arXiv:2506.08293v1 [q-bio.BM] 9 Jun 2025

Diffusion Sequence Models for Enhanced Protein Representation and Generation

Logan Hallee[1, 2], **Nikolaos Rafailidis**[1], **David B. Bichara**[2], **and Jason P. Gleghorn**[1, 2, 3 ✉]

[1]Center for Bioinformatics and Computational Biology, University of Delaware
[2]Synthyra, PBLLC
[3]Department of Biomedical Engineering, University of Delaware

Proteins are fundamental to biology, executing diverse functions through complex physicochemical interactions, and they hold transformative potential across medicine, materials science, and environmental applications. Protein Language Models (pLMs) aim to unlock insights from the vast space of unlabeled protein sequences by learning rich, semantic representations from primary sequences via masked language modeling. However, these models typically exhibit limited generative capacity. In this work, we introduce the Diffusion Sequence Model (DSM), a novel pLM trained with masked diffusion to enable both high-quality representation learning and generative protein design. DSM builds upon the ESM2 architecture by incorporating a masked forward diffusion process inspired by the LLaDA framework. After training, DSM is capable of generating diverse, biomimetic sequences that align with expected amino acid compositions, secondary structures, and predicted functions, even with 90% token corruption. Furthermore, DSM's learned representations match or exceed those of similarly sized pLMs on downstream tasks. We also introduce DSM$_{ppi}$, a variant fine-tuned to generate protein binders by attending to target sequences. We demonstrate DSM$_{ppi}$'s effectiveness on the challenging Bench-tested Binder Benchmark (BenchBB), where both DSM and DSM$_{ppi}$ produce candidates with superior predicted binding affinity compared to known binders. Our results establish masked diffusion as a powerful paradigm for unifying protein representation and generation in a single framework.

Mask Diffusion | Protein design | Protein binders | Protein Language Modeling | Annotation Vocabulary | Protein-Protein Interactions |

Correspondence: ***gleghorn@udel.edu***

1 Introduction

The evolution of protein systems occurs through gradual, mostly deleterious, and seemingly stochastic modifications to genetic sequences. The niche adaptation of biological systems via selection narrows the natural protein landscapes into a minuscule portion of the possible protein universe (1–3). In fact, exploring the total protein universe appears infeasible by any natural process with over 10^{1301} possible protein sequences of length 2,048. Even within the smaller space of known natural proteins, we understand a small portion mechanistically. A fraction of 1% of cataloged protein sequences have ever been expressed, and fewer have been well annotated (3–6). This disparity highlights the scarcity of experimental data, rendering protein design and functional annotation persistent challenges in the life sciences.

A deeper understanding of the protein universe is critical for advancing knowledge of disease mechanisms and biological systems (7–15). Moreover, designed proteins have other transformative impacts beyond biology, with potential to advance plastic degradation and recycling, carbon capture, and the creation of novel materials (16–18). Unfortunately, the experimental validation of protein properties remains time- and money-intensive; the same can be said about traditional protein design. Therefore, there is a crucial need for reliable computational systems that can annotate and generate protein sequences.

In response, the Protein Language Model (pLM) community has pursued these aims in tandem with the rapid advancement of natural language processing (NLP) systems. By treating proteins as a semantic language, various transformer-like neural networks are an effective choice to map gene products to meaningful numerical representations. Through amino acid, codon, nucleotide, or atomic inputs, researchers typically deploy semi-supervised masked language modeling (MLM) to pretrain models (19–25). The linear primary sequences of proteins fold into intricate 3D structures, so the global interaction relationships mapped with bidirectional attention have dominated as a modeling choice. The combination of bidirectional attention and MLM naturally encodes the protein structural information into the attention maps of pLMs when a large enough layer depth is used, offering an emergent phenomenon that enhances the representations from pLMs (26). However, while pLM representations can be used and fine-tuned for many downstream tasks, they do not inherently map well to abstract concepts like biological processes or fitness (3, 27). Additionally, they have poor generative modeling capabilities because they are typically only trained to denoise inputs from 15-20% noise (3, 28). While autoregressive (AR) models offer a solution to generative capabilities, AR-based pLMs typically have worse representations on downstream tasks (29, 30). Additionally, we argue that it is an ineffective modality for protein generation. While the linear left-to-right generative process mimics protein synthesis well, the phenotypic changes from mutagenesis and epistasis occur based on the conditional probability of global context. As such, implementation of bidirectional token information mixing is a beneficial choice (29–31).

To address the limitations of generative pLMs, we propose Diffusion Sequence Modeling (DSM), a framework that unifies biologically meaningful representation and generative modeling through masked diffusion. To achieve this, we modified the Large Language Diffusion Models (LLaDA) framework (32) to further pretrain popular pLMs using a masked-based diffusion process. DSM extends ESM2 (33) with a novel language modeling head and training objective, enabling robust denoising across high corruption rates and sequence generation with global context. After training, we observed low cross-entropy losses and accurate sequence reconstruction at low mask rates up to 90% corruption. With unconditional generative sampling, DSM models produced distinct sequence distributions that closely mimicked amino acid k-mers, predicted secondary structures, and predicted protein functions compared to natural sequences. In addition to their generative prowess, DSM models also produced high-quality protein sequence representations. DSM models match or outperform MLM-based and discrete diffusion pLMs (DPLM) of the same size as well as an AR pLM almost twice DSM's size. We further extended the DSM framework by finetuning DSM on sets of interacting protein pairs with the hope of leveraging target protein inputs to generate protein binders (DSM$_{ppi}$). Our case study with DSM$_{ppi}$ and DSM highlighted the potential of unconditional and conditional generation schemes for protein design, with both methods producing promising protein binders with higher predicted binding affinity than the best known binders in BenchBB (34). BenchBB provides a rigorous benchmark by featuring diverse protein targets representative of therapeutically relevant binder design, directly demonstrating DSM$_{ppi}$'s practical applicability (34). Together, these results establish masked diffusion as a natural fit for pLMs, offering a unified framework for high-quality representation and biologically coherent generation.

2 Background

2.1 LLaDa

The LLaDA framework introduces a diffusion-based alternative to standard AR generative modeling in NLP by training a transformer neural network with a masked diffusion forward process (32), offering bidirectional context utilization in contrast to AR models. This makes it particularly appealing for protein sequences, where functional properties often exhibit long-range dependencies in 1D but are actually close in 3D, only fully realized in forward passes with bidirectional context. More formally, LLaDa training assumes a neural network with parameters θ, $p_\theta(\cdot, x_t)$, that takes x_t as input and predicts all masked tokens M in a single forward pass. The objective is

$$\mathcal{L}(\theta) = -\mathbb{E}_{t,x_0,x_t}\left[\frac{1}{t+\epsilon}\sum_{i=1}^{L}\mathbf{1}\left[x_t^i = M\right]\log p_\theta\big(x_0^i \mid x_t\big)\right],$$

where x_0 is sampled from the data with sequence length L and $t \in [0, 1]$ is sampled uniformly, independently for each mini-batch. The indicator function $\mathbf{1}(\cdot)$ gathers only masked tokens, so that unmasked token logits are not included in the loss value (32). Importantly, the $\frac{1}{t}$ steers away from the traditional MLM training objective by penalizing the model more severely for low mask examples, scaling the loss consistently based on the difficulty of the input. A small ϵ prevents division by 0. Notably, this generative format complies with *Fisher consistency*, implying favorable scaling for larger models and datasets (32, 35).

After training, LLaDA models can perform a reverse process over a corpus of masked tokens. While the trained model unmasks all tokens every forward pass, some tokens can be remasked with various sampling methods to simulate a diffusion process. After fine-tuning, the eight billion parameter version of LLaDA performs on par with autoregressive large language models (LLMs) on several downstream benchmarks, including superior performance to LLaMA3 eight billion on MMLU, TruthfulQA, as well as math and coding tasks (32).

2.2 Annotation Vocabulary

The Annotation Vocabulary (AV) is a standardized combination of various protein-related ontologies organized into a machine-readable format. The original version utilizes Enzyme Commission (EC) numbers, Gene Ontologies (GO) for Biological Process (BP), Molecular Function (MF), and Cellular Compartment (CC), as well as Interpro (IP) and Gene3D (3D) domains (3). In this work, we expanded AV with UniProt keywords and cofactor information to further enhance the vocabulary (4). AV tokens are mapped to unique integers, enabling direct token embedding for various representation learning schemas. AV has previously been used to train protein representation and generation models (3). We annotated proteins in this work using the **Translator** model (**Supplemental A.1.2**), and used AV terms for downstream evaluation (3).

2.3 Alignment Score

Needleman-Wunsch global sequence alignment with evolutionarily informed scoring matrices is an ideal method for evaluating the similarity of two amino acid sequences, as it can recognize biologically similar motifs that are located at different positions. However, the score trends monotonically with the length of the inputs, providing an unscaled comparison that is not ideal for a standardized similarity metric. We previously developed a normalized Needleman-Wunsch score to design a more interpretable metric, called the Alignment Score (ASc) (3). Defined as $ASc(a, b) = \frac{l}{f(a,a)-f(a,b)+l}$, where $f(a, b)$ is the Needleman-Wunsch alignment score with BLOSUM62 between ground truth sequence a and generated sequence b, and l is the length of sequence a. The result is an error term between intra- and inter-sequence alignment in the denominator that reduces the score upon poor alignment. The score ranges from 0 (no similarity) to 1 (same sequence), with random sequence pairs typically scoring around 0.15, and similarity gradually approaching 1 (3). We use ASc throughout this work as a normalized measure of sequence recovery quality, especially at high corruption levels.

3 Methods

3.1 Computational resources

All experiments used a single A5000 GPU, A6000 GPU, GH200 GPU, or four A100 GPUs. The pretraining of DSM$_{650}$ was the most computationally intensive, requiring 12 days of 4xA100 time. The remaining experiments used 48 hours or less time on one of the mentioned hardware.

3.2 DSM pretraining

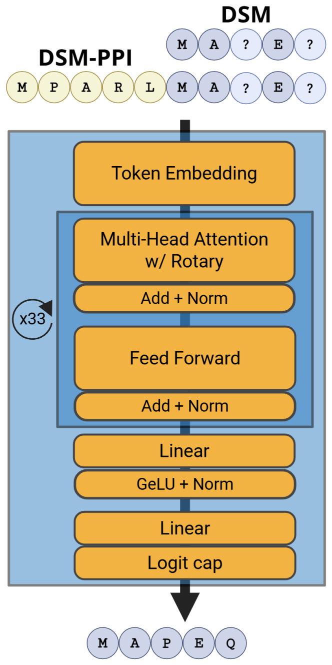

Fig. 1. Model architecture with DSM and DSM-PPI training scheme highlighted. Amino acids are masked uniformly and filled in completely with each forward pass, either with or without interacting protein context.

For DSM pretaining, we used OMG$_{prot50}$, a protein-only dataset containing over 207 million sequences after clustering the Open MetaGenomic dataset (OMG) at 50% sequence identity (36). We randomly removed 10,000 sequences for the validation and test sets each. We queried UniProt (release 2024_06) for newly deposited entries after 8/17/2024 (OMG release) with a minimum of transcript-level evidence. These 3,300 entries were added to the test set to accommodate for training sets with unknown composition before OMG use.

DSM models were extended from pretrained ESM2 checkpoints (33). To stabilize training and improve sampling quality, we modified the language modeling head by adding an additional linear transformation as well as a soft-logit cap (37, 38). We tied the final projection to the token embedding matrix during training, which has been shown to be beneficial for representation learning (39, 40).

During DSM training, we used a cosine learning rate scheduler with a 1,000-step linear warmup from zero (41). We trained two versions of DSM, the first to prototype our process from the ESM2 150 million parameter checkpoint (ESM2$_{150}$) for 100,000 steps with a batch size of 32, maximum sequence length of 512, and learning rate of $1e^{-4}$ using the AdamW optimizer (42). This resulted in the **DSM$_{150}$** checkpoint. The larger version of DSM used the ESM2$_{650}$ checkpoint with the same training scheme of DSM$_{150}$, except with a global batch size of 128 and a maximum sequence length of 2048 to produce **DSM$_{650}$**.

