Identifying promising sequences for protein engineering using a deep Transformer Protein Language Model

Trevor S. Frisby and Christopher James Langmead[1]
Computational Biology Department
Carnegie Mellon University
{tfrisby, cjl}@cs.cmu.edu

February 15, 2023

Abstract

Protein engineers aim to discover and design novel sequences with targeted, desirable properties. Given the near limitless size of the protein sequence landscape, it is no surprise that these desirable sequences are often a relative rarity. This makes identifying such sequences a costly and time-consuming endeavor. In this work, we show how to use a deep Transformer Protein Language Model to identify sequences that have the most *promise*. Specifically, we use the model's self-attention map to calculate a PROMISE SCORE that weights the relative importance of a given sequence according to predicted interactions with a specified binding partner. This PROMISE SCORE can then be used to identify strong binders worthy of further study and experimentation. We use the PROMISE SCORE within two protein engineering contexts— Nanobody (Nb) discovery and protein optimization. With Nb discovery, we show how the PROMISE SCORE provides an effective way to select lead sequences from Nb repertoires. With protein optimization, we show how to use the PROMISE SCORE to select site-specific mutagenesis experiments that identify a high percentage of improved sequences. In both cases, we also show how the self-attention map used to calculate the PROMISE SCORE can indicate which regions of a protein are involved in intermolecular interactions that drive the targeted property. Finally, we describe how to fine-tune the Transformer Protein Language Model to learn a predictive model for the targeted property, and discuss the capabilities and limitations of fine-tuning with and without knowledge transfer within the context of protein engineering.

***K*eywords** Protein design · Protein engineering · Attention · Transformer · Protein Language Model · Transfer learning · Fine-tuning

1 Introduction

1.1 Protein engineering

Protein engineering is a rapidly evolving field that aims to develop novel protein sequences with useful properties. Protein engineers can be found across many scientific domains, so these useful properties can have equally far-reaching applications, including those that are therapeutic [1, 2, 3, 4], industrial [5, 6, 7], and environmental [8, 9]. Regardless of application, a common issue among protein engineering pipelines is the sheer number of experiments required to identify these useful sequences. The combinatorics alone associated with altering the amino acid composition at different sites within a protein sequence quickly escalates beyond what is experimentally feasible. Developing general-purpose methods that can help protein engineers with experimental decision making is thus important and necessary.

[1] CJL is now the Director of Digital Biologics Discovery at Amgen

Protein engineering pipelines do not fit within a single mold. Just as the applications are varied, so too are the experimental requirements and focal points. While some workflows may focus specifically on a particular objective, many are multi-faceted and include different stages that each involve their own rounds of sequence selection or experimental decision making. A general workflow may be given as:

$$\texttt{Protein Discovery} \rightarrow \texttt{Protein Optimization} \rightarrow \texttt{Model Building}$$

As many protein engineering pipelines include at least one of these components, we will now briefly define each of these stages, and describe the types of experiments and decisions that are made within each.

Protein Discovery While the design of novel protein sequences lies at the heart of most protein engineering objectives, in many settings there is an initial discovery phase. Perhaps most notably, this includes the design of antibody-based therapeutics [10, 11, 12]. Antibodies (Abs) play an important role in initiating an immune response by specifically targeting and binding targeted antigens. An Ab's specificity for and ability to bind an antigen depends on the sequence identity at three complementarity-determining regions (CDR), as residues in these CDR will bind to the surface of an antigen. The Ab residues involved in antigen binding are known as paratopes, and those on the antigen as epitopes. Understanding the interactions between paratopes and epitopes is at the heart of many Ab therapeutics design problems [13, 14]. Recently, a similar class of proteins called nanobodies (Nbs) have also been used for their therapeutic properties [15, 16]. Nbs are shorter than Abs, which makes them easier and cheaper to produce, though are only found naturally in camelid species, which can make them harder to initially collect (whereas Abs can be obtained from common lab mice). Structurally, Abs and Nbs form Y-shaped proteins, where the CDR are located within hypervariable tips. The hypervariability means that sequences collected from an animal sample will have high diversity, and only a subset of the sequences collected will strongly bind a specified antigen target. Collecting this repertoire of diverse sequences is the crux of the protein discovery phase. These Abs or Nbs will undergo assays that quantify their ability to bind a specified antigen. From this, a panel of the best sequences, or *leads*, are identified and used in downstream engineering objectives that take advantage of the leads' strong therapeutic properties. Finding ways to identify strong leads while minimizing the required wet-lab experimentation is thus an important undertaking.

Protein Optimization Optimization is an integral component of all protein engineering campaigns. It is a stage where protein sequences are modified with the goal of identifying those that possess desirable properties. The optimization typically involves making a relatively small number of changes to a given parental protein sequence. For example, in *de novo* protein design [17, 18], computational models are used to propose novel sequences that are expected to have certain structural properties. These proposed sequences are typically experimentally refined (ie. optimized) to ensure they best match the desired properties. More commonly, naturally obtained sequences are optimized directly, such as leads selected from a protein discovery campaign. In any case, there are many types of experimental technologies that may be used to perform the sequence optimization. Traditionally, these approaches involve random mutagenesis at one or more selected sites [19]. Site-specific mutagenesis [20, 21, 22, 23] and deep-mutational scanning [24] further allow greater throughput and specificity as to where in a protein's sequence edits should be made. Emerging approaches based on CRISPR/Cas9 [25, 26] continue this trend. Given the costs associated with these experiments as well as the exponential number of *potential* edits that can be made to any protein sequence, it is important to identify ways that efficiently and effectively select at which sites to perform these mutagenesis experiments.

Model Building In conjunction with the novel sequences that may be designed and synthesized during a protein engineering campaign, it is often desirable to also obtain a structural model of the system under investigation. The model serves as a tool that can be used to better understand physical properties of a protein, such as how it folds or interacts with other proteins. Traditionally, these models have been obtained through experimental means, such as x-ray crystallography [27]. While x-ray crystallography can provide a high-resolution structure of a protein or protein complex, it requires very specific conditions that may preclude its applicability to certain systems. It also only provides a static, or fixed model for a protein's structure, whereas most interactions worth modeling depend on many moving parts. *Computational* models built on machine learning and artificial intelligence bridge this gap between the generalizability and applicability of purely experimental model building approaches. With respect to protein structure prediction, AlphaFold [28] is able to predict any protein's 3D structure, and is shown to have highest accuracy on the benchmark CASP challenges [29]. More generally, computational models can be used to make predictions on any protein property. In some settings, these models can even form a feedback loop, where predictions by a model are used to select experiments, and the experimental results are used to update the model [30, 31, 32]. Whether or not a model is used in-line with running protein engineering experiments, most protein engineering pipelines generate large amounts of data, which makes them well-suited for training machine learning models at the conclusion of all experimentation.

1.2 Representation learning and Transformer Protein Language Models

Protein engineers look to design novel proteins that have certain desired functions or structure. It is has been shown that these biological features are represented through the statistical dependencies between evolutionarily selected sequences that are found in nature [33, 34]. Finding ways to encode and exploit this evolutionary information could help to guide the sequence design choices made by protein engineers. Traditionally, these statistical dependencies have been conveyed through multiple sequence alignments (MSA). MSAs are used to infer patterns that exist between protein sequences, such as those that are conserved or co-vary, both of which can influence protein structure and function [35]. A machine learning technique known as *generative modeling* can allow one to sample new protein sequences that maintain the underlying statistics of a given MSA [36]. However, the computational complexity of sequence alignment can make it difficult to perform across large data sets, and most of the millions of publicly available protein sequences are unaligned and unlabeled [37].

The development of high-capacity generative language models that are trained on unlabeled data has been an area of keen focus within the field of Natural Language Processing (NLP). The Transformer [38, 39] is one such class of model that has been successfully applied to many NLP tasks [40, 41]. While the composition of protein sequences (ie. the 20 natural amino acids) is considerably different than text used in NLP tasks, recent work has shown that the Transformer has great promise when applied to protein sequences [42]. The authors show that, after training on over 250 million distinct protein sequences, their Transformer model, which they call the Evolutionary Scale Model (ESM-1b), successfully encodes biochemical properties of amino acids, biological variation and remote homology, and secondary structure and tertiary contacts.

ESM-1b, as well as other Transformers generally, works by learning a *representation* for an input sequence. This representation is a high dimensional, real-valued vector that encodes the statistical dependencies and patterns present in the sequences used during training. With ESM-1b, this representation is learned by means of a deep neural network. This network consists of 33 connected Transformer blocks, each of which consist of an attention mechanism that feeds into a feed-forward neural network. The attention mechanism [43] is what allows the model to capture the statistical dependencies across all sequences, as it accounts for all pairwise interactions between each position in a sequence. The resulting self-attention map, or matrix of all pairwise interactions, can thus be interpreted as a protein-protein interaction map.

Since ESM-1b is trained on many millions of protein sequences spanning practically all known protein families, it learns many general patterns that exist across all protein sequences. When model building, it is typical to want to learn how to predict a specific task (eg. how proteins from a specific family binds a specific target). Transformer models can provide a good starting point when obtaining these specific models. Transformers have been shown to be effective *few-shot learners* [44], meaning they can be used to predict specific tasks when given only few labeled examples. Developing exact mechanisms to train such models is an active area of research within NLP, and includes methods such as knowledge distillation [45], fine-tuning [46], and transfer learning [47].

In this work, we show how to use the deep transformer protein language model ESM-1b to help select experiments associated with the protein discovery and protein optimization phases of protein engineering pipelines. We then demonstrate how to apply the computational model building techniques of fine-tuning and transfer learning using the Transformer in each of these settings.

