## 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 $\lambda$ 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. The best model is selected based on the validation loss. The weights for the $\kappa$-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 $5\mathrm{e}^{-4}$, 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 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 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** | $\mathrm{w}_{\mathrm{aux}}$ |
|---|---|---|---|---|---|---|---|
| **1unc** | 23 | 2 | 36 | 1000 | 32 | 50 | 1 |
| **1fsd** | 37 | 4 | 28 | 1000 | 32 | 25 | 1 |
| **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.

| | **VAE** | **$\kappa$-prior (fixed)** | **$\kappa$-prior (learned)** | **Standard estimator** | **Flow** |
|---|---|---|---|---|---|
| **1unc** | 0.021 | 2.080 | **0.013** | 5.888 | 122.490 |
| **1fsd** | **0.585** | 13.949 | 12.732 | 9.666 | 107.052 |
| **1pga** | **0.040** | 3.654 | 1.709 | 1154.914 | 3157.693 |

## 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.

| | **VAE** | **$\kappa$-prior (fixed)** | **$\kappa$-prior (learned)** | **Standard estimator** | **Flow** |
|---|---|---|---|---|---|
| **cln025** | 0.456 | 0.539 | 0.606 | 0.686 | **0.194** |
| **2f4k** | 0.373 | 0.297 | 0.342 | 0.517 | **0.183** |

# C VAE SAMPLING

## C.1 COMPARING CONSTRAINTS TO ATOM FLUCTUATIONS ACROSS SAMPLES

As derived in Eq. (9), 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 400 separately sampled sets of Lagrange multipliers. Since the constraints are placed on non-superposed (i.e. not structurally aligned) protein structures[3], this plot shows the variance along the atom chain for non-superposed structures.

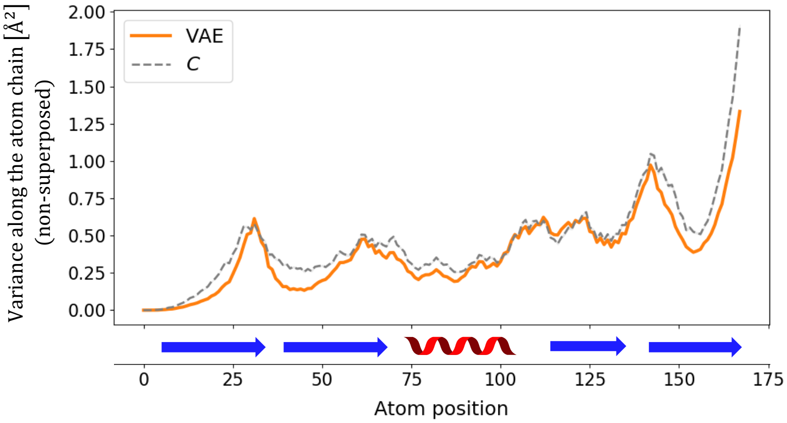

Figure A1: Variance along the atom chain for non-superposed 1pga structures sampled with the VAE (orange) compared to constraints $C$ calculated from predicted $\lambda$-values (grey dashed). Secondary structure element locations are indicated.

