DiAMoNDBack: Diffusion-denoising Autoregressive Model for Non-Deterministic Backmapping of Cα Protein Traces

Michael S. Jones,[†,‡] Kirill Shmilovich,[†,‡] and Andrew L. Ferguson[*,†]

†*Pritzker School of Molecular Engineering, University of Chicago, Chicago, Illinois 60637, United States*

‡*These authors contributed equally to this work.*

E-mail: andrewferguson@uchicago.edu

Abstract

Coarse-grained molecular models of proteins permit access to length and time scales unattainable by all-atom models and the simulation of processes that occur on long-time scales such as aggregation and folding. The reduced resolution realizes computational accelerations but an atomistic representation can be vital for a complete understanding of mechanistic details. Backmapping is the process of restoring all-atom resolution to coarse-grained molecular models. In this work, we report DiAMoND-Back (Diffusion-denoising Autoregressive Model for Non-Deterministic Backmapping) as an autoregressive denoising diffusion probability model to restore all-atom details to coarse-grained protein representations retaining only Cα coordinates. The autoregressive generation process proceeds from the protein N-terminus to C-terminus in a residue-by-residue fashion conditioned on the Cα trace and previously backmapped backbone and side chain atoms within the local neighborhood. The local and autoregressive nature of our model makes it transferable between proteins. The stochastic nature of the denoising diffusion process means that the model generates a realistic ensemble of backbone and side chain all-atom configurations consistent with the coarse-grained Cα trace. We train DiAMoNDBack over 65k+ structures from Protein Data Bank (PDB) and validate it in applications to a hold-out PDB test set, intrinsically-disordered protein structures from the Protein Ensemble Database (PED), molecular dynamics simulations of fast-folding mini-proteins from DE Shaw Research, and coarse-grained simulation data. We achieve state-of-the-art reconstruction performance in terms of correct bond formation, avoidance of side chain clashes, and diversity of the generated side chain configurational states. We make DiAMoNDBack model publicly available as a free and open source Python package.

1 Introduction

Coarse-grained molecular models of proteins can substantially reduce the cost of molecular dynamics simulations and permit access to time and length scales and direct simulation of long-time processes such as folding, aggregation, and self-assembly that are inaccessible to all-atom molecular simulations.[1–9] Coarse-graining achieves these accelerations by selectively clumping groups of atoms into super-atoms, or coarse-grained beads and deriving an associated coarse-grained force field with which to propagate the system dynamics. Eliminating degrees of freedom in the coarse-grained representation leads to computational accelerations associated with the reduced cost of tracking fewer particles, the possibility of larger numerical integration time steps, and smoothing of the underlying free energy landscape that accelerates phase space exploration of the coarse-grained system.[10–12] Applications of coarse-graining to biomolecular dynamics have a rich history commencing with the pioneering work of Levitt and Warshel[13] and have led to a plethora of modern-day coarse-grained force fields such as MARTINI,[11,14,15] SPICA,[16] Rosetta,[17] PACE,[18] CABS,[19] AWSEM,[20] and Upside.[21] In recent years, there has been an explosion of interest in machine-learned coarse-grained potentials[22–32] that can be constructed from all-atom simulation data in a bottom-up fashion by rigorous variational techniques such as force matching[33,34] or relative entropy minimization.[35]

The primary concession of coarse-graining is a loss of atomistic detail that can be important in many applications such as determining atomistic contacts in protein-protein or protein-ligand interactions[36] or in downstream *ab initio* calculations that require atomistic detail to compute properties such as dipole moments or NMR spectra.[37] Backmapping is the process of reintroducing the lost degrees of freedom into a coarse-grained representation. This procedure can be conceived as a super-resolution task going from a coarse-grained to an atomistic geometry. The intrinsic loss of resolution in constructing a coarse-grained model means that the backmapping operation is one-to-many, and a primary challenge for backmapping algorithms is the generation of one or more physically realistic atomistic con-

figurations associated with a particular coarse-grained structure. Contemporary backmapping methods can typically be categorized into either rules-based or data-driven approaches. Rules-based approaches use heuristics to produce an initial guess for the atomistic structure that is then refined using geometry optimization and/or energy minimization. The initial structures can be generated by querying fragment libraries,[38–40] using random arrangements,[41] or geometrically-guided initialization.[42–47] Subsequent structural refinement and/or energy minimization is often necessary as the initial structures generated with rules-based approaches introduce unphysical artifacts such as atomistic clashes and/or anomalous bonds and dihedrals.[48] The requirement to manually adjust each backmapped structure introduces significant computational cost while also inherently biasing the final atomistic structure towards the particular choice of minimization scheme.[36] Rules-based approaches also tend to be deterministic in the sense that a particular coarse-grained structure will yield a single all-atom backmapped configuration. This can be an undesirable property since they fail to capture the thermodynamic ensemble of atomistic structures faithful to a single coarse-grained representation.

Data-driven techniques seek to remedy shortcomings of rules-based approaches by training neural networks to produce atomic structures conditioned on the coarse-grained representation.[49–55] These methods can achieve higher throughput compared to rules-based approaches as the models are trained to produce better well-equilibrated geometries that do not require a second stage of refinement or energy minimization. The most successful data-driven approaches tend to employ generative models, such as Variational AutoEncoders (VAEs)[56] and Generative Adversarial Networks (GANs),[57] that produce atomistic structures conditioned on the coarse-grained structure as a model input and can learn to produce a distribution of backmapped atomistic structures.[49–51,53–55] While many of these data-driven techniques have demonstrated good performance when applied to relatively small biomolecules such as alanine dipeptide and chignolin,[51,53,54] they typically require training of bespoke models using atomistic training data and are not transferable to other molecules

outside of the training data. A lack of transferability strongly limits the broader applicability of a backmapping model since training costs for one molecule cannot be amortized over other systems, and models cannot be developed for systems for which only coarse-grained trajectories are accessible and atomistic training data is either unavailable or insufficient to train a robust model. Work by Stieffenhofer et al. demonstrated potential for a transferable model by training on small molecule data and applying to polymer systems with their monomer units corresponding to the small molecules.[49,50] More recently, Yang and Gómez-Bombarelli present the first instance of a chemically transferable backmapping model designed to backmap Cα traces into full-resolution atomistic protein structures using a VAE architecture operating in the internal coordinate representation (dihedrals, angles, bond-lengths) of the protein.[55] The authors train on data from the Protein Ensemble Database (PED),[58] which largely represents intrinsically disordered (IDP) proteins, and held out four PED proteins for testing and evaluation.

In this work, we present a transferable backmapping model for proteins termed DiAMoNDBack (Diffusion-denoising Autoregressive Model for Non-Deterministic Backmapping) (Fig. 1). The model is based on the recently popularized class of generative Denoising Diffusion Probabilistic Model (DDPM).[59,60] DDPMs have demonstrated impressive performance within a number of molecular domains such as protein-ligand docking,[61,62] generation of molecular conformers,[63,64] learning of coarse-grained potentials,[31] and protein structure generation.[65–69] Our model is tasked to backmap atomistic proteins from Cα traces by autoregressively generating atomistic structures in a residue-by-residue fashion from the N-terminus to C-terminus of the chain conditioned on the Cα trace and any previously backmapped residues within the local neighborhood. The full protein structure is assembled by stitching together the backmapped residues along the coarse-grained Cα backbone. Importantly, the local and autoregressive nature of our model makes it transferable between proteins, and the stochastic nature of the denoising diffusion process means that the model generates an ensemble of backbone and side chain configurations consistent with the coarse-grained Cα

trace. This means that we can both amortize the training cost of the model by applying it to arbitrary proteins outside of the training data, apply it to coarse-grained simulation trajectories for which no accompanying all-atom training data exists, and generate multiple physically-consistent realizations of the backbone and side chain configurations.

We train DiAMoNDBack over 65k+ structures curated from the ProteinNet [70,71] database containing structures from the Protein Data Bank (PDB) [72,73] to construct a general-purpose generative model for backmapping protein structures from Cα traces. We validate our model in applications to a hold-out PDB test set, and demonstrate the transferability of the model in applications to intrinsically-disordered protein structures from the Protein Ensemble Database (PED), [58] molecular dynamics simulations of fast-folding mini-proteins from DE Shaw Research, [74] and coarse-grained simulation data generated from bespoke coarse-grained potentials by Majewski et al. [27] We benchmark against the PULCHRA rules-based approach developed by Rotkiewicz and Skolnick [47] and the data-driven VAE-based GenZProt model developed by Yang and Gómez-Bombarelli. [55] We achieve state-of-the-art reconstruction performance in terms of (i) correct formation of bonds, (ii) avoidance of side chain steric clashes, and (iii) diversity of the generated side chain configurational states. Contrary to the rules-based approach we do not suffer from a deterministic backmapping to a single structure, [47] and we better reproduce the natural distribution of side chain configurational states by avoiding the mode-collapse associated with the VAE-based approaches. [55] One drawback of the model is that the DDPM generation process is relatively slow, making it approximately 50$\times$ slower than GenZProt and 1000-3000$\times$ slower than PULCHRA when backmapping an individual sequence. However, the model still generates structures at a rate of approximately 1 residue per second and thousands of sequences or frames can be backmapped in parallel as long as GPU memory is not exceeded. We make DiAMoNDBack model publicly available to the community as a free and open source Python package (see Data Availability Statement).

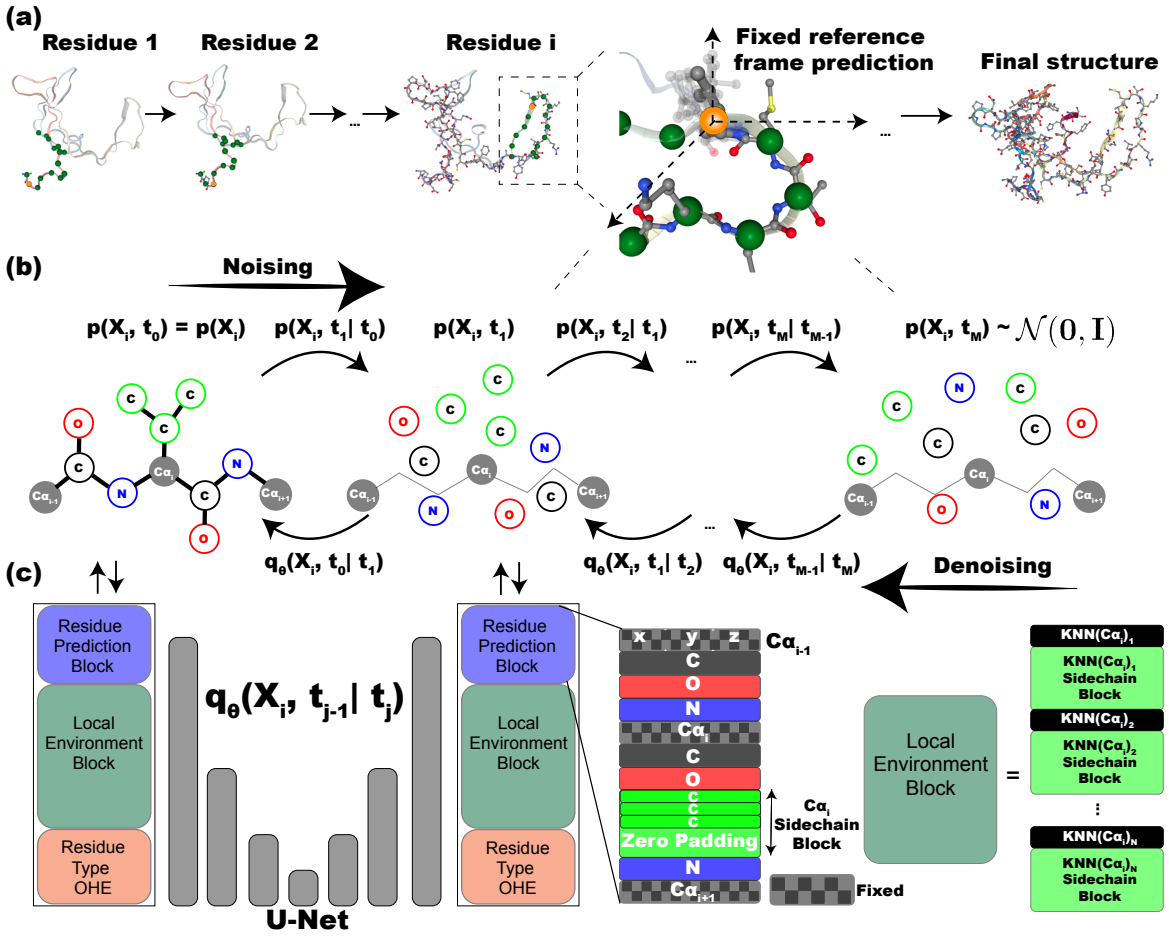

Figure 1: Schematic illustration of DiAMoNDBack (Diffusion-denoising Autoregressive Model for Non-Deterministic Backmapping). **(a)** Atomistic detail is restored from Cα traces by backmapping the structure residue-by-residue from the N-terminus to the C-terminus. Each residue prediction task is performed within a canonical reference frame and depends on the local environment around the target Cα (red sphere) comprising the N most spatially proximate residues to the target Cα (green spheres). **(b)** The residue prediction task is performed using a Denoising Diffusion Probabilistic Model (DDPM) that learns to reverse an M-step noising process applied to real samples. The trained DDPM model can then operate on random noise to recover realistic-looking samples. **(c)** Learning the denoising process involves training a U-net that is designed to predict the noise added to a corrupted sample. The input to the network are Cartesian coordinate representations of the target residue, the local environment, and a one-hot encoding of the residue identity. Conditioning is achieved by only noising and regressing on components of the representation that are allowed to change throughout the diffusion steps in the residue prediction block such as the backbone C, N and O atoms and the side chain atoms. Atoms comprising the partially-decoded local environment and the Cα backbone, along with the one-hot residue type encoding only serve as conditioning information passed to the network, are fixed throughout the diffusion steps, and do not contribute to the loss.