2 Materials and Methods

2.1 Encoding a sequence and its protein target as input for ESM-1b

We use the deep Transformer Protein Language Model ESM-1b [42] to jointly model interactions between a protein sequence, s_{binder}, and its protein binding target sequence, s_{target} (Figure 1). For our purposes, ESM-1b is a function $\texttt{ESM}(\cdot)$ that takes as input a length n (tokenized) protein sequence, s, and outputs a tuple $(\mathcal{R}, \mathcal{L}, \mathcal{A})$:

$$(\mathcal{R}, \mathcal{L}, \mathcal{A}) = \texttt{ESM}(s) \tag{1}$$

Here, $\mathcal{R} \in \mathbb{R}^{1280}$ is the sequence *representation*, or feature-rich learned encoding obtained from the final Transformer block. $\mathcal{L} \in \mathbb{R}^{n \times k}$ contains the *logits* returned by the Transformer model. These logits can be used to assign an evolutionary probability to each of k possible tokens (the 20 standard amino acids, nonstandard amino acids, and special tokens internal to ESM-1b) at each position of the input sequence. Finally, $\mathcal{A} \in \mathbb{R}^{n \times n}$ is the self-attention map returned by the final Transformer block's attention mechanism. This attention map is analogous to a predicted protein-protein interaction map, and is the principal output from ESM-1b that we use.

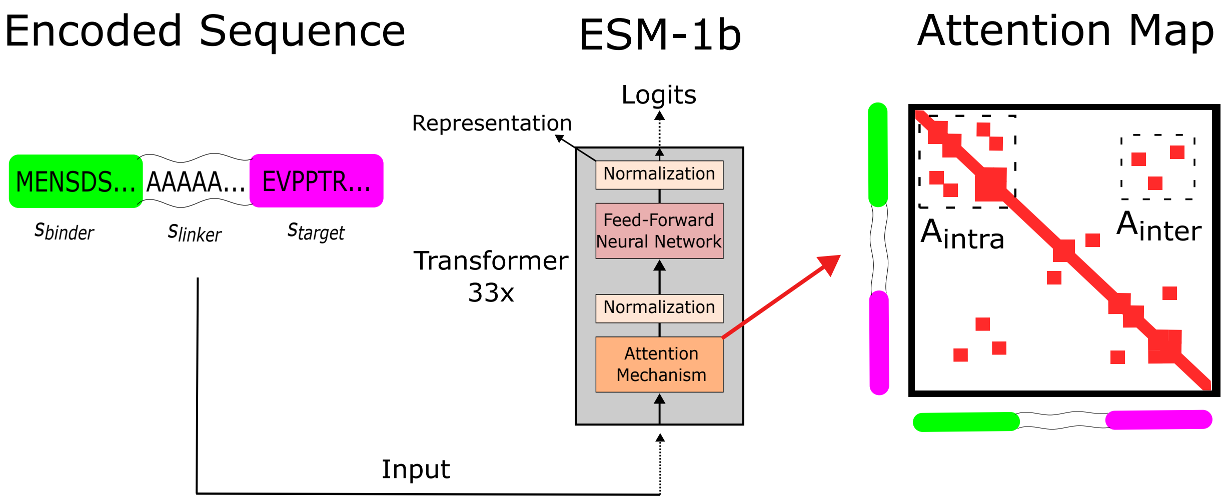

Figure 1: We use a deep Transformer Protein Language Model to identify *promising* sequences within protein engineering campaigns. Using Nb discovery and protein optimization as two test cases, we introduce the PROMISE SCORE, and show how it can help address the needle-in-the-haystack issue of identifying protein sequences with targeted properties. To calculate a PROMISE SCORE, we specifically encode a sequence and its binding target through concatenation with a linker, and use this adjoined sequence as input to the pre-trained ESM-1b model. This Transformer model uses an attention mechanism that captures all expected pairwise interactions between residues within the input sequence through an attention map. We use this map to quantify intermolecular and intramolecular interactions within each protein sequence. The PROMISE SCORE reflects these interactions, meaning stronger sequences should tend to have higher PROMISE SCORES than weaker sequences.

We want to use ESM-1b to identify interactions between s_{binder} and s_{target}. To do this, we construct input sequence s to encode both of these sequences jointly. In our experiments, we show that a simple concatenation scheme is effective. We encode each input sequence s as:

$$s = \texttt{concatenate}(s_{binder}, s_{linker}, s_{target}) \tag{2}$$

where s_{linker} is a sequence of one of the following:

1. **Alanine.** An amino acid. A poly-alanine linker represents a flexible loop in s.
2. **<mask>.** A special token used by ESM-1b.
3. **No linker.** s_{binder} and s_{target} are concatenated without a separating linker.

ESM-1b imposes the constraint that any input protein sequence must be no longer than 1024 residues. In our experiments, the combined lengths of s_{binder}, s_{linker} and s_{target} was always considerably less than 1024. We experimented with linker lengths 10, 50, and 100.

2.2 Formulating a "PROMISE SCORE"

In order to assess the promise of a given sequence s_{binder} as a focus for protein engineering, we use the attention map for s to calculate a "PROMISE SCORE". Treating the attention map as a predicted protein-protein interaction map, we use it to calculate a score that accounts for *intermolecular* interactions between s_{binder} and s_{target}, and, when appropriate, *intramolecular* interactions within s_{binder}. Let n, ℓ, and τ be the lengths of s_{binder}, s_{linker}, and s_{target}, respectfully, and let $N = n + \ell + \tau$. We accomplish this by first isolating sections of $\mathcal{A}$ that correspond to either intermolecular or intramolecular interactions:

$$\mathcal{A}_{inter} = \mathcal{A}_{1:n,n+\ell:N} \in \mathbb{R}^{n\times\tau} \qquad \mathcal{A}_{intra} = \mathcal{A}_{1:n,1:n} \in \mathbb{R}^{n\times n} \tag{3}$$

$\mathcal{A}_{inter}$ and $\mathcal{A}_{intra}$ represent intermolecular and intramolecular interaction profiles, which we use to calculate separate intermolecular and intramolecular scores. We obtain these profiles and scores for all sequences being considered for protein engineering. To standardize the numerical scaling of these profiles in order to ensure fair comparisons between different sequences, we apply min-max scaling to both as follows:

$$\mathcal{A}' = \frac{\mathcal{A} - \texttt{min}(\mathcal{A})}{\texttt{max}(\mathcal{A}) - \texttt{min}(\mathcal{A})} \tag{4}$$

The intermolecular interaction score S_{inter} is given as a sum over $\mathcal{A}_{inter}$ normalized by the length of s_{binder}. For all $a_{ij} \in \mathcal{A}_{inter}$:

$$\mathcal{S}_{inter} = \frac{1}{n} \sum_{i,j} a_{ij} \tag{5}$$

In our protein discovery experiments, we simply use PROMISE SCORE $\mathcal{S} = \mathcal{S}_{inter}$. In many real-life protein engineering pipelines, a subsequent protein optimization stage applies site-specific mutagenesis experiments to a functional but suboptimal baseline protein sequence $s_{baseline}$. This could correspond to a lead sequence selected from a previous protein discovery campaign, or a wildtype sequence that has been identified previously. We obtain intramolecular interaction score $\mathcal{S}_{intra}$ by calculating a difference between the intramolecular interaction profiles of $s_{baseline}$ and s_{binder}. Where $\mathcal{B}_{intra}$ is the intramolecular interaction profile for $s_{baseline}$, we obtain such a score as:

$$\mathcal{S}_{intra} = \|\mathcal{B}_{intra} - \mathcal{A}_{intra}\| \tag{6}$$

The larger $\mathcal{S}_{intra}$ is, the greater the difference between intramolecular interaction profiles. We assume that the functional property of $s_{baseline}$ is driven in part by the the intramolecular interactions reflected by $\mathcal{B}_{intra}$. This leads us to expect that a functional s_{binder} should have an intramolecular interaction profile that is more similar than dissimilar to $\mathcal{B}_{intra}$. This reasoning motivates our formulation of the PROMISE SCORE $\mathcal{S}$ during protein optimization— a linear combination of $\mathcal{S}_{inter}$ and $\mathcal{S}_{intra}$:

$$\mathcal{S} = \mathcal{S}_{inter} - \lambda \mathcal{S}_{intra} \tag{7}$$

A promising design will have many intermolecular interactions and an intramolecular interaction profile that is similar to a known functional baseline.

Hyperparameter $\lambda \geq 0$ governs the degree to which the intramolecular term impacts the overall score. In principle, any standard hyperparameter fitting method could be applied to set λ. We have devised a means to automatically select λ that takes advantage of the fact that S_{inter} and S_{intra} can be pre-computed for any sequence without the need for an external label. Let $\boldsymbol{\mathcal{S}}_{inter}$ and $\boldsymbol{\mathcal{S}}_{intra}$ be the set of all $\mathcal{S}_{inter}$ and $\mathcal{S}_{intra}$ computed for each sequence that is being evaluated. Our approach is to treat λ as a scaling factor between $\mathcal{S}_{inter}$ and $\mathcal{S}_{intra}$. Let $\overline{\boldsymbol{\mathcal{S}}_{inter}}$ refer to the mean of $\boldsymbol{\mathcal{S}}_{inter}$, and $\overline{\boldsymbol{\mathcal{S}}_{intra}}$ the mean of $\boldsymbol{\mathcal{S}}_{intra}$. We identify the scaling factor by taking a simple ratio, and use this approach to select λ in all of our experiments:

$$\lambda = \frac{\overline{\boldsymbol{\mathcal{S}}_{inter}}}{\overline{\boldsymbol{\mathcal{S}}_{intra}}} \tag{8}$$

2.3 Fine-tuning and knowledge transfer with ESM-1b

While we have so far used the pre-trained ESM-1b model as-is in order to calculate PROMISE SCORES, we now describe how it can be *fine-tuned* in order to learn a model that predicts the binding strength between s_{binder} and s_{target}. In order to imbue ESM-1b with the ability to provide such a prediction, we attach a three-layered neural network to the head of the model. In other words, ESM-1b's learned representation serves as input for the neural network head. Fine-tuning occurs during model training. During backpropogation, we fine-tune by updating the parameters of ESM-1b's final Transformer block in conjuction with the parameters of the neural network head. All other ESM-1b parameters are left fixed at their pre-trained state. In our experiments, the neural network head consists of three linear layers connected by rectified linear unit (ReLU) activation functions. The full model is trained for 30 epochs using PyTorch's AdamW optimizer [48, 49] with default parameter settings (learning rate $\gamma = 1\mathrm{e}{-6}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay coefficient $\lambda = 0.01$).

