Internal-Coordinate Density Modelling of Protein Structure: Covariance Matters

Marloes Arts *ma@di.ku.dk*
Department of Computer Science
University of Copenhagen

Jes Frellsen *jefr@dtu.dk*
Department of Applied Mathematics and Computer Science
Technical University of Denmark

Wouter Boomsma *wb@di.ku.dk*
Department of Computer Science
University of Copenhagen

Reviewed on OpenReview: *https://openreview.net/forum?id=9XRZtZRmEB*

Abstract

After the recent ground-breaking advances in protein structure prediction, one of the remaining challenges in protein machine learning is to reliably predict distributions of structural states. Parametric models of fluctuations are difficult to fit due to complex covariance structures between degrees of freedom in the protein chain, often causing models to either violate local or global structural constraints. In this paper, we present a new strategy for modelling protein densities in internal coordinates, which uses constraints in 3D space to induce covariance structure between the internal degrees of freedom. We illustrate the potential of the procedure by constructing a variational autoencoder with full covariance output induced by the constraints implied by the conditional mean in 3D, and demonstrate that our approach makes it possible to scale density models of internal coordinates to full protein backbones in two settings: 1) a unimodal setting for proteins exhibiting small fluctuations and limited amounts of available data, and 2) a multimodal setting for larger conformational changes in a high data regime.

1 Introduction

Proteins are macro-molecules that are involved in nearly all cellular processes. Most proteins adopt a compact 3D structure, also referred to as the *native state*. This structure is a rich source of knowledge about the protein, since it provides information about how the protein can engage biochemically with other proteins to conduct its function. The machine learning community has made spectacular progress in recent years on the prediction of the native state from the amino acid sequence of a protein (Jumper et al., 2021; Senior et al., 2020; Wu et al., 2022b; Baek et al., 2021; Wu et al., 2022a). However, the static picture of the structure of a protein is misleading: in reality a protein is continuously moving, experiencing both thermal fluctuations and larger conformational changes, both of which affect its function. One of the remaining challenges in machine learning for structural biology is to reliably predict these distributions of states, rather than just the most probable state. We discuss the state of the density modelling field in Section 5 (Related work).

Modelling the probability density of protein structure is non-trivial, due to the strong constraints imposed by the molecular topology. The specific challenges depend on the chosen structural representation: if a structure is represented by the 3D coordinates of all its atoms, these atom positions cannot be sampled independently without violating the physical constraints of, e.g., the bond lengths separating the atoms. In addition, an arbitrary decision must be made about how the structure is placed in a global coordinate system, which

implies that operations done on this representation should preferably be invariant or equivariant to this choice. An alternative is to parameterize the structure using internal coordinates, i.e. in terms of bond lengths, bond angles and dihedrals (rotations around the bonds). The advantage of this representation is that internal degrees of freedom can be sampled independently without violating the local bond constraints of the molecule. It also makes it possible to reduce the number of degrees of freedom to be sampled – for instance fixing the bond lengths to ideal values, since they fluctuate much less than the torsion angles and bond angles.

For the reasons given above, an internal coordinate representation would appear to be an attractive choice for density modelling. However, one important problem reduces the appeal: small fluctuations in internal coordinates will propagate down the chain, leading to large fluctuations remotely downstream in the protein. As a consequence, internal-coordinate density modelling necessitates careful modelling of the covariance structure between the degrees of freedom in order to ensure that small fluctuations in internal coordinates result in small perturbations of the 3D coordinates of the protein. Such covariance structures are typically highly complex, making direct estimation difficult.

In this paper, we investigate whether density modelling of full-size protein backbones in internal coordinates is feasible. We empirically demonstrate the difficulty in estimating the covariance structure of internal coordinates from data, and instead propose a technique for *inducing* the covariance structure by imposing constraints on downstream atom movement using the Lagrange formalism. Rather than estimating the covariance structure from scratch, we can instead modulate the covariance structure by choosing appropriate values for allowed fluctuations of downstream atoms (Fig. 1). We demonstrate the procedure in the context of a variational autoencoder.

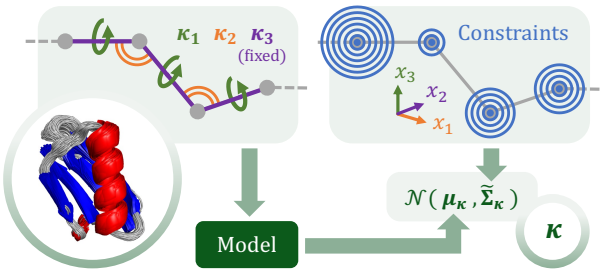

Figure 1: A protein structure ensemble is modelled in internal coordinate space (dihedrals in green, bond angles in orange, and fixed bond lengths in purple), while imposing constraints on atom fluctuations in Euclidean space. The resulting full covariance structure can be used to construct a multivariate Gaussian distribution, from which samples can be drawn that reflect the local and global fluctuations of the protein.

Given a prior on the internal coordinate fluctuations and a predicted mean, we impose constraints on the atom fluctuations in 3D space to obtain a full covariance structure over the internal coordinates. We show that this allows us to generate valid structures in terms of both internal and Cartesian coordinates. Our method is validated in two regimes: a low data regime for proteins that exhibit small, unimodal fluctuations, and a high data regime for proteins that exhibit multimodal behavior. We anticipate that this method could serve as a building block applicable more generally for internal-coordinate density estimation, for instance internal-coordinate denoising diffusion models.

Our main contributions are:

- We formulate a procedure for inducing full protein backbone covariance structure in internal coordinates, based on constraints on atom fluctuations in 3D space.
- Rather than predicting a full covariance matrix over internal coordinates, our proposed method only requires to predict one Lagrange multiplier for each atom, from which the full covariance matrix can be constructed. For M atoms with fixed bond lengths, this corresponds to a reduction from $(2 \times M - 5)^2$ to simply M predicted values.

- We design a variational autoencoder which models fluctuations for full-length protein backbones in internal coordinates. Even though constraints are formulated in Euclidean space, the model is not dependent on a global reference frame (i.e. it is rotationally invariant).
- We demonstrate that our model provides meaningful density estimates on ensemble data for proteins obtained from experiment and simulation, even when data is limited. For more complex densities and a large amount of data, we show that the expressiveness of the chosen base model plays a key role.

Scope. The focus of this paper is on modelling distributions of protein structure states in internal coordinates, with the emphasis on the low data regime. We are thus concerned with thermodynamic ensembles, rather than the detailed dynamics that a molecule undergoes or generalization across different proteins. Dynamics could potentially be modelled on top of our approach, for instance by fitting a discrete Markov model to describe transitions between states, and using our approach to model the thermal fluctuations within a state, but this is beyond the scope of the current work. The main contribution of this paper is a technique for modelling internal coordinate covariance structure, which has applicability beyond the proof of principle presented here. Another perspective on our approach is that we wish to describe the aleatoric uncertainty associated with a static structure.

2 Background

2.1 Cartesian vs internal coordinates

As stated above, Cartesian coordinates and internal coordinates each have advantages and disadvantages. Assume we have a 3D protein structure in Euclidean space with atom positions $\boldsymbol{x}$. Throughout this paper, we only consider backbone atoms N, C_α and C, which gives a total number of atoms $M = 3 \times L$, where L is the number of amino acids. The Euclidean setting thus results in $3 \times M$ coordinates. Even though in this setting each of the atoms can fluctuate without affecting other atoms in the backbone chain, there is no guarantee for chemical integrity, i.e. conservation of bond lengths and respecting van der Waals forces. This can lead to local backbone crossings and generally unphysical protein structures.

One way to ensure chemical integrity is to parameterize protein structure in internal coordinate space using dihedrals $\boldsymbol{\kappa_1}$, bond angles $\boldsymbol{\kappa_2}$ and bond lengths $\boldsymbol{\kappa_3}$. Here, dihedrals are torsional angles that twist the protein around the bond between two consecutive atoms, bond angles are angles within the plane that is formed by two consecutive bonds, and bond lengths are the distances between two consecutive backbone atoms (left panel Fig. 1). Since bond length distributions have very little variance (Creighton, 1993), we choose to fix them, thereby reducing the number of variables over which we need to estimate the covariance. We will refer to the remaining two internal coordinates together as $\boldsymbol{\kappa}$ to avoid notation clutter. As dihedrals are defined by four points (the dihedral is the angle between the plane defined by the first three points and the plane defined by the last three points) and bond angles are defined by three points, the resulting protein structure representation will have $(2 \times M) - 5$ coordinates. Not only does this result in less coordinates to determine a full covariance structure over, the coordinates are also automatically rotation and translation invariant, as opposed to Cartesian coordinates.

The remaining problem is that small changes in one internal coordinate can have large consequences for the global structure of the protein, since all atoms downstream of the internal coordinate will move together, acting as a rigid body. It is therefore challenging to preserve global structure while altering internal coordinates.

2.2 Standard precision estimators do not capture global fluctuations

Because of the limitations of internal coordinates mentioned in Section 2.1, it is a highly non-trivial task to capture a full covariance structure over $\boldsymbol{\kappa}$ which also conforms to constraints in Euclidean space that are inherent to the protein, especially when the amount of data is limited. As an example, we use a standard estimator to get a precision matrix (i.e. the inverse of the covariance matrix) over $\boldsymbol{\kappa}$ for a short molecular dynamics (MD) simulation of the B1 domain of protein G, based on PDB entry 1pga (Fig. 2). Details about the simulation can be found in Appendix A. We see that when we take samples from a multivariate Gaussian

over $\boldsymbol{\kappa}$ with estimated mean and precision matrix, the samples exhibit atom fluctuations that are much higher than the original simulation, and with a very different pattern along the chain.

Much of the covariance structure is a direct consequence of the internal coordinate representation. For instance, preserving a simple distance between two atoms in a chain requires correlated changes between all the dihedral angles separating them. However, given a specific structure of the chain, these correlations are predictable, and we can therefore attempt to express them collectively in terms of a much smaller number of parameters. Specifically, we will consider the correlations in internal coordinates that are induced by uncorrelated fluctuations in the atomic coordinates in Euclidean space.