3.3 Model probing

Models were evaluated with a linear probe (**Supplemental Section A.2**) after embedding supervised datasets (**Supplemental Section A.3**) with mean pooling over the last hidden state. Importantly, we gauged model performance against two controls. One was a negative control, gathered by replacing the pLM embeddings with randomly generated vectors. The second control was based on an ESM2$_{35M}$ copy that has randomized weights. We previously established that running similar inputs through a randomized transformer would still place them relatively close in the final latent space (3, 43), presumably due to the self-attention mechanism attending to similar regions of input tokens regardless of the starting values of the token embeddings. This is particularly powerful for pLMs, where protein sequences share a lot of functional relevance based on sequence homology. So we viewed the `Random Transformer` control as a benchmark for the correlation based on basic sequence homology. To easily view the gain over the negative control metric r, we linearly normalized the current score p against the best score q by calculating $\bar{p} = \frac{p-r}{q-r}$, where 1.0 is the best performer (**Figure 5**, raw scores in **Supplemental Figure S4**). In total, we benchmarked DSM against ESM2, GLM2, ProtBert, ProtCLM$_{1B}$ (an AR pLM from Biomap), ESMC (ESM++), DPLM, ANKH, and ProtT5 (19, 28, 33, 36, 44–47).

3.3.1 Secondary structure prediction

Secondary structure performance was screened as a token-wise multiclass classification problem using the same training schema as the linear probes, but with a single block transformer that utilized full-residue embeddings instead. We chose a hidden size of 512 for the vanilla transformer block with rotary embeddings (48, 49). The best-performing base pLM was fully fine-tuned with the same training parameters, except with low rank adaptation (LoRA) applied to the attention layers instead of an external probe (50).

3.4 Distribution comparisons

To analyze and contrast the distributions of amino acids, secondary structures, and AV terms, we used a χ^2 test for independence alongside the Jensen-Shannon divergence (JS) (51, 52). For amino acids and secondary structures, we independently compared the 1-mers, 2-mers, and 3-mers. For AV, we used only 1-mers due to the large vocabulary size. We hypothesized that *de novo* generated proteins would inhabit distributions that reject the χ^2 null due to the high sample count and degrees of freedom, but showcase a low JS. We believed this would imply biologically mimetic but distinct distributions given low χ^2 p-values yet high similarity implied from JS. For χ^2, we used the raw counts, and for calculating JS we converted to frequencies after additive smoothing to avoid division by zero (53).

3.5 Unconditional generation analysis

Unconditional generation of DSM models was accomplished by feeding in only mask tokens of a chosen length surrounded by `[CLS]` and `[EOS]`. This procedure simulates a reverse diffusion process via progressive denoising. The total number of forward steps was calculated based on the number of tokens and a step divisor s, with approximately s tokens unmasked each step. Technically, every forward pass should fill in every mask after pretraining, so s tokens were chosen to keep, and the rest were remasked. There are many possible strategies to choose which s tokens to keep, such as logit or confidence-based ranking, topk, and beam search, etc. We observed in local experiments that pLMs are poorly conditioned to choose which position should be unmasked next; therefore, we used a random choice approach without search. To balance speed and quality, we swept over values of s and sampling temperatures, optimizing for the lowest JS between the validation set and generated amino acid 3-mer distributions.

To analyze the trends of DSM generation, we chose the 10,000 validation sequences as a reference for natural proteins. A *de novo* corpus was generated by looping through the validation set and generating a protein unconditionally based on the length of the protein in the validation set. The result was two corpora of amino acid sequences of exactly the same length, 9,989 total after removing entries less than 20 or greater than 2,048 long. Then, we applied the distribution comparisons mentioned above (**Section 3.4**).

3.6 Sequence reconstruction evaluation

We define sequence reconstruction as the ability for a pLM to reconstruct noised inputs at various corruption levels. To evaluate DSM on sequence reconstruction, we masked the pretraining validation and test sets with 5%, 15%, 30%, 50%, 70%, and 90% mask rates, then recorded the cross-entropy loss and weighted F1 score over the masked positions and ground truth, as well as ASc. For comparison we did the same for ESM2$_8$, ESM2$_{35}$, ESM2$_{150}$, ESM2$_{650}$, and ESM2 three billion (ESM2$_{3B}$). Importantly, the same random seeds ensured that identical masked positions across models were used for each run, allowing for a fair comparison.

3.7 DSM binder generation

A new Protein-Protein Interaction (PPI) dataset was compiled by downloading the StringDB version 12 entries for model organisms, including *Homo sapiens*, *Mus musculus*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Danio rerio*, *Caenorhabditis elegans*, *Escherichia coli* (K12), and *Pseudomonas aeruginosa* (PAO1) (54). Entries with a combined score less than 900 (90% confidence) were discarded, and the set of unique sequences was put aside. CD-hit was run on the sequence set, clustering at 90% sequence identity, and the representative entries were saved (55). We only retained PPI examples from our compiled list where both sequences were representatives. We split a random section of the dataset for evaluation and removed sequences matching an example in the training set. The final dataset consisted of 646,000 high-quality PPI entries for training, a random split of 5,850 entries for validation, and a data-leakage-free test set of 1,320 entries.

After data compilation and pretraining, we fine-tuned the DSM models to generate a candidate interacting sequence (SeqB) conditioned on a fixed target sequence (SeqA). We reasoned that most protein interactions take place through some type of inter-protein binding, so this training process simulated the task of *de novo* binder design. This was achieved by masking the second sequence in the PPI entry. The format `[CLS]-SeqA-[EOS]-[MASKED SeqB]-[EOS]` was used for the same exact objective as pretraining. The DSM$_{150-ppi}$ and DSM$_{650-ppi}$ were trained with LoRA on the attention layers for one epoch of the PPI dataset (50). A LoRA r of 8, α of 32, and dropout of 0.01 were used. We trained a control version of DSM$_{150-ppi}$ with the same exact hyperparameters but with inputs `[CLS]-SeqB-[EOS]` only. The control model omits conditioning on SeqA, allowing us to confirm that improvements are due to context-awareness rather than simple additional fine-tuning. Importantly, the order of SeqA and SeqB was switched randomly throughout training, but not testing.

We confirmed the ability for DSM to utilize SeqA context to get better at reconstructing SeqB by applying the sequence reconstruction pipeline (3.6) of ESM2 and DSM models on the PPI test set at a 15% mask rate. ESM2 and base DSM models received only SeqB, DSM$_{150-ppi-control}$ also received only SeqB, and the DSM$_{ppi}$ versions received both sequences.

We established a pipeline for template-based design in an attempt to use DSM to create stronger binders. To explore well-known target-binder pairs, we conducted an extensive literature review for the targets in the brand-new BenchBB protein binder benchmark (34). BenchBB is a compilation of standardized protein targets (EGFR, IL-7Rα, MBP, PD-L1, BBF-14, BHRF1, Cas9) for rigorous and consistent evaluations in computational protein binder design. We sourced their current best publicly available binders, described in **Supplemental section A.4**.

To explore the relevant sequence space near the known binder templates, we conducted large-scale screens based on random masking. With 100,000 iterations for each protein target, we randomly masked its best-known binder between 0 and 100% uniformly. Half of the time, we also uniformly sampled a random section of the template to use instead of the entire sequence in an attempt to hone in on domains vital to binding activity. This design enables the exploration of both global and localized binding-relevant motifs, promoting diversity while maintaining biological plausibility. After masking, the template sequence was fed to a DSM model with $s = 100$ and a temperature of 1.0. Target proteins and their newly designed binders were sent to Synteract2 through the Synthyra API to predict their binding affinity **Supplemental section A.1.1** (56). `Unconditional` generation refers to DSM$_{650}$ designing a version of the template *without* reference to the target. `Conditional` generation refers to DSM$_{650-ppi}$ designing a version of the template *while* referencing the target.

4 Results and Discussion

4.1 DSM generated biomimetic but distinct sequence distributions

We conducted a grid search over sampling temperature and s, identifying that temperature = 1.0, no gumbel softmax, and filling in one token at a time ($s = 1$) yielded the lowest JS with natural sequences. However, we observed that the JS of amino acid 3-mers was almost unchanged between $s = 1$ and $s = 5$, so we chose $s = 5$ for future experimentation, resulting in five times faster throughput. After the hyperparameter search, we generated 10,000 sequences unconditionally based on the validation set. Then, we predicted the secondary structure and AV terms for the natural and generated sequences, leaving us with three types of distributions to compare between natural and our *de novo* proteins. While the χ^2 test rejects the null hypothesis of identical distributions (likely due to large sample size), the low JS indicates high similarity.

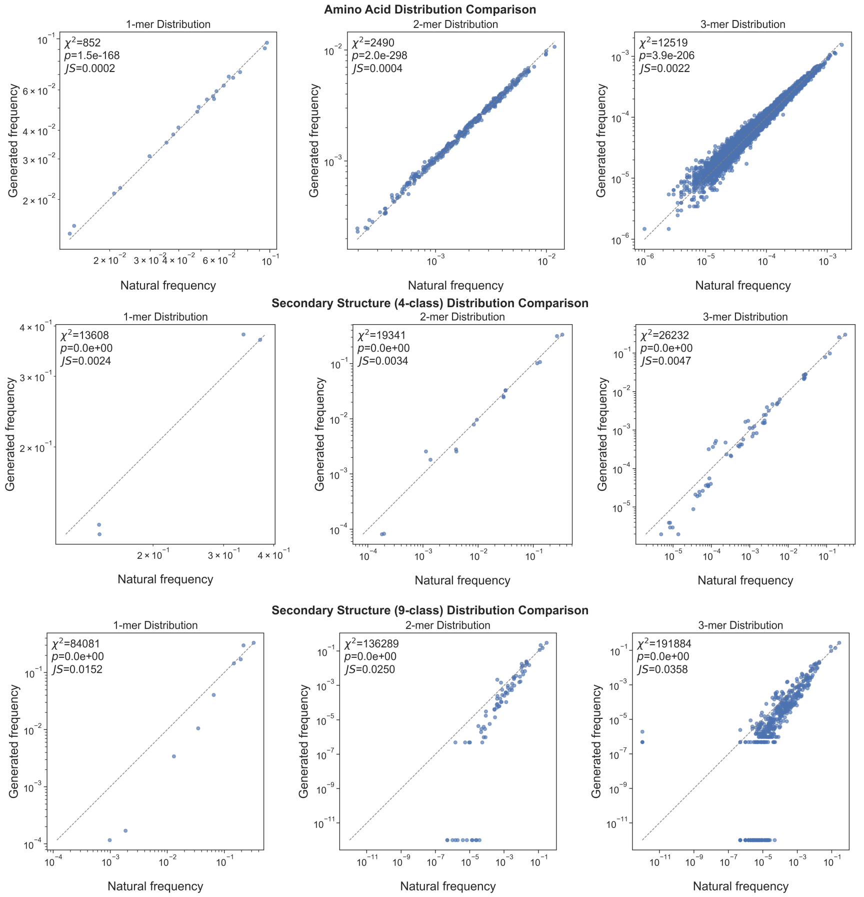

Fig. 2. Plot of natural vs. generated 1-mer, 2-mer, and 3-mer comparisons with associated χ^2 values and JS. Amino acid frequencies (top row), secondary structure [four-class] token frequencies (middle row), and secondary structure [nine-class] token frequencies (bottom row).