We expect that it would be beneficial to be able to take knowledge gained in one engineering campaign and apply it to that of another (related) campaign. To this end, we show the ability to transfer knowledge learned from one protein engineering objective to another within the discovery and optimization stages. Within protein discovery, we use models for protein binding learned from Nano-HSA data and show their ability to predict binding on Nano-GST data, and vice-versa. Similarly within protein optimization, we take models learned from BRCA1-BARD1 binding data and show their ability to predict binding between Spike and ACE2, and vice-versa.

Table 1: An overview of the data we used to evaluate the PROMISE SCORE. The sequence IDs are given as the DataBase:ID indicating where the sequence can be accessed, or are NA (not applicable).

Dataset	Baseline s_{binder}	s_{target}	Total sequences	"Strong" percentage
Nano-HSA	NA	UniProt:P02768	20,670	47.5
Nano-GST	NA	UniProt:P08515	51,330	31.1
BRCA1	PDB:1JM7	PDB:1JM7	1389	10
Spike	PDB:6M0J	PDB:6M0J	3651	10

2.4 Data

We use the PROMISE SCORE at two phases of the protein engineering pipeline— protein discovery and protein optimization. Table 1 provides an overview of the data we use in our experiments.

At the protein discovery phase, we show how to use the PROMISE SCORE to select lead sequences from two Nb repertoire data sets [50]. The repertoires were previously collected from an immunized llama using protein targets human serum albumin (Nano-HSA) and glutathione S-transferase (Nano-GST). The relative binding strengths of each Nb sequence to its target sequence was originally provided as a categorical label, where one of the three possible labels corresponds to strong binding sequences. We re-labeled each sequence as "strong binding" or "not strong binding" to obtain a binary descriptor for relative binding strength, which we use in our experiments. In total, the Nano-HSA data set contains 20,670 Nb sequences and the Nano-GST data set 51,330. The percentage of strong binders in Nano-HSA and Nano-GST was 47.5% and 31.1%, respectively.

When computing PROMISE SCORE $\mathcal{S}$ with these data, we let s_{binder} be the full Nb sequence. However, when obtaining $\mathcal{A}_{inter}$, we focused our attention on the CDR3 region of each Nb by only considering the region of attention map $\mathcal{A}$ that corresponds to intermolecular interactions between the CDR3 and target sequence.

At the protein optimization stage, we show how to use the PROMISE SCORE to select mutagenesis experiments for two different proteins— BRCA1 and Spike. We obtained BRCA1 sequences from single site mutational scans [51] that measured the downstream activity caused by interaction between BRCA1 and its binding target protein BARD1. These measurements were obtained from 1389 single site mutations. Spike sequences were similarly obtained from single site mutational scans [52] that measured the binding affinity between Spike and its binding target protein ACE2, and it included 3651 single site mutations. We obtained binary labels for both datasets by binning the raw experimental values, where the top 10% were labeled as "strong" sequences, and the rest "not strong".

3 Experiments and Results

3.1 The PROMISE SCORES of strong and weak binders are different

In order for the PROMISE SCORE to provide an effective means for selecting sequences for protein engineering, the scores of known strong binders should be distinguishable from those of weak binders. According to how the PROMISE SCORE is formulated, we further expect that the scores of strong binders should be generally higher than those of weak binders. For both Nb discovery and protein optimization, we computed PROMISE SCORES for all sequences using different linker types and lengths, and compare how strong binders were scored relative to weak ones. The raw scores are only comparable within a given linker type and length— the relative ability to discern strong from weak binders is what can be compared.

During Nb discovery (Figure 2 top), there is a significant difference in the distribution of PROMISE SCORES of strong and weak binders for each linker type and length, and the average score of strong scoring sequences is higher than that of weak binders. For Nano-GST, using no linker obtains PROMISE SCORES that have greatest ability to discern strong from weak binding Nb sequences on average. For Nano-HSA, a length 100 alanine linker provides the greatest ability to discern. For both Nano-HSA and Nano-GST, calculating a PROMISE SCORE using a mask linker of any length or a length 50 alanine also confers the ability to discern strong from weak binders (Figure S1 top).

During protein optimization (Figure 2 bottom), in all but two cases, there is again a significant difference in the distribution of PROMISE SCORES of strong and weak binders for each linker type and length, where the stronger sequences have a higher PROMISE SCORE on average than their weak counterparts. The exceptions are with BRCA1 using no linker and a length 10 mask linker, though they just barely misses the threshold for significance at $p = 0.066$ and $p = 0.054$, respectively. In general, the PROMISE SCORES for each linker type and length in protein discovery and

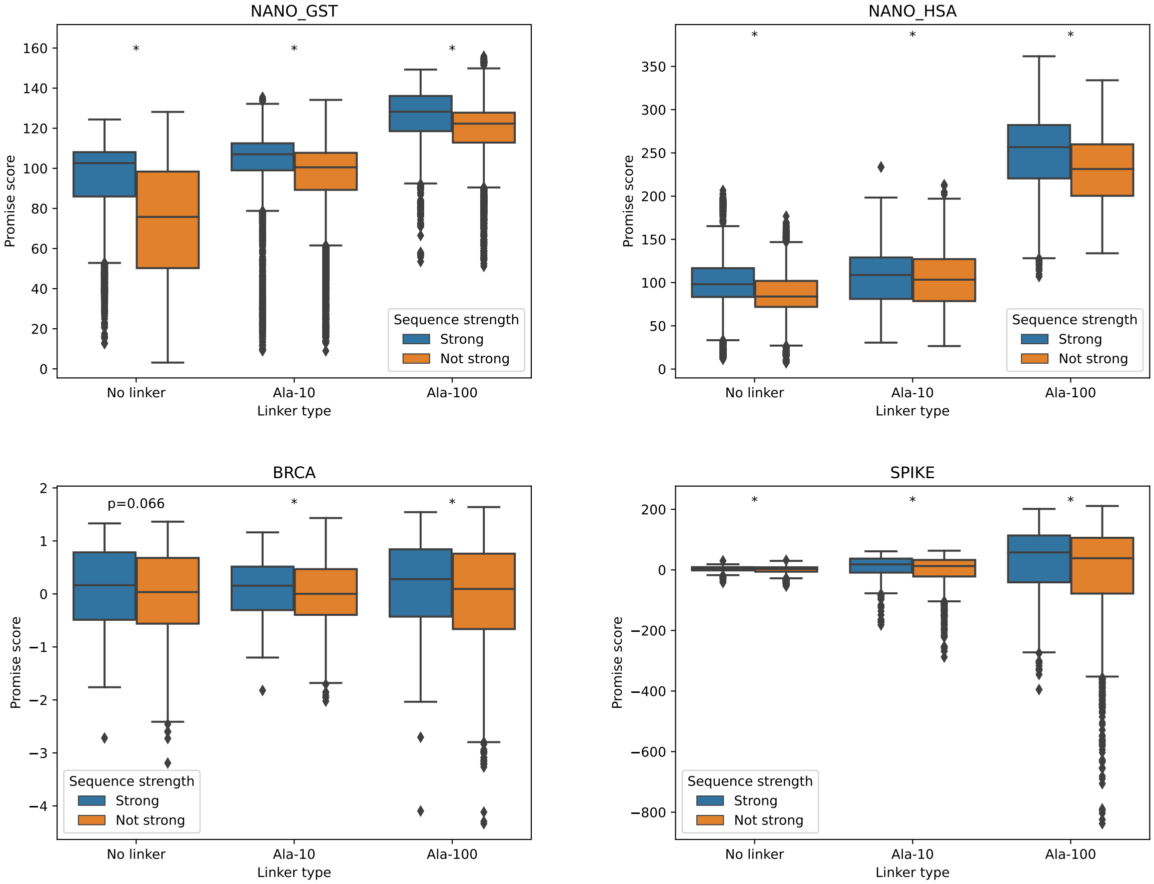

Figure 2: We calculated PROMISE SCORES for each sequence identified within two Nb discovery campaigns (Top) and two protein optimization campaigns (Bottom), and compare how "strong" sequences scored compared to those that were "not strong". In each case, strong binders had an average PROMISE SCORE greater than sequences that were not strong. These differences were either statistically significant ($p < 0.05$, Mann-Whitney U-test, denoted with *), or had a p-value just above the 0.05 threshold (BRCA no linker– $p = 0.066$).

protein optimization provides a narrow (but still real) ability to discern strong from weak binders. This also includes using a mask linker or length 50 alanine (Figure S1 bottom).

3.2 The PROMISE SCORE identifies beneficial experiments

Since strong binders tend to have higher PROMISE SCORES than weak binders, we can use the score to rank the sequences in a repertoire and expect that a set of top-ranked sequences will be enriched with strong binders. Figure 3 demonstrates this for different sized subsets obtained by taking the top $1, 2, \dots, 100$ sequences ranked by the PROMISE SCORE. Thus, the PROMISE SCORE may be useful as a means to identify binders.

During protein discovery, we compare the cumulative strong binder enrichment for the top 100 sequences obtained using the PROMISE SCORE calculated with no linker, 10 length alanine, and 100 length alanine linkers. The enrichment is given as the ratio of the observed strong binder frequency at a given sample size to the expected strong binder frequency (ie. the underlying true strong binder frequency). We compare to sequence-derived calculations commonly used to assess Nb sequences when identifying potential lead sequences. Though none of these metrics directly measure a protein's ability to bind a target, they measure qualities that are reasonably expected to correlate with this property. These metrics include:

1. **CDR3 Length.** The number of residues within the CDR3. While an optimal CDR3 length depends on the particular target sequence, previous analysis [50] of the Nb sequences used in our experiments determined that sequences with shorter CDR3 tended to be stronger binders.