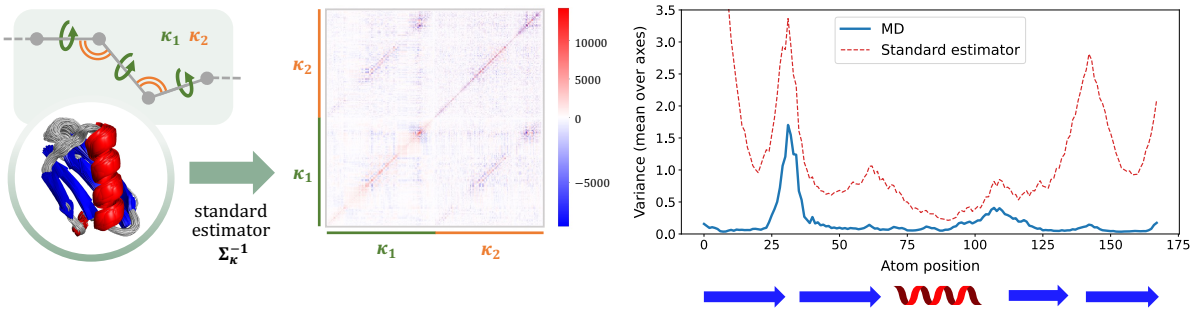

Figure 2: When a standard estimator is used to get the precision structure over internal coordinates, resulting atom fluctuations significantly deviate from MD simulations. Blue arrows and red helices represent secondary structural elements. The variance is calculated as the mean of the variances over the x, y and z axis, in Å^2.

3 Internal-coordinate density modelling with constraints

3.1 Setup

We parameterize a 3D protein structure in terms of internal coordinates (i.e. dihedrals and bond angles, while bond lengths are kept fixed due to their low variance), which together will be referred to as $\boldsymbol{\kappa}$. Our aim is to obtain a multivariate Gaussian distribution $p(\Delta\boldsymbol{\kappa})$ over the deviations $\Delta\boldsymbol{\kappa}$ from the mean, centered at zero, with a full precision structure. This target distribution is subject to constraints over atom fluctuations, enforcing the preservation of global structure. Our point of departure is to specify a Gaussian prior $q(\Delta\boldsymbol{\kappa})$, which we will call the κ-prior, over the internal coordinate distribution, where the mean is zero and the precision is a diagonal matrix with the diagonal filled by the inverse variance over all $\boldsymbol{\kappa}$ values $\sigma^{-2}_{\boldsymbol{\kappa},\text{data}}$. This prior sets the overall scale of fluctuations expected for the different degrees of freedom. For convenience, we will set this equal to the empirical variance of the degrees of freedom observed in the data [1]. The strength of the κ-prior can be tuned using hyperparameter a. The κ-prior is thus defined as

$$q(\Delta\boldsymbol{\kappa}) = \frac{1}{Z_q}\exp\left(-\frac{1}{2}\Delta\boldsymbol{\kappa}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\kappa},\text{prior}}\Delta\boldsymbol{\kappa}\right), \tag{1}$$

where $Z_q = \boldsymbol{\sigma}_{\boldsymbol{\kappa},\text{data}}\sqrt{2\pi a}$ is the normalization constant for the κ-prior distribution and $\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\kappa},\text{prior}} = a \cdot \text{diag}(\sigma^{-2}_{\boldsymbol{\kappa},\text{data}})$ is the precision matrix. Our approach will be to construct a new distribution p which is as close as possible to q, but which fulfills a constraint that prohibits the downstream 3D coordinates from fluctuating excessively. We thus wish to minimize the Kullback-Leibler divergence between the objective distribution and κ-prior:

$$\mathcal{D}_{\text{KL}}(p|q) = \int p(\Delta\boldsymbol{\kappa})\ln\frac{p(\Delta\boldsymbol{\kappa})}{q(\Delta\boldsymbol{\kappa})}\mathrm{d}\Delta\boldsymbol{\kappa}, \tag{2}$$

subject to constraints on the expected value over squared atom displacements of each downstream atom:

$$\mathbb{E}_{\Delta\boldsymbol{\kappa}\sim p(\Delta\boldsymbol{\kappa})}\left[\Delta x_m^2\right] = \int \Delta x_m^2 p(\Delta\boldsymbol{\kappa})\mathrm{d}\Delta\boldsymbol{\kappa} = C_m \tag{3}$$

[1] When used like this the κ-prior should be thought of as a base distribution, rather than a Bayesian prior, similar to how the prior concept is used in a variational autoencoder.

where Δx_m^2 is the squared displacement of atom m and C_m is a tolerance value on the expected value of this fluctuation (i.e. the variance $\sigma_{x_m}^2$ in Cartesian coordinates, assuming an isotropic Gaussian). We will see later that by tuning the C_m values (blue circles in Fig. 1), we can modulate the covariance structure in internal coordinates.

3.2 Lagrange formalism to incorporate constraints

Employing Jaynes' maximum entropy principle (Jaynes, 1957), we use the Lagrange formalism to incorporate a constraint for each of our M atoms, under the conditions that our probability density $p(\Delta\boldsymbol{\kappa})$ has zero mean and sums to one[2]. This leads to the Lagrangian

$$\tilde{\mathcal{D}}(p|q) = \int p(\Delta\boldsymbol{\kappa}) \ln \frac{p(\Delta\boldsymbol{\kappa})}{q(\Delta\boldsymbol{\kappa})} \mathrm{d}\Delta\boldsymbol{\kappa} + \lambda_0 \left(\int p(\Delta\boldsymbol{\kappa}) \mathrm{d}\Delta\boldsymbol{\kappa} - 1 \right) + \sum_{m=1}^{M} \lambda_m \left(\int \Delta x_m^2 p(\Delta\boldsymbol{\kappa}) \mathrm{d}\Delta\boldsymbol{\kappa} - C_m \right).$$

Next, we take the functional derivative and set it to zero: $\frac{\partial \tilde{\mathcal{D}}(p,q)}{\partial p(\Delta\boldsymbol{\kappa})} = 0$, leading to the following well-established result (Jaynes, 1957; Kesavan & Kapur, 1989)[3]:

$$0 = \ln \frac{p(\Delta\boldsymbol{\kappa})}{q(\Delta\boldsymbol{\kappa})} + 1 + \lambda_0 + \sum_{m=1}^{M} \lambda_m \Delta x_m^2 \;\Rightarrow\; p(\Delta\boldsymbol{\kappa}) = \frac{1}{Z_p} q(\Delta\boldsymbol{\kappa}) \exp\left(- \sum_{m=1}^{M} \lambda_m \Delta x_m^2 \right) \tag{4}$$

where $Z_p = \exp(-1 - \lambda_0)$ is the normalization constant of the target distribution. Note that since $\frac{\partial^2 \tilde{\mathcal{D}}(p,q)}{\partial p(\Delta\boldsymbol{\kappa})^2} = \frac{1}{p(\Delta\boldsymbol{\kappa})}$ is positive, we know our solution will indeed be a minimum.

3.3 First order approximation for atom fluctuations

In order to use Eq. (4) to satisfy the imposed constraints, we need to express Δx^2 in terms of $\Delta\boldsymbol{\kappa}$. Using a first order Taylor expansion (small angle approximation), we can express the 3D displacement vector $\Delta\boldsymbol{x}_m$ of each atom as a linear transformation of the angular perturbation:

$$\Delta\boldsymbol{x}_m \approx \sum_i \frac{\partial \boldsymbol{x}_m}{\partial \kappa_i} \Delta\kappa_i \tag{5}$$

where $\boldsymbol{x}_m$ is the position vector of the m^{th} atom, assuming that the location of atom m is downstream of the i^{th} internal coordinate in the direction in which the changes are propagated through the chain. The partial derivatives $\frac{\partial \boldsymbol{x}_m}{\partial \kappa_i}$ are known in closed form (Bottaro et al., 2012). From Eq. (5) it follows that the squared displacement distance can be approximated by

$$\Delta x_m^2 \approx \left\| \sum_i \frac{\partial \boldsymbol{x}_m}{\partial \kappa_i} \Delta\kappa_i \right\|^2 = \sum_{ij} \left(\frac{\partial \boldsymbol{x}_m}{\partial \kappa_i} \Delta\kappa_i \cdot \frac{\partial \boldsymbol{x}_m}{\partial \kappa_j} \Delta\kappa_j \right) = \Delta\boldsymbol{\kappa}^{\mathrm{T}} \boldsymbol{G}_m \Delta\boldsymbol{\kappa} \;, \tag{6}$$

where $\boldsymbol{G}_m^{i,j} = \frac{\partial \boldsymbol{x}_m}{\partial \kappa_i} \cdot \frac{\partial \boldsymbol{x}_m}{\partial \kappa_j}$ is a symmetric and positive-definite matrix.

Substituting Eq. (6) and our κ-prior expression from Eq. (1) into our target distribution from Eq. (4) gives a new Gaussian distribution:

$$p(\Delta\boldsymbol{\kappa}) \approx \frac{1}{\tilde{Z}} \exp\left(-\frac{1}{2} \Delta\boldsymbol{\kappa}^{\mathrm{T}} \left(\boldsymbol{\Sigma}_{\boldsymbol{\kappa},\text{prior}}^{-1} + \boldsymbol{\Sigma}_{\boldsymbol{\kappa},\text{constr}}^{-1} \right) \Delta\boldsymbol{\kappa} \right) = \mathcal{N}(0, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\kappa}}) \;, \tag{7}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\kappa},\text{constr}}^{-1} = 2 \sum_{m=1}^{M} \lambda_m \boldsymbol{G}_m$ and $\tilde{Z}$ is the new normalization constant.

[2]Even though throughout this derivation we have included the normalization constant for completeness, in practice we work with unnormalized densities and normalize post hoc, since we recognize the final result to be Gaussian.

