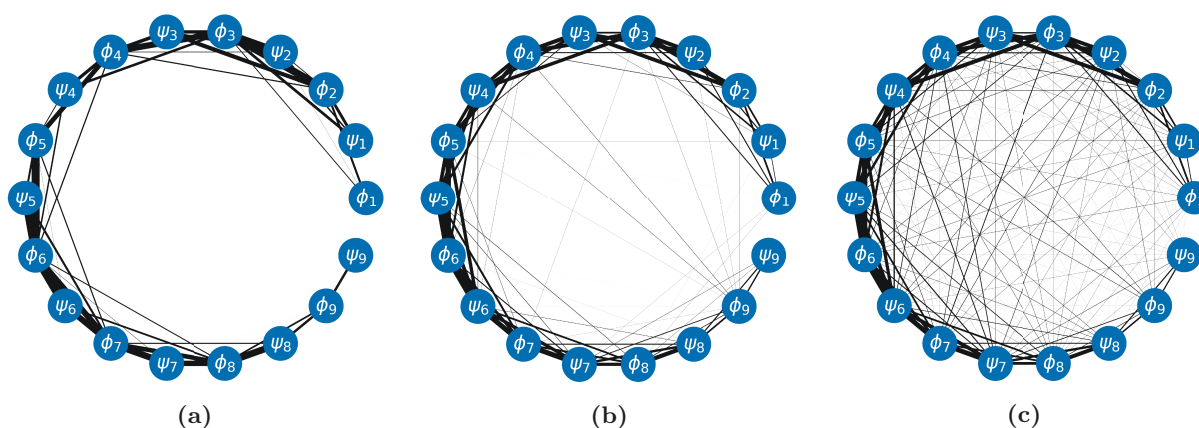


**FIG. 5:** (a)–(f) Trajectories of six OPs for Ala<sub>3</sub> for factor graph biased (blue) and unbiased (orange) simulations. It is clearly evident that the factor graph based biased simulations achieve spontaneous back-and-forth movement across high-dimensional configuration space. The acceleration relative to unbiased MD is quantified in (g), which shows the commute time for pathways between every pair of states for factor graph biased (blue) and unbiased (orange) simulations of capped Ala<sub>3</sub>. We use the state definitions described in Ref. 26. Points on the dotted line were not sampled and do not correspond to a commute time of  $10^5$ .

tions for capped Ala<sub>3</sub>, implementing the bias through PLUMED.

As can be seen from the various time series in Fig. 5, the simulations conducted using the static bias potential enhanced the fluctuations of all individual OPs. We also consider the transition rates for all of the  $\binom{8}{2} = 28$  transitions between the eight dominant metastable states, corresponding to the positive and negative regions of the Ramachandran plot for each of the three Alanine residues. These are shown in Fig. 5(g) as the average interconversion or commute time between different states.<sup>26</sup> When compared to the unbiased sampling, the factor graph biased simulations have significantly enhanced sampling (at least twice as fast, and up to 350 times faster among those sampled in both simulations) along 22 of the 28

transitions. The only transitions that are not enhanced are the ones that were fast to begin with except the notable exception of  $S_2$  to  $S_5$  which was not sampled in 150 ns of simulation with the factor graph bias. Interestingly, the low barrier transitions are not significantly slowed down due to biasing, which has been a concern in enhanced sampling.<sup>34</sup> Finally, in addition to the bias associated with the factor graph, we repeated both the bias generation and biased simulation for a factor graph without any edges which corresponds to biasing each OP in parallel without considering any correlations. These simulations are used for additional comparison in the SI.



**FIG. 6:** Markov random fields calculated using trajectories of Aib<sub>9</sub>'s dihedral angles with edge thickness showing the magnitude of the inverse covariance. (a) corresponds to a penalty parameter lower than that obtained with 5-fold cross validation. (b) corresponds to the cross-validated penalty parameter value. (c) corresponds to a penalty parameter lower than the cross-validated value.

### III.B Aib<sub>9</sub> Peptide

Aib<sub>9</sub> is a  $\alpha$ -aminoisobutyric acid-based  $3_{10}$ -helix<sup>15,16,27</sup> that displays conformational transitions at different timescales ranging from picoseconds to microseconds, and that may very well vary with solvent. Here we study Aib<sub>9</sub> in explicit water with 200 ns of simulation (see SI for details).<sup>35</sup> The most distinct of these is the transition between right- and left-handed helices, as shown in Fig. 1. Aib<sub>9</sub> requires nine  $\phi$  and nine  $\psi$  dihedrals for a complete description of the conformational dynamics. The trajectory of these 18 OPs was used to calculate Markov random fields shown in Fig. 6 which shows the change in graph structure as the graphical lasso penalty parameter is tuned.