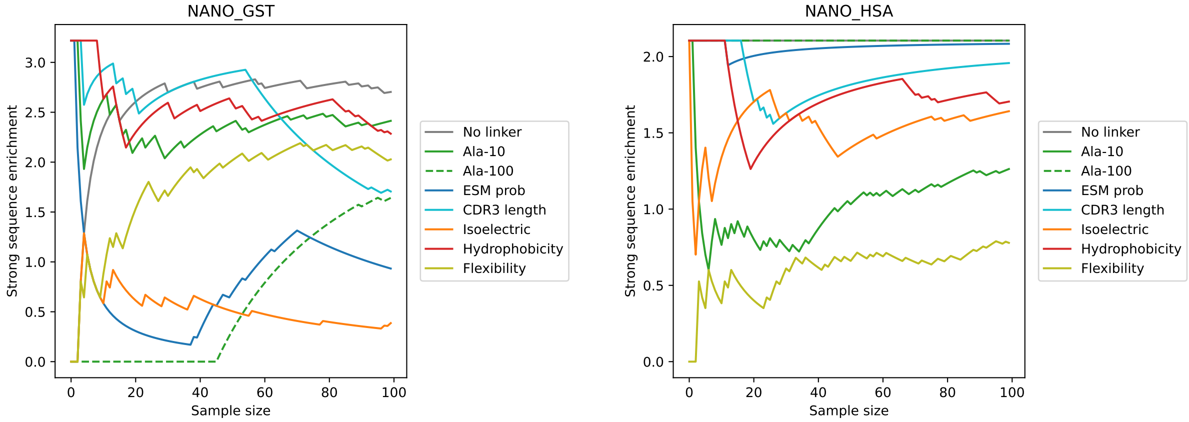

Figure 3: We used PROMISE SCORES to help identify potential lead sequences from two Nb discovery campaigns. We found that PROMISE SCORE derived methods identified sequences that were enriched with strong binders, and often identified a higher frequency of strong binders compared to other sequence-based calculation methods. In both Nb discovery campaigns, the best overall strategy used the PROMISE SCORE.

2. **Nb Isoelectric point (pI).** [53, 54] The pH where the Nb carries no net electrical charge. Low pI Nbs are more likely to aggregate, which can lead to inactivation and induces immunogenicity [55].

3. **CDR3 Hydrophobicity.** The frequency of hydrophobic residues within the CDR3. The prevalence of hydrophobic residues can increase the propensity for aggregation [56].

4. **Nb Flexibility.** [57] The relative mobility of Nb residues. An Nb with greater flexibility may have greater ability to bind its target.

We ranked all sequences according to these metrics and report the cumulative strong binder enrichment of the top 100 sequences. CDR3 length and hydrophobicity are ranked in ascending order, whereas Nb isoelectric point and flexibility are ranked in descending order. We also used the logits produced by ESM-1b to calculate a probability under the ESM-1b model for the CDR3 of each sequence. We do this by computing the softmax over the logits associated with each position of the CDR3. The probability is then given as the product over the probabilities assigned to each residue present in the sequence. We only used the Nb sequences as input for ESM-1b when calculating a probability, rather than the sequence concatenated to the target sequence, and report the cumulative strong binder enrichment of the top 100 most probable sequences.

With Nano-GST (Figure 3 left), the PROMISE SCORE computed with no linker and a length 10 alanine linker have a cumulative strong binder enrichment greater than two at almost every sample size up to 100. At sample size 100, they produce two of the highest cumulative strong binder enrichment values, along with a length 50 alanine linker (Figure S2 left). Using the PROMISE SCORE computed with a length 100 alanine linker struggles to identify strong binders for the top 50 scoring sequences, but by sample size 100 produces a cumulative enrichment greater than 1.5. Calculating a PROMISE SCORE using a mask linker tends to have a strong binder enrichment less than 2 (Figure S2 left). Among the sequence-based calculations, hydrophobicity tends to produce the highest cumulative enrichment, and typically ranks between no linker and alanine 10. The shortest CDR3 are enriched with strong binders, but the enrichment fades by sample size 100. The isoelectric point of Nb sequences is generally not enriched for strong binders at these sample sizes, and neither are the probable sequences under the ESM-1b model.

With Nano-HSA (Figure 3 right), the PROMISE SCORE computed with no linker and a length 100 alanine linker are consistently the two strategies that yield the greatest strong binder enrichment. Both have cumulative enrichment greater than two across all sample sizes up to 100. Using any length mask linker or 50 length alanine yields an enrichment greater than 1.5 for each sample size up to 100 (Figure S2 right). The most probable sequences under the ESM-1b model also fare well, with a strong binder enrichment consistently right around 2. The PROMISE SCORE calculated with a length 10 alanine linker has the lowest strong binder enrichment among the ESM-1b derived metrics. Each of CDR3 length, CDR3 hydrophobicity, and Nb isoelectric points tend to have cumulative strong binder enrichment that is lower than using no linker or a length 100 linker, but higher than the length 10 linker. Sequences predicted to have high flexibility are not enriched for strong binders.

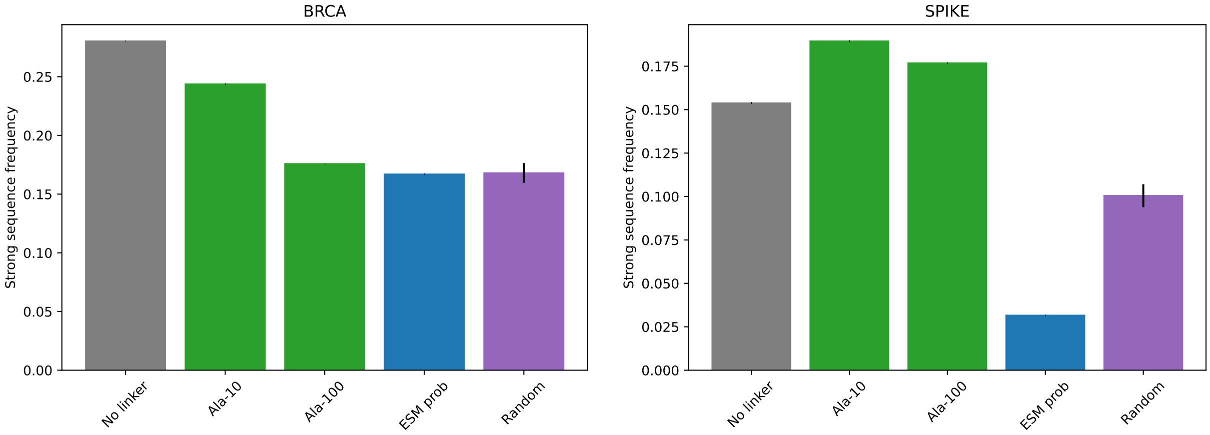

Figure 4: We used PROMISE SCORES to select 10 site-specific mutagenesis experiments on each of BRCA1 and Spike. For most linker types and lengths, the PROMISE SCORES identified sites that had a higher strong binder frequency than selecting randomly or using evolutionary sequence probabilities.

To summarize across both Nb discovery campaigns, the no linker PROMISE SCORES most consistently identified sequences enriched with strong binders. Among the comparison metrics, each of the isoelectric point, protein flexibility, and ESM-1b were inconsistent– sequences selected according to each were enriched with strong binders in one campaign and not in the other. While both hydrophobicity and CDR3 length consistently identified sequences enriched with strong binders, at most sample sizes, the enrichment was lower than that obtained by the no linker PROMISE SCORE.

We also show how the PROMISE SCORE can be used to select site-specific mutagenesis experiments during protein optimization. Rather than selecting from a set of discovered sequences, the task here is to select specific positions from a baseline sequence at which to conduct single-site mutagenesis. The intent is to select the positions that will cumulatively yield the highest percentage of strong binders. To do this, we first obtain the PROMISE SCORE for each individual (linked) variant sequence. We then take the top 500 scoring sequences, and identify at which position each sequence is variant relative to the baseline sequence. We treat this tally as a "vote", and the selected sites correspond to the positions that obtain the most votes. In our experiments, we identified the top 10 sites. Figure 4 shows the strong binder frequency obtained by this procedure on the BRCA1 and Spike protein optimization sequences.

With BRCA1 (Figure 2 left), the PROMISE SCORE calculated with no linker identified sites that yielded the highest percentage of strong binders. Alanine with length 10 linker also performed well and had the second highest strong binder frequency. The length 100 alanine linker had the third highest strong binder frequency. We generally found that short mask linkers were also effective (Figure S3 left). To better quantify how enriched for strong binders these experiments were, we simulated randomly selecting mutagenesis sites 50 times for comparison. Both no linker and a length 10 alanine linker have greater strong frequency enrichment relative to the random sampling. The length 100 alanine linker performed comparably to random sampling. We also calculated the probability of each variant sequence under the ESM-1b model, and used those probabilities to select single sites for mutagenesis. We used the same voting scheme as with the PROMISE SCORE, but instead used the sequence probabilities. The sites with most probable sequences yielded strong binders at frequency similar to random sampling.

With Spike (Figure 2 right), the PROMISE SCORE calculated with alanine linkers yield the highest strong binder frequency. The length 10 linker was highest and the length 100 linker second highest. Using no linker yielded the third highest frequency. All three of these approaches performed similarly to each other and better than random sampling. We found that using a mask linker was generally less effective than using an alanine linker (Figure S3 right). The sites selected by the most probable sequences under the ESM-1b model were not enriched for strong binders, and had a lower strong binder frequency than random sampling.

