DiffPIE: Guiding Deep Generative Models to Explore Protein Conformations under External Interactions

Yanbin Wang and Ming Chen*

Purdue University, Department of Chemistry, West Lafayette, 47906, USA

E-mail: chen4116@purdue.edu

Abstract

In recent years, many foundation generative models have been developed to predict structures of molecules and materials. Although these foundation models have achieved great success, it is challenging to collect enough data to train foundation generative models. One such example is to predict protein conformations with protein-environmental interactions (PEI), such as interactions introduced by organic linkers or material surfaces. We propose a physics-guided route to extrapolate foundation models beyond their training domain. Our method couples a pretrained deep generative model with explicit, physics-based interaction potentials for PEI, steering sampling toward conformations consistent with external constraints without any retraining or fine-tuning. We demonstrate accurate and efficient conformation prediction of (i) cyclic peptide with organic linkers and (ii) peptide adsorbed on the gold surface. The generated structures serve as high-quality initial conditions for downstream simulations, providing a general, systematic approach to extend foundation models to proteins under system-specific environmental interactions.

Introduction

In recent years, foundation machine learning (ML) models in the chemistry community have attracted much attention. These models aim to solve fundamental questions in chemistry with broad applications derived from the high generalizability of deep neural networks.[1] Many of these models have been developed to address challenging problems that include, but do not limit to, property prediction,[2–7] structure generations,[8–16] force field parameterization,[17–20] and chemical reaction prediction.[21,22] One type of these foundation models, named generative models, aims to generate atomic structures of small molecules,[23–30] biomolecules,[16,31–37] and materials.[38–44] Conventionally, structure generation has been a highly challenging task that can only be solved by optimization-based methods[45–54] or physics-based simulation approaches,[55–62] while the development of the foundation generative models provides highly efficient solutions to the problem of structure generation. These developments significantly advance research fields such as drug development,[24,25,30] protein structure prediction,[16,31–37] and material designs.[38–44] However, it is questionable whether a foundation generative model is capable of efficiently exploring individual chemical spaces of small organic molecules, materials, biomolecules, and others. A simple estimation of the number of organic molecules made by C, H, O, N, S, and with a molecular weight less than 500, is about 10^{60}.[63] The number of possible Wyckoff sequences, a way to classify crystal structures, is approximately 10^{18}.[64] An estimate of proteins that possibly have been present since the origin of life is about 10^{21}.[65] The huge size of an individual chemical space already poses significant challenges in collecting high-quality data to train foundation generative models to explore the space. In fact, the success of protein foundation models is highly dependent on the accumulated high-quality protein structure data over decades.

It is even more challenging to generate the atomic structures of chemical systems in interactions with the environment. Using proteins as examples, the protein-environment interaction (PEI) includes possible covalent and / or noncovalent interactions with organic molecules,[66] materials,[67,68] soluble polymers,[69] and membranes. The generation of protein-

with-PEI conformations is crucial to understanding protein functions in a wide variety of applications in biosensing, antifouling, drug delivery, and drug design.[70–72] Although highly important, methods for predicting protein-with-PEI conformations are limited. Although experimental methods such as X-ray crystallography,[73] NMR,[74] and cryo-EM[75] have achieved great success in determining the atomistic structures of proteins, the direct application of these structural determination methods to protein with PEI is challenging. These limited examples of structural studies focus on protein-protein complexes,[76] protein-nucleic acid complexes,[77,78] protein-drug complexes,[79] and cyclic peptides.[80] However, resolving protein structures with many other PEI types, such as protein-material interactions, is a challenge. More often, protein conformations with these PEI are measured using ensemble-based methods such as IR, circular dichroism, and SAXS where obtaining atomic structures is still an open question.[81–83] In addition to experimental approaches, theoretical modeling to resolve protein conformations is also challenging. Generic approaches to exploring proteins conformations under external interactions include all-atom (AA) molecular dynamic (MD) simulations and coarse-grained MD simulations.[84–89] All-atom MD simulations are often limited by simulation time scale and system length scale. Although CG models are accessible to larger system sizes and longer simulation timescales, improving the accuracy of CG models is still an open problem.[90,91] The challenges of determining protein-with-PEI conformations become significant obstacles in collecting data to train a foundational generative model of proteins with PEI. Therefore, current foundation models can only predict protein structures with limited classes of PEIs such as protein-biomolecule complex or protein-drug complex.[92,93] Thus, guiding existing deep learning models to explore protein conformations under external interactions represents a promising avenue for fully utilizing the potential of foundation models.

Recent studies have begun to explore similar directions,[94] particularly through guided sampling using diffusion models.[95–97] Diffusion models learn by progressively adding noise to data points and can generate new data through a reverse diffusion process.[98,99] In particular,

the reverse diffusion process, either its drift or its diffusion term, can be modified to manipulate the data points generated.[100,101] Examples of such modifications include incorporating restraint potentials to ensure that all-atom structures generated are consistent with experimental measurements,[101] enforcing the generation of structures containing specific motifs,[102] and guaranteeing correct stereochemistry in protein–ligand complexes.[25] Some studies have also used AA force fields to refine protein structure generation.[95,97] Building upon these ideas, we present a systematic framework, termed the "Diffusion Model for Proteins Interacting with the Environment (DiffPIE)", to integrate PEIs into an existing diffusion model. The core idea is that a diffusion model implicitly learns a "statistical potential" for protein structures without PEIs. By adding explicit physical interactions between the protein and the surrounding chemical species to this statistical potential, we construct a new effective potential energy landscape. Incorporating this modified potential into the diffusion process allows the model to generate protein structures that account for external interactions.

We will demonstrate the success of the proposed framework with two representative examples: a peptide denatured on a gold surface and a peptide cyclized with an organic linker. In both cases, we show that PEIs can be effectively modeled using MD simulations of fragments of the protein, rather than simulating the entire protein with its environment. Furthermore, we demonstrate that incorporating physics-based potentials effectively guides the structure generation process. Short plain MD simulations are then used to relax the generated structures, illustrating that the DiffPIE model can identify high-quality initial structures for proteins with PEIs.

Results

DiffPIE: Guiding Protein Structure Generation with Protein-environment Interaction

Assume that a protein structure ensemble can be generated using a pre-trained foundation score-based diffusion model, which learns through a forward noise-adding process described by:

$$d\mathbf{x} = f(\mathbf{x}, t)dt + g(t)d\mathbf{w},\ t \in [0, 1], \tag{1}$$

and generates new structures through a reverse denoising process:

$$d\mathbf{x} = [f(\mathbf{x}, t) - g^2(t)\mathbf{s}(\mathbf{x}, t)]dt + g(t)d\bar{\mathbf{w}}. \tag{2}$$

To simplify the discussion, we consider $\mathbf{x}$ in Eqs.(1) and (2) as representing atomic coordinates. As we will demonstrate later, DiffPIE can also operate with diffusion models that hybridize Cartesian coordinates and generalized coordinates. Here, $f(\mathbf{x}, t)$ and $g(t)$ define the noise schedule, while $\mathbf{s}(\mathbf{x}, t)$ is trained to approximate $\nabla_{\mathbf{x}} \log P_t(\mathbf{x})$, where $P_t(\mathbf{x})$ is the time-dependent marginal distribution in Eq.(2). The resulting generated structures follow a distribution $p(\mathbf{x})$ learned from the training dataset, which aims to approximate the Boltzmann distribution of a protein without PEI, denoted as $p_{\mathrm{B}}^{\mathrm{NPIE}}(\mathbf{x})$. However, as illustrated in Fig. 1(c), $p(\mathbf{x})$ can differ significantly from the true Boltzmann distribution of a protein under PEI, $p_{\mathrm{B}}^{\mathrm{PIE}}(\mathbf{x})$. Consequently, it is challenging to study the thermodynamics of $p_{\mathrm{B}}^{\mathrm{PIE}}(\mathbf{x})$ using structures generated solely from $p(\mathbf{x})$.