2 Methods

2.1 Diffusion-denoising Autoregressive Model for Non-Deterministic Backmapping (DiAMoNDBack)

Given a coarse-grained Cα trace of a protein $\mathbf{x} = [x_1, x_2, \ldots, x_{n-1}, x_n] \in \mathbb{R}^{n\times 3}$ with $x_i \in \mathbb{R}^{1\times 3}$ representing the spatial localization of each of the $i = 1 \ldots n$ Cα beads, the backmapping process seeks to learn the distribution of atomistic structures $p(\mathbf{X} \in \mathbb{R}^{N\times 3}|\mathbf{x})$ containing N atoms consistent with and conditioned on the coarse-grained structure $\mathbf{x}$. We frame this reconstruction task in an autoregressive formulation[49,50,75] where we generate the atomistic structure $\mathbf{X} = [X_1, X_2, \ldots, X_{n-1}, X_n]$ residue-by-residue, where $X_i \in \mathbb{R}^{p_i\times 3}$ represents the p_i atoms associated with residue i and $\sum_{i=1}^{n} p_i = N$. By decomposing the full distribution into a product of conditional distributions
$p(\mathbf{X}|\mathbf{x}) = p(X_1|\mathbf{x})p(X_2|\mathbf{x}, \{X_1\})p(X_3|\mathbf{x}, \{X_1, X_2\}) \ldots p(X_n|\mathbf{x}, \{X_1, X_2, \ldots, X_{n-2}, X_{n-1}\})$ we simplify the learning problem to backmapping residues $p(X_i|\mathbf{x}, \{X_1, X_2, \ldots, X_{i-2}, X_{i-1}\})$ in an autoregressive fashion rather than one-shot generation of the full protein structure (Fig. 1a). We solve this learning problem by training an autoregressive denoising diffusion probabilistic model (DDPM)[59,60,76–78] implemented within a conditional U-net architecture composed of 1D convolutional layers[79,80] (Fig. 1c). Conditioning for each step of the residue-by-residue autoregressive backmapping is based on the Cα trace, the N=14 most spatially proximate residues (i.e., have been brought into proximity by a secondary structural element, the tertiary fold, or quaternary complex, but which may be distantly separated in the backbone amino acid sequence), and a one-hot encoding of the residue type. Backmapping is performed by aligning each residue into a canonical alignment that permits us to directly generate the Cartesian coordinates of each backmapped atom in a rotationally and translationally invariant canonical reference frame that avoids the need for costly data augmentations otherwise required to implicitly learn insensitivity to rigid atomic rotations and translations.[54,81]

The DiAMoNDBack training protocol involves gathering an all-atom protein configuration from the training data, selecting a random residue index within the chain, selecting a random time step within the DDPM, sampling the commensurate degree of Gaussian noise to corrupt the sample, and training the U-net to predict the added noise and therefore learn to reverse the noising procedure (Fig. 1b). Once trained, the inference protocol generatively restores atomistic detail in a residue-wise fashion from the N to C-terminus of the Cα representation of the protein chain. Specifically, we pass through each amino acid in an N-to-C fashion using the trained U-net to transform random Gaussian noise into coordinates of the constituent atoms of the residue. These coordinates are then used to update the chain representation that is used to condition backmapping of subsequent residues. The backmapping procedure for each residue therefore comprises four steps: (i) alignment into the canonical reference frame, (ii) featurization to extract the conditioning variables, (iii) inference of the predicted atomic coordinates via the DDPM implemented within the trained U-net, and (iv) realignment of the decoded residue into the protein backbone and incorporation of the atomic coordinates into the updated chain representation. Due to the challenges in representing terminal residues within a canonical reference frame,[55] N- and C-terminal residues are handled separately after first backmapping all of the internal residues. Multi-chain proteins are backmapped in the same order in which they appear in the structure file using the same intrachain N-to-C ordering. Inter-chain residues are treated analogously to intra-chain residues when constructing the local environment conditioning. As in the case of single-chain proteins, N- and C-termini are backmapped once all internal residues have been placed in all chains.

Full details of our mathematical formalism, DDPM loss function, conditional U-net architecture, residue featurization, canonical alignment process, treatment of N- and C-terminal residues, hyperparameter tuning – including the choice of an N-to-C autoregressive ordering, use of a canonical Cartesian reference frame, and selection of N=14 most spatially proximate conditioning residues – are provided in the Supporting Information.

2.2 Data Curation

We collated four data sets for DiAMoNDBack training and testing: (i) PED – structural ensembles of primarily intrinsically disordered proteins,[58] (ii) PDB – structures drawn from the RCSB Protein Data Bank,[70–73] (iii) DES – D.E. Shaw Research molecular dynamics simulations of fast folding mini proteins,[74] and (iv) CG – coarse-grained simulations conducted by Majewski et al.[27] For each all-atom data set (i-iii) coarse-graining was performed by removing all atoms other than the Cα traces..

PED. For the purpose of comparison to the GenZProt model of Yang and Gómez-Bombarelli[55] we train over the Protein Ensemble Database (PED),[58] which contains structural ensembles of proteins including many intrinsically disordered proteins (IDPs). We discard three sequences – PED00125e000, PED00126e000, and PED00161e002 – that contain non-canonical amino acids, leaving us with 9228 structures comprising a total of 928,539 individual amino acid residue training samples. Following Yang and Gómez-Bombarelli,[55] we adopted four PED proteins – PED00151ecut0, PED00090e000, PED00055e000, and PED00218e000 – containing 20-140 frames and including one two-chain protein (PED00218e000) as our test set, and employed the remaining data as out training set. Since GenZProt does not support backmapping of terminal residues, to make head-to-head comparisons with this model we report all quantitative analyses restricted to internal residues only.

PDB. Our primary production-level model was trained over protein structures collated from the Protein Data Bank (PDB)[72,73] held in the SidechainNet[71] extension of ProteinNet[70] that itself builds on the data for the biennial Critical Assessment of protein Structure Prediction (CASP) challenges.[82] For this PDB training data set we retain a majority of configurations but filter according to a number of criteria. We discarded any configuration that had four or more disconnected chains or contained a chain less than five residues long, leaving 98,665/103,716 sequences. Next, we removed any structures that include incomplete side-chain coordinates for any non-terminal residues resulting in the elimination of an additional 32,403 structures. Finally, we eliminated 2,562 problematic structures containing one

or more malformed neighboring Cα-Cα distances lying outside the range of 2.7-4.1 Å, where our cutoffs were informed by collating histograms of neighboring Cα-Cα distance distributions to identify outliers (Fig. S1). This led us to retain a total 65,360 structures containing over 13M individual residue training samples (we note some structures were eliminated under multiple criteria and are not double-counted in the filtration). For the PDB test set, we employ the same test set as that provided by the ProteinNet database for the CASP12 blind structure prediction challenge.[83] We filter the test set consistent with the criterion we used to filter the training data set removing structures that were missing some portion of the side chain atoms. After data cleaning, we extracted 24 test set proteins ranging in size from 60-599 residues that includes eight multi-chain proteins.

DES. The PED and PDB training data comprise static protein structures derived predominantly from experimental structure determination. These training examples are expected to largely correspond to structures lying in local or global minima of the configurational free energy landscape. We were interested to test if the performance of our model would improve with additional fine tuning on all-atom molecular dynamics (MD) trajectories containing a greater diversity of configurations including metastable states and transition states. We refined the model trained over the PDB training data by subjecting it to additional training over MD trajectories of 11 fast-folding mini proteins conducted by D.E. Shaw Research (DES).[74] We aggregated one complete trajectory of each protein, eliminating villin (2F4K) that contains a non-canonical amino acid residue, and strided each trajectory into 10,000 equally spaced frames. Using the procedure described in Sidky et al.[84] we separated these frames into 100 contiguous chunks and randomly shuffled these chunks to form an 80/20 train/test split for each protein. In this way, the model is exposed to configurations across the full trajectory, but the test set retains regions that are temporally disjoint and distinct from the training data. To compile the fine-tuning data set we combined the training splits from each of the 11 proteins totaling 88,000 frames with an aggregate of 3,860,000 distinct residue training samples. When constructing our fine-tuning data set we find that

performance for terminal residue prediction can substantially improve by over-representing terminal residue training examples by repeating their occurrences in the training data set (Fig. S13). However, we find that there exists a trade-off where performance on internal residues begins to suffer if termini are too over-represented. For the fine-tuned models presented here, we employ a 5× augmentation of terminal residues, which we find to be a good balance in resolving terminal and internal residues with high fidelity.

CG. Finally, we collected Cα coarse-grained trajectories from the work of Majewski et al.[27] for three proteins of varying size: 1FME (28 residues), PRB (47 residues), and A3D (73 residues). In contrast to previous data sets, these are simulations carried out using a bespoke Cα-based coarse-grained force field. As such, there are no corresponding all-atom reference structures, so these data serve purely as testing data and a means to evaluate the out-of-domain generalization and transferability of our model. We performed even striding across all available 32 trajectories for each protein to obtain 2,000 frames for each system.

A visual comparison of the four data sets is presented in Fig. 2. In Fig. 2a we illustrate the distribution of sequence lengths. The PED training data comprises proteins of 13-260 residues in length with a total of 96 sequences and 9788 configurations. The DES data contains 11 sequences ranging from the small chignolin protein containing just 10 residues to the large λ-repressor containing 80 residues for a total of 88,000 structures. The PDB training data set contains the largest diversity of proteins of 5-2082 residues in length and comprises 65,360 structures. Alongside the sequence diversity, we represent the structural diversity of our training data sets by visualizing their distribution in the space of alpha-helical and beta-sheet content (Fig. 2b). The PED data tends toward relatively low alpha-helix and beta-sheet content, reflecting the intrinsically disordered nature of the data set. The PDB data spans a wide range of alpha-helix and beta-sheet content indicative of the more globular and ordered structures that originate from crystallographic data. The DES data, while containing the least sequence diversity, covers a wide range of structural diversity due to the sampling of both folded and unfolded configurations in the MD simulations. In

Fig. 2c, we illustrate the hold out test set proteins corresponding to each of the three training data sets in the space of their alpha-helix and beta-sheet content along with visualizations of selected structures. In Fig. S6 we present a residue-level comparison of the representation of the 20 natural amino acids within the PED, PDB, and DES training data.

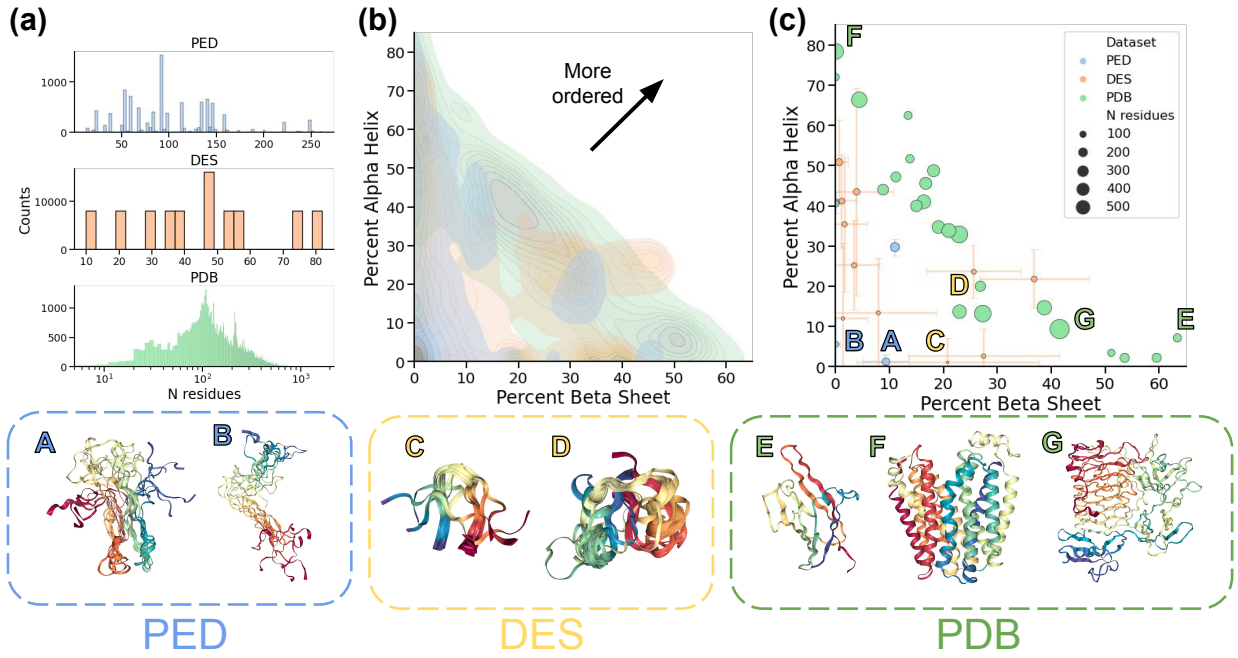

Figure 2: Visual analysis of protein sequences used for training and testing in the PED, PDB, and DES training sets. **(a)** Distribution of sequence lengths. **(b)** Distribution of structural diversity in the space of average percent alpha-helix and beta-sheet content. Disordered/unstructured structures reside close to the origin. Colors correspond to the histograms in panel (a): PED = blue, PDB = green, and DES = orange. Darker regions represent areas with higher probability density. **(c)** Test set proteins for each training data set shown in the same space of alpha-helix and beta-sheet content. The size of each point corresponds to the number of residues in the protein. For the PED and DES data, multiple frames are present for each protein and error bars indicate the standard deviation across frames. Selected representative training structures are labeled A-F and the corresponding structures rendered along the bottom of the figure.