3.3 The attention map provides insights into protein-target interactions

In addition to identifying strong binders for the purposes of protein engineering, PROMISE SCORES can also provide biological insights about the modes of interaction between a protein sequence and its target protein. To calculate any PROMISE SCORE, we use the attention map to quantify intermolecular interactions between a protein sequence s_{binder} and its target s_{target}, given by $\mathcal{A}_{inter}$. The rows of $\mathcal{A}_{inter}$ correspond to each position of s_{binder}, whereas the columns

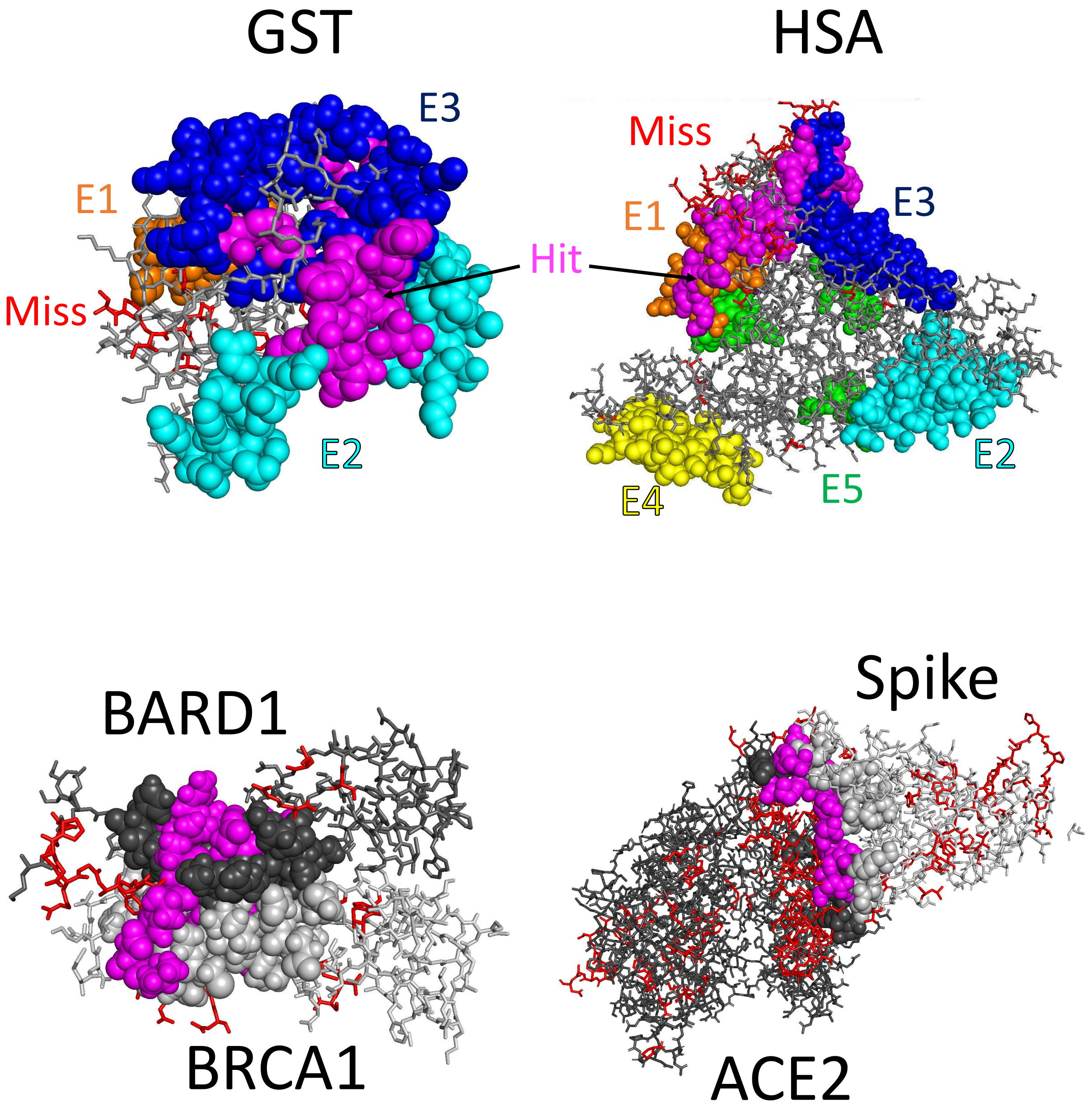

Figure 5: We use intermolecular contact map $\mathcal{A}_{inter}$ to calculate residue-specific scores to identify residues that contribute the most to intermolecular interactions. With Nb discovery (Top), we highlight overlap between our top scoring residues and epitopes previously validated through cross-linking mass spectrometry. The magenta spheres indicate this overlap. Each other colored sphere indicates a residue within a validated epitope region. The red sticks correspond to top scoring sites that did not fall in any validated region (denoted "Miss"), and the gray sticks indicate all other residues within the protein sequence. With protein optimization (Bottom), we show structures of each protein bound to its target. The spheres indicate residues known to play a role in protein-target binding. Those shaded magenta were identified by the per-residue scoring. Red sticks indicate residues outside this binding region that were also identified. Dark gray sticks indicate s_{target} residues, and light gray s_{binder} residues. The structures we used are given by PDB IDs 1DUG, 1AO6, 1JM7, and 6M0J.

correspond to each position of s_{target}. While calculating the PROMISE SCORE involves summing over all of $\mathcal{A}_{inter}$, we can sum over the columns to instead obtain a per-residue score for the target sequence. We use these per-residue scores to identify specific residues on s_{target} that are expected to interact the most with a given s_{binder}. To identify the target residues predicted to be most involved in intermolecular interactions, we took the top 10% of protein sequences according to PROMISE SCORE, and used these sequences to obtain an average per-residue score for the target residue. We then identified target residues that scored in the top 20 percentile as those most strongly believed to be involved in intermolecular interactions. Figure 5 shows 3D structures of each target protein tested with the top scoring residues highlighted (Figure S4 shows each of these structures rotated 180°).

In a Nb discovery campaign, identifying target residues involved in intermolecular interaction is akin to identifying Nb epitopes. With both Nano-GST and Nano-HSA, we compared the strongest scoring residues on target proteins to experimentally validated epitope regions. With Nano-GST (Figure 5 top left), we identified 42 GST residues predicted to be involved in intermolecular interactions. Of these identified residues, 12 of them are located within the experimentally validated E3 epitope. E3 was the strongest validated epitope, as 50% of the tested Nbs bound to this region. Seven of the 12 residues are singletons within the E3 region located at residues 158-200, and the other five residues are at the contiguous E3 region at residues 213-217. Of the remaining identified residues, one fell within the E2 epitope region (residue 125), and the rest did not fall within one of the validated epitope regions.

With Nano-HSA (Figure 5 top right), we identified 98 HSA residues predicted to be involved in intermolecular interactions. These predicted interacting residues overlap most strongly with validated epitopes E1 and E3. For E1, these include the contiguous region at residues 23-25, 113-114, and 126-127, as well as the singleton residues 28, 111, 116, 118, 123, and 138. For E3, these include the contiguous regions at residues 5-13 and 93-94 as well as singletons at residues 96 and 98. 5% of the validated Nbs bound to epitope E1 and 20% at epitope E3. There were two sporadic singleton residues identified in other validated epitopes, residue 172 in E5, and residue 301 in E2. The remaining selected residues did not fall in one of the validated epitope regions, though blanketed the protein region surrounding epitopes E1, E3, and E5.

We performed similar analyses with both of the protein optimization protein sequences. In addition to computing per-residue scores for s_{target}, we also computed per-residue scores for s_{binder} by summing over the rows of $\mathcal{A}_{inter}$. As with the s_{target} per-residue scores, we identified the top 20% of residue scores as those most involved in intermolecular interactions. We found 3D structures for both the BRCA1-BARD1 (Figure 5 bottom left) and Spike-ACE2 complexes (Figure 5 bottom right), and highlight the residues on s_{binder} and s_{target} that had highest scores indicative of intermolecular interactions. In general, we found that the largest contiguous cluster of identified residues were located at the protein-target binding interface.

Binding between BRCA1 and BARD1 is known to be driven by interactions between BRCA1 residues 8-22 and 81-96 and BARD1 residues 36-48 and 101-116 [58]. Of the 23 high scoring BRCA1 residues we identified, 10 of them fell within these regions (residues 14-15, 18-19, 21-22, 86, 93, and 95-96), and the remaining high scoring residues flanked each of these regions (residues 24, 26-27, 41-42, 79, and 97-103). Similarly, 10 out of the 24 BARD1 residues we identified fell within these ranges (residues 36-38, 42, 45, 102, 105, 108-109, and 112), and the rest also flanked these regions (residues 27-35, 49, 51, 54, 99, and 134).

Spike-ACE2 binding is facilitated through specific interactions between 17 different Spike residues and 20 different ACE2 residues [59]. We identified 39 Spike residues expected to be most involved in intermolecular interaction, two of which overlap with the known interacting residues. While some of the non-overlapping residues are in the vicinity of these binding regions, we found that identified residues were widely spread across the entire Spike protein. With ACE2, we identified 121 residues, including 13 of the known interacting residues. While many of the non-overlapping residues were again in the vicinity of the known binding region, many were also spread throughout the ACE2 protein.

In order to demonstrate that the per-residue patterns we have described above did not arise by chance, we simulated *randomly* selecting residues for each protein system. In each case, we randomly selected a number of residues equal to the amount identified by the per-residue PROMISE SCORE. Over 50 replicates, we identified the frequency with which these randomly selected residues fell within a known binding area, and compared these frequencies to those obtained with the PROMISE SCORE selected residues (Figure S5). Indeed, for each of the regions that we identified substantial overlap between the residues selected by the per-residue PROMISE SCORES and a known binding region, we found that the Promise score selected residues within that region at a rate higher than if selected at random.