When comparing natural vs. generated frequencies, we expected entries to fall close to $y = x$. Indeed, we saw that the amino acid k-mers followed $y = x$ closely with very low JS values < 0.01, yet low χ^2 p-values (**Figure 2**). This demonstrates that DSM has a strong understanding of biological amino acid usage. The secondary structure comparison followed a similar trend, with low JS and low p-values (**Figure 2**). We reasoned that the strict alignment of the four-class secondary structure implied that DSM produced sequences with natural-like or convincing structural features. Lastly, the nine-class secondary structure followed similar trends (**Figure 2**). However,

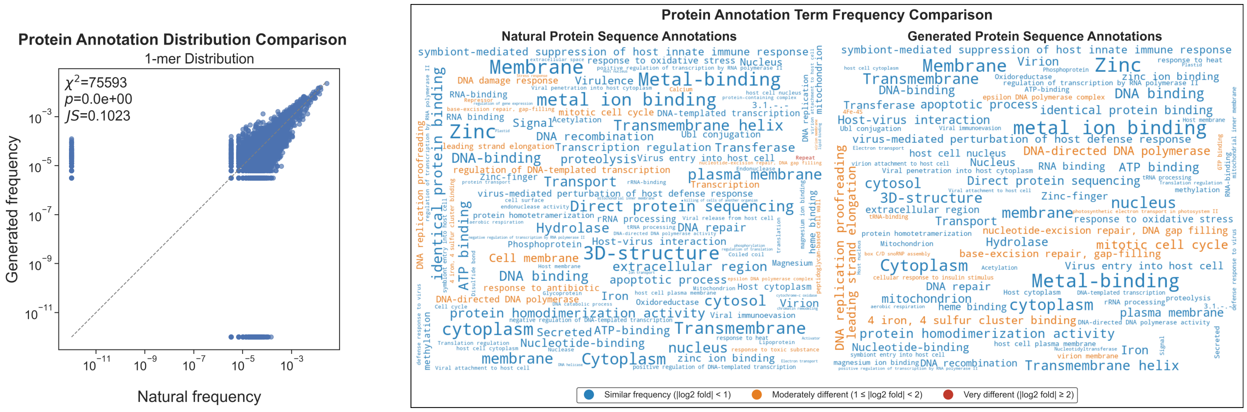

Fig. 3. (Left) Plot of natural vs. generation AV token 1-mers with χ^2 values and JS. (Right) Word cloud of natural and generated sequence annotations. Higher frequency terms have a bigger font, and terms are colored based on similarity between natural and generated frequencies via $\log_2$ fold change.

there is some clear nuance in the distribution of unconditional DSM outputs. Whereas a few secondary structure k-mers are overrepresented, the vast majority are underused. We viewed the nine-class 2-mers and 3-mers as fairly niche structural regions that may be linked to a highly specific biochemical function. With this in mind, these data imply that DSM produces 'generic' possible sequences, at least without steering towards a specific type of protein.

When the AV terms were compared, we saw that natural and generated sequences were described by many unique AV terms, with the rest much closer to $y = x$. This resulted in a very low p-value but higher JS value of 0.102 (**Figure 3A**). Given that the AV vocabulary exceeds 88,000 terms, with some overlapping usage, we hypothesized that increased sequence sampling would further reduce JS. However, such increases in sample size would likely exacerbate χ^2 test sensitivity, making statistical significance less informative. Closer inspection of the predicted functions shared between the natural and generated proteins (**Figure 3B**) demonstrated high overlap among common terms, combined with low p-values and JS across amino acid, secondary structure, and AV term k-mers. This suggests that DSM generates biomimetic sequences that occupy a distinct yet biologically plausible distribution when unconditional generation is used. This balance, capturing the latent structure of biochemical systems, may be ideal for fine-tuning towards specific design tasks.

The secondary structure model screening (**Supplemental Figure S6**) and final production model performance (**Supplemental Figure S7**) along with protein annotation performance (**Supplemental Section A.1.2**) can be explored in the Supplemental Material.

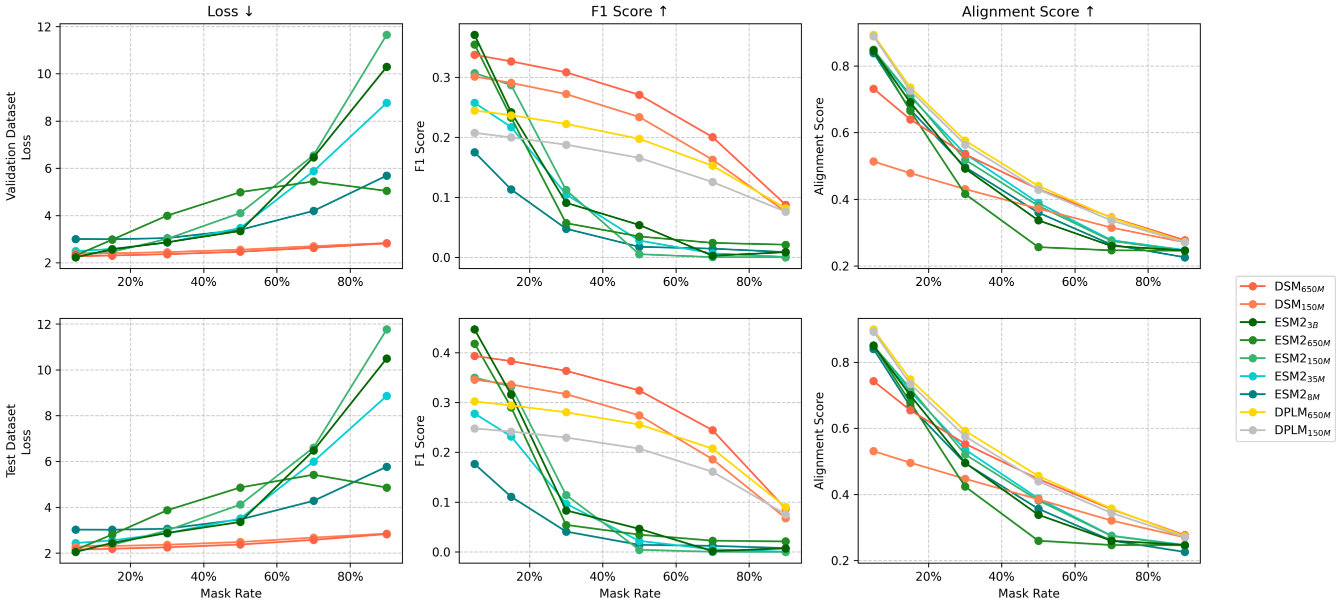

Fig. 4. Loss, weighted F1, and ASc for DSM, DPLM, and ESM2 models across a range of mask percentages for the validation (top row) and test sets (bottom row).

4.2 DSM demonstrated substantially improved sequence reconstruction compared to MLM counterparts

In line with more direct comparisons to existing MLM-based pLMs, we examined sequence reconstruction with various mask rates, filling in masked sequences with a single forward pass. We evaluated DSM against MLM-based ESM2 models, ranging from eight million to three billion parameters, as well as DPLM models, which employ a discrete diffusion process. Of note, the DPLM `generate` function does not output raw logits, so they were omitted from the loss analysis.

We observed that DSM models exhibited worse performance compared to ESM2 models at low mask percentages (5-15%); however, across higher mask rates, they maintained a low loss and high reconstruction metrics compared to MLM-only training (**Figure 4**). While the DPLM models had similar performance to DSM models, even matching alignment scores at high mask rates, DSM models showed a considerable gain in the F1 scores, ranging from 2.4% to 37.8% higher based on the mask percentage.

Interestingly, the diffusion-based pLMs showcased a distinct sequence reconstruction scaling law that became clear when contrasting DSM and DPLM compared to ESM2 (**Figure 4**). Even at 90% masking, the diffusion models produced highly similar sequences to the ground truth with an ASc ~0.27; DSM slightly outperformed DPLM (0.2771 vs. 0.2721 ASc on validation, 0.2766 vs. 0.2734 ASc on test). Both models demonstrated impressive reconstruction ability, as 0.27 ASc is over four standard deviations above the mean of randomly paired natural protein sequences. Importantly, these sequence reconstruction metrics show that DSM could generate realistic protein sequences from a *single* forward pass. This efficiency positions DSM as a compelling alternative to AR or discrete diffusion models, especially in large-scale or real-time design settings.

4.3 DSM produced high-quality embeddings in addition to generative capabilities

Another important use case for pLMs is protein annotation based on direct supervised learning or vector search. We chose to evaluate the propensity for annotation by probing frozen versions of the pLMs and assessing the intrinsic correlation between the embeddings and valuable downstream tasks (datasets described in **Supplemental Figure S3**). DSM_{650} produced the highest quality embeddings among similarly sized pLMs, generating consistently high F1 scores across a wide variety of tasks, only overtaken by the much larger ProtT5 on average (**Figure 5**). Importantly, we saw a boost in performance over its base weights of $\mathrm{ESM2}_{650}$, although all the model performances were much better than random and similar (**Supplemental Figure S4**). We observed a striking difference between the autoregressive pLM tested, ProtCLM, and the diffusion-based pLMs, DPLM and DSM, suggesting that diffusion offers a promising approach to unify generative capabilities and representation quality.

4.4 DSM can be fine-tuned to design protein binders

We confirmed that DSM models could successfully use interacting protein data to reconstruct sequences. Whereas $\mathrm{DSM}_{150-ppi-control}$ exhibited a small increase in reconstruction accuracy compared to DSM_{150} (31.4% $\rightarrow$ 31.9%), $\mathrm{DSM}_{150-ppi}$ had 33.0% accuracy. We saw similar trends in all the reported metrics, and much lower losses between ESM2 and DSM base models relative to DSM_{ppi} versions. In particular, $\mathrm{DSM}_{650-ppi}$ achieved a cross-entropy loss of 1.989, the lowest among all tested models (full metrics in **Supplemental Table S3**).

To apply these generative abilities, we generated 1.4 million potential interactors, 100,000 for each BenchBB target and template-guided scheme, with the goal of producing proteins with high binding affinities. Interestingly, predicted pKd (ppKd) values remained stable across mask rates, suggesting DSM's robustness to input corruption in template-based binder design. Top binder designs by ppKd for each target and method can be found in **Supplemental Figure S9**; the success rates and average

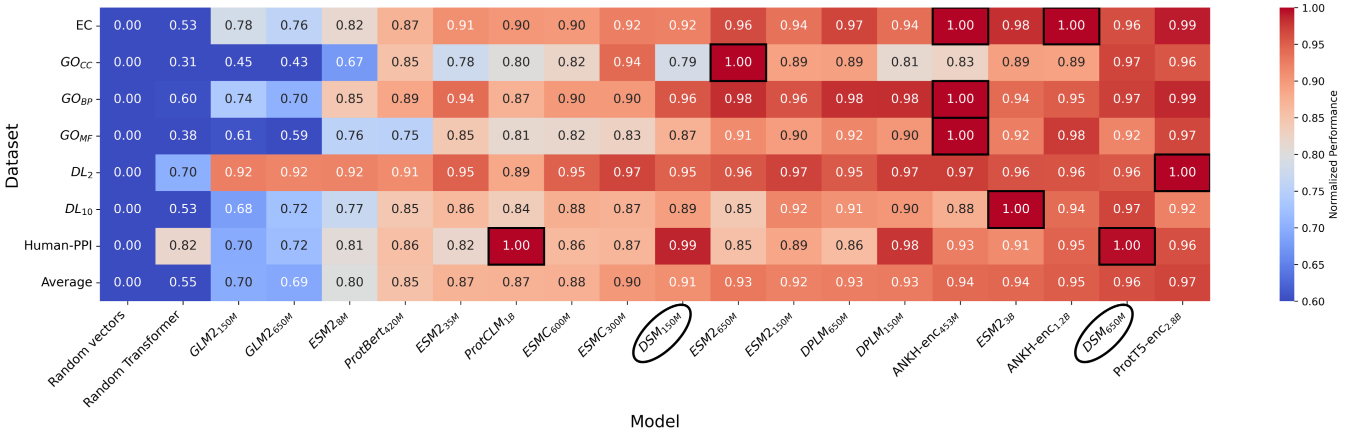

Fig. 5. Protein representation probe scores, reported and colored by their relative F1 score increase over the random vector control.

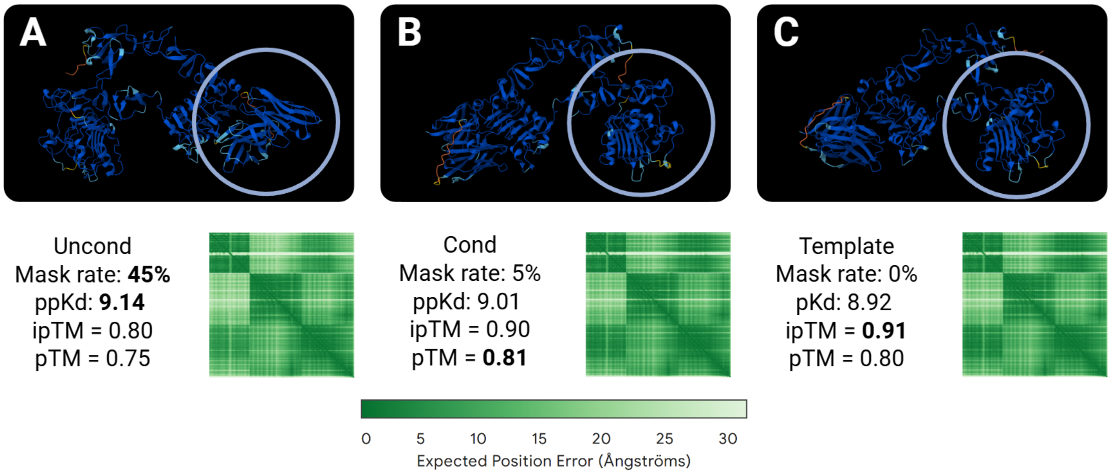

Fig. 6. Designed EGFR binders folded alongside EGFR using AlphaFold3, associated plots from web server rendered (57). (A) Unconditional generation at a high mask rate with a very high ppKd and compliance with AlphaFold3. (B) Conditional generation with the highest recorded AlphaFold3 pTM of folded samples from either method. (C) Adapytv competition winner with the highest ipTM of folded designs (34)

.

ppKd for each target-method combination can be found in **Supplemental Table S2**. Some compelling EGFR binders (**Figure 6**) showcase two notable designs folded alongside EGFR using AlphaFold3 (57). One design (**Figure 6A**) exhibits higher predicted binding affinity (ppKd) than the best known binder (**Figure 6C**) from the EGFR competition while maintaining high ipTM and pTM scores, despite having 45% of the original template hidden from DSM. The protein in **Figure 6B** highlights the only folded design that exhibits a higher pTM compared to the original template, while also having a higher ppKd.