2.3 Evaluation metrics

We define three metrics to evaluate the quality of our backmapping predictions that measure (1) the quality of the reconstructed atomic bonds, (2) the degree of non-bonded steric clashes

between residues, and (3) the diversity of atomistic structures produced.

Bond score (↑). The quality of the bond graph for our backmapped structures is determined by calculating the percentage of bonded atom distances that lie within 10% of the bond distance in the reference atomistic structure. For the coarse-grained trajectories in the CG test data that do not possess corresponding atomistic reference trajectories, we use the average values for bond lengths from the test set DES data for the associated protein as the reference. The bond score varies between 0-100% with higher values indicating better performance.

Clash score (↓). A major failure mode for backmapping is the the construction of side-chain placements that result in unphysical steric clashes. We quantify the degree of steric clashes by computing the fraction of residues in which one or more atoms lie within 1.2 Å of the atoms of another residue. For neighboring residues in the protein backbone, only clashes between the side chain atoms are considered. The 1.2 Å threshold was selected based on Yang and Gómez-Bombarelli[55] adopting this cutoff for defining atomic clashes. The clash score varies between 0-100% with lower values indicating better performance.

Diversity score (↓). A single coarse-grained Cα trace is consistent with multiple all-atom configurations. A desirable trait of a backmapping model is its capacity to generate an ensemble of all-atom predictions each of which is compatible with the Cα trace and itself contains well-formed bonds and avoids steric clashes. Previous work has used the root-mean squared distance (RMSD) between samples as a metric for generative diversity, where a high RMSD between samples indicates high diversity.[53] However, a low RMSD between the samples and the atomistic reference has also been used to show good adherence to conditioning and faithful reconstruction.[55] We combine these two desiderata to posit that the single all-atom reference structure should be indistinguishable from the distribution of generated configurations. As a corollary, the average pairwise RMSDs between (i) all generated samples and the reference and (ii) all generated samples with themselves should

be approximately equal. This leads us to define a generative diversity score DIV as,

$$RMSD_{ref} = \frac{1}{G}\sum_{i}^{G} RMSD(\mathbf{X}_i^{gen}, \mathbf{X}^{ref}) \tag{1}$$

$$RMSD_{gen} = \frac{2}{G(G-1)}\sum_{i}^{G}\sum_{j}^{(i-1)} RMSD(\mathbf{X}_i^{gen}, \mathbf{X}_j^{gen}) \tag{2}$$

$$DIV = 1 - \frac{RMSD_{gen}}{RMSD_{ref}} \tag{3}$$

where G is the number of generated samples $\{\mathbf{X}_1^{gen}, \ldots, \mathbf{X}_G^{gen}\}$ conditioned on the Cα trace possessing a single reference configuration $\mathbf{X}^{ref}$.

As shown in the Supporting Information, $RMSD_{gen} \approx \frac{2}{(G-1)\sqrt{G}}\sum_i^G RMSD(\mathbf{X}_i^{gen}, \overline{\mathbf{X}^{gen}})$, where $\overline{\mathbf{X}^{gen}} = \frac{1}{G}\sum_i^G \mathbf{X}_i^{gen}$ is the mean generated configuration and we assume the inequalities used in the derivation to be tight. Approximating $\frac{2}{(G-1)\sqrt{G}} \approx \frac{1}{G}$ allows us to interpret $RMSD_{gen}$ as the average RMSD of the generated configurations around their own mean. This construction is instructive as it allows us to then approximately interpret the diversity score DIV as a comparison of the average RMSD of the generated configurations around the reference configuration $\mathbf{X}^{ref}$ relative to that around their own mean $\overline{\mathbf{X}^{gen}}$. In general, one would anticipate a tighter distribution around the mean of a distribution than any other imposed point, such that we would expect $RMSD_{gen}$<$RMSD_{ref}$. This, however, is not a strict inequality and one may expect it to be violated, particularly for small values of G that produces noisy estimates of $\overline{\mathbf{X}^{gen}}$. Nevertheless, our empirical calculations show that in the present applications this inequality generally holds with only occasional violations, and that our calculated diversity scores approximately lie on the interval [0,1]. Diversity scores of unity are obtained for deterministic backmapping since $RMSD_{gen}$=0, while diversity scores close to zero are achieved for $RMSD_{ref}$≈$RMSD_{gen}$ and are indicative of better model performance. We note that although the diversity score provides a useful statistical measure of the generated ensemble, a score near zero does not guarantee that structures are physically plausible unless satisfactory bond and clash scores have also been achieved. Additionally,

we note that the diversity metric cannot be calculated for the coarse-grained data that lack reference atomistic structures, but we can still compute $RMSD_{gen}$ as a proxy metric for generative diversity.

3 Results and Discussion

We train and evaluate DiAMoNDBack on four train/test data sets and benchmark its performance against GenZProt as a state-of-the-art the VAE-based deep generative model[55] and the PULCHRA rules-based approach that generates an approximate structure using heuristics and atomic fragments followed by rotating side chain dihedrals to resolve steric clashes.[47] In the main results, we choose not to perform any energy minimization or molecular mechanics relaxations for any of the three models to compare and evaluate the backmapping approaches independent of any molecular force-field and to avoid the high computational cost associated with these operations for large proteins. However, we do conduct a limited analysis of the effect of energy minimization on model performance to illustrate what gains might be realized. We also note that while we provide routines for handling prediction of terminal residues, for the purposes of comparison to GenZProt that does not support termini prediction all analyses herein are performed on protein structures stripped of terminal residues. Details of our usage and application of PULCHRA and GenZProt and an analysis of the quality of DiAMoNDBack terminal residue backmapping is provided in the Supporting Information.

3.1 PED training and evaluation

Following Yang and Gómez-Bombarelli,[55] we first train DiAMoNDBack on the PED training data set comprising conformational ensembles of intrinsically disordered proteins.[58] A numerical comparison of DiAMoNDBack and GenZProt performance on the set of four held-out PED test examples is presented in Table 1. In terms of bond score, performance of

DiAMoNDBack and GenZProt are both excellent, lying above 95% correctly formed bonds in all cases and with essentially indistinguishable performance. In terms of clash score, DiAMoNDBack outperforms GenZProt with an overall improvement of ∼50% fewer clashing residues across the four structures and superior clash scores to GenZProt on all test proteins except PED00151ecut0 where our mean clash score is slightly better than GenZProt but not outside of error. In terms of diversity score, DiAMoNDBack shows a substantial improvement in generative diversity relative to GenZProt, achieving diversity scores of 0.23 or better on all four test examples while GenZProt scores are no better than 0.85. (We recall from Sec 2.3 that lower diversity scores are indicative of superior model performance – deterministic models with no conformational diversity possess diversity scores of unity, whereas models in which the average diversity between generated configurations matches that of the generated configurations with the reference have diversity scores of zero.) As discussed further in Sec. 3.3, the improved conformational diversity of DiAMoNDBack relative to GenZProt appears to be at least partially attributable to the VAE-based model suffering from mode collapse and failing to generate a high diversity of side chain dihedral angles.

3.2 PDB and DES training and evaluation

The PED data set is both relatively small and the intrinsically disordered character of the constituent proteins means that models trained on these data are not representative of globular and folded structures typically associated with functional proteins. We therefore trained our production-level DiAMoNDBack model, termed DiAMoNDBack (PDB), over 65k+ structures collated from the Protein Data Bank (PDB)[70–73] comprising 680× more sequences, 6.7× more configurations, and 15× more individual residue training examples than PED. The PDB structures predominantly reside in local or global minima of the configurational free energy landscape and may therefore underrepresent transient and metastable states. As such, we also fine-tuned our PDB trained model over the DES data set comprising long simulation trajectories of 11 fast-folding mini-proteins generated by D.E. Shaw

Table 1: Comparison of DiAMoNDBack and GenZProt trained over the PED training data and evaluated on the four held-out PED test examples. Bond scores enumerate the fraction of correctly formed atomistic bond lengths with higher values on the 0-100% range associated with superior performance. Clash scores enumerate the fraction of residues engaged in physically unrealistic steric clashes with lower values on the 0-100% range associated with superior performance. Diversity scores measure the configurational diversity of the generated atomistic configurations and approximately lie on a [0,1] range. Lower diversity scores are associated with superior performance: deterministic models with no configurational diversity possess a diversity score of unity, whereas models producing an average diversity between generated configurations matching that of the generated configurations with the reference have diversity scores of zero. Standard deviations in the reported values are estimated using five-fold block averaging for the bond and clash scores and using jackknife resampling for the diversity scores. The model exhibiting superior performance in any category outside of error bars is indicated in **bold**.

	Test protein			
	PED00055e000	PED00090e000	PED00151ecut0	PED00218e000
	Bond (↑) [%]	Bond (↑) [%]	Bond (↑) [%]	Bond (↑) [%]
GenZProt	97.55±0.02	95.30±0.02	**97.05±0.01**	98.06±0.01
DiAMoNDBack (PED)	97.70±0.08	**97.94±0.07**	96.76±0.05	98.07±0.10
	Clash (↓) [%]	Clash (↓) [%]	Clash (↓) [%]	Clash (↓) [%]
GenZProt	4.67±0.12	9.27±0.13	0.34±0.07	5.42±0.28
DiAMoNDBack (PED)	**2.82±0.18**	**5.04±0.68**	0.30±0.03	**1.76±0.38**
	Diversity (↓)	Diversity (↓)	Diversity (↓)	Diversity (↓)
GenZProt	0.9048±0.0004	0.888±0.002	0.8533±0.0007	0.862±0.001
DiAMoNDBack (PED)	**0.221±0.002**	**0.208±0.003**	**0.148±0.001**	**0.187±0.006**

Research[74] to produce our fine-tuned production-level model DiAMoNDBack (PDB;DES-FT). We hypothesized that the fine-tuned model should also be better calibrated to the DES force field and represent an example of developing a bespoke force field-specific variant of the baseline DiAMoNDBack (PDB) model using modest amounts of all-atom simulation data.

In Table 2 we present a comparison of the performance of PULCHRA,[47] GenZProt,[55] DiAMoNDBack (PDB), and DiAMoNDBack (PDB;DES-FT) on the 24 proteins comprising the held-out PDB test set and the held-out test split for the 11 DES all-atom simulation trajectories. We report the bond, clash, and diversity scores averaged over the test sets and also the best and worst performing systems as judged by the clash score to illustrate the performance range.

The bond scores averaged over the PDB and DES test sets are better than 94% for all four models. GenZProt slightly underperforms the other three models by 2-3 percentage points, but the difference is rather small.

The clash scores expose a more significant performance spread among the models. For GenZProt, 8.43% (PDB) and 6.01% (DES) of residues are positioned in unphysical steric clashes when averaged over the test sets. DiAMoNDBack (PDB) performs more than an order of magnitude better with clash scores of 0.57% (PDB) and 0.33% (DES), and the fine-tuned DiAMoNDBack (PDB;DES-FT) model is better still at 0.52% (PDB) and 0.18% (DES). The rules-based PULCHRA model is almost as good as the fine-tuned DiAMoND-Back on the DES data at 0.20% (DES) but is superior on the PDB at 0.15% (PDB). The superior performance of DiAMoNDBack (PDB;DES-FT) relative to DiAMoNDBack (PDB) on the DES test set is expected, and illustrates the value of fine-tuning a bespoke model for a particular force field. The small performance improvement on the PDB test data, or at least the absence of any performance degradation within error bars, was more surprising and suggests that the fine-tuned model is not overfitting to the DES data and maintaining a transferable and generic backmapping model. The extremely good clash score performance

Table 2: Comparison of backmapping performance on PDB data and MD trajectories between the rules-based PULCHRA approach,[47] GenZProt,[55] DiAMoNDBack trained on the PDB data set DiAMoNDBack (PDB), and DiAMoNDBack fine-tuned on MD trajectory data DiAMoNDBack (PDB;DES-FT). The "PDB overall" and "DES overall" columns report aggregate metrics averaged over all samples and frames. Two additional columns reports metrics on the best- and worst-performing systems as determined by clash score to give an appreciation for the performance range. Standard deviations in the reported values for DiAMoNDBack and GenZProt are estimated using five-fold block averaging for the bond and clash scores and using jackknife resampling for the diversity scores. As a deterministic algorithm, the values reported for PULCHRA do not have associated uncertainties and the diversity scores for this model are, by definition, unity. We note that overall metrics for PULCHRA on the PDB test set are only evaluated on the 16/24 single-chain proteins, as the software failed to operate on multi-chain systems. The model exhibiting superior performance in any category outside of error bars in the PDB overall and DES overall tasks is indicated in **bold**.