To address this, DiffPIE generates a biased distribution $p_{\mathrm{bias}}(\mathbf{x})$ that approximates $p_{\mathrm{B}}^{\mathrm{PIE}}(\mathbf{x})$, enabling rapid relaxation toward an ensemble representative of the true protein-environment system. DiffPIE achieves this by modifying the reverse diffusion process:

$$d\mathbf{x} = [f(\mathbf{x}, t) - g^2(t)(\mathbf{s}(\mathbf{x}, t) + \lambda(t)\mathbf{F}(\mathbf{x}, t)]dt + g(t)d\bar{\mathbf{w}}. \tag{3}$$

Here, $\mathbf{F}(\mathbf{x}, t)$ is a biasing drift term that encodes protein-environment coupling, and $\lambda(t)$, which changes with t, controls the strength y of the applied bias. A larger λ leads to stronger bias away from $p(\mathbf{x})$. The details of designing $\lambda(t)$ will be discussed in the "Method" section. In DiffPIE, $\mathbf{F}(\mathbf{x}, t)$ is constructed by propagating a biasing force $\mathbf{F}_0(\mathbf{x})$, modeled by fragmentation of the studied protein, back into the diffusion model, following an approach similar to the manifold constraint algorithms.[103,104] Users have flexibility to design $\mathbf{F}(\mathbf{x}, t)$ suitable for a wide range of PEI, as illustrated in Fig. 1(a). Different choices of $\mathbf{F}_0(\mathbf{x})$ allow the generation of protein structures consistent with different PEI scenarios. It is important to emphasize that DiffPIE does not require any retraining or fine-tuning of the foundation diffusion model. Instead, it only modifies the generation process by introducing the biasing drift $\mathbf{F}(\mathbf{x}, t)$ during sampling. The DiffPIE framework is illustrated in Fig.1(b). We begin by designing a customized biasing force, with specific examples provided in following sections. In this study, we use Str2Str[16] as the foundation diffusion model; however, in principle, any diffusion model can be incorporated. A key advantage of Str2Str is its flexibility in controlling the generated ensemble by selecting different seed structures and tuning hyperparameters. From the generated structures, we select representative conformations which can serve as new seeds or inform further hyperparameter adjustments. This iterative process is repeated until no additional new conformations are generated. Finally, the selected structures are subjected to relaxation through either structure minimization or short MD simulations, depending on the application.

Sampling Cyclic Peptide Structures Containing Organic Linkers

P3-F is a cyclic peptide (Fig. 2(a)) that can interrupt interactions between the Nuclear Factor Erythroid-2-related factor (Nrf2) and the E3 ligase Kelch-like EH-associated protein 1 (Keap1) to counteract oxidative stress.[105] The sequence of P3-F is CDPETGECL. Two cysteines are cyclized by divinylpyrimidine with Michael addition between the vinyl group and cysteine (CYS) thiols.[105] Here, we use P3-F as an example to demonstrate how to use

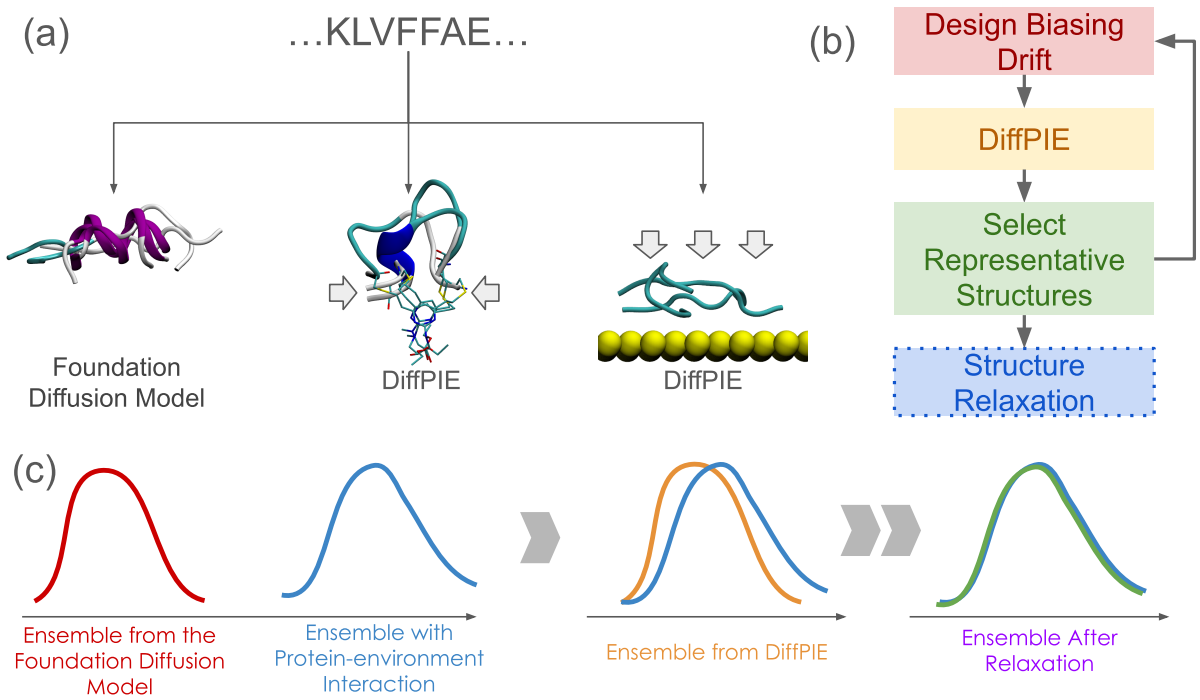


Figure 1: Panel (a) illustrates the concept of biasing protein structure generation using $\mathbf{F}_0(\mathbf{x})$. The left bundle of structures was generated by the foundation diffusion model, Str2Str, with $\mathbf{F}_0(\mathbf{x}) = 0$, corresponding to unbiased generation. The middle bundle shows structures generated with a $\mathbf{F}_0$ designed to mimic the restraint forces imposed by an organic linker. The right bundle presents structures generated with a $\mathbf{F}_0$ simulating adsorption forces, such as those from a surface interaction. Panel (b) summarizes the overall workflow of DiffPIE. Once a biasing force is designed, DiffPIE can be applied without any retraining. The distribution of generated structures can, in turn, inform adjustments to the hyperparameters in DiffPIE or refinements in the design of the biasing force. Although structure relaxation is not the primary focus of this paper, we demonstrate examples using short MD simulations for further refinement. Panel (c) illustrates how DiffPIE alters the ensemble distribution. The ensemble generated by the foundation diffusion model without bias (red) can significantly differ from the true protein ensemble under PEI (blue). By introducing the biasing force through DiffPIE, the generated ensemble distribution (orange) is shifted closer to the true ensemble. Additional MD relaxation can further improve the ensemble (green) to better approximate the true distribution.

DiffPIE to generate peptide structures constrained by organic linkers.Fig. 2(b) demonstrates how to construct $\mathbf{F}_0$. We first used MD to simulate a system of two cysteine dipeptides cross-linked by the linker molecule. The thermodynamic mean force of six distances (labeled with colored dashed lines in Fig. 2(b)) has been used as $\mathbf{F}_0$. Fig. 2(c) highlights a one-dimension free energy surface (FES) of the distance between two α-carbons (Cα) in two cysteines. It is clear that the FES is highly anharmonic with multiple minima, suggesting that the conventional manifold constraint method, which uses a harmonic potential to apply restrictions to generated structures, is not applicable in this case. Fig.2(c) also highlights that the Cα–Cα distance in the PDB structure is not a minimum in the FES, suggesting that stable structures in this cyclic peptide should balance the internal energy of the peptide modeled by the foundation diffusion model and the bias energy of the linker modeled by $\mathbf{F}_0$. Fig.2(d) shows the impact of applying a bias drift. Structures generated by Str2Str without $\mathbf{F}_0$, DiffPIE, and a metadynamic[106] simulation of the entire peptide are mapped in the space of AlphaRMSD.[107] Str2Str generated a large number of helix-like structures, while these structures are almost not seen in the metadynamics simulation. DiffPIE significantly reduces the population of helix structures and highlights the conformation with a low AlphaRMSD, which is consistent with the conformations from metadynamics. We also examine the diversity of the generated structures in Fig. 2(e). The background FES of the Uniform Manifold Approximation and Projection (UMAP)[108] CVs was generated after unbiasing the metadynamics simulation.[108] We map the generated structures onto the same UMAP CVs as the scatter points. This clearly demonstrates that DiffPIE is capable of generating structures that visit most of the stable conformations. The PDB structure, denoted by the red point, corresponds to a metastable conformation. Since the PDB structure was resolved for the Keap1–P3-F complex, it indicates that the protein–peptide interaction is the main driving force stabilizing the P3-F PDB structure. Even though the PDB structure is not the most stable conformation in the metadynamics simulation, DiffPIE successfully generates similar structures, highlighting a possible application of DiffPIE in predicting cyclic peptide drug

