Chemically Transferable Generative Backmapping of Coarse-Grained Proteins

Soojung Yang [1] **Rafael Gómez-Bombarelli** [2]

Abstract

Coarse-graining (CG) accelerates molecular simulations of protein dynamics by simulating sets of atoms as singular beads. Backmapping is the opposite operation of bringing lost atomistic details back from the CG representation. While machine learning (ML) has produced accurate and efficient CG simulations of proteins, fast and reliable backmapping remains a challenge. Rule-based methods produce poor all-atom geometries, needing computationally costly refinement through additional simulations. Recently proposed ML approaches outperform traditional baselines but are not transferable between proteins and sometimes generate unphysical atom placements with steric clashes and implausible torsion angles. This work addresses both issues to build a fast, transferable, and reliable generative backmapping tool for CG protein representations. We achieve generalization and reliability through a combined set of innovations: representation based on internal coordinates; an equivariant encoder/prior; a custom loss function that helps ensure local structure, global structure, and physical constraints; and expert curation of high-quality out-of-equilibrium protein data for training. Our results pave the way for out-of-the-box backmapping of coarse-grained simulations for arbitrary proteins.

1. Introduction

Protein dynamics ranges from large microsecond-scale movements of protein domains to small fast fluctuations of side chain atoms within protein pockets, and is connected to essential biological functions such as signaling, enzyme catalysis, and molecular machines (Salvatella, 2014). Despite the importance of dynamics and the large success of ML for prediction of protein structure, research on conformational ensembles started to accelerate only recently, mainly because data were scarce. Very flexible proteins (intrinsically disordered proteins, IDPs) or protein regions are better understood as conformational ensembles rather than static structures. Experimental structure determination methods observed either one frozen structure or the average of the conformational ensemble and thus are not suitable for describing individual dynamic states (Miller & Phillips Jr., 2021). Conformational ensembles are thus mainly generated using simulations such as Molecular Dynamics (MD) simulations or statistical sampling. Following the simulation, a representative subset can be selected from the pool of sampled conformers to match the properties and constraints derived from experimental measurements (Orellana, 2019; Salvatella, 2014).

Atomistic simulations are often too computationally expensive for the time and length scale of protein dynamics. An effective way to overcome these limitations is to use coarse-grained (CG) simulations with simplified particles. Representing systems in a reduced number of degrees of freedom provides access to much larger spatiotemporal scales (Kmiecik et al., 2016). However, the speedup comes at the cost of atomistic details, which are essential in determining protein biochemical functions. For example, identifying specific atom-level contacts at a protein-protein interaction surface or a ligand binding pocket is crucial to understand molecular recognition, signaling, or ligand binding (Badaczewska-Dawid et al., 2020). Thus, *backmapping*, or restoring all-atom structures from CG structures, can be required to get a complete picture of protein function, especially for drug and protein design practices (Śledź & Caflisch, 2018; Huang et al., 2016).

Current popular backmapping methods involve two steps: 1) the generation of initial structures based on a set of geometric rules (Lombardi et al., 2016), libraries of protein fragments (Heath et al., 2007), or random placements (Rzepiela et al., 2010), 2) the refinement of the generated structures by Monte Carlo relaxation or MD simulations. The second step is necessary because these rule-based sampling methods usually result in poor initial structures (Roel-Touris & Bonvin, 2020). However, the optimization step requires of an exhaustive computation and can be biased towards the choice of the scoring function and relaxation methods

[1]Computational and Systems Biology, MIT, Cambridge, MA, United States [2]Department of Material Science and Engineering, MIT, Cambridge, MA, United States. Correspondence to: Rafael Gómez-Bombarelli <rafagb@mit.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

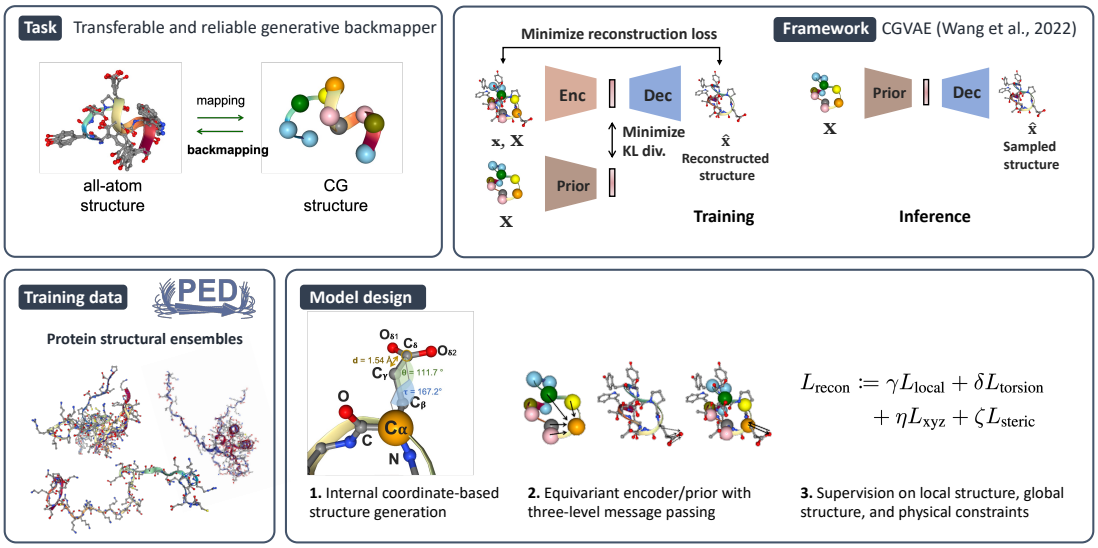

Figure 1. Overview. We aim to build a transferable and reliable backmapping tool for proteins. Our method builds on a VAE framework (Wang et al., 2022). We train the VAE model on the protein structural ensemble data curated from PED. Our model can be characterized with three components : internal coordinate-based representation, equivariant encoding, and physics-informed learning objectives.

(Badaczewska-Dawid et al., 2020).

Recently, data-driven methods have been proposed to achieve both efficiency and successful restoration of lost details through generative approaches. Li et al. (2020); Stieffenhofer et al. (2021); Wang et al. (2022); Shmilovich et al. (2022) that learn the distribution of all-atom conformers conditioned on the CG structures. While those methods show promising performances on simple systems like alanine dipeptide and mini-protein chignolin, most methods cannot generalize beyond the chemistry on which they are trained (Li et al., 2020; Wang et al., 2022; Shmilovich et al., 2022). Stieffenhofer et al. (2021) shows the possibility of chemical transferability by training the model on two small molecules and testing the model on a polymer whose monomers encompass each of the two small molecules. Still, no prior methods have been tested on structures that have high structural complexity and a wide range of flexibility as in large protein molecules.

Here, we propose a deep generative backmapping tool that has transferability across protein space. Specifically, our model reconstructs the protein all-atom structure from the alpha carbon of each amino acid. We build the model on the framework of Wang et al. (2022), where a Variational Auto-Encoder (VAE) model approximates the 3D spatial distribution of all-atom structures conditioned on CG structures. We achieve the transferability by training on structures from the Protein Ensemble Database (PED) (Lazar et al., 2021), which is a database of experimentally validated structural ensembles of IDPs and IDP-globular protein complexes. We hypothesize that a deep generative model, can learn the complex spatial interdependence of atoms and residues trained on a variety of geometries and chemical environments. We name our model **GenZProt**, as the model generates Z-matrix, a set of internal coordinates that defines a 3D molecular structure, for all-atom protein structures. GenZProt utilizes an equivariant encoder/prior that encodes residue-wise spatial information, and shows improved performance compared to its invariant counterpart and the ability to perform inference on arbitrary proteins outside the training dataset.

Naive rule- or ML-based backmapping strategies may fail to capture physical and chemical constraints, such as preserving the molecular connectivity of the all-atom representation, avoiding steric clashes, and reconstructing long-range interactions between side chains. GenZPort is constructed to preserve the topology by generating structures based on internal coordinates—bond length, bond angle, and torsion angle—instead of explicitly predicting Cartesian coordinates of atoms. Therefore, the training procedure relies on a loss function that optimizes local structure (bond length and bond angle), global structure (torsion angle and reconstruction in Cartesian space), as well as novel physical constraints (avoiding steric clashes). These design choices are proven to be crucial to achieve high-quality samples through ablation studies. We provide an overview of our method in Figure 1.

Our contributions can be summarized as follows:

- We propose the first data-driven generative backmapper that is transferable across the entire protein space. We achieve the transferability by training on compu-

tationally generated, experimentally validated diverse structural ensemble data.

- We propose a model design to achieve high-quality backmapping, relying on internal coordinates, an equivariant encoder, and loss functions that enforce physical constraints and preserve chemical connectivity.

2. Methods

2.1. Data

PED (Lazar et al., 2021) hosts 227 entries of protein structural ensembles, mostly computationally generated and experimentally constrained. Experimental validation reduces the potential bias introduce by errors in the sampling method, such as approximations in the force fields and thus provides better training statistics. From PED, we selected 84 proteins for training and four proteins for testing. **Appendix D** details the curation of training and testing set.

2.2. CG mapping scheme

We choose alpha Carbon (C_α) mapping for coarse-graining—every amino acid residue is represented as one bead centered at its C_α. C_α atoms are explicitly present in popular medium resolution coarse-grained models, such as CABS (Kolinski, 2004) or MARTINI (Monticelli et al., 2008). As a result, the majority of backmapping algorithms starts from the C_α trace level Badaczewska-Dawid et al. (2020).

2.3. Internal Coordinate-based Structure Generation

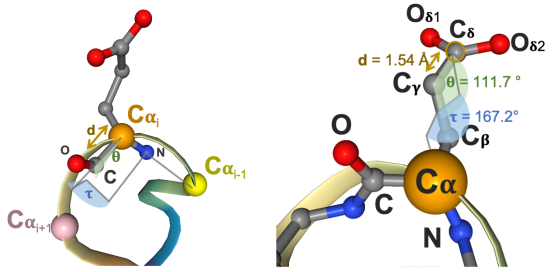

Figure 2. Internal coordinate-based reconstruction. **(a)** Backbone atoms N_i, C_i are placed using adjacent three C_α as anchors. **(b)** Side chain atoms are placed using adjacent three atoms within the same residue.

Relying on internal coordinates makes it easier to preserve the bond topology, since bond lengths and angles, which are very sensitive to small distortions can be kept within a physical range. However, correctly predicting atomic placements and interactions in 3D space is as important as preserving the topology (Lee et al., 2023). Instead of attempting to reconstruct Cartesian coordinates, GenZProt achieves faithful reconstruction of the bond topology by generating internal coordinate representation of each atom directly as model outputs.

GenZProt generates a set of internal coordinates (so-called Z-matrix), which is then converted to Cartesian coordinates through a rule-based algorithm. The placement of an atom A in 3D space can be determined from three anchor atoms B, C, D and a set of internal coordinates, bond length d_{AB}, bond angle θ_{ABC}, and torsion angle τ_{ABCD}, as shown in Figure 2.

