EIGENFOLD: GENERATIVE PROTEIN STRUCTURE PREDICTION WITH DIFFUSION MODELS

Bowen Jing,*[1] Ezra Erives,[1] Peter Pao-Huang,[3] Gabriele Corso,[1]
Bonnie Berger,[1 2] Tommi Jaakkola[1]
[1]CSAIL, Massachusetts Institute of Technology
[2]Dept. of Mathematics, Massachusetts Institute of Technology
[3]Dept. of Computer Science, University of Illinois

ABSTRACT

Protein structure prediction has reached revolutionary levels of accuracy on single structures, yet distributional modeling paradigms are needed to capture the conformational ensembles and flexibility that underlie biological function. Towards this goal, we develop EIGENFOLD, a diffusion generative modeling framework for sampling a distribution of structures from a given protein sequence. We define a diffusion process that models the structure as a system of harmonic oscillators and which naturally induces a cascading-resolution generative process along the eigenmodes of the system. On recent CAMEO targets, EIGENFOLD achieves a median TMScore of 0.84, while providing a more comprehensive picture of model uncertainty via the ensemble of sampled structures relative to existing methods. We then assess EIGENFOLD's ability to model and predict conformational heterogeneity for fold-switching proteins and ligand-induced conformational change. Code is available at `https://github.com/bjing2016/EigenFold`.

1 INTRODUCTION

The development of accurate methods for protein structure prediction such as AlphaFold2 (Jumper et al., 2021) has revolutionized *in silico* understanding of protein structure and function. However, while such methods are designed to model static experimental structures from crystallography or cryo-EM, proteins *in vivo* adopt dynamic structural ensembles featuring conformational flexibility, change, and even disorder to effect their biological functions (Teague, 2003; Wright & Dyson, 2015). These aspects represent the next frontier towards a more complete understanding of protein structure and function (Lane, 2023). Accordingly, there is increasing need for *generative* models for protein structure prediction that can produce a distribution of conformations for a single protein sequence.

Meanwhile, generative modeling in molecular machine learning has undergone a renaissance driven by the paradigm of *diffusion models* (Sohl-Dickstein et al., 2015; Song et al., 2021). When applied to problems such as protein design (Watson et al., 2022), molecular docking (Corso et al., 2022), and ligand design (Schneuing et al., 2022), such models have displayed impressive distributional modeling. These capabilities make diffusion models compelling tools for understanding protein structural ensembles given a fixed sequence, but they have yet to be explored for this purpose.

To bridge this gap, we develop EIGENFOLD, the first diffusion generative modeling framework for protein structure (and structural ensemble) prediction. To do so, we formulate a novel diffusion process—*harmonic diffusion*—that models the molecule as a system of harmonic oscillators. The structure is projected onto the eigenmodes (or normal modes) of the system during the forward diffusion, such that the corresponding reverse diffusion can be viewed as a cascading-resolution generative process—first sampling the rough global structure before refining local details. This enables EIGENFOLD to accurately sample protein structures with as few as 100 inference steps.

The EIGENFOLD framework can be used in isolation or in conjunction with pretrained embeddings from existing structure prediction models. In this work, we train EIGENFOLD using edge and node

*Correspondence to `bjing@mit.edu`

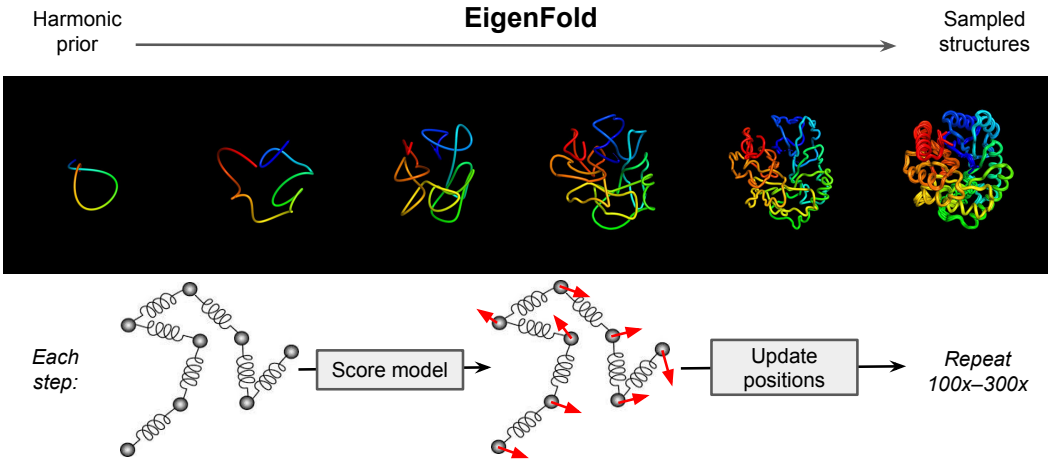

Figure 1: Overview of generative structure prediction with EIGENFOLD. Samples from the harmonic diffusion prior (left) are gradually transformed into complete structures via a cascading-resolution generative process. One complete trajectory and five final sampled structures for CAMEO target `7dz2.C` are shown here. In each denoising step, the score model predicts update "forces," which are combined with harmonic constraints to update the residue positions. Depending on protein size, this procedure is repeated 100 to 300 times.