## 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.

---

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

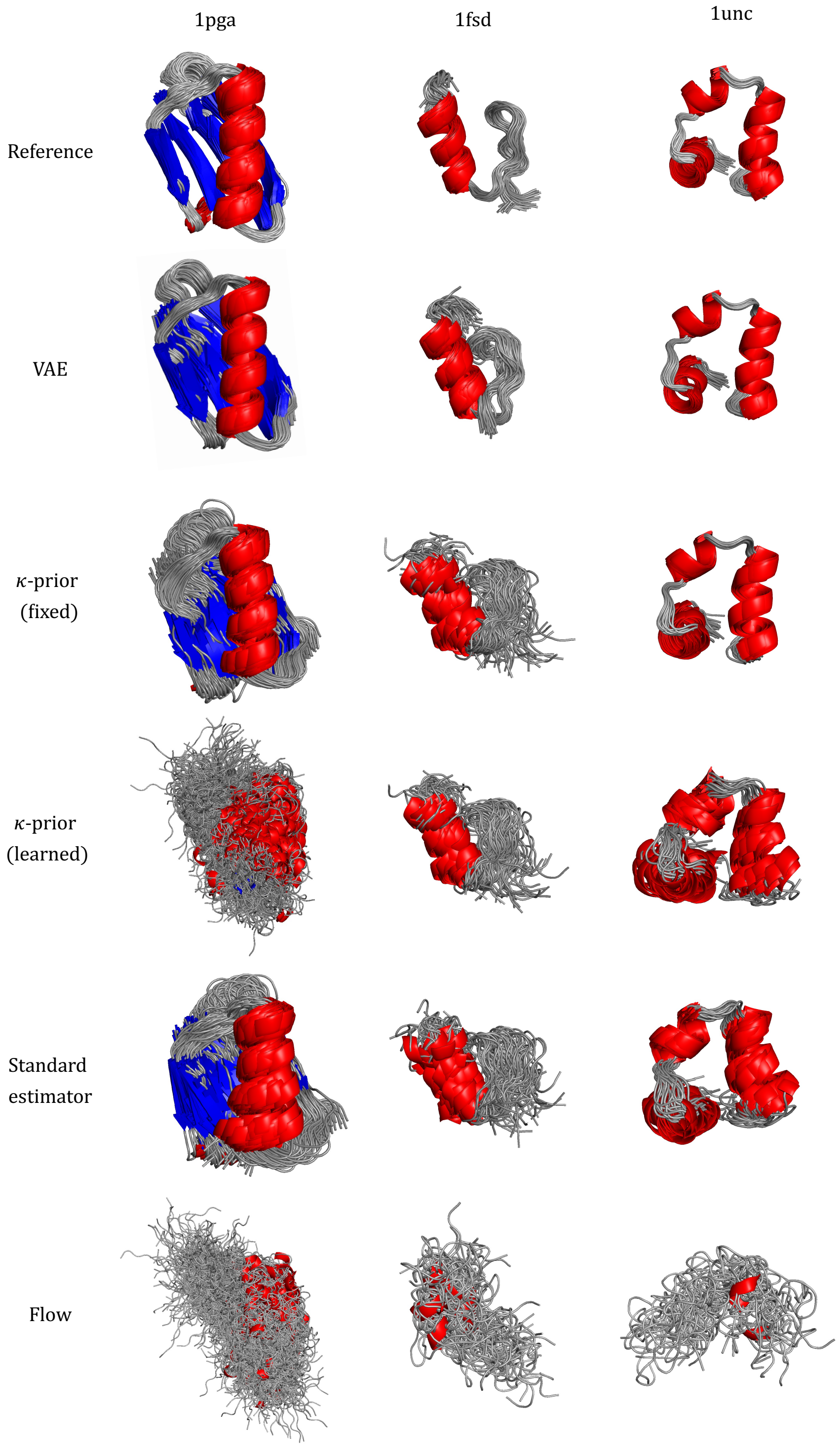

Figure A2: Visualization of ensembles for reference data, the VAE model and baselines for 1pga, 1fsd and 1unc. Number of samples is equal to the reference ensemble (400, 41 and 25 for 1pga, 1fsd and 1unc, respectively.)

### 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 $\tilde{\Sigma}$, 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.