5 Conclusion

We introduced Diffusion Sequence Modeling (DSM), a simple yet powerful way to retrofit any MLM-based pLM with a diffusion objective. With only minor modifications to the masking scheme, loss calculation, and custom logit-cap head, DSM turns transformer encoders into protein designers. DSM achieves strong sequence reconstruction performance on unseen data across high corruption levels and generates biomimetic proteins that capture the statistical and functional properties of natural sequences. Additionally, on diverse representation learning benchmarks, DSM outperforms state-of-the-art MLM, autoregressive, and discrete diffusion pLMs. Fine-tuning DSM for conditional binder generation further enables template-guided design of interacting proteins, showcasing strong results on the new BenchBB benchmark.

While our study relied on *in-silico* proxies for secondary structure, function, and binding affinity, the framework is agnostic to the supervision source; integrating wet-lab feedback or structure-aware losses from crystal structures is a natural next step. We also envision DSM variants for mutagenesis studies, multimodal conditioning (including structure, small molecules, text, and AV tokens), and active learning loops to iteratively refine the model (3, 58, 59). Other biological sequence modalities, such as codons, could further refine organism-specific protein production optimization through codon usage refinement (20, 60). We also believe architectural changes, for example, mixture-of-experts systems with modality-based routing, could also improve the embedding and generation quality (61).

It is important to note that pLMs have the potential to revolutionize numerous economic sectors, including healthcare and environmental sciences, while also carrying associated risks. In our work, we proactively mitigate unintended risks by excluding disease-related annotation terms and actively avoid training models on proxies for virulence, with the intention of not directly enabling types of bioterrorism. This approach aligns with broader concerns in the scientific community about the dual-use nature of AI in biology, where models could potentially lower barriers for malicious actors designing harmful biological entities (62).

In summary, DSM closes the gap between *understanding* and *creating* proteins in a single architecture, enabled by a few Python line changes, providing a foundation for rapid, scalable, and biologically grounded protein design.

DATA AND CODE AVAILABILITY

Selected datasets, code, and model weights can be found at github.com/Gleghorn-Lab/DSM.

COMPETING FINANCIAL INTERESTS

LH, DB, and JPG are co-founders of and have an equity stake in Synthyra, PBLLC.

AUTHOR CONTRIBUTIONS

Conceptualization (LH, JPG), Synthyra API (LH, DB), Model architectures (LH), Data Curation (LH), Investigation (LH, NR), Formal Analysis (LH, NR, JPG), Writing – Original Draft (LH, JPG), Writing – Review & Editing (LH, NR, DB, JPG), Supervision (JPG), Project Administration (JPG), Funding acquisition (LH, JPG).

ACKNOWLEDGEMENTS

The authors thank Katherine M. Nelson, Ph.D., for reviewing and commenting on drafts of the manuscript. This work was partly supported by the University of Delaware Graduate College through the Unidel Distinguished Graduate Scholar Award (LH), the National Science Foundation through NAIRR pilot 240064 (JPG), and the National Institutes of Health through NIGMS T32GM142603 (NR), R01HL133163 (JPG) and R01HL145147 (JPG).

6 References

1. Xukang Shen, Siliang Song, Chuan Li, and Jianzhi Zhang. Synonymous mutations in representative yeast genes are mostly strongly non-neutral. *Nature*, 606(7915): 725–731, June 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04823-w.
2. Laurence Loewe and William G. Hill. The population genetics of mutations: good, bad and indifferent. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1544):1153–1167, April 2010. ISSN 0962-8436. doi: 10.1098/rstb.2009.0317.
3. Logan Hallee, Niko Rafailidis, Colin Horger, David Hong, and Jason P. Gleghorn. Annotation vocabulary (might be) all you need. *bioRxiv*, page 2024.07.30.605924, July 2024. doi: 10.1101/2024.07.30.605924.
4. The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, January 2023. ISSN 0305-1048. doi: 10.1093/nar/gkac1052.
5. Craig E Jones, Alfred L Brown, and Ute Baumann. Estimating the annotation error rate of curated GO database sequence annotations. *BMC Bioinformatics*, 8:170, 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-170.
6. Sabrina de Azevedo Silveira, Raquel Cardoso de Melo-Minardi, Carlos Henrique da Silveira, Marcelo Matos Santoro, and Wagner Meira Jr. Enzymap: Exploiting protein annotation for modeling and predicting ec number changes in uniprot/swiss-prot. *PLOS ONE*, 9(2):e89162, February 2014. ISSN 1932-6203. doi: 10.1371/journal.pone.0089162.
7. Braun Markus, Gruber Christian C, Krassnigg Andreas, Kummer Arkadij, Lutz Stefan, Oberdorfer Gustav, Siirola Elina, and Snajdrova Radka. Accelerating biocatalysis discovery with machine learning: A paradigm shift in enzyme engineering, discovery, and design. *ACS Catalysis*, 13(21):14454–14469, November 2023. doi: 10.1021/acscatal.3c03417.
8. Catherine S. Millar-Haskell, Allyson M. Dang, and Jason P. Gleghorn. Coupling synthetic biology and programmable materials to construct complex tissue ecosystems. *MRS communications*, 9(2):421–432, June 2019. ISSN 2159-6859. doi: 10.1557/mrc.2019.69.
9. Vonetta L. Edwards, Elias McComb, Jason P. Gleghorn, Larry Forney, Patrik M. Bavoil, and Jacques Ravel. Three-dimensional models of the cervicovaginal epithelia to study host-microbiome interactions and sexually transmitted infections. *Pathogens and Disease*, 80(1):ftac026, August 2022. ISSN 2049-632X. doi: 10.1093/femspd/ftac026.
10. John P. DeLong, Maitham A. Al-Sammak, Zeina T. Al-Ameeli, David D. Dunigan, Kyle F. Edwards, Jeffry J. Fuhrmann, Jason P. Gleghorn, Hanqun Li, Kona Haramoto, Amelia O. Harrison, Marcia F. Marston, Ryan M. Moore, Shawn W. Polson, Barbra D. Ferrell, Miranda E. Salsbery, Christopher R. Schvarcz, Jasmine Shirazi, Grieg F. Steward, James L. Van Etten, and K. Eric Wommack. Towards an integrative view of virus phenotypes. *Nature Reviews Microbiology*, 20(2):83–94, February 2022. ISSN 1740-1534. doi: 10.1038/s41579-021-00612-w.
11. Katherine M. Nelson, N'Dea Irvin-Choy, Matthew K. Hoffman, Jason P. Gleghorn, and Emily S. Day. Diseases and conditions that impact maternal and fetal health and the potential for nanomedicine therapies. *Advanced Drug Delivery Reviews*, 170:425–438, March 2021. ISSN 1872-8294. doi: 10.1016/j.addr.2020.09.013.
12. Rachel M. Gilbert and Jason P. Gleghorn. Connecting clinical, environmental, and genetic factors point to an essential role for vitamin a signaling in the pathogenesis of congenital diaphragmatic hernia. *American Journal of Physiology. Lung Cellular and Molecular Physiology*, 324(4):L456–L467, April 2023. ISSN 1522-1504. doi: 10.1152/ajplung.00349.2022.
13. Yuanjun Shen and Jason P. Gleghorn. Class iii phosphatidylinositol-3 kinase/vacuolar protein sorting 34 in cardiovascular health and disease. *Journal of Cardiovascular Translational Research*, 18(2):392–407, April 2025. ISSN 1937-5395. doi: 10.1007/s12265-024-10581-z.
14. Richard J. Roberts, Logan Hallee, and Chi Keung Lam. The potential of hsp90 in targeting pathological pathways in cardiac diseases. *J Pers Med*, 11(12):1373, 2021. ISSN 2075-4426. doi: 10.3390/jpm11121373.
15. Sujoita Sen, Logan Hallee, and Chi Keung Lam. The potential of gamma secretase as a therapeutic target for cardiac diseases. *J Pers Med*, 11(12):1294, 2021. doi: 10.3390/jpm11121294. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.
16. Enrique Herrero Acero, Doris Ribitsch, Anita Dellacher, Sabine Zitzenbacher, Annemarie Marold, Georg Steinkellner, Karl Gruber, Helmut Schwab, and Georg M. Guebitz. Surface engineering of a cutinase from thermobifida cellulosilytica for improved polyester hydrolysis. *Biotechnology and Bioengineering*, 110(10):2581–2590, 2013. ISSN 1097-0290. doi: 10.1002/bit.24930.
17. Ju-Jiun Pang, Jong-Shik Shin, and Si-Yu Li. The catalytic role of rubisco for in situ co2 recycling in escherichia coli. *Frontiers in Bioengineering and Biotechnology*, 8, November 2020. ISSN 2296-4185. doi: 10.3389/fbioe.2020.543807.
18. Ali Miserez, Jing Yu, and Pezhman Mohammadi. Protein-based biological materials: Molecular design and artificial production. *Chemical Reviews*, 123(5):2049–2111, March 2023. ISSN 0009-2665. doi: 10.1021/acs.chemrev.2c00621.
19. Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Trans Pattern Anal Mach Intell*, 44 (10):7112–7127, 2022. ISSN 1939-3539. doi: 10.1109/TPAMI.2021.3095381.
20. Logan Hallee, Nikolaos Rafailidis, and Jason P. Gleghorn. cdsBERT - extending protein language models with codon awareness. *bioRxiv*, 2023. doi: 10.1101/2023.09.15.558027. Pages: 2023.09.15.558027 Section: New Results.
21. Sizhen Li, Saeed Moayedpour, Ruijiang Li, Michael Bailey, Saleh Riahi, Milad Miladi, Jacob Miner, Dinghai Zheng, Jun Wang, Akshay Balsubramani, Khang Tran, Minnie Zacharia, Monica Wu, Xiaobo Gu, Ryan Clinton, Carla Asquith, Joseph Skalesk, Lianne Boeglin, Sudha Chivukula, Anusha Dias, Fernando Ulloa Montoya, Vikram Agarwal, Ziv Bar-Joseph, and Sven Jager. CodonBERT: Large language models for mRNA design and optimization. *bioRxiv*, 2023. doi: 10.1101/2023.09.09.556981. Pages: 2023.09.09.556981 Section: New Results.
22. Zilin Ren, Lili Jiang, Yaxin Di, Dufei Zhang, Jianli Gong, Jianting Gong, Qiwei Jiang, Zhiguo Fu, Pingping Sun, Bo Zhou, and Ming Ni. CodonBERT: a BERT-based architecture tailored for codon optimization using the cross-attention mechanism. *Bioinformatics*, page btae330, 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae330.
23. Eric Nguyen, Michael Poli, Matthew G. Durrant, Armin W. Thomas, Brian Kang, Jeremy Sullivan, Madelena Y. Ng, Ashley Lewis, Aman Patel, Aaron Lou, Stefano Ermon, Stephen A. Baccus, Tina Hernandez-Boussard, Christopher Ré, Patrick D. Hsu, and Brian L. Hie. Sequence modeling and design from molecular to genome scale with evo. *bioRxiv*, 2024. doi: 10.1101/2024.02.27.582234. Publisher: Cold Spring Harbor Laboratory _eprint: https://www.biorxiv.org/content/early/2024/02/27/2024.02.27.582234.full.pdf.
24. Kangjie Zheng, Siyu Long, Tianyu Lu, Junwei Yang, Xinyu Dai, Ming Zhang, Zaiqing Nie, Wei-Ying Ma, and Hao Zhou. Esm all-atom: Multi-scale protein language model for unified molecular modeling. *arXiv*, June 2024. doi: 10.48550/arXiv.2403.12995. arXiv:2403.12995 [q-bio].
25. Garyk Brixi, Matthew G. Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A. Gonzalez, Samuel H. King, David B. Li, Aditi T. Merchant, Mohsen Naghipourfar, Eric Nguyen, Chiara Ricci-Tam, David W. Romero, Gwanggyu Sun, Ali Taghibakshi, Anton Vorontsov, Brandon Yang, Myra Deng, Liv Gorton, Nam Nguyen, Nicholas K. Wang, Etowah Adams, Stephen A. Baccus, Steven Dillmann, Stefano Ermon, Daniel Guo, Rajesh Ilango, Ken Janik, Amy X. Lu, Reshma Mehta, Mohammad R. K. Mofrad, Madelena Y. Ng, Jaspreet Pannu, Christopher Ré, Jonathan C. Schmok, John St John, Jeremy Sullivan, Kevin Zhu, Greg Zynda, Daniel Balsam, Patrick Collison, Anthony B. Costa, Tina Hernandez-Boussard, Eric Ho, Ming-Yu Liu, Thomas McGrath, Kimberly Powell, Dave P. Burke, Hani Goodarzi, Patrick D. Hsu, and Brian L. Hie. Genome modeling and design across all domains of life with evo 2. *bioRxiv*, page 2025.02.18.638918, February 2025. doi: 10.1101/2025.02.18.638918.
26. Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 118(15):e2016239118, April 2021. ISSN 0027-8424. doi: 10.1073/pnas.2016239118.
27. Francesca-Zhoufan Li, Ava P. Amini, Yisong Yue, Kevin K. Yang, and Alex X. Lu. Feature reuse and scaling: Understanding transfer learning with protein language models. *bioRxiv*, 2024. doi: 10.1101/2024.02.05.578959.
28. Ahmed Elnaggar, Hazem Essam, Wafaa Salah-Eldin, Walid Moustafa, Mohamed Elkerdawy, Charlotte Rochereau, and Burkhard Rost. Ankh: Optimized protein language model unlocks general-purpose modelling. *arXiv*, January 2023. doi: 10.48550/arXiv.2301.06568. arXiv:2301.06568 [cs].
29. Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion language models are versatile protein learners. *arXiv*, October 2024. doi: 10.48550/arXiv.2402.18567. arXiv:2402.18567 [cs].
30. Quentin Fournier, Robert M. Vernon, Almer van der Sloot, Benjamin Schulz, Sarath Chandar, and Christopher James Langmead. Protein language models: Is scaling necessary? *bioRxiv*, page 2024.09.23.614603, September 2024. doi: 10.1101/2024.09.23.614603.