Since the topology of a residue is fully determined from its amino acid type, we use a predefined set of anchor atoms per-residue. However, the choice of the predefined set of anchors for C_α trace-to-all-atom backmapping task is not trivial. We devise a hierarchical atomic placement algorithm, where the backbone atoms are placed using C_αs as anchors and the side chain atoms are placed sequentially.

In Lombardi et al. (2016) the authors postulate that the backbone peptide bond is perpendicular to the plane defined by three adjacent C_α atoms. Based on this assumption, we hypothesize that a machine learning model can learn to predict the placement of the backbone atoms of the i^{th} residue, N_i, C_i, relative to three adjacent C_α atoms, $C_{\alpha_{i-1}}, C_{\alpha_i}, C_{\alpha_{i+1}}$. Once we obtain the placement of C_{α_i}, N_i, C_i, we define three anchors within the i^{th} residue to place a remaining backbone atom O_i and side chain atoms. Atoms are then sequentially added to 3D space—for example, when the positions of C_{α_i}, N_i, C_i are known, C_{β_i} is placed from the anchors C_{α_i}, N_i, C_i, and with the C_{β_i} position known, C_{γ_i} is placed from anchors $C_{\beta_i}, C_{\alpha_i}, N_i$. We describe the transformation method in Figure 2.

Despite the sequential transformation, our model has a short inference time since our decoder generates all internal coordinates simultaneously in one shot. Refer to **Appendix C.2** for more details on Z-matrix to 3D coordinate conversion.

2.4. VAE framework

We build our model on the VAE framework introduced in (Wang et al., 2022). In this framework, stochastic backmapping is formulated as a modeling task of the distribution of all-atom structure x conditioned on CG structure X. The conditional distribution $p(x|X)$ is factorized as a latent variable model with a prior $P_\theta(z|X)$ and decoder $q_\psi(x|z, X)$, formulated as $p(x|X) = \int q_\psi(x|z, X)P_\theta(z|X)dz$. The encoder $p_\phi(z|x, X)$ is introduced to train the learnable prior and decoder.

During training, the CG latent variable z is sampled from encoder $p_\phi(z|\mathbf{x}, \mathbf{X})$ as $z = \mu_\phi + \sigma_\phi \circ \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(0, I)$. During sampling, given the coarse structure X, we sample the latent variable from the prior ($z \sim P_\theta(z|\mathbf{X})$). Latent

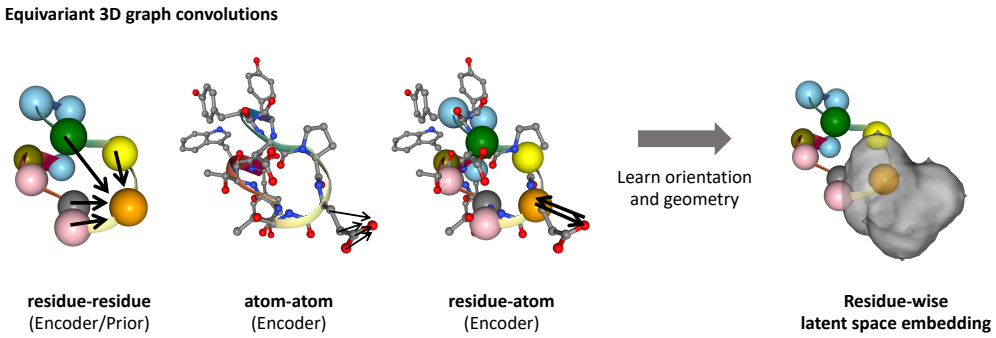

Figure 3. The three levels of equivariant 3D graph message passing operations in encoder and prior.

representation z is then passed to the decoder to generate the all-atom structure $\hat{\mathbf{x}}$.

2.5. Model Architecture

Equivariant encoder and prior. We introduce an equivariant encoder and prior architecture designed to learn the spatial interdependence of atom and residue placements. Since it is intuitive to model molecular structures as graphs, we perform message passing operations on graphs where residues and atoms are the nodes. The orientation and geometry of the residues surrounding an atom are crucial to determine their 3D placement. Thus, we use geometric tensors to represent the node attributes and use SE(3)-equivariant neural networks to perform message passing on the nodes. This equivariant message passing neural network module was implemented with the `e3nn` library (Geiger et al., 2022), mainly referring to the score model of DiffDock (Corso et al., 2022), which was used to predict docked poses of ligands in protein binding pockets.

We digitize the protein molecular graph by assigning residue and atom identity as initial node attributes. In our model design, the encoder performs message passing at three levels: atom-atom pair within the cutoff distance $9\mathring{A}$, atom-residue pair for every atom in a residue, and residue-residue pair within the cutoff distance $21\mathring{A}$. The three levels of graph convolution are illustrated in Figure 3. The prior performs message passing at residue level only.

Invariant decoder. We present a decoder that transforms residue-wise latent variables into internal coordinates necessary for atom placement within each residue. To accurately model the joint distribution of internal coordinates, it is ideal to allow flexibility within the physical range. Since bond lengths follow a constrained single-mode Gaussian distribution with small variance, we utilize a lookup table (implemented as a PyTorch `nn.Embedding`) based on residue type. We permit full flexibility in torsion angles.

While bond angles are correlated with torsion angles (Karplus & Kushick, 1981), we opt to enable full flexibility only for backbone bond angles and employ a lookup table for side chain bond angles. As the statistics of side chain angles can vary significantly depending on the computational sampling method used to generate the ensemble, to ensure the model learns from the correct statistics, we eliminate stochasticity in angles. Our backbone atom placement is based on adjacent C_α atoms, utilizing the angles $\theta_{NC_{\alpha_i}C_{\alpha_{i-1}}}$ and $\theta_{CC_{\alpha_i}C_{\alpha_{i+1}}}$. As these angles exhibit higher variance compared to side chain angles, we provide them with greater flexibility.

To predict flexible backbone bond angles and torsion angles, we utilize message passing and pooling operations on node-wise feature vectors, followed by Multi-Layer Perceptron (MLP) layers. More detailed discussion and an ablation study on the integration of side chain bond angle flexibility can be found in **Appendix C.1**.

Loss functions. The VAE model is trained to minimize the Evidential Lower Bound (ELBO) objective, which includes the reconstruction term to train the encoder and decoder and the Kullback–Leibler (KL) divergence term to minimize the difference between the prior and the encoder (Kingma & Welling, 2013), namely $L_{\text{ELBO}} := L_{\text{recon}} + \beta L_{\text{KL}}$.

To learn geometry and interactions at the atomic level while ensuring the validity of the generated structures, we supervise the model on both topology and atom placements in 3D space. Topology reconstruction is measured by a Mean-Squared-Error (MSE) loss term on bond lengths (L_{bond}) and a periodic angular loss term for angles (L_{angle}). We define L_{local} as a sum of L_{bond} and L_{angle} with $\epsilon = 10^{-7}$:

$$\underbrace{\frac{1}{|B|}\sum_{b\in B}(b-\hat{b})^2}_{L_{\text{bond}}} + \underbrace{\frac{1}{|A|}\sum_{\theta\in A}\sqrt{2(1-\cos(\theta-\hat{\theta}))+\epsilon}}_{L_{\text{angle}}},$$

where B is a set of all bonds, b and $\hat{b}$ are ground truth and

Table 1. **Ablation study on the model architecture**. **m1** : Our proposed model with equivariant encoder and invariant Z-matrix decoder. **m2** : Invariant encoder and Z-matrix decoder. **m3** : Equivariant encoder and Cartesian coordinate decoder. **m4** : Invariant encoder and Cartesian coordinate decoder. **m5** : **m1** trained with PED00151 only. **m6** : **m4** trained with PED00151 only.

	Method	PED00055	PED00090	PED00151	PED00218
RMSD (Å; ↓)	**m1 (GenZProt)**	**0.457±0.002**	**0.550±0.005**	**0.557±0.001**	**0.496±0.001**
	m2	0.578±0.004	0.787±0.002	0.648±0.005	0.565±0.003
	m3	2.432±0.035	2.475±0.026	2.798±0.011	2.393±0.043
	m4 (CGVAE)	2.244±0.001	2.355±0.002	2.901±0.040	2.241±0.004
	m5 (GenZProt, single)	-	-	0.832±0.001	-
	m6 (CGVAE, single)	-	-	2.072±0.000	-
GED (↓)	**m1 (GenZProt)**	**0.002±0.000**	**0.006±0.000**	**0.000±0.000**	**0.001±0.000**
	m2	0.007±0.000	0.017±0.000	0.005±0.000	0.003±0.000
	m3	0.349±0.035	0.431±0.010	0.405±0.002	0.339±0.008
	m4 (CGVAE)	0.246±0.002	0.382±0.004	0.308±0.003	0.208±0.002
	m5 (GenZProt, single)	-	-	0.084±0.001	-
	m6 (CGVAE, single)	-	-	0.140±0.000	-
Steric clash ratio (%; ↓)	**m1 (GenZProt)**	**0.140±0.003**	**0.142±0.002**	**0.211±0.008**	**0.190±0.003**
	m2	0.173±0.000	0.180±0.003	0.267±0.002	0.204±0.002
	m3	2.880±0.622	3.517±0.731	3.584±0.362	3.088±0.351
	m4 (CGVAE)	1.880±0.075	2.646±0.046	3.027±0.063	1.909±0.012
	m5 (GenZProt, single)	-	-	1.090±0.164	-
	m6 (CGVAE, single)	-	-	2.032±0.060	-

Table 2. **Ablation study on the reconstruction loss definition**. **m1** : Our proposed model with L_{recon} defined in Equation (3). **m7** : Trained without L_{torsion}. **m8** : Trained without L_{xyz}. **m8** : Trained without L_{steric}

	L_{recon}	PED00055	PED00090	PED00151	PED00218
RMSD (Å; ↓)	**m1 (GenZProt)**	**0.457±0.002**	**0.550±0.005**	**0.557±0.001**	**0.496±0.001**
	m7 ($-L_{\text{torsion}}$)	0.495±0.002	0.582±0.003	0.571±0.001	0.509±0.000
	m8 ($-L_{\text{xyz}}$)	1.910±0.251	1.905±0.136	2.025±0.337	1.754±0.198
	m9 ($-L_{\text{steric}}$)	0.467±0.005	0.573±0.013	0.570±0.005	0.524±0.003
GED (↓)	**m1 (GenZProt)**	0.002±0.000	0.006±0.000	**0.000±0.000**	**0.001±0.000**
	m7 ($-L_{\text{torsion}}$)	**0.001±0.000**	**0.004±0.000**	**0.000±0.000**	**0.001±0.000**
	m8 ($-L_{\text{xyz}}$)	0.046±0.000	0.057±0.001	0.026±0.000	0.033±0.000
	m9 ($-L_{\text{steric}}$)	0.002±0.000	0.006±0.000	0.003±0.000	**0.001±0.000**
Steric clash ratio (%; ↓)	**m1 (GenZProt)**	0.140±0.003	0.142±0.002	**0.211±0.008**	0.190±0.003
	m7 ($-L_{\text{torsion}}$)	**0.135±0.002**	**0.131±0.001**	0.236±0.013	0.181±0.003
	m8 ($-L_{\text{xyz}}$)	0.147±0.005	0.221±0.009	0.253±0.041	**0.144±0.007**
	m9 ($-L_{\text{steric}}$)	0.156±0.001	0.157±0.004	0.266±0.002	0.199±0.001

predicted bond length respectively. A is a set of all angles, θ and $\hat{\theta}$ are ground truth and predicted angle in radian. Further elaboration on the choice of the periodic angular loss term can be found in **Appendix F.1**.