	PDB overall	PDB best (TBM#T0922)	PDB worst (TBM-hard#T0912)	DES overall	DES best (NTL9)	DES worst (UVF)
	Bond (↑) [%]	Bond (↑) [%]	Bond (↑) [%]	Bond (↑) [%]	Bond (↑) [%]	Bond (↑) [%]
PULCHRA	98.91	99.82	98.41	98.45	98.7	99.2
GenZProt	96.253±0.005	97.71±0.22	94.20±0.024	94.855±0.002	96.41±0.003	95.905±0.002
DiAMoNDBack (PDB)	**99.18±0.04**	99.78±0.18	98.97±0.09	97.981±0.002	98.24±0.01	97.22±0.01
DiAMoNDBack (PDB;DES-FT)	98.99±0.06	99.56±0.09	98.61±0.25	**98.725±0.004**	98.88±0.01	98.26±0.01
	Clash (↓) [%]	Clash (↓) [%]	Clash (↓) [%]	Clash (↓) [%]	Clash (↓) [%]	Clash (↓) [%]
PULCHRA	**0.15**	0	1.01	0.20	0.26	0.24
GenZProt	8.43±0.22	10.56±1.11	14.81±0.43	6.01±0.04	1.43±0.01	6.96±0.04
DiAMoNDBack (PDB)	0.57±0.09	0.00±0.00	1.07±0.44	0.33±0.01	0.12±0.02	0.58±0.04
DiAMoNDBack (PDB;DES-FT)	0.52±0.16	0.00±0.00	0.91±0.47	**0.175±0.007**	0.04±0.01	0.38±0.02
	Diversity (↓)	Diversity (↓)	Diversity (↓)	Diversity (↓)	Diversity (↓)	Diversity (↓)
PULCHRA	1	1	1	1	1	1
GenZProt	0.865±0.002	0.84±0.01	0.884±0.001	0.8316±0.0001	0.8758±0.0001	0.8962±0.0001
DiAMoNDBack (PDB)	**0.037±0.004**	0.16±0.03	-0.004±0.015	0.0329±0.0002	0.0706±0.0004	0.0230±0.0002
DiAMoNDBack (PDB;DES-FT)	0.064±0.004	0.09±0.03	0.076±0.008	**0.0223±0.0002**	0.0235±0.0002	0.0206±0.0005

of PULCHRA is unsurprising since this algorithm attempts to explicitly resolve steric clashes by rotating side chain dihedrals after placement of the residues. It is particularly encouraging, therefore, that DiAMoNDBack (PDB;DES-FT) is competitive with and/or superior to PULCHRA in this metric.

To explore the influence of energy minimization upon the clash score, we performed energy minimization of the PDB test set using the Generalized Amber Forcefield implemented in OpenBabel v2.3 employing a conjugate gradient minimization with a maximum of 1000 steps and the default convergence criteria of 10^{-6} (Figures S9-S10).[85] Before energy minimization, 59/120 DiAMoNDBack samples from the PDB dataset had a clash score above zero. After energy minimization, the clash score was reduced for all but three samples, the average clash score was reduced from 0.57% to 0.07%, and only 11/120 samples had a clash score greater than zero. For all proteins in the test set, clashes were completely eliminated from at least 3/5 generated samples. When performing energy minimization on the GenZProt generated samples, we found that, although the average clash score decreases from 8.18% to 2.43%, all generated samples had some clashes before energy minimization and 84/120 generated samples still had clashing residues afterwards. This shows that DiAMoNDBack produces superior backmapped structures that already possess very low clash scores but which are amenable to further refinement by energy minimization.

The diversity score of the deterministic rules-based PULCHRA model that generates a single backmapped configuration is, by definition, unity. GenZProt improves upon this slightly to achieve scores of 0.87 (PDB) and 0.83 (DES) but the proximity of these values to unity indicates that the preponderance of configurations are structurally very similar and the configurational diversity one would expect to be present within the ensemble of atomistic configurations consistent with the coarse-grained Cα trace is not well represented. In contrast, DiAMoNDBack (PDB) – 0.037 (PDB) and 0.033 (DES) – and DiAMoNDBack (PDB;DES-FT) – 0.064 (PDB) and 0.022 (DES) – achieve diversity scores very close to the ideal value of zero, indicating that the distribution of configurational diversity between the

backmapped atomistic configurations matches that of the generated configurations around the reference ground truth. The high bond scores and low clash scores for the DiAMoND-Back models indicate that despite the high configurational diversity, all of these various configurations are physically realistic with well-formed bonds and few steric collisions. The small performance boost in the DES test set using the DES fine-tuned model is indicative of a slight improvement in diversity resulting from the additional within-sample training, but the effect is almost negligible. Similarly, the small, but nearly negligible, performance degradation on the PDB test set indicates that the DES fine-tuned model is not overfitting.

We illustrate the diversity of generated all-atom configurations in Fig. 3 where we visualize five randomly generated backmappings for the best and worst-performing PDB and DES test set sequences reported in Table 2. As anticipated, we tend to see greater diversity of side chain configurations on the solvent-exposed exterior of the protein relative to those in the more tightly-packed hydrophobic core (Fig. S11-S12). Interestingly, the N-to-C ordering of autoregressive decoding does not result in any detectable trends in the bond, clash, or diversity scores as a function of residue position in the protein chain (Fig. S8). This suggests that the model has been well trained and achieves equally good backmapping quality irrespective of primary structure (i.e., sequence position) or location within the tertiary fold.

In Table S1 we provide mean RMSD values between the generated and reference configurations for the PDB test set. The one-to-many nature of coarse-grained to all-atom backmapping means that the RMSD values alone do not reflect the diversity of equally valid backmapped all-atom configurations consistent with a particular coarse-grained structure. Indeed, a very low RMSD score may be indicative of a pathology in the model to have memorized the training data and a failure to generate the full ensemble of physically valid all-atom configurations consistent with a particular Cα trace. With this caveat in mind, we do observe better RMSD values for both DiAMoNDBack models compared to GenZProt and PULCHRA, indicating a more faithful reproduction of the reference structures. Taken together with the high diversity of DiAMoNDBack generated structures (Table 2, Figure

3), this lends further confidence that the model is producing physically plausible all-atom configurations.

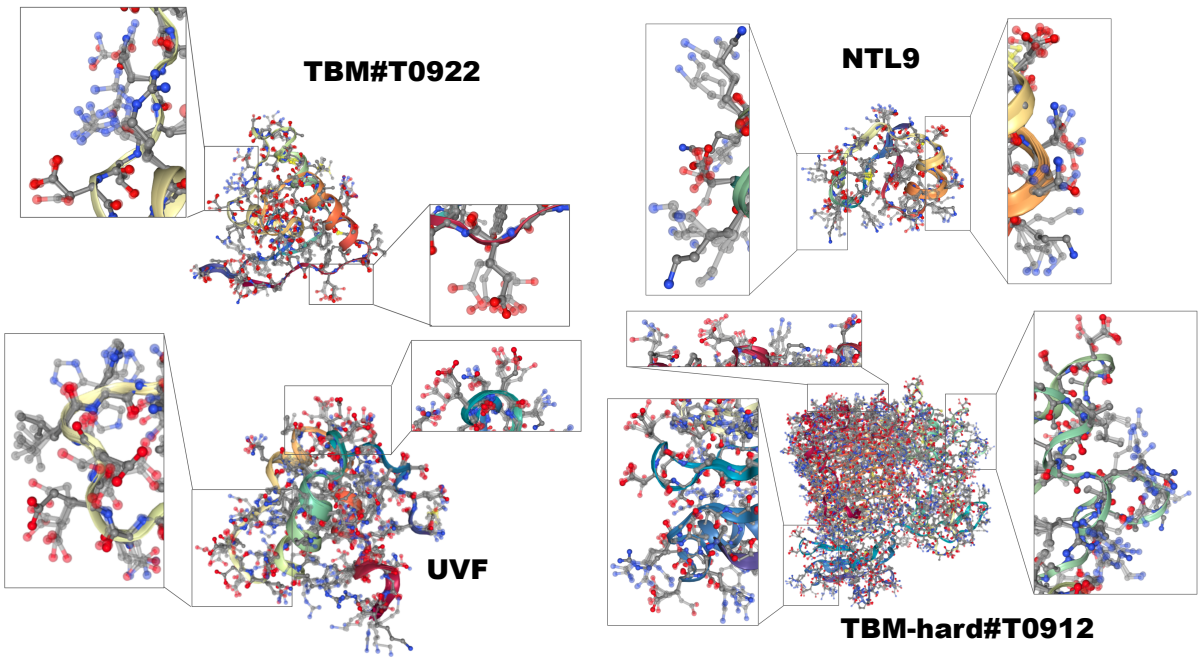

Figure 3: Illustrative visualization of five DiAMoNDBack (PDB) atomistic backmappings for the best and worst performing examples from the PDB test set (best: TBM#T0922; worst: TBM-hard#T0912) and the DES test set (best: NTL9; worst: UVF) according to clash score. The five backmapped structures conditioned on the coarse-grained Cα trace are shown as translucent structures and the single atomistic reference structure shown as opaque. Insets for each protein zoom into particular regions to highlight the generative diversity in side chain placements.

To further explore the impact of DES fine-tuning on the model performance we expose in Fig. 4a the distribution of bond and clash scores over five independently generated backmappings over the test set of 11 proteins in the DES data for the DiAMoNDBack (PDB) and DiAMoNDBack (PDB;DES-FT) models. The small but statistically significant improvement in the bond score upon fine tuning from (97.981±0.002)% to (98.725±0.004)% (Table 2) is visually apparent from the shift in probability mass in the violin plots towards 100%. The clash score improvement from (0.33±0.01)% to (0.175±0.007)% (Table 2) is also statistically significant but less visually apparent in a shift in the per-sequence distributions due to the large population of frames with zero clashing residues. An average of ∼74% frames across

all sequences are generated with no clashes in all five samples for the PDB-trained model, which improves to ∼84% frames with no clashes for the fine-tuned model. In Fig. 4b, we illustrate the improvement in model performance on each of the 11 proteins the DES test set by projecting the bond and clash scores of DiAMoNDBack (PDB) and DiAMoNDBack (PDB;DES-FT) into the plane. The migration of all points towards the upper-left of the plot after fine tuning indicates an across-the-board improvement in the bond and clash scores. An analogous analysis for the 24 proteins in the PDB test in Fig. 4c set shows a slightly different trend – changes in bond and clash scores are mixed on the PDB test set after fine tuning on the DES training data, with a percent improvement of 8.77% in clash score from (0.57±0.09)% to (0.52±0.16)% and a percent degradation of only 0.03% in bond score from (99.18±0.04)% to (98.99±0.06)% (Table 2). These minor changes indicate that the fine-tuned model is not strongly overfit and the change in clash score actually lies within error bars.

Finally, comparing the performance of the baseline DiAMoNDBack (PDB) model, we observe generally better bond scores and poorer clash scores for the PDB test set compared to the DES test set. We can attribute the superior PDB bond scores to improved in-sample performance of the model fitted to the PDB training data. The inferior clash scores are more difficult to account for, but we suggest that they are likely a result of the PDB structures being much larger and more globular, while the MD simulation data represents smaller sequences with many frames each that undergo many folding transitions and spend time in more extended configurations that are less susceptible to clashes.