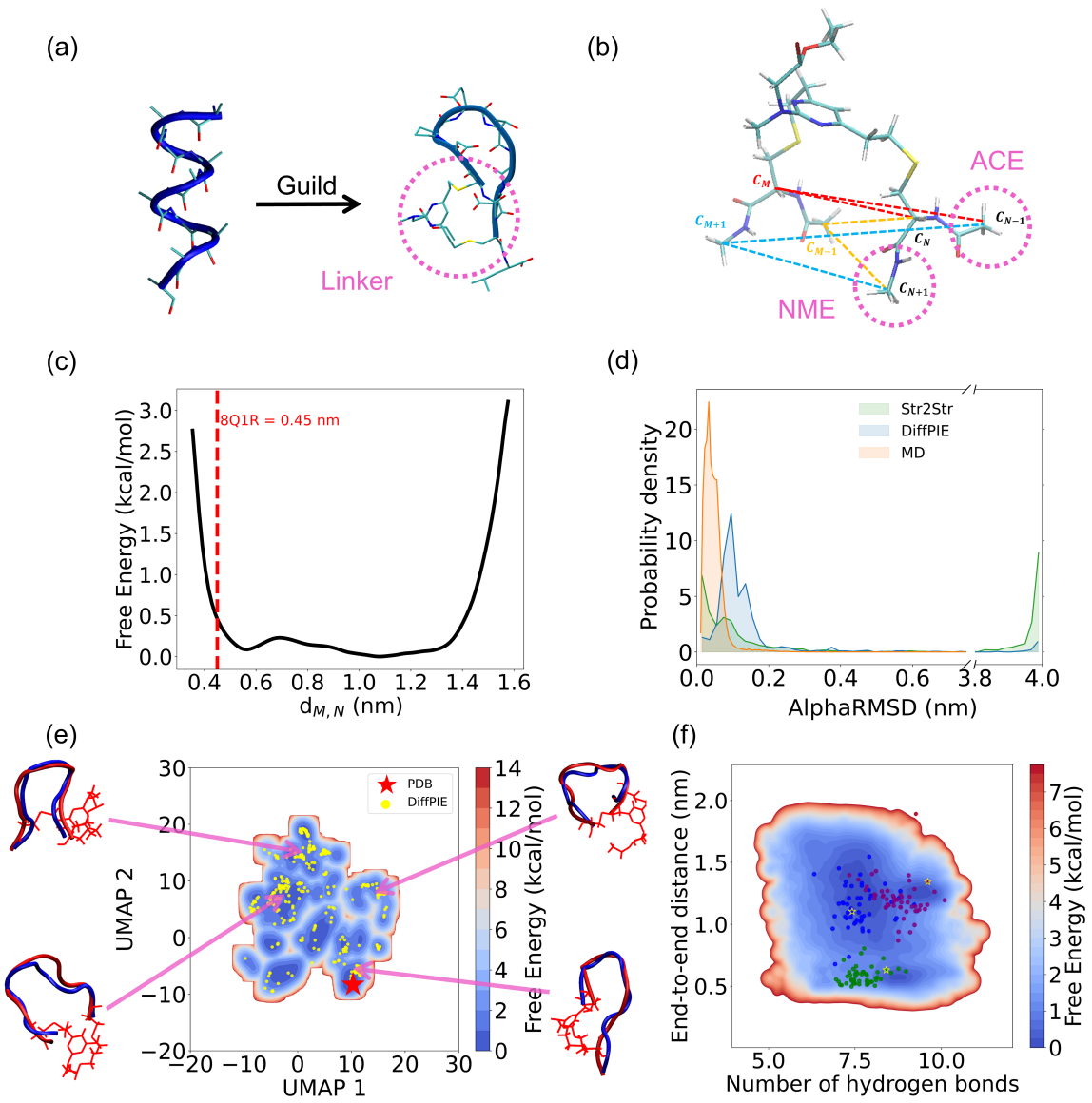


Figure 2: (a) DiffPIE is designed to guild the peptide (P3-F) sampling towards the structure with geometric restraint from linker molecule (PDB ID: 8Q1R). (b) Model molecule for $\mathbf{F}_0$: a staple with two CYS residues capped by acetyl (ACE) and N-methylamide (NME) to form dipeptide-like ends. Six dashed-line distances serve as collective variables (CVs), with mean forces yielding $\mathbf{F}_0$. (c) Free-energy surface (FES) of the Cα–Cα distance between the two CYS in the model, with the red dashed line from the PDB structure. (d) Effect of biasing force on sample distribution, shown as AlphaRMSD for Str2Str, DiffPIE, and metadynamics MD. AlphaRMSD measures the α-helix similarity of generated structures. Small AlphaRMSD means the structure is dissimilar to α-helix while large the structure is similar to α-helix. (f) Background FES from metadynamics configurations mapped via UMAP, with PDB (red star) and DiffPIE samples (yellow) projected onto the same space; representative structures from simulation (red) and DiffPIE (blue) are compared. (g) 10-ns unbiased MD relaxation from low-energy DiffPIE clusters, with initial structures (stars) and trajectories (dots) projected onto an FES of end-to-end distance and backbone hydrogen bonds.

binding.[109] Fig. 2(e) also shows that multiple structures generated by DiffPIE are comparable with stable structures sampled by metadynamics. In addition, we examine the quality of the generated structures and their captivity as initial structures for downstream sampling. First, stable conformations are included in the structural library generated by DiffPIE, as shown in Fig.2(f). Short MD relaxations initiated from the and purple structures stabilized within their respective conformations. In contrast, Fig. 2(f) shows a non-optimal green structure generated by DiffPIE that, although suboptimal, was able to rapidly relax to a similar conformation.

Sampling Peptide Structures on Gold Surface

Gold nanoparticles have been widely used in biochemistry, with applications such as surface plasmon-enhanced spectroscopy[110] and drug delivery.[111] Although protein denaturation is well known to occur on naked gold surfaces, predicting the resulting denatured protein structures remains a challenge.[112,113] In this section, we demonstrate how to use DiffPIE to predict possible structures of proteins that bind to a gold surface. The system we study is amyloid β_{16-22} (Aβ_{16-22}) with the KLVFFAE sequence on the surface of Au(111) (Fig. 3(a)), which has been used as a model system to study the environmental impacts on protein fibrillation.[114] Additionally, the small size of this system allows us to exhaustively sample all possible conformations with metadynamics to validate DiffPIE. The biasing force $F_0(\mathbf{x})$ in this example is the thermodynamic force associated with binding Aβ_{16-22} to the gold surface. The force is approximated by the contributions of each residue. We perform metadynamics simulations of a dipeptide molecule for each type of amino acid in Aβ_{16-22} to generate a free energy surface (FES) based on three distances, d_1, d_2, and d_3, as shown in Fig. 3(b). The detailed description of the construction of $F_0(\mathbf{x})$ will be presented in the "Methods" section. For residue i in Aβ_{16-22}, we generate the force on its Cα, while the Cα–surface distances of residue $(i-1)$ and residue $(i+1)$ match d_1 and d_3, respectively. The design of $F_0(\mathbf{x})$ highlights the importance of the orientation of the residue, as reflected by the different