Defining good reconstruction of atom placements in 3D space is not trivial for a backmapping task. A trivial solution for our internal coordinate-based generation setting would be a periodic angular loss term for torsion angles. However, a torsion angle can have a larger effect on the overall structure than other torsion angles. For example, a rotation near C_α would change the residue geometry more than a rotation at the end of the side chain. However, a simple regression would place an equal weight on every torsion angle. Thus, we additionally introduce a root-mean-squared distance (RMSD) loss term in Cartesian coordinate space:

$$
\begin{aligned}
L_{\text{torsion}} &:= \frac{1}{|T|} \sum_{\tau \in T} \sqrt{2 \times (1 - \cos(\tau - \hat{\tau})) + \epsilon} \\
L_{\text{xyz}} &:= \frac{1}{|N|} \sum_{\mathbf{x} \in N} ||\mathbf{x} - \hat{\mathbf{x}}||_2^2
\end{aligned}
\tag{1}
$$

where T is a set of all torsion angles, τ and $\hat{\tau}$ are ground truth and predicted torsion angle, respectively. N is a set of all atoms, $\mathbf{x}$ and $\hat{\mathbf{x}}$ are ground truth and predicted Cartesian coordinates of an atom, respectively.

To put further constraints on the chemical validity of the structures, we introduce steric clash loss, L_{steric}, as an auxil-

iary learning objective, defined as:

$$L_{\text{steric}} \coloneqq \sum_{\mathbf{x}\in\mathcal{N}} \sum_{\mathbf{y}\in\mathcal{B}_r(\mathbf{x})} \max(2.0 - ||\mathbf{x}-\mathbf{y}||_2^2, 0.0) \quad (2)$$

where $\mathcal{B}_r(\mathbf{x})$ is a set of atoms within the cutoff distance $r = 5.0$ Å with atom $\mathbf{x}$. Minimizing L_{steric} keeps the distance between any two nonbonded atom pairs larger than 2.0 Å.

The reconstruction term then becomes:

$$L_{\text{recon}} \coloneqq \gamma L_{\text{local}} + \delta L_{\text{torsion}} + \eta L_{\text{xyz}} + \zeta L_{\text{steric}}. \quad (3)$$

Hyperparameters $\gamma, \delta, \eta, \zeta$ are set to 1.0, 1.0, 1.0, 3.0, respectively. We explore different hyperparameter settings in our ablation study.

3. Experiments

In our experiments, we perform ablation studies on the model architecture and loss functions, and compare our model with the baseline, **CGVAE**. **CGVAE** was partially modified to take multiple proteins as training data. For each experiment, we perform five random seed experiments and report the mean and variance of the metrics. We refer the structures decoded from the encoder-sampled latent variables as *reconstructed* and the structures generated from the prior sampling as *sampled* structures.

3.1. Test Proteins

We test our model with four proteins of varying flexibility and compactness: **PED00055** (87 residues), **PED00090** (92 residues), **PED00151** (46 residues), **PED00218** (129 residues). **PED00055** and **PED00090** are mostly globular with short disordered tails, **PED00151** is an IDP, and **PED00218** is a complex of a globular protein and an IDP.

3.2. Metrics

We evaluate the model performance with three metrics: Root Mean Squared Distance (RMSD), Graph Edit Distance (GED), and Steric clash score.

Root Mean Squared Distance (RMSD). To evaluate the reconstruction, we report the $RMSD$ value of ground truth and reconstructed structures for each model.

Graph Edit Distance (GED). The sample quality is evaluated by measuring how well the generated geometries preserve the original chemical bond graph, which is quantified by the graph edit distance ratio $\lambda(G_{gen}, G_{true})$ between generated graph and the ground truth graph.

Steric clash score. In addition to GED, we report the ratio of steric clash occurrence in all atom-atom pairs within a 5.0 Å distance as a metric to measure the sample quality. For each atom-atom pair, distance smaller than 1.2 Å is considered a steric clash.

4. Results

4.1. Ablation Studies

Transferability and model architecture.

Table 1 shows how changing the model architecture affects the model performance (**m1-m6**). **m1-m4** are transferable models trained with 84 protein ensembles. **m5** and **m6** are single-chemistry models trained with **PED00151** alone.

Our proposed model with equivariant encoder/prior and a Z-matrix decoder, **m1**, shows the best performance for every metric. **m1** performs better than the model with an invariant encoder/prior (**m2**), implying the importance of the encoder/prior equivariance. Models with a Cartesian coordinate decoder (**m3, m4**) fail to give high-quality reconstructions for our large test proteins. As shown in Figure 4, reconstructions from **m3** and **m4** have many broken bonds and inaccurate topologies. Note that **m4** is equivalent to CGVAE, except that we modified its node definition to make it trainable for many proteins. We conclude that internal coordinate-based decoding coupled with equivariant encoder/prior can faithfully keep the topology while reconstructing high-quality structures with low RMSD and steric clash rates.

We also analyze the effect of training on a large protein dataset compared to training on a single protein structure. **m5** has a model architecture identical to **m1** (GenZProt) while **m6** is identical to **m4** (CGVAE), except that **m5** and **m6** are trained with **PED00151** structures only (284 frames). **m1** performs better than **m5**, even though the training set does not include **PED00151**. Such a result proves that a generalized model could be a better choice for a structure with few data points than a single-chemistry model. **m6** performs better than its transferable version but still performs worse than internal coordinate-based models. Figure 5 is the visualization of the reconstructed structures from **m1**, **m4**, **m5**, and **m6**.

Learning objectives. In Table 2, we evaluate the model performance as we change the learning objective. From **m7-m9**, the model architecture is identical to **m1**. To maintain topology and minimize steric clash, keeping the distance between nonbonded atom pairs was crucial, which was achieved through 3D coordinate-based losses, such as L_{xyz} and L_{steric}. Consequently, models **m8** and **m9** showed higher steric clash ratio compared to models **m1** and **m7**. Also, a higher RMSD observed in **m7** implies the importance of accurate torsion angle predictions in reconstructing precise 3D geometry.

4.2. Qualitative analysis

Generated structures. Figure 9 shows reconstructed structures and sampled structures from **m1** (GenZProt) for four

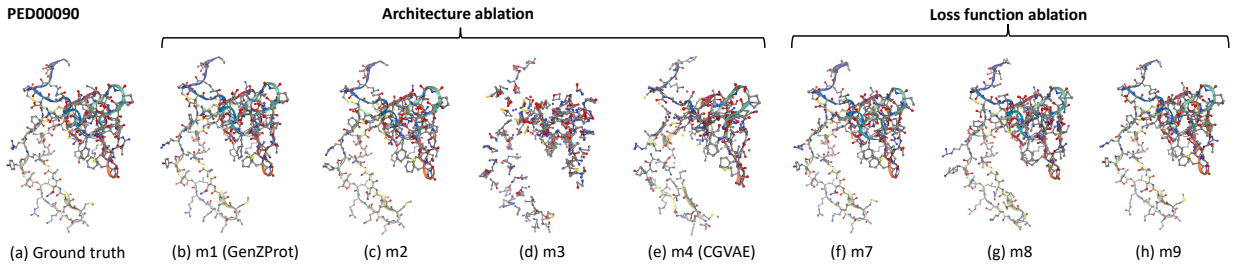

Figure 4. Reconstruction of PED00090 from **m1-m4**, **m7-m9**.

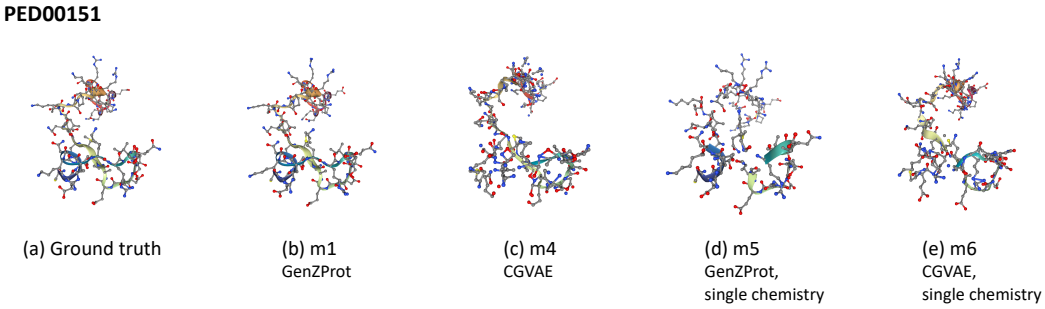

Figure 5. Reconstruction of PED00151 from transferable models **m1, m4** and single-chemistry models **m5, m6**.

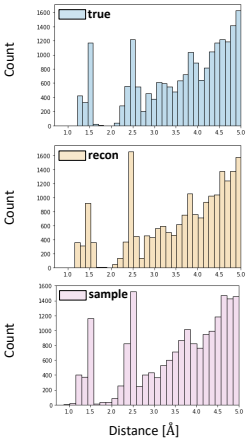

Figure 6. Atom-atom pairwise distances of ground truth, reconstructed, and sampled structures of PED00218.

test proteins. Both reconstructed and sampled structures recover the topology faithfully and do not show any notable steric clashes.

Atom-atom distance distribution. Figure 6 shows all atom-atom pairwise distances < 5 Å in the ground truth structures ("true"), reconstructed structures ("recon"), and sampled structures ("sample") of **PED00218**, generated

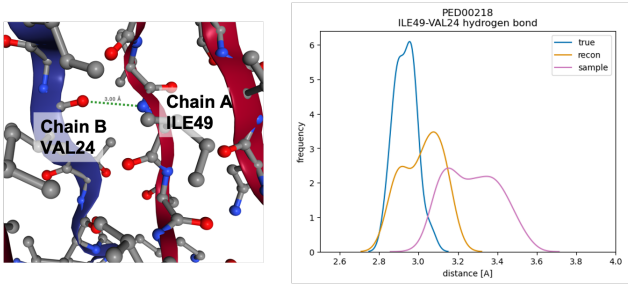

Figure 7. KDE plot of distance between N of ILE49 (chain A) and atom O of VAL24 (chain B) in **PED00218**

from **m1**. The ground truth distribution is well reconstructed by the encoder or the prior. 5 Å is the higher cutoff for attractive London-van der Waals interactions (Sengupta & Kundu, 2012). Encoder-generated reconstructions completely avoid steric clashes (< 1.2 Å), while prior-generated samples have few steric clashes. Atom-atom pairs with distance 3.3 Å $< d <$ 4.0 Å are likely hydrophobic interactions (van der Waals interactions), which is implied by a peak around 3.7 Å. Both reconstructed and sampled structures have a peak at 3.7 Å in a density similar to the ground truth, hinting that long-range interactions are preserved.

