Regularized indirect learning improves phage display ligand discovery

Joseph S. Brown,[1†] Yitong Tseo,[2†] Michael A. Lee,[1] Jeffrey Y.-K. Wong,[1] Soojung Yang,[2] Yehlin Cho,[1,2] Chae Rin Kim,[1] Andrei Loas,[1] Ratmir Derda,[3] Rafael Gomez-Bombarelli,[2*] Bradley L. Pentelute[1,4-6*]

[1] Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States
[2] Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States
[3] Department of Chemistry, University of Alberta, Edmonton, AB T6G 2G2, Canada
[4] The Koch Institute for Integrative Cancer Research, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, United States
[5] Center for Environmental Health Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States
[6] Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, United States
[†]Denotes equal contribution
*Email: rafagb@mit.edu, blp@mit.edu

Abstract

Phage display is commonly employed for the discovery of high affinity ligands to biomolecular targets. However, ranking the discovered ligands for their affinity and specificity to the target is obscured by genetic amplification bias and amplification of target-unrelated phage, resulting in inefficient experimental validation and potentially intractable discovery. Here, we describe the use of indirect machine learning (ML) to improve the efficient discovery of target-specific peptide ligands from next-generation sequencing (NGS) data. We combine peptide sequence information (input) with experimental fitness scores (output) of the individual peptide performance across the rounds of bio-panning in a bidirectional long short-term memory (BiLSTM) architecture. Because the fitness scores contain bias, we use regularization to facilitate limited indirect learning and effectively process the peptide sequence information, while still using the predicted fitness scores to rank the peptides. Peptides containing high-affinity binding motifs to our target were ranked by the regularized model more than threefold higher, compared to any combination of experimental fitness scores. Baseline models of random forest (RF) and -nearest neighbor (KNN) demonstrated slightly lower performance but also demonstrated the importance of regularization. However, the BiLSTM model emerged as the most robust, as it was less sensitive to the peptide representation and the specific fitness score used. Shapley residue analysis generated interpretable structure-activity-relationship (SAR) by providing insight into predicted affinity-driving residues and physicochemical properties across the entire peptide and as well as at motif-specific positions. We expect that this approach will elucidate high-affinity ligands against a multitude of targets, vastly improving the discovery capability of phage display.

Introduction

Phage display is a robust method to perform genetically-encoded peptide or protein ligand discovery against biomolecular targets.[1–7] Peptides and peptidomimetics are a growing therapeutic modality that i) provide sufficient surface area to bind protein-protein interfaces, ii) have recently experienced higher clinical trial success rates than small molecules, and iii) remain less expensive to produce than protein biologics.[8–12] Phage display is accessible, inexpensive, and can serve as a first-line approach to perform peptide ligand discovery from several commercially-available libraries. Furthermore, engineered phage libraries have enabled macrocyclization and covalent pharmacophore modification to improve discovery outcomes and pharmacokinetic properties. [4–7,13–15] Due to these advantages, phage display led to the discovery of clinically investigated peptidomimetics as targeted chemotherapeutic conjugates.[15–18] Phage display has found broad utility in a variety of contexts including in vivo,[2] ex vivo, and in vitro (on-cell) bio-panning.[3] The robust use of phage display has generally been enabled by the phage capsule's protection of the genetic material and by the rise of sensitive next-generation sequencing (NGS). The sensitivity achieved by NGS has enabled several other ligand discovery technologies including DNA-encoded libraries,[19,20] mRNA display,[14,21] and yeast display.[22,23]

However, phage display ligand discovery faces several fundamental challenges that obscure the ranking of the discovered ligands by their affinity and specificity to the target.[24–26] Traditionally, three to five rounds of bio-panning are used to enrich target-specific high-affinity peptides, each of which can introduce bias from the bacterial-based amplification.[1] Specifically, during amplification, target-unrelated phage (TUP) variants can propagate due to their mutations that confer a growth advantage in biological amplification in host bacteria.[24–26] The isolation of target-specific phage can be especially challenging if the biomolecular target is not well suited to drive the affinity selection (i.e., due to its disorder or allostery)[27] resulting in the predominant isolation of target-unrelated phage variants. In addition, the efficiency of bacterial amplification can be varied with high-affinity candidates being disfavored, resulting in low representation.[26] Due to the high sensitivity of the NGS, these biases obfuscate the ranking of peptides for their affinity and specificity toward the target. Additionally, with up to a million peptides revealed per sample, the peptides from NGS data must be ranked in some manner for feasible experimental validation while handling these biases.

Several bioinformatic scores attempt to overcome these challenges by assessing the fitness performance of peptides in the bio-panning process. The fitness performance of each peptide can be understood by their observed: (i) protein selectivity, (ii) enrichment, and (iii) similarity to other identified peptide ligands (Figure 1B). First, peptide enrichment through bio-panning rounds can be quantified by enrichment ratio (ER).[19,26] However, enrichment does not include any measure of protein specificity and thus relies on the experiment design and/or additional analysis to remove nonspecific peptides from the ranking. Second, specificity can be quantified by the comparison of the intensity of the peptide across the bio-panned targets. This specificity can be represented as a pairwise fold change (FC) similar to other bioinformatic analysis. Along with its calculated statistical confidence (p-value), a volcano plot describes the confidence in peptide target specificity.[4,5,28] However, utilizing FC with its p-value alone may fail to identify high-affinity peptides that do not exhibit rapid increases in enrichment due to any detrimental amplification bias. Moreover, identical aliquots of the phage library must be used to start all bio-panning; otherwise, the fold change comparison between slightly different subsets

would inherently lead to the false positive identification.[26] Third, clustering analysis works by grouping the peptides by their chemical similarity with the goal to reveal frequently-appearing motifs that may have driven high-affinity binding in the bio-panning. However, determining the optimal number of clusters and measuring peptide similarity lack an objective approach for optimization, and may lead to over- or underfitting.[29] Many clustering methods also make assumptions about the distribution of the data including *K*-means clustering, which assumes spherically-shaped datasets. The high dimensionality of the input data can lead to issues with local versus global feature relevance.[30,31] Lastly, the presence of nonspecific and parasitic peptides within the input data can further obfuscate clustering efforts.

Machine learning (ML) is set to facilitate a paradigm shift in drug discovery and development for its ability to reveal underlying or nonobvious patterns beyond statistical analysis. Thus, it has been deployed for the discovery of antimicrobial, cell-penetrant, or immunogenic peptides.[32–34] For phage display, ML has improved discovery outcomes by being trained directly on the NGS data,[35] on curated sequences for classification,[28] or on fitness scores (e.g., FC, ER).[36] However, approaches using only one type of data (e.g., ER) may still be affected by the isolation of target-unrelated phage and amplification bias. In contrast, the combination of all fitness scores and sequence-based information may provide a more complete data set to elucidate high-affinity peptides from phage display. However, to our knowledge, a combined model has not been developed.

Here, we describe the use of indirect learning to accelerate the discovery of high-affinity peptide ligands from phage display bio-panning (Figure 1C). Specifically, our model ranks peptides for their likelihood to be high-affinity, target-specific ligands by learning underlying connections between the experimental fitness scores from input peptide sequences. Because they contain bias, the experimental fitness scores serve as a proxy for the desired ranking based on affinity and specificity of the peptide to the target. We use a bidirectional long short-term memory (BiLSTM) model to parse peptide sequence information. To evaluate our model, we used a calculated “hit rate” for the ranking of peptide “hits,” which contain known target-specific motifs, above non-motif containing peptides. Strict regularization was found to be important to highly rank peptide hits from the NGS data and resulted in over three-fold higher ranking of peptide hits by the BiLSTM versus only using experimental fitness scores. We additionally baseline the BiLSTM model against other appropriate models for this task including random forest (RF) and K-nearest neighbors (KNN), which demonstrated slightly lower performance but reinforced the importance of regularization. Furthermore, we examine the structure-activity relationship (SAR) and investigate the peptide motifs using Shapley additive analysis. To our knowledge, this is the first work to combine both FC and ER fitness scores as proxies to indirectly learn the unbiased ranking of peptides for their target affinity and specificity from genetically-encoded affinity selection. From this framework, future work will evaluate the binding affinity of prioritized peptides across a wider range of biomolecular targets to assess generality of the approach.