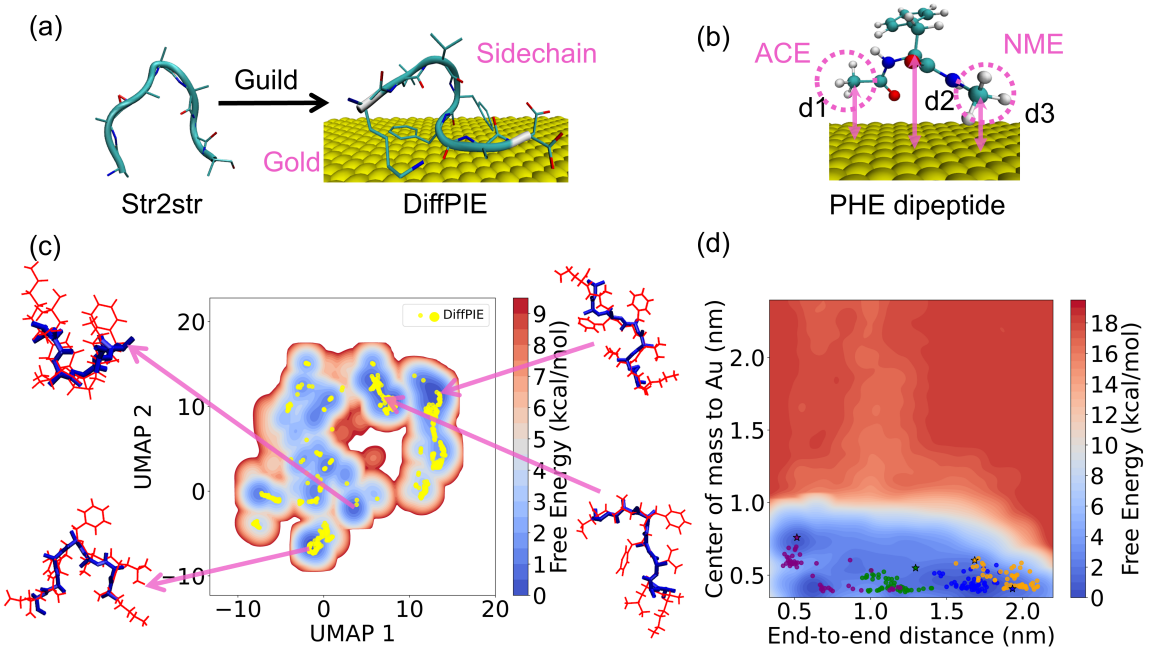


Figure 3: (a) presents the system of $A\beta_{16-22}$ on Au(111) surface. (b) demonstrates a model dipeptide molecule (PHE in this case) binding on the gold surface to construct the biasing force $\mathbf{F}_0$. The binding free energy along three CVs, d_1, d_2, and d_3, was evaluated by a metadynamics simulation. Here, d_1, d_2, and d_3 are defined as the distances from the gold surface to the methyl carbon of the acetyl cap, the α-carbon, and the methyl carbon of the N-methylamide cap, respectively. (c) shows the comparison between metadynamics simulations and DiffPIE. The background FES was generated by projecting the metadynamics simulation trajectory onto UMAP features. The DiffPIE samples are projected onto the same UMAP feature space as yellow dots. We further present the top-down view of representative simulated structures (red) from low free energy clusters and similar DiffPIE samples (blue), with the gold surface oriented parallel to the plane of the paper. (d) shows short MD relaxations of three structures selected randomly from DiffPIE structure clusters. The starting structures (stars) and their relaxation trajectories (dots) are mapped onto the FES of the $A\beta_{16-22}$ end-to-end distance and the $A\beta_{16-22}$ center-to-gold surface distance.

values of d_1 and d_3. Unlike the previous example, $\mathbf{F}_0$ in this case cannot be expressed as the gradient of an energy function, highlighting the flexibility of designing $\mathbf{F}_0$. Similarly to the previous example, we project the structures generated by DiffPIE onto UMAP features trained from the benchmark metadynamics simulation. As shown in Fig. 3(c), the clusters of structures generated by DiffPIE overlap well with low-free-energy regions on the FES of metadynamics, suggesting that the populations of structures generated by DiffPIE are qualitatively consistent with conformation stability. DiffPIE is also capable of finding structures similar to representative conformations from metadynamics simulations. We emphasize that while some configurations are highly flexible and DiffPIE can only recover structures with similar overall shapes, for conformations that are energetically stabilized with low free energy, DiffPIE can generate structures that nearly overlap with the simulated structures. Due to strong interactions between amino acid and gold surfaces, these representative structures are not sampled by the original Str2Str algorithm (see Supporting Information). Again, generating structures with DiffPIE only took less than one hour, while metadynamics simulations may take days to complete. Finally, we tested the relaxation of generated structures by randomly selecting three structures as starting points for short MD simulations. Both the starting structures and the relaxation trajectories are presented in Fig. 3(d). For some starting structures, the relaxation leads to the same final conformation, suggesting that representative structures of stable conformations are included in our predictions. Meanwhile, high-free-energy structures quickly relax, indicating that even short MD simulations can significantly aid structure refinement.

Discussion and Conclusion

The two examples presented in this study demonstrate that the introduction of PEI into the foundation generative model (DiffPIE) is capable of exploring protein-with-PEI conformations, which is an impossible mission with the foundation generative model. We highlight

that DiffPIE does not require any further training or fine-tuning of the foundation generative model. In both examples, our objective was to achieve a trade-off between the accuracy of PEI models and the computational cost of building them. Thus, approximations were used in both cases. In the first example, we ignored the nonbonded interactions between the organic linker and residues other than cysteines. In the second example, the presence of the gold surface disrupts the solvation shell of residues near the surface, leading to changes in the solvation free energy that are non-additive. The generated structures suggest that such approximations do not significantly affect the downgrade quality of structures generated by DiffPIE, which suggests that DiffPIE does not require an exact PEI model for the generation of peptide structures. Moreover, different designs of the biasing force in two examples highlight the flexibility of designing PEI models in DiffPIE. The generated structures can also be used as starting points for MD simulations or enhanced sampling simulations.[115,116] As suggested in recent studies, the introduction of enhanced sampling simulations starting from multiple important conformations predicted by machine learning models can significantly improve sampling efficiency.[117,118] We expect that DiffPIE can be combined with enhanced sampling methods or other free-energy calculation approaches to further improve structural exploration and thermodynamic analysis.

Ideally, we hope that DiffPIE will generate structures following the Boltzmann distribution of a peptide with PEI. However, in practice, it is limited by several theoretical and technical challenges. First, generating structures that follow the Boltzmann distribution of a protein requires that the foundation generative model be able to generate protein structures following the Boltzmann distribution without PEI. The foundation generative model used in this study, Str2Str, is known to be biased towards generating more helix-like structures. As shown in the "Results" section, the bias force in DiffPIE can help correct this bias but cannot fully eliminate it. Some recent studies have proposed methods to correct Str2Str[95] or to develop new foundation generative models capable of generating peptide ensembles without PEI.[119,120] We expect that improved foundation generative models will also enhance

the performance of DiffPIE. Second, DiffPIE is limited by the accuracy of the PEI models. Therefore, we do not expect the current version of DiffPIE to generate structures that fully follow the Boltzmann distribution of proteins with PEI. Improving the quality of the DiffPIE-generated ensemble remains an important goal for future work.

Another challenge in DiffPIE is how to efficiently identify structures that work best as initial structures for sampling. The current solution is to screen structures using short MD simulations, which provides the most reliable, but also the most computationally expensive, screening approach. Other methods, such as energy minimization in vacuum or with an implicit solvent, work well for the cyclic peptide example but perform poorly for the peptide-on-gold example. Alternative energy scoring functions for proteins, such as the Rosetta energy functions,[121] are typically tuned for proteins without PEI and are not directly applicable to either example without careful reparameterization. Therefore, the development of efficient and accurate screening methods for proteins with PEI remains an important area for future research.

In summary, we have developed DiffPIE, a framework that guides a foundation diffusion model with physics-based interactions to generate protein structures with PEI. This framework does not require additional training and can be adapted to various types of PEI. In addition to the examples discussed in this work, DiffPIE has the potential to greatly broaden the applications of protein foundation diffusion models, including protein–ligand binding, proteins at interfaces, and proteins in external fields.

Methods

Our proposed method is general for any score-based diffusion model (SBDM). A SBDM perturbs data distribution to noise prior with a diffusion process over a unit time by a linear stochastic differential equation (SDE):

$$d\mathbf{x} = f(\mathbf{x}, t)dt + g(t)d\mathbf{w},\ t \in [0, 1], \tag{4}$$

where $f(\mathbf{x}, t), g(t)$ are user-defined drift and diffusion functions of the SDE and $\mathbf{w}$ denotes a standard Wiener process. With a carefully designed SDE, the marginal probability of $\mathbf{x}$ at diffusion time t, $P_t(\mathbf{x})$, changes from the data distribution $P(\mathbf{x})$ to approximately a simple Gaussian distribution. In this paper, we use an SDE with drift term $f(\mathbf{x}, t) = 0$ and diffusion function $g(t) = \sqrt{\mathrm{d}\sigma^2(t)/\mathrm{d}t}$, where $\sigma(t)$ represents the noise scale. For any diffusion process in Eq.(4), it has a corresponding reverse-time SDE[99,122] :

$$d\mathbf{x} = [f(\mathbf{x}, t) - g^2(t)\nabla_{\mathbf{z}} \log P_t(\mathbf{x})]dt + g(t)d\bar{\mathbf{w}}, \tag{5}$$

with $\bar{\mathbf{w}}$ a standard Wiener process in reverse time. The trajectories of the reverse SDE have the same marginal densities as those of the forward SDE. Thus, the reverse-time SDE can gradually convert noise to data. The SBDM parameterizes the time-dependent score function $\nabla_{\mathbf{x}} \log P_t(\mathbf{x})$ in reverse SDE with a neural network $\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}(t), t)$.