We further investigate one long-range interaction in

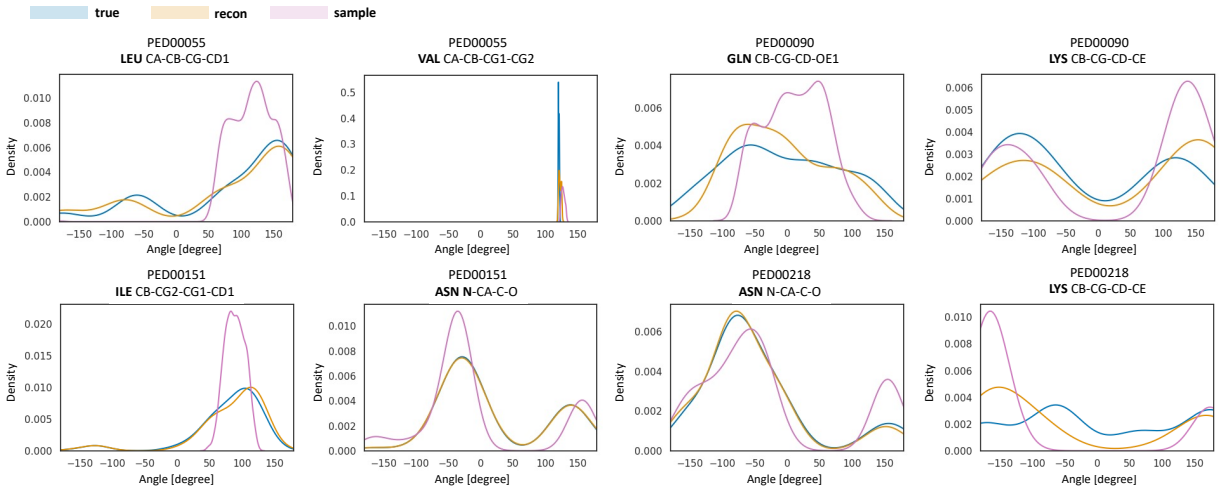

Figure 8. KDE plots of torsion angles from the structures generated from **m1**.

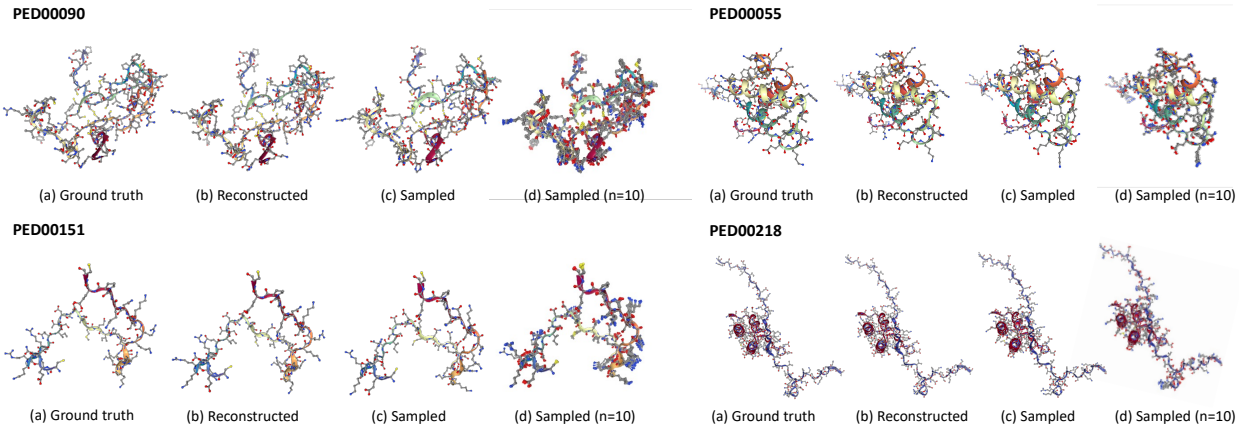

Figure 9. Structures reconstructed from **m1** for all four test proteins: PED00055, PED00090, PED00151, PED00218.

PED00218. **PED00218** is a peptide-protein complex, where a long IDP (chain B) is binding to a globular protein (chain A). The binding surface involves a hydrogen bond between a backbone nitrogen of the chain A ILE49 and a backbone oxygen of the chain B VAL24. Figure 7 is the a kernel density estimate (KDE) plot of a distance between these two interacting atoms. The length of hydrogen bonds typically ranges in 2.7 Å $< d <$ 3.3 Å (McRee, 2012). We can find reconstructed and sampled structures within the range of hydrogen bonds although the distributions are shifted to the right.

Torsion angle distribution. Figure 8 shows the torsion angle distribution of ground truth, reconstructed, and sampled structures. Both the encoder- and prior-generated structures recover ground truth distributions well. However, as shown in the KDE plot for LYS of **PED00218**, the prior sometimes fails to find all modes of the distribution. This learning problem might be an inherent problem with VAE since its learning objective, a reverse KL divergence, can be minimized even when the prior fits to only one mode. As a result, the learned prior distribution would not spread out to low probability regions (Murphy, 2012). We propose to apply a diffusion model on the latent space of GenZProt as future work. Latent space diffusion, or stable diffusion, has been recently highlighted for achieving an expressive prior while retaining the generation quality (Rombach et al., 2021).

4.3. Quantitative evaluation

Detailed quantitative evaluation of the sampled structures, including Earth Mover's Distances (EMD) between the ground truth and sampled torsion angle distributions and sample quality metrics such as GED and steric clash ratio,

can be found in **Appendix A.1**. We compare our model to the baseline CGVAE and demonstrate superior performance across all metrics. Furthermore, we compare our model to two non-ML deterministic baselines, CG2AA (Lombardi et al., 2016) and MODELLER (Webb & Sali, 2016) in **Appendix A.3**, demonstrating our model's well-balanced speed and reliability.

5. Conclusion

We introduce GenZProt, a transferable and reliable backmapper that can be used out-of-the-box for any arbitrary protein. We achieved chemical transferability by training on a protein conformational ensemble dataset curated from PED. Plus, we achieved reliability by employing physics-informed training objectives and devising an internal coordinate-based local structure construction method.

As our model seamlessly handles an arbitrary number of peptide chains, our model can be utilized to repack side chains of protein binding interfaces. We showed the potential of using our model for binding surface reconstruction by testing on the protein-peptide complex PED00218. Upon binding or complex formation, protein side chain conformations can significantly change, and accounting for side chain flexibility can substantially improve protein-protein docking (Gray et al., 2003).

Furthermore, in principle, our framework should be applicable to any family of polymers with a fixed number of building blocks. For future work, we propose applying our model to nucleic acids and nucleic acid-protein complexes.

Software and Data

Code and dataset for training and inference are available at `https://github.com/learningmatter-mit/GenZProt`.

Acknowledgements

We acknowledge the support from Novo Nordisk and Ilju Overseas Ph.D. scholarship. We thank Wujie Wang and Simon Axelrod for insightful discussion. We also thank Alexander Hoffman, Sihyun Yu, Akshay Subramanian, Yitong Tseo, and Lucia Vina Lopez for valuable feedback on the manuscript.

References

Badaczewska-Dawid, A. E., Kolinski, A., and Kmiecik, S. Computational reconstruction of atomistic protein structures from coarse-grained models. *Computational and Structural Biotechnology Journal*, 18:162–176, 2020. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2019.12.007. URL https://www.sciencedirect.com/science/article/pii/S2001037019305537.

Corso, G., Stärk, H., Jing, B., Barzilay, R., and Jaakkola, T. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2022.

Feldman, H. J. and Hogue, C. W. Probabilistic sampling of protein conformations: New hope for brute force? *Proteins: Structure, Function, and Bioinformatics*, 46(1):8–23, 2002. doi: https://doi.org/10.1002/prot.1163. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.1163.

Geiger, M., Smidt, T., M., A., Miller, B. K., Boomsma, W., Dice, B., Lapchevskyi, K., Weiler, M., Tyszkiewicz, M., Batzner, S., Madisetti, D., Uhrin, M., Frellsen, J., Jung, N., Sanborn, S., Wen, M., Rackers, J., Rød, M., and Bailey, M. Euclidean neural networks: e3nn, April 2022. URL https://doi.org/10.5281/zenodo.6459381.

Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A., and Baker, D. Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *Journal of Molecular Biology*, 331(1):281–299, 2003. ISSN 0022-2836. doi: https://doi.org/10.1016/S0022-2836(03)00670-3. URL https://www.sciencedirect.com/science/article/pii/S0022283603006703.

Gō, N. and Scheraga, H. A. On the use of classical statistical mechanics in the treatment of polymer chain conformation. *Macromolecules*, 9(4):535–542, 1976. doi: 10.1021/ma60052a001. URL https://doi.org/10.1021/ma60052a001.

Heath, A. P., Kavraki, L. E., and Clementi, C. From coarse-grain to all-atom: toward multiscale analysis of protein landscapes. *Proteins*, 68(3):646–661, August 2007.

Huang, P.-S., Boyken, S. E., and Baker, D. The coming of age of de novo protein design. *Nature*, 537(7620):320–327, Sep 2016. ISSN 1476-4687. doi: 10.1038/nature19946. URL https://doi.org/10.1038/nature19946.

Jing, B., Corso, G., Chang, J., Barzilay, R., and Jaakkola, T. Torsional diffusion for molecular conformer generation, 2022. URL https://arxiv.org/abs/2206.01729.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Karplus, M. and Kushick, J. N. Method for estimating the configurational entropy of macromolecules. *Macromolecules*, 14(2):325–332, 1981. doi: 10.1021/ma50003a019. URL https://doi.org/10.1021/ma50003a019.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes, 2013. URL https://arxiv.org/abs/1312.6114.

Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A. E., and Kolinski, A. Coarse-grained protein models and their applications. *Chemical Reviews*, 116(14):7898–7936, 2016. doi: 10.1021/acs.chemrev.6b00163. URL https://doi.org/10.1021/acs.chemrev.6b00163. PMID: 27333362.

Kolinski, A. Protein modeling and structure prediction with a reduced representation. *Acta Biochim. Pol.*, 51(2):349–371, 2004.

