A APPENDIX

A.1 ADDITIONAL BACKGROUND

Related Work on Modeling Dynamics For completeness, we discuss recent work modeling temporal aspects of protein dynamics here. Wang et al. (2020) learns the dynamics between a particular starting conformation and a particular target conformation by training a control Hamiltonian represented by a graph neural network. From the perspective of Hamiltonian systems, Chen et al. (2020) introduces Symplectic RNNs, which leverages sympletic integration for learning the dynamics of a physical system. Ingraham et al. (2019b) learns dynamics via a differentiable simulator using Langevin dynamics.

Mardt et al. (2018) introduces VAMPnet, which maps molecular coordinates to Markov states, to capture molecular kinetics. Lee et al. (2019) augments molecular dynamics simulation with deep learning to improve the sampling of the folded states of proteins. Tsai et al. (2020) uses a LSTM to predict protein dynamics.

Another line of work involves enhanced sampling with neural net approximations of slow variables (Chen & Ferguson, 2018; Chen et al., 2019; Ribeiro et al., 2018). Deep and adversarial learning have been employed for such enhanced sampling (Bonati et al., 2019; Zhang et al., 2019). Noé et al. (2019) uses a deep generative neural net to directly sample the equilibrium distribution of a many-body system defined by an energy function, without using molecular simulation.

A.2 CANONICAL CORRELATION ANALYSIS

Given the two embeddings, $\boldsymbol{X} \in \mathbb{R}^{n \times m_1}$ and $\boldsymbol{Y} \in \mathbb{R}^{n \times m_2}$, CCA finds two linear transformations, $\boldsymbol{A} \in \mathbb{R}^{m_1 \times m_1}$ and $\boldsymbol{B} \in \mathbb{R}^{m_2 \times m_2}$, such that

$$\boldsymbol{A}, \boldsymbol{B} = \operatorname*{arg\,min}_{\boldsymbol{A},\boldsymbol{B}} ||\boldsymbol{X}\boldsymbol{A} - \boldsymbol{Y}\boldsymbol{B}||_F^2 \quad \text{s.t.} \quad \boldsymbol{X}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{X} = \boldsymbol{I}_{m_1}, \boldsymbol{Y}^T\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{Y} = \boldsymbol{I}_{m_2}. \tag{10}$$

Equation 10 has a closed form solution given by SVD. It follows that $\max diag(\boldsymbol{X}^T\boldsymbol{A}^T\boldsymbol{B}\boldsymbol{Y})$ denotes the highest correlation between any axis in each embedding.

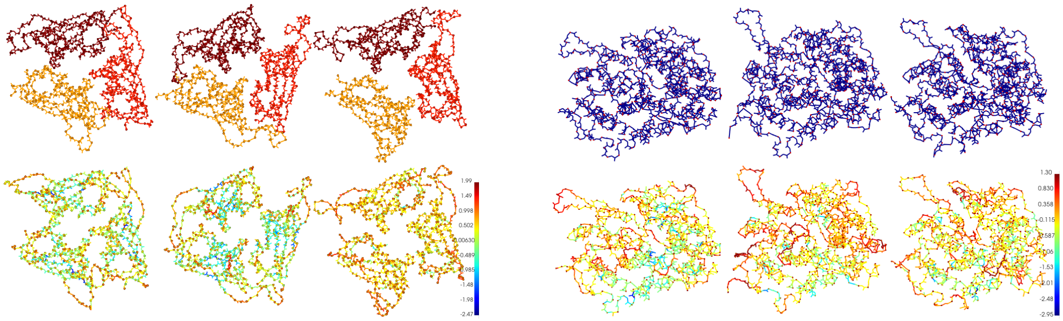
A.3 NETWORK HYPERPARAMETERS

Here we specify the hyperparameters of the networks used in conducting our experiments. Each encoder contains 5 layers with filter sizes, $\{12, 24, 48, 96, 96\}$. The decoder structure is mirrored with filter sizes, $\{128, 128, 64, 32, 16, 3\}$. Each graph attention layer has 4 heads of attention. The dimensions of the intrinsic and extrinsic latent spaces are set to 16 and 32 respectively.

For training, we use ADAM with a learning rate of 1E-3 (Kingma & Ba, 2014). Learning rate decays at a rate of 0.995 per epoch. We train models with a weight decay penalty of 5E-5. The models are trained 100 epochs, which is enough to achieve convergence, with a batch size of 64. Additionally, we set $\lambda_{\mathcal{R}} = $ 5E-1 for the bond length penalty.The neighborhood radius for defining the sparsity of the graph attention layer is set to 2.5 Å in the first layer. This radius is scaled at each layer with the stride of the previous convolution.

Table 6: Average atom-wise L_2 error and bond length error on the test sets using ProGAE. We compare models trained on only intrinsic data, only extrinsic data, and both intrinsic and extrinsic data. It follows that the extrinsic information is needed for accurate reconstruction. Whereas the intrinsic information aids in improving the validity of reconstruction seen in Table 4. We note that the intrinsic-only model can still accurately reconstruct bond lengths, i.e. intrinsic geometry.

	$\mathbf{L_2}$ **Error** (Å)	**Bond Length Error** (Å)
S Protein		
Intrinsic Only	5.27	0.16
Extrinsic Only	1.55	0.39
Intrinsic and Extrinsic	1.56	0.39
hACE2		
Intrinsic Only	3.07	0.06
Extrinsic Only	0.89	0.16
Intrinsic and Extrinsic	0.90	0.16



(a) 3 RBDs of S protein

(b) Ectodomain of Human ACE2

Figure 6: Reconstructions of protein frames from test data using ProGAE. The top row displays the ground truth, while the bottom row displays the corresponding structure generated by the network. Color in the top row denotes separate protein chains, while color in the bottom row indicates the log of atom-wise L_2 error. Color of the bonds indicates the average of the constituent atoms.