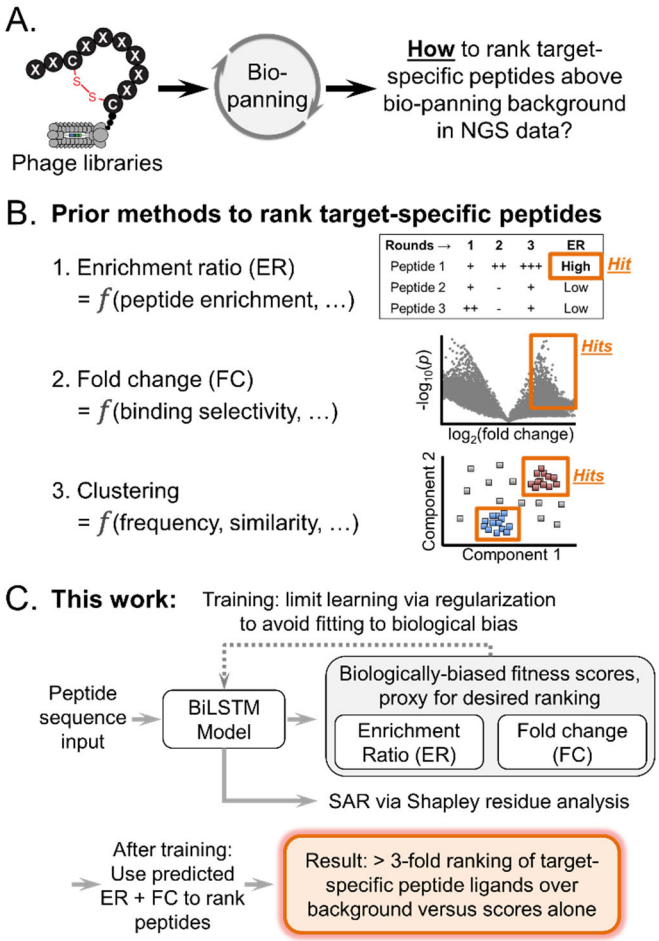


Figure 1. This work uses biologically-biased fitness scores from phage bio-panning as a proxy objective for indirect supervised machine learning model for to rank peptides for their affinity and specificity to the target above background. **A.** A significant challenge of *de novo* discovery with phage display is ranking target-specific peptide binders above the bio-panning background to improve research efficiency and maximize discovery success. **B.** Fitness scores can rank peptide sequences from next-generation sequencing (NGS) data for experimental synthesis and validation, including enrichment ratio (ER), which quantifies the round-to-round change in individual peptide enrichment, and fold change (FC), which quantifies the peptide-protein selectivity. Clustering analysis has also been performed to parse and group the isolated peptides based on chemical similarity. However, each are susceptible to detrimental biological amplification bias. **C.** This work uses the FC + ER fitness scores as a proxy for the desired ranking of peptides by their affinity and specificity from NGS data. Thus, the training of a bidirectional long short-term memory (BiLSTM) model on peptide sequence information is limited via regularization to avoid fitting to the biological bias. The resulting BiLSTM model provides a >3-fold increase in accuracy of ranking peptide hits (motif-containing peptides) from the NGS background, termed “hit rate.”

Results and discussion

Peptide fitness scores from phage panning provide partially orthogonal information well-suited for model development.

Multiplexed phage display libraries with linear and macrocyclic peptides were selected against mouse double minute 2 (MDM2) and anti-hemagglutinin antibody (12ca5, Figure 2A, see SI: Section 2-8). Briefly, three rounds of phage display panning were completed using an automated protocol, with the protein target pre-immobilized on magnetic beads. Bio-panning was completed in multiplex format by mixing linear peptide libraries (X_{12} and X_7) and macrocyclic peptide libraries ACX_7C and AX_MCX_NC (M+N=6) all together for panning. The mixture of all libraries was incubated first with unlabeled magnetic beads to remove high-affinity bead binders. Nonspecific binding was blocked with 2% non-fat milk for 1 hour before incubation of the protein target with depleted pooled phage library together for 1 hour. Overall, these measures to limit nonspecific binding were successful, with only 0.06% of peptide sequences containing the off-target streptavidin binding motif HPQ (SI Section 5.1.1).

Within this work, we evaluated the accuracy of our approach to rank high-affinity peptides above background by whether peptides contain known target-specific motifs. Specifically, the known 12ca5 binding motif is D**DY(A/S)[37–39] and the known MDM2 binding motif from prior phage display work is F**ΦΦ, where Φ are the hydrophobic amino acids phenylalanine, tryptophan, leucine, isoleucine, valine and tyrosine.[40–44] As 12ca5 is a peptide-binding antibody, we placed more focus on validating our methodology to isolate motif-containing sequences against MDM2, with 12ca5 serving as a control.

In our data, FC and ER fitness scores appeared partially orthogonal, though neither score readily ranked peptides “hits,” defined to have the known binding motif, above background sequences likely due to biases in bio-panning (Figure 2). Both FC and ER scores are driven by the affinity selection process in bio-panning and each has led to the identification of high-affinity binders from genetically encoded libraries.[4,5,19,26] However, both fitness scores may be susceptible to amplification bias and isolation of target-unrelated phage. Common to all affinity selections, a weaker signal may be observed if the biomolecular target does not strongly drive the selection through affinity-ranked interactions. In comparison to 12ca5, MDM2 appeared less able to drive a strong affinity selection as seen in the volcano plot of −ln(*p*-value) vs ln(FC) (Figure 2B). The peptide hits colored in orange would be poorly isolated from an FC-based approach. Relating FC to ER, we see that additional information emerges (Figure 2C) with the appearance of three peptide groups. The peptide hits appear predominantly in the high FC and high-to-modest ER region, as expected. However, only 2.8% of the peptides in this region (high FC (> 2.5)[4,5] and high-to-modest ER (> 0)) contain the desired, high-affinity MDM2 motif. Another high FC population demonstrated lower ER, suggestive of target-selectivity but weak enrichment. However, this population contains few motif-containing peptide hits. Lastly, there was a population that demonstrated low FC and ER likely from the off-target control. Combining these together, the clearest localization of the desired motif-containing peptides can be seen in a three-dimensional overlay (Figure 2D), where most of the hits appear to have a high FC and ER and a modest *p*-value. Overall, these data suggest that the combination of FC and ER may benefit any analysis approach toward the identification of specific high-affinity peptides from phage display data (Figure 2B,C,D).