3.4 Using ESM-1b to learn models for protein binding

Having used ESM-1b to prioritize sequences expected to strongly bind their target, we now show to what extent we can use ESM-1b to learn predictive classification models for this targeted property. With Nb discovery, we randomly selected 5000 sequences from Nano-GST and Nano-HSA to serve as two separate training sets, and another 15,000

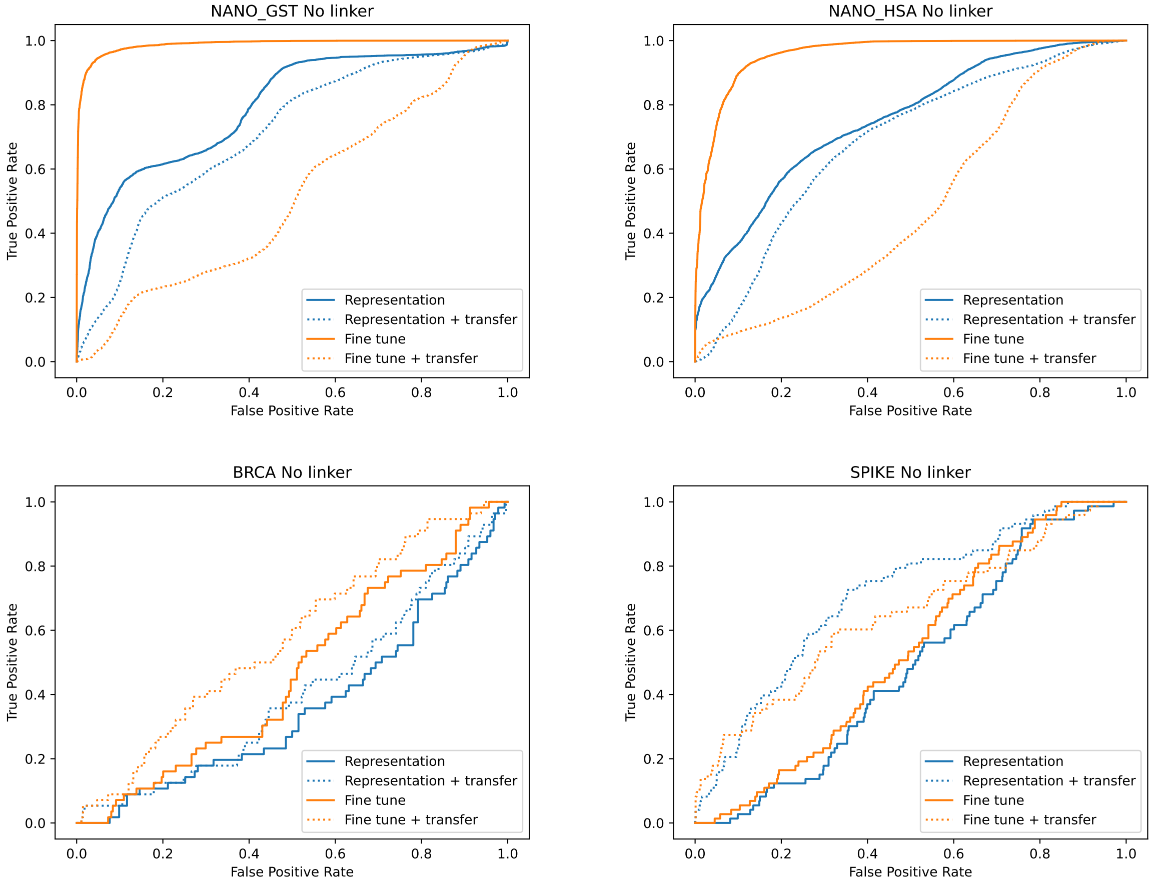

Figure 6: ROC curves for models trained to classify strong and weak binders for both Nb discovery (top) and protein optimization (bottom). Orange curves correspond to models that used fine-tuning, and blue those that did not. Solid curves correspond to models applied in a traditional machine learning paradigm, whereas dashed lines used transfer learning. With Nb discovery, fine-tuned models used in the traditional setting were very strong. While all models learned in protein optimization were of lower quality, we saw evidence that transfer learner led to model improvement.

sequences from each as testing sets. Within a protein engineering context, the training sequences are analogous to a preliminary round of experimentation carried out for the purpose of collecting initial data. We then used fine-tuning to train two classifiers, one that was trained with Nano-GST, and one that was trained with Nano-HSA. We then used these models in a traditional (ie. used to predict binding strength to the same target sequence used during training) and transfer learning manner. As a point for comparison, we also trained classifiers without fine-tuning. That is, during training, we only updated the parameters of the neural network head, and left all ESM-1b parameters fixed. This demonstrates the ability to simply use the pre-trained representations when learning a model for protein binding. We also used these models in both the transfer and traditional learning settings.

We obtained comparable results with both Nano-GST and Nano-HSA (Figure 6 top). Fine-tuned models used traditionally were far and away the best at predicting binding strength. The model used to predict Nb binding strength with GST had an AUC of 0.988, and the model used to predict Nb binding strength to HSA an AUC of 0.958. The non-fine-tuned models that used pre-trained representations were second best, though clearly less effective than their fine-tuned counterparts. The AUC obtained were 0.795 with GST and 0.753 with HSA. We found that knowledge transfer hurt the predictive capability of both the fine-tuned and non-fine-tuned models. The non-fine-tuned models were hurt less, and still achieved an AUC of 0.712 with GST and 0.683 with HSA. Fine tuning combined with transfer learning essentially yielded random classifiers, as the models has AUC of 0.514 and 0.474 with GST and HSA, respectively. We found similar results for each linker type used in the input sequence encoding (Figure S6).

We next used ESM-1b to train classifiers using the two sets of protein optimization sequences. For both BRCA1 and Spike, we randomly selected 75% of the single residue sites in each protein. We used the sequences previously obtained via single-site mutagenesis at these selected positions as training data. The mutatations at all other sites were then used as test data. We again trained models with and without fine-tuning, and used both in a traditional and transfer setting.

While we were able to obtain at least one reliable classification model for protein binding using fine-tuning with the Nb discovery data, we found that it was much more difficult to do so within the context of protein optimization (Figure 6 bottom). With BRCA, the best model used fine-tuning with transfer, but only had an AUC of 0.573, meaning the model does not effectively classify strong and weak binders. Interestingly, the ROC curves provide some evidence that transfer learning may have some positive effect relative to traditional learning, though again none of the models score highly according to AUC. With Spike, we again found that the models that used transfer learning outperformed those that used traditional learning. We obtained the best model using no fine-tuning with transfer, which had an AUC of 0.708. The model with fine-tuning plus transfer had an AUC of 0.646. The traditional learning models performed no better than random classifiers. As with Nb discovery, we found that using different linkers had minimal effect on the resulting models (Figure S7).

4 Discussion

We have shown how to use the deep Transformer Protein Language Model ESM-1b to calculate a PROMISE SCORE that can prioritize specific sequences and be used to guide protein engineers towards designs that are more likely to bind to a given target. We believe this can lead to more efficient and successful protein engineering campaigns.

4.1 The role of linkers when computing PROMISE SCORES

In our experiments, we hypothesized that self-self attention maps would reveal, in a coarse-grained fashion, the interactions between a sequence s_{binder}, and a binding partner s_{target}. We tested this hypothesis by first concatenating s_{binder} and s_{target} with a separating linker, and use this as input to the ESM-1b model. Strictly speaking, ESM-1b was not designed with this use case in mind. Rather, the model was trained using millions of *monomer* sequences, and has been used to describe secondary structures or make downstream predictions for *singular* protein sequences. And yet, when we use the attention map generated with our concatenated protein-target sequence, we are able to calculate PROMISE SCORES that tend to favor strong binders. From a biological perspective, this suggests that the types of interactions present in the monomers used to train ESM-1b are informative enough to identify the intermolecular interaction between a protein and its binding partner. From a modeling perspective, this suggests that representation learning approaches trained on monomers are strong enough to identify the ways that multiple sequences interact with each other.

Where we used linkers, we experimented with two types — a polyalanine, and the special <mask> character that is internal to ESM-1b. Alanine's simple methyl side chain makes it a flexible amino acid. The idea was that the model may interpret a chain of alanines as a "tether" connecting the protein sequence and its target protein sequence. The flexibility would allow the two protein sequences to interact as if they were two unconnected proteins in close proximity. The <mask> character is used when training ESM-1b. Random residues are replaced with the <mask> character, and the model learns to predict which residue belongs where the <mask> is placed. In our use, the <mask> character denotes space between sequences without specifically committing to any particular residue type that may bias the interactions between residues.

Generally speaking, we found that using no linker always produced a good result. We found that using a linker could occasionally perform better than no linker, but not overwhelmingly so. When comparing the two linker types, we found that the natural amino acid alanine produced more consistent results, whereas there was much greater variance in the performance of PROMISE SCORES calculated with <mask> linkers. For future work, we would recommend using no linker or a short alanine linker.

4.2 The generalizability of PROMISE SCORES

We first showed how to use PROMISE SCORES to select lead sequences using Nbs obtained from two different Nb discovery campaigns. Notably, we showed that PROMISE SCORE selected sequences are often stronger than sequences selected by a number of sequence-based scoring metrics. This again shows the power of representation-learning approaches like ESM-1b. Despite not being trained to know anything about the biophysical properties of a given protein sequence, we show that we can use the model to make inferences about the relative strength of a protein with greater effect than using biophysical calculations.

We also showed how to use PROMISE SCORES to select single site mutagenesis experiments during protein optimization. The primary task associated with protein optimization is considerably different than that of Nb discovery. Rather than identifying a subset of promising sequences from a naturally large and diverse repertoire, the goal is to identify specific sites on a baseline sequence to perform single-site saturation mutagenesis. As a result, the set of sequences under consideration will be smaller and more uniform than those encountered during Nb discovery, both of which add to the challenge of this task. We overcame this challenge by adding an extra term ($\mathcal{S}_{intra}$) to the calculation of PROMISE SCORE $\mathcal{S}$ that accounted for intramolecular interactions. The formulation of $\mathcal{S}_{intra}$ is consistent with protein optimization— we want to identify sequences that enhance an already existing feature of a baseline sequence. So while we have shown the effectiveness of computing particular PROMISE SCORES for Nb discovery and protein optimization, the more general point is that we have shown an ability to formulate scoring functions using the attention map of a Transformer Protein Language Model that are tailored to specific protein engineering objectives. We believe that our work can serve as a baseline for different protein engineering objectives that have different constraints and assumptions.

While we have demonstrated how the PROMISE SCORE may be used in isolation to prioritize protein sequences during Nb discovery and protein optimization, we know that most protein engineering pipelines are multi-facteted. They often include different types of analyses that have somewhat orthogonal yet complementary focuses. These may include genomic assays that identify DNA-protein interactions [60], molecular modeling that predict molecular dynamics and structure [61], or phylogenetic analyses that uncover genomic relationships between related species [62]. We want to emphasize that using PROMISE SCORES is similarly complementary, as they can be used to prioritize sequences at any and all stages of complex protein engineering pipelines.