31. Krithik Ramesh, Sameed M. Siddiqui, Albert Gu, Michael D. Mitzenmacher, and Pardis C. Sabeti. Lyra: An efficient and expressive subquadratic architecture for modeling biological sequences. *arXiv*, March 2025. doi: 10.48550/arXiv.2503.16351. arXiv:2503.16351 [cs].
32. Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv*, February 2025. doi: 10.48550/arXiv.2502.09992. arXiv:2502.09992 [cs].
33. 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, March 2023. doi: 10.1126/science.ade2574.
34. Tudor-Stefan Cotet, Igor Krawczuk, Filippo Stocco, Noelia Ferruz, Anthony Gitter, Yoichi Kurumida, Lucas de Almeida Machado, Francesco Paesani, Cianna N. Calia, Chance A. Challacombe, Nikhil Haas, Ahmad Qamar, Bruno E. Correia, Martin Pacesa, Lennart Nickel, Kartic Subr, Leonardo V. Castorina, Maxwell J. Campbell, Constance Ferragu, Patrick Kidger, Logan Hallee, Christopher W. Wood, Michael J. Stam, Tadas Kluonis, Süleyman Mert Ünal, Elian Belot, Alexander Naka, and Adaptyv Competition Organizers. Crowdsourced protein design: Lessons from the adaptyv egfr binder competition. *bioRxiv*, page 2025.04.17.648362, April 2025. doi: 10.1101/2025.04.17.648362.
35. R. A. Fisher and Edward John Russell. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368, 1922. doi: 10.1098/rsta.1922.0009.
36. Andre Cornman, Jacob West-Roberts, Antonio Pedro Camargo, Simon Roux, Martin Beracochea, Milot Mirdita, Sergey Ovchinnikov, and Yunha Hwang. The omg dataset: An open metagenomic corpus for mixed-modality genomic language modeling. *bioRxiv*, August 2024. doi: 10.1101/2024.08.14.607850. bioRxiv 2024.08.14.607850.
37. Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju-yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. Gemma 2: Improving open language models at a practical size. *arXiv*, October 2024. doi: 10.48550/arXiv.2408.00118. arXiv:2408.00118 [cs].
38. Irwan Bello, Hieu Pham, Quoc V. Le, Mohammad Norouzi, and Samy Bengio. Neural combinatorial optimization with reinforcement learning. *arXiv*, January 2017. doi: 10.48550/arXiv.1611.09940. arXiv:1611.09940 [cs].
39. Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv*, March 2023. doi: 10.48550/arXiv.2111.09543. arXiv:2111.09543 [cs].
40. Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv*, December 2024. doi: 10.48550/arXiv.2412.13663. arXiv:2412.13663 [cs].
41. Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv*, May 2017. doi: 10.48550/arXiv.1608.03983. arXiv:1608.03983 [cs].
42. Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv*, January 2019. doi: 10.48550/arXiv.1711.05101. arXiv:1711.05101 [cs].
43. Ziqian Zhong and Jacob Andreas. Algorithmic capabilities of random transformers. *arXiv*, October 2024. doi: 10.48550/arXiv.2410.04368. arXiv:2410.04368v1.
44. Xingyi Cheng, Bo Chen, Pan Li, Jing Gong, Jie Tang, and Le Song. Training compute-optimal protein language models. *bioRxiv*, 2024. doi: 10.1101/2024.06.06.597716. Pages: 2024.06.06.597716 Section: New Results.
45. ESM Team. Esm cambrian: Revealing the mysteries of proteins with unsupervised learning. *Evolutionary Scale Blog*, December 2024. Evolutionary Scale, blog post.
46. Logan Hallee, David Bichara, and Jason P. Gleghorn. Esm++: Efficient and hugging face compatible versions of the esm cambrian models. *Hugging Face*, 2024. doi: 10.57967/hf/3726.
47. Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion language models are versatile protein learners. *arXiv*, October 2024. doi: 10.48550/arXiv.2402.18567.
48. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv*, 2017. doi: 10.48550/arXiv.1706.03762. Number: arXiv:1706.03762.
49. Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv*, November 2023. arXiv:2104.09864 [cs].
50. Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv*, October 2021. doi: 10.48550/arXiv.2106.09685. arXiv:2106.09685 [cs].
51. Karl Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, July 1900.
52. Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, January 1991.
53. Christopher Manning, Prabhakar Raghavan, and Hinrich Schuetze. Introduction to information retrieval. *Cambridge UP*, 2009.
54. Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, Farrokh Mehryary, Radja Hachilif, Annika L. Gable, Tao Fang, Nadezhda T. Doncheva, Sampo Pyysalo, Peer Bork, Lars J. Jensen, and Christian von Mering. The string database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51(D1):D638–D646, January 2023. ISSN 1362-4962. doi: 10.1093/nar/gkac1000.
55. Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics (Oxford, England)*, 22 (13):1658–1659, July 2006. ISSN 1367-4803. doi: 10.1093/bioinformatics/btl158.
56. Logan Hallee and Jason P. Gleghorn. Protein-protein interaction prediction is achievable with large language models. *bioRxiv*, page 2023.06.07.544109, June 2023. doi: 10.1101/2023.06.07.544109.
57. 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, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w.
58. Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. SaProt: Protein language modeling with structure-aware vocabulary. *bioRxiv*, 2023. doi: 10.1101/2023.10.01.560349. Pages: 2023.10.01.560349 Section: New Results.
59. Xiaohan Lin, Zhenyu Chen, Yanheng Li, Zicheng Ma, Chuanliu Fan, Ziqiang Cao, Shihao Feng, Yi Qin Gao, and Jun Zhang. Protokens: Probabilistic vocabulary for compact and informative encodings of all-atom protein structures. *bioRxiv*, page 2023.11.27.568722, July 2024. doi: 10.1101/2023.11.27.568722.
60. Logan Hallee and Bohdan B. Khomtchouk. Machine learning classifiers predict key genomic and evolutionary traits across the kingdoms of life. *Scientific Reports*, 13(1): 2088, 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-28965-7.
61. Logan Hallee, Rohan Kapur, Arjun Patel, Jason P. Gleghorn, and Bohdan B. Khomtchouk. Contrastive learning and mixture of experts enables precise vector embeddings in biological databases. *Scientific Reports*, 15(1):14953, April 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-98185-8.
62. Sarah R. Carter, Samuel Curtis, Claudia Emerson, Jeffrey Gray, Ian C. Haydon, Andrew Hebbeler, Claire Qureshi, Nicholas Randolph, Alexander Rives, and Lynda Stuart. Community values, guiding principles, and commitments for the responsible development of ai for protein design, March 2024.

63. Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021. doi: 10.21105/joss.03021.
64. Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv*, February 2019. doi: 10.48550/arXiv.1803.08375. arXiv:1803.08375 [cs].
65. Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv*, July 2012. doi: 10.48550/arXiv.1207.0580. arXiv:1207.0580 [cs].
66. J. A. Cuff and G. J. Barton. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins*, 34(4):508–519, March 1999. ISSN 0887-3585. doi: 10.1002/(sici)1097-0134(19990301)34:4<508::aid-prot10>3.0.co;2-4.
67. Yuedong Yang, Jianzhao Gao, Jihua Wang, Rhys Heffernan, Jack Hanson, Kuldip Paliwal, and Yaoqi Zhou. Sixty-five years of the long march in protein secondary structure prediction: the final stretch? *Briefings in Bioinformatics*, 19(3):482–494, May 2018. ISSN 1477-4054. doi: 10.1093/bib/bbw129.
68. Martin Pacesa, Lennart Nickel, Joseph Schmidt, Ekaterina Pyatova, Christian Schellhaas, Lucas Kissling, Ana Alcaraz-Serna, Yehlin Cho, Kourosh H. Ghamary, Laura Vinué, Brahm J. Yachnin, Andrew M. Wollacott, Stephen Buckley, Sandrine Georgeon, Casper A. Goverde, Georgios N. Hatzopoulos, Pierre Gönczy, Yannick D. Muller, Gerald Schwank, Sergey Ovchinnikov, and Bruno E. Correia. Bindcraft: one-shot design of functional protein binders. *bioRxiv*, October 2024. doi: 10.1101/2024.09.30.615802. bioRxiv 2024.09.30.615802.
69. Vinicius Zambaldi, David La, Alexander E. Chu, Harshnira Patani, Amy E. Danson, Tristan O. C. Kwan, Thomas Frerix, Rosalia G. Schneider, David Saxton, Ashok Thillaisundaram, Zachary Wu, Isabel Moraes, Oskar Lange, Eliseo Papa, Gabriella Stanton, Victor Martin, Sukhdeep Singh, Lai H. Wong, Russ Bates, Simon A. Kohl, Josh Abramson, Andrew W. Senior, Yilmaz Alguel, Mary Y. Wu, Irene M. Aspalter, Katie Bentley, David L.V. Bauer, Peter Cherepanov, Demis Hassabis, Pushmeet Kohli, Rob Fergus, and Jue Wang. De novo design of high-affinity protein binders with alphaproteo. *arXiv*, September 2024. doi: 10.48550/arXiv.2409.08022. arXiv:2409.08022v1.
70. Iwan Zimmermann, Pascal Egloff, Cedric A.J. Hutter, Fabian M. Arnold, Peter Stohler, Nicolas Bocquet, Melanie N. Hug, Sylwia Huber, Martin Siegrist, Lisa Hetemann, Jennifer Gera, Samira Gmür, Peter Spies, Daniel Gygax, Eric R. Geertsma, Roger J.P. Dawson, and Markus A. Seeger. Synthetic single domain antibodies for the conformational trapping of membrane proteins. *eLife*, page e34317, May 2018. doi: 10.7554/eLife.34317.
71. C. D. Suraweera, M. G. Hinds, and M. Crystal Kvansakul. Crystal structures of epstein–barr virus bcl-2 homolog bhrf1 bound to bid and puma bh3 motif peptides. *Viruses*, 14(2222), 2022. ISSN 10. doi: 10.3390/v14102222.
72. Anthony L. Desbien, John W. Kappler, and Philippa Marrack. The epstein–barr virus bcl-2 homolog, bhrf1, blocks apoptosis by binding to a limited amount of bim. *Proceedings of the National Academy of Sciences*, 106:5663–5668, 2009. ISSN 14. doi: 10.1073/pnas.0901036106.
73. Jinlong Zhou, Yue Xiao, Quian Tang, Yunjun Yan, Dongqi Liu, and Houjin Zhang. De novo design protein binders for mbp and gst tags. *Biochemical and Biophysical Research Communications*, 748:151322, February 2025. ISSN 0006-291X. doi: 10.1016/j.bbrc.2025.151322. Publisher: Elsevier.
74. Wei Yang, Derrick R. Hicks, Agnidipta Ghosh, Tristin A. Schwartze, Brian Conventry, Inna Goreshnik, Aza Allen, Samer F. Halabiya, Chan Johng Kim, Cynthia S. Hinck, David S. Lee, Asim K. Bera, Zhe Li, Yujia Wang, Thomas Schlichthaerle, Longxing Cao, Buwei Huang, Sarah Garrett, Stacey R. Gerben, Stephen Rettie, Piper Heine, Analisa Murray, Natasha Edman, Lauren Carter, Lance Stewart, Steven C. Almo, Andrew P. Hinck, and David Baker. Design of high-affinity binders to immune modulating receptors for cancer immunotherapy. *Nature Communications*, February 2025. doi: 10.1038/s41467-025-57192-z.
75. E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–845, September 1988. ISSN 0006-341X.
76. Xu Sun and Weichao Xu. Fast implementation of delong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, 21(11):1389–1393, November 2014. ISSN 1558-2361. doi: 10.1109/LSP.2014.2337313.