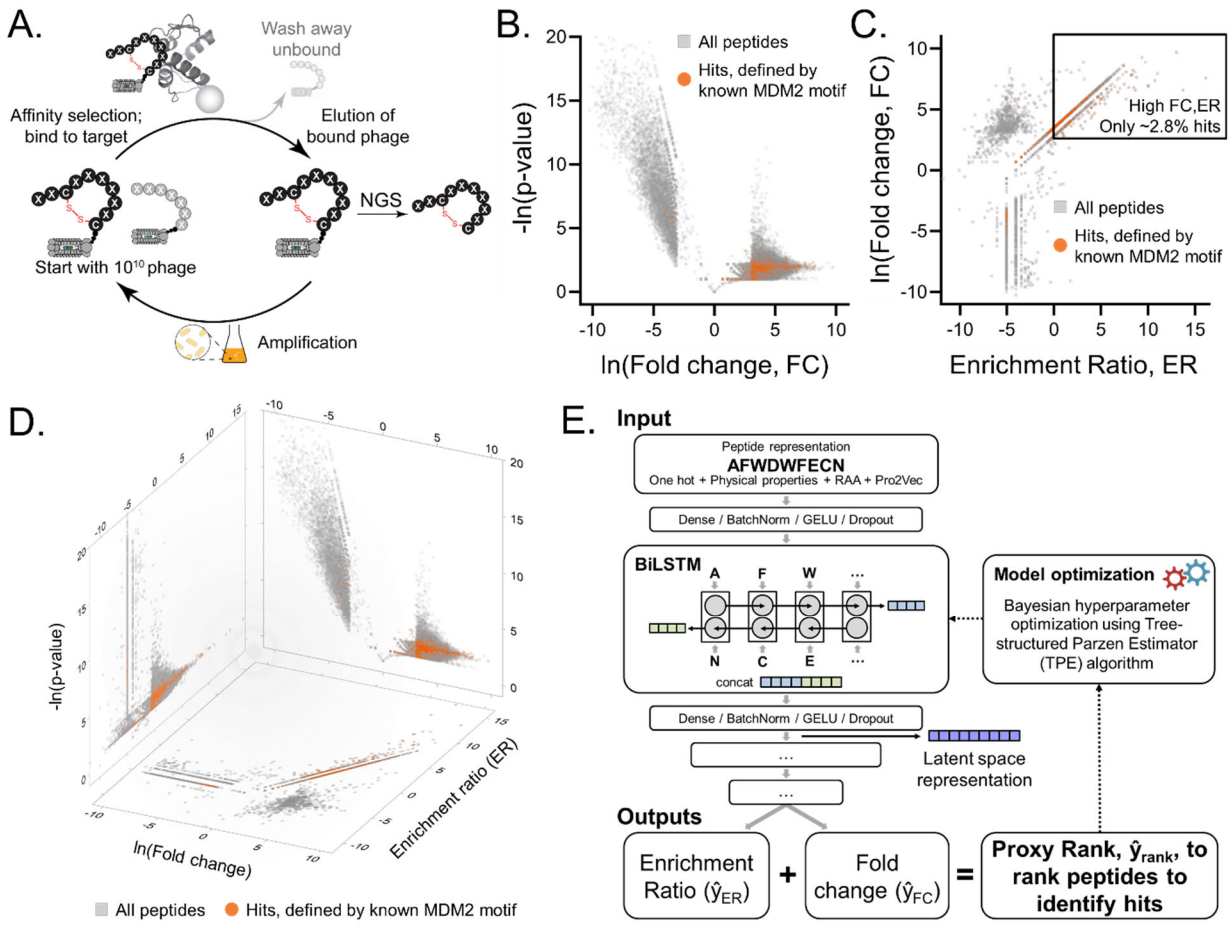


Figure 2. Experimental fitness scores of fold change (FC) and enrichment ratio (ER) provided partially orthogonal information, which when combined may improve the ranking of target-specific peptide ligands from phage display. **A.** Phage display bio-panning isolated peptide ligands by iterative affinity selection and inherently biased biological amplification of bound phage. **B.** Volcano plot of −ln(*p*-value) vs ln(FC) demonstrated most peptide hits (orange) show modest FC and *p*-value. **C.** The combination of FC and ER partially reveals motif-containing peptide hits as a population described by both high FC and high ER. However, only 2.8% of the peptide sequences within the high FC (>2.5) and high-to-modest ER (>0) region are hits and contain the high-affinity MDM2 motif. **D.** The projection of the data in 3-dimensions visualized all three criteria (ER, FC and its associated p-value) aids in identifying the location of the peptide hits. Orange points in plots **B** through **D** represent peptides that contain the MDM2 motif. **E**. Our approach utilized a bi-directional long short-term memory (BiLSTM) model to learn on FC and ER as a proxy objective to output the desired high ranking of target-specific peptide hits over the NGS background.

Clustering of peptides from multiplex phage panning did not clearly identify hit peptides.

Before employing a more powerful model, we first determined that two common clustering methods, *k*-means clustering[30,45] and Cluster Database at High Identity with Tolerance (CD-HIT),[29] were insufficient to identify groups of peptide hits distinct from background in the MDM2 phage display data. For *k*-means clustering, the peptides were encoded by amino acid, each

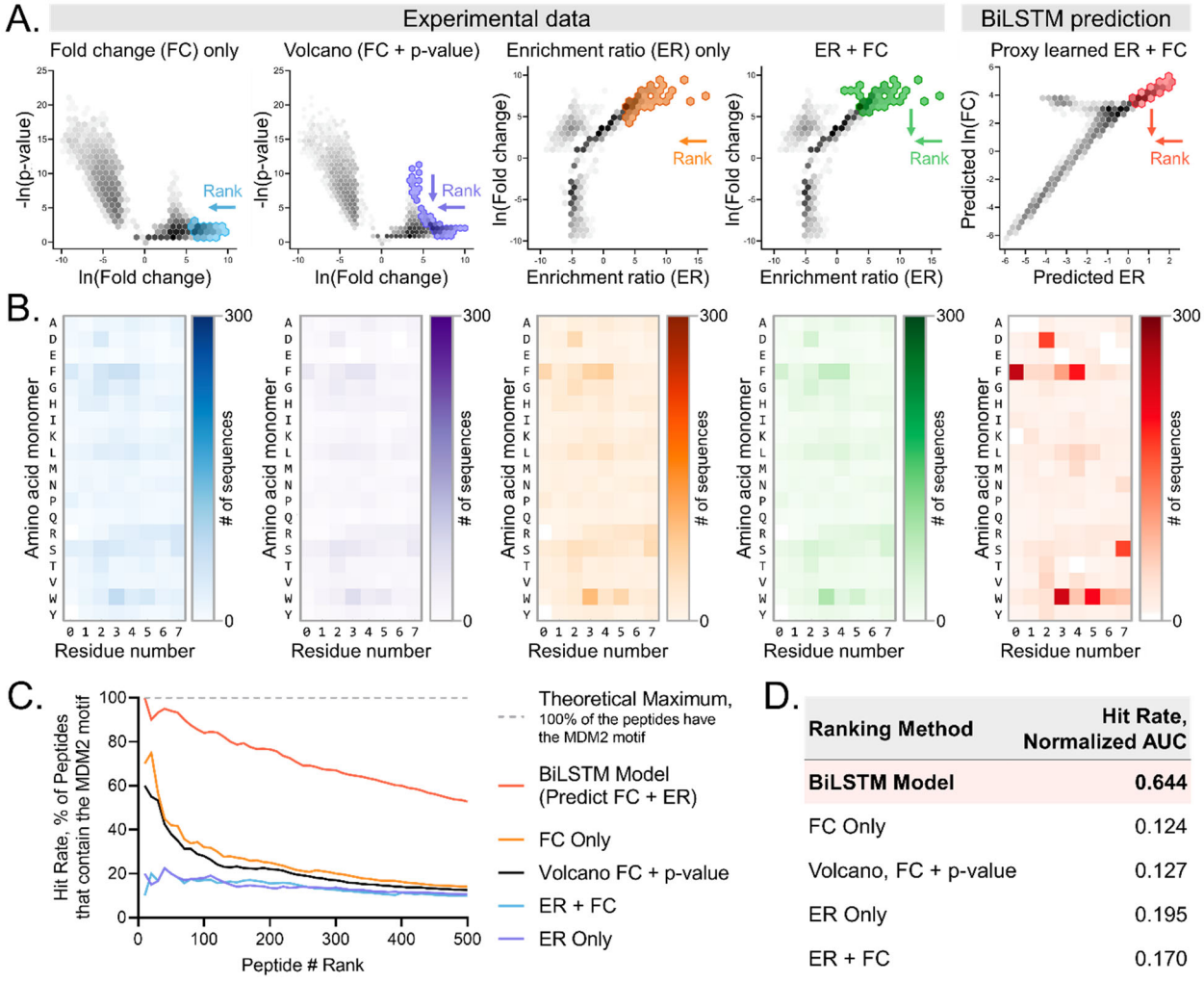
represented as a 36-length vector from one-hot encoding, relative propensity for binding score,[46] DELPHI predicted protein interaction score,[47] and 14 physicochemical property descriptors.[46,48] Residue-based encoding directly improves the ability to perform SAR analysis, ranging from amino acid-specific (one-hot) to generalizable (physical property) contributions toward binding affinity. Encoded peptides were decomposed using dimensionality reduction with Uniform Manifold Approximation and Projection (UMAP)[49] and principal component analysis (PCA). The data was then clustered using the *k*-means algorithm after optimization of the number of clusters *k* using the elbow method, and a logo plot was generated for each cluster (Figure S3A). Additionally, we calculated the ER or FC of each cluster in an attempt to guide the determination of target-selective clusters (Figure S3D).

