PROTEIN GENERATION WITH EMBEDDING LEARNING FOR MOTIF DIVERSIFICATION

Kevin Michalewicz[1,2,3*] **Chen Jin**[1] **Philip Teare**[1] **Tom Diethe**[1] **Mauricio Barahona**[2]

Barbara Bravi[2] **Asher Mullokandov**[1†]

[1]Centre for AI, Data Science & Artificial Intelligence, Biopharma R&D, AstraZeneca, UK
[2]Department of Mathematics, Imperial College London, UK
[3]Instituto de Ingeniería Biomédica, Universidad de Buenos Aires, Argentina

ABSTRACT

A fundamental challenge in protein design is the trade-off between generating structural diversity while preserving motif biological function. Current state-of-the-art methods, such as partial diffusion in RFdiffusion, often fail to resolve this trade-off: small perturbations yield motifs nearly identical to the native structure, whereas larger perturbations violate the geometric constraints necessary for biological function. We introduce Protein Generation with Embedding Learning (PGEL), a general framework that learns high-dimensional embeddings encoding sequence and structural features of a target motif in the representation space of a diffusion model's frozen denoiser, and then enhances motif diversity by introducing controlled perturbations in the embedding space. PGEL is thus able to loosen geometric constraints while satisfying typical design metrics, leading to more diverse yet viable structures. We demonstrate PGEL on ten representative cases, which include a cancer-related transcription factor complex, an antibody-antigen complex, and an enzyme. PGEL achieves greater structural diversity, better designability, and improved self-consistency, as compared to partial diffusion. Our results establish PGEL as a general strategy for embedding-driven protein generation allowing for systematic, viable diversification of functional motifs.

1 INTRODUCTION

Designing proteins that achieve precise biological functions while allowing for structural diversity has long been a central goal in computational protein design. Recent advances in structure prediction models like AlphaFold (Jumper et al., 2021; Abramson et al., 2024), RoseTTAFold (Baek et al., 2021), ESMFold (Lin et al., 2023) and Boltz (Wohlwend et al., 2024; Passaro et al., 2025) have revolutionized protein generative models and paved the way for improved diffusion models in protein design. Among them, RFdiffusion (Watson et al., 2023), which results from fine-tuning RoseTTAFold, has shown strong performance in both unconditional and conditional generation.

Yet, targeted local modification remains a challenge. A common approach is *partial diffusion* in RFdiffusion, in which a native or designed structure undergoes only a few denoising steps to induce diversification (Watson et al., 2023; Vázquez Torres et al., 2024). However, this method faces a fundamental diversity-fidelity trade-off: small structural perturbations keep near-native conformations, but lack diversity, while larger perturbations induce excessive geometric drift that disrupts functional features (Lin et al., 2024). Overcoming this limitation requires rethinking how diffusion models can introduce controlled variation while still anchoring designs to essential geometric constraints.

A promising direction comes from recent advances in conditional image generation. Models such as Stable Diffusion and Latent Diffusion Models (LDMs) generate images from noise guided by text prompts (Ho et al., 2020; Rombach et al., 2022). Beyond standard prompting, textual inversion

*For correspondence: kmichalewicz@fi.uba.ar
†For correspondence: asher.mullokandov@astrazeneca.com

learns new prompt embeddings to represent unseen visual concepts (Gal et al., 2022; Jin et al., 2024). Once learned, these embeddings can be diversified to generate outputs that preserve the original concept while exploring novel variations. Here we adopt this embedding-centric view in the context of protein generation.

We present Protein Generation with Embedding Learning (PGEL), a general framework representing the first adaptation of textual inversion principles to protein diffusion models. PGEL introduces two key approaches with broad applicability: (1) learning high-dimensional embeddings that capture the sequence and structural characteristics of target protein regions of interest, thus shifting the paradigm from coordinate-space to embedding-space perturbations, and (2) relaxing evolutionary and structural constraints by masking embeddings. Although we present our work here using RFdiffusion's representation space, our method is general and readily adaptable to other protein diffusion models, and can thus leverage the rich representational capacity of pre-trained diffusion models without expensive retraining or fine-tuning.

In this work we formalize a new design task, *motif diversification*: given an experimentally characterized structure with a functional motif embedded in a larger scaffold, we aim to generate alternative motif conformations and poses while keeping the rest of the protein fixed in real space. Diversification is defined at the backbone level, meaning that motif residues are allowed to move and rearrange relative to each other and to the scaffold, but without pre-specifying side-chain identities or chemistry. Instead, we require that downstream sequence design and structure prediction can recover sequences that realize each diversified backbone. Biologically, this corresponds to exploring families of alternative structural realizations of an underlying functional motif (*e.g.*, a binding epitope or active site) that remain compatible with the scaffold, enabling the enhancement of properties such as affinity, specificity, or stability while preserving the overall protein architecture. Some existing approaches address related challenges, but differ in scope and implementation: structure inpainting methods (*e.g.*, masked region generation) fully marginalize a region by masking and regenerating it *de novo*, discarding the specific native geometry (Zhang et al., 2023), whereas flexible backbone loop remodeling in Rosetta (KIC/Next-Generation KIC) samples local conformations under explicit geometric and energetic restraints to achieve high-fidelity but relatively localized exploration (Mandell et al., 2009; Stein & Kortemme, 2013; Leman et al., 2020). Hence these tools do not explicitly target controlled exploration of a *neighborhood* around an existing functional motif while keeping a surrounding scaffold nearly fixed.

We, therefore, compare chiefly to partial diffusion in RFdiffusion, the prevailing stochastic baseline for local variation which has been recently applied in therapeutically relevant design settings, including *de novo* creation of high-affinity peptide binders and venom toxin neutralizers (Vázquez Torres et al., 2024; 2025). Across ten representative scenarios, eight proposed in Watson et al. (2023), with two additional cases involving an antibody bound to Alzheimer's disease's amyloid beta peptide and the adenylate kinase enzyme, PGEL (1000 samples) produces more designable structures in nine cases (motif pLDDT $\geq$ 70, scRMSD $\leq$ 1Å, mRMSD $\leq$ 2Å) than partial diffusion. PGEL also yields more structurally diverse TM-score clusters distinguishable from native, and shows better self-consistency after inverse folding and refolding (meeting mRMSD and pAE thresholds), while maintaining predicted binding affinities comparable to native and exceeding those obtained with partial diffusion. Our results support embedding learning combined with masking as a general, efficient strategy for systematic motif diversification.

2 BACKGROUND

Functional motifs. Conditional generation around functional residues, often framed as *motif scaffolding*, has been a focal point for recent protein design methods. In that setting, the motif and scaffold are defined as disjoint subsets with the scaffold varied while the motif geometry is preserved. Approaches like RFdiffusion (where the motif coordinates are fixed), the Monte Carlo-based Twisted Diffusion Sampler (Wu et al., 2023) applied to FrameDiff, and Genie2 (Lin et al., 2024) have made progress on this task, though performance remains task-dependent and can yield few or no backbones meeting success criteria in specific cases. In our motif diversification task, the scaffold is held fixed and the motif is diversified to explore multiple, function-preserving geometric realizations, enabling improvements in *e.g.*, affinity, specificity or stability, while maintaining the broader structural context.

Protein embeddings. The limited availability of structural data motivated the development of models that transform sequences into sequence embeddings that encode structural information. These embeddings have been employed for various tasks such as property prediction using a Gaussian Process regression model (Yang et al., 2018) and residue-residue contact prediction via a Bidirectional Long Short-Term Memory (BiLSTM) architecture (Bepler & Berger, 2019). Transformers (Vaswani et al., 2017) have been used in generating sequence embeddings, including for antibody-specific applications like paratope prediction (Leem et al., 2022). Transfer learning has also been shown to significantly improve performance across architectures by enabling the use of pre-trained embeddings that capture fundamental sequence-structure relationships (Detlefsen et al., 2022). Other approaches explicitly include structural information (Ali et al., 2024), such as contact maps-derived embeddings, and have shown enhanced performance in particular downstream tasks such as structure similarity assessment (Kandathil et al., 2025), structure searching (Greener & Jamali, 2024), property prediction (Blaabjerg et al., 2024; Danner et al., 2025), and domain classification (Lau et al., 2024). Similarly, protein function annotation and local flexibility prediction have benefited from Graph Convolutional Networks, which combine structure-derived graphs to propagate contextual signals from protein sequence embeddings obtained with pre-trained models (Gligorijević et al., 2021; Michalewicz et al., 2025).

Diffusion models for proteins. Earlier works adapted Denoising Diffusion Probabilistic Models (DDPMs) to protein design by conditioning on local structural elements or coarse fold constraints (Wu et al., 2024; Anand & Achim, 2022; Trippe et al., 2023; Luo et al., 2022) yet, while encouraging, they produced few sequences that refolded to target backbones. RFdiffusion subsequently emerged as the diffusion approach that reliably yields designable structures and sequences that recover the intended geometry. In RFdiffusion, a highly accurate protein structure prediction method (RoseTTAFold (Baek et al., 2021)) is fine-tuned to undo random perturbations of atomic coordinates introduced via 3D Gaussian noise (*i.e.*, to denoise). RFdiffusion can be constrained to specific binding targets, or symmetry specifications, and once trained it can be viewed as a *frozen denoiser*. RoseTTAFold/AlphaFold-style models (including RFdiffusion) learn so-called *state* and *pair* embeddings (related to per-residue and residue-residue properties of the protein structure, respectively) and MSA embeddings related to multiple sequence alignment (Jumper et al., 2021).

Textual inversion. Gal et al. (2022) builds on LDMs (Rombach et al., 2022), a specific class of DDPMs, to perform textual inversion. In the context of text-to-image models, let x represent an image, s a text prompt, ϵ_θ a pre-trained denoising network, and ε an image encoder. LDMs aim to minimize the following loss:

$$\mathcal{L}_{\text{LDM}} := \mathbb{E}_{z\sim\varepsilon(x),s,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c_\theta(s))\|_2^2\right] \tag{1}$$

Here, $c_\theta(s)$ represents a pre-trained text encoder that conditions the denoiser ϵ_θ based on the text prompt s, and z_t is a noised version of the image embedding z at timestep t. The goal of textual inversion is to learn a new text embedding v_* corresponding to a particular concept s_* such that it minimizes the LDM loss (equation 1). This means conditioning ϵ_θ on v_* so the generated image $\tilde{x}$ closely resembles the original image x. Neutral prompts, such as "A photo of s_*" or "A portrait of s_*" are used while keeping ϵ_θ and c_θ frozen. Multi-Concept Prompt Learning (Jin et al., 2024) extends this idea to handle multiple concepts by incorporating three regularization techniques: attention masking, bind adjective, and prompts contrastive loss.