The secondary structure of Aib<sub>9</sub> causes these OPs to be correlated in larger groups which are not amenable to our procedure limiting the factor graph to at most triplet interactions. This means that additional techniques<sup>36–39</sup> will be required to reduce the dimensionality of the OPs associated with larger factors in more complex systems which will be the subject of future investigation. Nevertheless, we can infer interesting conclusions regarding the underlying network governing the conformational dynamics of Aib<sub>9</sub> from Fig. 6. Firstly, irrespective of the value of the penalty parameter  $\lambda$  in Eq. 3, we never find the two ends of the peptide to be connected with a significant inverse covariance, in contrast to the MRF we found for the smaller capped Ala<sub>3</sub> peptide. Secondly, the interior residues of the peptide display thicker edges than the ones on the outside, demonstrating that inner residues move in a more correlated and conditionally dependent manner. Thirdly, the MRF is roughly symmetric as one approaches the center from either end i.e. clockwise from  $\Psi_9$  or anticlockwise from  $\Phi_1$ . This is indeed in agreement with measurements from sophisticated enhanced sampling and much longer (i.e. at least sev-

eral microsecond) simulations, which suggest that Aib<sub>9</sub> is achiral and both the left- and right-handed helices have same free energies.<sup>16,27</sup> We also find that applying the Girvan-Newman algorithm<sup>40</sup> to the sparse and cross-validated MRFs yields two communities which are split by the center of the protein. These communities suggest the edges crossing the two sides of the protein are involved in many shortest paths between nodes and in turn the dynamics of the central residues are crucial for conformational change. It is interesting to note that MRF trained on the first 80 ns or longer gives the same result as the trajectory has sufficiently sampled both right- and left-handed helices.

## IV DISCUSSION

In this work, we have developed and applied a probabilistic graphical models based framework for making sense of high-dimensional data arising in molecular simulations. This makes it possible to learn reliable approximations to high-dimensional molecular distributions in a tractable manner. Our approximations here are restricted up to a family of single, pair and triplet interactions but could be easily extended further if needed.

We validated this framework through detailed analysis of two biomolecular test systems. First, we used it to learn a factor graph structure for the small peptide capped Ala<sub>3</sub> that captures inter-residue correlations relevant to various conformational transitions in this molecule. Inspired by recent methods<sup>41,42</sup> which conduct enhanced sampling biasing multiple independent components, we demonstrate how our framework can be used to divide a large number of order parameters (OPs) into groups for biasing. As such, we use our factor graph structure along with belief propagation to construct a static bias for enhanced sampling. Using this bias, with

no time dependence, we were able to enhance the fluctuations along each OP as well as accelerate the sampling of almost all state-to-state transitions with spontaneous back-and-forth movement between states. While this static bias approach is arguably not a practical enhanced sampling method, it illustrates the importance of including correlations in the biasing variables as well as the possibility to use factorized models with more sophisticated enhanced sampling methods.<sup>39,41</sup> For this system we also demonstrate an intervention procedure<sup>17</sup> that directly tests the predictions of inter-residue correlations learned in our framework. As a second illustrative system we also model the peptide Aib<sub>9</sub> using a Markov field constructed with graphical lasso. This allows us to highlight the interpretability of probabilistic graphical models applied to molecular simulations of this peptide as it undergoes an interesting left-handed to right-handed helix transition. This graph structure highlights features of Aib<sub>9</sub> such as the difference in flexibility between central and outer residues, and the symmetry of correlations between residues.

We conclude by noting that despite the limitations of our simple illustrative strategies, this work can also be combined with sampling methods that are used to determine effective sets of basis functions for describing molecular systems<sup>43</sup> or those that are used for dimensionality reduction.<sup>36,39,44</sup> We envision that these methods will be used to develop an automated framework for sampling biomolecules that takes a starting structure and runs enhanced sampling simulations without the necessity for expert knowledge. This combination should allow simulation of new systems where traditional low-dimensional biasing variables are unsuccessful in a high-throughput manner, for instance biomolecular complexes with more than one constituent. Finally, we highlight that the framework developed in this manuscript is also closely related to the topic of causal inference.<sup>45,46</sup> The graphs learned for these molecular systems can be converted into directed Bayesian networks when combined with information about whether past states of an order parameter are predictive of future states of another. This would further improve the ability to understand molecular systems as it allows us to make not just correlational but causal and mechanistic claims.