Lazar, T., Martínez-Pérez, E., Quaglia, F., Hatos, A., Chemes, L. B., Iserte, J. A., Méndez, N. A., Garrone, N. A., Saldaño, T. E., Marchetti, J., Rueda, A. J. V., Bernadó, P., Blackledge, M., Cordeiro, T. N., Fagerberg, E., Forman-Kay, J. D., Fornasari, M. S., Gibson, T. J., Gomes, G.-N. W., Gradinaru, C. C., Head-Gordon, T., Jensen, M. R., Lemke, E. A., Longhi, S., Marino-Buslje, C., Minervini, G., Mittag, T., Monzon, A. M., Pappu, R. V., Parisi, G., Ricard-Blum, S., Ruff, K. M., Salladini, E., Skepö, M., Svergun, D., Vallet, S. D., Varadi, M., Tompa, P., Tosatto, S. C. E., and Piovesan, D. PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Res.*, 49(D1):D404–D411, January 2021.

Lee, J. H., Yadollahpour, P., Watkins, A., Frey, N. C., Leaver-Fay, A., Ra, S., Cho, K., Gligorijević, V., Regev, A., and Bonneau, R. Equifold: Protein structure

prediction with a novel coarse-grained structure representation. *bioRxiv*, 2023. doi: 10.1101/2022.10.07.511322. URL https://www.biorxiv.org/content/early/2023/01/02/2022.10.07.511322.

Leung, H. T. A., Bignucolo, O., Aregger, R., Dames, S. A., Mazur, A., Bernèche, S., and Grzesiek, S. A rigorous and efficient method to reweight very large conformational ensembles using average experimental data and to determine their relative information content. *Journal of Chemical Theory and Computation*, 12(1):383–394, 2016. doi: 10.1021/acs.jctc.5b00759. URL https://doi.org/10.1021/acs.jctc.5b00759. PMID: 26632648.

Li, W., Burkhart, C., Polińska, P., Harmandaris, V., and Doxastakis, M. Backmapping coarse-grained macromolecules: An efficient and versatile machine learning approach. *The Journal of Chemical Physics*, 153 (4):041101, 2020. doi: 10.1063/5.0012320. URL https://doi.org/10.1063/5.0012320.

Lombardi, L. E., Martí, M. A., and Capece, L. CG2AA: backmapping protein coarse-grained structures. *Bioinformatics*, 32(8):1235–1237, April 2016.

MacKerell, A. D. J., Bashford, D., Bellott, M., Dunbrack, R. L. J., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., Joseph-McCarthy, D., Kuchnir, L., Kuczera, K., Lau, F. T. K., Mattos, C., Michnick, S., Ngo, T., Nguyen, D. T., Prodhom, B., Reiher, W. E., Roux, B., Schlenkrich, M., Smith, J. C., Stote, R., Straub, J., Watanabe, M., Wiórkiewicz-Kuczera, J., Yin, D., and Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The Journal of Physical Chemistry B*, 102(18):3586–3616, 1998. doi: 10.1021/jp973084f. URL https://doi.org/10.1021/jp973084f. PMID: 24889800.

McRee, D. E. *Practical protein crystallography*. Elsevier Science, 2012.

Miller, M. D. and Phillips Jr., G. N. Moving beyond static snapshots: Protein dynamics and the protein data bank. *Journal of Biological Chemistry*, 296, Jan 2021. ISSN 0021-9258. doi: 10.1016/j.jbc.2021.100749. URL https://doi.org/10.1016/j.jbc.2021.100749.

Monticelli, L., Kandasamy, S. K., Periole, X., Larson, R. G., Tieleman, D. P., and Marrink, S.-J. The martini coarse-grained force field: Extension to proteins. *Journal of Chemical Theory and Computation*, 4(5):819–834, 2008. doi: 10.1021/ct700324x. URL https://doi.org/10.1021/ct700324x. PMID: 26621095.

Murphy, K. P. *Machine learning: A probabilistic perspective*. The MIT Press, 2012.

Orellana, L. Large-scale conformational changes and protein function: Breaking the in silico barrier. *Frontiers in Molecular Biosciences*, 6, 2019. ISSN 2296-889X. doi: 10.3389/fmolb.2019.00117. URL https://www.frontiersin.org/articles/10.3389/fmolb.2019.00117.

Ozenne, V., Bauer, F., Salmon, L., Huang, J.-r., Jensen, M. R., Segard, S., Bernadó, P., Charavay, C., and Blackledge, M. Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics*, 28(11):1463–1470, 05 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/bts172. URL https://doi.org/10.1093/bioinformatics/bts172.

Pele, O. and Werman, M. Fast and robust earth mover's distances. In *2009 IEEE 12th International Conference on Computer Vision*, pp. 460–467. IEEE, September 2009.

Roel-Touris, J. and Bonvin, A. M. Coarse-grained (hybrid) integrative modeling of biomolecular interactions. *Computational and Structural Biotechnology Journal*, 18:1182–1190, 2020. ISSN 2001-0370. doi: https://doi.org/10.1016/j.csbj.2020.05.002. URL https://www.sciencedirect.com/science/article/pii/S2001037020302658.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models, 2021. URL https://arxiv.org/abs/2112.10752.

Rzepiela, A. J., Schäfer, L. V., Goga, N., Risselada, H. J., De Vries, A. H., and Marrink, S. J. Reconstruction of atomistic details from coarse-grained structures. *Journal of Computational Chemistry*, 31 (6):1333–1343, 2010. doi: https://doi.org/10.1002/jcc.21415. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.21415.

Salvatella, X. *Understanding Protein Dynamics Using Conformational Ensembles*, pp. 67–85. Springer International Publishing, Cham, 2014. ISBN 978-3-319-02970-2. doi: 10.1007/978-3-319-02970-2_3. URL https://doi.org/10.1007/978-3-319-02970-2_3.

Sengupta, D. and Kundu, S. Role of long- and short-range hydrophobic, hydrophilic and charged residues contact network in protein's structural organization. *BMC Bioinformatics*, 13(1):142, Jun 2012. ISSN 1471-2105. doi: 10.1186/1471-2105-13-142. URL https://doi.org/10.1186/1471-2105-13-142.

Shmilovich, K., Stieffenhofer, M., Charron, N. E., and Hoffmann, M. Temporally coherent backmapping

of molecular trajectories from coarse-grained to atomistic resolution. *The Journal of Physical Chemistry A*, 126(48):9124–9139, 2022. doi: 10.1021/acs.jpca.2c07716. URL https://doi.org/10.1021/acs.jpca.2c07716. PMID: 36417670.

Stieffenhofer, M., Bereau, T., and Wand, M. Adversarial reverse mapping of condensed-phase molecular structures: Chemical transferability. *APL Materials*, 9(3): 031107, 2021. doi: 10.1063/5.0039102. URL https://doi.org/10.1063/5.0039102.

Vaidehi, N. and Jain, A. Internal coordinate molecular dynamics: A foundation for multiscale dynamics. *The Journal of Physical Chemistry B*, 119(4):1233–1242, 2015. doi: 10.1021/jp509136y. URL https://doi.org/10.1021/jp509136y. PMID: 25517406.

Wang, W., Xu, M., Cai, C., Miller, B. K., Smidt, T. E., Wang, Y., Tang, J., and G'omez-Bombarelli, R. Generative coarse-graining of molecular conformations. In *International Conference on Machine Learning*, 2022.

Webb, B. and Sali, A. Comparative protein structure modeling using modeller. *Current Protocols in Bioinformatics*, 54(1):5.6.1–5.6.37, 2016. doi: https://doi.org/10.1002/cpbi.3. URL https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/cpbi.3.

Word, J. M., Lovell, S. C., Richardson, J. S., and Richardson, D. C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.*, 285(4):1735–1747, January 1999.

Śledź, P. and Caflisch, A. Protein structure-based drug design: from docking to molecular dynamics. *Current Opinion in Structural Biology*, 48:93–102, 2018. ISSN 0959-440X. doi: https://doi.org/10.1016/j.sbi.2017.10.010. URL https://www.sciencedirect.com/science/article/pii/S0959440X17301100. Folding and binding in silico, in vitro and in cellula • Proteins: An Evolutionary Perspective.

Šali, A. and Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *Journal of Molecular Biology*, 234(3):779–815, 1993. ISSN 0022-2836. doi: https://doi.org/10.1006/jmbi.1993.1626. URL https://www.sciencedirect.com/science/article/pii/S0022283683716268.

A. Additional Experimental Results

A.1. Sampled Structure Quality and Diversity

Here, we present an analysis of the metrics obtained from the sampled structures. We compare GED and steric clash ratio between the reconstructed structures that are generated using encoder-decoder framework and the sampled structures that are generated from a prior-decoder framework. We provide the results in Table 3. While the metrics of the sampled structures are worse than those from reconstructed structures, they still outperform the reconstructed structures generated by the baseline CGVAE (m4).

In addition, we report the diversity of the generated structures. The generation of diverse structures is crucial as it allows us to capture a wide range of conformers that are representative of the current conformational space. To assess the diversity of the generated structures, we employ $RMSD_{gen}$, a diversity metric introduced in the work of Wang et al. (2022). Higher values of $RMSD_{gen}$ indicate a greater diversity in the generated structures, reflecting the ability of our model to capture a more extensive range of conformations.

We also conducted a quantitative comparison of the torsion angle distributions among the ground truth structures, reconstructed structures, and sampled structures. To measure the distance between two distributions, we employ the Earth Mover's Distance (EMD) as a metric. In simple terms, EMD represents the minimum cost required to transform one histogram into another (Pele & Werman, 2009). Since torsion angles exhibit periodicity, we project the angles onto a unit circle $(\cos\theta, \sin\theta)$. Subsequently, we compute the 2D EMD of the transformed Euclidean coordinates using the Python package `PyEMD`, which implements the algorithm proposed in Pele & Werman (2009).

For a given random seed experiment, we calculate the EMD for each amino acid type and compute the average across the present amino acid types. Then, the mean and variance of the amino acid-averaged EMD values across five random seed experiments are computed and reported in Table 4. Our model (m1) consistently outperforms the baseline CGVAE in all cases. It is worth mentioning that when considering chi1 ($\chi 1$) angles, the EMD of the ground truth torsion angles and the torsion angles sampled from our model is significantly higher compared to that of reconstructed torsion angles, while CGVAE model (m4) exhibits similar EMD values for both reconstruction and sampling. This observation suggests that our model's prior is not capturing the full complexity of the encoder's learned latent space. We hypothesize that employing a more expressive prior, by employing hierarchical VAE or diffusion model, or acquiring a larger dataset, would be beneficial in addressing this limitation.

Table 3. Performance metrics of the structures generated from our model (m1). 'Recon' and 'Sample' stand for reconstructed structures and sampled structures, respectively. For the sampling process, ten conformers were generated per one C_α trace.