[3]Note that $\frac{\partial}{\partial y} \left(y \ln y/q + \lambda_0 (y-1) + \sum_{m=1}^{M} \lambda_m \left(\Delta x_m^2 y - C_m \right) \right) = \ln \frac{y}{q} + y \cdot \frac{1}{y} + \lambda_0 + \sum_{m=1}^{M} \lambda_m \Delta x_m^2$

3.4 Satisfying the constraints

The final step in the constrained optimization is to establish the values for the Lagrange multipliers. A closed form solution for this is not readily available, but using the findings from Section 3.3, we can now rewrite the constraints from Eq. (3) as

$$C_m = \mathbb{E}_{\Delta\boldsymbol{\kappa}\sim\mathcal{N}(0,\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\kappa}})}\left[\Delta\boldsymbol{\kappa}^{\mathrm{T}}\boldsymbol{G}_m\Delta\boldsymbol{\kappa}\right] = \mathrm{tr}(\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\kappa}}\boldsymbol{G}_m) \tag{8}$$

where the last simplification step comes from standard expectation calculus on a quadratic form ($\Delta\boldsymbol{\kappa}^{\mathrm{T}}\boldsymbol{G}_m\Delta\boldsymbol{\kappa}$), where $\Delta\boldsymbol{\kappa}$ has zero mean (Eq. 378 in Petersen et al., 2008). Although it is non-trivial to express the Lagrange multipliers λ in terms of atom fluctuations C, we thus see that it is possible to *evaluate* C *given* a set of Lagrange multipliers λ. In the following, we will therefore construct our models such that our networks predict λ, directly. The predicted set of Lagrange multipliers λ will serve as a proxy for the atom fluctuations C_m (as demonstrated in Appendix C.1). We thereby predict the scale for the fluctuation levels allowed for all atoms in Cartesian coordinates, which in turn induces a full covariance structure over internal coordinates.

3.5 VAE pipeline

VAE model architecture. To demonstrate how our method works within a modelling context, we use a variational autoencoder (VAE), as illustrated in Fig. 3. We chose this model because of its simplicity, inspectable latent space, and stable training behavior, making modelling in a low data regime feasible. The VAE has a simple linear encoder that takes internal coordinates $\boldsymbol{\kappa}$ (dihedrals and bond angles, bond lengths are kept fixed) as input and maps to latent space $\boldsymbol{z}$, where we have a standard Gaussian as a prior on the latent space, which we call the z-prior to avoid confusion with the κ-prior. The decoder outputs the mean over $\boldsymbol{\kappa}$, which is converted into Cartesian coordinates using pNeRF (AlQuraishi, 2018). This mean structure in 3D coordinates is used for two purposes. First, using the structure we can evaluate the partial derivatives of atom positions with respect to the individual κ as in Eq. (6). Second, the predicted mean structure is used to obtain a pairwise distance matrix $\boldsymbol{d}$. This rotationally and translationally invariant representation reflects the steric constraints between non-covalently bonded atoms, and was therefore chosen as the input to a U-Net (Ronneberger et al., 2015), from which we estimate values for the Lagrange multipliers for each constraint. This allows the variational autoencoder, conditioned on the latent state $\boldsymbol{z}$, to modulate the allowed fluctuations. Implementation-wise, the U-net is concluded with an average pooling operation that for each row-column combination computes one Lagrange multiplier λ. Together with our fixed-variance κ-prior over $\boldsymbol{\kappa}$ and hyperparameter a determining the strength of this κ-prior, a new precision matrix is formed according to Eq. (7). We train a new VAE for each protein, where each trained model can generate new structures through simple ancestral sampling: first generating $\boldsymbol{z}$ from the standard normal z-prior, and subsequently sampling from a multivariate Gaussian distribution with the decoded mean and the constructed precision matrix. We refer to Appendix A for more details on data handling, model architecture, and other model settings.

Loss. Following standard practice, we optimize the evidence lower bound (ELBO) using the Gaussian likelihood on the internal degrees of freedom as constructed above. While this loss encourages a fit to the target κ density, it does not ensure that the predicted Lagrange multipliers are within the range within which our first order approximation of the fluctuations is valid. To ensure this, we add the inverse of the L1 norm of the Lagrange multipliers as an auxiliary regularizing loss, which prevents the κ-prior from dominating. By tuning the weight w_{aux} on the auxiliary loss, we can influence the strength of the constraints. Our total loss $\mathcal{L}$ is thus defined as

$$\mathcal{L} = -\mathbb{E}_{q_{\boldsymbol{\phi}}}\left[\log p_{\boldsymbol{\theta}}(\Delta\boldsymbol{\kappa}|\boldsymbol{z})\right] + \mathcal{D}_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\Delta\boldsymbol{\kappa}) \;||\; p(\boldsymbol{z})\right) + w_{\mathrm{aux}} \cdot |\boldsymbol{\lambda}|^{-1}, \tag{9}$$

where $\mathcal{D}_{\mathrm{KL}}$ is the Kullback–Leibler divergence, and $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are the parameters of the encoder and decoder, respectively.

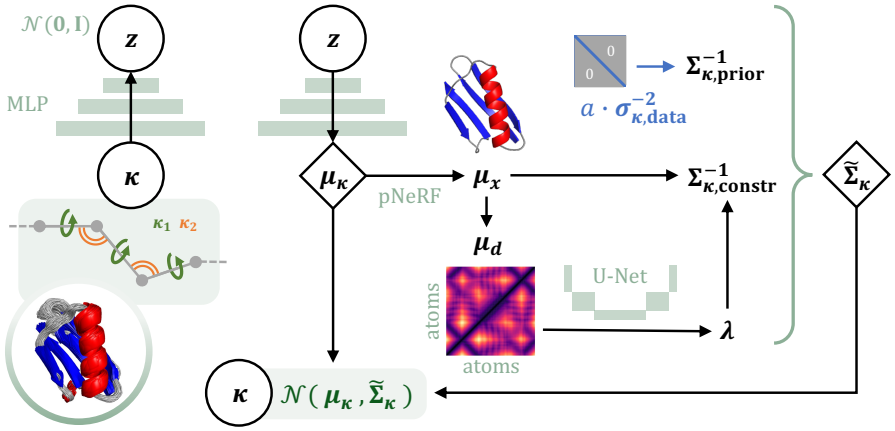

Figure 3: Model overview. The encoder (left) embeds internal coordinates into the latent space. The decoder (right) predicts a mean, from which constraints are extracted to obtain a precision matrix. Together with the κ-prior over the precision matrix based on the input data, a new precision matrix is formed which can be used to sample from a multivariate Gaussian.

4 Experiments

4.1 Test cases

Unimodal setting in low data regime. We consider three test proteins for small fluctuations in a low data regime: 1unc, 1fsd, and 1pga. 1unc corresponds to the solution structure of the human villin C-terminal headpiece subdomain. This protein contains 36 residues with 108 backbone (N, C_α and C) atoms. This solution nuclear magnetic resonance (NMR) dataset is freely available from the Protein Data Bank and contains 25 conformers. 1fsd, a beta beta alpha (BBA) motif, is also a freely available NMR dataset containing 41 structures. This system has 28 residues with 84 backbone atoms. 1pga, corresponding to B1 immunoglobulin-binding domain protein G, is a 56 amino acid long protein with 168 backbone atoms. We have a short in-house molecular dynamics (MD) simulation, which is 20ns long and structures were saved at a 50ps interval, resulting in 400 structures for this protein. See Appendix A for more details about the simulation.

Multimodal setting in high data regime. We also include two test cases for larger fluctuations following a multimodal distribution in a high data regime. Both are known as "fast-folders" and the MD datasets are available upon request from Lindorff-Larsen et al. (2011). We refer the reader to this work for detailed descriptions of the simulations. Chignolin (cln025) is a peptide with a hairpin structure, containing 10 residues and thus 30 backbone atoms. The simulation is 106 µs long, saved at a 200 ps interval, resulting in 534.743 data points. The second test case, 2f4k, is the chicken villin headpiece, with 35 residues and 105 backbone atoms. The simulated trajectory is 125 µs and also saved every 200 ps, yielding 629.907 structures.

4.2 Metrics

For the unimodal setting, we choose two simple measures of local and global structure, respectively. To evaluate local structure fluctuations, we show Ramachandran plots, a well-known visualization tool in the context of protein structures, where ϕ and ψ dihedrals, which are the torsional angles around the $N - C_\alpha$ and $C_\alpha - C$ bonds, are plotted against each other. As a global measure to evaluate the overall 3D structural fluctuations, we report the variance over atom positions, averaged over three dimensions, across superposed (i.e. structurally aligned) samples.

For the multimodal setting, we report free energy landscapes, parameterized by the first two components of time-lagged independent component analysis (TICA) (Molgedey & Schuster, 1994). Similar to e.g. PCA, TICA fits a linear model to map a high-dimensional input to a lower-dimensional output, but TICA also

incorporates the time axis. In this case, the high-dimensional representation consists of all dihedrals and pairwise distances within the protein backbone. The resulting components are ranked according to their capacity to explain the slowest, low-frequency, modes of motion. Taking the first two components corresponds to selecting reaction coordinates that underlie the slowest protein conformational changes, which is highly correlated with (un)folding behavior. This projection to a lower dimensional space can be considered a free energy landscape that reflects which conformational changes are favorable over others. We fit the TICA model on the time-ordered MD data, and pass samples from the VAE and baselines through the fitted model to create free energy landscapes.

Baselines. Apart from comparing the generated samples from our model to reference distributions from MD or NMR, we include four baselines. The first two baselines are VAE-based, using an encoder-decoder architecture identical to our own VAE, but without incorporating the meaningful 3D constraints. The first baseline, named "κ-prior (fixed)" is a VAE trained to predict $\boldsymbol{\mu}_{\boldsymbol{\kappa}}$ given a fixed covariance matrix that is equal to $\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\kappa},\text{prior}}$. In other words, this is the same as our full VAE setup, but omitting the imposed 3D constraints. The second baseline is "κ-prior (learned)", which corresponds to a more standard VAE-setting where the decoder directly outputs a mean and a variance (i.e. a diagonal covariance matrix). The third baseline does not involve a VAE, but samples structures from a multivariate Gaussian with a mean based on the dataset and a precision matrix computed by a standard estimator. This is an emperical estimator for MD datasets, and an Oracle Approximating Shrinkage (OAS) estimator (Chen et al., 2010) for NMR datasets, since emperical estimators do not converge for such low amounts of samples. Finally, we include the flow-based model from Köhler et al. (2023) as a fourth baseline, which, to the best of our knowledge, is the current state of the art in density modelling for internal coordinate representations.