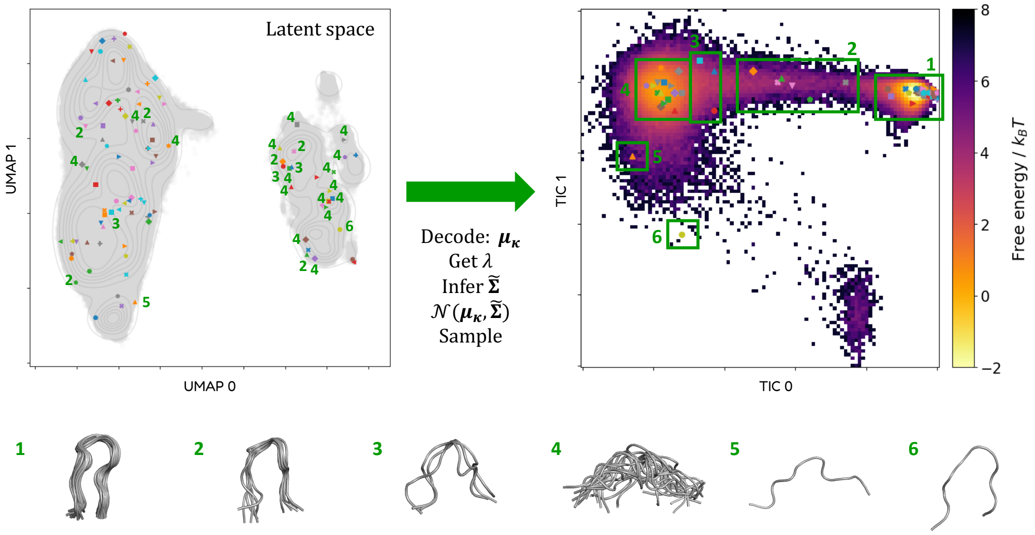

Figure A3: Top left: UMAP reduction to 2D of the originally 16-dimensional VAE latent space, with a 100 samples shown in random shapes and colors. The grey scatterplot depicts the aggregated posterior, with the KDE of the aggregated posterior as grey lines. Annotated green numbers correspond to boxes 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 $\kappa$-prior $a$, and the weight of the regularizing loss $w_{\mathrm{aux}}$. 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 $w_{\mathrm{aux}} = [1, 25, 50]$.

## D.1 UNIMODAL

Fig. A4 shows results for the ablation on $a$ and $w_{\mathrm{aux}}$ in the unimodal setting. Increasing the strength of the $\kappa$-prior through $a$ while keeping $w_{\mathrm{aux}}$ constant corresponds to narrower distributions in the Ramachandran plot and bond angle distributions. A higher weight $w_{\mathrm{aux}}$ 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 $\kappa$-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 $\kappa$-prior, there is interplay between $a$ and $w_{\mathrm{aux}}$, 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.