The clustering primarily produced separate clusters for each library utilized (i.e., separating linear and macrocyclic libraries) due to the lack of sequence alignment and afforded no meaningful information about potential peptide hits or target-selective clusters within each library. Only a single cluster containing the 12ca5-based aspartic acid motif could be identified, with no clusters containing the MDM2 motif in any form. This result was made evident when the location of 12ca5- and MDM2-motifs were overlaid on the clusters revealing dispersion, indicating that our clustering approach did not isolate any motif-containing hit peptides. Clustering using CD-HIT was also attempted across a range of similarity metrics. CD-HIT uses greedy incremental clustering to estimate sequence similarity first without an alignment at similarity threshold, and then with sequence alignment if the similarity falls below the threshold.[29] However, even with alignment its alignment, CD-HIT was unable to identify any cluster of peptides larger than three members in the phage display data. Overall, our efforts indicated that isolation of motif-containing peptides can be challenging with clustering, warranting the use of more powerful tools to combine the partially orthogonal information from the peptide sequence, ER, FC and its associated p-value.

Regularized learning on fitness scores as a proxy objective improves the ranking of peptide hits over background.

We proposed regularized supervised learning to perform indirect learning on biologically-biased fitness scores of enrichment (ER) and target selectivity (FC, *p*-value) as a proxy objective to reveal an improved ranking of peptide hits above background sequences. Peptides that exhibit high FC and ER have a higher fraction of hits (defined by their target-specific motif). However, only 2.8% of these high FC and high ER peptides were observed to be hits to MDM2 (Figure 2). We hypothesized that the biological bias that contributes to the NGS background and obscures the straightforward ranking of the peptide hits by FC and ER was similar to other types of noise in experimental labels. Thus, the use of regularization to limit the model from fitting to data could reveal a broader underlying pattern and potentially enable the improved ranking of peptide hits from the NGS data. In this way, we sought to perform indirect learning to highly rank peptide hits by using regularized supervised learning from encoded peptide sequence inputs. The resulting predicted outputs of ER ($\hat{y}_{ER}$) and FC ($\hat{y}_{FC}$) could then be summed together to provide a single ranking ($\hat{y}_{rank}$, Figure 2E) and evaluated by whether they contain the MDM2-binding motif. From the input peptide sequence, we employed a BiLSTM model to predict the $\hat{y}_{rank}$ ranking (Figure 2E).[50] The BiLSTM architecture was chosen for its capacity to preserve sequence order, represent peptide libraries of multiple lengths, and handle cases of motif frame shift and macrocycle bidirectionality.

In the BiLSTM model, the greatest training performance to identify and efficiently rank MDM2 and 12ca5 peptide hits was achieved using multiple types of regularization. Regularization limits overfitting during training. We hypothesize regularization limits the model to learning broader underlying sequence patterns that drive binding, rather than nuances between the performance of individual peptides, which may be more strongly affected by biological bias. The importance of regularization was revealed during analysis of the effects of hyperparameter optimization on the ranking of peptide hits using Bayesian hyperparameter optimization with the Tree-Structured Parzen Estimator algorithm[51] (Figure S4). Specifically, high dropout ($\lambda_{Dropout}$ = 0.5), substantial weight decay penalty (λ_{L2} = 0.01), low learning rate (α = 0.0005), shallow model depth (depth = 7), narrow model width (width = 64), and substantial batch sizes (n = 128) were found to improve the hit rate of discovery (Figure S4 for further detail). Each of these parameters affect the model's learning, likely only forming broad connections between the peptide features with experimental FC and ER during the training process. This stringent regularization improved the model's ability to highly rank peptide hits above background at the cost of reducing prediction accuracy (Figure S5), following the classic variance-bias tradeoff.[52]



Ranking Method	Hit Rate, Normalized AUC
BiLSTM Model	**0.644**
FC Only	0.124
Volcano, FC + p-value	0.127
ER Only	0.195
ER + FC	0.170

Figure 3. BiLSTM model efficiently ranks MDM2 motif-containing peptide hits above NGS background >3-fold better than any combination of experimental fitness scores. Ranking peptides is required due to the large number of potential peptide ligands identified to prioritize the investment of synthesis and experimental binding validation. **A.** Hexbin projections with highlighted zones corresponding to the top 500 peptides as determined by the different strategies to rank the peptides as potential peptide hits. Arrows shown in the bottom right display the direction of ranking (x-direction, y-direction, or both). **B.** Positional frequency matrix of the top 500 identified peptides. The macrocyclic 9-mer AX_NCX_MC (N + M = 6) library contained most of the peptide hits, and outperformed the other libraries. Thus, the positional frequency matrices of the top 500 show the 9-mer variable region of the 9-mer library (cysteine not shown on y-axis for clarity). **C.** The BiLSTM model outperformed all other experimental methods to efficiently rank hit peptides (motif-containing) above background. This result is shown by the hit rate, which is defined as the number of motif-containing peptide divided by their rank shown as a percentage. Only 527 peptides contained the MDM2 motif, and thus, the top 500 peptides ranked by each method was examined. For example, if the top 100 peptides were synthesized, 84% of the BiLSTM ranked peptides are expected to be hits as they contain the MDM2-binding motif, versus ~30% from other methods. Considering the area under the curve up to 500 peptides, the BiLSTM performs >3-fold better at highly ranking the motif-containing peptides. **D.** Calculation of the area under the hit rate curve in C indicates that 64% of the top 500 BiLSTM ranked peptides contain the MDM2 motif.

The regularized BiLSTM model highly ranked motif-containing peptide (hits) over background.

Our BiLSTM model highly ranked peptides containing F**ΦΦ motif known to drive high-affinity peptide binding to MDM2[40–44] above background (Figure 3A,B). Additional confidence can be placed in ligand discovery when compounds containing a similar set of critical features or residues (i.e., a "motif") are enriched, forming the base utility of clustering.[8,9,53–56] In the context of affinity selection, these motifs are generally assumed to facilitate high-affinity interactions. Within our NGS data, 527 peptides contain the MDM2 motif. Thus, we sought to compare the number of peptides that contain the MDM2 motifs within the top 500 ranked peptides from the 10-fold cross-validated model $\hat{y}_{rank}$ versus combinations of the experimental fitness scores including FC-only, FC with its associated p-value (volcano plot),[4,5,28] ER-only, and FC+ER (Figure 3A). We assessed their motif pattern by using a positional frequency matrix seen in Figure 3B. Only the BiLSTM model showed a clear motif pattern closely matching the F**ΦΦ motif known to drive high-affinity peptide binding to MDM2.[40–44] The other approaches showed no clear discernable pattern and appeared random.