Metric		PED00055	PED00090	PED00151	PED00218
GED	recon	0.002 (0.000)	0.006 (0.000)	0.000 (0.000)	0.001 (0.000)
	sample	0.054 (0.000)	0.070 (0.000)	0.019 (0.000)	0.030 (0.000)
Steric clash ratio (%)	recon	0.140 (0.003)	0.142 (0.002)	0.211 (0.008)	0.190 (0.003)
	sample	0.473 (0.022)	0.682 (0.020)	0.637 (0.023)	0.456 (0.009)
$RMSD_{gen}$ (Å)	sample	1.871 (0.002)	0.029 (0.001)	1.711 (0.006)	1.727 (0.002)

A.2. Speed Analysis

Table 5 shows the average observed runtime of sampling of our model. Our model shows fast sampling speeds of approximately 0.009 seconds per frame when tested with batch size $= 8$. The sampling time can be proportionally reduced as we increase the batch size.

A.3. Non-ML Benchmarks

We compare our model to two non-machine learning benchmarks, namely **CG2AA** (Lombardi et al., 2016) and **MODELLER** (Webb & Sali, 2016; Šali & Blundell, 1993). CG2AA employs geometric reasoning to build backbone and side chain atoms on C_α traces. For instance, as mentioned in our main text, it assumes that the plane formed by three adjacent C_α atoms is perpendicular to the peptide bond, enabling the placement of backbone atoms. CG2AA does not involve energy minimization stage. On the other hand, MODELLER utilizes comparative protein structure modeling techniques to generate

Table 4. GenZProt refers to our m1 model and CGVAE refers to m4 model. The phi(ϕ) angles correspond to torsion angles for the N-C_α rotatable bond, the psi(ψ) angles represent torsion angles for the C_α-C rotatable bond, and the chi1($\chi 1$) angles denote the torsion angle for the C_α-C_β rotatable bond. The metric reported here is the Earth Mover's Distance (EMD) of the radial distribution. 'Recon' and 'Sample' stand for reconstructed structures and sampled structures, respectively.

PED00055	Torsion	Recon	Sample
GenZProt	Phi	0.017±0.000	0.065±0.000
	Psi	0.020±0.000	0.073±0.000
	Chi1	0.055±0.001	0.268±0.001
CGVAE	Phi	0.276±0.004	0.279±0.001
	Psi	0.196±0.002	0.194±0.001
	Chi1	0.437±0.033	0.424±0.026

PED00090	Torsion	Recon	Sample
GenZProt	Phi	0.031±0.000	0.089±0.000
	Psi	0.027±0.000	0.087±0.000
	Chi1	0.062±0.000	0.322±0.001
CGVAE	Phi	0.210±0.000	0.284±0.002
	Psi	0.164±0.001	0.199±0.000
	Chi1	0.375±0.011	0.433±0.014

PED00151	Torsion	Recon	Sample
GenZProt	Phi	0.011±0.000	0.034±0.000
	Psi	0.010±0.000	0.053±0.000
	Chi1	0.045±0.001	0.262±0.000
CGVAE	Phi	0.109±0.001	0.179±0.004
	Psi	0.143±0.002	0.146±0.001
	Chi1	0.361±0.003	0.414±0.011

PED00218	Torsion	Recon	Sample
GenZProt	Phi	0.017±0.000	0.076±0.001
	Psi	0.018±0.000	0.071±0.000
	Chi1	0.056±0.001	0.340±0.004
CGVAE	Phi	0.239±0.001	0.351±0.000
	Psi	0.208±0.000	0.286±0.001
	Chi1	0.372±0.015	0.383±0.006

Table 5. Approximate inference times of GenZProt.

protein	sequence length	runtime [sec/frame]
PED00055	87	0.006
PED00090	92	0.010
PED00151	46	0.006
PED00218	129	0.012

3D protein structures that satisfy a set of spatial restraints. In the context of backmapping, C_α positions are used as restraints to construct all-atom structures. MODELLER involves optimization in the 3D Cartesian space using conjugate gradients and CHARMM22 force field (MacKerell et al., 1998) molecular dynamics with simulated annealing. Here, we present a benchmark of our model against CG2AA and MODELLER using our PED00055 test dataset.

Table 6. Benchmark of the model performance with two non-ML methods. For MODELLER, we utilized two different options of the `AutoModel` object: slow MD option (`md_level=refine.slow`) for more refined structure generation and fast MD (`md_level=refine.very_fast`) for faster generation. The mean and variance values were calculated across 52 frames of the PED00055 dataset. The reported scores for our model were obtained from sampled structures generated by the GenZProt prior-decoder.

	runtime [sec/frame]	steric clash ratio [%]	GED	CG RMSD [Å]
GenZProt	0.006 (0.000)	0.473 (0.022)	0.054 (0.000)	0.000 (0.000)
CG2AA	0.016 (0.000)	1.642 (0.003)	-	0.656 (0.038)
MODELLER (slow MD)	4.100 (0.012)	0.000 (0.000)	0.008 (0.000)	0.696 (0.040)
MODELLER (fast MD)	1.806 (0.002)	0.000 (0.000)	0.018 (0.000)	0.381 (0.002)

Among all methods, our model shows the shortest runtime per frame. Considering that our model's runtime can be even shortened if we increase the batch size, it gives our model a potential to be used simulataneously with molecular dynamics simulation or other costly operations. Compared to MODELLER which uses energy optimization, our model is roughly 300 $\sim$ 700 times faster. Moreover, our model is end-to-end differentiable, so it can be jointly trained with other deep learning models for downstream tasks such as flexible protein-ligand docking.

The CG2AA method shows a poorly reconstructed topology that would necessitate additional computationally expensive relaxation, as evidenced by a high steric clash ratio. GED scores could not be measured for CG2AA, as the reconstructed protein structure had a much smaller number of atoms ($N = 566$) compared to the original structure ($N = 658$). On the

other hand, structures generated by MODELLER show zero steric clashes and yield better GED scores than our model. However, we observe a trade-off between the GED score and the CG RMSD score, which measures the RMSD between the ground truth and the modelled C_α tracess. This suggests that MODELLER could achieve lower GED scores by relaxing C_α structures, which is violating the geometrical consistency of the backmapping. Notably, CG2AA also does not preserve the C_α trace precisely, whereas our model strictly preserves the CG structure.

An important distinction between our model and CG2AA or MODELLER lies in the generation of all-atom conformer distribution for a given C_α trace. While CG2AA and MODELLER deterministically produce a single structure, our model generates multiple possible conformers. Additionally, our model possesses the advantage of being end-to-end differentiable, enabling its joint training with other deep learning models for various downstream tasks, including flexible protein-ligand docking.

B. Related Work

Our work builds on Wang et al. (2022), among other recent studies on generative models for backmapping. Wang et al. (2022) provides a principled probabilistic formulation of the backmapping problem and proposes CGVAE, a Variational Auto-Encoder (VAE) model that approximates the 3D spatial distribution of all-atom structures conditioned on CG structures. Compared to Wang et al. (2022), our work shows several significant advancements, including generalization to arbitrary proteins and faithful reconstruction of a protein's topology.

Our work can be connected to protein structure prediction tasks. AlphaFold2 (Jumper et al., 2021) showed that learning-based methods could give robust predictions for protein structures. However, AlphaFold2 is trained on crystallography-based structural data, and the model is limited to a single structure prediction.

To capture the ensemble of structures that characterizes a flexible biomolecule, one would need either new ML architectures trained on out-of-equilibrium data or MD simulations over large time scales. Our work explores both directions as we train our model on a database of IDP ensembles and test on backmapping tasks to assist CG MD simulation-based studies.

While not specifically designed for backmapping, generative models have been used for small molecule conformer generation tasks. Jing et al. (2022) connects to our work with its internal coordinate-based conformer generation framework, where bond lengths and angles are constrained, and torsion angles are predicted with a diffusion model. Note that we cannot directly use such models for protein backmapping tasks since a backmapper needs to be conditioned on CG structures. Also, small molecules have less complexity and fewer long-range interactions than macromolecules, meaning learning tasks for small molecules could be simpler.

C. Model Design

C.1. Flexibility of the Side Chain Bond Angles

Our model is designed following the chemical assumptions that bond lengths and angles are essentially constrained given the high energy cost associated with their distortion, while torsion angles are mostly responsible for conformational freedom in amino acid side chains (Gō & Scheraga, 1976). Crystallographic studies of protein structures validate the little variability in bond lengths and angles of amino acid side chains (Karplus & Kushick, 1981). Furthermore, some of the ensembles we trained on were generated from methods that enforce fixed values for bonds and angles, and those ensembles are likely to deviate from the natural angle distribution. Thus, we utilized a lookup table to store the mean value of each bond angle across the training dataset.

Indeed, given the significant correlation between bond angles and torsion angles, modeling the joint distribution of the angles and torsion angles would improve the model performance. We anticipate learning bond lengths and angles to be a manageable task, as it primarily involves a mostly harmonic potential with a Gaussian distribution. In order to offer comprehensive insights, we utilize the available data to train the model with side chain bond angle flexibility and present the empirical results here. We trained a model (m10) that predicts side chain bond angles using MLP layers unlike the original model (m1) that relied on a lookup table. Then, we assessed the RMSD, GED, and steric clash ratio values across five random seeds and four test proteins. m10 demonstrated reduced performance compared to m1, with changes from 0.515 Å to 0.554 Å for RMSD, 0.002 to 0.002 for GED, and 0.170% to 0.216% for steric clash ratio, respectively. However, the average bond angle prediction error was significantly decreased in m10 – m10 had an average bond angle error of 1.89° for PED00218, while m1 had an error of 11.7°. For PED00090, m10 and m1 had average bond angle errors of 1.529° and

13.099°, respectively.

Although using fewer restraints on bond angles led to an increase in unphysical geometries, our findings reveal a significant enhancement in angle prediction accuracy. However, additional data gathering may be necessary to ensure that the current training ensembles accurately represent the statistics of angles. We contend that allowing flexibility on the side chain bond angles would further improve our model's performance in the existence of a more extensive dataset with accurate angle statistics.

C.2. Molecular Geometry and Internal Coordinate System

One possible representation of the molecular geometry is to list Cartesian coordinates of each atom. However, bond length, bond angle, and torsion angle are a more natural representation of proteins than the Cartesian coordinates since a topology of a molecule does not change unless it goes through a chemical reaction. In addition, since bond length, bond angle, and torsion angle have different frequencies of degrees of freedom, it could be easier to manipulate geometry and perform a conformational search with internal coordinate representation (Vaidehi & Jain, 2015).

To fully specify molecular geometry with Cartesian coordinates, $3N$ values are needed for a system of N atoms (i.e., x, y, z for each atom). For internal coordinate-based representation, it is a convention to specify a molecular geometry with Z-matrix. Each line of the Z-matrix defines a position of an atom: i, atom type, $j, d_{ij}, k, \theta_{ijk}, l, \tau_{ijkl}$, where i is the index of the current atom whose position is being defined, and j, k, l are the indices of adjacent atoms whose positions are already defined. The positions of the atoms j, k, l are used as anchors to place the atom i. d, θ, and τ are distance, angle, and torsion angle, respectively. Thus, our decoder outputs three values per atom i, $d_{ij}, \theta_{ijk}, \tau_{ijkl}$, where indices j, k, l are predefined given the residue type. During training, a fully differentiable Algorithm 1 is used to convert the Z-matrix to Cartesian coordinates. Then, L_{xyz} and L_{steric} are computed from the reconstructed Cartesian coordinate.