Finally, we believe it is important to note that PROMISE SCORES can be computed using any Transformer Protein Language Model, not just ESM-1b. While we found ESM-1b to be an effective and convenient model with which to showcase this ability, any model that uses self-attention could be used in its place. Our intent is not to champion one Transformer model over the other, but to demonstrate that this class of model can be used to prioritize protein sequences within the context of protein engineering. We see this as a strength of the approach, as it can be easy to implement using new and improved models and they become available. Indeed, even the author's of ESM-1b have released other versions of Transformer Protein Language Models with differing designs and specifications since we began working with ESM-1b [63, 64, 65]. We leave formal comparisons between Transformer Protein Language Models applied towards this goal to future work.

4.3 ESM-1b as a coarse-grained model for intermolecular interactions

While the PROMISE SCORE is a singular numeric value that reflects the promise of a sequence in its entirety, we showed that we can also easily obtain per-residue scores for s_{binder} or s_{target}. Using these per-residue scores, we show that we can identify specific residues that are known to be involved in intermolecular interactions. More generally, we believe these results demonstrate that ESM-1b can be viewed as a coarse-grained model for intermolecular interactions. With each set of sequences that we used, we observed that the "hits" and "misses" tended to congregate in regions of each protein that have highest concentration of known interacting residues.

In the case of Nb discovery, we used these per-residue scores to identify potential epitopes, and compared these findings to experimentally validated epitope sites. It is important to note that the experimental validation consisted of cross-linking models at sites chosen by computational molecular docking. The validation thus does not provide any evidence for or against potential epitopes beyond the sites that were tested. Additionally, the experimental validation only used 24 Nb sequences, meaning the reported binding percentages are noisy estimates of the true underlying values. Still, across two sets of Nb discovery sequences, the per-residue scores identified sites that overlapped three out of eight total validated epitope regions, as well as identifying many sites outside these regions. We believe that this is an encouraging sign that Protein Language Models like ESM-1b are able to identify specific residues that are expected to contribute to intermolecular interactions. Such an ability could allow protein engineers to better pinpoint the regions in a protein of interest that should be investigated within an protein optimization campaign. Of course, we also made simplifying assumptions in our work that contribute to disagreement with the validation analyses. For one, we only considered interactions involving CDR3, so any effect caused by CDR1 or CDR2 are unaccounted for entirely. Additionally, we arbitrarily used the top 20 percentile sequences to calculate the per-residue scores, where conceivably there is some other optimal threshold. As such, we see these analyses as a jumping off point for future work that identify epitope, or even paratope sequences.

Applied to protein optimization, we investigated not only finding per-residue scores for s_{target}, but also doing so for s_{binder}. For the BRCA1-BARD1 complex, nearly half of the residues that we identified fell within the known binding interface. This strongly suggests that high per-residue scores obtained from $\mathcal{A}_{inter}$ do indeed identify residues involved in intermolecular interactions. In the case of Spike-ACE2, we identified more than half of the known interacting ACE2 residues, though only two out of the 17 known interacting Spike residues. This highlights an important distinction behind how we compute scores for each s_{target} (ACE2) and s_{binder} (Spike). For each sequence used, s_{target} is unchanged, whereas each s_{binder} differs from all others at exactly one residue (due to each Spike sequence having been obtained from a single site mutation scan). Thus, the distribution of high scores throughout the Spike sequence indicates that the model believes different mutations affect the residues that interact at the binding interface. This capability could be used to generate hypotheses about how mutations affect binding patterns between a protein and its partner. Still, that the vast majority of known ACE2 interacting residues were identified indicates that the Transformer accurately identifies interacting residues within the Spike-ACE2 complex.

4.4 Model learning within the context of protein engineering

Having a predictive model for the targeted property in a protein engineering campaign could serve as an invaluable asset, as it can serve as a tool that helps with sequence design decisions. For this reason, we investigated how well we could learn such models using ESM-1b. First, we compared fine-tuning ESM-1b to using the sequence representations from the pre-trained model, and found that fine-tuning worked very well with the Nb discovery setting. With protein optimization, we found it difficult to learn reliable models with or without fine-tuning. We believe this ability to learn reliable models with Nb discovery but not with protein optimization can be explained by differences between these data. As discussed in section 4.2, the Nb discovery data is both more diverse and contains more sequences than the protein optimization data. Diversity and data quantity are well-known factors that influence the ability to train machine learning models [66]. This highlights an important consideration for the experimental design choices within protein engineering campaigns. Our work suggests that single-site mutagenesis experiments across relatively short spans of a protein sequence may not yield informative enough data for accurate model learning. If obtaining a reliable predictive model is imperative or desirable when using single-site mutagenesis, protein engineers may want to consider ways to inject greater diversity into their experiments.

In addition to fine-tuning, we also investigated the viability of knowledge transfer when training models. The ability to use knowledge transfer would allow protein engineers to take what they learn during one engineering campaign and apply it to another. With the Nb discovery data, we found that knowledge transfer was less harmful compared to not using knowledge transfer if using the pre-trained representations rather than a fine-tuned model. This makes sense— a fine-tuned model is adapted to work with a specific s_{binder}-s_{target} pair, so using it on different data would be counter productive. Interestingly, with the protein optimization data, we found that using knowledge transfer yielded models with higher AUC relative to using no knowledge transfer. One thought as to why we may see an improvement in these cases again relates back to the issue of diversity. Perhaps being trained on a set of sequences that were different entirely than those being tested on actually provided a greater degree of diversity that led to slightly better performance. As a final point, we only tried a very basic form of transfer learning in each of these cases, where we use one data set to train a model (eg. Nano-HSA), and make prediction on a separate data set (eg. Nano-GST). It is conceivable that having access to many different data sets to use during training could improve the model's performance.

5 Conclusion

We have shown how to use the deep Transformer Protein Language Model ESM-1b to identify promising sequences during protein engineering campaigns. By jointly encoding a protein sequence s_{binder} and its binding target s_{target}, we use ESM-1b's self-attention mechanism to identify intermolecular and intramolecular interactions between these sequences. We use these interactions to formulate what we refer to as the PROMISE SCORE, and show how this score can be tailored to prioritize protein sequences in two distinct protein engineering domains— protein discovery and protein optimization. With protein discovery, we show how to use PROMISE SCORES to effectively select lead sequences from two separate Nb repertoires. And with protein optimization, we show how to use PROMISE SCORES to identify single-site mutagenesis experiments that successfully identify strong binders. In both cases, we show how to also compute per-residue scores that indicate those expected to undergo intermolecular interactions. We showcase that high scoring residues on Nb target proteins correspond to known epitopes, and those within the BRCA1-BARD1 and Spike-ACE2 complexes correspond to known interacting residues. Finally, we demonstrate that ESM-1b can be fine-tuned to learn accurate models for Nb binding strength, and discuss the limitations of model learning in single-site mutagenesis protein engineering campaigns. We believe this work is highly adaptable and directly applicable to many protein engineering pipelines, so can help to make protein engineering more efficient and effective.

Acknowledgments

We would like to thank AMD for the donation of critical hardware and support resources from its HPC Fund that made this work possible.

References

[1] Paul J Carter. Introduction to current and future protein therapeutics: a protein engineering perspective. *Experimental cell research*, 317(9):1261–1269, 2011.

[2] Mark L Chiu, Dennis R Goulet, Alexey Teplyakov, and Gary L Gilliland. Antibody structure and function: the basis for engineering therapeutics. *Antibodies*, 8(4):55, 2019.

[3] Ruei-Min Lu, Yu-Chyi Hwang, I-Ju Liu, Chi-Chiu Lee, Han-Zen Tsai, Hsin-Jung Li, and Han-Chung Wu. Development of therapeutic antibodies for the treatment of diseases. *Journal of biomedical science*, 27(1):1–30, 2020.

[4] Brian Kelley. Developing therapeutic monoclonal antibodies at pandemic pace. *Nature biotechnology*, 38(5):540–545, 2020.

[5] Shivcharan Prasad and Ipsita Roy. Converting enzymes into tools of industrial importance. *Recent Pat. Biotechnol.*, 12(1):33–56, 2018.

[6] Jing Mu, Liangcan He, Peng Huang, and Xiaoyuan Chen. Engineering of nanoscale coordination polymers with biomolecules for advanced applications. *Coord. Chem. Rev.*, 399(213039):213039, November 2019.

[7] Grazia M Borrelli and Daniela Trono. Recombinant lipases and phospholipases and their use as biocatalysts for industrial applications. *Int. J. Mol. Sci.*, 16(9):20774–20840, September 2015.

[8] Baotong Zhu, Dong Wang, and Na Wei. Enzyme discovery and engineering for sustainable plastic recycling. *Trends in biotechnology*, 40(1):22–37, 2022.

[9] Hyeoncheol Francis Son, In Jin Cho, Seongjoon Joo, Hogyun Seo, Hye-Young Sagong, So Young Choi, Sang Yup Lee, and Kyung-Jin Kim. Rational protein engineering of thermo-stable petase from ideonella sakaiensis for highly efficient pet degradation. *ACS Catalysis*, 9(4):3519–3526, 2019.

[10] George J Weiner. Building better monoclonal antibody-based therapeutics. *Nature Reviews Cancer*, 15(6):361–370, 2015.

[11] Louis M Weiner, Joseph C Murray, and Casey W Shuptrine. Antibody-based immunotherapy of cancer. *Cell*, 148(6):1081–1084, 2012.

[12] Dennis R Goulet and William M Atkins. Considerations for the design of antibody-based therapeutics. *Journal of pharmaceutical sciences*, 109(1):74–103, 2020.

[13] Rahmad Akbar, Philippe A Robert, Milena Pavlović, Jeliazko R Jeliazkov, Igor Snapkov, Andrei Slabodkin, Cédric R Weber, Lonneke Scheffer, Enkelejda Miho, Ingrid Hobæk Haff, Dag Trygve Tryslew Haug, Fridtjof Lund-Johansen, Yana Safonova, Geir K Sandve, and Victor Greiff. A compact vocabulary of paratope-epitope interactions enables predictability of antibody-antigen binding. *Cell Rep.*, 34(11):108856, March 2021.