Similarly, the BiLSTM ranked peptide hits from the NGS data with an >3-fold higher hit rate than the experimental fitness scores it indirectly learns from, concentrating the likelihood of success for initial synthesis validation attempts (Figure 3C,D). The ranking the peptides from bio-panning NGS data is required to efficiently prioritize the investment of experimental validation toward peptides with the highest likelihood of being high-affinity hits. Similar to the motif identification, the BiLSTM rank ($\hat{y}_{rank}$) from the 10-fold cross-validated model was assessed for its efficiency to highly rank peptide hits in the top 500 peptides from the NGS data. Specifically, we determined the percentage of peptides that contain the MDM2 motifs as a function of their rank within the top 500 ranked peptides as a "hit rate" (Figure 3C). A theoretically perfect ranking would

rank all peptides hits above background (gray dashed line in Figure 3C). Thus, per peptide invested in experimental validation, the hit rate assesses how efficiently hits are identified. Overall, the BiLSTM model hit rate significantly outperforms all fitness score-based approaches by >3-fold, with ER only providing the next best performance.

Figure 4. Baseline of other indirect learning models across peptide representation and proxy objective (ER, FC, or ER+FC) of the desired ranking of peptide hits over background. The efficiency of each model was assessed by its hit rate, defined as the number of motif-containing peptide divided by their rank shown here as a fraction. on both the MDM2 and 12ca5 target protein systems. All results reported on the hit rate of a 20% holdout dataset taken before hyperparameter optimization (n = 84 maximum theoretical MDM2 hits, n = 166 maximum theoretical 12ca5 hits).

Other suitable models perform similarly, but the BiLSTM appears the most robust.

Next, we benchmarked the performance of the BiLSTM model against baseline models including random forest (RF) and K-Nearest Neighbor (KNN, which are also suitable for this analysis of NGS-based discovery data using the hit rate.[19,57] Hyperparameter optimization of the RF model prioritized shallow tree depth (max depth of 10) and greater number of estimators (200) presumably to increase regularization by averaging multiple decision trees that individually suffer from high variance, resulting in the elucidation of underlying data patterns that are robust to noise in the dataset. Similarly, during hyperparameter optimization, regularization in KNN models was implicitly controlled by adjusting for large neighbor set sizes (35). Ultimately, all three model architectures achieved comparable performance, with a slight advantage to the BiLSTM. Notably, all three model architectures across all peptide representations and indirect training objectives (ER, FC, or ER+FC) achieved significantly higher hit rates relative to the experimental fitness scores alone. This result further underscored the vital importance of regularization, likely for its role in limiting the model to only learn and rank broader, underlying sequence patterns (i.e., the MDM2 motif) and provide an efficient ranking of peptide hits.

The BiLSTM model demonstrated greater robustness for the type of peptide encoding input (One-Hot, Physicochemical) and proxy objective (ER, FC) used relative to RF and KNN models (Figure 4). In the two distinct protein systems examined, the fitness score used as the proxy objective (ER, FC, and p-value) resulted in different effectiveness to rank peptide hits. Specifically, for MDM2, ER provided the most valuable information as the proxy objective for ranking peptide hits across the model baselines. Whereas for 12ca5, FC and p-value proved to be the most valuable as the proxy objective for ranking peptide hits. Similarly, the RF and KNN showed improved performance for specific encodings (physicochemical for RF and one-hot

encoding for KNN) rather than the combined encoding. However, across both proteins, the BiLSTM showed the best performance with the combined peptide encoding (both One-Hot and physicochemical) and both (FC+ER) proxy objectives. Thus, the application of the BiLSTM is the most streamlined and robust, without the need to consider the optimal proxy objective and/or encoding combinations.

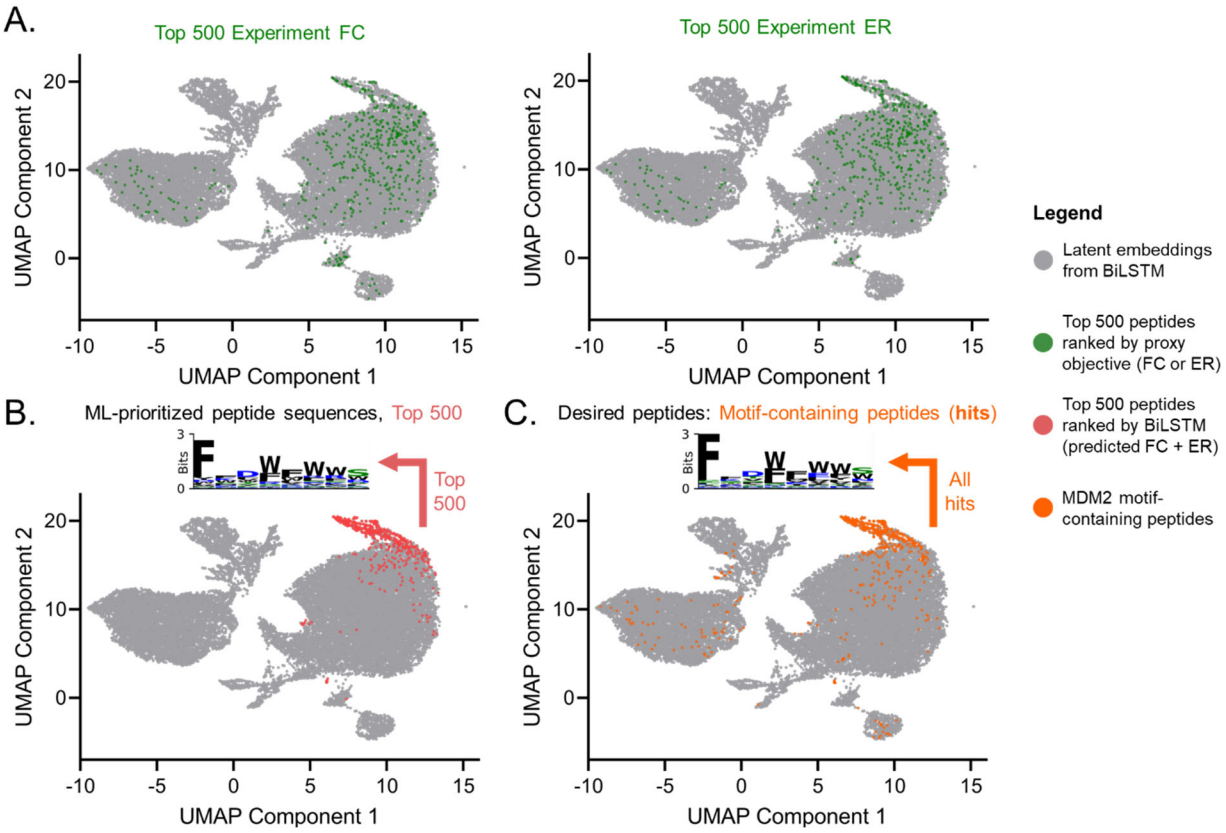


Figure 5. The UMAP decompositions of the BiLSTM learned latent features demonstrated the effect of regularization on proxy objective learning to result in the consolidation of peptide hits. The UMAP transformation was obtained from the penultimate layer of the BiLSTM model with multiple color-coded points overlayed. **A.** The 500 peptides with the top experimental FC and ER peptides highlighted in dark green. **B.** The 500 peptides with the top predicted BiLSTM ranking ($\hat{y}_{rank} = \hat{y}_{ER} + \hat{y}_{FC}$) highlighted in red were observed to be grouped together for MDM2. A logo plot exhibits multiple frameshifts of the hydrophobic motif. **C.** For comparison, all MDM2 motif-containing sequences (F**ΦΦ) are highlighted in orange, indicating nearly the same group of peptides prioritized in the top 500 by the BiLSTM ranking in **B**. Cysteine residues are excluded from the logo plots for clarity of the motif.