4.3 Internal-coordinate density modelling results

Unimodal setting in low data regime. The unimodal, low data regime test cases exhibit small fluctuations around the native protein structure, where the largest fluctuations correspond to the loops connecting different secondary structure elements. Fig. 4 demonstrates that 1unc, 1fsd and 1pga structures sampled from the VAE conform to local and global constraints, with valid Ramachandran distributions compared to the reference as well as improved atom position variance along the chain with respect to the baselines (see quantitative results in Table A2). Even in the extremely low data regime of 25 and 41 data points for 1unc and 1fsd, respectively (top two rows in Fig. 4), the VAE is able to estimate a full covariance matrix that approximates the distribution better than the baselines, especially in loop regions. Consequently, the baselines show a higher tendency to sample unphysical structures with backbone crossings, even though the local structure is preserved. These effects can also be observed in the 3D visualization of the sampled ensembles in Appendix C.2.

The third test case, 1pga (bottom row in Fig. 4), has a more complex structure with two β-strands at the N-terminus forming a sheet together with two β-strands from the C-terminus. These global constraints are not captured well by the baselines in this low data regime (400 data points), resulting in very high fluctuations in loop regions which violate the native structure (additional visualizations can be found in Appendix C.2). For our VAE, we see the benefits of imposing global constraints, resulting in much better density estimation compared to the baselines. Moreover, we can control the interplay between local and global constraints by adjusting the hyperparameters of our model, as exemplified in Appendix D.1. However, the complexity of this protein prevents perfect density estimation in a low data regime. Interestingly, we show in Appendix C.1 that the variance of the atom positions highly correlates to the imposed constraints C calculated from a set of predicted Lagrange multipliers using Eq. (7).

Multimodal setting in high data regime. Here, we explore the use of our approach for modelling more complex behavior in a high data regime. Fig. 5 shows the free energy landscape in terms of the two first TICA components for cln025 and 2f4k. When comparing the VAE and the baselines to the MD reference (see also quantitative results in Table A3), it is clear that the learned prior and standard estimator do not capture all modes in the free energy landscape. The flow-based model performs best, suggesting that in this multimodal setting with plenty of available data, our proof-of-concept VAE is not as expressive as this state-of-the-art model. Moreover, the benefit of imposing 3D constraints on top of the fixed κ-prior (see baseline) seems

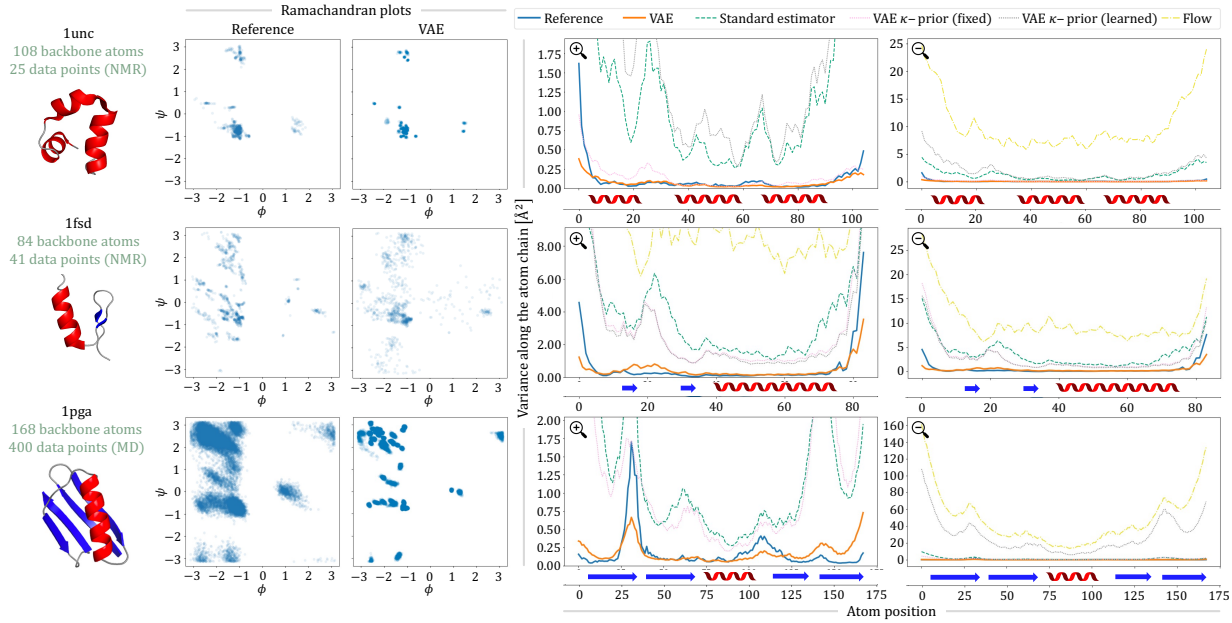

Figure 4: Modelling fluctuations in the unimodal setting for 1pga, 1fsd, and 1unc. Left: structure visualization, with α-helices in red and β-sheets as blue arrows. Middle: Ramachandran plots for the MD reference and VAE samples. Right: variance along the atom chain for VAE samples, MD reference, and baselines. Secondary structure elements are indicated along the x-axis.

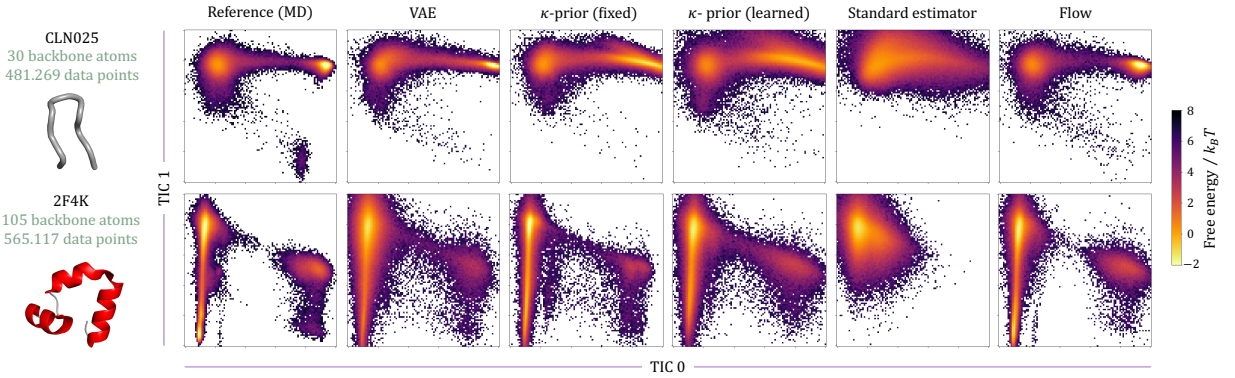

Figure 5: Modelling (un)folding behavior in the multimodal setting for cln025 and 2f4k. Left: structure visualization. Right: TICA free energy landscapes for MD reference, VAE, and baselines.

beneficial for cln025, but the effect is not as strong for the α-helical 2f4k, where local constraints might dominate (a similar effect can be seen when comparing the VAE to the fixed κ-prior baseline for 1unc in the unimodal setting). However, our simple VAE setup is evidently able to model large conformational changes through its latent space, demonstrating how our general-purpose method for modeling fluctuations using 3D constraints can be incorporated into more expressive models to model complex behavior.

Similar to the unimodal case, there is a tradeoff between local and global constraints which we can modulate using hyperparameters, as demonstrated in Appendix D.2. In addition, we show in Appendix C.3 how distinct regions of the VAE latent space map to different clusters in the TICA free energy landscape, and visualize the corresponding structures.

5 Related work

There is a large body of work on models for analyzing trajectories of molecular dynamics simulations, either through Markov state models (Chodera & Noé, 2014; Singhal & Pande, 2005; Sarich et al., 2013; Schütte et al., 1999; Prinz et al., 2011), or more complex modelling strategies (Mardt et al., 2018; Hernández et al., 2018; Sultan et al., 2018; Mardt et al., 2020; Xie et al., 2019). Typically, these focused on dimensionality reduced representations of the molecular structures, and are therefore not density models from which samples can be drawn. Similarly, interesting work has been conducted on generative models for contact maps, which are not directly realizable as 3D structures (Anand & Huang, 2018).

To our knowledge, the first generative density model of full protein coordinates was the Boltzmann generator (Noé et al., 2019), a normalizing flow over the Cartesian coordinates of protein ensembles. An extension of this approach was later used to estimate C_α-only coarse-grained force fields for molecular dynamics simulations (Köhler et al., 2023). This method, which uses internal coordinate inputs, demonstrated the ability of augmented flows to sample structural ensembles for proteins up to 35 amino acids . A range of works have explored the use of latent variable models. One example is the IG-VAE (Eguchi et al., 2022), which generates structures in 3D coordinates but expresses the loss in terms of distances and internal coordinates to maintain SE(3) invariance. Similar approaches have been used to analyze cryo-EM data, where the task is to generate ensembles of structures given the observed cryo-EM image data. Since cryo-EM data provides information at slightly lower resolution than the full-atomic detail we discuss here, the output of these approaches are often density maps in 3D space (Zhong et al., 2021; Punjani & Fleet, 2021). One example of atomic-level modelling in this space is Rosenbaum et al. (2021), which decodes deterministically into 3D coordinates, but describes the variance in image space. Finally, diffusion models and flow-matching have recently provided a promising new approach to density modelling, with impressive examples of density modelling at the scale of full-size proteins (Watson et al., 2022; Ingraham et al., 2022; Anand & Achim, 2022; Arts et al., 2023; Schreiner et al., 2023; Jing et al., 2024).