embeddings from OmegaFold (Wu et al., 2022c), in effect transforming the latter into a generative model. When trained on the PDB and evaluated on CAMEO benchmarks, samples from this model are comparable with existing methods such as RoseTTAFold (Baek et al., 2021) in terms of single-structure accuracy. However, unlike existing methods, EIGENFOLD provides a *distribution* of structures rather than scalar predicted errors, providing more insights into model uncertainty. In particular, we show that the *variation* among the sampled structures is highly indicative of the model *error* relative to the ground truth, for many metrics of model accuracy.

We then benchmark EIGENFOLD's ability to model conformational change and flexibility using two datasets: one of fold-switching proteins (Chakravarty & Porter, 2022) and one of conformational changes associated with binding (Saldaño et al., 2022). The analysis yields mixed results, in which properties of EIGENFOLD sampled structures are moderately correlated with properties of the ground-truth conformation but do not predict them to high accuracy. While not quite bridging the gap between single-structure prediction and structural ensemble prediction, these results and methodology lay the foundation for many possible directions for improvement in future work.

2 BACKGROUND AND RELATED WORK

Protein structure prediction. The problem of predicting an experimental-level protein structure from sequence is widely considered to have been solved by AlphaFold2 (Jumper et al., 2021) in CASP 14. Since then, alternative models such as RoseTTAFold (Baek et al., 2021), ESMFold (Lin et al., 2022), and OmegaFold (Wu et al., 2022c) have replicated or approached similar levels of performance. All are developed and trained as *deterministic* maps from input (sequence or MSA) to output (structure), making them suboptimal for modeling structural ensembles (Lane, 2023; Chakravarty & Porter, 2022; Saldaño et al., 2022). MSA subsampling and clustering (i.e., varying the input) in conjunction with AlphaFold2 has recently been shown to reveal alternate conformations (Wayment-Steele et al., 2022; Stein & Mchaourab, 2022; Del Alamo et al., 2022), but the generality and reliability of these techniques remains unclear.

Diffusion models learn an iterative, stochastic generative process that transforms samples from a simple prior to the data distribution. This generative process is trained to be the reverse of a *forward* diffusion transforming the data to the prior (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021). To obtain the generative process from the forward process, it is necessary and sufficient to learn a *score model* to approximate $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ for all values of diffusion time t; we refer to Song et al. (2021) for a more comprehensive methodological overview.

The flexible formulation and strong performance of diffusion models have made them increasingly popular in generative machine learning. While such models have traditionally used isotropic Gaussian noise as the forward process, applications for molecule structure have increasingly featured non-isotropic or non-Euclidean processes that exploit the reduced degrees of freedom and chemical priors in molecular structure (Jing et al., 2022; Ingraham et al., 2022). Our work proceeds in a similar spirit and seeks to define a suitable diffusion process for protein structure prediction.

As generative models, diffusion models have been natural choices for inverse (design) problems. However, they have also been applied productively to forward problems such as ligand-protein docking (Corso et al., 2022) and molecular simulation (Wu et al., 2022a). In particular, diffusion models operating over internal coordinates hold state-of-the-art performance on small-molecule *conformer generation* (Jing et al., 2022). Generative protein structure prediction can be regarded as the macromolecular analogue of conformer generation; however, the significantly larger molecular graphs in protein structure call for different considerations and formulations.

Protein structure diffusion. Several works have fruitfully applied diffusion modeling to the broadly defined task of *protein structure design*. Early works formulated variations of isotropic Euclidean diffusion of residue coordinates (Anand & Achim, 2022; Trippe et al., 2022) or backbone dihedral angles (Wu et al., 2022b). Later works demonstrated impressive experimental results (Watson et al., 2022) and programmability (Ingraham et al., 2022). On the other hand, there have been significantly fewer diffusion models developed for *forward* problems involving protein structures (i.e., where the protein sequence is known) (Qiao et al., 2022; Nakata et al., 2022). Both of these address the task of flexible protein-ligand docking, but have been limited by a dependence on contact maps and by protein size, respectively.