The effect of regularized indirect learning of the proxy objectives (FC and ER) was revealed by UMAP of the BiLSTM latent space.

From visualizing the BiLSTM latent space, the effect of regularization or enact indirect learning resulted in the cluster-like consolidation of motif-containing peptides as seen by UMAP (Figure 5). Examining the latent space BiLSTM embeddings can reveal the influence of the learning toward the proxy objective FC and ER versus the model's ability to parse and/or

consolidate similarities of the peptide features. The UMAP decomposition into two dimensions of all latent peptide embeddings from the penultimate layer of the BiLSTM is shown throughout Figure 5 with various color-coded overlays. The peptides with the top 500 predicted ER and FC highlighted to indicate the location of peptides that were highly ranked by the model.

The clear consolidation of highly ER- and FC-ranked peptides indicated that the BiLSTM model placed significant weight toward understanding and grouping key motifs or patterns from the NGS data background. Moreover, the partial overlap of the top 500 predicted and top 500 experimental peptides (Figure 5A vs B) demonstrated the lack of complete model accuracy, likely due to regularization. The evidence for limited learning and accuracy was further confirmed by parity plots (Figure S5 and S7). For MDM2 (Figure 4), we see one region in the UMAP plot that contains motif-containing peptide sequences that strongly overlaps with those predicted to have high ER and FC, which are summed to give the ranking. For 12ca5, the top 500 peptides exhibiting the highest predicted 12ca5 ER and 12ca5 FC scores also exhibit a significant consolidation within the UMAP projection. In addition to successfully identifying 12ca5 motif-containing sequences (*DYA*), the BiLSTM model prioritized a similar set of anionic peptides containing motifs including D**DY*, which is highly similar to the known motif (D**DYA), and LE*E, which has not been reported before. Overall, these findings underscore the effect of regularization to consolidate similar peptides from NGS data across while considering their FC and ER fitness scores.

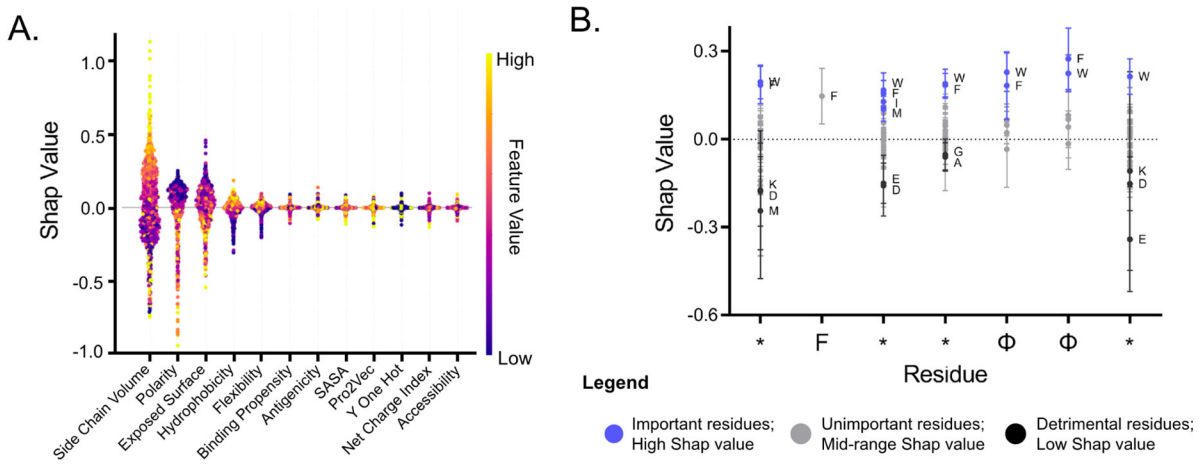


Figure 6. Built-in model interpretability using Shapley analysis provided amino acid and property based SAR. **A.** Shapley feature importance across representation features as calculated by the 10-model ensemble trained via cross validation splitting on a test set of 500 randomly sampled peptides. This result indicated that high volume, low polarity, high hydrophobicity, and high flexibility are predicted to improve MDM2 binding propensity. **B.** Positional Shapley feature importance across residue identities as calculated by the 10-model ensemble on the set of all 558 MDM2 motifs containing hits within the dataset. Sequences are aligned by motif position, and error bars are calculated according to the standard deviation of Shapley values per residue across all peptides and all models. This result underlied the importance of hydrophobic amino acids to drive binding and the potential for small or polar amino acids to disrupt peptide binding to MDM2.

Shapley analysis provided amino-acid and positional structure activity relationships.

In addition to potentiating discovery, we can analyze the BiLSTM model results and prioritizations at the individual amino acid level to infer structure activity relationship (SAR) information using Shapley Additive Explanation analysis.[58] We have demonstrated that the BiLSTM model highly ranks peptides hits. From the ranking, the valuable sequence motifs, down to the specific amino acid level, can be revealed using Shapley Additive Explanation analysis of our 10-fold cross-validated BiLSTM model ensemble. Shapley analysis uses coalition game theory to calculate the contribution of each encoded feature (for this work amino acid and physicochemical property) to the final model prediction.[58] Thus, this analysis identifies the importance of each representation feature (Figure 6A) for its influence in driving MDM2 binding.

For MDM2, high volume, low polarity (P_{12} polarizability), low exposed surface area, and high (H_{12}) hydrophobicity descriptors[46,48] were found to be the most indicative characteristics of residues to drive high-affinity binding (Figure 5A). These parameters match well with the properties of canonical uncharged aromatic amino acids, including tyrosine, tryptophan, and phenylalanine, which are known to be a part of the MDM2 motif that drives high affinity binding. Other features such as low exposure, high flexibility, and median side chain net charge index according to the cross validated model ensemble correlate with the MDM2 binding likelihood. From the same analysis, favorable properties to drive 12ca5 can be inferred as well from the “low” Shapley values or the stand-alone analysis (Figure S9). High polarity, high exposed surface area, low flexibility, and low hydrophobicity were seen to likely drive 12ca5 binding, consistent with the D**DYA motif. For both proteins, the two pretrained descriptors of relative binding propensity and DELPHI protein interaction scores were less connected to peptide binding activity. In addition, one-hot descriptors show relatively low Shapley importance ranges, which suggested that the model ensemble eschewed specific categorical understanding in favor of deeper physicochemical understanding.

For additional interpretability, we summed the Shapley values across the representation dimension to determine positional importance, also referred to as Positional Shapley (PoSHAP) (as illustrated in Figure 6B).[59] Positional Shapley analysis of peptide hits aligned by the theoretical MDM2 motif (Figure 6B) allowed the quantitative comparison of residue importance at different positions. Our findings revealed that uncharged aromatic amino acids had the most influence on the model performance, with the highest contributions according to our proxy metric. Hydrophobe 1 (Φ_1 in F**$\Phi_1\Phi_2$) is often seen to be tyrosine in literature,[40–44] but seen to be highly weighted as tryptophan by our model. The negatively charged residues aspartic and glutamic acid in position 3 or 7 (relative to the start of the motif) were recognized to significantly reduce propensity to bind by our model ensemble in addition to other small or polar amino acids. These polar amino acids likely prefer to be solvated rather than bound to the hydrophobic MDM2 patch surface. For 12ca5 (Figure S9) PoSHAP revealed the higher importance of the aspartic acid residues within the motif (D**DYA) for binding than the tryptophan and alanine residues. Overall, we envisage that the integration of PoSHAP within BiLSTM improves the interpretability of the model and the SAR information gained from affinity selection and bio-panning discovery.