A Supplementary Material

A.1 Synthyra models

Synthyra is a Public Benefit Company (PBLLC) that offers protein annotation services with deep learning models. While the Synthyra models used throughout this study are closed-source, anyone can use free credits to query the models at https://Synthyra.com. Here, we present various implementation details and performance metrics to further inform the results of this study.

A.1.1 Synteract2

Synteract2 is the second generation in a line of LLMs built to model protein-protein interactions (PPI). Synteract2 jointly models the probability of PPI, plausible binding sites, and the binding affinity (pKd) between two input amino acid sequences. This study extensively used the binding affinity track of Synteract2 to screen potential designed binders. The binding affinity track of Synteract2 is extended from a BERT-like pLM trained with custom parameter-efficient fine-tuning methods and trained on the binding affinity dataset from the APPT project https://github.com/Bindwell/APPT - a processed version can be found here: https://huggingface.co/datasets/Synthyra/ProteinProteinAffinity. We highlight its leading performance on the Haddock and Affinityv5.5 benchmarks over previous state-of-the-art methods **Supplemental Figure S1**.

A.1.2 Translator

Translator is a transformer-like network that directly translates protein primary sequence to AV terms. It was used in this study to estimate the protein functions of natural and generated sequences. It outputs any predicted "aspect" information, including Enzyme Commission (EC) numbers, Gene Ontologies (GO) for Biological Process (BP), Molecular Function (MF), and Cellular Compartment (CC), in addition to Interpro (IP) domains, Gene3D (3D) domains, UniProt keywords, and cofactors. To support the reliability of the closed-source *Translator* model in our work, we report performance metrics on the Translator test set and a recent case study on its performance. The test set was a non-redundant split of 1,000 sequences and annotations from high-quality UniProt entries that the model was not trained on. Additionally, the case study consisted of 656 UniProt entries deposited after *Translator* was trained that have experimentally verified annotations. **Supplemental Table S1** reports an aspect-by-aspect performance for the test set and case study set using the default settings in the Synthyra API. We also examined the same performance averaged together, varying the internal parameter top-k, which allowed *Translator* to be more exploratory with better recall or more precision (**Supplemental Figure S2**). The default parameter was $k = 3$, which had the highest F1 score. We tracked the minimum confidence in prediction such that every prediction above that confidence was correct for each k.

A.2 Protify

Protify (https://github.com/Synthyra/Protify) is an open-source project that offers low-code pLM analysis. We used it extensively to benchmark the representations of DSM models compared to all pLMs evaluated, as well as to train the secondary structure predictor for our analysis of natural versus generated distributions.

Within Protify, linear neural networks were trained with the architecture:

$$y = \sigma(\mathrm{LN}(\sigma(\sigma(\mathrm{LN}(x)W_1 + b_1)W_2 + b_2))W_3 + b_3)W_4 + b_4,$$

mapping input $x \in \mathbb{R}^d$ to output $y \in \mathbb{R}^o$ where o was the number of classes in the dataset. $W_1 \in \mathbb{R}^{d\times h}$, $W_2 \in \mathbb{R}^{h\times h}$, $W_3 \in \mathbb{R}^{h\times p}$,

Table S1. Performance of each *aspect* of *Translator* outputs. (a) Test set metrics. (b) Case study metrics.

(a) Test set metrics

Aspect	Precision	Recall	F1	Accuracy
EC	0.21	0.85	0.33	0.85
BP	0.16	0.75	0.27	0.75
CC	0.25	0.77	0.38	0.77
MF	0.37	0.73	0.49	0.73
IP	0.78	0.76	0.77	0.76
3D	0.82	0.81	0.81	0.81
Keywords	0.41	0.86	0.56	0.86
Cofactor	0.38	0.81	0.52	0.81

(b) Case study metrics

Aspect	Precision	Recall	F1	Accuracy
EC	0.41	0.84	0.55	0.84
BP	0.15	0.56	0.23	0.56
CC	0.26	0.77	0.39	0.77
MF	0.39	0.70	0.50	0.70
IP	0.95	0.81	0.88	0.81
3D	0.98	0.89	0.93	0.89
Keywords	0.56	0.84	0.67	0.84
Cofactor	0.71	0.67	0.69	0.67

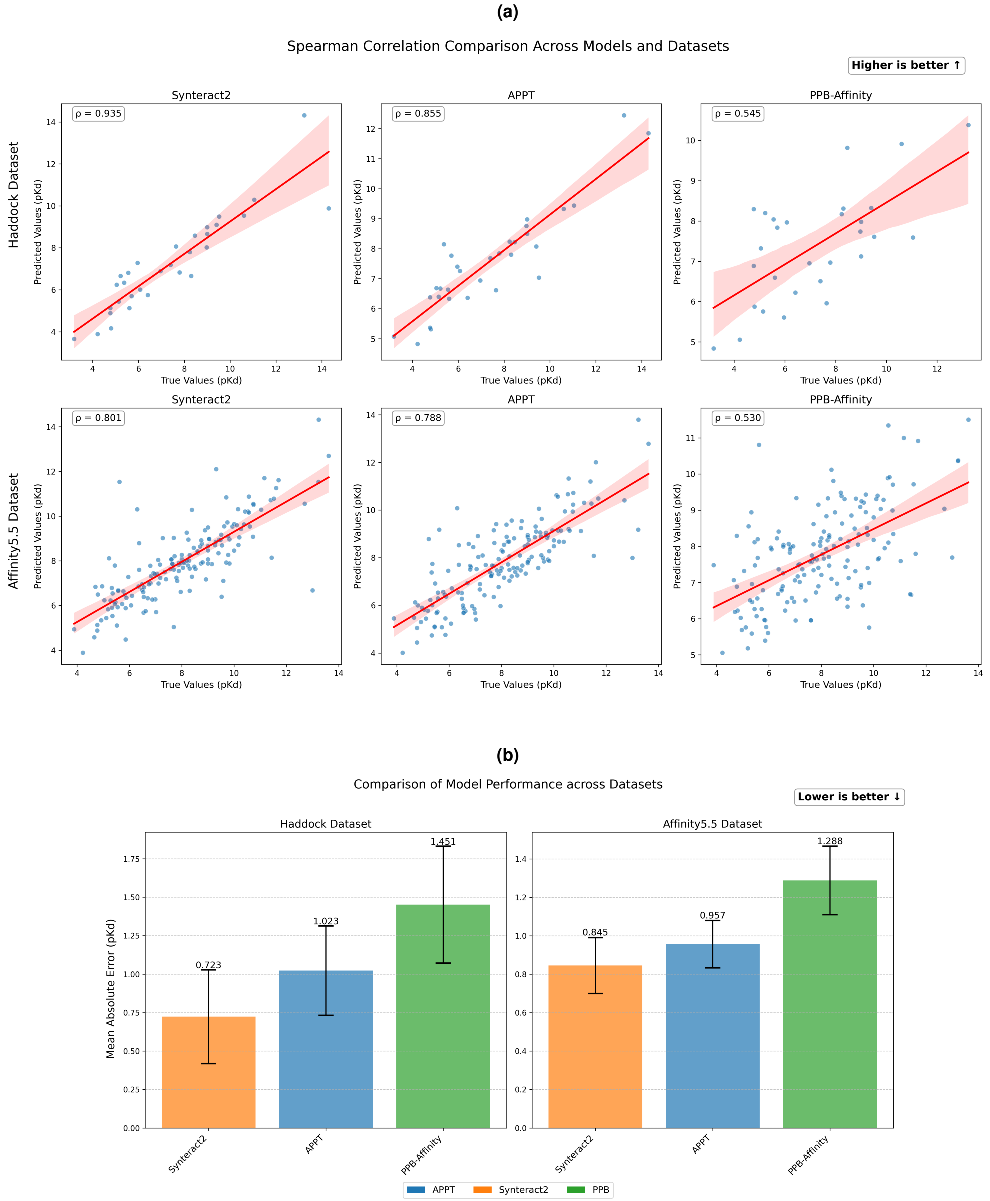

Fig. S1. Predicted binding affinity performance, comparing Synteract2, APPT, and PDB-affinity on the Haddock and Affinityv5.5 benchmarks. (a) Spearman ρ reported with line of best fit and 95% confidence intervals are calculated with `seaborn regplot`, which employs bootstrapping (63). (b) Mean absolute error with 95% confidence intervals calculated via t-test.

- EC: 13.1k train, 1.5k valid, 1.6k test - multi-label classification - 585 classes
- CC: 26k train, 3k valid, 3.4k test - multi-label classification - 320 classes
- MF: 26k train, 3k valid, 3.4k test - multi-label classification - 489 classes
- BP: 26k train, 3k valid, 3.4k test - multi-label classification - 1943 classes
- DL2: 5.5k train, 1.3k valid, 1.7k test - Binary classification
- DL10: 8,7k train, 2.2k valid, 2.8k test - Multiclass classification - 10 classes
- MB: 5k train, 662 valid, 665 test - Binary classification
- HPPI: 26.3k train, 234 valid, 180 test - Binary classification of two input sequences
- SS4: 10.8k train, 626 valid, 50 test - Token-wise classification - 4 classes
- SS9: 10.8k train, 626 valid, 50 test - Token-wise classification - 9 classes

Fig. S3. Supervised datasets used to probe model performance. EC, CC, MF, BP, DL2, DL10, MB, and HPPI were from the SaProt repository (58). SS4 and SS9 were modified from Proteinea (28).

$W_4 \in \mathbb{R}^{p \times o}$, b corresponded to the associated bias term in the linear layer, σ was ReLU (64), h was 8,192, p was calculated with

$$\left\lfloor \frac{2 \cdot o + 255}{256} \right\rfloor \cdot 256,$$

and a 20% random dropout was applied before every MatMul (except the first one) (65). Cross-entropy loss was used for multiclass and binary datasets, and binary-cross-entropy was used for multi-label datasets. Our training scheme followed 1,000 linear warmup steps from zero to a learning rate of $1e^{-4}$, a cosine learning rate scheduler (41), a batch size of 64, with evaluation on the validation set every epoch. Training occurred until a patience of 10 was exceeded for the validation loss. AdamW was used for the optimizer (42). Weighted F1 scores were reported for the test sets in the results.