3 METHODS

We now present our method, *Protein Generation with Embedding Learning (PGEL)*, and describe how we learn the embedding representation of a motif in Section 3.1. In Section 3.2, we propose an approach to increase motif diversity, and Section 3.3 details the evaluation metrics.

3.1 PROTEIN GENERATION WITH EMBEDDING LEARNING (PGEL)

We generalize the notion of textual inversion with LDMs to proteins, treating the structure as analogous to an image, and the sequence as analogous to a text prompt. Let R_* be a region of interest, or *motif*, defined as a continuous or discontinuous set of L_* amino acids within a protein. The motif has structure x_* and sequence s_*, where the coordinates of x_* are obtained from an experimental

Protein Data Bank (PDB) entry, and the sequence s_* is *masked* when passed as an input to PGEL, *i.e.*, the amino acid range of the motif is specified, but not its exact composition.

PGEL learns a representation of R_* in embedding space, which we denote as v_*. The remainder of the protein constitutes the *scaffold*, with structure x_c and sequence s_c of length L_c, from which the protein LDM frozen ENCODER computes an embedding representation v_c.

The procedure (see Figure 1 and Algorithm 1) starts by building a noised protein structure in which the scaffold coordinates are retained while the motif coordinates are subjected to T rounds of Gaussian noise injection, following Trippe et al. (2023). At each timestep t, the protein LDM frozen DENOISER predicts a denoised motif structure $\hat{x}_*^{(0)}$, conditioned jointly on the learnable motif embedding v_* and the fixed embedding v_c. These embeddings include state, pair and MSA embeddings. Then, by using structure $x^{(t)}$ and the intermediate structure $[x_c, \hat{x}_*^{(0)}]$, a reverse diffusion step REVERSESTEP, which does not contain any learnable parameters, yields $x^{(t-1)}$ (see Algorithm 3). In practice, we employ pre-trained building blocks of RFdiffusion for both the ENCODER and DENOISER, though alternative models could be substituted if desired.

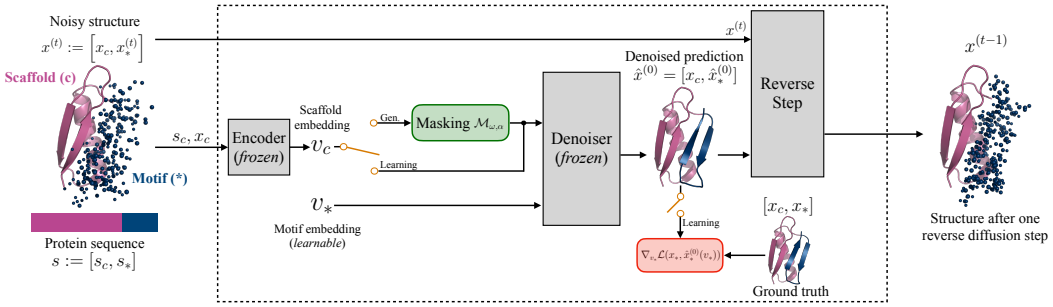


Figure 1: Outline of the PGEL learning and generation procedures during one reverse diffusion step.

Embedding optimization. The embedding v_* is learned by minimizing:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda_{\text{DM}}\mathcal{L}_{\text{DM}} + \lambda_{\text{torsion}}\mathcal{L}_{\text{torsion}} \tag{2}$$

This loss function is composed of three terms, described hereafter, which compare different features of the ground truth structure x_* and the predicted structure $\hat{x}_*^{(0)}(v_*)$ of the motif with the coefficients $\lambda_{\text{DM}}, \lambda_{\text{torsion}} \in \mathbb{R}_{\geq 0}$ controlling the relative weight of the terms.

Data fidelity term (backbone atoms). For each motif residue $i \in R_*$ we consider the $A = 4$ backbone atoms (nitrogen N, α-carbon C_α, carbon C, oxygen O). Let $\hat{x}_{i,a}^{(0)} \in \mathbb{R}^3$ denote the predicted position of atom a in residue i and $x_{i,a} \in \mathbb{R}^3$ its ground truth counterpart. We then compute the mean squared error (MSE) between the backbone atoms of the ground truth and predicted motif:

$$\mathcal{L}_{\text{MSE}}(x_*, \hat{x}_*^{(0)}(v_*)) = \frac{1}{AL_*} \sum_{i \in R_*} \sum_{a \in \{\text{N},\text{C}_\alpha,\text{C},\text{O}\}} \left\| \hat{x}_{i,a}^{(0)}(v_*) - x_{i,a} \right\|^2 \tag{3}$$

Distance matrix between α-carbons. Let $\hat{x}_{i,\text{C}_\alpha}^{(0)} \in \mathbb{R}^3$ denote the predicted position of the α-carbon atom in residue i, and $x_{i,\text{C}_\alpha} \in \mathbb{R}^3$ its ground truth counterpart. We define the following loss term based on α-carbon Distance Matrices (DM), inspired by the distrogram notion (Senior et al., 2020):

$$\mathcal{L}_{\text{DM}}(x_*, \hat{x}_*^{(0)}(v_*)) = \frac{1}{L_*^2} \sum_{i \in R_*} \sum_{j \in R_*} \left(\left\| \hat{x}_{i,\text{C}_\alpha}^{(0)}(v_*) - \hat{x}_{j,\text{C}_\alpha}^{(0)}(v_*) \right\| - \| x_{i,\text{C}_\alpha} - x_{j,\text{C}_\alpha} \| \right)^2 \tag{4}$$

In contrast to $\mathcal{L}_{\text{MSE}}$, $\mathcal{L}_{\text{DM}}$ is invariant under rigid motions (translations and rotations), thus encouraging global shape consistency.

Backbone torsion angles. Let $\hat{\phi}_i$ and $\hat{\psi}_i$ denote the predicted backbone torsion angles at residue i, computed from $\hat{x}_*^{(0)}(v_*)$, and let ϕ_i and ψ_i be the corresponding ground truth values (Ramachandran

et al., 1963). We impose a constraint on angular torsions through a cosine-based loss term akin to that of AlphaFold (Jumper et al., 2021):

$$\mathcal{L}_{\text{torsion}}(x_*, \hat{x}_*^{(0)}(v_*)) = \frac{1}{L_* - 2} \sum_{i=2}^{L_*-1} \left[1 - \cos\Big(\hat{\phi}_i(v_*) - \phi_i\Big) + 1 - \cos\Big(\hat{\psi}_i(v_*) - \psi_i\Big)\right] \quad (5)$$

This term penalizes sterically implausible geometries, helping improve performance under the predicted local distance difference test (pLDDT). We do not include the third backbone angle ω as it is typically considered fixed at 180 degrees (Cutello et al., 2006).

Algorithm 1 PGEL – Embedding learning

Input: region of interest/motif R_* with masked sequence s_* and structure x_*, fixed scaffold with sequence s_c and structure x_c, pre-trained ENCODER and DENOISER.
Output: learned embedding v_* for region R_*.
initialize v_* with zeros.
while not converged **do**
  Build noised structure $x^{(T)} := [x_c, x_*^{(T)}]$ with associated sequence $s := [s_c, s_*]$.
  $v_c = \text{ENCODER}(s_c, x_c)$
  for $t = T$ **down to** 1 **do**
    $\hat{x}_*^{(0)} = \text{DENOISER}(v_c, v_*)$
    $x^{(t-1)} = \text{REVERSESTEP}(x^{(t)}, [x_c, \hat{x}_*^{(0)}])$
    Update v_* by taking a gradient step $\nabla_{v_*}\mathcal{L}(x_*, \hat{x}_*^{(0)}(v_*))$
  end for
end while
Return v_*

Once the embeddings are learned for a particular protein, we employ Algorithm 2 to generate novel proteins with diversified region of interest R_* (see Figure 1). We demonstrate the necessity of learning motif embeddings by comparing with simpler baselines in Appendix D.

Algorithm 2 PGEL – Generation with embedding masking

Input: region of interest/motif R_* with learned embeddings v_*, fixed scaffold with sequence s_c and structure x_c, pre-trained ENCODER and DENOISER.
Output: generated structure.
Build noised structure $x^{(T)} := [x_c, x_*^{(T)}]$.
Draw at random the sample masking type $\omega \sim \text{Ber}(\frac{1}{2})$ (row if 0, column if 1).
Sample masking rate $\alpha \sim \mathcal{U}[0, 1]$.
Define $\mathcal{M}_{\omega,\alpha}(\cdot)$ as a zero mask with type ω and rate α.
$v_c = \text{ENCODER}(s_c, x_c)$
for $t = T$ **down to** 1 **do**
  $\hat{x}_*^{(0)} = \text{DENOISER}(\mathcal{M}_{\omega,\alpha}(v_c), v_*)$
  $x^{(t-1)} = \text{REVERSESTEP}(x^{(t)}, [x_c, \hat{x}_*^{(0)}])$
end for
Return $x^{(0)}$

3.2 ENHANCING THE DIVERSITY OF GENERATED MOTIFS

MSA embeddings in sequence-to-structure predictors contain evolutionary covariation information about residues, thereby capturing geometric constraints such as residue proximity. In RFdiffusion, however, such embeddings are derived solely from the input sequence $s := [s_c, s_*]$ rather than from a full stack of aligned sequences, and can be represented as a $d_{\text{MSA}} \times L$ matrix, where d_{MSA} is the depth of the MSA embeddings and $L := L_* + L_c$ the total protein length. With PGEL, we show that the diversity of generated structures can be increased by applying perturbations to the scaffold MSA embeddings $v_c^{\text{MSA}} \in \mathbb{R}^{d_{\text{MSA}} \times L_c}$. These embeddings couple through attention mechanisms with state and pair embeddings produced by an internal RFdiffusion encoder, and also interact with

the learned motif embedding v_*, which provides an independent conditioning signal for the frozen denoiser (see Appendix A for more details).

Embedding masking. We studied the effect of applying zero masks, *i.e.* masks zeroing specific elements, to the scaffold MSA embeddings during generation (see Algorithm 2). *Row masking* corresponds to masking specific features for all residues, whereas *column masking* zeroes out all features of specific residues. Both strategies lift some constraints on inter-residue distances, and modulate which co-variation patterns remain accessible. We sample $\omega \sim \mathrm{Ber}(\frac{1}{2})$ to choose the masking mode ($\omega = 0$ for row masking and $\omega = 1$ for column masking) and $\alpha \sim \mathcal{U}[0, 1]$, the masking rate, to set the fraction of rows or columns masked. This defines the operator $\mathcal{M}_{\omega,\alpha}(\cdot)$, which implements zero masking with type ω and rate α. As such, masking v_c relaxes the geometric constraints of the generated motif (see Appendix B), while its omission results in no diversity, as shown in Appendix D. We also assessed the correlation between motif structural diversity and rate α in Appendix E.