3 METHOD

3.1 HARMONIC DIFFUSION

Consider a structure graph $G = (\mathcal{V}, \mathcal{E})$ embedded in 3D space with coordinates $\mathbf{x} \in \mathbb{R}^{3n}$, where $n = |\mathcal{V}|$. When G represents a protein with a specific sequence, the *generative protein structure prediction* problem can be framed as learning G-dependent probability distributions $p_G(\mathbf{x})$. We now consider diffusion modeling of $p_G(\mathbf{x})$ under a forward diffusion process $d\mathbf{x} = -\frac{1}{2}\mathbf{H}\mathbf{x}\,dt + d\mathbf{w}$ where $\mathbf{H}$ is symmetric positive semi-definite.

Naively, one may choose $\mathbf{H}$ to be proportional to the identity (diffusing to an isotropic Gaussian), as is universally done for images and previously done for molecular conformer generation (Xu et al., 2021). However, such a diffusion does not take into account the chemical graph structure and quickly disassembles the molecule into highly implausible states. Instead, we observe that a choice of $\mathbf{H}$ corresponds to a choice of an arbitrary Gaussian *stationary distribution* of the diffusion:

$$\lim_{t\to\infty} p(\mathbf{x}_t) \propto \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x}\right) \tag{1}$$

We can re-interpret this distribution as the Boltzmann distribution $p(\mathbf{x}) \propto e^{-E(\mathbf{x})}$ of an arbitrary *quadratic potential* $E(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x}$. Similarly, we may re-interpret the forward diffusion as the Brownian motion of a particle under the same time-independent quadratic potential: $d\mathbf{x} = -\frac{1}{2}\nabla_{\mathbf{x}}E(\mathbf{x})\,dt + d\mathbf{w}$. In the physics literature, such motion is known as *overdamped Langevin* or *Brownian dynamics* (Erban, 2014) and is known to converge to the Boltzmann distribution of the potential, consistent with the formulation here.

This Brownian dynamics perspective on forward diffusions provides clear guidance on how to choose the drift term $\mathbf{H}$: we choose it so that undesired, chemically implausible structures have high energy $E(\mathbf{x})$. Conceptually, this accomplishes two main objectives: (1) samples from the prior distribution are automatically consistent with the encoded chemical constraints; and (2) the forward (and later, reverse) diffusions maintain these constraints such that highly implausible structures are never reached. In *harmonic diffusion*, we choose $E(\mathbf{x})$ as the sum of quadratic or *harmonic* constraints for each edge in $\mathcal{E}$, meaning:

$$E(\mathbf{x}) = \frac{\alpha}{2}\sum_{(i,j)\in\mathcal{E}} ||\mathbf{x}_i - \mathbf{x}_j||^2 \tag{2}$$

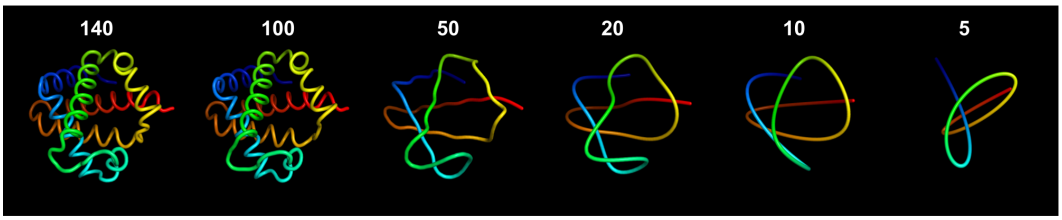

Figure 2: Eigenmode projections of the backbone structure of PDB ID `1a3n.A` onto progressively smaller numbers of modes. Due to symmetry, each mode is 3-fold degenerate. The full structure is described with 140 unique modes since the protein length is 141 (the last set of modes corresponds to translations). For ease of comparison, the structure is projected without perturbation; in the actual forward diffusion, noise is also injected into the remaining modes.

Here, $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^3$ are the coordinates of the ith and jth nodes and $\alpha > 0$ is interpreted as the strength of the edge. This potential is quadratic in the coordinates $\mathbf{x}$ and therefore can be written in the form $\frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x}$, where $\mathbf{H}$ depends on the graph. Intuitively, this potential and the consequent drift term $\mathbf{H}\mathbf{x}$ constrain adjacent nodes to be nearby in 3D space, resolving the most noticeable shortcoming (i.e., molecular disassembly) of previous isotropic diffusions.

For protein structures, one option is to define $\mathcal{V}$ to be the heavy atoms, $\mathcal{E}$ the set of bonds between them, and construct $\mathbf{H}$ based on a harmonic potential defined using those bonds. However, in this work, we focus on sampling residue-level protein structures, in which $\mathcal{V}$ is the set of residues and $\mathcal{E}$ the edges connecting neighboring residues. That is, a protein with m residues is represented by a line graph G of length m, in which $\mathbf{x}$ represents the coordinates of the alpha carbons. To construct the harmonic potential $\mathbf{H}$, we set $\alpha = 3/3.8^2\,\text{Å}^{-2}$ to enforce a RMS distance of 3.8 Å between adjacent alpha carbons (Chakraborty et al., 2013).