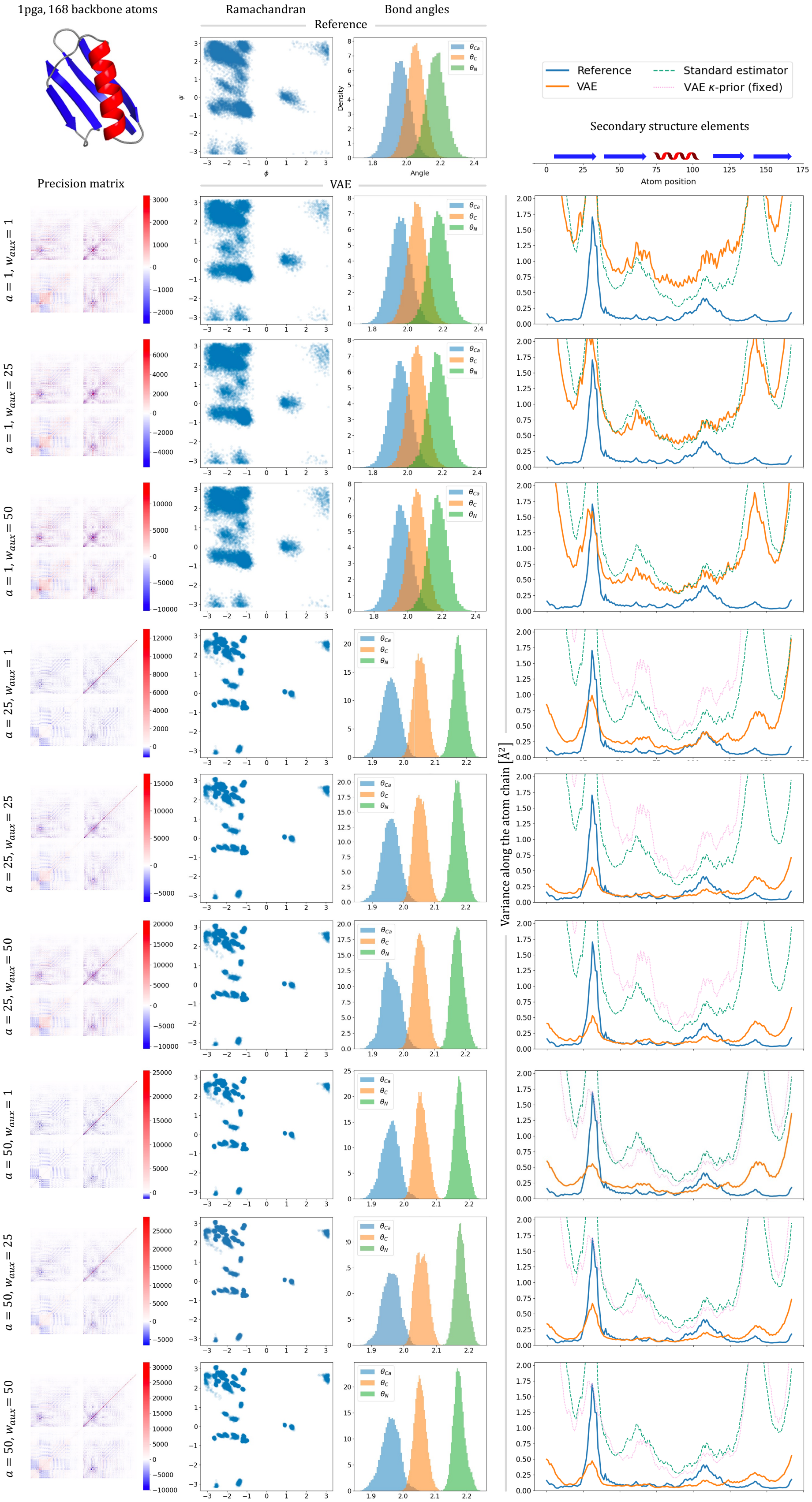

Figure A4: Ablation of $a$ and $w_{\text{aux}}$ for protein G (1pga, structure shown at top left). From left to right: precision matrix example predicted by the VAE, Ramachandran plot, bond angle distributions, fluctuations along the atom chain (secondary structure elements indicated, VAE $\kappa$-prior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

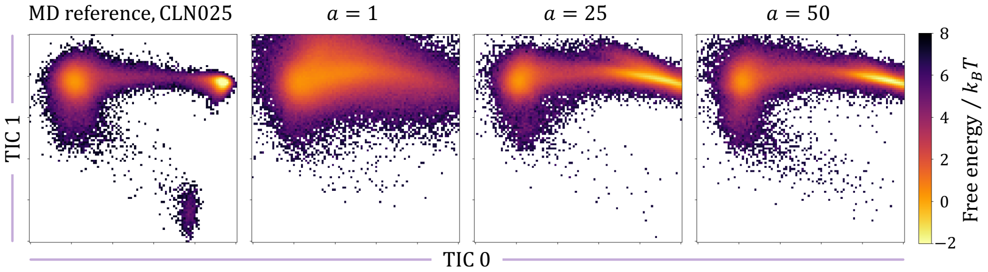

Figure A5: Influence of hyperparameter $a$ on samples drawn from the VAE with a fixed $\kappa$-prior (without imposing constraints) for chignolin (cln025), visualized in TIC space.

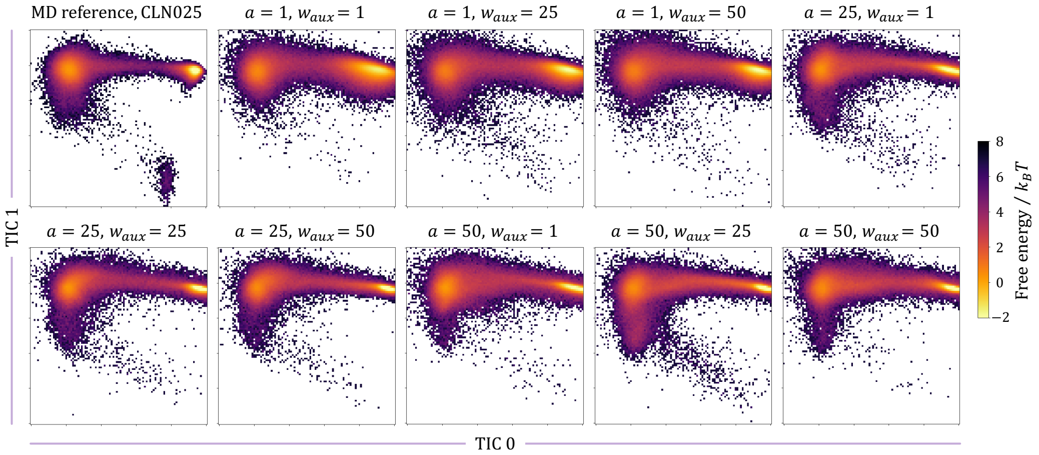

Figure A6: Hyperparameter ablation of $a$ and $w_{\text{aux}}$ VAE samples for chignolin (cln025), visualized in TIC space.