## Notes

The code needed to reproduce the models used in this work will be made available at [github.com/tiwarylab](https://github.com/tiwarylab). The input files necessary to reproduce the simulations done in this work will be made available on PLUMED NEST.<sup>47</sup>

## ACKNOWLEDGMENTS

This work was supported by the National Science Foundation, Grant No. CHE-2044165. ZS was also supported by University of Maryland COMBINE program NSF award DGE-1632976. This work used XSEDE Bridges through allocation TG-CHE180053, which is supported by National Science Foundation grant number ACI-1548562. We also thank UMD's Deepthought2 and MARCC's Bluecrab HPC clusters for computing resources. We would like to thank Shams Mehdi for the Aib<sub>9</sub> trajectory. We would also like to thank Pavan Ravindra, Yihang Wang, Michael Strobel, and Nathan Zimmerberg for discussion and feedback about the direction of this project.

- <sup>1</sup>J. G. Kirkwood and E. M. Boggs, *The Journal of Chemical Physics* **10**, 394 (1942).
- <sup>2</sup>R. Kikuchi, *Physical review* **81**, 988 (1951).
- <sup>3</sup>J. D. Weeks, D. Chandler, and H. C. Andersen, *The Journal of Chemical Physics* **55**, 5422 (1971).
- <sup>4</sup>D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques* (MIT press, 2009).
- <sup>5</sup>P. Tiwary and A. van de Walle, "A review of enhanced sampling approaches for accelerated molecular dynamics," in *Multiscale Materials Modeling for Nanomechanics*, edited by C. R. Weinberger and G. J. Tucker (Springer International Publishing, Cham, 2016) pp. 195–221.
- <sup>6</sup>O. Valsson, P. Tiwary, and M. Parrinello, *Annual review of physical chemistry* **67**, 159 (2016).
- <sup>7</sup>S. Somani, B. J. Killian, and M. K. Gilson, *The Journal of chemical physics* **130**, 04B603 (2009).
- <sup>8</sup>R. M. Donovan-Maiye, C. J. Langmead, and D. M. Zuckerman, *Journal of chemical theory and computation* **14**, 426 (2018).
- <sup>9</sup>Forensic Science International: Genetics Supplement Series **3**, e119 (2011), progress in Forensic Genetics 14.
- <sup>10</sup>R. J. McEliece, D. J. C. MacKay, and Jung-Fu Cheng, *IEEE Journal on Selected Areas in Communications* **16**, 140 (1998).
- <sup>11</sup>J. Moussouris, *Journal of Statistical Physics* **10**, 11 (1974).
- <sup>12</sup>J. Friedman, T. Hastie, and R. Tibshirani, *Biostatistics* **9**, 432 (2007), <https://academic.oup.com/biostatistics/article-pdf/9/3/432/17742149/kxm045.pdf>.
- <sup>13</sup>J. Pearl, in *AAAI* (1982) pp. 133–136.
- <sup>14</sup>O. Valsson and M. Parrinello, *Journal of Chemical Theory and Computation* **11**, 1996 (2015), pMID: 26574405, <https://doi.org/10.1021/acs.jctc.5b00076>.
- <sup>15</sup>V. Botan, E. H. Backus, R. Pfister, A. Moretto, M. Crisma, C. Toniolo, P. H. Nguyen, G. Stock, and P. Hamm, *Proceedings of the National Academy of Sciences* **104**, 12749 (2007).
- <sup>16</sup>M. Biswas, B. Lickert, and G. Stock, *The Journal of Physical Chemistry B* **122**, 5508 (2018), pMID: 29338243, <https://doi.org/10.1021/acs.jpcc.7b11800>.
- <sup>17</sup>J. Pearl, *Statistical Science* **8**, 266 (1993).
- <sup>18</sup>T. M. Cover, *Elements of information theory* (John Wiley & Sons, 1999).
- <sup>19</sup>T. Minka, "Divergence measures and message passing," *Tech. Rep. MSR-TR-2005-173* (2005).
- <sup>20</sup>J. Yedidia, W. Freeman, and Y. Weiss, in *Exploring Artificial Intelligence in the New Millennium*, edited by G. Lakemeyer and B. Nebel (Morgan Kaufmann Publishers, 2003) Chap. 8, pp. 239–236.
- <sup>21</sup>M. Rosenblatt, *The Annals of Mathematical Statistics* **27**, 832 (1956).
- <sup>22</sup>E. Parzen, *The Annals of Mathematical Statistics* **33**, 1065 (1962).