Taken together, this analysis shows that the DiAMoNDBack model trained over the PDB training data is capable of achieving competitive or superior accuracy in bond and clash scores to GenZProt and PULCHRA while also recapitulating a diverse ensemble of atomistic structures faithful to a particular Cα coarse-graining. Fine tuning the model over the DES training data results in a slightly improved model for the DES test set without significant performance degradation over the PDB test set, and suggests a route to bespoke

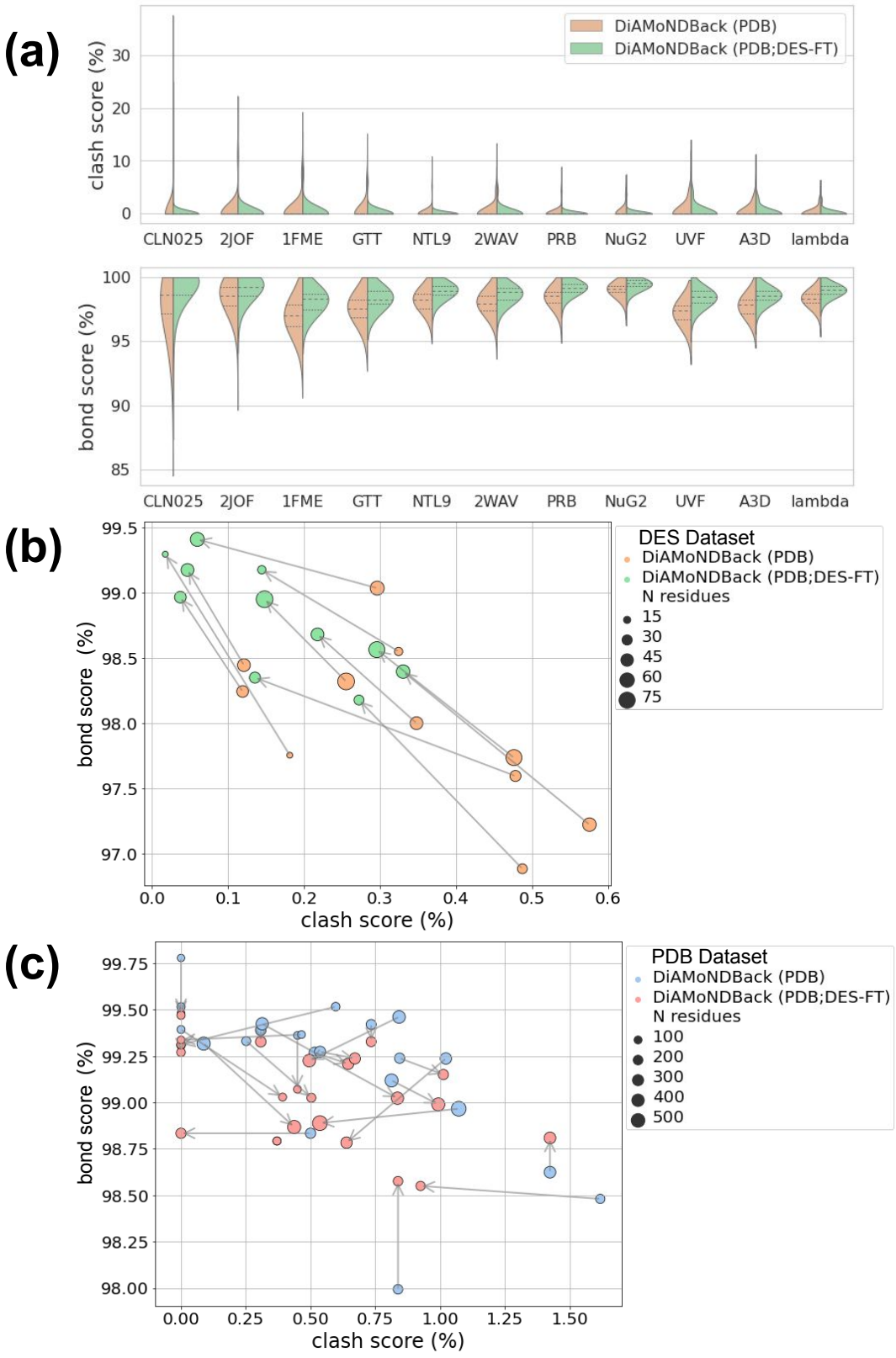
(a)
DiAMoNDBack (PDB)
DiAMoNDBack (PDB;DES-FT)
clash score (%)
bond score (%)
CLN025
2JOF
1FME
GTT
NTL9
2WAV
PRB
NuG2
UVF
A3D
lambda
(b)
DES Dataset
DiAMoNDBack (PDB)
DiAMoNDBack (PDB;DES-FT)
N residues
15
30
45
60
75
bond score (%)
clash score (%)
(c)
PDB Dataset
DiAMoNDBack (PDB)
DiAMoNDBack (PDB;DES-FT)
N residues
100
200
300
400
500
bond score (%)
clash score (%)

Figure 4: Comparison of DiAMoNDBack (PDB) and DiAMoNDBack (PDB;DES-FT) models to assess the influence of DES fine-tuning upon performance. **(a)** Split violin plots of the distribution of clash scores (top) and bond scores (bottom) over five independently generated backmapped atomistic structures over the test set of 11 proteins in the DES data for the DiAMoNDBack (PDB) and DiAMoNDBack (PDB;DES-FT) models. As expected, both the bond scores and the clash scores improve with DES fine tuning. Scatter plots showing the change in bond score and clash score in applications of the DiAMoNDBack (PDB) and DiAMoNDBack (PDB;DES-FT) models to **(b)** the 11 proteins in the DES test set and **(c)** the 24 proteins in the PDB test set. Fine tuning on the DES training data results in an across-the-board improvement in bond and clash scores on the DES test set – grey tie lines linking the results for the DiAMoNDBack (PDB) model applied to the DES test set (orange points) to those for the DiAMoNDBack (PDB;DES-FT) model applied to the DES test set (green points) illuminate a migration towards improved bond scores (up) and improved clash scores (left). Fine tuning results on the PDB test set show mixed results in the improvement/degradation of bond and clash scores – grey tie lines link the results for the DiAMoNDBack (PDB) model applied to the PDB test set (blue points) to those for the DiAMoNDBack (PDB;DES-FT) model applied to the PDB test set (red points). Symbol size indicates the size of the test protein measured by number of residues. The lone red point that appears to not be unconnected to a blue point is due to the two points lying nearly on top of one another.

model training for particular molecular force fields.

3.3 Analysis of side chain dihedral angles in generated atomistic structures

To further evaluate the structural fidelity of our backmapped structures we compared the distribution of side chain C-Cα-Cβ-Cγ dihedral angles of generated configurations relative to that collated from the test set simulation trajectories for the 11 DES proteins (Fig. 5). We conduct this analysis for 17/20 amino acids – these dihedrals are not present in the small Gly and Ala side chains and none of the DES proteins contain Cys residues. Dihedral angle distributions for the generated configurations are calculated by generating five backmapped configurations for each frame of each protein in the DES test set molecular dynamics trajectories and collating normalized histograms of the dihedral angle distribution. An analogous procedure is used to compute the reference distribution directly from the test set simulation

trajectories and we quantify the similarity of the two distributions using the Jensen-Shannon distance metric (JSD), which is the square root of the Jensen-Shannon divergence.[86] Employing a base two logarithm bounds the JSD to the range [0,1], with the lower bound of zero achieved for identical distributions.

In Fig. 5a we compare the calculated JSD values for PULCHRA, GenZProt, DiAMoNDBack (PDB), and DiAMoNDBack (PDB;DES-FT). The DiAMoNDBack models exhibit superior performance for all amino acid residues with the fine-tuned model enjoying a small benefit in performance for all residues. We present a comparison of the dihedral angle distributions for three selected residues in Fig. 5b, with the remaining residues presented in Fig. S14. The DiAMoNDBack (PDB) and DiAMoNDBack (PDB;DES-FT) models produce distributions in excellent agreement with the molecular dynamics reference and accurately recapitulate the multimodal nature of these distributions. The rules-based PULCHRA model is able to reproduce this multimodality but tends to produce longer tails and exhibits significantly poorer agreement to the reference distribution. PULCHRA operates by first generating an approximate structure using heuristics and atomic fragments followed by rotation of side chain dihedrals to resolve steric clashes.[47] We attribute the tails to this second step that overpopulates regions of side chain dihedral space that are unrepresented in the reference data. On average, we observe 26% lower JSD scores compared to PULCHRA for DiAMoNDBack (PDB) and this improves to 49% for DiAMoNDBack (PDB;DES-FT). The GenZProt model exhibits the poorest agreement to the reference data and seemingly encounters challenges in mimicking the multimodal distributions that we hypothesize may be attributable to mode collapse within the GenZProt VAE.[87] An analysis of internal energies in the backmapped configurations demonstrates that the DiAMoNDBack models also perform well in reproducing these distributions (Fig. S15-S16), but it is important to exercise caution in interpreting these results for the purposes of a structural comparison due to the high sensitivity of energy potentials to minor structural changes. We also recall that *post-hoc* energy minimization could be conducted to yield low-energy configurations as demonstrated

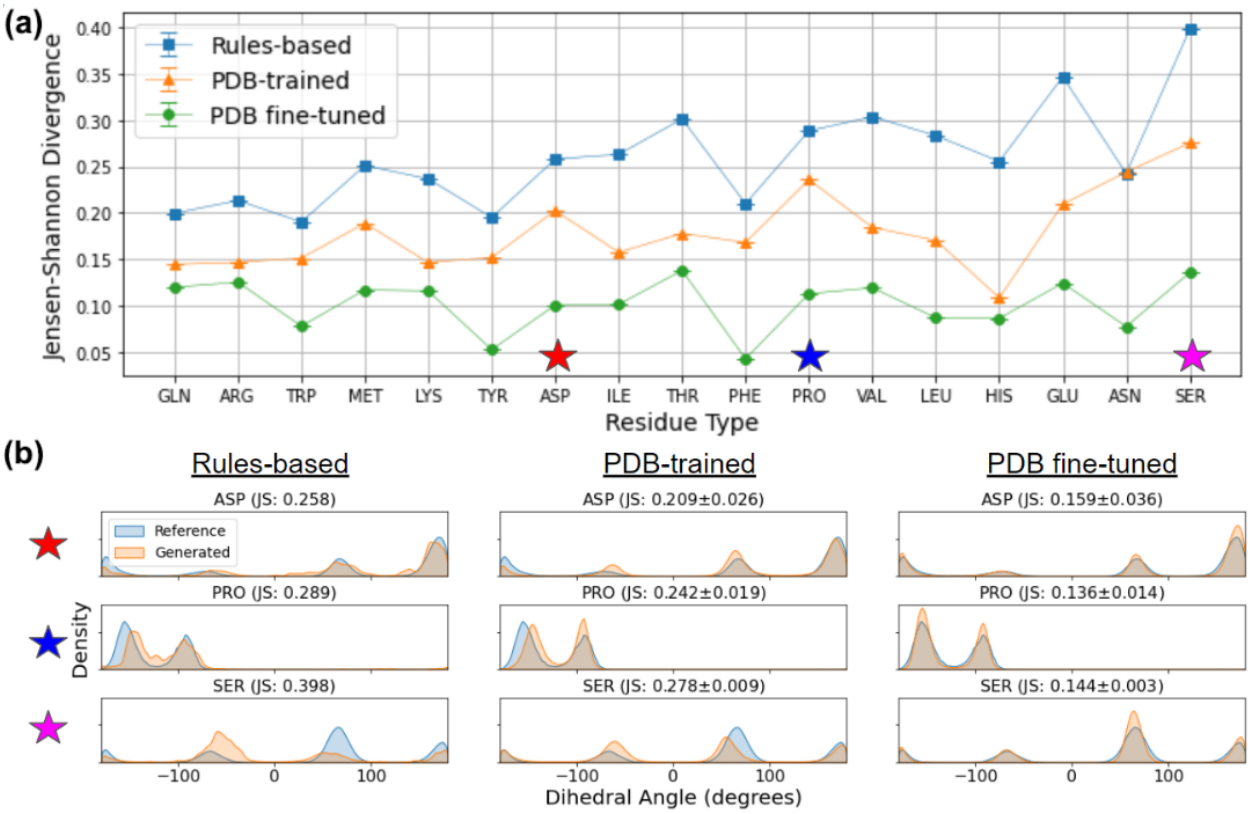

Figure 5: Comparison of side chain C-Cα-Cβ-Cγ dihedral angles distributions over the test set simulation trajectories for the 11 DES proteins. **(a)** Jensen-Shannon divergences (JSD) are computed on a residue-wise basis between the molecular dynamics reference data and the backmapped configurational ensembles generated by PULCHRA, GenZProt, DiAMoNDBack (PDB), and DiAMoNDBack (PDB;DES-FT). Data are not reported for Gly and Ala, which do not possess this dihedral angle, and Cys, which is not represented in any of the 11 DES proteins. Error bars in the reported JSD values correspond to standard deviations and are estimated for the GenZProt and DiAMoNDBack models by averaging across five independent generations. The rules-based PULCHRA model is deterministic and has no associated standard deviation as all generations produced by this method are identical. **(b)** Comparison of the dihedral angle distributions for three selected residues Asp (red star), Pro (blue star), and Ser (pink star). Plots for all 17 residues are presented in Fig. S12. The DiAMoNDBack (PDB) and DiAMoNDBack (PDB;DES-FT) models accurately recapitulate the multimodal dihedral angle distributions. The rules-based PULCHRA model tends to produce longer tails that overpopulate dihedral angles rarely visited in the reference data. The VAE-based GenZProt model encounters challenges in fitting the distributions possibly associated with mode collapse.

in Fig. S9-S10 but that we elect not to perform any refinement in our main results for the purposes of restricting our comparisons to the backmapping methodology alone without biasing to a particular molecular force-field and avoiding the large computational cost associated

with these operations for large proteins.