3.2 Eigenmode Projections

The harmonic potential describes a forward SDE which can be used to train a score model and reversed via the Euler-Maruyama approach as described in Song et al. (2021). However, the resulting reverse SDE is very stiff and requires a large number of reverse diffusion steps. To understand and solve this issue, we now study the behavior of this forward diffusion in more detail and propose an effective diffusion projection scheme.

Let $\mathbf{H}$ be decomposed as $\mathbf{H} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^T$ for orthogonal $\mathbf{P}$ and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \dots \lambda_{3n})$ all nonnegative. Drawing an analogy with normal mode analysis in mechanics, we call eigenvectors of $\mathbf{H}$ (i.e. columns of $\mathbf{P}$) the *normal modes* or *eigenmodes* of the system, λ_i the *strength* of the modes, and $\mathbf{y} \equiv \mathbf{P}^T\mathbf{x}$ the coordinates along these modes. The diffusion kernel $p_{t|0}$ and stationary distribution p_∞ then both become uncorrelated (albeit nonspherical) Gaussians along the normal mode coordinates $\mathbf{y}$. The KL divergence between the perturbation kernel and the stationary distribution can then be decomposed as the sum of divergences in the coordinates along each mode:

$$D_{\text{KL}}(p_{t|0}||p_\infty) = \sum_{i=1}^{3n} \left[e^{-\lambda_i t}\left(E_i - \frac{1}{2}\right) - \frac{1}{2}\log\left(1 - e^{-\lambda_i t}\right)\right] \tag{3}$$

where $E_i \equiv \lambda_i(\mathbf{y}_0)_i^2/2$ is the initial energy in the ith mode.

The above expression fully describes the convergence of the forward diffusion towards the stationary distribution. In particular, the divergence along each mode decays with rate constant λ_i, which can vary by many orders of magnitude for different modes. Thus, the diffusion kernel will quickly converge to the stationary distribution along strong (i.e., large λ_i) modes, but will take much longer to converge along weak modes. This is analogous to the Born-Oppenheimer approximation in physics, where certain degrees of freedom equilibrate so rapidly that they are effectively stationary on timescales relevant to other degrees of freedom. Indeed, we can characterize which degrees of freedom are "active" at any given diffusion time, starting with $3n$ at $t = 0$ and reaching zero when we have converged to the joint stationary distribution. Nevertheless, regardless of the number of active modes, the stiffnes $\lambda_{\text{max}}/\lambda_{\text{min}}$ of the SDE remains very large, necessitating small step sizes for the entire duration of the sampled trajectory.

We now propose that, in both the forward and reverse diffusions, the structure is *projected down to* only the modes that are still active at the given point in the diffusion (Figure 2). That is, we set $(\mathbf{y}_t)_i = 0$ for all modes i such that $\lambda_i t > \tau$ for some threshold τ, thereby projecting the structure onto the subspace spanned by the remaining modes for which $\lambda_i t \leq \tau$. By construction, this reduces the stiffness $\lambda_{\text{max}}/\lambda_{\text{min}}$ of the SDE. At inference time, we start by sampling $(\mathbf{y}_T)_i \sim p_\infty$ from the stationary distribution only for the k weakest (smallest eigenvalue) modes, where k is a hyperparameter. Then, during the reverse diffusion, we successively sample $(\mathbf{y}_t)_j \sim p_\infty$ from the stationary distribution of the jth eigenmode just as it is about to become active, i.e., at $t_j = \tau/\lambda_j$. Similar to cascaded diffusion modeling of images (Ho et al., 2022), this process induces a cascading-resolution generative process of the molecular structure, with the global structure being determined before local details (Figure 1). Altogether, this procedure enables the sampling of large macromolecular structures with 100 or fewer Euler solver steps—significantly fewer than the 5000 steps required by Xu et al. (2021) for much smaller molecules.

3.3 Score Model Architecture

Next, we develop a score model architecture $\mathbf{s}_\theta(\mathbf{x}, t)$ suitable for learning $\nabla_{\mathbf{x}} \log p(\mathbf{x}_t)$. We construct graph neural networks with tensor-product message-passing layers in `e3nn` (Thomas et al., 2018; Geiger & Smidt, 2022). The message-passing graph is constructed as a complete graph of size m, such that message-passing occurs between all pairs of residues. In addition to the residue coordinates, the score network is provided with node and edge features obtained by running OmegaFold on the input sequence and extracting the node and pair embeddings from the Geoformer stack. In this sense, our score model can be viewed as substituting for the deterministic structure module which usually operates on the Geoformer outputs. The score model is $SE(3)$-equivariant; since the stationary density (Equation 2) is $SE(3)$-invariant, this ensures that the final model density will also be $SE(3)$-invariant (Xu et al., 2021).