3.3 Evaluation metrics

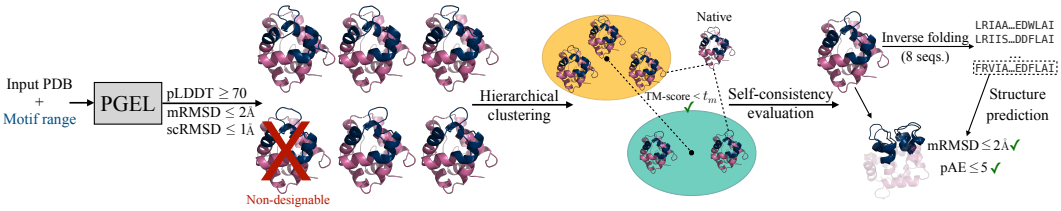


Figure 2: Summary of the evaluation metrics. PGEL takes as input a PDB entry and the amino acid range corresponding to the motif. From 1000 PGEL-generated backbones, designable candidates are filtered by root mean square deviation (RMSD) and pLDDT thresholds, and structural diversity is assessed via hierarchical clustering. Backbones are also required to be *distinguishable* from the native. Cluster representatives undergo self-consistency evaluation: sequences assigned to the designable backbones with ProteinMPNN are refolded, and at least one predicted structure must satisfy set mRMSD and predicted alignment error (pAE) conditions relative to the generated backbone.

Designability. To quantify designability in the motif diversification task, we first require a motif $\mathrm{pLDDT} \geq 70$ as computed by an RFdiffusion internal block, following the threshold adopted for related tasks by Lin et al. (2024). We also require the scaffold RMSD to be $\mathrm{scRMSD} \leq 1$Å, to ensure that the residues surrounding the motif remain fixed. Finally, we set the threshold for the motif RMSD to $\mathrm{mRMSD} \leq 2$Å, to allow for structural diversification of the motif.

Diversity. To quantify structural diversity among generated proteins (Figure 2), we compute the pairwise TM-scores (Zhang & Skolnick, 2004) across all designable candidates, and employ hierarchical clustering (Lin et al., 2024) with several linkage thresholds t_m to group similar backbones under this score. Diversity is measured by the number of clusters. We evaluate also the TM-score with respect to the native motif: a cluster is considered *distinguishable* from native if, for TM-score threshold $t_m \in [0, 1]$, at least one cluster member exhibits lower similarity than t_m relative to the native. This analysis ensures that we capture the structural distinctiveness of the backbones.

Self-consistency. For the distinguishable backbones we use the procedure in Trippe et al. (2023) based on inverse folding to assess self-consistency between generated and predicted structures. Specifically, we use ProteinMPNN with default parameters (Dauparas et al., 2022) to assign 8 plausible sequences to each backbone, followed by a sequence-to-structure model, here AlphaFold3 (Abramson et al., 2024), to predict 8 full proteins. A designed backbone is deemed self-consistent if it satisfies for at least one of the 8 predicted structures: $\mathrm{mRMSD} \leq 2$Å and $\mathrm{pAE} \leq 5$ (Figure 2). Previous studies included this procedure under the designability assessment (Watson et al., 2023; Lin et al., 2024). However, this is computationally expensive and, when prioritizing diversity, often inefficient: many backbones either fail the initial scRMSD or mRMSD filters or exhibit negligible structural diversity. In motif diversification, diversity among generated proteins and distinguishability with respect to the native are decisive. We therefore invert the pipeline to enforce these criteria first and reserve the costly self-consistency evaluation only for diverse candidates.

Binding affinity. For protein-protein complexes, we run PRODIGY (Vangone & Bonvin, 2015; Xue et al., 2016) to estimate the binding affinity ΔG expressed in kcal/mol, with larger $|\Delta G|$ values indicating stronger binding. It is desirable that new designs present binding affinity values comparable to, or larger in magnitude than, those of the native complex. Note that learning is not optimized to enhance binding affinity; rather, this serves as an *a posteriori* assessment.

4 EXPERIMENTS

We focus on ten representative test cases. Eight were proposed for different tasks by Watson et al. (2023), and two were added in this work: a TAP01 family antibody in complex with an amyloid beta peptide, which is related to Alzheimer's disease (van Dyck, 2018), and a mutated version of the adenylate kinase enzyme. We present the results in Figure 3 and Tables 3 and 4, and we describe in detail the cases of a monomer and a p53-MDM2 complex.

4.1 PROTOCOL

We established a protocol to systematically compare our method with RFdiffusion's partial diffusion using the metrics introduced in Section 3.3. For partial diffusion, we generated 1000 protein backbones by uniformly sampling the number of diffusion timesteps, $T \sim \mathcal{U}\{2, 3, \ldots, 49\}$, as $T = 50$ corresponds to the full diffusion process in RFdiffusion. In this way, we cover a spectrum of structural perturbations ranging from near-native backbones to unrelated ones. For PGEL, we performed the learning of v_* with Stochastic Gradient Descent with learning rate $l_r = 4 \times 10^{-4}$ and momentum $p = 0.9$, $\lambda_{\mathrm{DM}} = 0.01$ and $\lambda_{\mathrm{torsion}} = 0.05$ (Algorithm 1), and we then generated 1000 protein backbones (Algorithm 2).

For both sets of 1000 generated structures, we evaluated designability and, among those deemed designable, we computed TM-scores between all generated motifs and with respect to the native structure. We then plotted the number of clusters as a function of the TM-score. For structures that were designable and diverse according to a typical TM-score threshold $t_m = 0.6$ (Lin et al., 2024), we performed inverse folding through ProteinMPNN to generate compatible sequences, followed by AlphaFold3 inference to assess whether the predicted sequences refolded into the intended backbones, fulfilling the self-consistency requirement defined in Section 3.3.

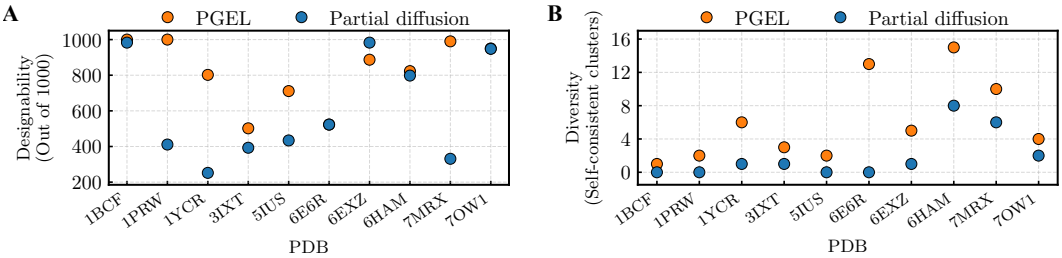


Figure 3: Comparison of partial diffusion in RFdiffusion and PGEL. (A) Number of viable structures out of 1000. (B) Number of self-consistent clusters (diversity), computed at TM-score threshold $t_m = 0.6$.

4.2 EXAMPLE 1: MONOMER

Calmodulin (PDB entry: 1PRW), a monomeric protein containing a double EF-hand motif spanning residues 16-35 and 52-71 (structured multi-motif), was considered as the representative test case for single-chain proteins. All the 1000 backbones candidates generated by PGEL resulted to be designable, well exceeding the 411 obtained by partial diffusion (Figure 3A, Table 3). All of the backbones generated by partial diffusion satisfied the pLDDT constraint, consistent with the fact that RFdiffusion's training favors high-confidence local structures, but 589 of them failed to meet the expected motif RMSD threshold. In these cases, the added noise during diffusion excessively perturbed the initial backbone, leading to conformations that no longer preserved the intended geometry of the EF-hand motif.

We then evaluated the structural diversity of the designable backbones, recording the number of clusters as a function of the TM-score threshold (Figure 4A). PGEL consistently produced a higher number of clusters across thresholds, demonstrating that embedding perturbations through masking introduce greater variability in backbone conformations. On the other hand, partial diffusion yielded structures too similar to the native backbone, and hence not distinguishable from it.

We carried out the self-consistency assessment at $t_m = 0.6$, as per protocol. For PGEL, the two clusters had backbones that successfully refolded into the intended conformations after sequence design and AlphaFold3 inference. Figure 4B illustrates these two successful cases, along with examples of backbones generated by partial diffusion that either did not satisfy the mRMSD metric condition or the distinguishability from native.

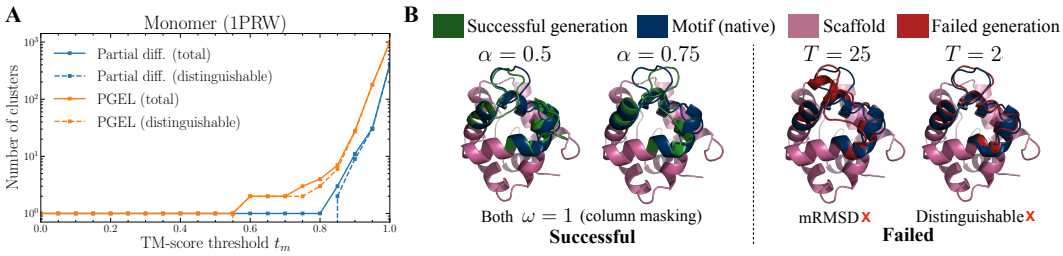


Figure 4: Results for a monomer (PDB: 1PRW). (A) Number of clusters, both total and distinguishable from native, as a function of the TM-score threshold t_m for PGEL and partial diffusion. (B) *Left:* two successful PGEL designs at $t_m = 0.6$ using column masking with rates $\alpha = 0.5$ and $\alpha = 0.75$. *Right:* two partial diffusion failed backbones at $t_m = 0.6$, one obtained with $T = 25$ timesteps that violates the motif RMSD constraint, and one with $T = 2$ timesteps that is not distinguishable from the native.