[14] Inbal Sela-Culang, Vered Kunik, and Yanay Ofran. The structural basis of antibody-antigen recognition. *Frontiers in Immunology*, 4, 2013.

[15] Emily Y. Yang and Khalid Shah. Nanobodies: Next generation of cancer diagnostics and therapeutics. *Frontiers in Oncology*, 10, 2020.

[16] Ivana Jovčevska and Serge Muyldermans. The therapeutic potential of nanobodies. *BioDrugs*, 34(1):11–26, February 2020.

[17] Ivan V Korendovych and William F DeGrado. De novo protein design, a retrospective. *Q. Rev. Biophys.*, 53(e3):e3, February 2020.

[18] Xingjie Pan and Tanja Kortemme. Recent advances in de novo protein design: Principles, methods, and applications. *Journal of Biological Chemistry*, 296:100558, 2021.

[19] Frances H Arnold. Directed evolution: Bringing new chemistry to life. *Angew. Chem. Int. Ed Engl.*, 57(16):4143–4148, April 2018.

[20] Rodrigo MP Siloto and Randall J Weselake. Site saturation mutagenesis: Methods and applications in protein engineering. *Biocatalysis and Agricultural Biotechnology*, 1(3):181–189, 2012.

[21] M M Ling and B H Robinson. Approaches to DNA mutagenesis: an overview. *Anal. Biochem.*, 254(2):157–178, December 1997.

[22] J Braman, C Papworth, and A Greener. Site-directed mutagenesis using double-stranded plasmid DNA templates. *Methods Mol. Biol.*, 57:31–44, 1996.

[23] W Wang and B A Malcolm. Two-stage PCR protocol allowing introduction of multiple mutations, deletions and insertions using QuikChange Site-Directed mutagenesis. *Biotechniques*, 26(4):680–682, April 1999.

[24] Douglas M. Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nature Methods*, 11(8):801–807, Aug 2014.

[25] Leyuan Ma, Jeffrey I. Boucher, Janet Paulsen, Sebastian Matuszewski, Christopher A. Eide, Jianhong Ou, Garrett Eickelberg, Richard D. Press, Lihua Julie Zhu, Brian J. Druker, Susan Branford, Scot A. Wolfe, Jeffrey D. Jensen, Celia A. Schiffer, Michael R. Green, and Daniel N. Bolon. Crispr-cas9–mediated saturated mutagenesis screen predicts clinical drug resistance with improved accuracy. *Proceedings of the National Academy of Sciences*, 114(44):11751–11756, 2017.

[26] Lucas F. Ribeiro, Liliane F. C. Ribeiro, Matheus Q. Barreto, and Richard J. Ward. Protein engineering strategies to expand crispr-cas9 applications. *International Journal of Genomics*, 2018:1652567, Aug 2018.

[27] Miles Congreve, Christopher W. Murray, and Tom L. Blundell. Keynote review: Structural biology and drug discovery. *Drug Discovery Today*, 10(13):895–907, 2005.

[28] 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.

[29] 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, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Applying and improving AlphaFold at CASP14. *Proteins*, 89(12):1711–1721, December 2021.

[30] Zachary Wu, S. B. Jennifer Kan, Russell D. Lewis, Bruce J. Wittmann, and Frances H. Arnold. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proceedings of the National Academy of Sciences*, 116(18):8852–8858, 2019.

[31] Trevor S. Frisby and Christopher James Langmead. Bayesian optimization with evolutionary and structure-based regularization for directed protein evolution. *Algorithms for Molecular Biology*, 16(1):13, Jul 2021.

[32] Trevor S Frisby, Zhiyun Gong, and Christopher James Langmead. Asynchronous parallel bayesian optimization for AI-driven cloud laboratories. *Bioinformatics*, 37(Suppl_1):i451–i459, July 2021.

[33] Charles Yanofsky, Virginia Horn, and Deanna Thorpe. Protein structure relationships revealed by mutational analysis. *Science*, 146(3651):1593–1594, 1964.

[34] D. Altschuh, T. Vernet, P. Berti, D. Moras, and K. Nagai. Coordinated amino acid changes in homologous protein families*. *Protein Engineering, Design and Selection*, 2(3):193–199, 09 1988.

[35] Gregory M Cooper and Christopher D Brown. Qualifying the relationship between sequence conservation and molecular function. *Genome Res.*, 18(2):201–205, February 2008.

[36] Sivaraman Balakrishnan, Hetunandan Kamisetty, Jaime G. Carbonell, Su-In Lee, and Christopher James Langmead. Learning generative models for protein fold families. *Proteins: Structure, Function, and Bioinformatics*, 79(4):1061–1078, 2011.

[37] The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, 11 2016.

[38] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[39] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[40] Anthony Gillioz, Jacky Casas, Elena Mugellini, and Omar Abou Khaled. Overview of the transformer-based models for nlp tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pages 179–183, 2020.

[41] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv*, 2019.

[42] 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*, 118(15):e2016239118, 2021.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz 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, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[44] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[45] Yen-Chun Chen, Zhe Gan, Yu Cheng, Jingzhou Liu, and Jingjing Liu. Distilling the knowledge of BERT for text generation. *CoRR*, abs/1911.03829, 2019.

[46] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[47] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S Song. Evaluating protein transfer learning with TAPE. *Adv. Neural Inf. Process. Syst.*, 32:9689–9701, December 2019.

[48] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[49] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2017.

[50] Yufei Xiang, Zhe Sang, Lirane Bitton, Jianquan Xu, Yang Liu, Dina Schneidman-Duhovny, and Yi Shi. Integrative proteomics identifies thousands of distinct, multi-epitope, and high-affinity nanobodies. *Cell Systems*, 12(3):220–234.e9, 2021.

[51] L. M. Starita, D. L. Young, M. Islam, J. O. Kitzman, J. Gullingsrud, R. J. Hause, D. M. Fowler, J. D. Parvin, J. Shendure, and S. Fields. Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics*, 200(2):413–422, Jun 2015.

[52] Tyler N. Starr, Allison J. Greaney, Sarah K. Hilton, Daniel Ellis, Katharine H.D. Crawford, Adam S. Dingens, Mary Jane Navarro, John E. Bowen, M. Alejandra Tortorici, Alexandra C. Walls, Neil P. King, David Veesler, and Jesse D. Bloom. Deep mutational scanning of sars-cov-2 receptor binding domain reveals constraints on folding and ace2 binding. *Cell*, 182(5):1295–1310.e20, Sep 2020.

[53] Bengt Bjellqvist, Graham J. Hughes, Christian Pasquali, Nicole Paquet, Florence Ravier, Jean-Charles Sanchez, Séverine Frutiger, and Denis Hochstrasser. The focusing positions of polypeptides in immobilized ph gradients can be predicted from their amino acid sequences. *ELECTROPHORESIS*, 14(1):1023–1031, 1993.

[54] Bengt Bjellqvist, Bodil Basse, Eydfinnur Olsen, and Julio E. Celis. Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a ph scale where isoelectric points correlate with polypeptide compositions. *ELECTROPHORESIS*, 15(1):529–539, 1994.

[55] Zhenwei Zhong, Yue Yang, Xiaorui Chen, Zhen Han, Jincai Zhou, Bohua Li, and Xiaowen He. Positive charge in the complementarity-determining regions of synthetic nanobody prevents aggregation. *Biochemical and Biophysical Research Communications*, 572:1–6, 2021.

[56] Mario S Valdés-Tresanco, Andrea Molina-Zapata, Alaín González Pose, and Ernesto Moreno. Structural insights into the design of synthetic nanobody libraries. *Molecules*, 27(7):2198, March 2022.

[57] Mauno Vihinen, Esa Torkkila, and Pentti Riikonen. Accuracy of protein flexibility predictions. *Proteins: Structure, Function, and Bioinformatics*, 19(2):141–149, 1994.

[58] Peter S. Brzovic, Ponni Rajagopal, David W. Hoyt, Mary-Claire King, and Rachel E. Klevit. Structure of a brca1–bard1 heterodimeric ring–ring complex. *Nature Structural Biology*, 8(10):833–837, Oct 2001.

[59] Jun Lan, Jiwan Ge, Jinfang Yu, Sisi Shan, Huan Zhou, Shilong Fan, Qi Zhang, Xuanling Shi, Qisheng Wang, Linqi Zhang, and Xinquan Wang. Structure of the sars-cov-2 spike receptor-binding domain bound to the ace2 receptor. *Nature*, 581(7807):215–220, May 2020.

[60] Trong Nguyen-Duc, Eveline Peeters, Serge Muyldermans, Daniel Charlier, and Gholamreza Hassanzadeh-Ghassabeh. Nanobody(R)-based chromatin immunoprecipitation/micro-array analysis for genome-wide identification of transcription factor DNA binding sites. *Nucleic Acids Res.*, 41(5):e59, March 2013.

[61] Tomer Cohen, Matan Halfon, and Dina Schneidman-Duhovny. Nanonet: Rapid and accurate end-to-end nanobody modeling by deep learning. *Frontiers in Immunology*, 13, 2022.

[62] Alex Klarenbeek, Khalil El Mazouari, Aline Desmyter, Christophe Blanchetot, Anna Hultberg, Natalie de Jonge, Rob C Roovers, Christian Cambillau, Sylvia Spinelli, Jurgen Del-Favero, Theo Verrips, Hans J de Haard, and Ikbel Achour. Camelid ig V genes reveal significant human homology not seen in therapeutic target genes, providing for a powerful therapeutic antibody platform. *MAbs*, 7(4):693–706, 2015.

[63] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29287–29303. Curran Associates, Inc., 2021.

[64] 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. *bioRxiv*, 2022.

[65] Chloe Hsu, Robert Verkuil, Jason Liu, Zeming Lin, Brian Hie, Tom Sercu, Adam Lerer, and Alexander Rives. Learning inverse folding from millions of predicted structures. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8946–8970. PMLR, 17–23 Jul 2022.

[66] Zhiqiang Gong, Ping Zhong, and Weidong Hu. Diversity in machine learning. *IEEE Access*, 7:64323–64350, 2019.