3.4 Ranking Sampled Structures

The trained EigenFold score model can sample multiple structures for a given protein sequence; however, it is often desirable to identify a single best structure prediction. For this purpose, we compute approximate lower bounds to model log-likelihoods for all sampled structures and select the one with the highest lower bound—i.e., the one most likely to have been sampled by EigenFold. Specifically, let $t_1, \ldots, t_K = T$ be a discretization of the forward and reverse SDEs. Then for any structure $\mathbf{x}$ (sampled or otherwise), we can compute a lower bound to the model log-likelihood as follows (Sohl-Dickstein et al., 2015):

$$\log p(\mathbf{x}_0) \geq \mathbb{E}_{\mathbf{x}_{t_1 \ldots t_K} \sim q} \log \left[p_\infty(\mathbf{x}_t) \prod_{k=1}^{K} \frac{p(\mathbf{x}_{t_{k-1}} \mid \mathbf{x}_{t_k})}{q(\mathbf{x}_{t_k} \mid \mathbf{x}_{t_{k-1}})} \right] \tag{4}$$

Thus, by sampling a forward trajectory $\mathbf{x}_{t_1} \ldots \mathbf{x}_{t_K}$ starting from any given $\mathbf{x}_0$, we obtain a Monte-Carlo estimate of the evidence lower-bound (ELBO) for that structure. Notably, this insight removes the need to train a separate model to rank samples, as previously done by Corso et al. (2022).

4 Experiments

We train EigenFold on all structures deposited in the PDB on or before Apr 30, 2020 and validate on structures deposited between May 1, 2020 and Nov 30, 2020. To reduce training time, we train (and validate) only on structures with residue lengths between 20 and 256, for a total of 230,520 (14,128) training (validation) structures. To assess single-structure prediction accuracy, we make predictions for all CAMEO targets released between Aug 1, 2022 and Oct 31, 2022. After excluding targets with 750 or more residues for which OmegaFold embeddings could not be generated, the final test set consists of 183 CAMEO targets.

To assess the ability of EigenFold to model conformational diversity, we collect and filter two datasets from previous works. First, we collect 77 pairs of PDB IDs corresponding to fold-switching proteins from Chakravarty & Porter (2022). Second, we collect 90 pairs of apo/holo PDB IDs corresponding to ligand-induced conformational change from Saldaño et al. (2022). For each dataset,

Table 1: Single-structure prediction accuracy of EIGENFOLD and baseline methods on CAMEO targets under 750 residues from Aug 1–Oct 31, 2022. All metrics are reported as mean / median.

	$\text{RMSD}_{C\alpha}$ ↓	TMScore ↑	GDT-TS ↑	$\text{lDDT}_{C\alpha}$ ↑
ALPHAFOLD2	3.30 / 1.64	0.87 / 0.95	0.86 / 0.91	0.90 / 0.93
ESMFOLD	3.99 / 2.03	0.85 / 0.93	0.83 / 0.88	0.87 / 0.90
OMEGAFOLD	5.26 / 2.62	0.80 / 0.89	0.77 / 0.84	0.83 / 0.89
ROSETTAFOLD	5.72 / 3.17	0.77 / 0.84	0.71 / 0.75	0.79 / 0.82
EIGENFOLD	7.37 / 3.50	0.75 / 0.84	0.71 / 0.79	0.78 / 0.85

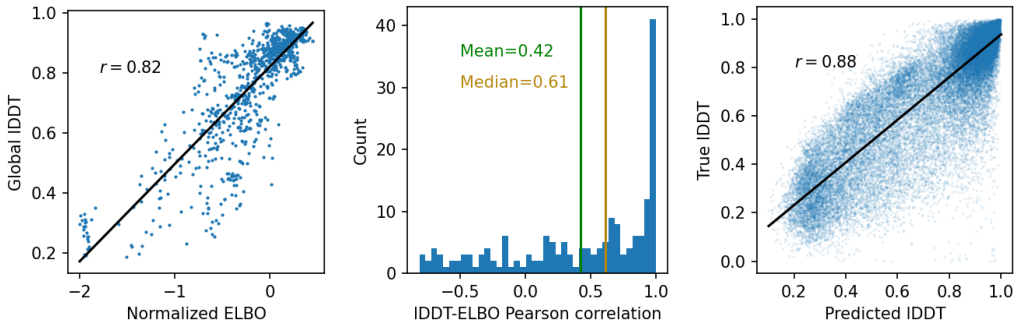

Figure 3: *Left:* Scatterplot of normalized ELBO (i.e., divided by $(3n-1)$ where n is the protein length) v.s. global lDDT. *Center:* Histogram of per-target Pearson correlations between the normalized ELBO and lDDT. *Right:* Scatterplot of predicted lDDT (average lDDT between sampled pairs) and true lDDT for all residues in the CAMEO test set.