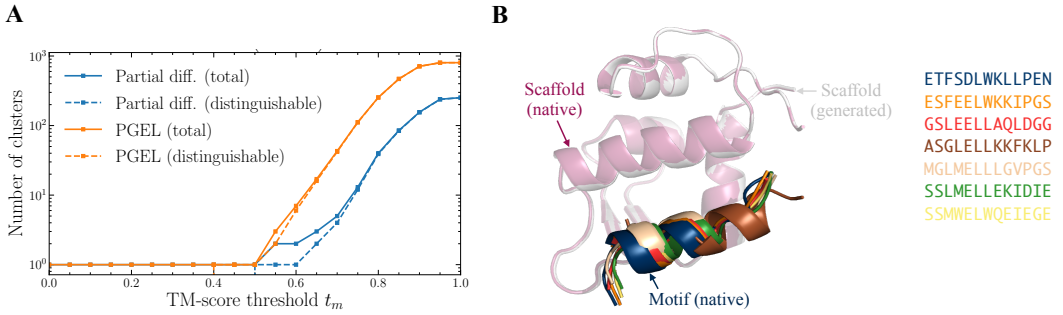


Figure 5: Results for a p53-MDM2 complex (PDB: 1YCR). (A) Number of clusters identified by PGEL and partial diffusion, both total and distinguishable from native, as a function of the TM-score threshold t_m. (B) Generated binding site backbones (overlaid with native), alongside the native sequence and sequences that refold to self-consistent structures.

4.3 Example 2: complex

The interaction between the transcription factor p53 and its negative regulator MDM2 is a key molecular process in cancer progression. Specifically, pharmacological disruption of the p53-MDM2 complex restores p53 activity and has been proven beneficial in cancer therapy (Hu et al., 2021).

Starting from PDB entry 1YCR, we addressed the motif diversification task by redesigning the complete p53 under the RMSD constraints described in Section 3.3. PGEL generated 802 designable backbones out of 1000 trials, with most non-designable cases attributable to low pLDDT confidence scores (Table 3). Partial diffusion produced only 252 designable backbones with considerably reduced structural diversity (a single cluster at TM-score threshold $t_m = 0.6$, see Figure 5B). PGEL, by comparison, gave six clusters at $t_m = 0.6$, all of which passed the self-consistency checks.

When assessing the binding affinity *a posteriori*, five out of six representatives of PGEL clusters exhibited lower affinity compared to the sole valid instance of partial diffusion (Table 5). Notably, sequence SSMWELWQEIEGE (see Figure 5B), designed with PGEL in combination with ProteinMPNN, folded, as predicted by AlphaFold3, into a structure with a binding affinity comparable to that of the native structure, despite sharing only around 15% of sequence identity. This result highlights PGEL's ability to generate backbones that can accommodate sequences unrelated to the native while refolding into structures that preserve function.

Computational remarks. Across the ten case studies, column masking was more effective than row masking (see Table 4, with over 60% of successful outcomes with $\omega = 1$). The time required per timestep during the generation process is nearly indistinguishable between PGEL and partial diffusion: average timestep $0.7\,s$ for partial diffusion and $0.71\,s$ for PGEL on a single NVIDIA GeForce RTX 3090 GPU with 24GB of memory.

5 Limitations and future work

PGEL inherits the biases and limitations of the underlying frozen RFdiffusion denoiser, including its training data distribution and architectural constraints. Moreover, learning embeddings requires additional optimization time, which ranged in our examples from 2 (PDB: 7MRX) to 20 minutes (PDB: 1YCR) on a single GPU, with a trade-off between speed and improved results (more details in Appendix F). Our experiments were limited to motifs of up to 40 residues, with practical limits of around 50 residues given available memory, though scaling to longer motifs should be feasible with larger hardware or engineering optimization. Beyond this, our evaluation is entirely *in silico* (pLDDT/RMSD/TM-score filtering, AlphaFold3 refolding, and PRODIGY ΔG) and thus predictive rather than experimental.

Future work will investigate alternative ways of perturbing embeddings, as this strategy for motif diversification remains largely unexplored, as well as different strategies for sampling the masking parameters ω and α. For instance, instead of sampling α uniformly between 0 and 1, one could bias it toward smaller masking rates (*e.g.*, using a Poisson distribution with rate λ, where λ tunes how conservative or aggressive the masking is), thus providing finer control over structural perturbations. A more systematic mapping between PGEL's ω and α and partial diffusion's T would also clarify the relationship between the diversity-fidelity trade-off in both methods. Finally, experimental validation will be pursued in follow-up work.

Code availability. Upon publication, we will release code and configurations at `github.com/AstraZeneca/PGEL` to facilitate reproducibility.

Author Contributions

Study concept and design: K.M., C.J., A.M.; Development of source code: K.M.; Analysis and interpretation of data: K.M., A.M.; Writing and revision of the manuscript: K.M., C.J., P.T., T.D., M.B., B.B. and A.M.; Study supervision: C.J., M.B., B.B. and A.M.

Acknowledgments

K.M. acknowledges support from the President's PhD Scholarship at Imperial College London. M.B. acknowledges support by the Engineering and Physical Sciences Research Council (EPSRC) under grant EP/N014529/1 funding the EPSRC Centre for Mathematics of Precision Healthcare at Imperial College London. All authors are grateful to Anshul Kanakia, Dino Oglic, Lorena Roldán Martín and Talip Uçar for helpful discussions.

References

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie,

Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, Jun 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL https://doi.org/10.1038/s41586-024-07487-w.

Sarwan Ali, Prakash Chourasia, and Murray Patterson. When protein structure embedding meets large language models. *Genes*, 15(1), 2024. ISSN 2073-4425. doi: 10.3390/genes15010025. URL https://www.mdpi.com/2073-4425/15/1/25.

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, Claudia Millán, Hahnbeom Park, Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues, Alberdina A. van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E. Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David Baker. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021. doi: 10.1126/science.abj8754. URL https://www.science.org/doi/abs/10.1126/science.abj8754.

Tristan Bepler and Bonnie Berger. Learning protein sequence embeddings using information from structure. In *International Conference on Learning Representations*, ICLR'19, 2019.

Lasse M. Blaabjerg, Nicolas Jonsson, Wouter Boomsma, Amelie Stein, and Kresten Lindorff-Larsen. Ssemb: A joint embedding of protein sequence and structure enables robust variant effect predictions. *Nature Communications*, 15(1):9646, Nov 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-53982-z. URL https://doi.org/10.1038/s41467-024-53982-z.

Vincenzo Cutello, Giuseppe Narzisi, and Giuseppe Nicosia. A multi-objective evolutionary approach to the protein structure prediction problem. *Journal of The Royal Society Interface*, 3(6):139–151, 2006. doi: 10.1098/rsif.2005.0083. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2005.0083.

Martin Danner, Matthias Begemann, Miriam Elbracht, Ingo Kurth, and Jeremias Krause. Utilizing protein structure graph embeddings to predict the pathogenicity of missense variants. *NAR Genomics and Bioinformatics*, 7(3):lqaf097, 07 2025. ISSN 2631-9268. doi: 10.1093/nargab/lqaf097. URL https://doi.org/10.1093/nargab/lqaf097.

J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, 2022. doi: 10.1126/science.add2187. URL https://www.science.org/doi/abs/10.1126/science.add2187.

Nicki Skafte Detlefsen, Søren Hauberg, and Wouter Boomsma. Learning meaningful representations of protein sequences. *Nature Communications*, 13(1):1914, Apr 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-29443-w. URL https://doi.org/10.1038/s41467-022-29443-w.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. URL https://arxiv.org/abs/2208.01618.

Vladimir Gligorijević, P. Douglas Renfrew, Tomasz Kosciolek, Julia Koehler Leman, Daniel Berenberg, Tommi Vatanen, Chris Chandler, Bryn C. Taylor, Ian M. Fisk, Hera Vlamakis, Ramnik J. Xavier, Rob Knight, Kyunghyun Cho, and Richard Bonneau. Structure-based protein function

prediction using graph convolutional networks. *Nature Communications*, 12(1):3168, May 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-23303-9. URL https://doi.org/10.1038/s41467-021-23303-9.

Joe G Greener and Kiarash Jamali. Fast protein structure searching using structure graph embeddings. *Bioinformatics Advances*, 5(1):vbaf042, 03 2024. ISSN 2635-0041. doi: 10.1093/bioadv/vbaf042. URL https://doi.org/10.1093/bioadv/vbaf042.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jiahao Hu, Jiasheng Cao, Win Topatana, Sarun Juengpanich, Shijie Li, Bin Zhang, Jiliang Shen, Liuxin Cai, Xiujun Cai, and Mingyu Chen. Targeting mutant p53 for cancer therapy: direct and indirect strategies. *Journal of Hematology & Oncology*, 14(1):157, Sep 2021. ISSN 1756-8722. doi: 10.1186/s13045-021-01169-0. URL https://doi.org/10.1186/s13045-021-01169-0.

Chen Jin, Ryutaro Tanno, Amrutha Saseendran, Tom Diethe, and Philip Alexander Teare. An image is worth multiple words: Discovering object level concepts using multi-concept prompt learning. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=F3x6uYILgL.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, Aug 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

Yogesh Kalakoti and Björn Wallner. Afsample2 predicts multiple conformations and ensembles with alphafold2. *Communications Biology*, 8(1):373, Mar 2025. ISSN 2399-3642. doi: 10.1038/s42003-025-07791-9. URL https://doi.org/10.1038/s42003-025-07791-9.

Shaun M Kandathil, Andy M Lau, Daniel W A Buchan, and David T Jones. Foldclass and merizosearch: scalable structural similarity search for single- and multi-domain proteins using geometric learning. *Bioinformatics*, 41(5):btaf277, 05 2025. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaf277. URL https://doi.org/10.1093/bioinformatics/btaf277.

Andy M. Lau, Nicola Bordin, Shaun M. Kandathil, Ian Sillitoe, Vaishali P. Waman, Jude Wells, Christine A. Orengo, and David T. Jones. Exploring structural diversity across the protein universe with the encyclopedia of domains. *Science*, 386(6721):eadq4946, 2024. doi: 10.1126/science.adq4946. URL https://www.science.org/doi/abs/10.1126/science.adq4946.

Jinwoo Leem, Laura S. Mitchell, James H.R. Farmery, Justin Barton, and Jacob D. Galson. Deciphering the language of antibodies using self-supervised learning. *Patterns*, 3(7):100513, 2022. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2022.100513. URL https://www.sciencedirect.com/science/article/pii/S2666389922001052.

Julia Koehler Leman, Brian D Weitzner, Steven M Lewis, Jared Adolf-Bryfogle, Nawsad Alam, Rebecca F Alford, Melanie Aprahamian, David Baker, Kyle A Barlow, Patrick Barth, et al. Macromolecular modeling and design in rosetta: recent methods and frameworks. *Nature methods*, 17 (7):665–680, 2020.