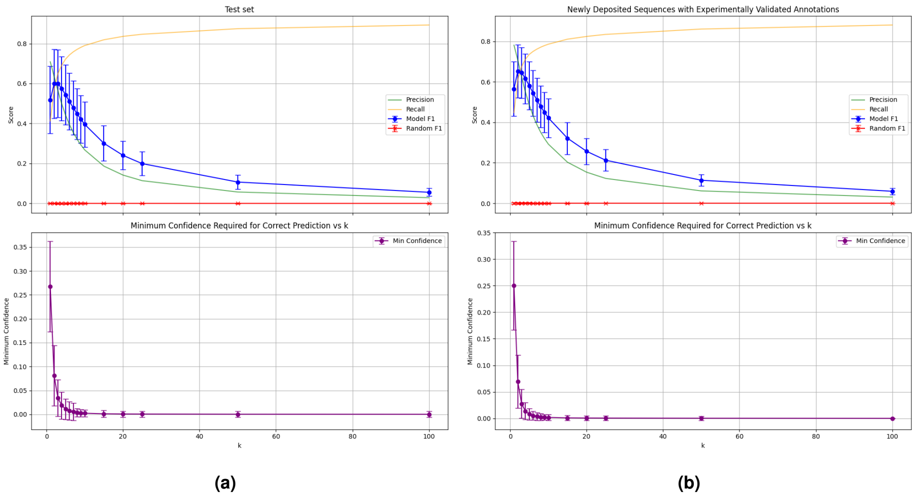

Fig. S2. Reported precision, recall, and F1 scores for *Translator* while varying k (logit-retrieval) from 1-100, as well as the minimum confidence for correctness. (a) Test set performance. (b) Case study performance.

A.3 Supervised datasets

We used the supervised datasets without modification to the splits or labels except for secondary structure (SS). For SS3 and SS8 we used the Proteinea training set for training (28), CB513 and TS115 for validation (66, 67), and CASP12, CASP13, and CASP14 for testing. Instead of the common approach to not use intrinsically disordered residue labels, we created a new label for those residues (D) (3). As such, there were four and nine classification options per residue for SS3 and SS8, respectively. We sometimes referred to them as **SS4** and **SS9**. EC, CC, MF, BP, DL2, DL10, and MB were downloaded from the SaProt repository (58).

A.4 Literature review for known binders of BenchBB

We conducted an extensive literature review to identify the current highest-affinity binders for each of the BenchBB targets. Binder sequence, binding residues, and affinity data come from the sources reported in **Supplementary Table S2**, unless otherwise specified. Sources included Adaptyv Bio's crowdsourced protein design preprint (34), AlphaProteo (69), and BindCraft (68). Although Adaptyv's materials provided Protein Data Bank (PDB) IDs for each target, exact target protein sequences were not explicitly defined. We used sequences for binding prediction using the following rationale:

- **EGFR:** We used the target sequence directly from Adaptyv's Protein Design Competition.

Table S2. Literature-derived target-binder attributes and corresponding binder sequences. Binding affinity is pKd.

Target	Target Length	Binding Residues	Binder Length	pK$_d$	Binder Source	Binder Sequence
EGFR	621	S11, N12, K13, T15, Q16, L17, G18, S356, S440, G441	241	8.92	(34)	QVQLQQSGPGLVQPSQSLSITCTVSGFSLTNYGVHWVRQSPGKGLEWLGVIWSGGNTDYNTPFTSRLSISRDTSKSQVFFKMNSLQTDDTAIYYCARALTYYDYEFAYWGQGTLVTVSAGGGGSGGGGSGGGGSDILLTQSPVILSVSPGERVSFSCRASQSIGTNIHWYQQRTNGSPKLLIRYASESISGIPSRFSGSGSGTDFTLSINSVDPEDIADYYCQQNNNWPTTFGAGTKLELK
BBF-14	122	–	147	7.68	(68)	SPIQEEIQKKVRELLEKLIEYLEELKEKAKPPFKEKLEEVIEGLERLKEEVDKVQLNMNLIVFEGLEVDEEGRVWFIVKEMLHATTEEEALENMDKFLESWEKVFKELLEYHFEHNDTSPTFDFFLDFLWWQLYGEPMPKGSHHHHH
BHRF1	159	E89, L98, G99, R100	125	8.07	(69)	MPSAFQIGLALVAAALDRALPEPYRGLALAIAAELSGLPEEELRRLVEAAEKAASADLPFEQQVGLALARIAAAVAGVGLARRAPSLPPEELLAAIREAIEEGGRIAAKALTRSGALEPVLAELP
SpCas9	1368	T360	116	6.42	(68)	SEEEKKEILYFIMEKLFDLDFNFKWPRNSPEEYTKAIEEFKAEVAKIVLETKEKFPEISPEELVELLEEAVYRVHRITHHWASYYVAREVIYELKKLKEKGWKAIEEYTESIISKV
IL7Rα	219	V58, L80, Y139	64	10.08	(69)	MTKVEEAKELVDKIMEAAKAKDLEKVNKLRTEFFELVNSLSLEEAEEVRKYADKKGEEWYKEQL
MBP	370	Y90, P91, F92, Y171, Y176, P315, M321, I329	126	7.62	(70)	SQVQLVESGGGSVQAGGSLRLSCVASGDIKYISYLGWFRQAPGKEREGVAALYTSTGRTYYADSVKGRFTVSLDNAKNTVYLQMNSLKPEDTALYYCAAAEWGSQSPLTQWFYRYWGQGTQVTVSA
PD-L1	221	I54, Y56, E58, N63, Q66, V76, R113, M115, S117, A121, Y123	120	10.38	(69)	SAEEKILANLEAMKAKALAAKTEEEKLFYAKALLAVAISYAIRGDYELARRAAELAVEVIKSLSKEEQKKVMDFLINIIKNITDPEDREKAIELAIAIAERLDEEVREEALKKIEELKKE

- **SpCas9:** We used the full canonical sequence from UniProt (Q99ZW2).
- **BBF-14:** A de-novo designed protein. We sourced the sequence from PDB (9HAG).
- **BHRF1:** The target sequence was obtained from PDB (2WH6). While the binder sequence with the lowest K_d came from (69), the binding residues reported in the table came from (71, 72). The binding residues reported in (69) are F65, T74, E77, D82, S85, R93.
- **IL7Rα:** The sequence was obtained from PDB (3DI3). (69) specified that the synthesized protein used in their work was residues 21-239. Upon comparing the PDB entry with the canonical UniProt sequence (P16871), we observed that they were identical from residue 21 to the C-terminus. The FASTA sequence reported in 3DI3 contained a cloning artifact at the N-terminus ("GSHM"), which was removed.
- **MBP:** We used the canonical UniProt sequence (P0AEX9), removing the initial 26 amino acids to remain consistent with the binding residue information reported in (73).
- **PD-L1:** We obtained the sequence from PDB (4Z18). To remain consistent with the binding residue information, retrieved from (74), we remove the first amino acid.

Full details are shown in **Supplementary Table S2**.

A.5 Raw and supporting data

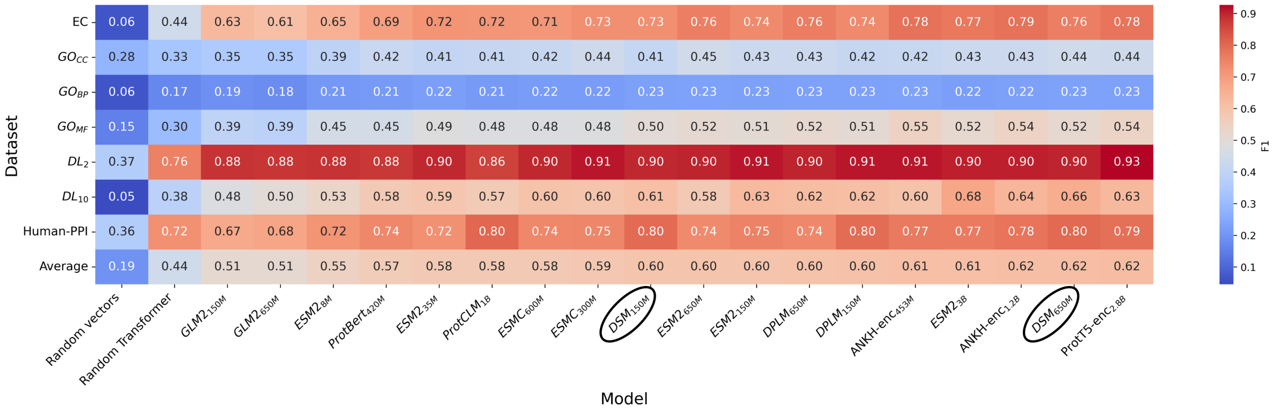

Fig. S4. Raw weighted F1 scores ($F1_{max}$ for multilabel problems) of linear probes for pLMs and datasets evaluated for representation quality.

Table S3. Reported sequence reconstruction metrics on the PPI test set. DSM_{ppi} has SeqA + SeqB inputs, while the remainder only receives SeqB. Metrics reported for a 15% mask rate.

Model	Cross entropy↓	MCC↑	F1↑	Recall↑	Precision↑	Accuracy↑
$ESM2_8$	3.047	0.073	0.112	0.118	0.223	11.8%
$ESM2_{35}$	2.498	0.179	0.221	0.233	0.286	23.3%
$ESM2_{150}$	2.202	0.276	0.318	0.328	0.366	32.2%
$ESM2_{650}$	2.407	0.260	0.307	0.307	0.378	30.7%
$ESM2_{3B}$	2.234	0.286	0.334	0.331	0.412	33.1%
DSM_{150}	2.225	0.268	0.311	0.314	0.353	31.4%
DSM_{650}	2.046	0.325	0.365	0.367	0.398	36.7%
$DSM_{150-ppi-control}$	2.199	0.274	0.315	0.319	0.358	31.9%
$DSM_{150-ppi}$	2.165	0.285	0.326	0.330	0.367	33.0%
$DSM_{650-ppi}$	**1.989**	**0.342**	**0.382**	**0.383**	**0.417**	**38.3%**

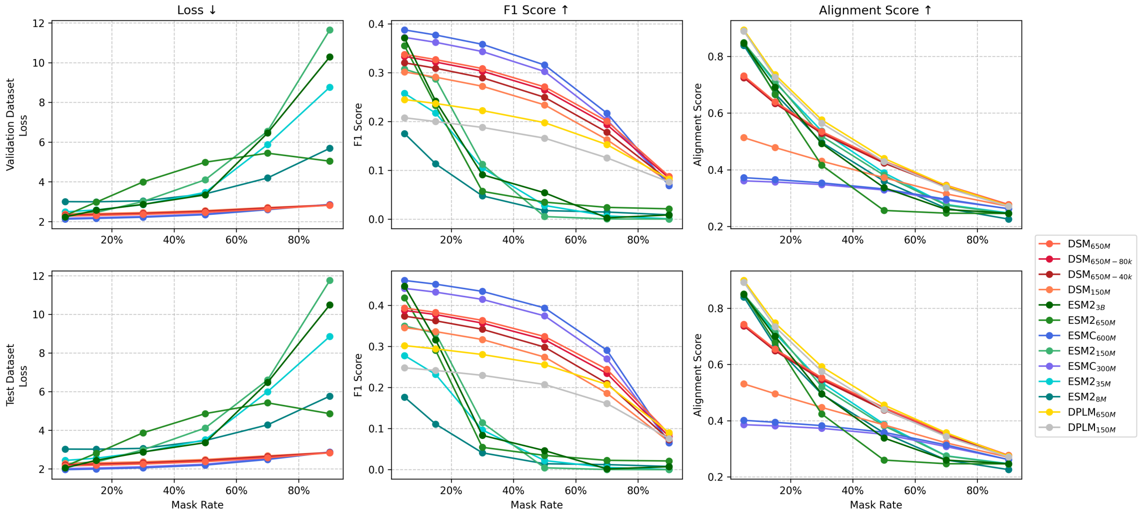

Fig. S5. Identical mask rate comparison plot to **Figure 4** except with checkpoints for DSM_{650} and ESMC models added (45, 46). The 40k and 80k step checkpoints demonstrate the gradual increase in performance of DSM over the training period. The ESMC models show exceedingly high performance across the board, although DSM still produces sequences with better ASc. However, this did not surprise us: while not much is known about ESMC training, it was pretrained on extensive meta-genomic data (45), and it is likely that it was trained on our evaluation sets. Additionally, the scaling laws of ESMC sequence reconstruction performance look much closer to the diffusion models than the MLM-based ESM2. This leads us to postulate that ESMC had a dramatically altered training scheme vs. MLM, perhaps directly a diffusion process or some higher or varied mask-rate MLM.