we sample structures using the SEQRES entries of the PDB IDs with the shorter sequence or designated as "Apo," respectively. Both sets are filtered to remove pairs where the two sequences differ significantly in length or where the sequence used for sampling is 750 residues or longer.

4.1 SINGLE STRUCTURE PREDICTION

For each of the CAMEO test targets, we sample five structures from EIGENFOLD and compute the approximate ELBO for each. The top-ranked structure is considered the final prediction and is compared to the ground truth via standard metrics. Table 1 compares the quality of these predictions relative to established methods RoseTTAFold, OmegaFold, ESMFold, and AlphaFold2. EIGENFOLD samples are comparable in quality to those from RoseTTAFold, but fall short of the best results from AlphaFold2 and ESMFold. The approximate structural ELBO is well-correlated with absolute structural accuracy and thus serves as a good measure of model confidence (Figure 3 (*left*, *center*)). In particular, the positive per-target correlations (i.e., correlating only within the five samples for each target) on most targets justifies the use of the approximate ELBO as a means of ranking samples within a target.

Next, we find that the variability of the sampled ensemble is highly informative about the model error and can be interpreted as revealing model uncertainty. To measure this variability, we define, for any global measure of structural deviation f, an f-variation:

$$f_{\text{var}} = \mathbb{E}_{\mathbf{y}_1, \mathbf{y}_2 \sim \text{EIGENFOLD}} \left[f(\mathbf{y}_1, \mathbf{y}_2) \right] \tag{5}$$

where the expectation is approximated with the five samples. For example, if $f = \text{TM}$, then TM_{var} measures the diversity of the sampled ensemble in terms of the average pairwise TMScore. We compare this quantity with f_{exp}, the average f of the sampled structures relative to the ground truth. As illustrated in Table 2, these measures are highly correlated across the CAMEO targets; thus, when the ground truth structure is unknown, the f_{var} computed from sampled structures can be interpreted as a well-calibrated *prediction* of f_{exp} relative to the ground-truth.

Table 2: Pearson correlations between f_{var} and f_{exp} for various metrics f of structural deviation. For residue-level or pairwise metrics, we compute a global correlation as well as a correlation for each target, reported as mean / median.

	Global	Per-Target
Protein-level metrics		
TM	0.88	–
GDT-TS	0.90	–
$\text{RMSD}_{\text{C}\alpha}$	0.85	–
$\text{lDDT}_{\text{C}\alpha}$	0.86	–
Residue-level metrics		
$\text{lDDT}_{\text{C}\alpha}$	0.88	0.73 / 0.81
Aligned position error	0.80	0.68 / 0.75
Pairwise metrics		
Distance error	0.75	0.69 / 0.72

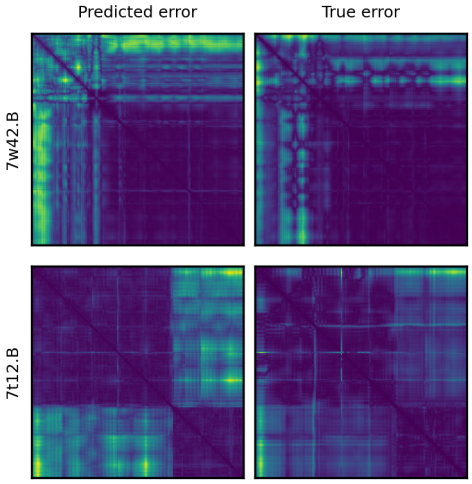

Figure 4: Predicted (left) and actual (right) pairwise distance errors for two CAMEO targets. Domains and inter-domain uncertainty are visibly and correctly predicted.

The correlation between f_{var} and f_{exp} also holds for residue-level and pairwise accuracy metrics. In particular, we compute an expected lDDT for any given residue between pairs of sampled structures, and find that this is well-correlated with the lDDT for that residue between sampled structures and the ground truth (Figure 3 (*right*) and Table 2). In this manner, we have access to a well-calibrated pLDDT for each residue, similar to the outputs of the confidence heads of existing structure prediction methods. Unlike existing methods, however, we can easily apply this framework to predict arbitrary error metrics without a bespoke confidence head. For example, in Table 2, we illustrate that the aligned residue position error (i.e., error in residue position after RMSD alignment) and absolute pairwise distance error can be similarly predicted. Furthermore, the residue-level and pairwise metrics have high per-target correlations, indicating that they can be used to interpret the relative model confidence in different parts of the protein and their spatial relationships (Figure 4).