Yeqing Lin, Minji Lee, Zhao Zhang, and Mohammed AlQuraishi. Out of many, one: Designing and scaffolding proteins at the scale of the structural universe with genie 2, 2024. URL https://arxiv.org/abs/2405.15489.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

Shitong Luo, Yufeng Su, Xingang Peng, Sheng Wang, Jian Peng, and Jianzhu Ma. Antigen-specific antibody design and optimization with diffusion-based generative models for protein structures. *Advances in Neural Information Processing Systems*, 35:9754–9767, 2022.

Daniel J. Mandell, Evangelos A. Coutsias, and Tanja Kortemme. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. *Nature Methods*, 6:551–552, 2009. doi: 10.1038/nmeth0809-551.

Kevin Michalewicz, Mauricio Barahona, and Barbara Bravi. Integrating protein sequence embeddings with structure via graph-based deep learning for the prediction of single-residue properties, 2025. URL https://arxiv.org/abs/2502.17294.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards accurate and efficient binding affinity prediction. *bioRxiv*, 2025. doi: 10.1101/2025.06.14.659707.

Valerio Piomponi, Alberto Cazzaniga, and Francesca Cuturello. Evolutionary constraints guide alphafold2 in predicting alternative conformations and inform rational mutation design. *Journal of Chemical Information and Modeling*, 65(18):9459–9468, 2025. doi: 10.1021/acs.jcim.5c01090. URL https://doi.org/10.1021/acs.jcim.5c01090. PMID: 40902999.

Gopalasamudram N. Ramachandran, Chandrasekharan Ramakrishnan, and Viswanathan Sasisekharan. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7(1):95–99, 1963. ISSN 0022-2836. doi: https://doi.org/10.1016/S0022-2836(63)80023-6. URL https://www.sciencedirect.com/science/article/pii/S0022283663800236.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, Jan 2020. ISSN 1476-4687. doi: 10.1038/s41586-019-1923-7. URL https://doi.org/10.1038/s41586-019-1923-7.

Amelie Stein and Tanja Kortemme. Improvements to robotics-inspired conformational sampling in rosetta. *PLoS ONE*, 8(5):e63090, 2013. doi: 10.1371/journal.pone.0063090.

Brian L. Trippe, Jason Yim, Doug Tischer, David Baker, Tamara Broderick, Regina Barzilay, and Tommi S. Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6TxBxqNME1Y.

Christopher H. van Dyck. Anti-amyloid-β monoclonal antibodies for alzheimer's disease: Pitfalls and promise. *Biological Psychiatry*, 83(4):311–319, 2018. ISSN 0006-3223. doi: https://doi.org/10.1016/j.biopsych.2017.08.010. URL https://www.sciencedirect.com/science/article/pii/S0006322317318978. Mechanisms of Alzheimer's Disease and Treatment.

Anna Vangone and Alexandre M. J. J. Bonvin. Contacts-based prediction of binding affinity in protein–protein complexes. *eLife*, 4:e07454, jul 2015. ISSN 2050-084X. doi: 10.7554/eLife.07454. URL https://doi.org/10.7554/eLife.07454.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Susana Vázquez Torres, Philip J. Y. Leung, Preetham Venkatesh, Isaac D. Lutz, Fabian Hink, Huu-Hien Huynh, Jessica Becker, Andy Hsien-Wei Yeh, David Juergens, Nathaniel R. Bennett, Andrew N. Hoofnagle, Eric Huang, Michael J. MacCoss, Marc Expòsit, Gyu Rie Lee, Asim K. Bera, Alex Kang, Joshmyn De La Cruz, Paul M. Levine, Xinting Li, Mila Lamb, Stacey R. Gerben, Analisa Murray, Piper Heine, Elif Nihal Korkmaz, Jeff Nivala, Lance Stewart, Joseph L. Watson, Joseph M. Rogers, and David Baker. De novo design of high-affinity binders of bioactive helical peptides. *Nature*, 626(7998):435–442, Feb 2024. ISSN 1476-4687. doi: 10.1038/s41586-023-06953-1. URL https://doi.org/10.1038/s41586-023-06953-1.

Susana Vázquez Torres, Melisa Benard Valle, Stephen P Mackessy, Stefanie K Menzies, Nicholas R Casewell, Shirin Ahmadi, Nick J Burlet, Edin Muratspahić, Isaac Sappington, Max D Overath, et al. De novo designed proteins neutralize lethal snake venom toxins. *Nature*, 639(8053):225–231, 2025.

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, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, Aug 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06415-8. URL https://doi.org/10.1038/s41586-023-06415-8.

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Noah Getz, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Liam Atkinson, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, and Regina Barzilay. Boltz-1: Democratizing biomolecular interaction modeling. *bioRxiv*, 2024. doi: 10.1101/2024.11.19.624167.

Kevin E Wu, Kevin K Yang, Rianne van den Berg, Sarah Alamdari, James Y Zou, Alex X Lu, and Ava P Amini. Protein structure generation via folding diffusion. *Nature communications*, 15(1): 1059, 2024.

Luhuan Wu, Brian L Trippe, Christian A. Naesseth, David M Blei, and John P Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. *arXiv preprint arXiv:2306.17775*, 2023.

Li C. Xue, João P. G. L. M. Rodrigues, Panagiotis L. Kastritis, Alexandre M. J. J. Bonvin, and Anna Vangone. Prodigy: a web server for predicting the binding affinity of protein–protein complexes. *Bioinformatics*, 32(23):3676–3678, 08 2016. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw514. URL https://doi.org/10.1093/bioinformatics/btw514.

Kevin K Yang, Zachary Wu, Claire N Bedbrook, and Frances H Arnold. Learned protein embeddings for machine learning. *Bioinformatics*, 34(15):2642–2648, 03 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty178. URL https://doi.org/10.1093/bioinformatics/bty178.

Cheng Zhang, Adam Leach, Thomas Makkink, Miguel Arbesú, Ibtissem Kadri, Daniel Luo, Liron Mizrahi, Sabrine Krichen, Maren Lang, Andrey Tovchigrechko, Nicolas Lopez Carranza, Uğur Şahin, Karim Beguir, Michael Rooney, and Yunguan Fu. FrameDiPT: SE(3) Diffusion Model for Protein Structure Inpainting. *bioRxiv*, 2023. doi: 10.1101/2023.11.21.568057. URL https://www.biorxiv.org/content/10.1101/2023.11.21.568057v2.

Yang Zhang and Jeffrey Skolnick. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, 57(4):702–710, 2004. doi: https://doi.org/10.1002/prot.20264. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.20264.

A REPRESENTATION AND CONDITIONING ARCHITECTURE

RFdiffusion, like RoseTTAFold, maintains three representation "tracks": an MSA track v^{MSA}, a pair track v^{pair}, and a per-residue state track v^{state} (Baek et al., 2021; Watson et al., 2023). For a protein of length L, we write these as

$$v^{\text{MSA}} \in \mathbb{R}^{d_{\text{MSA}} \times L}, \quad v^{\text{pair}} \in \mathbb{R}^{d_{\text{pair}} \times L \times L}, \quad v^{\text{state}} \in \mathbb{R}^{d_{\text{state}} \times L}. \tag{6}$$

In all experiments we instantiate PGEL's denoiser on the released RFdiffusion "Base" model (`Base_ckpt.pt`), whose public configuration specifies $d_{\text{state}} = 16$, $d_{\text{MSA}} = 256$ and $d_{\text{pair}} = 128$; we keep all model weights and these dimensions fixed.

We partition the sequence positions into motif residues R_* with $|R_*| = L_*$ and scaffold residues R_c with $|R_c| = L_c$, so that $L = L_c + L_*$. Running the frozen RFdiffusion encoder on the scaffold with sequence s_c and structure x_c yields scaffold embeddings

$$v_c^{\text{MSA}} \in \mathbb{R}^{d_{\text{MSA}} \times L_c}, \quad v_c^{\text{pair}} \in \mathbb{R}^{d_{\text{pair}} \times L_c \times L_c}, \quad v_c^{\text{state}} \in \mathbb{R}^{d_{\text{state}} \times L_c}. \tag{7}$$

PGEL introduces learnable motif embeddings

$$v_* = \big(v_*^{\text{MSA}},\, v_*^{\text{pair}},\, v_*^{\text{state}}\big), \tag{8}$$

where $v_*^{\text{MSA}} \in \mathbb{R}^{d_{\text{MSA}} \times L_*}$ and $v_*^{\text{state}} \in \mathbb{R}^{d_{\text{state}} \times L_*}$. $v_*^{\text{pair}} \in \mathbb{R}^{d_{\text{pair}} \times L \times L_*}$ collects all pairwise features involving at least one motif residue (motif-motif and motif-scaffold blocks). The only trainable parameters in PGEL are the entries of v_*; all RFdiffusion parameters are frozen.

At each optimization step, we construct composite tracks that align with the RFdiffusion denoiser inputs by combining scaffold and motif embeddings. For the MSA and state tracks we concatenate along the sequence dimension:

$$v^{\text{MSA}} = \text{concat}\big(v_c^{\text{MSA}},\, v_*^{\text{MSA}}\big) \in \mathbb{R}^{d_{\text{MSA}} \times L}, \quad v^{\text{state}} = \text{concat}\big(v_c^{\text{state}},\, v_*^{\text{state}}\big) \in \mathbb{R}^{d_{\text{state}} \times L}. \tag{9}$$

The full pair tensor $v^{\text{pair}} \in \mathbb{R}^{d_{\text{pair}} \times L \times L}$ is obtained by keeping the scaffold-scaffold block equal to v_c^{pair} and inserting v_*^{pair} into all rows/columns corresponding to motif positions. The frozen RFdiffusion denoiser then takes the noisy coordinates $x(t)$ together with $\big(v^{\text{MSA}}, v^{\text{pair}}, v^{\text{state}}\big)$ and produces a prediction $\hat{x}^{(0)}(v_*)$. We use the motif slice $\hat{x}_*^{(0)}(v_*)$ in the PGEL loss, while the scaffold coordinates are clamped to x_c when applying the reverse diffusion update from $x(t)$ to $x(t-1)$.

Information exchange between the learned motif embedding and the fixed scaffold embedding occurs entirely through the existing RFdiffusion blocks: global self-attention in the MSA track, MSA-pair and pair-state couplings, and the $SE(3)$-equivariant refinement network. Gradients from the structural and torsion losses propagate through these layers to update v_*, while v_c and all network weights remain fixed.