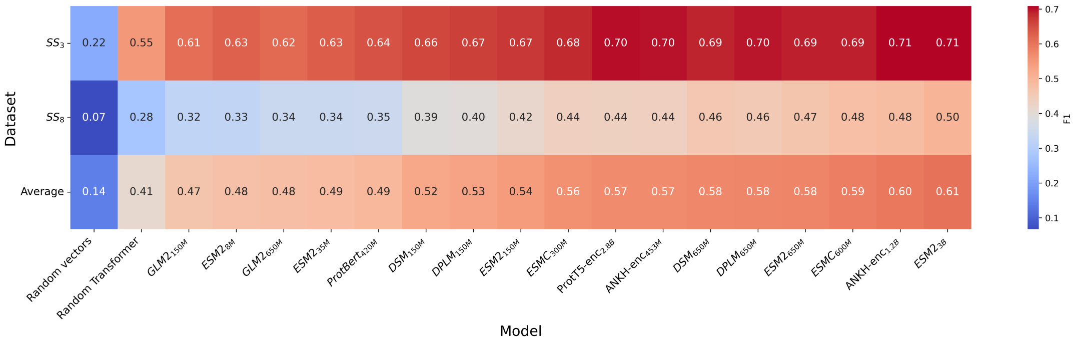

Fig. S6. Transformer probe probing of various pLMs on secondary structure datasets, weighted F1 scores reported.

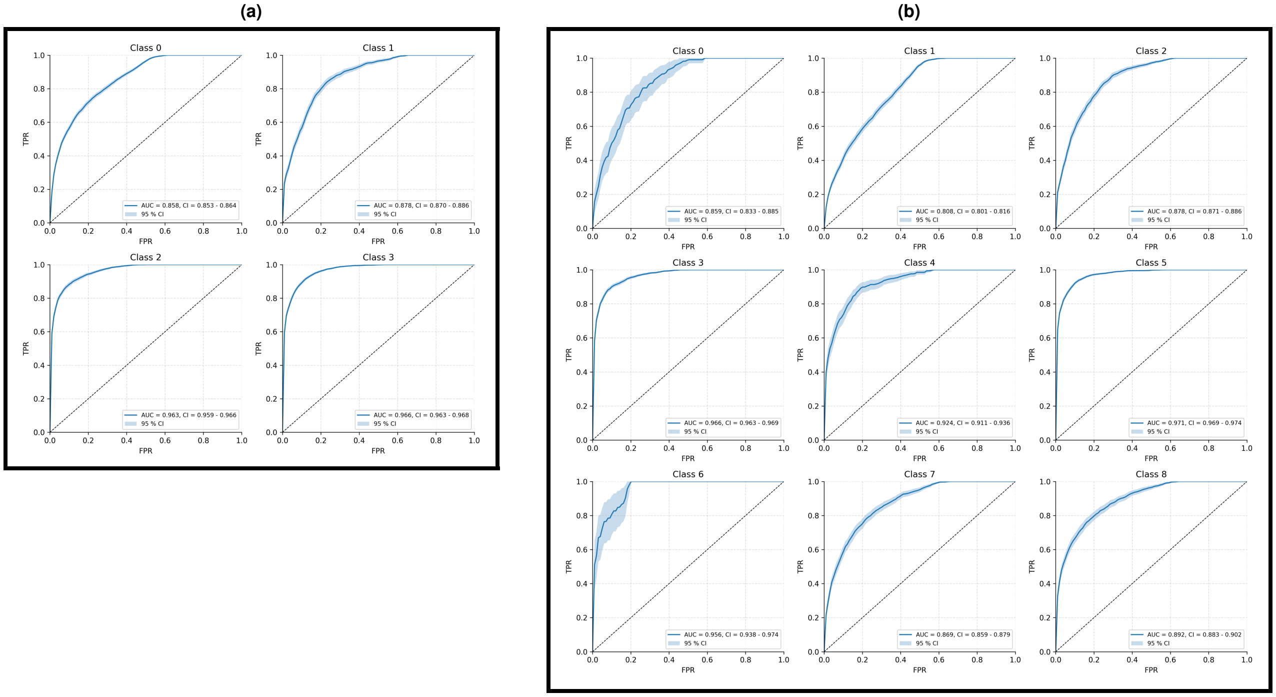

Fig. S7. We selected the ESMC-600 (ESM++ large) (45, 46) as the final model for the base of our production model, balancing performance with throughput using **Supplemental Figure S6**. Shown are AUC plots with statistically sound 95% confidence intervals around an ROC curve added using DeLong's test in the pAUC package (75, 76). For reference, the classes in each model represent the following secondary structures (DSSP conventions with D denoting disordered regions): **SS4**: 0 = C (*coil/loop*); 1 = D (*disordered*); 2 = E (*beta strand*); 3 = H (*alpha helix*), and **SS9**: 0 = B (*beta bridge*); 1 = C (*coil/loop*); 2 = D (*disordered*); 3 = E (*beta strand*); 4 = G (*3$_{10}$ helix*); 5 = H (*alpha helix*); 6 = I (*pi helix*); 7 = S (*bend*); 8 = T (*turn*). **(a)** Four-class secondary structure (SS4) production model (test set). **(b)** None-class secondary structure (SS9) production model (test set).

Table S4. Trends in binder generation of DSM$_{650}$ (Unconditional) and DSM$_{650-ppi}$ (Conditional). Reported is the average predicted binding affinity (ppKd), success rate (percentage higher than best known binder ppKd), and the percentage of designs with a ppKd higher than the best known binder.

Target	Template ppKd error	Conditional		Unconditional	
		Success rate (%)	Avg ppKd	Success rate (%)	Avg ppKd
EGFR	0.09	0.59	8.05	3.55	8.25
BBF-14	1.91	4.78	8.37	7.85	8.66
BHRF1	1.52	88.20	7.05	87.45	7.12
IL-7Rα	2.99	1.75	6.17	5.90	6.33
MBP	0.24	31.70	7.19	38.40	7.27
Cas9	1.11	44.63	7.52	59.35	7.54
PD-L1	1.22	12.64	8.51	12.75	8.47

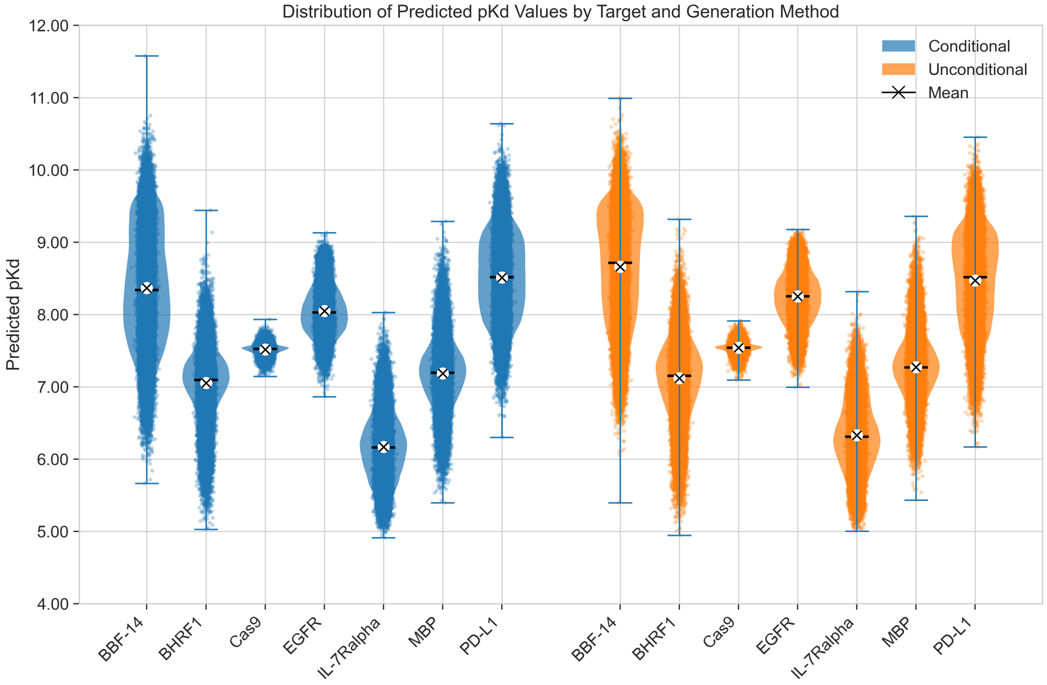

Fig. S8. ppKd trends for each protein target for 100,000 designed sequences via Unconditional and Conditional methods.

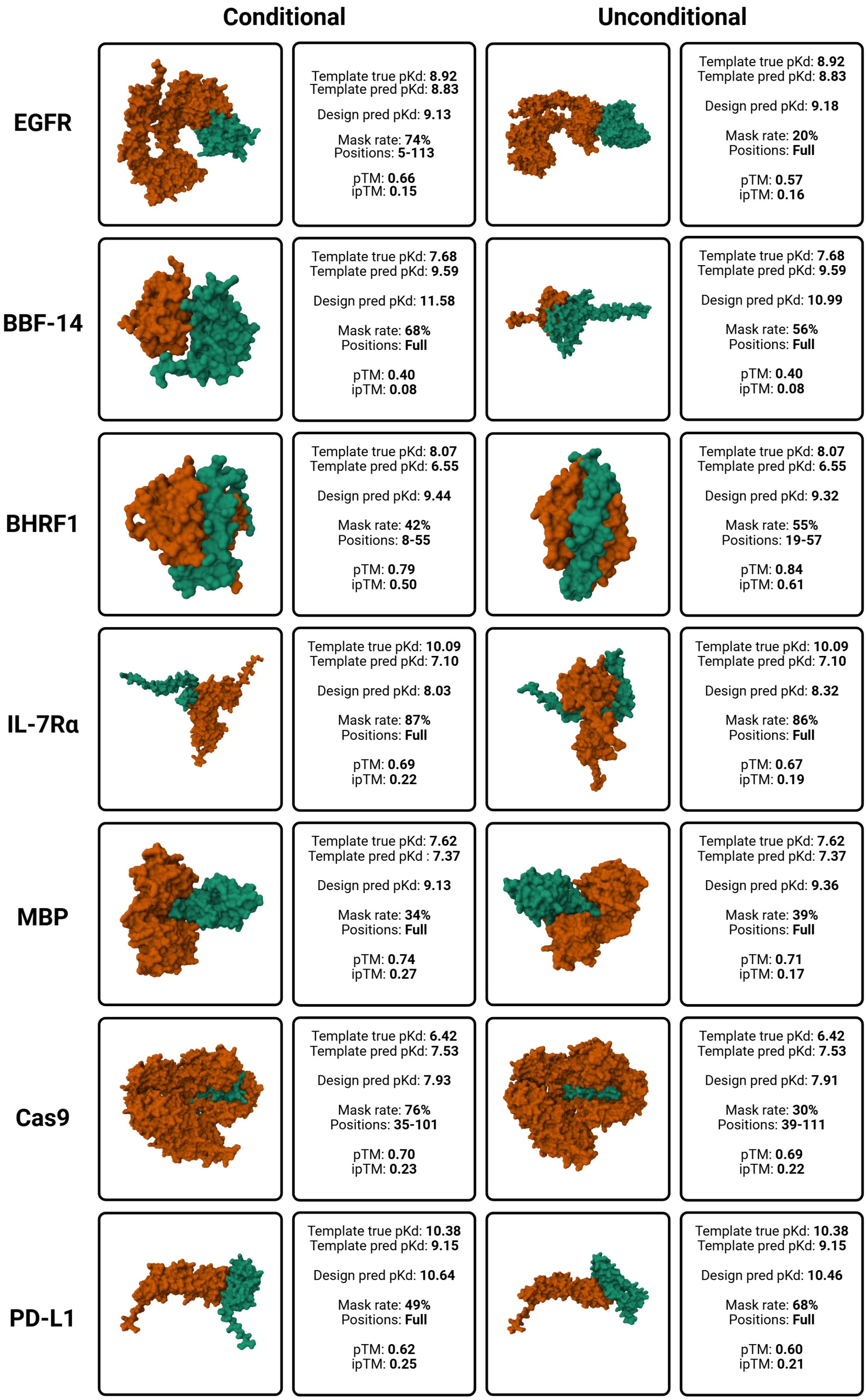

Fig. S9. AlphaFold3 folded dimers of BenchBB targets and the best ppKd out of the 100,000 designs via Synteract2 (57). Template statistics and AlphaFold3 metrics are reported.