4.2 Conformational Diversity

To assess how well EigenFold can model protein conformational diversity, we sample five structures for each of the fold-switching and apo/holo pairs, and investigate the following questions:

1. How well do the predicted structures model both conformations on a global level?
2. Is the level of sample diversity predictive of the magnitude of the conformational change?
3. Is the residue-level variation among samples predictive of true residue flexibility?

To answer the first question, we define an ensemble TM-score as follows:

$$\text{TM}_{\text{ens}}(\mathbf{x}_1, \mathbf{x}_2, \{\mathbf{y}_i\}) = \frac{1}{2}\left[\max_i \text{TM}(\mathbf{y}_i, \mathbf{x}_1) + \max_i \text{TM}(\mathbf{y}_i, \mathbf{x}_2)\right] \tag{6}$$

where $\mathbf{x}_1, \mathbf{x}_2$ are the two ground truth structures and $\{\mathbf{y}_i\}$ are the EigenFold samples. This measures how well the sampled structures cover both ground-truth conformational states. Figure 5 (*left*) illustrates that EigenFold samples generally are a poor model of the two ground truth conformations, in the sense that they offer no improvement over a hypothetical baseline that always predicts a single conformation. Furthermore, the samples—even if different from each other—are generally very similar in terms of their deviation from the two ground truth structures, and tend to either heavily favor a single structure or model both structures relatively poorly.

Next, to address the second and third questions, for each pair we compute the TM_{var} (average pairwise TMScore) using the five sampled structures as a measure of sample diversity, and compare

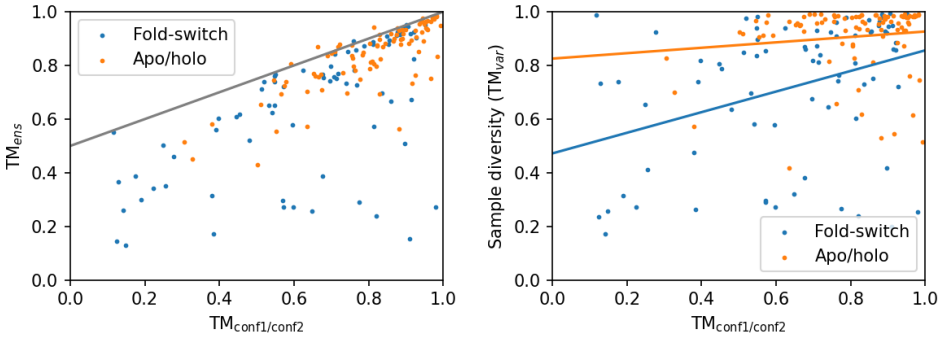

Figure 5: *Left:* Coverage of the two ground truth conformations (TM_{ens}) plotted against the TM-score between the two conformations (i.e., the most dissimilar pairs are on the left). The gray line indicates a hypothetical baseline which always perfectly predicts one of the two conformations. *Right:* Scatterplot of the TM-score between the two true conformations versus the average TM-score between sampled conformations.

Table 3: Pearson correlations between sample diversity and ground-truth diversity, measured in terms of TMScore (protein-level metric) or residue flexibility (i.e., absolute deviation after RMSD alignment). For the latter, we report global correlations and mean/median per-target correlations.

	Fold-switch	Apo/Holo
TM	0.36	0.12
Residue flexibility (global)	0.23	0.13
Residue flexibility (per-target)	0.28 / 0.26	0.41 / 0.40

with the TMScore between the two ground truth conformations ($TM_{conf1/conf2}$). As shown in Figure 5 (*right*) and Table 3, TM_{var} and $TM_{conf1/conf2}$ are moderately correlated. At the residue-level, we examine whether the flexibility of a residue within the sampled structures (i.e., average positional difference post-RMSD alignment) is predictive of the true flexibility of that residue under the conformational change. Table 3 shows that both the global and per-target correlations are also positive but moderate. Altogether, while the conformational diversity and residue-level flexibility within sampled EIGENFOLD structures are somewhat informative of underlying conformational changes, the magnitude or residue-level localization of such changes are not modelled to high accuracy.

5 CONCLUSION

In this work, we developed EIGENFOLD, the first diffusion generative model for predicting protein structures from a fixed protein sequence. In doing so, we built the first bridges between the rapidly advancing fields of diffusion modeling for molecules and modern structure prediction frameworks. Our model matches the performance of established methods on CAMEO targets and reveals model uncertainty via an ensemble of structural predictions, enabling customizeable ways to estimate and understand prediction error. We anticipate that these capabilities will prove important in downstream applications in which the relevant error could otherwise not be estimated using existing methods.