B MASKING

Multiple-sequence-alignment (MSA) based structure predictors, such as AlphaFold, RoseTTAFold and RFdiffusion, derive a set of MSA embeddings that encode evolutionary covariation between residue positions and therefore impose strong geometric constraints on the predicted structure. In RFdiffusion these embeddings can be represented as a matrix $v_c \in \mathbb{R}^{d_{\text{MSA}} \times L_c}$, where d_{MSA} is the MSA embedding depth and L_c is the number of scaffold residues. Motivated by recent work that uses MSA column masking in AlphaFold-based pipelines to weaken coevolutionary constraints and increase conformational diversity (Kalakoti & Wallner, 2025; Piomponi et al., 2025), we introduce a simple masking operator $\mathcal{M}_{\omega,\alpha}(\cdot)$ that perturbs the scaffold MSA embedding during PGEL generation.

Specifically, given a masking type $\omega \sim \text{Ber}(\frac{1}{2})$ and a masking rate $\alpha \sim \mathcal{U}[0, 1]$ we construct a binary mask on v_c:

- Row masking ($\omega = 0$) selects a fraction α of rows of v_c (MSA embedding features) and sets them to zero for all scaffold residues. This reduces the effective dimensionality of the MSA feature space while preserving which residues are constrained.

- Column masking ($\omega = 1$) selects a fraction α of columns of v_c (scaffold residue positions) and sets all their MSA features to zero. This locally removes the evolutionary context of specific scaffold residues while leaving the remaining scaffold positions fully constrained.

In both cases, masking weakens the coevolutionary signals that couple the scaffold to the motif and thereby relaxes the induced geometric constraints. In our implementation, masking is applied only to the scaffold MSA embedding: the scaffold state and pair embeddings, as well as the learned motif embedding v_*, are left unmodified so any changes in the motif under different masks are mediated by the way EvoFormer attention blocks mix MSA, state, and pair embeddings. The frozen RFdiffusion denoiser therefore receives two conditioning signals at each reverse step: the perturbed scaffold MSA embedding $\mathcal{M}_{\omega,\alpha}(v_c)$ and the learned motif embedding v_*.

Empirically, we observe that column masking ($\omega = 1$) is substantially more effective than row masking in generating diverse yet designable motifs (around 80% of successful generations in Table 4 employ column masking). This is consistent with the interpretation above: while row masking removes global features that are shared across all residues and therefore tends to preserve the relative pattern of constraints, zeroing entire columns selectively weakens constraints arising from a subset of scaffold residues that are strongly coevolving with the motif and which enforce the native motif geometry. (This local relaxation allows the motif to move relative to the scaffold frame, while the rest of the scaffold remains tightly constrained by the unmasked columns). An ablation where we disable masking altogether ($\mathcal{M}_{\omega,\alpha} \equiv \text{Id}$) shows that, while PGEL still produces designable structures, the diversity collapses to a single TM-score cluster per system at threshold $t_m = 0.6$, confirming that masking is necessary to move beyond the native motif basin while maintaining designability (Appendix D).

C REVERSESTEP ALGORITHM

Let $x^{(t)} = \{(r_l^{(t)}, u_l^{(t)})\}_{l=1}^L$ denote the noisy protein backbone structure at diffusion step t, where each residue l is represented by a rotation $r_l^{(t)} \in SO(3)$, with $SO(3)$ the special orthogonal group in three dimensions, and a translation $u_l^{(t)} \in \mathbb{R}^3$. Let $\hat{x}^{(0)} = \{(\hat{r}_l^{(0)}, \hat{u}_l^{(0)})\}_{l=1}^L$ denote the predicted denoised structure. Let $\{\beta^{(t)}\}_{t=1}^T$ be a variance schedule with $\gamma^{(t)} = 1 - \beta^{(t)}$ and $\bar{\gamma}^{(t)} = \prod_{s=1}^t \gamma^{(s)}$. For translations, let $u_l^{(t-1)}$ be sampled from a Gaussian distribution with covariance $\beta^{(t)} I_3$. For rotations, let s_l denote the score approximation presented in Watson et al. (2023), $\epsilon_{l,d}$ isotropic Gaussian perturbations and $\{f_d\}_{d=1}^3$ a basis of the Lie algebra $SO(3)$.

Algorithm 3 REVERSESTEP function (Watson et al., 2023)

Input: noisy structure $x^{(t)}$, denoised prediction $\hat{x}^{(0)}$.
Output: updated structure $x^{(t-1)}$.
for $l = 1, \ldots, L$ **do**
$(r_l^{(t)}, u_l^{(t)}) = x_l^{(t)}$
$(\hat{r}_l^{(0)}, \hat{u}_l^{(0)}) = \hat{x}_l^{(0)}$
$u_l^{(t-1)} \sim \mathcal{N}\left(\frac{\sqrt{\bar{\gamma}^{(t-1)}}\beta^{(t)}}{1-\bar{\gamma}^{(t)}} \hat{u}_l^{(0)} + \frac{\sqrt{\gamma^{(t)}}(1-\bar{\gamma}^{(t-1)})}{1-\bar{\gamma}^{(t)}} u_l^{(t)},\ \beta^{(t)} I_3 \right)$
// Updating rotations below
$s_l = \text{ROTATIONSCOREAPPROXIMATION}(r_l^{(t)}, \hat{r}_l^{(0)}, \sigma_t^2)$
$\epsilon_{l,1}, \epsilon_{l,2}, \epsilon_{l,3} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$
$r_l^{(t-1)} = r_l^{(t)} \exp_{I_3}\left\{ (\sigma_t^2 - \sigma_{t-1}^2) r_l^{(t)\top} s_l + \sqrt{\sigma_t^2 - \sigma_{t-1}^2} \sum_{d=1}^{3} \epsilon_{l,d} f_d \right\}$
$x_l^{(t-1)} = (r_l^{(t-1)}, u_l^{(t-1)})$
end for
Return $x^{(t-1)}$

D BASELINES

To demonstrate the necessity of both learned embeddings and zero masking, we consider simpler baselines in which one of these steps is omitted.

We first show that the presence of a fix scaffold, which informs the frozen denoiser via computed scaffold embeddings v_c (see Algorithm 1 and Appendix A), is insufficient to produce constraint-satisfying structures if the motif embeddings v_* are not learned. To test this, we skip Algorithm 1 and proceed directly to the generation with embedding masking (Algorithm 2). At each timestep, we now obtain the predicted motif structure $\hat{x}_*^{(0)}$ employing a masked version of the scaffold embeddings $\mathcal{M}_{\omega,\alpha}(v_c)$ and with v_* either set to zero or to random values drawn from a standard normal distribution. In both scenarios, we observe a total of 0 successful generations across all experiments. For zero embeddings, we obtain a mean mRMSD of 3.11Å and a mean pLDDT of 49 across the ten examples, while these values are 3.42Å and 44, respectively, for random embeddings.

It could be argued that the stochasticity of RFdiffusion's denoiser, even with frozen parameters, is sufficient to generate diversified motifs after embedding learning. Nevertheless, when computing at each timestep the predicted motif structure as $\hat{x}_*^{(0)} = \text{DENOISER}(v_c, v_*)$, only a single self-consistent cluster is obtained at $t_m = 0.6$ per experiment.

E MASKING RATE AND ITS IMPACT ON STRUCTURAL DIVERSITY

We studied the effect of the masking rate $\alpha \in [0, 1]$ on the resulting structural diversity, quantified by the TM-score between each generated motif and the native structure. We plotted $1 -$ TM-score as a function of α for all the experiments, obtaining moderate correlations as measured by the Pearson correlation score r with statistical significance (Figure 6).

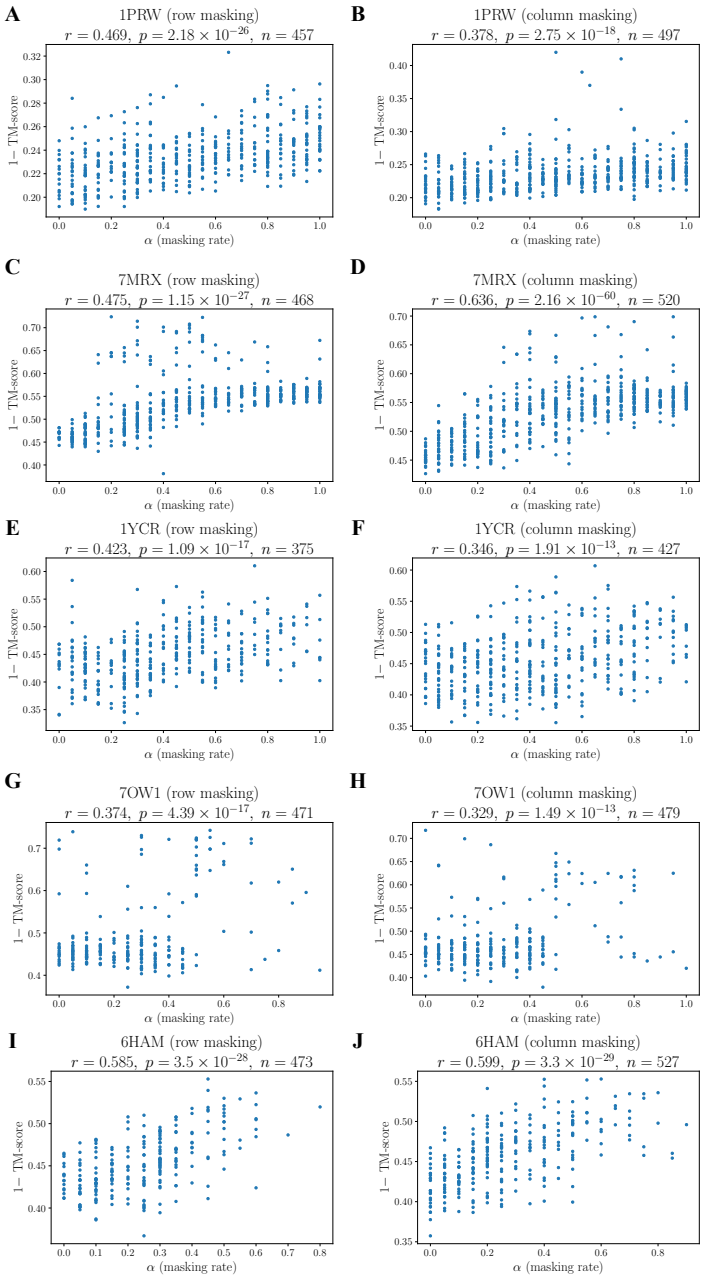


Figure 6: $1 -$ TM-score of each generated structure deemed designable (with respect to the native motif) as a function of the masking rate α, for row and column masking and five examples. For each case, we report the Pearson correlation coefficient r, its associated p-value, and the sample size.

F COMPUTATIONAL PERFORMANCE: STANDALONE *vs.* LINKED MOTIFS

We observed that the learning process described in Algorithm 1 requires more time when the designed motif is a standalone protein, *i.e.*, a full chain. The results indicate that the computational performance appears largely independent of the motif length L_* (Table 1).