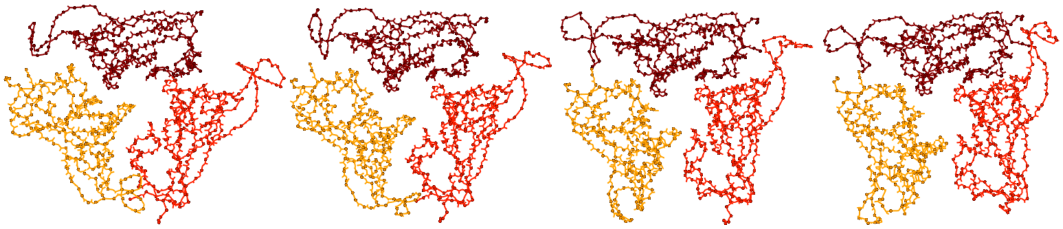
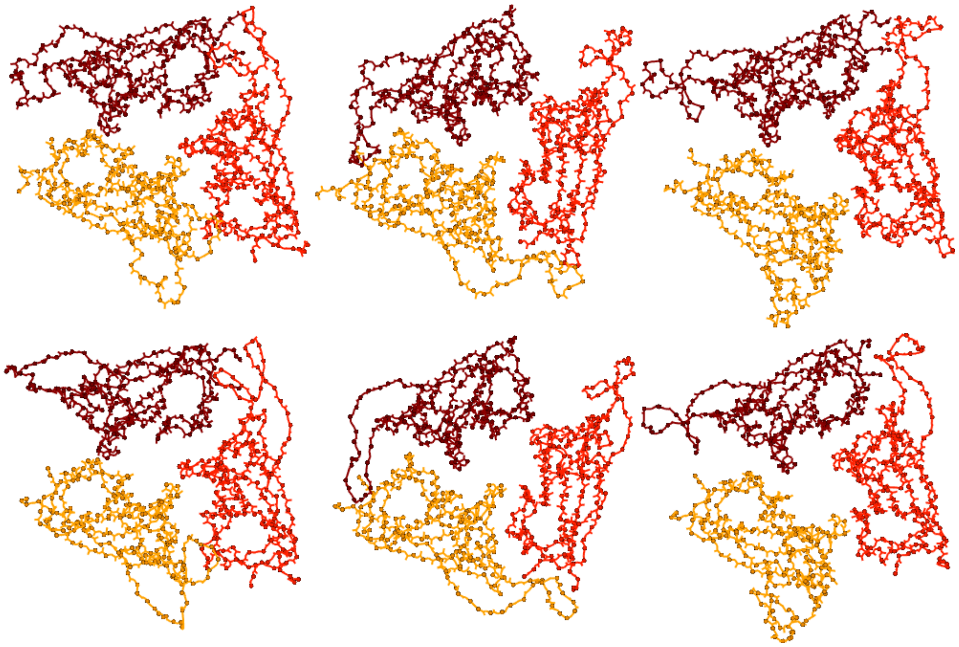
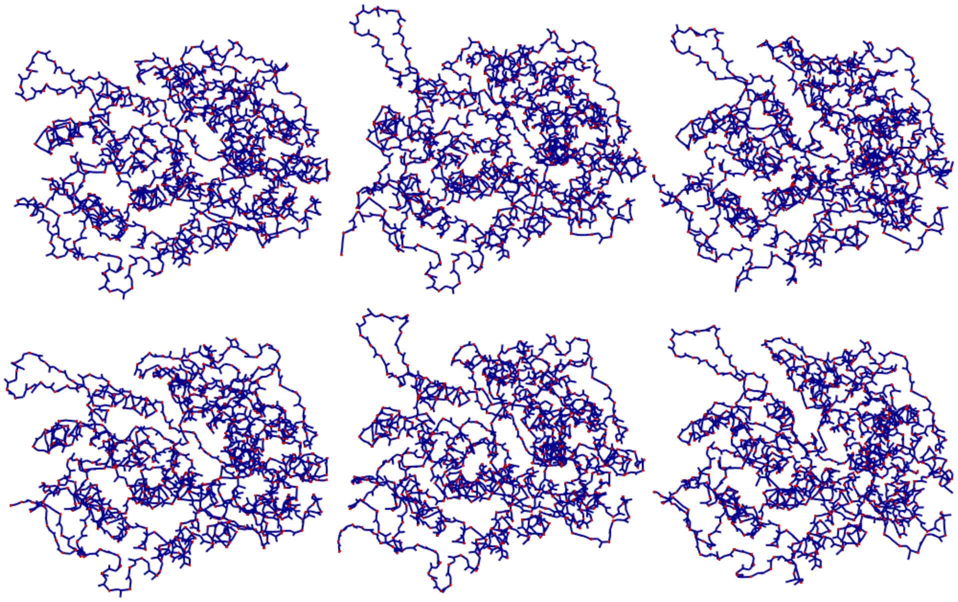


Figure 7: A sample latent interpolation between S proteins from different trajectories. This shows the smooth transition along the latent path analyzed in Figure 5.



(a) 3 RBDs of S protein



(b) Ectodomain of Human ACE2

Figure 8: The protein reconstructions from Figures 6a and 6b. The color in the bottom row indicates different protein fragments, instead of the log of atom-wise L_2 error.