3.4 Analysis of residue-wise performance

We next sought to explore whether particular residue types were more prone to produce poorly formed bonds and be involved in unphysical steric collisions. Identifying residue-level trends in these performance metrics can help expose potential failure modes for our model. In Fig. 6 we present scatter plots of the bond and clash scores broken down on a per residue basis. Application of DiAMoNDBack (PDB) to the PDB test set shows that all residues possess clash scores of 3% or less, and all but three residues – Arg, Lys, and Trp – possess bond scores better than 98% (Fig. 6a). This trend is maintained in application of DiAMoNDBack (PDB;DES-FT) to the DES test set, with all residues possessing good clash scores of 1% or better, and all but Arg, Lys, and Trp possessing bond scores better than 98%. The three outliers in each case still possess good bond scores better than 93%, but it is informative to understand this relatively poorer behavior. Arg and Lys both possess long, charged side chains that are exceptionally dynamic, can rapidly exchange their protons with water, and are challenging to resolve experimentally.[88,89] This observation is consistent with our observation during our data cleaning procedure that Arg and Lys residues tended to possess significantly more incomplete side chains within the PDB data set compared to other residues, potentially indicating lower confidence in experimental resolution of side chain atomic coordinates. Trp is both the bulkiest amino acid, and therefore potentially the most susceptible to steric clashes. It is also the least represented within the PDB training data (Fig. 6c) with the relatively smaller number of training examples meaning that the model may be less well trained on this residues and less able to generalize to unfamiliar configurational environments.

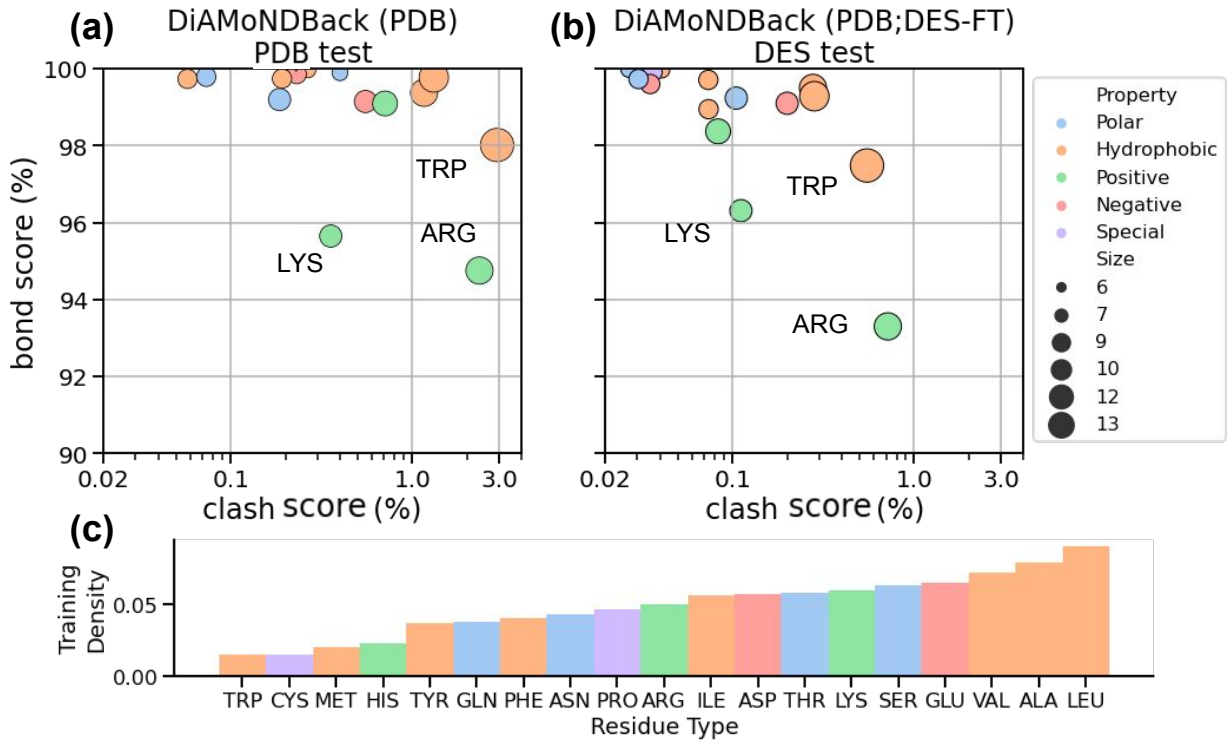

Figure 6: Analysis of DiAMoNDBack residue-wise bond and clash scores. Scatterplots illustrating the residue-wise bond and clash scores for **(a)** the DiAMoNDBack (PDB) model evaluated on the PDB test set and **(b)** the DiAMoNDBack (PDB;DES-FT) model evaluated on the DES test set. Markers are sized according to the number of atoms in the corresponding side chain and colored by phsyicochemical grouping: polar, hydrophobic, positively charged, negatively charged, and special (Pro and Cys). **(c)** Probability distribution illustrating representation of each residue type within the PDB training data. The bar heights are normalized to sum to unity.

3.5 Evaluation on CG trajectories

To illustrate a realistic application of the backmapping models to coarse grained molecular trajectories, we test the capability of PULCHRA, GenZProt, and DiAMoNDBack to restore atomistic detail to simulation trajectories of three small proteins – 1FME (28 residues), PRB (47 residues), and A3D (73 residues) – generated by Majewski et al. using bespoke Cα-based coarse-grained potentials[27] (Table 3). Compared to our prior analyses, these data represent an out-of-distribution test case generated by a coarse-grained force field to which the model was never exposed during training and for which there is no atomistic ground truth. The bond scores for all four models are excellent and on par with those for the in-sample PDB

and DES tests reported in Table 2. The clash scores are slightly poorer but for PULCHRA and DiAMoNDBack are still very good, achieving 2% or fewer clashes for all three proteins. The GenZProt clash score is quite poor at 8-15%. Without atomistic structures we cannot use Eqn. 3 to compute the diversity score, so instead compute $RMSD_{gen}$ using Eqn. 2 as a proxy for generative diversity. The deterministic PULCHRA model generates zero diversity, while GenZProt, DiAMoNDBack (PDB), and DiAMoNDBack (PDB;DES-FT) generate $RMSD_{gen}$ values of, respectively, 0.212 nm, 1.69 nm, and 1.57 nm on average. The baseline DiAMoNDBack (PDB) model produces an average $RMSD_{gen}$ increase over GenZProt of nearly 8× across these three sequences. Finally, in Fig. 7 we present illustrative visualizations of the atomistic backmappings for these three coarse-grained proteins generated by DiAMoNDBack (PDB). These results illustrate that DiAMoNDBack is capable of performing physically realistic and diverse atomistic backmappings for out-of-distribution coarse grained simulation trajectories produced by a Cα-based coarse-grained model.

4 Conclusions

In this work, we present DiAMoNDBack (Diffusion-denoising Autoregressive Model for Non-Deterministic Backmapping) as a transferable approach for backmapping Cα protein traces into atomistic structures using Denoising Diffusion Probabilistic Models (DDPMs). Our approach builds the protein structure in an autoregressive manner residue-by-residue, at each instance predicting the Cartesian coordinates of the target residue aligned to a canonical reference frame conditioned on all previously decoded residues within a local neighborhood and the coarse-grained Cα trace. We train models on a corpus of 65k+ PDB structures exposing our model to a rich variety of local residue environments to establish a general-purpose model for backmapping generic Cα traces. Evaluating our approach on held-out PDB structures and all-atom molecular dynamics simulations reveals excellent performance in terms of the quality of the reconstructed atomic bonds, the degree of non-bonded steric

Table 3: Application of backmapping to bespoke coarse-grained trajectories generated by Majewski et al.[27] Comparisons are presented for three sequences spanning a range of lengths: 1FME (28 residues), PRB (47 residues), and A3D (73 residues). Standard deviations in the bond and clash scores reported for DiAMoNDBack and GenZProt are estimated using five-fold block averaging for the bond and clash scores. As a deterministic algorithm, the values reported for PULCHRA do not have associated uncertainties. In the absence of an atomistic ground truth, the average pairwise RMSD between generated samples is reported as a proxy for generative diversity.

	CG 1FME	CG PRB	CG A3D
	Bond (↑) [%]	Bond (↑) [%]	Bond (↑) [%]
PULCHRA	98.93	99.18	99.60
GenZProt	95.57±0.01	97.210±0.003	96.930±0.001
DiAMoNDBack (PDB)	96.72±0.01	98.23±0.01	98.25±0.01
DiAMoNDBack (PDB;DES-FT)	97.81±0.03	98.97±0.02	98.64±0.01
	Clash (↓) [%]	Clash (↓) [%]	Clash (↓) [%]
PULCHRA	1.29	0.673	0.298
GenZProt	14.01±0.03	8.20±0.05	11.88±0.01
DiAMoNDBack (PDB)	1.96±0.04	1.12±0.05	1.06±0.04
DiAMoNDBack (PDB;DES-FT)	1.75±0.06	0.97±0.03	1.08±0.04
	RMSD (↑) [nm]	RMSD (↑) [nm]	RMSD (↑) [nm]
PULCHRA	0	0	0
GenZProt	0.23±0.09	0.225±0.070	0.18±0.02
DiAMoNDBack (PDB)	1.96±0.29	1.43±0.16	1.67±0.17
DiAMoNDBack (PDB;DES-FT)	1.80±0.27	1.30±0.14	1.57 ±0.15

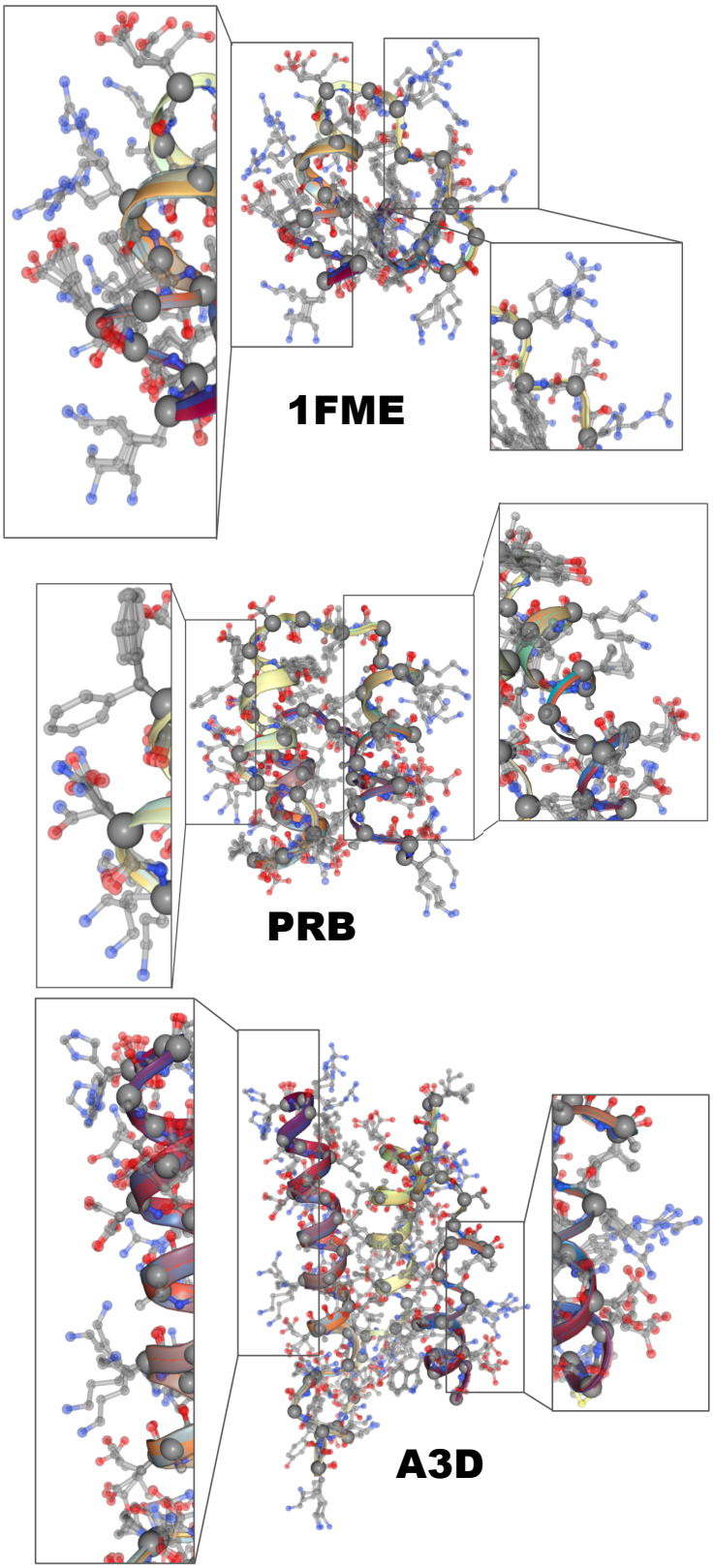
1FME
PRB
A3D

Figure 7: Illustrative DiAMoNDBack (PDB) atomistic backmappings for coarse-grained simulation trajectories of the 1FME (28 residues), PRB (47 residues), and A3D (73 residues) proteins generated by Majewski et al. using bespoke Cα-based coarse-grained potentials.[27] Five indepnedent backmapped structures conditioned on the coarse-grained Cα trace are shown as translucent and the single Cα trace shown as opaque. Insets for each protein zoom into particular regions to highlight the generative diversity in side chain placements.

clashes between residues, and the capacity to generate a diverse ensemble of atomistic structures consistent with a particular Cα trace. While we find improved structural metrics compared to the previous transferable backmapping work of Yang and Gómez-Bombarelli[55] (GenZProt) and are competitive with rules-based approaches[47] (PULCHRA), the generative diversity is where our model excels in producing a substantially more diverse ensemble of atomistic reconstructions consistent with the coarse-grained Cα trace. Analysis of side chain dihedral angles in reconstructed structures also reveal our model generates more physically plausible distributions of internal coordinates compared to the rules-based approach that involves manually adjusting side chain dihedral angles to resolve clashes. We demonstrate a deployment of our model to coarse-grained simulation trajectories generated by bespoke Cα-based coarse-grained force fields, and show that it can generate high-quality bonds ($\sim$98% bond lengths within 10% of reference data) and low fractions of clashing residues ($\sim$1.25%). We also demonstrate fine-tuning of our baseline PDB-trained DiAMoNDBack model on limited numbers of all-atom simulation data to develop a force field-specific model with slightly improved performance on those data. We believe that DiAMoNDBack offers a generic, transferable, and accurate backmapping tool of value to the community and we have made it freely available as an open source Python package (see Data Availability Statement).