Table 1: Learning time comparison for standalone *vs.* linked motif designs.

PDB	Standalone	L_*	Learning time (min.)
1PRW	×	42	3.20
7MRX	×	22	2.06
7OW1	×	8	4.18
6HAM	×	22	2.48
1YCR	✓	13	19.63
7WT5	✓	8	34.87
3MZW	✓	58	23.41
1ATP	✓	20	32.12

G STUDY OF REGULARIZATION HYPERPARAMETERS

We studied which combinations of λ_{DM} and $\lambda_{\mathrm{torsion}}$ minimize the mRMSD at convergence of Algorithm 1 (see equation 2). Further, we examined the contribution of individual loss terms $\mathcal{L}_{\mathrm{DM}}$ and $\mathcal{L}_{\mathrm{torsion}}$ by performing ablation experiments in which either the distance matrix term (equation 4) or the backbone torsion angle term (equation 5) were removed.

Although all combinations in Table 2 satisfy the $\mathrm{mRMSD} < 2$Å constraint, the configurations with $\lambda_{\mathrm{torsion}} = 0.05$ and $\lambda_{\mathrm{DM}} \in \{0.01, 0.05, 0.1\}$, as well as $\lambda_{\mathrm{torsion}} = \lambda_{\mathrm{DM}} = 0.1$ appear to provide the best performance. While these results offer some guidance on the appropriate orders of magnitude of the regularization strengths, the full validation pipeline should be run for each case to assess the number of successes.

Regarding the ablation study, in the three examples the pLDDT values are no longer constraint-satisfying at convergence, with a mean value of 65 across the three experiments, when $\mathcal{L}_{\mathrm{torsion}}$ is set to zero. When $\mathcal{L}_{\mathrm{DM}} = 0$, mRMSD values are worse at convergence (as also reflected in the first column of Table 2, and the center of mass suffers a progressive deviation from that of the native structure as the iterations progress.

Table 2: Grid search over regularization strengths λ_{DM} and $\lambda_{\mathrm{torsion}}$. Each cell reports the motif RMSD at convergence, averaged across three test cases (PDB IDs: 1PRW, 7MRX and 6HAM).

λ_{DM} \ $\lambda_{\mathrm{torsion}}$	0.005	0.01	0.05	0.1	0.5
0.005	0.95 ± 0.32	0.99 ± 0.29	0.93 ± 0.19	0.94 ± 0.18	1.02 ± 0.21
0.01	1.82 ± 0.40	1.02 ± 0.29	0.77 ± 0.21	1.11 ± 0.30	1.04 ± 0.22
0.05	1.58 ± 0.37	0.98 ± 0.33	0.81 ± 0.24	0.96 ± 0.27	0.99 ± 0.24
0.1	1.03 ± 0.36	1.00 ± 0.31	0.78 ± 0.28	0.80 ± 0.29	0.83 ± 0.25
0.5	1.19 ± 0.42	1.15 ± 0.34	0.90 ± 0.17	0.87 ± 0.25	0.92 ± 0.22

H PGEL EVALUATION RESULTS

Table 3: Number of designable backbones (out of 1000) and self-consistent clusters (diversity) for PGEL and partial diffusion across benchmarking tasks.

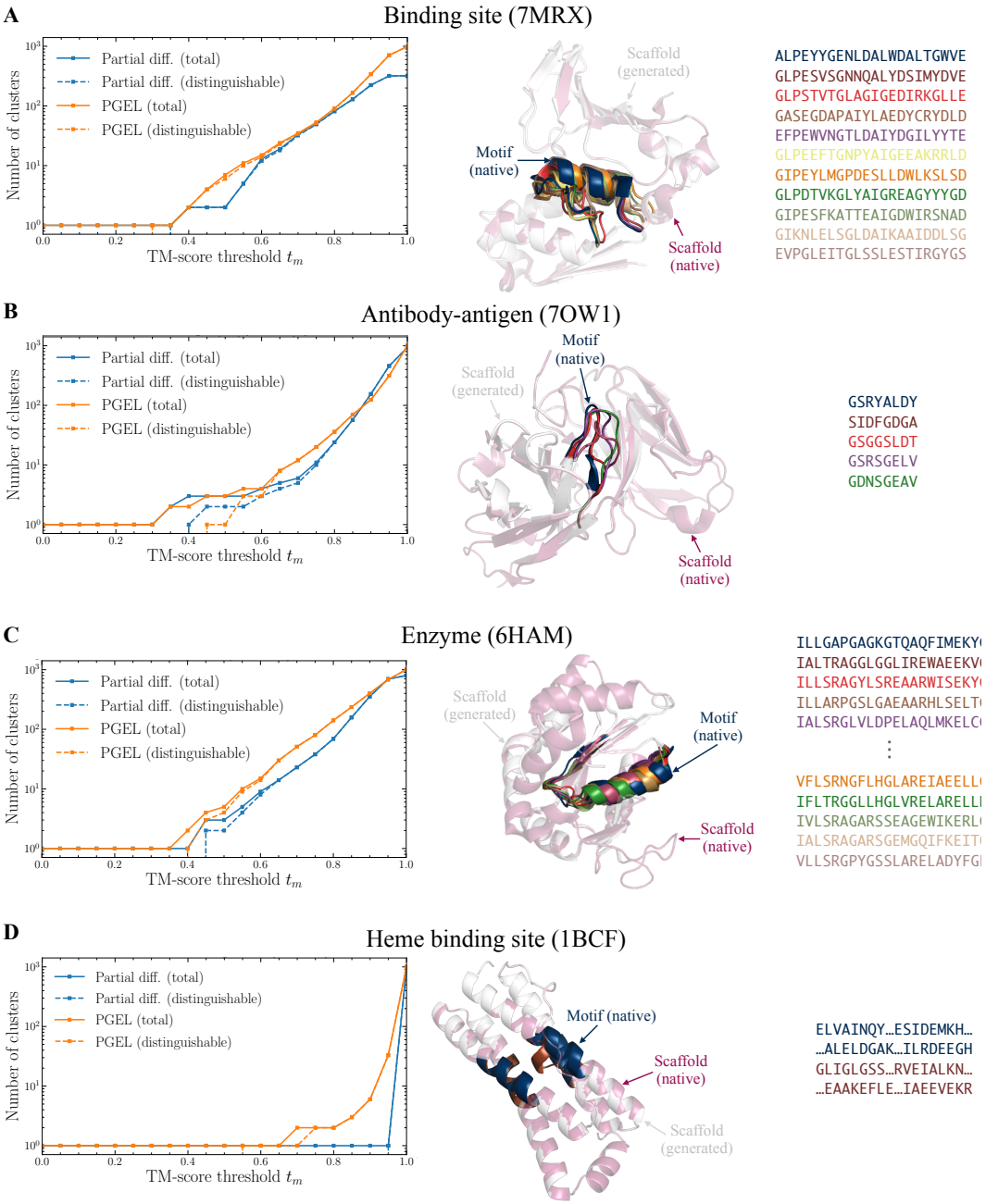
PDB	Designability		Diversity	
	Partial diff.	PGEL	Partial diff.	PGEL
1BCF	983	**1000**	0	**1**
1PRW	411	**1000**	0	**2**
1YCR	252	**802**	1	**6**
3IXT	393	**502**	1	**3**
5IUS	434	**711**	0	**2**
6E6R	523	523	0	**13**
6EXZ	**983**	887	1	**5**
6HAM	798	**823**	8	**15**
7MRX	331	**990**	6	**10**
7OW1	948	**950**	2	**4**

Table 4: Detailed results of PGEL successes.

PDB & design ID	α	ω	mRMSD (Å)	Motif pLDDT	Sequence	mRMSD AF3 (Å)	pAE
1BCF							
699	0.9	0	1.99	76	GLIGLGSS…RVEIALKN…EAAKEFLE…IAEEVEKR	1.97	2.17
1PRW							
17	0.75	1	1.78	79	FRVIAGGEDGLVTLEQLARY/VRRVAGRGGRLISFEDFLAI	1.54	4.43
860	0.5	1	1.96	81	ARWLDKGGSGAVFGEQLGEF/VAAALEGGKEAKLEEWFLNY	1.25	4.84
1YCR							
14	0.3	1	1.36	75	GSLEELLAQLDGG	1.56	2.88
36	0.7	1	1.09	72	MGLMELLLGVPGS	1.25	3.19
285	0.3	1	1.59	72	ASGLELLKKFKLP	1.60	3.53
334	0.65	1	1.03	72	SSLMELLEKIDIE	1.25	2.88
619	0.2	1	0.59	87	ESFEELWKKIPGS	1.70	2.45
695	0.25	1	1.04	74	SSMWELWQEIEGE	0.86	2.42
3IXT							
112	0.85	0	1.23	75	NLTAVTAYFNSGVTPDELRALARN	0.97	3.49
118	0.1	0	1.03	74	NEELYKEISKTPFNGEEWIKLARN	1.74	4.09
542	0.4	1	1.14	70	NAQKAKEALDSGFNGEELRKVFSN	1.83	3.82
5IUS							
348	0.75	1	1.8	71	YSLQVKKTLPDGTVVTVATT…SRFTVTLIDFNGNIRLLESEPA	0.71	3.61
782	0.4	1	1.53	72	PVALCKRVLPNSSEYISGQT…GPARCVYIDLVPGGSFQEGGTA	1.67	4.03
6E6R							
16	0.35	1	1.90	79	TEEEVLGEKKGVC	1.03	2.55
95	0.25	0	1.85	73	TEEECLKEKVGVC	1.97	1.62
97	0.25	0	1.72	81	KEEEVLKEKLKVC	1.99	1.96
121	0.6	0	1.02	75	KGWEELRKKVGLC	0.98	1.88
125	0.05	0	1.58	84	EPEKCLTEKLGVC	1.08	1.68
126	0.15	0	1.69	74	KEEECLKEKLKVC	1.90	1.70
201	0.4	0	1.93	75	TEEEILSGKYGLC	1.35	1.98
205	0.05	1	1.62	80	KGDECLKEKIGVC	1.82	1.66
306	0.1	0	1.86	82	TPEEVLTKVLGLC	0.89	1.80
522	0.45	0	0.93	70	KTFEEIKEKLGCF	0.94	1.84