- <sup>23</sup>G. Elidan, I. McGraw, and D. Koller, in *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'06 (AUAI Press, Arlington, Virginia, USA, 2006) p. 165–173.
- <sup>24</sup>G. Bussi, A. Laio, and P. Tiwary, “Metadynamics: A unified framework for accelerating rare events and sampling thermodynamics and kinetics,” in *Handbook of Materials Modeling : Methods: Theory and Modeling*, edited by W. Andreoni and S. Yip (Springer International Publishing, Cham, 2018) pp. 1–31.
- <sup>25</sup>A. Barducci, G. Bussi, and M. Parrinello, *Physical Review Letters* **100**, 020603 (2008).
- <sup>26</sup>S.-T. Tsai, Z. Smith, and P. Tiwary, “Sgoop-d: Estimating kinetic distances and reaction coordinate dimensionality for rare event systems from biased/unbiased simulations,” (2021), arXiv:2104.13560 [physics.comp-ph].
- <sup>27</sup>A. Perez, F. Sittel, G. Stock, and K. Dill, *Journal of Chemical Theory and Computation* **14**, 2109 (2018), pMID: 29547695, <https://doi.org/10.1021/acs.jctc.7b01294>.
- <sup>28</sup>P. Tiwary, *The Journal of Physical Chemistry B* **121**, 10841 (2017).
- <sup>29</sup>C. Abrams and G. Bussi, *Entropy* **16**, 163 (2014).
- <sup>30</sup>S. Kullback and R. A. Leibler, *The Annals of Mathematical Statistics* **22**, 79 (1951).
- <sup>31</sup>G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi, *Comput. Phys. Commun.* **185**, 604 (2014).
- <sup>32</sup>M. Bonomi, G. Bussi, C. Camilloni, G. A. Tribello, P. Banáš, A. Barducci, M. Bernetti, P. G. Bolhuis, S. Bottaro, D. Branduardi, and et. al., *Nat. Methods* **16**, 670 (2019).
- <sup>33</sup>M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, J. H. M. Bonvin, D. Lindahl, *SoftwareX* **1-2**, 19 (2015).
- <sup>34</sup>R. A. Miron and K. A. Fichthorn, *Physical review letters* **93**, 128301 (2004).
- <sup>35</sup>S. Mehdi, D. Wang, and P. Tiwary, Manuscript under preparation.
- <sup>36</sup>P. Tiwary and B. J. Berne, *Proc. Natl. Acad. Sci. U. S. A.* **113**, 2839 (2016), <https://www.pnas.org/content/113/11/2839.full.pdf>.
- <sup>37</sup>P. Tiwary and B. Berne, *The Journal of chemical physics* **147**, 152701 (2017).
- <sup>38</sup>Z. Smith, D. Pramanik, S.-T. Tsai, and P. Tiwary, *J. Chem. Phys.* **149**, 234105 (2018), <https://doi.org/10.1063/1.5064856>.
- <sup>39</sup>Y. Wang, J. M. L. Ribeiro, and P. Tiwary, *Nature communications* **10**, 1 (2019).
- <sup>40</sup>M. Girvan and M. E. J. Newman, *Proceedings of the National Academy of Sciences* **99**, 7821 (2002), <https://www.pnas.org/content/99/12/7821.full.pdf>.
- <sup>41</sup>J. Pfendtner and M. Bonomi, *Journal of Chemical Theory and Computation* **11**, 5062 (2015), pMID: 26574304, <https://doi.org/10.1021/acs.jctc.5b00846>.
- <sup>42</sup>P. Shaffer, O. Valsson, and M. Parrinello, *Proceedings of the National Academy of Sciences* **113**, 1150 (2016), <https://www.pnas.org/content/113/5/1150.full.pdf>.
- <sup>43</sup>P. Ravindra, Z. Smith, and P. Tiwary, *Mol. Syst. Des. Eng.* **5**, 339 (2020).
- <sup>44</sup>D. Wang and P. Tiwary, *The Journal of Chemical Physics* **154**, 134111 (2021).
- <sup>45</sup>J. Pearl, *Statist. Surv.* **3**, 96 (2009).
- <sup>46</sup>J. Pearl, *Causality* (Cambridge university press, 2009).
- <sup>47</sup>M. Bonomi, G. Bussi, C. Camilloni, G. A. Tribello, and The PLUMED consortium, *Nature Methods* **16**, 670 (2019).