In future work we would like to investigate a number of innovations to further enhance the quality of our atomistic reconstructions. Our model currently generates atomistic residues from Cα traces conditioned on the N-nearest neighboring residues. Our current data representation exposes the Cartesian coordinates of the local environment and imposes practical limits on the conditioning size due to increasing dimensionality slowing training and infer-

ence efficiencies. Residues outside the purview of this local neighborhood can therefore be excluded from the conditioning effectively hidden from the model and potentially leading to clashes. A more comprehensive conditioning scheme that incorporates the full protein structure, for instance by using a graph neural network to assemble the conditioning information, could potentially improve the quality of our generated structures. One significant drawback of our model is that the DDPM generation process is relatively slow, making DiAMoND-Back approximately 50× slower than GenZProt and 1000-3000× slower than PULCHRA (Fig. S7). Speed-ups could be achieved by treating non-interacting regions of a structure independently and decoding residues in parallel when possible to accelerate upon sequential N-to-C decoding. While in this work we focus on backmapping from Cα traces, we observe that with minor architectural changes and retraining the framework is extensible to any resolution of coarse-graining and could readily be applied to multi-site per residue models such as MARTINI[11,14,15] and AWSEM[20] or even multi-residue per site ultra coarse-grained models.[90,91] Furthermore, growing repositories of nucleic acid protein complexes[92] can be used to train a model that backmaps from DNA-protein coarse-grained forcefields such as AWSEM-3SPN.2[93] or GENESIS-CG.[94]

Data and Code Availability

We make our backmapping model publicly available by releasing pre-trained models and code for use at `https://github.com/Ferg-Lab/DiAMoNDBack`. We also make available all the data splits used to train and test the models reported in this work available via Zenodo at DOI:10.5281/zenodo.8169238.[95]

Acknowledgement

This material is based on work supported by the National Science Foundation under Grant No. CHE-2152521. K.S. was supported by a fellowship from the Molecular Sciences Soft-

ware Institute under the National Science Foundation Grant No. CHE-2136142. This work was completed in part with resources provided by the University of Chicago Research Computing Center. We gratefully acknowledge computing time on the University of Chicago high-performance GPU-based cyberinfrastructure supported by the National Science Foundation under Grant No. DMR-1828629. We thank Adrià Pérez for their help in sharing the coarse-grained trajectories. We are grateful to D.E. Shaw Research for sharing the protein simulation trajectories. Some of the material presented in this manuscript first appeared in Dr. Shmilovich's PhD dissertation, presented at the University of Chicago in July 2023.[96]

Supporting Information Available

Additional information on the mathematical formalism, DDPM loss function, conditional U-net architecture, residue featurization, canonical alignment process, treatment of N- and C-terminal residues, model hyperparameter tuning, residue representation in our training data, usage of the PULCHRA and GenZProt models, analysis of reconstructed terminal residue quality, dihedral angle distributions for all residues, error as a function of residue position along the protein chain, and analysis of internal energies.

Conflict of Interest Statement

A.L.F. is a co-founder and consultant of Evozyne, Inc. and a co-author of US Patent Applications 16/887,710 and 17/642,582, US Provisional Patent Applications 62/853,919, 62/900,420, 63/314,898, 63/479,378, and 63/521,617, and International Patent Applications PCT/US2020/035206 and PCT/US2020/050466.

References

(1) Clementi, C. Coarse-grained models of protein folding: toy models or predictive tools? *Curr. Opin. Struct. Biol.* **2008**, *18*, 10–15.

(2) Noid, W. G. Perspective: Coarse-grained models for biomolecular systems. *J. Chem. Phys.* **2013**, *139*, 09B201_1.

(3) Saunders, M. G.; Voth, G. A. Coarse-graining methods for computational biology. *Annu. Rev. Biophys.* **2013**, *42*, 73–93.

(4) Kmiecik, S.; Gront, D.; Kolinski, M.; Wieteska, L.; Dawid, A. E.; Kolinski, A. Coarse-grained protein models and their applications. *Chem. Rev.* **2016**, *116*, 7898–7936.

(5) Mohr, B.; Shmilovich, K.; Kleinwächter, I. S.; Schneider, D.; Ferguson, A. L.; Bereau, T. Data-driven discovery of cardiolipin-selective small molecules by computational active learning. *Chem. Sci.* **2022**, *13*, 4498–4511.

(6) Shmilovich, K.; Mansbach, R. A.; Sidky, H.; Dunne, O. E.; Panda, S. S.; Tovar, J. D.; Ferguson, A. L. Discovery of self-assembling π-conjugated peptides by active learning-directed coarse-grained molecular simulation. *J. Phys. Chem. B.* **2020**, *124*, 3873–3891.

(7) Kim, Y. C.; Hummer, G. Coarse-grained models for simulations of multiprotein complexes: application to ubiquitin binding. *J. Mol. Biol.* **2008**, *375*, 1416–1433.

(8) Scott, K. A.; Bond, P. J.; Ivetac, A.; Chetwynd, A. P.; Khalid, S.; Sansom, M. S. Coarse-grained MD simulations of membrane protein-bilayer self-assembly. *Structure* **2008**, *16*, 621–630.

(9) Lequieu, J.; Córdoba, A.; Moller, J.; De Pablo, J. J. 1CPN: A coarse-grained multi-scale model of chromatin. *J. Chem. Phys.* **2019**, *150*, 215102.

(10) Marrink, S. J.; Tieleman, D. P. Perspective on the Martini model. *Chem. Soc. Rev.* **2013**, *42*, 6801–6822.

(11) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; De Vries, A. H. The MARTINI force field: coarse grained model for biomolecular simulations. *J. Phys. Chem. B.* **2007**, *111*, 7812–7824.

(12) Fritz, D.; Koschke, K.; Harmandaris, V. A.; van der Vegt, N. F.; Kremer, K. Multiscale modeling of soft matter: scaling of dynamics. *Phys. Chem. Chem. Phys.* **2011**, *13*, 10412–10420.

(13) Levitt, M.; Warshel, A. Computer simulation of protein folding. *Nature* **1975**, *253*, 694–698.

(14) Monticelli, L.; Kandasamy, S. K.; Periole, X.; Larson, R. G.; Tieleman, D. P.; Marrink, S. J. The MARTINI coarse-grained force field: extension to proteins. *J. Chem. Theory Comput.* **2008**, *4*, 819–834.

(15) Souza, P. C. T.; Alessandri, R.; Barnoud, J.; Thallmair, S.; Faustino, I.; Grünewald, F.; Patmanidis, I.; Abdizadeh, H.; Bruininks, B. M. H.; Wassenaar, T. A.; Kroon, P. C.; Melcr, J.; Nieto, V.; Corradi, V.; Khan, H. M.; Domanski, J. J.; Javanainen, M.; Martinez-Seara, H.; Reuter, N.; Best, R. B.; Vattulainen, I.; Monticelli, L.; Periole, X.; Tieleman, D. P.; de Vries, A. H.; Marrink, S. J. Martini 3: a general purpose force field for coarse-grained molecular dynamics. *Nat. Methods* **2021**, *18*, 382 – 388.

(16) Seo, S.; Shinoda, W. SPICA force field for lipid membranes: domain formation induced by cholesterol. *J. Chem. Theory Comput.* **2018**, *15*, 762–774.

(17) Das, R.; Baker, D. Macromolecular modeling with Rosetta. *Annu. Rev. Biochem.* **2008**, *77*, 363–382.

(18) Han, W.; Wan, C.-K.; Jiang, F.; Wu, Y.-D. PACE force field for protein simulations. 1. Full parameterization of version 1 and verification. *J. Chem. Theory Comput.* **2010**, *6*, 3373–3389.

(19) Kolinski, A. Protein modeling and structure prediction with a reduced representation. *Acta Biochim. Pol.* **2004**, *51 2*, 349–71.

(20) Davtyan, A.; Schafer, N. P.; Zheng, W.; Clementi, C.; Wolynes, P. G.; Papoian, G. A. AWSEM-MD: protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *J. Phys. Chem. B.* **2012**, *116*, 8494–8503.

(21) Jumper, J. M.; Faruk, N. F.; Freed, K. F.; Sosnick, T. R. Accurate calculation of side chain packing and free energy with applications to protein molecular dynamics. *PLoS Comput. Biol.* **2018**, *14*, e1006342.

(22) Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W. DeePCG: Constructing coarse-grained models via deep neural networks. *J. Chem. Phys.* **2018**, *149*, 034101.

(23) Wang, J.; Olsson, S.; Wehmeyer, C.; Pérez, A.; Charron, N. E.; De Fabritiis, G.; Noé, F.; Clementi, C. Machine learning of coarse-grained molecular dynamics force fields. *ACS Cent. Sci.* **2019**, *5*, 755–767.

(24) Wang, W.; Gómez-Bombarelli, R. Coarse-graining auto-encoders for molecular dynamics. *npj Comput. Mater.* **2019**, *5*, 125.

(25) Husic, B. E.; Charron, N. E.; Lemm, D.; Wang, J.; Pérez, A.; Majewski, M.; Krämer, A.; Chen, Y.; Olsson, S.; de Fabritiis, G.; Noé, F.; Clementi, C. Coarse graining molecular dynamics with graph neural networks. *J. Chem. Phys.* **2020**, *153 19*, 194101.

(26) Chennakesavalu, S.; Toomer, D. J.; Rotskoff, G. M. Ensuring thermodynamic consistency with invertible coarse-graining. *J. Chem. Phys.* **2023**, *158*, 124126.

(27) Majewski, M.; Pérez, A.; Thölke, P.; Doerr, S.; Charron, N. E.; Giorgino, T.; Husic, B. E.; Clementi, C.; Noé, F.; De Fabritiis, G. Machine Learning Coarse-Grained Potentials of Protein Thermodynamics. *arXiv preprint arXiv:2212.07492* **2022**,

(28) Ding, X.; Zhang, B. Contrastive learning of coarse-grained force fields. *J. Chem. Theory Comput.* **2022**, *18*, 6334–6344.

(29) Durumeric, A. E.; Charron, N. E.; Templeton, C.; Musil, F.; Bonneau, K.; Pasos-Trejo, A. S.; Chen, Y.; Kelkar, A.; Noé, F.; Clementi, C. Machine learned coarse-grained protein force-fields: Are we there yet? *Curr. Opin. Struct. Biol.* **2023**, *79*, 102533.

(30) Kohler, J.; Chen, Y.; Krämer, A.; Clementi, C.; Noé, F. Flow-matching: Efficient coarse-graining of molecular dynamics without forces. *J. Chem. Theory Comput.* **2023**, *19*, 942–952.

(31) Arts, M.; Satorras, V. G.; Huang, C.-W.; Zuegner, D.; Federici, M.; Clementi, C.; Noé, F.; Pinsler, R.; Berg, R. v. d. Two for One: Diffusion Models and Force Fields for Coarse-Grained Molecular Dynamics. *arXiv preprint arXiv:2302.00600* **2023**,

(32) Krämer, A.; Durumeric, A. P.; Charron, N. E.; Chen, Y.; Clementi, C.; Noé, F. Statistically Optimal Force Aggregation for Coarse-Graining Molecular Dynamics. *arXiv preprint arXiv:2302.07071* **2023**,

(33) Izvekov, S.; Voth, G. A. A multiscale coarse-graining method for biomolecular systems. *J. Phys. Chem. B.* **2005**, *109*, 2469–2473.