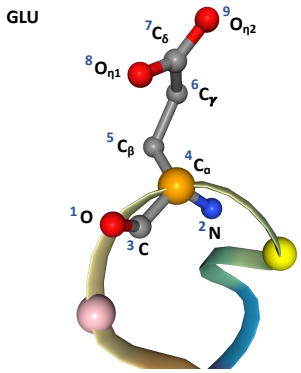

Figure 10. Structure of a glutamic acid.

Atoms in a residue are placed sequentially. For example, as shown in Figure 10, the beta carbon (C_β), $i = 5$, is constructed from atoms $j = 4, k = 3, l = 2$, which are the alpha carbon, C, and N, respectively. Similarly, the gamma carbon (C_γ, $i = 6$) is constructed from atoms $j = 5, k = 4, l = 3$, which correspond to the beta carbon, alpha carbon, and C, respectively.

However, adding atoms one by one will require N steps for a protein with N atoms, which will be extremely time-consuming. Thus, we reconstruct all residues at once in a parallel manner. For the ith step of the conversion, ith atoms of all residue are placed simultaneously. The order of the atoms is predefined (e.g., L =[O, N, C, CA, CB, CG, CD, OE1, OE2] for GLU). For any protein, 13 conversion steps are executed, as the maximum number of atoms in a residue is 13 except the already known C_α.

Algorithm 1 A pseudocode for the reconstruction of the list of Cartesian coordinates of side chain atoms, L, for a residue with m side chain atoms.

Input: $L = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4]$, # $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$, correspond to O, C, N, CA, respectively
for $i = 5$ **to** $m + 4$ **do**
  Input: row i of the Z-matrix $d_{ij}, \theta_{ijk}, \tau_{ijkl}$
  Let $j = i - 1, k = i - 2, l = i - 3$
  Compute $\mathbf{v}_{jk} \coloneqq L[j] - L[k]$
  Compute $\mathbf{v}_{kl} \coloneqq L[k] - L[l]$
  Compute $\mathbf{v} \coloneqq d_{ij}\mathbf{v}_{jk}/\|\mathbf{v}_{jk}\|_2^2$ # a vector of length d_{ij} pointing from j to k
  Compute $\mathbf{n} \coloneqq \mathbf{v}_{jk} \times \mathbf{v}_{kl}$ # a vector normal to the plane defined by j, k, l
  $\mathbf{v} \leftarrow \mathbf{R}(\theta_{ijk})\mathbf{v}$ # Rotate $\mathbf{v}$ around n by θ_{ijk}
  $\mathbf{v} \leftarrow \mathbf{R}(\tau_{ijkl})\mathbf{v}$ # Rotate $\mathbf{v}$ around $\mathbf{v}_{jk}$ by τ_{ijkl}
  $L[i] = \mathbf{v} + L[j]$ # Cartesian coordinate of ith atom
end for
Return L

D. Training and Test Dataset

D.1. Protein Conformational Ensembles

Proteins are built from up to 20 different amino acids in Table **??**. In a protein chain, amino acids are connected to their neighbors by peptide bonds: an amide group of an amino acid forms a peptide bond (CO-NH) with a carboxyl group of an adjacent amino acid. Peptide bonds and the alpha carbons together form a continuous chain of atoms called *backbone*. An individual amino acid connected to a peptide chain in a protein is called a *residue*. Each residue has a chemical group attached to the alpha carbon, called *side chain*.

Proteins do not exist in a static snapshot but rather exist in ensembles of conformations. The common procedure of ensemble calculation involves the generation of a starting pool of conformations using sampling programs such as Flexible-Meccano (FM) (Ozenne et al., 2012), TRaDES (Feldman & Hogue, 2002), or MD simulations. Then, a subset of conformers whose computed values fit the measurements from NMR or Small-Angle X-ray Scattering (SAXS) is selected as a representative structural ensemble. Each structure in a conformational ensemble is called a model or a frame.

Table 7. Amino acid abbreviation chart

Glycine	G, GLY	Proline	P, PRO
Alanine	A, ALA	Valine	V, VAL
Leucine	L, LEU	Isoleucine	I, ILE
Methionine	M, MET	Cysteine	C, CYS
Phenylalanine	F, PHE	Tyrosine	Y, TYR
Typtophan	W, TRP	Histidine	H, HIS
Lysine	K, LYS	Arginine	R, ARG
Glutamine	Q, GLN	Asparagine	N, ASN
Glutamic Acid	E, GLU	Aspartic Acid	D, ASP
Serine	S, SER	Threonine	T, THR

Our training and test data are from the protein structural ensemble database PED (Lazar et al., 2021). In this section, we discuss how we chose the entries for training and testing and provide analysis and statistics of the data.

We split the train and test set by protein entries (i.e., models never see the test protein entries during training). The validation set is identical to the test set, and the learning rate reduction and early stopping are controlled based on the validation loss.

D.2. Training Proteins

From 227 total entries of PED, we use 84 entries for training, four entries for validation, and four entries for testing.

The list of training entries are :

PED00003, PED00004, PED00006, PED00011, PED00013, PED00022, PED00024, PED00025, PED00032, PED00033,

PED00034, PED00036, PED00040, PED00041, PED00044, PED00045, PED00046, PED00050, PED00051, PED00052, PED00053, PED00054, PED00062, PED00072, PED00073, PED00074, PED00077, PED00078, PED00080, PED00085, PED00086, PED00087, PED00088, PED00092, PED00093, PED00094, PED00095, PED00097, PED00098, PED00099, PED00100, PED00101, PED00102, PED00104, PED00107, PED00109, PED00111, PED00112, PED00113, PED00114, PED00115, PED00117, PED00118, PED00120, PED00121, PED00123, PED00124, PED00125, PED00126, PED00132, PED00135, PED00141, PED00143, PED00145, PED00148, PED00150, PED00155, PED00156, PED00157, PED00158, PED00159, PED00160, PED00161, PED00175, PED00180, PED00181, PED00185, PED00190, PED00192, PED00193, PED00220, PED00217, PED00225, PED00227

The list of validation entries are : PED00175, PED00023, PED00043, PED00119

These proteins were excluded from the train and test set for the following reasons :

- Metal ion-binding complexes : PED00009, PED00026, PED00035, PED00037, PED00038, PED00039, PED00058, PED00059, PED00063, PED00068, PED00069, PED00106, PED00108, PED00110, PED00131, PED00134, PED00136
- Nucleotide-binding complexes : PED00057, PED00129, PED00130, PED00147
- Cofactor-binding complexes : PED00075, PED00089, PED00091, PED00133, PED00222
- PTM-including proteins except phosphorylation and oxidation : PED00014, PED00015, PED00047, PED00049, PED00064, PED00096, PED00127, PED00128
- D-amino acid protein : PED00103
- Proteins simulated or experimentally measured in unnatural conditions (e.g., denatured proteins, SDS or micelle containing solutions) : PED00060, PED00061, PED00065, PED00066, PED00067, PED00081, PED00116, PED00144, PED00146, PED00147, PED00149, PED00152, PED00205

We included proteins with phosphorylation and oxidation PTM since they much more frequently appear than the other PTMs.

Among 84 training entries, 23 entries were computed from the MD simulation. Sixty-one entries used sampling methods such as Flexible-Meccano, an all-atom structural optimization and sampling method for IDPs, based on amino acid-specific conformational potentials and volume exclusion (Ozenne et al., 2012).

D.3. Test Proteins

In this section, we introduce our four test proteins : PED00055, PED00090, PED00151, and PED00218. Structural ensemble **PED00055**, the N-terminal domain of DNA polymerase β, is sampled with an X-PLOR *ab initio* simulation and constrained with CHARMM parameters and NMR measurements. **PED00090** is a structural ensemble of the human chorionic gonadotropin alpha subunit sampled with X-PLOR and constrained with NMR measurements. **PED00151** is a structural ensemble of a Nuclear Localization Signal (NLS 99-140) peptide, sampled with MD simulation package CAMPARI and reweighted to match the experimental measurement from smFRET and SAXS. **PED00218** is a structure ensemble of a complex Taf14ET-Sth1EBMC, and its structures were derived from MD simulation and fit to NMR measurements. PED provides 55, 27, 29,598, 20 frames for entries PED00055, PED00090, PED00151, and PED00218, respectively. We use all frames for PED00055, PED00090, and PED00218 as testing set. For PED00151, we randomly sample 140 frames from the ensemble PED00151e000.

D.4. Single Chemistry Experiments

We perform the single chemistry experiments with entry PED00151. PED provides three ensembles for PED00151 : PED00151e000 (9,746 frames), PED00151e001 (9,924 frames), and PED00151e002 (9,928 frames). Each ensemble is reweighted with the COPER program (Leung et al., 2016) to match the experimental FRET efficiency and R_g values. To reduce the training time, we randomly sample 140, 142, and 142 samples from the ensemble PED00151e000, PED00151e001, and PED00151e002, respectively. We use PED00151e001 and PED00151e002 samples (284 frames) as the train and validation set. We randomly select 224 frames as training set and use the remaining 60 frames for validation. PED00151e000 (140 frames) is used as the test set.

D.5. Data Statistics

This section provides a quantitative analysis of the train and the test set. Our training set includes $\sim$ 10,000 frames, and the test set includes $\sim$ 500 frames. Our training and test proteins have 9,562 and 354 residues in total, respectively. In other words, the model has seen $\sim$ 10,000 different residue environments. The distribution of protein sequence length and the number of frames are shown in Figure 11 plot **(b)** and **(d)**. Figure 11 plot **(c)** shows the distribution of amino acid counts in all training entries. The amino acids are well distributed, except for tryptophans (TRP; W) and cysteines (CYS; C).

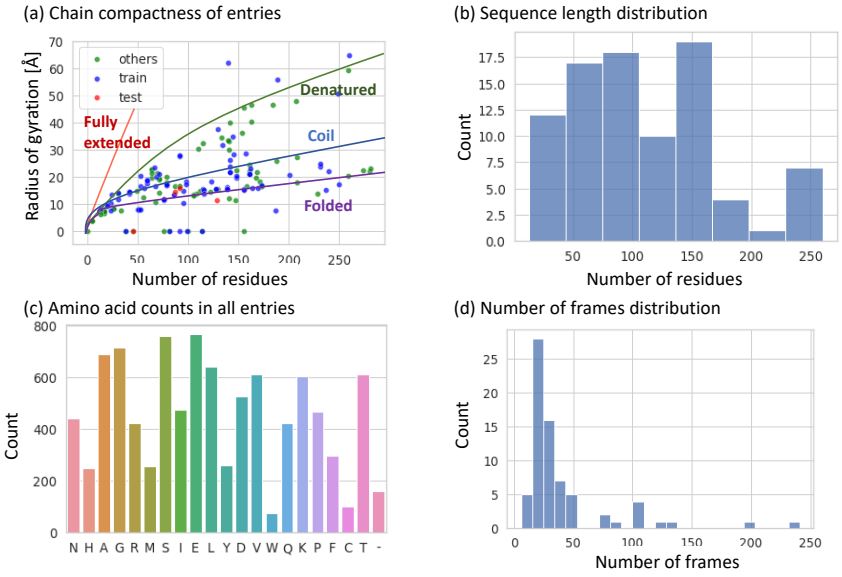

Figure 11. **(a)** Compactness plot. Train set, test set, and excluded entries are colored in blue, red, and green, respectively. Large proteins (number of residues $>$ 300) are omitted.