Although a generative modeling paradigm opens the door towards modeling conformational diversity and change, we find that EIGENFOLD is currently unable to model these aspects of protein structure with high accuracy. Instead, it appears that the distribution of predicted structures is largely reflective of model uncertainty rather than underlying (i.e., aleatoric) uncertainty arising from flexibility. There may be several reasons for this gap: the model may not be accurate enough to resolve conformational changes of small magnitude; the training set consists largely of crystal structures that show little conformational flexibility; and the use of OmegaFold embeddings—without fine-tuning—may inject a bias towards the single-structure output of OmegaFold. Improving these aspects, with more tailored training settings or more expressive end-to-end score network architectures, could serve as promising directions of future work.

ACKNOWLEDGEMENTS

We thank Hannes Stärk, Jason Yim, Jeremy Wohlwend, and Rohit Singh for helpful feedback and discussions. This work was supported by the NIH NIGMS under grant #1R35GM141861 and a Department of Energy Computational Science Graduate Fellowship.

REFERENCES

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

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

Sandeep Chakraborty, Ravindra Venkatramani, Basuthkar J Rao, Bjarni Asgeirsson, and Abhaya M Dandekar. Protein structure quality assessment based on the distance profiles of consecutive backbone cα atoms. *F1000Research*, 2:Article–ID, 2013.

Devlina Chakravarty and Lauren L Porter. Alphafold2 fails to predict protein fold switching. *Protein Science*, 31(6):e4353, 2022.

Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking. *arXiv preprint arXiv:2210.01776*, 2022.

Diego Del Alamo, Davide Sala, Hassane S Mchaourab, and Jens Meiler. Sampling alternative conformational states of transporters and receptors with alphafold2. *Elife*, 11:e75751, 2022.

Radek Erban. From molecular dynamics to brownian dynamics. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 470(2167):20140036, 2014.

Mario Geiger and Tess Smidt. e3nn: Euclidean neural networks. *arXiv preprint arXiv:2207.09453*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23: 47–1, 2022.

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

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

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

Thomas J Lane. Protein structure prediction has reached the single-structure frontier. *Nature Methods*, pp. 1–4, 2023.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, and Alexander Rives. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *arXiv*, 2022.

Shuya Nakata, Yoshiharu Mori, and Shigenori Tanaka. End-to-end protein-ligand complex structure generation with diffusion-based generative models. *bioRxiv*, 2022.

Zhuoran Qiao, Weili Nie, Arash Vahdat, Thomas F Miller III, and Anima Anandkumar. Dynamic-backbone protein-ligand structure prediction with multiscale generative diffusion models. *arXiv preprint arXiv:2209.15171*, 2022.

Tadeo Saldaño, Nahuel Escobedo, Julia Marchetti, Diego Javier Zea, Juan Mac Donagh, Ana Julia Velez Rueda, Eduardo Gonik, Agustina García Melani, Julieta Novomisky Nechcoff, Martín N Salas, et al. Impact of protein conformational diversity on alphafold predictions. *Bioinformatics*, 38(10):2742–2748, 2022.

Arne Schneuing, Yuanqi Du, Charles Harris, Arian Jamasb, Ilia Igashov, Weitao Du, Tom Blundell, Pietro Lió, Carla Gomes, Max Welling, et al. Structure-based drug design with equivariant diffusion models. *arXiv preprint arXiv:2210.13695*, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Richard A Stein and Hassane S Mchaourab. Speach_af: Sampling protein ensembles and conformational heterogeneity with alphafold2. *PLOS Computational Biology*, 18(8):e1010483, 2022.

Simon J Teague. Implications of protein flexibility for drug discovery. *Nature reviews Drug discovery*, 2(7):527–541, 2003.

Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint*, 2018.

Brian L Trippe, Jason Yim, Doug Tischer, Tamara Broderick, David Baker, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. *arXiv preprint arXiv:2206.04119*, 2022.

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

Hannah K Wayment-Steele, Sergey Ovchinnikov, Lucy Colwell, and Dorothee Kern. Prediction of multiple conformational states by combining sequence clustering with alphafold2. *bioRxiv*, 2022.

Peter E Wright and H Jane Dyson. Intrinsically disordered proteins in cellular signalling and regulation. *Nature reviews Molecular cell biology*, 16(1):18–29, 2015.

Fang Wu, Qiang Zhang, Xurui Jin, Yinghui Jiang, and Stan Z Li. A score-based geometric model for molecular dynamics simulations. *arXiv preprint arXiv:2204.08672*, 2022a.

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

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

Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2021.