(34) Noid, W. G.; Chu, J.-W.; Ayton, G. S.; Krishna, V.; Izvekov, S.; Voth, G. A.; Das, A.; Andersen, H. C. The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.* **2008**, *128*, 244114.

(35) Shell, M. S. The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *J. Chem. Phys.* **2008**, *129*, 144108.

(36) Badaczewska-Dawid, A. E.; Kolinski, A.; Kmiecik, S. Computational reconstruction of atomistic protein structures from coarse-grained models. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 162–176.

(37) McQuarrie, D. A.; Simon, J. D. *Physical chemistry: a molecular approach*; University science books Sausalito, CA, 1997; Vol. 1.

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

(39) Hess, B.; León, S.; Van Der Vegt, N.; Kremer, K. Long time atomistic polymer trajectories from coarse grained simulations: bisphenol-A polycarbonate. *Soft Matter* **2006**, *2*, 409–414.

(40) Peter, C.; Kremer, K. Multiscale simulation of soft matter systems–from the atomistic to the coarse-grained level and back. *Soft Matter* **2009**, *5*, 4357–4366.

(41) Rzepiela, A. J.; Schäfer, L. V.; Goga, N.; Risselada, H. J.; De Vries, A. H.; Marrink, S. J. Reconstruction of atomistic details from coarse-grained structures. *J. Comput. Chem.* **2010**, *31*, 1333–1343.

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

(43) Wassenaar, T. A.; Pluhackova, K.; Böckmann, R. A.; Marrink, S. J.; Tieleman, D. P. Going backward: a flexible geometric approach to reverse transformation from coarse grained to atomistic models. *J. Chem. Theory Comput.* **2014**, *10*, 676–690.

(44) Gopal, S. M.; Mukherjee, S.; Cheng, Y.-M.; Feig, M. PRIMO/PRIMONA: a coarse-grained model for proteins and nucleic acids that preserves near-atomistic accuracy. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1266–1281.

(45) Brocos, P.; Mendoza-Espinosa, P.; Castillo, R.; Mas-Oliva, J.; Pineiro, A. Multiscale molecular dynamics simulations of micelles: coarse-grain for self-assembly and atomic resolution for finer details. *Soft Matter* **2012**, *8*, 9005–9014.

(46) Machado, M. R.; Pantano, S. SIRAH tools: mapping, backmapping and visualization of coarse-grained models. *Bioinformatics* **2016**, *32*, 1568–1570.

(47) Rotkiewicz, P.; Skolnick, J. Fast procedure for reconstruction of full-atom protein models from reduced representations. *J. Comput. Chem.* **2008**, *29*, 1460–1465.

(48) Nicholson, D. N.; Greene, C. S. Constructing knowledge graphs and their biomedical applications. *Comput. Struct. Biotechnol. J.* **2020**, *18*, 1414–1428.

(49) Stieffenhofer, M.; Bereau, T.; Wand, M. Adversarial reverse mapping of condensed-phase molecular structures: Chemical transferability. *APL Mater.* **2021**, *9*, 031107.

(50) Stieffenhofer, M.; Wand, M.; Bereau, T. Adversarial reverse mapping of equilibrated condensed-phase molecular structures. *Mach. Learn.: Sci. Technol.* **2020**, *1*, 045014.

(51) Li, W.; Burkhart, C.; Polińska, P.; Harmandaris, V.; Doxastakis, M. Backmapping coarse-grained macromolecules: An efficient and versatile machine learning approach. *J. Chem. Phys.* **2020**, *153*, 041101.

(52) An, Y.; Deshmukh, S. A. Machine learning approach for accurate backmapping of coarse-grained models to all-atom models. *Chem. Commun.* **2020**, *56*, 9312–9315.

(53) Wang, W.; Xu, M.; Cai, C.; Miller, B. K.; Smidt, T.; Wang, Y.; Tang, J.; Gómez-Bombarelli, R. Generative coarse-graining of molecular conformations. *arXiv preprint arXiv:2201.12176* **2022**,

(54) Shmilovich, K.; Stieffenhofer, M.; Charron, N. E.; Hoffmann, M. Temporally coherent backmapping of molecular trajectories from coarse-grained to atomistic resolution. *J. Phys. Chem. A* **2022**, *126*, 9124–9139.

(55) Yang, S.; Gómez-Bombarelli, R. Chemically transferable generative backmapping of coarse-grained proteins. *arXiv preprint arXiv:2303.01569* **2023**,

(56) Kingma, D. P.; Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* **2013**,

(57) Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144.

(58) Lazar, T.; Martínez-Pérez, E.; Quaglia, F.; Hatos, A.; Chemes, L. B.; Iserte, J.; 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. D.; Ricard-Blum, S.; Ruff, K. M.; Salladini, E.; Skepö, M.; Svergun, D. I.; Vallet, S. D.; Váradi, M.; Tompa, P.; Tosatto, S. C. E.; Piovesan, D. PED in 2021: A major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Res.* **2021**, *49*, D404 – D411.

(59) Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.

(60) Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. International Conference on Machine Learning. 2015; pp 2256–2265.

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

(62) Schneuing, A.; Du, Y.; Harris, C.; Jamasb, A. R.; Igashov, I.; Du, W.; Blundell, T. L.; Li'o, P.; Gomes, C.; Welling, M.; Bronstein, M. M.; Correia, B. E. Structure-based drug design with equivariant diffusion models. *arXiv preprint arXiv:2210.13695* **2022**,

(63) Jing, B.; Corso, G.; Chang, J.; Barzilay, R.; Jaakkola, T. Torsional diffusion for molecular conformer generation. *arXiv preprint arXiv:2206.01729* **2022**,

(64) Igashov, I.; Stärk, H.; Vignac, C.; Satorras, V. G.; Frossard, P.; Welling, M.; Bronstein, M.; Correia, B. Equivariant 3d-conditional diffusion models for molecular linker design. *arXiv preprint arXiv:2210.05274* **2022**,

(65) Wu, K. E.; Yang, K. K.; Berg, R. v. d.; Zou, J. Y.; Lu, A. X.; Amini, A. P. Protein structure generation via folding diffusion. *arXiv preprint arXiv:2209.15611* **2022**,

(66) Trippe, B. L.; Yim, J.; Tischer, D.; Broderick, T.; Baker, D.; Barzilay, R.; Jaakkola, T. Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem. *arXiv preprint arXiv:2206.04119* **2022**,

(67) Watson, J. L.; Juergens, D.; Bennett, N. R.; Trippe, B. L.; Yim, J.; Eisenach, H. E.; Ahern, W.; Borst, A. J.; Ragotte, R. J.; Milles, L. F.; Wicky, B. I. M.; Hanikel, N.; Pellock, S. J.; Courbet, A.; Sheffler, W.; Wang, J.; Venkatesh, P.; Sappington, I.; Torres, S. V.; Lauko, A.; Bortoli, V. D.; Mathieu, E.; Barzilay, R.; Jaakkola, T.; DiMaio, F.; Baek, M.; Baker, D. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. *bioRxiv* **2022**, 2022–12.

(68) Qiao, Z.; Nie, W.; Vahdat, A.; Miller III, T. F.; Anandkumar, A. Dynamic-Backbone Protein-Ligand Structure Prediction with Multiscale Generative Diffusion Models. *arXiv preprint arXiv:2209.15171* **2022**,

(69) Ingraham, J.; Baranov, M.; Costello, Z.; Frappier, V.; Ismail, A.; Tie, S.; Wang, W.; Xue, V.; Obermeyer, F.; Beam, A.; Grigoryan, G. Illuminating protein space with a programmable generative model. *bioRxiv* **2022**, 2022–12.

(70) AlQuraishi, M. ProteinNet: A standardized data set for machine learning of protein structure. *BMC Bioinf.* **2019**, *20*, 1–10.

(71) King, J. E.; Koes, D. R. SidechainNet: An all-atom protein structure dataset for machine learning. *Proteins: Struct., Funct., Bioinf.* **2021**, *89*, 1489–1496.

(72) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

(73) Berman, H.; Henrick, K.; Nakamura, H. Announcing the worldwide protein data bank. *Nat. Struct. Mol. Biol.* **2003**, *10*, 980–980.

(74) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How fast-folding proteins fold. *Science* **2011**, *334*, 517–520.

(75) Stieffenhofer, M.; Scherer, C.; May, F.; Bereau, T.; Andrienko, D. Benchmarking coarse-grained models of organic semiconductors via deep backmapping. *Front. Chem.* **2022**, *10*, 982757.

(76) Luo, C. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970* **2022**,

(77) Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Shao, Y.; Zhang, W.; Cui, B.; Yang, M.-H. Diffusion models: A comprehensive survey of methods and applications. *arXiv preprint arXiv:2209.00796* **2022**,

(78) Dhariwal, P.; Nichol, A. Diffusion models beat GANs on image synthesis. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 8780–8794.

(79) Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention. 2015; pp 234–241.

(80) Wang, Y.; Herron, L.; Tiwary, P. From data to noise to data for mixing physics across

temperatures with generative artificial intelligence. *Proc. Natl. Acad. Sci. U. S. A.* **2022**, *119*, e2203656119.

(81) Shmilovich, K.; Willmott, D.; Batalov, I.; Kornbluth, M.; Mailoa, J.; Kolter, J. Z. Orbital Mixer: Using atomic orbital features for basis-dependent prediction of molecular wavefunctions. *J. Chem. Theory Comput.* **2022**, *18*, 6021–6030.

(82) Moult, J.; Fidelis, K.; Kryshtafovych, A.; Schwede, T.; Tramontano, A. Critical assessment of methods of protein structure prediction (CASP)—round x. *Proteins: Struct., Funct., Bioinf.* **2014**, *82*, 1–6.

(83) Schaarschmidt, J.; Monastyrskyy, B.; Kryshtafovych, A.; Bonvin, A. M. Assessment of contact predictions in CASP12: co-evolution and deep learning coming of age. *Proteins: Struct., Funct., Bioinf.* **2018**, *86*, 51–66.

(84) Sidky, H.; Chen, W.; Ferguson, A. L. High-resolution Markov state models for the dynamics of Trp-cage miniprotein constructed over slow folding modes identified by state-free reversible VAMPnets. *J. Phys. Chem. B.* **2019**, *123*, 7999–8009.

(85) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 1–14.

(86) Lin, J. Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **1991**, *37*, 145–151.

(87) Lucas, J.; Tucker, G.; Grosse, R. B.; Norouzi, M. Understanding Posterior Collapse in Generative Latent Variable Models. DGS@ICLR. 2019.

(88) Nguyen, D.; Chen, C.; Pettitt, B. M.; Iwahara, J. NMR methods for characterizing the basic side chains of proteins: electrostatic interactions, hydrogen bonds, and conformational dynamics. *Methods Enzymol.* **2019**, *615*, 285–332.

(89) Esadze, A.; Li, D.-W.; Wang, T.; Brüschweiler, R.; Iwahara, J. Dynamics of lysine side-chain amino groups in a protein studied by heteronuclear 1H- 15N NMR spectroscopy. *J. Am. Chem. Soc.* **2011**, *133*, 909–919.

(90) Dama, J. F.; Sinitskiy, A. V.; McCullagh, M.; Weare, J.; Roux, B.; Dinner, A. R.; Voth, G. A. The theory of ultra-coarse-graining. 1. General principles. *J. Chem. Theory Comput.* **2013**, *9*, 2466–2480.

(91) Trylska, J. Coarse-grained models to study dynamics of nanoscale biomolecules and their applications to the ribosome. *J. Phys.: Condens. Matter* **2010**, *22*, 453101.

(92) Sagendorf, J. M.; Markarian, N.; Berman, H. M.; Rohs, R. DNAproDB: an expanded database and web-based tool for structural analysis of DNA–protein complexes. *Nucleic Acids Res.* **2020**, *48*, D277–D287.

(93) Lu, W.; Bueno, C.; Schafer, N. P.; Moller, J.; Jin, S.; Chen, X.; Chen, M.; Gu, X.; Davtyan, A.; de Pablo, J. J.; Wolynes, P. G. OpenAWSEM with Open3SPN2: A fast, flexible, and accessible framework for large-scale coarse-grained biomolecular simulations. *PLoS Comput. Biol.* **2021**, *17*, e1008308.

(94) Tan, C.; Jung, J.; Kobayashi, C.; Torre, D. U. L.; Takada, S.; Sugita, Y. Implementation of residue-level coarse-grained models in GENESIS for large-scale molecular dynamics simulations. *PLoS Comput. Biol.* **2022**, *18*, e1009578.

(95) Jones, M. S.; Shmilovich, K.; Ferguson, A. L. Supporting data for:"DiAMoNDBack: Diffusion-denoising Autoregressive Model for Non-Deterministic Backmapping of Cα Protein Traces". 2023; `https://doi.org/10.5281/zenodo.8169239`.

(96) Shmilovich, K. Data-Driven Approaches for Molecular Design and Simulation: From Self-Assembling Peptides to Enhanced Sampling Techniques and Atomistic Structure Generation. Ph.D. thesis, The University of Chicago, 2023.

Graphical TOC Entry