Suppose that the SBDM baseline aims to predict protein structures that follow the Boltzmann distribution, i.e. $P(\mathbf{x}) \approx p^{\mathrm{NPIE}}\mathrm{B}(\mathbf{x}) \propto \exp\{-\beta U(\mathbf{x})\}$ where $U(\mathbf{x})$ is a protein energy function that is similar to a "statistical potential". $\beta = 1/k_{\mathrm{B}}T$ where k_{B} is the Boltzmann constant and T is temperature. PEI can be written as an extra energy term $U_{\mathrm{ext}}(\mathbf{x})$. The Boltzmann distribution of protein structures with PEI is $p_{\mathrm{B}}^{\mathrm{PIE}}(\mathbf{x}) \propto \exp\{-\beta(U(\mathbf{x})+U_{\mathrm{ext}}(\mathbf{x}))\}$ It has been proven that sampling $P_{\mathrm{bias}}(\mathbf{x}) \approx p_{\mathrm{B}}^{\mathrm{PIE}}(\mathbf{x})$ requires a reverse diffusion process with time-dependent marginal distribution satisfying: $P_{\mathrm{bias},t}(\mathbf{x}(t)) \propto P_t(\mathbf{x}(t)) \exp\{-\beta U_t(\mathbf{x}(t))\}$ where U_t is a time-dependent energy function defined as[95]

$$U_t(\mathbf{x}(t)) = -k_{\mathrm{B}}T \log \mathbb{E}_{p(\mathbf{x}(0)|\mathbf{x}(t))} \left(e^{-\beta U_{\mathrm{ext}}(\mathbf{x}(0))}\right). \tag{6}$$

$p(\mathbf{x}(0)|\mathbf{x}(t))$ in Eq.(6) is the probability of $\mathbf{x}(0)$ conditional on the location of $\mathbf{x}(t)$ in the unmodified revert diffusion process. To further simplify the expression of U_t, we adopted an

approximation used in the manifold constraint algorithm for the diffusion model:

$$p(\mathbf{x}(0)|\mathbf{x}(t)) \approx \delta(\mathbf{x}(0) - \hat{\mathbf{x}}(0)), \tag{7}$$

where $\delta(\cdot)$ is the Dirac delta function and $\hat{\mathbf{x}}(0) = \mathbb{E}(\mathbf{x}(0)|\mathbf{x}(t))$. In certain types of diffusion models such as DDPM, $\hat{\mathbf{x}}_0$ is an explicit function of $\mathbf{x}(t)$ and t: $\hat{\mathbf{x}}_0 = \mathbf{h}(\mathbf{x}(t), t)$.[98] With the approximated $p(\mathbf{x}(0)|\mathbf{x}(t)$, Eq.(6) can be simplified to $U_t(\mathbf{x}(t)) \approx U_{\text{ext}}(\mathbf{h}(\mathbf{x}(t), t))$. The approximation of using Eq.(7) works better when t is close to 0. Therefore, care must be taken at the initial stage of backward diffusion. In this study, we scaled $\mathbf{F}(\mathbf{x}(t), t)$ by a time-dependent factor $\lambda(t)$. $\lambda(t)$ is closed to 0 when t is closed to 1 while $\lambda(t)$ increases to 1 when t approaches 0. The detailed designs of $\lambda(t)$ are presented in the supporting information. With $U_{\text{ext}}(\hat{\mathbf{x}}(0))$, we can calculate the potential U_{ext} and we can push the potential pack to the diffusion model: $\mathbf{F}(\mathbf{x}(t), t) = -\nabla_{\mathbf{x}(t)} U_{\text{ext}}(\mathbf{h}(\mathbf{x}(t), t))$. Although using U_{ext} provides a clear picture of how to connect DiffPIE to the Boltzmann distribution of the protein with PEI. We can relax the requirement of using an interaction potential U_{ext} for the structure generation task. This provides a more flexible tool for users to design $\mathbf{F}$, such as $\mathbf{F}$ representing non-equilibrium perturbations from the environment. Therefore, in general, we will express $\mathbf{F} = \mathbf{F}[U_{\text{ext}}](\mathbf{x}(t), t)$ to highlight the dependency of $\mathbf{F}$ on U_{ext}.

In this work, two examples of $\mathbf{F}$ are used: one with an explicitly defined $\mathbf{F}(\mathbf{x}(t), t) = -\nabla_{\mathbf{x}(t)} U_{\text{ext}}(\mathbf{h}(\mathbf{x}(t), t))$, and another that uses $\mathbf{F}$ depending on U_{ext} in a more complicate way, as describe later in this paragraph. In the example of the cyclic peptide P3-F, we use the FES of six distances (Fig. 2(b)) as U_{ext}. This six-dimensional potential was fitted using a logarithmic function of a mixture of Gaussians. We emphasize that the selected collective variables (CVs) provide joint information on both the distance and orientation between the two cysteines. In the second example, $\mathbf{F}$ is designed as follows. First, we use metadynamics to generate a binding FES, A_i, for a dipeptide molecule that represents the residue i of the protein on the gold surface. A_i is a function of three distances, pre-

sented in Fig. 3(b), and can be viewed as a function of $\mathbf{y}_1$, $\mathbf{y}_2$, and $\mathbf{y}_3$, representing the Cα atom ($\mathbf{y}_2$) and the two methyl carbons ($\mathbf{y}_1$ and $\mathbf{y}_3$) of the two capping groups. Therefore, we can express A_i as $A_i(\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3)$. We define the biasing drift on the i-th C$_\alpha$, $\mathbf{F}^{(i)}$, as $\mathbf{F}^{(i)}(\mathbf{x}(t), t) = -\mathbf{P}^i \nabla_{\mathbf{x}(t)} A_i(\hat{\mathbf{x}}_0^{i-1}(\mathbf{x}(t), t), \hat{\mathbf{x}}_0^{i}(\mathbf{x}(t), t), \hat{\mathbf{x}}_0^{i+1}(\mathbf{x}(t), t))$ where $\hat{\mathbf{x}}_0^{i-1}$, $\hat{\mathbf{x}}_0^{i}$, and $\hat{\mathbf{x}}_0^{i+1}$ are the coordinates of the $(i-1)$-th, i-th, and $(i+1)$-th Cα atoms in $\hat{\mathbf{x}}_0$, respectively. $\mathbf{P}^i$ is a projection operator that only keeps $\nabla_{\mathbf{x}(t)} A_i$ on the the i-th C$_\alpha$. Please refer to the Supporting Information for details on the generation of $\mathbf{F}$.

Str2Str builds protein structures by applying translations and rotations in residue-specific local frames. Biasing forces are applied only to Cα Cartesian coordinates, not orientations, since Tweedie's formula does not extend to $SO(3)$,[104] and projecting forces onto the $SO(3)$ tangent space can cause numerical instability.[95] Generating side chain conformations consistent with both the backbone structure and PEI is essential, as side-chain-environment interactions strongly influence the binding affinity to material surfaces. Str2Str provides only backbone atoms, so side chain backmapping typically uses FASPR.[123] However, FASPR alone is insufficient under PEI, as it can create steric clashes with surfaces or cannot build the required linkers. To overcome this, we developed an algorithm to reconstruct missing components (side chains and linkers) from a backbone structure A generated by DiffPIE, which corresponds to MD-derived structures used in U_{ext}. In the first example, the P3-F CYS side chains and the organic linker were reconstructed by aligning the backbone atoms in A with MD snapshots. In the second, the side chains near the gold surface were rebuilt by selecting dipeptide conformations from MD that (1) align with the backbone frame of A and (2) match residue–surface displacement. The remaining side chains were completed using FASPR.

Acknowledgement

Y.W. and M.C. gratefully acknowledge the support of the American Chemical Society Petroleum Research Fund (Grant Number 67307).

Supporting Information Available

All codes and scripts used in this study are available at GitHub (DiffPIE). The molecular dynamics (MD) trajectories and generated results can be accessed at Google Drive.

The following files are available free of charge.

- Supporting information — Extrapolating Foundation Generative Models with Physics: a Case Study of Exploring Peptide Conformations under Protein-environment Interactions: addtional details of biasing force preparation, simulation setup, and DiffPIE structure generations.

References

(1) Choi, J.; Nam, G.; Choi, J.; Jung, Y. A Perspective on Foundation Models in Chemistry. *JACS Au* **2025**, *5*, 1499–1518.

(2) You, Y.; Chen, T.; Sui, Y.; Chen, T.; Wang, Z.; Shen, Y. Graph Contrastive Learning with Augmentations. Advances in Neural Information Processing Systems. 2020; pp 5812–5823.

(3) Wang, Y.; Wang, J.; Cao, Z.; Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence* **2022**, *4*, 279–287.

(4) Liu, S.; Wang, H.; Liu, W.; Lasenby, J.; Guo, H.; Tang, J. Pre-training Molecular

Graph Representation with 3D Geometry. International Conference on Learning Representations. 2022.