Conclusion

Here, regularized indirect learning on biologically biased fitness scores improved the ranking of high-affinity, target-specific peptide hits from phage bio-panning. Indirectly learning connections from peptide fitness (FC and ER) to peptide features appeared to overcome the biological bias or “noise” from target-unrelated phage and/or weak affinity-driven enrichment. The indirect learning was critically enforced by regularization in the form of high dropout ($\lambda_{Dropout} = 0.5$), high L2 penalty ($\lambda_{L2} = 0.01$), low learning rate ($\alpha = 0.0005$), shallow model depth (depth = 7), narrow model width (width = 64), and substantial batch sizes (n = 128) for the BiLSTM architecture, and led to the cluster-like consolidation of sequence feature information (Figure 5) guided by experimental FC and ER. This regularized approach led to the high ranking of peptides hits to MDM2 at a >3-fold improved hit rate relative any combination of the fitness scores used for training (Figure 3). Compared to baseline models of RF and KNN, the BiLSTM model demonstrated similar or slightly improved hit rate. However, the BiLSTM appeared more robust for its ability to provide good performance from the combine proxy objective (FC and ER) and all amino acid descriptors for both MDM2 and 12ca5 (Figure 4), meaning optimal encoding or proxy objective need not be explored. In comparison, RF and KNN showed variability in their optimal encoding method and target objective (e.g., FC, FC+ER), whereas the BiLSTM model improved with the more diverse information input. Lastly, the addition of Shapley Additive Explanation analysis allows for SAR-level information to be isolated from the ligand discovery experiment directly. From initial discovery experiments, Shapley analysis holds potential to guide the importance of peptide amino acid composition (Figure 6A) as well as with respect to sequence (Figure 6B), informing derivatization efforts.

Next, we will seek to apply this indirect learning approach toward peptide fitness to phage display against novel targets including experimental validation. This future direction will reveal the connection between predicted hit rate against these model protein targets (12ca5 and MDM2) as well as establish a true experimental hit rate against more challenging targets. We expect that this indirect learning approach to learn proxy objectives will generally improve the hit rate and discovery of high-affinity peptide ligands against biomolecular targets, offering a useful computational tool to streamline and enhance the pre-clinical development pipeline of next-generation peptidomimetic therapeutics.

Data availability

Data supporting the findings of this work are available within the Supplementary Information, which contains phage display biopanning, titering, and amplification protocols; details on MDM2 chemical synthesis; clustering of NGS data; BiLSTM hyperparameter optimization; parity plots; and additional sequence analysis of the latent space. All data utilized in this work is available at https://github.com/YitongTseo/ml_phage_display.

Code availability

All the code utilized in this work is available at https://github.com/YitongTseo/ml_phage_display. Tutorial Jupyter notebooks are also in the repository along with all notebooks to generate data necessary for figure preparation.

Acknowledgements

Funding for this work was provided by Novo Nordisk A/S. We thank Dr. Thomas E. Nielsen and Dr. Uli Stilz for their helpful discussions in support of our work. J.S.B. acknowledges support from the Pharmaceutical Research and Manufacturers of America (PhRMA) Foundation through the Postdoctoral Fellowship in Drug Discovery. M.A.L. acknowledges support from the MIT Dean of Science Fellowship.

Competing interests

B.L.P. is a co-founder and/or member of the scientific advisory board of several companies focusing on the development of protein and peptide therapeutics.

References

1. Smith, G. P. & Petrenko, V. A. Phage display. *Chem Rev* **97**, 391–410 (1997).
2. Pasqualini, R. & Ruoslahti, E. Organ targeting in vivo using phage display peptide libraries. *Nature* **380**, 364–366 (1996).
3. Philpott, D. N. *et al.* Rapid On-Cell Selection of High-Performance Human Antibodies. *ACS Cent Sci* **8**, 102–109 (2022).
4. Wong, J. Y. K. *et al.* Genetically-encoded discovery of proteolytically stable bicyclic inhibitors for morphogen NODAL. *Chem Sci* **12**, 9694–9703 (2021).
5. Ekanayake, A. I. *et al.* Genetically Encoded Fragment-Based Discovery from Phage-Displayed Macrocyclic Libraries with Genetically Encoded Unnatural Pharmacophores. *J Am Chem Soc* **143**, 5497–5507 (2021).
6. Oppewal, T. R., Jansen, I. D., Hekelaar, J. & Mayer, C. A Strategy to Select Macrocyclic Peptides Featuring Asymmetric Molecular Scaffolds as Cyclization Units by Phage Display. *J Am Chem Soc* **144**, 3644–3652 (2022).
7. Kong, X. D. *et al.* De novo development of proteolytically resistant therapeutic peptides for oral administration. *Nat Biomed Eng* **4**, 560–571 (2020).
8. Henninot, A., Collins, J. C. & Nuss, J. M. The Current State of Peptide Drug Discovery: Back to the Future? *J Med Chem* **61**, 1382–1414 (2018).
9. Muttenthaler, M., King, G. F., Adams, D. J. & Alewood, P. F. Trends in peptide drug discovery. *Nat Rev Drug Discov* **20**, 309–325 (2021).
10. Wells, J. A. & McClendon, C. L. Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature* **450**, 1001–1009 (2007).
11. Cunningham, A. D., Qvit, N. & Mochly-Rosen, D. Peptides and peptidomimetics as regulators of protein–protein interactions. *Curr Opin Struct Biol* **44**, 59–66 (2017).
12. Lubell, W. D. Peptide-Based Drug Development. *Biomedicines 2022, Vol. 10, Page 2037* **10**, 2037 (2022).
13. Heinis, C. Drug discovery: Tools and rules for macrocycles. *Nat Chem Biol* **10**, 696–698 (2014).
14. Vinogradov, A. A., Yin, Y. & Suga, H. Macrocyclic Peptides as Drug Candidates: Recent Progress and Remaining Challenges. *J Am Chem Soc* **141**, 4167–4181 (2019).
15. Giordanetto, F. & Kihlberg, J. Macrocyclic drugs and clinical candidates: What can medicinal chemists learn from their properties? *J Med Chem* **57**, 278–295 (2014).
16. Rigby, M. *et al.* BT8009; A Nectin-4 Targeting Bicycle Toxin Conjugate for Treatment of Solid Tumors. *Mol Cancer Ther* **21**, 1747–1756 (2022).
17. Rigby, M. *et al.* BT8009; A Nectin-4 Targeting Bicycle Toxin Conjugate for Treatment of Solid Tumors. *Mol Cancer Ther* **21**, 1747–1756 (2022).
18. Bendell, J. C. *et al.* BT5528-100 phase I/II study of the safety, pharmacokinetics, and preliminary clinical activity of BT5528 in patients with advanced malignancies associated with EphA2 expression. *https://doi.org/10.1200/JCO.2020.38.15_suppl.TPS3655* **38**, TPS3655–TPS3655 (2020).
19. McCloskey, K. *et al.* Machine learning on DNA-encoded libraries: A new paradigm for hit finding. *J Med Chem* **63**, 8857–8866 (2020).
20. Kómár, P. & Kalinić, M. Denoising DNA Encoded Library Screens with Sparse Learning. *ACS Comb Sci* **22**, 410–421 (2020).