Our dataset includes proteins of various levels of compactness. Protein compactness can be characterized by the radius of gyration (R_g) as a function of the chain length (Lazar et al., 2021). Figure 11 plot **(a)** plots R_g of protein chains against the chain length. Each dot represents a chain in an entry. The trend lines in Figure 11 plot **(a)** are taken from Figure 2 of Lazar et al. (2021): completely flexible, rod-like chains follow a linear trend since the size of the protein will be proportional to the sequence length, folded proteins approximately follow a known scaling law, and disordered proteins fall in between. As shown in the plot, our test set does not include proteins with extreme disorderedness. However, since one of our test proteins, PED00151, is a disordered IDP with partial coils, we assume that testing on PED00151 would be enough to show the model performance on flexible proteins.

D.6. Data Preprocessing

Hydrogen removal. Since residues can be in many protonation states, we remove all the hydrogens from the train and test structures to reduce the number of building block representations. Moreover, in practice, protonation and hydrogen placement software such as REDUCE (Word et al., 1999) have been reliably used. Thus, we only consider heavy atoms for our reconstruction and sampling tasks.

Handling terminal residues and multiple chains. Since we reconstruct backbone nitrogens and carbons with three alpha carbons ($C_{\alpha_{i-1}}, C_{\alpha_i}, C_{\alpha_{i+1}}$ where $N > i \geq 1$) as anchors, we cannot reconstruct atomistic positions for terminal residues. Therefore, we mask the $i = 0$ and $i = N$ residues for training and inference. Also, when the entry is a protein complex with multiple chains, two terminal residues exist for each chain. In such cases, we mask all the terminal residues.

Handling PTMs. We treat phosphorylated Threonine (TPO) and phosphorylated Serine (SPO) as individual building blocks in addition to 20 canonical amino acids. We include proteins with oxidated residues (OXT) in our training and test sets. However, we do not treat oxidized residues as separate building blocks since oxidation appears in many amino acid types. Instead, we remove all the additional oxygen atoms added by oxidation PTM.

Sampling a subset from large entries. For the entries with a large number of frames (Number of frames > 500), we use the sampled subset of the entry to avoid the model overrepresenting those entries. We sample so that the number of frames per entry does not exceed 500. Following is the list of the large entries : PED00003, PED00006, PED00011, PED00022, PED00024, PED00025, PED00143, PED00145, PED00148, PED00150, PED00155, PED00180, PED00181.

E. Experimental Details

E.1. GenZProt

Our proposed model and the ablation models are trained with the hyperparameters defined in Table 8. Models were trained with Xeon-G6 GPU nodes until convergence, with a maximum runtime of 20 hours. Five random seeds—123, 321, 12345, 42, 24—were used. In our hyperparameter search, we prioritized minimizing steric clash by assiging a larger coefficient to the steric class loss term (ζ), while setting other coefficients to 1.0 without further consideration. We observed a trade-off between steric clash ratio and RMSD as a function of ζ: increasing the ζ value led to lower steric clash ratios and higher RMSD.

Table 8. A list of hyperparameters. **m1-m9** are defined in the main text.

Hyperparameter	m1-m6	m7	m8	m9
Node-wise latent variable dimension	36	36	36	36
Atom neighbor cutoff [Å]	9.0	9.0	9.0	9.0
Residue neighbor cutoff [Å]	21.0	21.0	21.0	21.0
Encoder convolution depth	3	3	3	3
Decoder convolution depth	4	4	4	4
Maximum training hours [hr]	20	20	20	20
Batch size	4	4	4	4
Learning rate	1e-3	1e-3	1e-3	1e-3
β coefficient for KL divergence	0.05	0.05	0.05	0.05
γ coefficient for L_{local}	1.0	1.0	1.0	1.0
δ coefficient for $L_{torsion}$	1.0	0.0	1.0	1.0
η coefficient for L_{xyz}	1.0	1.0	0.0	1.0
ζ coefficient for L_{steric}	3.0	3.0	3.0	0.0

E.2. Baseline - CGVAE

We modify the original version of CGVAE (Wang et al., 2022) to make it trainable for multiple chemical systems. Original CGVAE's encoder operates with atom-wise feature vectors, while GenZProt's encoder operates with residue-wise feature vectors. For a protein with N residues and n atoms, Original CGVAE's invariant encoder initializes n node attributes with the atom identity. Then, it performs message passing operations through atom-atom pairs within a cutoff distance and pools the atom-wise information to obtain a CG bead-wise latent variable. Unlike GenZProt, the CGVAE encoder does not perform CG bead-CG bead pair message passing. CGVAE prior operates at the CG level—the prior initializes node feature vectors with the index of the corresponding CG bead and performs CG bead-CG bead pair message passing operations. When the model is trained for a single chemistry, the index alone would have provided enough information for all-atom reconstruction. However, for a transferable model, we provide additional information by initializing the node feature vector with the residue identity. For the encoder, we concatenate the residue identity with the atom identity to initialize the atom-wise feature vector. For the prior, we use residue identity to initialize the feature vector.

E.3. Metrics

Root Mean Squared Distance (RMSD). The reconstruction task evaluates the model's capacity to encode and reconstruct given structures. We report the $RMSD$ value of ground truth and reconstructed structures for each model. The lower the $RMSD$, the closer the generated structure is to the ground truth structure.

Graph Edit Distance (GED). The sample quality is evaluated by measuring how well the generated geometries preserve the original chemical bond graph, which is quantified by the graph edit distance ratio $\lambda(G_{gen}, G_{true})$ between generated graph and the ground truth graph. G_{gen} is deduced from the coordinates by connecting bonds between pair-wise atoms where the distances are within a threshold defined by an atomic covalent radius cutoff used in (Wang et al., 2022). The lower

the λ, the better G_{gen} resembles G_{true}.

Steric clash score. In addition to GED, we report the ratio of steric clash occurrence in all atom-atom pairs within a 5.0 Å distance as a metric to measure the sample quality. For each atom-atom pair, distance smaller than 1.2 Å is considered a steric clash.

F. Learning Objectives

F.1. Periodic Angular Loss

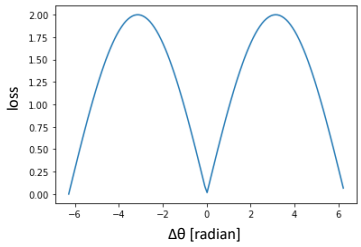

Figure 12. Periodic angular loss.

Periodic loss for angles introduced in Equation 2.5 is defined as:

$$L_{\text{angle}} = \frac{1}{|A|} \sum_{\theta \in A} \sqrt{2(1 - \cos(\theta - \hat{\theta})) + \epsilon} \tag{4}$$

.

This loss function is minimized at $\Delta\theta = \theta - \hat{\theta} = 0, 2\pi$, and maximized at $\Delta\theta = \pi, 1.5\pi]$. Figure 12 shows the angle loss term value as a function of $\Delta\theta$.

An alternative angular loss function would be Von Mises negative log-likelihood (NLL) loss, which can be computed as $-\cos\theta - \hat{\theta}$ at concentration parameter 1. Empirically, our proposed loss function showed better results – the model trained with NLL loss yielded an average RMSD of 0.781 Å and an average GED of 0.008 across five random seeds and four test proteins, while ours had an average RMSD of 0.515 Å and an average GED of 0.002. The observed discrepancy may be due to the different scaling between the angle loss and the NLL loss, as the NLL loss exhibits a sharper decrease in the loss value as $\Delta\theta$ decreases.

F.2. Interaction Score

We devised an interaction score to evaluate the model's ability to learn long-range interactions. Interactions were identified based on atom-atom pairwise distances, as the distance is the most determining variable of intermolecular interactions: force field terms such as Lennard-Jones potential or electrostatic energy are computed as a function of distance.

We tested adding the interaction score to our training objective, but the interaction score loss did not affect the model performance in reconstructing the long-range interactions. Thus, we introduce the score as a metric and not as a loss function.

Identification of the interactions. We considered two classes of interactions.

1. Hydrogen bonds, ion-ion interactions, dipole-dipole interactions : We identify heteroatom pairs within the distance of 3.3 Å.
2. Pi-pi stacking : We identify a pair of aromatic rings (PHE, TYR, TRP, HIS) that the distance between their ring centers is smaller than 5.5 Å.

The interaction score is defined as:

$$
\begin{aligned}
L_{\text{atom-pair}} &\coloneqq \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{A}} \max(||\mathbf{x}-\mathbf{y}||_2^2 - 4.0, 0.0) \\
L_{\text{pi-pair}} &\coloneqq \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{P}} \max(||\mathbf{x}-\mathbf{y}||_2^2 - 6.0, 0.0)
\end{aligned}
\tag{5}
$$

where $\mathcal{P}$ is a set of pair of atoms that are identified as type 1 interacting pair ($d_{xy} < 3.5$ Å), and $\mathcal{P}$ is a set of pair of aromatic rings that are identified as type 2 interacting pair ($d_{xy} < 5.5$ Å). The smaller the $L_{\text{atom-pair}}$ and $L_{\text{pi-pair}}$, the better the long-range interactions are reconstructed.

Here, we report the interaction scores tested from different model architectures.

Table 9. **Interaction scores**. **m1** : Our proposed model with equivariant encoder and invariant Z-matrix decoder. **m2** : Invariant encoder and Z-matrix decoder. **m3** : Equivariant encoder and Cartesian coordinate decoder. **m4** : Invariant encoder and Cartesian coordinate decoder. **m5** : **m1** trained with PED00151 only. **m6** : **m4** trained with PED00151 only.

	Method	PED00055	PED00090	PED00151	PED00218
Interaction score (↓)	**m1 (GenZProt)**	**0.025**±0.000	**0.069**±0.002	**0.057**±0.000	**1.270**±0.000
	m2	0.128±0.003	0.282±0.018	0.213±0.003	1.332±0.002
	m3	2.527±0.165	1.539±0.014	2.139±0.085	2.412±0.006
	m4 (CGVAE)	1.416±0.202	1.141±0.043	1.797±0.555	1.593±0.215
	m5 (GenZProt, single)	-	-	0.221±0.001	-
	m6 (CGVAE, single)	-	-	1.574±0.016	-