(5) Hu, W.; Liu, B.; Gomes, J.; Zitnik, M.; Liang, P.; Pande, V.; Leskovec, J. Strategies for Pre-training Graph Neural Networks. International Conference on Learning Representations. 2020.

(6) Rong, Y.; Bian, Y.; Xu, T.; Xie, W.; WEI, Y.; Huang, W.; Huang, J. Self-Supervised Graph Transformer on Large-Scale Molecular Data. Advances in Neural Information Processing Systems. 2020; pp 12559–12571.

(7) Wang, S.; Guo, Y.; Wang, Y.; Sun, H.; Huang, J. SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction. Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. New York, NY, USA, 2019; p 429–436.

(8) Ross, J.; Belgodere, B.; Chenthamarakshan, V.; Padhi, I.; Mroueh, Y.; Das, P. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence* **2022**, *4*, 1256–1264.

(9) Chang, J.; Ye, J. C. Bidirectional generation of structure and properties through a single molecular foundation model. *Nature Communications* **2024**, *15*, 2323.

(10) Zeni, C. et al. A generative model for inorganic materials design. *Nature* **2025**, *639*, 624–632.

(11) Ross, J.; Belgodere, B.; Hoffman, S. C.; Chenthamarakshan, V.; Navratil, J.; Mroueh, Y.; Das, P. Gp-molformer: A foundation model for molecular generation. *arXiv preprint arXiv:2405.04912* **2024**,

(12) Gruver, N.; Sriram, A.; Madotto, A.; Wilson, A. G.; Zitnick, C. L.; Ulissi, Z. W. Fine-

Tuned Language Models Generate Stable Inorganic Materials as Text. The Twelfth International Conference on Learning Representations. 2024.

(13) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; others Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.

(14) Ruff, K. M.; Pappu, R. V. AlphaFold and implications for intrinsically disordered proteins. *Journal of molecular biology* **2021**, *433*, 167208.

(15) Beal, D. J. ESM 2.0: State of the art and future potential of experience sampling methods in organizational research. *Annu. Rev. Organ. Psychol. Organ. Behav.* **2015**, *2*, 383–407.

(16) Lu, J.; Zhong, B.; Zhang, Z.; Tang, J. Str2str: A score-based framework for zero-shot protein conformation sampling. *arXiv preprint arXiv:2306.03117* **2023**,

(17) Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Physical review letters* **2018**, *120*, 143001.

(18) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nature communications* **2017**, *8*, 13890.

(19) Unke, O. T.; Chmiela, S.; Gastegger, M.; Schütt, K. T.; Sauceda, H. E.; Müller, K.-R. SpookyNet: Learning force fields with electronic degrees of freedom and nonlocal effects. *Nature communications* **2021**, *12*, 7273.

(20) Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications* **2022**, *13*, 2453.

(21) Deng, Y.; Zhao, X.; Sun, H.; Chen, Y.; Wang, X.; Xue, X.; Li, L.; Song, J.; Hsieh, C.-Y.; Hou, T.; Pan, X.; Alomar, T. S.; Ji, X.; Wang, X. RSGPT: a generative transformer model for retrosynthesis planning pre-trained on ten billion datapoints. *Nature Communications* **2025**, *16*, 7012.

(22) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science* **2019**, *5*, 1572–1583, PMID: 31572784.

(23) Xu, M.; Yu, L.; Song, Y.; Shi, C.; Ermon, S.; Tang, J. Geodiff: A geometric diffusion model for molecular conformation generation. *ICLR* **2022**,

(24) Schneuing, A.; Harris, C.; Du, Y.; Didi, K.; Jamasb, A.; Igashov, I.; Du, W.; Gomes, C.; Blundell, T. L.; Lio, P.; Welling, M.; Bronstein, M.; Correia, B. Structure-based drug design with equivariant diffusion models. *Nature Computational Science* **2024**, *4*, 899–909.

(25) Corso, G.; Stärk, H.; Jing, B.; Barzilay, R.; Jaakkola, T. Diffdock: Diffusion steps, twists, and turns for molecular docking. *ICLR* **2023**,

(26) Chen, L.; Li, Y.; Ma, Y.; Gao, L.; Yu, L. Multiscale graph equivariant diffusion model for 3D molecule design. *Science Advances* **2025**, *11*, eadv0778.

(27) Gebauer, N. W.; Gastegger, M.; Hessmann, S. S.; Müller, K.-R.; Schütt, K. T. Inverse design of 3d molecular structures with conditional generative neural networks. *Nature communications* **2022**, *13*, 973.

(28) Jolicoeur-Martineau, A.; Baratin, A.; Kwon, K.; Knyazev, B.; Zhang, Y. Any-Property-Conditional Molecule Generation with Self-Criticism using Spanning Trees. 2025; https://arxiv.org/abs/2407.09357.

(29) Zeng, C.; Jin, J.; Karypis, G.; Transtrum, M.; Tadmor, E. B.; Hennig, R. G.; Roitberg, A.; Martiniani, S.; Liu, M. PropMolFlow: Property-guided Molecule Generation with Geometry-Complete Flow Matching. 2025; https://arxiv.org/abs/2505.21469.

(30) Zhung, W.; Kim, H.; Kim, W. Y. 3D molecular generative framework for interaction-guided drug design. *Nature Communications* **2024**, *15*, 2688.

(31) Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; others Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **2024**, *630*, 493–500.

(32) Watson, J. L. et al. De novo design of protein structure and function with RFdiffusion. *Nature* **2023**, *620*, 1089–1100.

(33) Krishna, R. et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. *Science* **2024**, *384*, eadl2528.

(34) Yu, H.; Yang, H.; Sun, W.; Yan, Z.; Yang, X.; Zhang, H.; Ding, Y.; Li, K. An interpretable RNA foundation model for exploring functional RNA motifs in plants. *Nature Machine Intelligence* **2024**, *6*, 1616–1625.

(35) Dalla-Torre, H.; Gonzalez, L.; Mendoza-Revilla, J.; Lopez Carranza, N.; Grzywaczewski, A. H.; Oteri, F.; Dallago, C.; Trop, E.; de Almeida, B. P.; Sirelkhatim, H.; Richard, G.; Skwark, M.; Beguir, K.; Lopez, M.; Pierrot, T. Nucleotide Transformer: building and evaluating robust foundation models for human genomics. *Nature Methods* **2025**, *22*, 287–297.

(36) Lewis, S. et al. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science* **2025**, *389*, eadv9817.

(37) Chu, A. E.; Kim, J.; Cheng, L.; El Nesr, G.; Xu, M.; Shuai, R. W.; Huang, P.-S. An all-atom protein generative model. *Proceedings of the National Academy of Sciences* **2024**, *121*, e2311500121.

(38) Merchant, A.; Batzner, S.; Schoenholz, S. S.; Aykol, M.; Cheon, G.; Cubuk, E. D. Scaling deep learning for materials discovery. *Nature* **2023**, *624*, 80–85.

(39) Zeni, C.; Pinsler, R.; Zügner, D.; Fowler, A.; Horton, M.; Fu, X.; Wang, Z.; Shysheya, A.; Crabbé, J.; Ueda, S.; others A generative model for inorganic materials design. *Nature* **2025**, *639*, 624–632.

(40) Antunes, L. M.; Butler, K. T.; Grau-Crespo, R. Crystal structure generation with autoregressive large language modeling. *Nature Communications* **2024**, *15*, 10570.

(41) Luo, X.; Wang, Z.; Gao, P.; Lv, J.; Wang, Y.; Chen, C.; Ma, Y. Deep learning generative model for crystal structure prediction. *npj Computational Materials* **2024**, *10*, 254.

(42) Yang, Z.; Zhao, Y.-M.; Wang, X.; Liu, X.; Zhang, X.; Li, Y.; Lv, Q.; Chen, C. Y.-C.; Shen, L. Scalable crystal structure relaxation using an iteration-free deep generative model with uncertainty quantification. *Nature Communications* **2024**, *15*, 8148.

(43) Ye, C.; Wang, Y.; Xie, X.; Zhu, T.; Liu, J.; He, Y.; Zhang, L.; Zhang, J.; Fang, Z.; Wang, L.; Liu, Z.; Weng, H.; Wu, Q. Materials discovery acceleration by using condition generative methodology. 2025; https://arxiv.org/abs/2505.00076.

(44) Rønne, N.; Aspuru-Guzik, A.; Hammer, B. Generative diffusion model for surface structure discovery. *Phys. Rev. B* **2024**, *110*, 235427.

(45) Glass, C. W.; Oganov, A. R.; Hansen, N. USPEX—Evolutionary crystal structure prediction. *Computer Physics Communications* **2006**, *175*, 713–720.