The primary objective in our paper is to investigate whether internal-coordinate density modelling with a full covariance structure is feasible using a simple, parsimonious setup. Internal-coordinate probabilistic models of proteins have traditionally focused on protein local structure, i.e. correct modelling of angular distributions of the secondary structure elements in proteins. Early work was based on hidden Markov models of small fragments (Camproux et al., 1999; 2004; de Brevern et al., 2000; Benros et al., 2006). The discrete nature of the fragments meant that these models did not constitute a complete probabilistic model of the protein structure. Later models solved this issue by modelling local structure in internal coordinates, using different sequential models and angular distributions (Edgoose et al., 1998; Bystroff et al., 2000; Hamelryck et al., 2006; Boomsma et al., 2008; 2014; Thygesen et al., 2021). Due to the downstream effects of small internal-coordinate fluctuations, these models are not by themselves capable of modelling the distribution of entire protein structures, but they are useful as proposal distributions in Markov chain Monte Carlo (MCMC) simulations of proteins (Irbäck & Mohanty, 2006; Boomsma et al., 2013). Using deep learning architectures to model the sequential dependencies in the protein chain, recent work has pushed the maximum length of fragments that can be reliably modelled to stretches of 15 residues (Thygesen et al., 2021), where the fragment size is limited due to the challenges in estimating the necessary covariance structure.

Our work was inspired by methods used for constrained Gaussian updates in MCMC simulation, first introduced by Favrin et al. (2001), and later extended by Bottaro et al. (2012). Our method generalizes the approach to global updates of proteins, derives the relationship between the Lagrange multipliers and corresponding fluctuations in Euclidean space, and uses neural networks to govern the level of fluctuations in order to modulate the induced covariance structures.

Recent work has demonstrated that internal-coordinate modelling can also be done using diffusion models (Jing et al., 2022). So far this method has been demonstrated only on small molecules. We believe the method we introduce in this paper might help scale these diffusion approaches to full proteins. In Cartesian space, the Chroma model (Ingraham et al., 2022) demonstrated the benefits of correlated diffusion arising from simple constraints between atoms, e.g. to maintain the structural integrity of the chain. Our method builds on similar considerations in internal coordinate space, where the challenge is to ensure global integrity rather than local integrity.

6 Discussion

Although protein structure prediction is now considered a solved problem, fitting the density of structural ensembles remains an active field of research. Many recent activities in the field focus on diffusion models in the Cartesian coordinate representation of a protein. In this paper, we take a different approach and investigate how we can describe small-scale fluctuations in terms of a distribution over the internal degrees of freedom of a protein. The main challenge in this context is the complex covariance between different parts of the chain. Failing to model this properly results in models that produce disruptive changes to the global structure, even for fairly minor fluctuations in the internal coordinates. Instead of estimating the covariance matrix from data, we show that it can be induced by imposing constraints on the Cartesian fluctuations. In a sense, this represents a natural compromise between internal and Cartesian coordinates: we obtain samples that are guaranteed to fulfill the physical constraints of the local protein topology (e.g. bond lengths, and bond angles), while at the same time producing meaningful fluctuations globally.

We implement the idea in the decoder of a variational autoencoder on two protein systems. This is primarily a proof of concept, and this implementation has several limitations. First of all, the standard deviations in the κ-prior of the internal degrees of freedom are currently set as a hyperparameter. These could be estimated from data, either directly within the current VAE setup, or using a preexisting model of protein local structure. Another limitation of the current model is that the produced fluctuations are generally too small to fully cover the individual modes of the target densities, mostly due to the first-order approximation we make for the atom fluctuations. This also comes into play in the multimodal setting, where our simple VAE falls short compared to the more complex flow-based method when enough data is available. Latent variable models, such as VAEs, can capture a complex distribution by decomposing it into a continuous mixture of Gaussians. While our proof-of-concept VAE model has demonstrated its ability to capture multimodal distributions, it lacks the expressivity to fully cover all individual modes. This obstacle could potentially be overcome by constructing a hierarchical VAE, where samples are constructed as a multi-step process, similar to the generation process in diffusion models. In fact, we believe that our fundamental approach of induced covariance matrices could be a fruitful way to make diffusion models in internal coordinates scale to larger systems, by allowing for larger non-disruptive steps. We leave these extensions for future work.

7 Code and data availability

The code base, NMR datasets and in-house generated MD data are available at this github repository: https://github.com/mearts/VAE_covariance_matters.

8 Acknowledgements

The work was supported by the Novo Nordisk Foundation through project grant NNF18OC0052719 and through the Center for Basic Machine Learning Research in Life Science (MLLS, grant NNF20OC0062606). We also acknowledge support from the Pioneer Centre for AI (DRNF grant number P1). Finally, we thank Tone Bengtsen for running the in-house MD simulation for 1pga.

References

Mohammed AlQuraishi. pnerf: Parallelized conversion from internal to cartesian coordinates. *bioRxiv*, pp. 385450, 2018.

Namrata Anand and Tudor Achim. Protein structure and sequence generation with equivariant denoising diffusion probabilistic models. *arXiv preprint arXiv:2205.15019*, 2022.

Namrata Anand and Possu Huang. Generative modeling for protein structures. *Advances in neural information processing systems*, 31, 2018.

Marloes Arts, Victor Garcia Satorras, Chin-Wei Huang, Daniel Zuügner, Marco Federici, Cecilia Clementi, Frank Noé, Robert Pinsler, and Rianne van den Berg. Two for one: Diffusion models and force fields for coarse-grained molecular dynamics. *Journal of Chemical Theory and Computation*, 19(18):6151–6159, 2023.

Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

C. Benros, A.G. de Brevern, C. Etchebest, and S. Hazout. Assessing a novel approach for predicting local 3D protein structures from sequence. *Proteins*, 62:865–880, 2006.

W. Boomsma, K.V. Mardia, C.C. Taylor, J. Ferkinghoff-Borg, A. Krogh, and T. Hamelryck. A generative, probabilistic model of local protein structure. *Proc Natl Acad Sci USA*, 105(26):8932–8937, 2008.

Wouter Boomsma, Jes Frellsen, Tim Harder, Sandro Bottaro, Kristoffer E Johansson, Pengfei Tian, Kasper Stovgaard, Christian Andreetta, Simon Olsson, Jan B Valentin, et al. Phaistos: A framework for markov chain monte carlo simulation and inference of protein structure. *Journal of computational chemistry*, 34 (19):1697–1705, 2013.

Wouter Boomsma, Pengfei Tian, Jes Frellsen, Jesper Ferkinghoff-Borg, Thomas Hamelryck, Kresten Lindorff-Larsen, and Michele Vendruscolo. Equilibrium simulations of proteins using molecular fragment replacement and nmr chemical shifts. *Proceedings of the National Academy of Sciences*, 111(38):13852–13857, 2014.

Sandro Bottaro, Wouter Boomsma, Kristoffer E. Johansson, Christian Andreetta, Thomas Hamelryck, and Jesper Ferkinghoff-Borg. Subtle monte carlo updates in dense molecular systems. *Journal of Chemical Theory and Computation*, 8(2):695–702, 2012.

C. Bystroff, V. Thorsson, and D. Baker. HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. *J Mol Biol*, 301(1):173–190, 2000.

AC Camproux, P. Tuffery, JP Chevrolat, JF Boisvieux, and S. Hazout. Hidden Markov model approach for identifying the modular framework of the protein backbone. *Protein Eng Des Sel*, 12(12):1063–1073, 1999.

AC Camproux, R. Gautier, and P. Tufféry. A hidden Markov model derived structural alphabet for proteins. *J Mol Biol*, 339(3):591–605, 2004.

Yilun Chen, Ami Wiesel, Yonina C Eldar, and Alfred O Hero. Shrinkage algorithms for mmse covariance estimation. *IEEE transactions on signal processing*, 58(10):5016–5029, 2010.

John D Chodera and Frank Noé. Markov state models of biomolecular conformational dynamics. *Current opinion in structural biology*, 25:135–144, 2014.

Thomas E Creighton. *Proteins: structures and molecular properties*. Macmillan, 1993.

AG de Brevern, C. Etchebest, and S. Hazout. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins*, 41(3):271–287, 2000.

Peter Eastman, Jason Swails, John D Chodera, Robert T McGibbon, Yutong Zhao, Kyle A Beauchamp, Lee-Ping Wang, Andrew C Simmonett, Matthew P Harrigan, Chaya D Stern, et al. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology*, 13(7): e1005659, 2017.

T. Edgoose, L. Allison, and DL Dowe. An MML classification of protein structure that knows about angles and sequence. *Pac Symp Biocomput*, 3:585–596, 1998.

Raphael R Eguchi, Christian A Choe, and Po-Ssu Huang. Ig-vae: Generative modeling of protein structure by direct 3d coordinate generation. *PLoS computational biology*, 18(6):e1010271, 2022.

G. Favrin, A. Irbäck, and F. Sjunnesson. Monte Carlo update for chain molecules: Biased Gaussian steps in torsional space. *J. Chem. Phys.*, 114:8154–8158, 2001.

T. Hamelryck, J.T. Kent, and A. Krogh. Sampling realistic protein conformations using local structural bias. *PLoS Comput Biol*, 2(9):e131, 2006.

Carlos X Hernández, Hannah K Wayment-Steele, Mohammad M Sultan, Brooke E Husic, and Vijay S Pande. Variational encoding of complex dynamics. *Physical Review E*, 97(6):062412, 2018.

Moritz Hoffmann, Martin Konrad Scherer, Tim Hempel, Andreas Mardt, Brian de Silva, Brooke Elena Husic, Stefan Klus, Hao Wu, J Nathan Kutz, Steven Brunton, and Frank Noé. Deeptime: a python library for machine learning dynamical models from time series data. *Machine Learning: Science and Technology*, 2021.

John Ingraham, Max Baranov, Zak Costello, Vincent Frappier, Ahmed Ismail, Shan Tie, Wujie Wang, Vincent Xue, Fritz Obermeyer, Andrew Beam, et al. Illuminating protein space with a programmable generative model. *bioRxiv*, pp. 2022–12, 2022.

A. Irbäck and S. Mohanty. Profasi: a monte carlo simulation package for protein folding and aggregation. *J. Comput. Chem.*, 27(13):1548–1555, 2006.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

Bowen Jing, Gabriele Corso, Jeffrey Chang, Regina Barzilay, and Tommi Jaakkola. Torsional diffusion for molecular conformer generation. *arXiv preprint arXiv:2206.01729*, 2022.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Alphafold meets flow matching for generating protein ensembles. *arXiv preprint arXiv:2402.04845*, 2024.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Hiremaglur K Kesavan and Jagat Narain Kapur. The generalized maximum entropy principle. *IEEE Transactions on systems, Man, and Cybernetics*, 19(5):1042–1052, 1989.