21. Wilson, D. S., Keefe, A. D. & Szostak, J. W. The use of mRNA display to select high-affinity protein-binding peptides. *Proc Natl Acad Sci U S A* **98**, 3750–3755 (2001).
22. Gai, S. A. & Wittrup, K. D. Yeast surface display for protein engineering and characterization. *Curr Opin Struct Biol* **17**, 467–473 (2007).
23. Boder, E. T. & Wittrup, K. D. Yeast surface display for screening combinatorial polypeptide libraries. *Nature Biotechnology 1997 15:6* **15**, 553–557 (1997).
24. Thomas, W. D., Golomb, M. & Smith, G. P. Corruption of phage display libraries by target-unrelated clones: Diagnosis and countermeasures. *Anal Biochem* **407**, 237–240 (2010).
25. Matochko, W. L., Cory Li, S., Tang, S. K. Y. & Derda, R. Prospective identification of parasitic sequences in phage display screens. *Nucleic Acids Res* **42**, 1784–1798 (2014).
26. Ito, T. *et al.* Selection of target-binding proteins from the information of weakly enriched phage display libraries by deep sequencing and machine learning. *MAbs* **15**, 2168470 (2023).
27. Menendez, A. & Scott, J. K. The nature of target-unrelated peptides recovered in the screening of phage-displayed random peptide libraries with antibodies. *Anal Biochem* **336**, 145–157 (2005).
28. Tjhung, K. F. *et al.* Silent Encoding of Chemical Post-Translational Modifications in Phage-Displayed Libraries. *J Am Chem Soc* **138**, 32–35 (2016).
29. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **28**, 3150–3152 (2012).
30. Coates, A. & Ng, A. Y. Learning feature representations with K-means. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **7700 LECTURE NO**, 561–580 (2012).
31. Saarela, M. & Jauhiainen, S. Comparison of feature importance measures as explanations for classification models. *SN Appl Sci* **3**, 1–12 (2021).
32. Schissel, C. K. *et al.* Deep learning to design nuclear-targeting abiotic miniproteins. *Nat Chem* **13**, 992–1000 (2021).
33. Torres, M. D. T., Melo, M. C. R., Crescenzi, O., Notomista, E. & de la Fuente-Nunez, C. Mining for encrypted peptide antibiotics in the human proteome. *Nature Biomedical Engineering 2021 6:1* **6**, 67–75 (2021).
34. Li, G., Iyer, B., Prasath, V. B. S., Ni, Y. & Salomonis, N. DeepImmuno: deep learning-empowered prediction and generation of immunogenic peptides for T-cell immunity. *Brief Bioinform* **22**, 1–10 (2021).
35. Saka, K. *et al.* Antibody design using LSTM based deep generative model from phage display library for affinity maturation. *Scientific Reports 2021 11:1* **11**, 1–13 (2021).
36. Mason, D. M. *et al.* Optimization of therapeutic antibodies by predicting antigen specificity from antibody sequence via deep learning. *Nature Biomedical Engineering 2021 5:6* **5**, 600–612 (2021).
37. Quartararo, A. J. *et al.* Ultra-large chemical libraries for the discovery of high-affinity peptide binders. *Nat Commun* **11**, 3183 (2020).
38. Rini, J. M., Schulze-Gahmen, U. & Wilson, I. A. Structural evidence for induced fit as a mechanism for antibody-antigen recognition. *Science (1979)* **255**, 959–965 (1992).
39. Houghten, R. A. *et al.* Generation and use of synthetic peptide combinatorial libraries for basic research and drug discovery. *Nature 1991 354:6348* **354**, 84–86 (1991).

40. Chang, Y. S. *et al.* Stapled α−helical peptide drug development: A potent dual inhibitor of MDM2 and MDMX for p53-dependent cancer therapy. *Proceedings of the National Academy of Sciences* **110**, E3445–E3454 (2013).
41. Phan, J. *et al.* Structure-based design of high affinity peptides inhibiting the interaction of p53 with MDM2 and MDMX. *Journal of Biological Chemistry* **285**, 2174–2183 (2010).
42. Bernal, F., Tyler, A. F., Korsmeyer, S. J., Walensky, L. D. & Verdine, G. L. Reactivation of the p53 tumor suppressor pathway by a stapled p53 peptide. *J Am Chem Soc* **129**, 2456–2457 (2007).
43. Zondlo, S. C., Lee, A. E. & Zondlo, N. J. Determinants of specificity of MDM2 for the activation domains of p53 and p65: Proline27 disrupts the MDM2-binding motif of p53. *Biochemistry* **45**, 11945–11957 (2006).
44. Ye, X. *et al.* Binary combinatorial scanning reveals potent poly-alanine-substituted inhibitors of protein-protein interactions. *Communications Chemistry 2022 5:1* **5**, 1–10 (2022).
45. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
46. Chen, K.-H., Hu, Y.-J., Chen, K.-H. ; & Hu, Y.-J. Residue–Residue Interaction Prediction via Stacked Meta-Learning. *International Journal of Molecular Sciences 2021, Vol. 22, Page 6393* **22**, 6393 (2021).
47. Li, Y., Brian Golding, G. & Ilie, L. DELPHI: accurate deep ensemble model for protein interaction sites prediction. *Bioinformatics* **37**, 896–904 (2021).
48. Zamyatnin, A. A. & Anokhin, P. K. Amino Acid, Peptide, and Protein Volume in Solution. *https://doi.org/10.1146/annurev.bb.13.060184.001045* **13**, 145–165 (2003).
49. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv* (2018) doi:10.48550/arxiv.1802.03426.
50. Huang, Z., Research, B., Xu, W. & Baidu, K. Y. Bidirectional LSTM-CRF Models for Sequence Tagging. (2015).
51. Bergstra, J., Bardenet, R., Bengio, Y. & Kégl, B. Algorithms for Hyper-Parameter Optimization. *Adv Neural Inf Process Syst* **24**, (2011).
52. Luxburg, U. von & Schölkopf, B. Statistical Learning Theory: Models, Concepts, and Results. *Handbook of the History of Logic* **10**, 651–706 (2011).
53. Kusumoto, Y. *et al.* Highly Potent and Oral Macrocyclic Peptides as a HIV-1 Protease Inhibitor: mRNA Display-Derived Hit-to-Lead Optimization. *ACS Med Chem Lett* (2022) doi:10.1021/ACSMEDCHEMLETT.2C00310.
54. Iskandar, S. E. & Bowers, A. A. mRNA Display Reaches for the Clinic with New PCSK9 Inhibitor. *ACS Med Chem Lett* (2022) doi:10.1021/ACSMEDCHEMLETT.2C00319.
55. Lau, J. L. & Dunn, M. K. Therapeutic peptides: Historical perspectives, current development trends, and future directions. *Bioorg Med Chem* **26**, 2700–2707 (2018).
56. Rogers, J. M., Passioura, T. & Suga, H. Nonproteinogenic deep mutational scanning of linear and cyclic peptides. *Proc Natl Acad Sci U S A* **115**, 10959–10964 (2018).
57. Loh, W. Y. Fifty Years of Classification and Regression Trees. *International Statistical Review* **82**, 329–348 (2014).
58. Lundberg, S. M., Allen, P. G. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. *Adv Neural Inf Process Syst* **30**, (2017).

59. Dickinson, Q. & Meyer, J. G. Positional SHAP (PoSHAP) for Interpretation of machine learning models trained from biological sequences. *PLoS Comput Biol* **18**, e1009736 (2022).