(46) Zhu, Q.; Hattori, S. Automated high-throughput organic crystal structure prediction via population-based sampling. *Digital Discovery* **2025**, *4*, 120–134.

(47) Prentiss, M. C.; Wales, D. J.; Wolynes, P. G. The energy landscape, folding pathways and the kinetics of a knotted protein. *PLoS computational biology* **2010**, *6*, e1000835.

(48) Chen, M.; Yu, T.-Q.; Tuckerman, M. E. Locating landmarks on high-dimensional free energy surfaces. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 3235–3240.

(49) Pracht, P.; Grimme, S.; Bannwarth, C.; Bohle, F.; Ehlert, S.; Feldmann, G.; Gorges, J.; Müller, M.; Neudecker, T.; Plett, C.; others CREST—A program for the exploration of low-energy molecular chemical space. *The Journal of Chemical Physics* **2024**, *160*.

(50) Wang, Y.; Lv, J.; Zhu, L.; Ma, Y. CALYPSO: A method for crystal structure prediction. *Computer Physics Communications* **2012**, *183*, 2063–2070.

(51) Terayama, K.; Yamashita, T.; Oguchi, T.; Tsuda, K. Fine-grained optimization method for crystal structure prediction. *npj Computational Materials* **2018**, *4*.

(52) Zhang, J.; Dolg, M. ABCluster: the artificial bee colony algorithm for cluster global optimization. *Physical Chemistry Chemical Physics* **2015**, *17*, 24173–24181.

(53) Wales, D. J.; Miller, M. A.; Walsh, T. R. Archetypal energy landscapes. *Nature* **1998**, *394*, 758–760.

(54) Mousseau, N.; Barkema, G. T. Traveling through potential energy landscapes of disordered materials: The activation-relaxation technique. *Phys. Rev. E* **1998**, *57*, 2419–2424.

(55) Gill, S. C.; Lim, N. M.; Grinaway, P. B.; Rustenburg, A. S.; Fass, J.; Ross, G. A.; Chodera, J. D.; Mobley, D. L. Binding modes of ligands using enhanced sampling (BLUES): rapid decorrelation of ligand binding modes via nonequilibrium candidate Monte Carlo. *The Journal of Physical Chemistry B* **2018**, *122*, 5579–5598.

(56) Poole, W. G.; Samways, M. L.; Branduardi, D.; Taylor, R. D.; Verdonk, M. L.; Essex, J. W. Accelerating fragment-based drug discovery using grand canonical nonequilibrium candidate Monte Carlo. *Nature Communications* **2025**, *16*, 6198.

(57) Rodriguez, G. J.; others A New Protein–Protein Docking Scoring Function Based on Interface Structure Similarity. *PLoS ONE* **2015**, *10*, e0125941.

(58) Knight, J. L.; Brooks, C. L. I. Multisite Dynamics for Simulated Structure–Activity Relationship Studies. *Journal of Chemical Theory and Computation* **2011**, *7*, 2728–2739, PMID: 22125476.

(59) Laio, A.; Parrinello, M. Escaping free-energy minima. *Proceedings of the National Academy of Sciences* **2002**, *99*, 12562–12566.

(60) Maragliano, L.; Vanden-Eijnden, E. A temperature accelerated method for sampling free energy and determining reaction pathways in rare events simulations. *Chem. Phys. Lett.* **2006**, *426*, 168–175.

(61) Abrams, J. B.; Tuckerman, M. E. Efficient and Direct Generation of Multidimensional Free Energy Surfaces via Adiabatic Dynamics without Coordinate Transformations. *The Journal of Physical Chemistry B* **2008**, *112*, 15742–15757.

(62) Yu, T.-Q.; Tuckerman, M. E. Temperature-Accelerated Method for Exploring Polymorphism in Molecular Crystals Based on Free Energy. *Phys. Rev. Lett.* **2011**, *107*, 015701.

(63) Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: A molecular modeling perspective. *Medicinal Research Reviews* **1996**, *16*, 3–50.

(64) Hornfeck, W.; Červený, K. On the combinatorics of crystal structures. II. Number of

Wyckoff sequences of a given subdivision complexity. *Acta Crystallographica Section A* **2023**, *79*, 280–294.

(65) Dryden, D. T.; Thomson, A. R.; White, J. H. How much of protein sequence space has been explored by life on Earth? *Journal of The Royal Society Interface* **2008**, *5*, 953–956.

(66) Pantelejevs, T.; Zuazua-Villar, P.; Koczy, O.; Counsell, A. J.; Walsh, S. J.; Robertson, N. S.; Spring, D. R.; Downs, J. A.; Hyvönen, M. A recombinant approach for stapled peptide discovery yields inhibitors of the RAD51 recombinase. *Chemical Science* **2023**, *14*, 13915–13923.

(67) Harter, R. D.; Stotzky, G. Formation of clay-protein complexes. *Soil Science Society of America Journal* **1971**, *35*, 383–389.

(68) Liu, J.; Peng, Q. Protein-gold nanoparticle interactions and their possible impact on biomedical applications. *Acta biomaterialia* **2017**, *55*, 13–27.

(69) Wu, Y.; Ng, D. Y.; Kuan, S. L.; Weil, T. Protein–polymer therapeutics: a macromolecular perspective. *Biomaterials science* **2015**, *3*, 214–230.

(70) Chen, H.; Yuan, L.; Song, W.; Wu, Z.; Li, D. Biocompatible polymer materials: role of protein–surface interactions. *Progress in Polymer Science* **2008**, *33*, 1059–1087.

(71) Fry, D. C. Protein–protein interactions as targets for small molecule drug discovery. *Peptide Science: Original Research on Biomolecules* **2006**, *84*, 535–552.

(72) Jao, D.; Xue, Y.; Medina, J.; Hu, X. Protein-based drug-delivery materials. *Materials* **2017**, *10*, 517.

(73) Drenth, J. *Principles of protein X-ray crystallography*; Springer Science & Business Media, 2007.

(74) Mittermaier, A. K.; Kay, L. E. Observing biological dynamics at atomic resolution using NMR. *Trends in biochemical sciences* **2009**, *34*, 601–611.

(75) Cheng, Y. Single-particle cryo-EM—How did it get here and where will it go. *Science* **2018**, *361*, 876–880.

(76) Yip, K. M.; Fischer, N.; Paknia, E.; Chari, A.; Stark, H. Atomic-resolution protein structure determination by cryo-EM. *Nature* **2020**, *587*, 157–161.

(77) Bonilla, S. L.; Kieft, J. S. The promise of cryo-EM to explore RNA structural dynamics. *Journal of Molecular Biology* **2022**, *434*, 167802, Regulatory Mechanisms through RNA Conformational Switching and Dynamics.

(78) Clore, G. M.; Gronenborn, A. M. Determining the structures of large proteins and protein complexes by NMR. *Trends in Biotechnology* **1998**, *16*, 22–34.

(79) Renaud, J.-P.; Chari, A.; Ciferri, C.; ti Liu, W.; Rémigy, H.-W.; Stark, H.; Wiesmann, C. Cryo-EM in drug discovery: achievements, limitations and prospects. *Nature Reviews Drug Discovery* **2018**, *17*, 471–492.

(80) Kessler, H. Conformation and Biological Activity of Cyclic Peptides. *Angewandte Chemie International Edition in English* **1982**, *21*, 512–523.

(81) Bernadó, P.; Mylonas, E.; Petoukhov, M. V.; Blackledge, M.; Svergun, D. I. Structural Characterization of Flexible Proteins Using Small-Angle X-ray Scattering. *Journal of the American Chemical Society* **2007**, *129*, 5656–5664.

(82) Adochitei, A.; Drochioiu, G. Rapid Characterization of Peptide Secondary Structure by FT-IR Spectroscopy. *Revue Roumaine de Chimie* **2011**, *56*, 783–791.

(83) Yang, J. T.; Wu, C.-S. C.; Martinez, H. M. *Enzyme Structure Part K*; Methods in Enzymology; Academic Press, 1986; Vol. 130; pp 208–269.

(84) 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. *Nature Communications* **2023**, *14*, 5739.

(85) Pan, A. C.; Jacobson, D.; Yatsenko, K.; Sritharan, D.; Weinreich, T. M.; Shaw, D. E. Atomic-level characterization of protein–protein association. *Proceedings of the National Academy of Sciences* **2019**, *116*, 4244–4249.

(86) Ishraaq, R.; Das, S. All-atom molecular dynamics simulations of polymer and polyelectrolyte brushes. *Chemical Communications* **2024**, *60*, 6093–6129.