Jonas Köhler, Yaoyi Chen, Andreas Krämer, Cecilia Clementi, and Frank Noé. Flow-matching: Efficient coarse-graining of molecular dynamics without forces. *Journal of Chemical Theory and Computation*, 19 (3):942–952, 2023.

Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, and David E Shaw. How fast-folding proteins fold. *Science*, 334(6055):517–520, 2011.

James A Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E Hauser, and Carlos Simmerling. ff14sb: improving the accuracy of protein side chain and backbone parameters from ff99sb. *Journal of chemical theory and computation*, 11(8):3696–3713, 2015.

Andreas Mardt, Luca Pasquali, Hao Wu, and Frank Noé. Vampnets for deep learning of molecular kinetics. *Nature communications*, 9(1):1–11, 2018.

Andreas Mardt, Luca Pasquali, Frank Noé, and Hao Wu. Deep learning markov and koopman models with physical constraints. In *Mathematical and Scientific Machine Learning*, pp. 451–475. PMLR, 2020.

Lutz Molgedey and Heinz Georg Schuster. Separation of a mixture of independent signals using time delayed correlations. *Physical review letters*, 72(23):3634, 1994.

Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

Jan-Hendrik Prinz, Hao Wu, Marco Sarich, Bettina Keller, Martin Senne, Martin Held, John D Chodera, Christof Schütte, and Frank Noé. Markov models of molecular kinetics: Generation and validation. *The Journal of chemical physics*, 134(17):174105, 2011.

Ali Punjani and David J Fleet. 3d flexible refinement: structure and motion of flexible proteins from cryo-em. *BioRxiv*, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Dan Rosenbaum, Marta Garnelo, Michal Zielinski, Charlie Beattie, Ellen Clancy, Andrea Huber, Pushmeet Kohli, Andrew W Senior, John Jumper, Carl Doersch, et al. Inferring a continuous distribution of atom coordinates from cryo-em images using vaes. *arXiv preprint arXiv:2106.14108*, 2021.

Marco Sarich, Ralf Banisch, Carsten Hartmann, and Christof Schütte. Markov state models for rare events in molecular dynamics. *Entropy*, 16(1):258–286, 2013.

Mathias Schreiner, Ole Winther, and Simon Olsson. Implicit transfer operator learning: Multiple time-resolution surrogates for molecular dynamics. *arXiv preprint arXiv:2305.18046*, 2023.

Schrödinger. *The PyMOL Molecular Graphics System.* version 2.5.2.

Ch Schütte, Alexander Fischer, Wilhelm Huisinga, and Peter Deuflhard. A direct approach to conformational dynamics based on hybrid monte carlo. *Journal of Computational Physics*, 151(1):146–168, 1999.

Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

Nina Singhal and Vijay S Pande. Error analysis and efficient sampling in markovian state models for molecular dynamics. *The Journal of chemical physics*, 123(20):204909, 2005.

Mohammad M Sultan, Hannah K Wayment-Steele, and Vijay S Pande. Transferable neural networks for enhanced sampling of protein dynamics. *Journal of chemical theory and computation*, 14(4):1887–1894, 2018.

Christian B Thygesen, Christian Skjødt Steenmans, Ahmad Salim Al-Sibahi, Lys Sanz Moreta, Anders Bundgård Sørensen, and Thomas Hamelryck. Efficient generative modelling of protein structure fragments using a deep markov model. In *International Conference on Machine Learning*, pp. 10258–10267. PMLR, 2021.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *bioRxiv*, pp. 2022–12, 2022.

Kevin E Wu, Kevin K Yang, Rianne van den Berg, James Y Zou, Alex X Lu, and Ava P Amini. Protein structure generation via folding diffusion. *arXiv preprint arXiv:2209.15611*, 2022a.

Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan Wu, Qi Xie, Bonnie Berger, et al. High-resolution de novo structure prediction from primary sequence. *BioRxiv*, pp. 2022–07, 2022b.

Tian Xie, Arthur France-Lanord, Yanming Wang, Yang Shao-Horn, and Jeffrey C Grossman. Graph dynamical networks for unsupervised learning of atomic scale dynamics in materials. *Nature communications*, 10(1): 1–9, 2019.

Ellen D Zhong, Tristan Bepler, Bonnie Berger, and Joseph H Davis. Cryodrgn: reconstruction of heterogeneous cryo-em structures using neural networks. *Nature methods*, 18(2):176–185, 2021.

A Experimental details

Model. See Fig. 3 for an overview of the model. The encoder and the decoder of the VAE are simple three-layer MLPs (multilayer perceptrons). Given a protein with M backbone atoms, the encoder takes $(2 \times M - 5) \times 2$ inputs, corresponding to $(M - 3)$ dihedrals and $(M - 2)$ bond angles which are fed in as pairs of $(\sin, \cos)$ inputs to avoid periodicity issues. Similarly, the decoder yields $(2 \times M - 5) \times 2$ outputs, which can be converted to angles in the $[-\pi, \pi]$ interval using the 2-argument arctangent (`atan2`). The MLP linear layer sizes of the encoder are $[128, 64, 32]$, mapping to a 16-dimensional latent space, and layer sizes of the decoder are $[32, 64, 128]$ (reverse of the encoder).

We use a standard U-Net (Ronneberger et al., 2015) to predict atom fluctuation constraints λ from the mean predictions that the decoder outputs. The predicted mean for the internal degrees of freedom $\boldsymbol{\mu_\kappa}$ is translated into a mean structure $\boldsymbol{\mu_x}$ using pNeRF (AlQuraishi, 2018), from which we calculate a pairwise distance matrix. This $M \times M$ matrix with a single channel serves as the input to the U-Net, which scales the number of channels up to 1024 in four steps before scaling back down to one channel in four steps.

All datasets were split 90%-10% into a training and validation set, with the same split for our VAE and all baselines. The best model is selected based on the validation loss. As we are handling very small datasets and will eventually evaluate per-atom fluctuations, it is not feasible to set aside a reliable test set that reflects the same distribution. Since we are in a generative modelling scenario rather than one where we are in need of predictive accuracy, the main focus is to fit a parametric model to a given data distribution. We therefore evaluate our generative performance based on the reference data, corresponding to the entire dataset, following related work (Köhler et al., 2023; Arts et al., 2023; Schreiner et al., 2023).

The weights for the κ-prior and auxiliary loss were explored with grid search (see Appendix D), values chosen for the models reported in the main paper are shown in Table A1 together with other experimental details. The model training starts with a warm-up phase in two different ways: 1) predicting $\boldsymbol{\mu_\kappa}$ only, with $\boldsymbol{\Sigma} = \mathbf{I}$ and 2) linearly increasing the weight of the KL-term from 0 to 1. Proteins in the low data regime (unimodal setting) have a 100 epoch mean-only warm-up and a 200 epoch KL warm-up, while proteins in the high data regime (multimodal setting) have a 3 epoch mean-only warm-up and an 8 epoch KL warm-up. All models were trained using an Adam optimizer with a learning rate of $5e^{-4}$, on a Nvidia Quadro RTX (48GB) GPU.

Final metrics are calculated on structures sampled from the model. For the evaluation in the unimodal setting, the number of samples was chosen to be equal to the total number of data points (25, 41 and 400 for 1unc, 1pga and 1fsd, respectively) for all models. For the multimodal cases, 400.000 samples were drawn for TIC analysis. TICA was done using the `Deeptime` library (Hoffmann et al., 2021), using a lagtime of 100 and reducing the high-dimensional input, consisting of dihedrals and pairwise distances, to two dimensions. The TICA model is fit on the reference data (ordered in time), from which the resulting linear map is stored and applied to sampled structures from the VAE and baselines. All structure visualizations were done using PyMOL (Schrödinger, version 2.5.2).

Table A1: Experimental details for test cases.

	# train	# validation	# residues	# epochs	batch size	a	w_{aux}
1unc	23	2	36	1000	32	50	1
1fsd	37	4	28	1000	32	1	25
1pga	360	40	56	1000	32	50	25
cln025	481269	53474	10	50	64	25	50
2f4k	565117	62790	35	50	32	50	1

Molecular dynamics details. The molecular dynamics simulation for 1pga was done in OpenMM (Eastman et al., 2017), using an Amber forcefield (Maier et al., 2015), water type TIP3P, box geometry "rhombic dodecahedron" and a padding of 1 nm on each side of the solvated protein (i.e. 2 nm in total). The simulation is 20ns in total with a 50ps time lag, giving 400 structures. For MD details on cln025 and 2f4k we refer the reader to Lindorff-Larsen et al. (2011).

B Quantitative results

B.1 Unimodal setting, low data regime

Table A2: MSE (lower is better) to reference for atom fluctuations, unimodal setting, mean ± standard deviation over 5 sampling runs.

	VAE	**κ-prior (fixed)**	**κ-prior (learned)**	**Standard estimator**	**Flow**
1unc	**0.020 ± 0.005**	3.268 ± 0.791	**0.015 ± 0.003**	4.219 ± 0.990	142.297 ± 24.191
1fsd	**0.496 ± 0.051**	16.148 ± 3.049	10.339 ± 1.675	9.583 ± 1.954	112.031 ± 16.897
1pga	**0.039 ± 0.002**	3.464 ± 0.200	1.550 ± 0.112	1040.001 ± 60.128	3070.602 ± 109.194

B.2 Multimodal setting, high data regime

Table A3: Jensen-Shannon distance (lower is better) between binned Boltzmann distributions, i.e. $\exp\left(-\frac{\text{free energy}}{k_B T}\right)$, comparing VAE and baselines to the reference, multimodal setting. Mean ± standard deviation over 5 sampling runs.