Continued on next page

PDB & design ID	α	ω	mRMSD (Å)	Motif pLDDT	Sequence	mRMSD AF3 (Å)	pAE
600	0.05	0	0.63	88	KSFEEVNKKLGIT	0.51	1.50
695	0.25	0	1.16	70	RTLKCVFEKLGIL	0.95	1.53
776	0.05	0	0.95	83	KTFEEVKKKLPGL	0.48	2.32
6EXZ							
42	0.75	0	1.25	76	LYSLFPGGFKVTLLF	1.35	3.58
166	0.6	1	1.11	85	LSTEIKGGVEYILKL	1.06	3.30
325	0.15	0	1.35	84	LYTLAPGLLTLCIAL	1.02	3.73
448	0.05	1	0.98	83	LSELFPGGVKLVVAL	0.89	4.71
699	0.5	1	1.82	74	LWQHAPGAGGLLLVG	1.98	4.56
6HAM							
40	0.45	1	1.29	76	IALTRAGGLGGLIREWAEEKVG	1.09	4.08
64	0.15	1	0.80	87	ILLSRAGYLSREAARWISEKYG	1.40	4.33
104	0.15	0	1.23	71	IALTNAGWLDGLIAEFMKEKTG	1.88	3.45
115	0.05	1	1.10	86	ILLARPGSLGAEAARHLSELTG	0.90	4.64
131	0.55	0	1.94	78	IALSRGLVLDPELAQLMKELCG	1.98	3.71
257	0.25	1	1.28	71	IALERAGYRDRLIKELGKELLG	1.40	4.14
313	0.25	1	1.78	71	IALSTKGGLSGLIADFAKEVLG	1.43	4.51
578	0.05	1	1.88	81	VLLHRPGAADELARWLAEKVGG	0.83	3.75
653	0.6	0	1.85	74	ICLSRAGVFSGLFREIAEEFGK	0.89	3.88
658	0.1	1	1.16	81	IVLSRPGANGSVAREYAKEKLG	0.79	4.67
669	0.3	1	1.75	72	VFLSRNGFLHGLAREIAEELLG	1.46	4.38
694	0.5	0	1.89	73	IFLTRGGLLHGLVRELARELLK	1.53	4.31
716	0.05	1	1.08	81	IVLSRAGARSSEAGEWIKERLG	1.22	4.19
839	0.15	1	1.28	81	IALSRAGARSGEMGQIFKEITG	1.19	4.30
937	0.4	1	1.80	71	VLLSRGPYGSSLARELADYFGL	1.49	4.92
7MRX							
0	0.55	0	0.74	79	GLPESVSGNNQALYDSIMYDVE	0.89	3.17
31	0.4	0	1.35	78	GLPDTVKGLYAIGREAGYYYGD	0.83	3.30
100	0.95	1	1.26	73	GLPSTVTGLAGIGEDIRKGLLE	1.78	3.73
114	0.75	1	0.83	75	GASEGDAPAIYLAEDYCRYDLD	1.22	3.79
145	0.4	0	0.78	78	EFPEWVNGTLDAIYDGILYYTE	0.69	4.93
308	0.5	1	1.57	74	GIPESFKATTEAIGDWIRSNAD	1.25	4.97
352	0.85	1	1.53	75	GLPEEFTGNPYAIGEEAKRRLD	1.98	3.81
730	0.8	1	1.85	72	GIPEYLMGPDESLLDWLKSLSD	1.42	4.90
744	0.8	1	0.88	75	EVPGLEITGLSSLESTIRGYGS	1.96	2.42
814	0.3	1	0.77	78	GIKNLELSGLDAIKAAIDDLSG	1.13	3.79
7OW1							
136	0.2	1	1.92	70	SIDFGDGA	1.93	4.92
287	0.05	1	0.53	91	GSGGSLDT	0.62	3.54
637	0.35	0	1.00	77	GSRSGELV	0.77	3.59
837	0.15	1	1.32	77	GDNSGEAV	0.99	4.26


A
Binding site (7MRX)
Partial diff. (total)
Partial diff. (distinguishable)
PGEL (total)
PGEL (distinguishable)
Number of clusters
TM-score threshold t_m
Scaffold (generated)
Motif (native)
Scaffold (native)
ALPEYYGENLDALWDALTGWVE
GLPESVSGNNQALYDSIMYDVE
GLPSTVTGLAGIGEDIRKGLLE
GASEGDAPAIYLAEDYCRYDLD
EFPEWVNGTLDAIYDGILYYTE
GLPEEFTGNPYAIGEEAKRRLD
GIPEYLMGPDESLLDWLKSLSD
GLPDTVKGLYAIGREAGYYYGD
GIPESFKATTEAIGDWIRSNAD
GIKNLELSGLDAIKAAIDDLSG
EVPGLEITGLSSLESTIRGYGS
B
Antibody-antigen (7OW1)
GSRYALDY
SIDFGDGA
GSGGSLDT
GSRSGELV
GDNSGEAV
C
Enzyme (6HAM)
ILLGAPGAGKGTQAQFIMEKY
IALTRAGGLGGLIREWAEEKV
ILLSRAGYLSREAARWISEKY
ILLARPGSLGAEAARHLSELT
IALSRGLVLDPELAQLMKELC
VFLSRNGFLHGLAREIAEELL
IFLTRGGLLHGLVRELARELL
IVLSRAGARSSEAGEWIKERL
IALSRAGARSGEMGQIFKEIT
VLLSRGPYGSSLARELADYFG
D
Heme binding site (1BCF)
Motif (native)
Scaffold (native)
Scaffold (generated)
ELVAINQY…ESIDEMKH…
…ALELDGAK…ILRDEEGH
GLIGLGSS…RVEIALKN…
…EAAKEFLE…IAEEVEKR

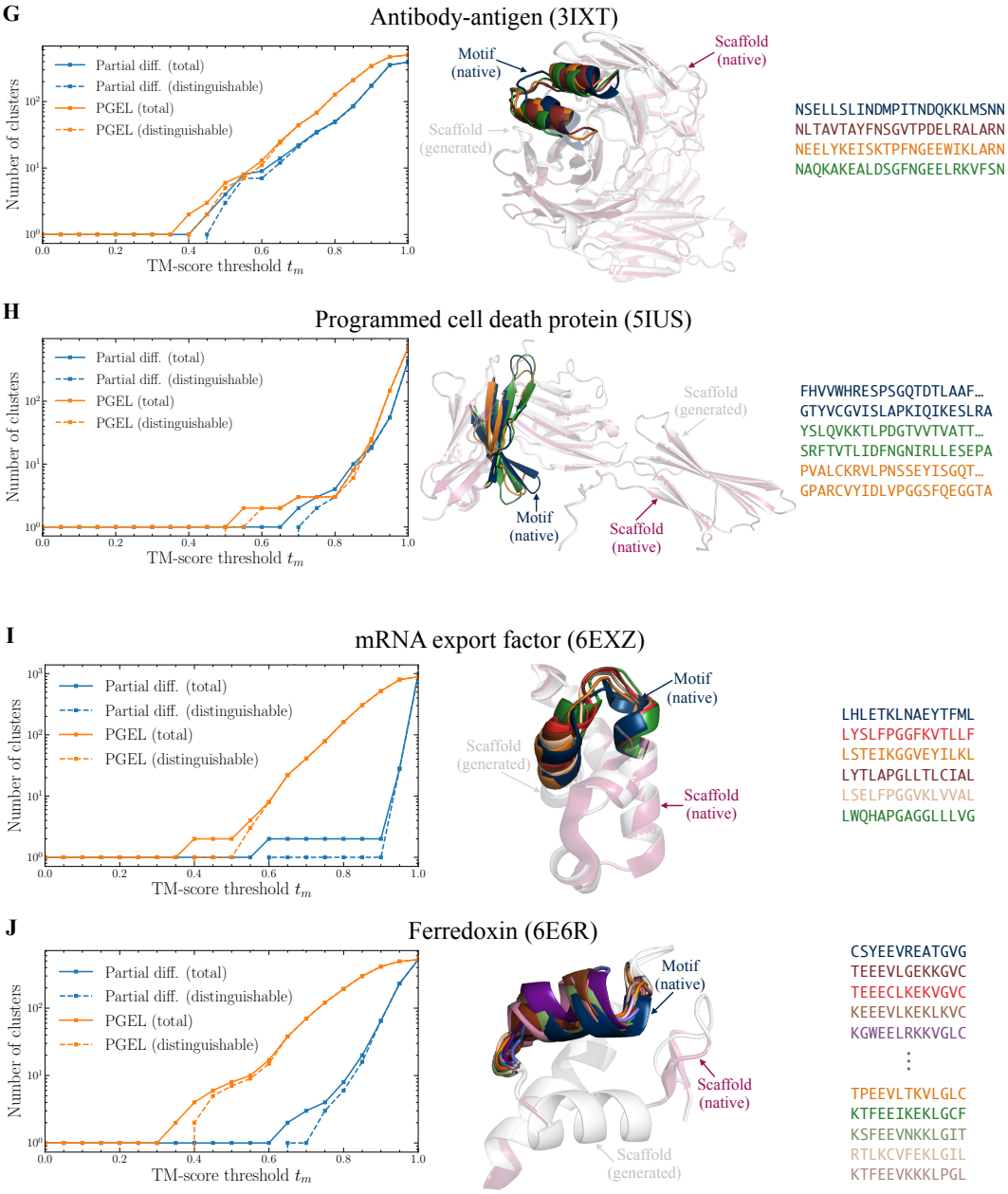


Figure 7: Same as Figure 5 but for the remaining examples. *Left:* Number of clusters identified by PGEL and partial diffusion, both total and distinguishable from native, as a function of the TM-score threshold t_m. *Right:* Generated binding site backbones (overlaid with native), alongside the native sequence and sequences that refold to self-consistent structures.

I BINDING AFFINITY RESULTS

Table 5: PRODIGY-predicted binding affinities for the three examples that correspond to protein-protein complexes (PDBs 1YCR, 5IUS and 7MRX).

PDB & design ID	Method	ΔG (kcal/mol)
1YCR		
Native	–	-7.7
14	PGEL	-7.1
36	PGEL	-5.9
285	PGEL	-6.8
334	PGEL	-6.8
619	PGEL	-6.6
695	PGEL	-7.6
101	Partial diff.	-6.4
5IUS		
Native	–	-9.7
348	PGEL	-10.0
782	PGEL	-9.7
7MRX		
Native	–	-11.4
0	PGEL	-9.3
31	PGEL	-11.4
100	PGEL	-10.8
114	PGEL	-9.5
145	PGEL	-10.1
308	PGEL	-10.4
352	PGEL	-12.8
730	PGEL	-11.6
744	PGEL	-9.5
814	PGEL	-9.7
1	Partial diff.	-9.5
18	Partial diff.	-10.9
307	Partial diff.	-8.5
327	Partial diff.	-8.1
513	Partial diff.	-9.7
780	Partial diff.	-9.4