(87) Souza, P. C.; Thallmair, S.; Conflitti, P.; Ramírez-Palacios, C.; Alessandri, R.; Raniolo, S.; Limongelli, V.; Marrink, S. J. Protein–ligand binding with the coarse-grained Martini model. *Nature communications* **2020**, *11*, 3714.

(88) Dandekar, B. R.; Mondal, J. Capturing protein–ligand recognition pathways in coarse-grained simulation. *The Journal of Physical Chemistry Letters* **2020**, *11*, 5302–5311.

(89) Shi, R.; Qian, H.-J.; Lu, Z.-Y. Coarse-grained molecular dynamics simulation of polymers: Structures and dynamics. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **2023**, *13*, e1683.

(90) Noid, W. G. Perspective: Coarse-grained models for biomolecular systems. *The Journal of chemical physics* **2013**, *139*.

(91) Jin, J.; Voth, G. A. Statistical mechanical design principles for coarse-grained interactions across different conformational free energy surfaces. *The Journal of Physical Chemistry Letters* **2023**, *14*, 1354–1362.

(92) Mock, M.; Langmead, C. J.; Grandsard, P.; Edavettal, S.; Russell, A. Recent advances in generative biology for biotherapeutic discovery. *Trends in Pharmacological Sciences* **2024**, *45*, 255–267.

(93) Strokach, A.; Kim, P. M. Deep generative modeling for protein design. *Current opinion in structural biology* **2022**, *72*, 226–236.

(94) Yu, Z.; Liu, Y.; Lin, G.; Jiang, W.; Chen, M. ESMAdam: a plug-and-play all-purpose protein ensemble generator. *bioRxiv* **2025**,

(95) Wang, Y.; Wang, L.; Shen, Y.; Wang, Y.; Yuan, H.; Wu, Y.; Gu, Q. Protein Conformation Generation via Force-Guided SE (3) Diffusion Models. *arXiv preprint ARXIV.2403.14088* **2024**,

(96) Hwang, M.; Lee, Z.; Kim, K.-S.; Kim, K.; Yang, E. HybridLinker: Topology-Guided Posterior Sampling for Enhanced Diversity and Validity in 3D Molecular Linker Generation. *arXiv preprint arXiv:2502.17349* **2025**,

(97) Kulytė, P.; Vargas, F.; Mathis, S. V.; Wang, Y. G.; Hernández-Lobato, J. M.; Liò, P. Improving antibody design with force-guided sampling in diffusion models. *arXiv preprint arXiv:2406.05832* **2024**,

(98) Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *NeurIPS* **2020**, *33*, 6840–6851.

(99) Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456* **2020**,

(100) Lamtyugina, A.; Behera, A. K.; Nandy, A.; Floyd, C.; Vaikuntanathan, S. Score-based generative diffusion with" active" correlated noise sources. *arXiv preprint arXiv:2411.07233* **2024**,

(101) Liu, Y.; Yu, Z.; Lindsay, R. J.; Lin, G.; Chen, M.; Sahoo, A.; Hanson, S. M. ExEnDiff: An Experiment-guided Diffusion model for protein conformational Ensemble generation. *bioRxiv* **2024**, 2024–10.

(102) Cao, L.; Coventry, B.; Goreshnik, I.; Huang, B.; Sheffler, W.; Park, J. S.; Jude, K. M.; Marković, I.; Kadam, R. U.; Verschueren, K. H.; others Design of protein-binding proteins from the target structure alone. *Nature* **2022**, *605*, 551–560.

(103) Chung, H.; Sim, B.; Ryu, D.; Ye, J. C. Improving diffusion models for inverse problems using manifold constraints. *Advances in Neural Information Processing Systems* **2022**, *35*, 25683–25696.

(104) Chung, H.; Kim, J.; Mccann, M. T.; Klasky, M. L.; Ye, J. C. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687* **2022**,

(105) Iegre, J.; Krajcovicova, S.; Gunnarsson, A.; Wissler, L.; Käck, H.; Luchniak, A.; Tångefjord, S.; Narjes, F.; Spring, D. R. A cell-active cyclic peptide targeting the Nrf2/Keap1 protein–protein interaction. *Chemical Science* **2023**, *14*, 10800–10805.

(106) Barducci, A.; Bussi, G.; Parrinello, M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* **2008**, *100*, 020603.

(107) Pietrucci, F.; Laio, A. A Collective Variable for the Efficient Exploration of Protein Beta-Sheet Structures: Application to SH3 and GB1. *Journal of Chemical Theory and Computation* **2009**, *5*, 2197–2201.

(108) McInnes, L.; Healy, J.; Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* **2018**,

(109) Zhang, C.; Wang, F.; Zhang, T.; Yang, Y.; Wang, L.; Zhang, X.; Lai, L. De Novo Design of Cyclic Peptide Binders Based on Fragment Docking and Assembling. *Journal of Chemical Information and Modeling* **2025**,

(110) Amendola, V.; Pilot, R.; Frasconi, M.; Maragò, O. M.; Iatì, M. A. Surface plasmon resonance in gold nanoparticles: a review. *Journal of physics: Condensed matter* **2017**, *29*, 203002.

(111) Daraee, H.; Eatemadi, A.; Abbasi, E.; Fekri Aval, S.; Kouhi, M.; Akbarzadeh, A. Application of gold nanoparticles in biomedical and drug delivery. *Artificial cells, nanomedicine, and biotechnology* **2016**, *44*, 410–422.

(112) Jahan Sajib, M. S.; Sarker, P.; Wei, Y.; Tao, X.; Wei, T. Protein corona on gold nanoparticles studied with coarse-grained simulations. *Langmuir* **2020**, *36*, 13356–13363.

(113) Wang, A.; Vangala, K.; Vo, T.; Zhang, D.; Fitzkee, N. C. A three-step model for protein–gold nanoparticle adsorption. *The Journal of Physical Chemistry C* **2014**, *118*, 8134–8142.

(114) Bellucci, L.; Ardèvol, A.; Parrinello, M.; Lutz, H.; Lu, H.; Weidner, T.; Corni, S. The interaction with gold suppresses fiber-like conformations of the amyloid β (16–22) peptide. *Nanoscale* **2016**, *8*, 8737–8748.

(115) Ribeiro, J. M. L.; Bravo, P.; Wang, Y.; Tiwary, P. Reweighted autoencoded variational Bayes for enhanced sampling (RAVE). *J. Chem. Phys.* **2018**, *149*, 072301.

(116) Vani, B. P.; Aranganathan, A.; Wang, D.; Tiwary, P. Alphafold2-rave: From sequence to boltzmann ranking. *Journal of chemical theory and computation* **2023**, *19*, 4351–4354.

(117) Vani, B. P.; Aranganathan, A.; Tiwary, P. Exploring kinase Asp-Phe-Gly (DFG) loop conformational stability with AlphaFold2-RAVE. *Journal of chemical information and modeling* **2023**, *64*, 2789–2797.

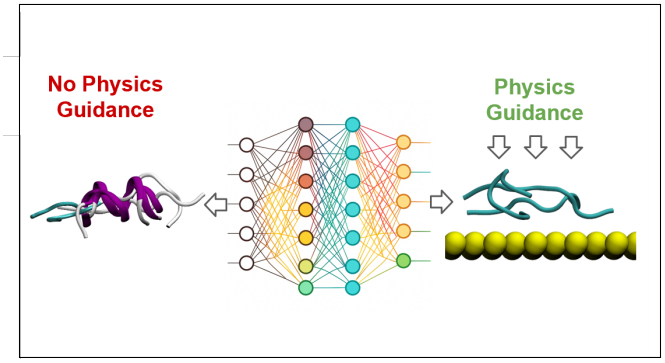
(118) Lamim Ribeiro, J. M.; Tiwary, P. Toward achieving efficient and accurate ligand-protein unbinding with deep learning and molecular dynamics through RAVE. *Journal of chemical theory and computation* **2018**, *15*, 708–719.

(119) 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.; others De novo design of protein structure and function with RFdiffusion. *Nature* **2023**, *620*, 1089–1100.

(120) Wu, K. E.; Yang, K. K.; van den Berg, R.; Alamdari, S.; Zou, J. Y.; Lu, A. X.; Amini, A. P. Protein structure generation via folding diffusion. *Nature communications* **2024**, *15*, 1059.

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

(122) Anderson, B. D. Reverse-time diffusion equation models. *Stoch. Process. Their Appl.* **1982**, *12*, 313–326.

(123) Huang, X.; Pearce, R.; Zhang, Y. FASPR: an open-source tool for fast and accurate protein side-chain packing. *Bioinformatics* **2020**, *36*, 3758–3765.

TOC Graphic