	VAE	**κ-prior (fixed)**	**κ-prior (learned)**	**Standard estimator**	**Flow**
cln025	0.4555 ± 0.0003	0.5384 ± 0.0004	0.6060 ± 0.0001	0.6861 ± 0.0002	**0.1931 ± 0.0006**
2f4k	0.3740 ± 0.0005	0.2967 ± 0.0001	0.3418 ± 0.0003	0.5179 ± 0.0003	**0.1833 ± 0.0003**

C VAE sampling

C.1 Comparing constraints to atom fluctuations across samples

As derived in Eq. (7), we can evaluate the constraint value C_m for each atom m given a set of Lagrange multipliers. These constraints were placed on the squared atom displacements, which is equivalent to the variance along the atom chain. Fig. A1 demonstrates that the isotropic fluctuations of 400 1pga samples drawn from the VAE are indeed quite close to C calculated from separately sampled sets of Lagrange multipliers. Since the constraints are placed on non-superposed (i.e. not structurally aligned) protein structures[4], this plot shows the variance along the atom chain for non-superposed structures.

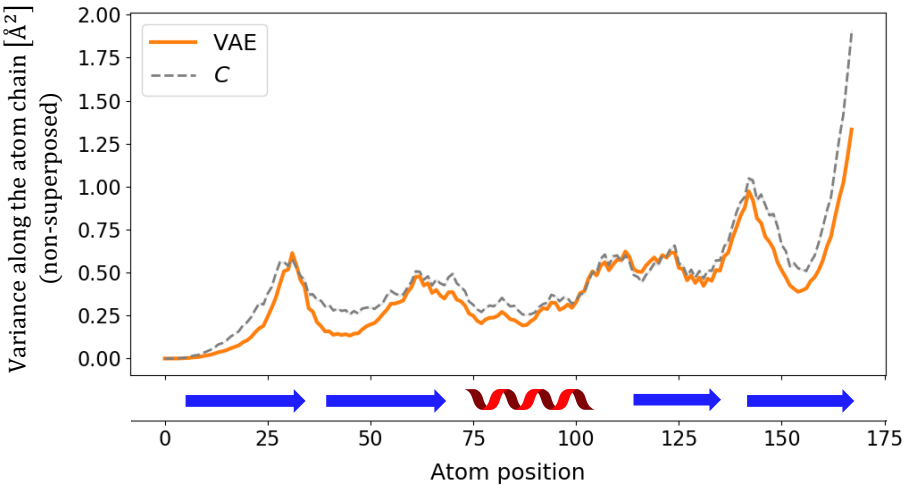

Figure A1: Variance along the atom chain for non-superposed 1pga structures sampled with the VAE (orange) compared to constraints C calculated from predicted λ-values (grey dashed). Secondary structure element locations are indicated.

C.2 Visualization of sampled structures in the unimodal setting

Fig. A2 shows sampled superposed ensembles for our model and baselines, as well as the MD/NMR reference. This demonstrates that VAE samples, where global constraints were enforced, generally have globally consistent fluctuations compared to the reference data. In contrast, the baselines tend to exhibit fluctuations that are too large, which can lead to unphysical structures containing crossings and, in some cases, lacking secondary structure elements.

C.3 Latent space visualization in the multimodal setting

In this section, we visualize the VAE latent space in the multimodal setting (cln025) in Fig. A3. Moreover, we demonstrate how 100 random samples from latent space map to structure samples in the TICA free energy landscape, and show the 3D structures that correspond to these samples. Transitions from the native state to more unfolded conformations can be observed when going from the cluster in the top right of TICA space towards the left. Depending on $\tilde{\mathbf{\Sigma}}$, fluctuations around the means (which are decoded from the latent space samples) can vary in size. Therefore, means that are close together in terms of latent space location do not necessarily lead to sampling similar 3D structures. Moreover, we used a UMAP to reduce the number of latent dimensions from 16 to 2, and this simplified representation might not capture the full complexity of the latent space. Nonetheless, it is apparent that more unfolded structures largely originate from the rightmost cluster in latent space.

[4]Sampled protein structures are built using pNeRF(AlQuraishi, 2018), which builds the chain step-by-step, thereby corresponding to our post-rotational constraints.

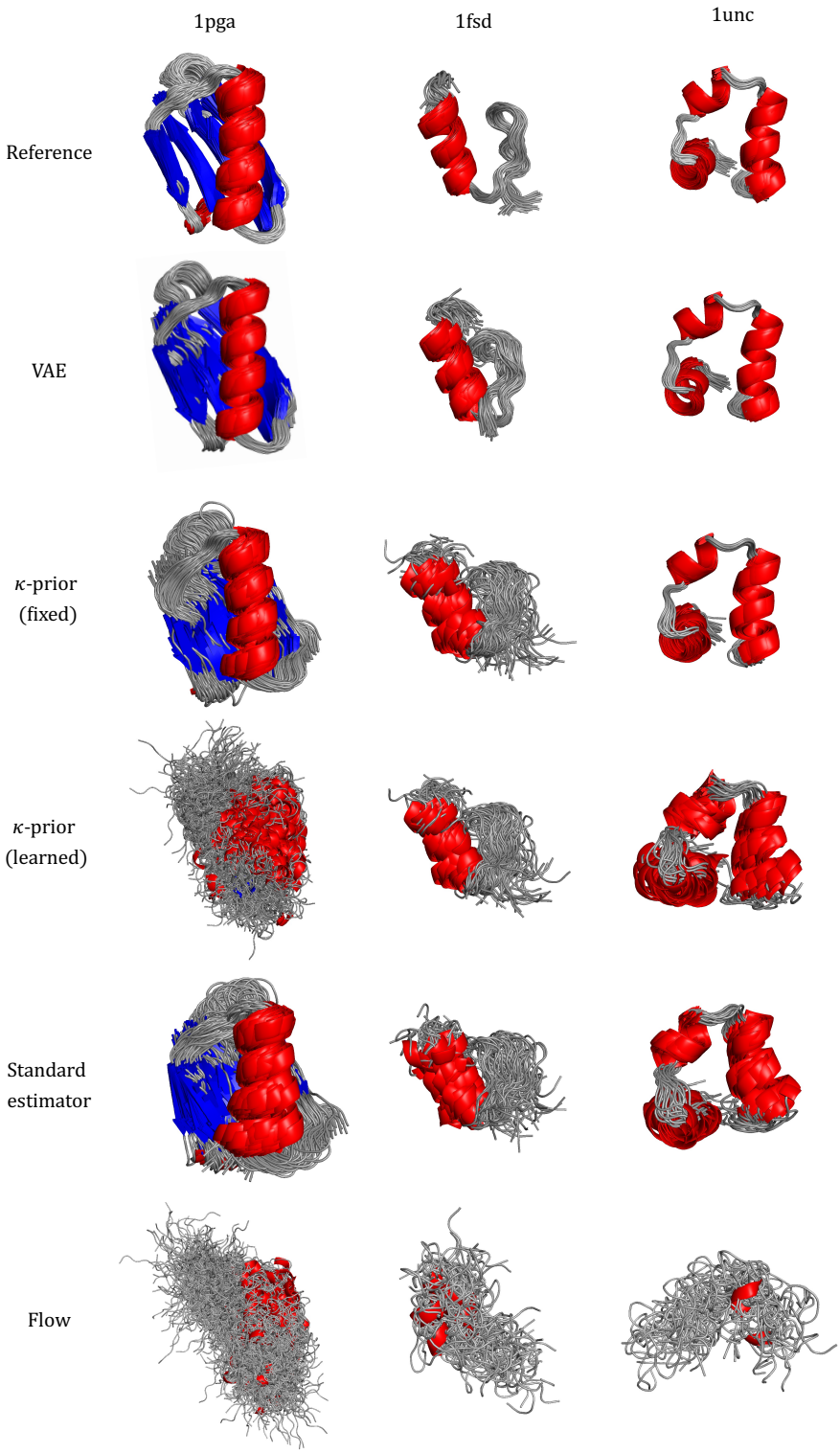

Figure A2: Visualization of ensembles for reference data, the VAE model and baselines for 1pga, 1fsd and 1unc. Number of samples is equal to the reference ensemble (400, 41 and 25 for 1pga, 1fsd and 1unc, respectively.)

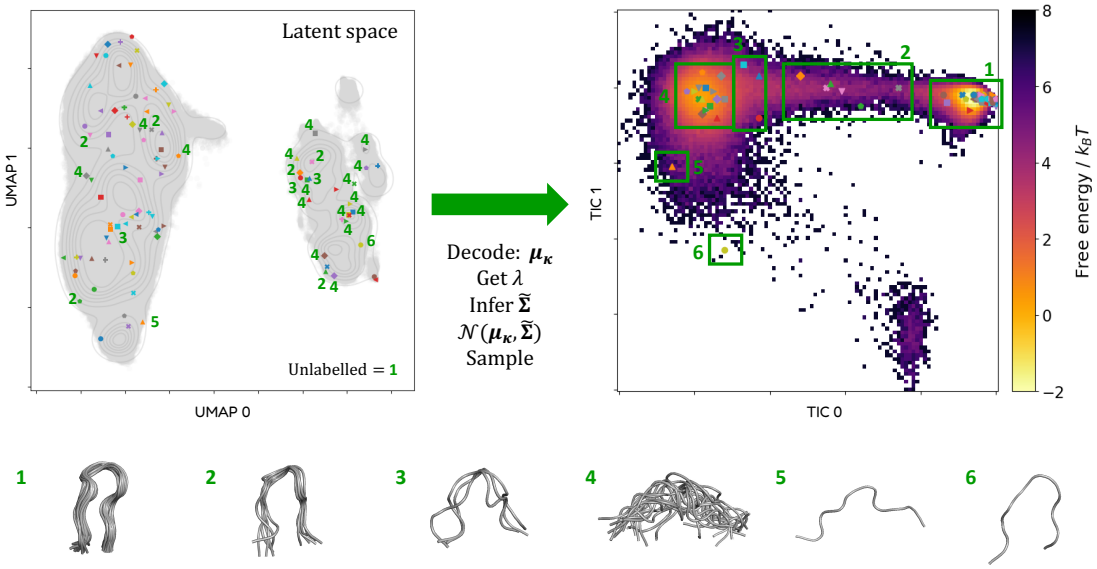

Figure A3: Top left: UMAP reduction to 2D of the originally 16-dimensional VAE latent space, with a 100 samples shown in random shapes and colors. The grey scatterplot depicts the aggregated posterior, with the KDE of the aggregated posterior as grey lines. Annotated green numbers correspond to boxes in the TICA free energy landscape (all structures corresponding to box 1 are left unlabelled to avoid clutter). Top right: structure samples corresponding to latent space samples visualized in TICA space with the same symbols and colors as the latent space samples. Samples are grouped together in green numbered boxes. Bottom row: 3D structures corresponding to the different numbered boxes in the TICA plot.

D Ablation for hyperparameters

The two main hyperparameters that need to be chosen in the VAE setting are the strength of the κ-prior a, and the weight of the regularizing loss w_{aux}. These two weights can be set to prioritize local or global constraints in different ways. We demonstrate the effect on a unimodal case (protein G, 1pga) and a multimodal case (chignolin, cln025). In both cases, results are shown for a gridsearch over $a = [1, 25, 50]$ and $w_{\text{aux}} = [1, 25, 50]$.

D.1 Unimodal

Fig. A4 shows results for the ablation on a and w_{aux} in the unimodal setting. Increasing the strength of the κ-prior through a while keeping w_{aux} constant corresponds to narrower distributions in the Ramachandran plot and bond angle distributions. A higher weight w_{aux} for a constant a leads to stronger global constraints, as demonstrated by the fluctuations along the atom chain.

D.2 Multimodal

To understand the impact of hyperparameters in the multimodal setting, we first consider the impact on samples drawn from the VAE that was trained with a fixed κ-prior, which depends on hyperparameter a. Fig. A5 illustrates how the distribution in the TIC free energy landscape changes when strengthening the prior. For $a = 1$, there is a preference towards the metastable cluster on the top left, while increasing the value of a leads to a stronger preference for the lowest energy cluster on the top right.

When sampling from the VAE, where constraints are imposed on top of the κ-prior, there is interplay between a and w_{aux}, as shown in Fig. A6. Even though the exact trend is less clear here, the relative values of the hyperparameters have an observable influence on e.g. the width of the "bridge" between the topmost two clusters, the size of the higher-energy downward extrusion of the top left cluster, and the spread towards the less populated cluster on the bottom right.

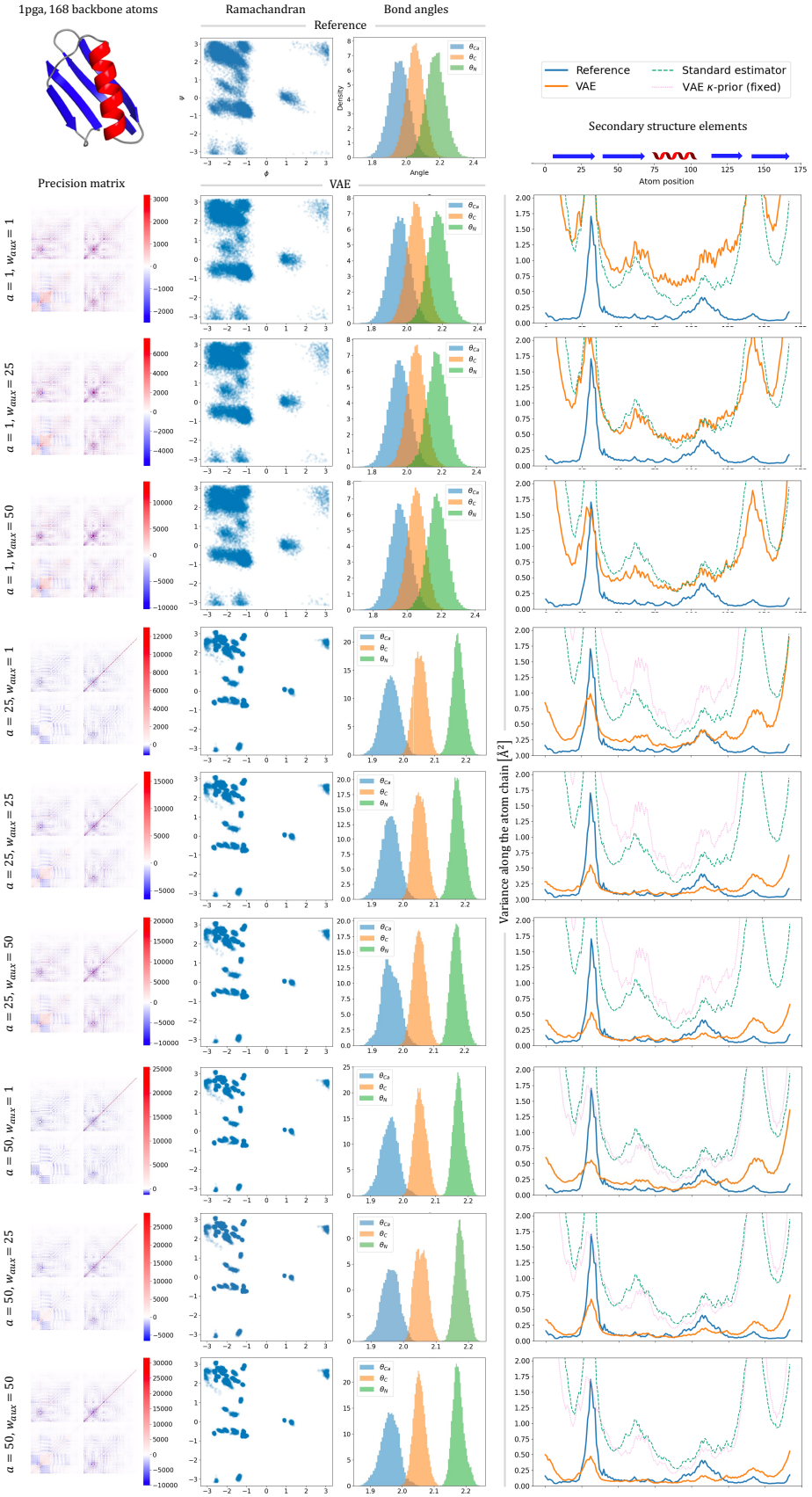

Figure A4: Ablation of a and w_{aux} for protein G (1pga, structure shown at top left). From left to right: precision matrix example predicted by the VAE, Ramachandran plot, bond angle distributions, fluctuations along the atom chain (secondary structure elements indicated, VAE κ-prior (fixed) out of scale for $a = 1$).

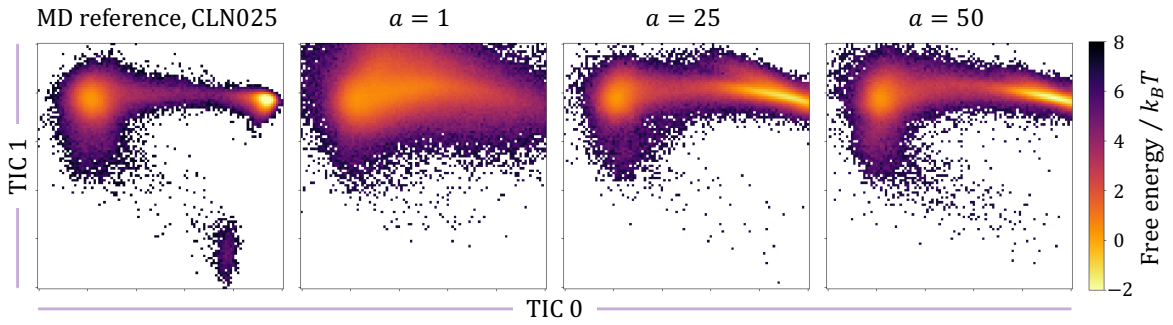

Figure A5: Influence of hyperparameter a on samples drawn from the VAE with a fixed κ-prior (without imposing constraints) for chignolin (cln025), visualized in TIC space.

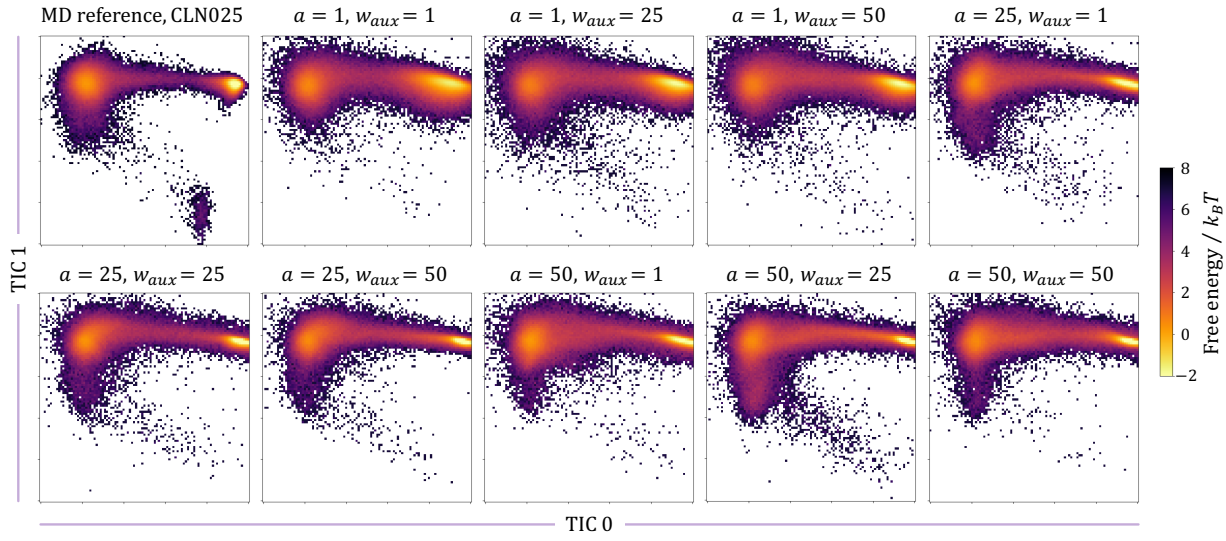

Figure A6: Hyperparameter ablation of a and w_{aux} VAE samples for chignolin (cln025), visualized in TIC space.