communications biology

A Nature Portfolio journal

Article

https://doi.org/10.1038/s42003-025-09160-y

Decoding protein binding plasticity via integrated deep ribosome display and deep learning

Tang Mengtong[1,2,3,5], Li Jiawei[1,2,3,5], Li Zhixi[1,2], Cui Jingsong[4,6] ✉ & Qi Hao[1,2,6] ✉

The plasticity in protein interaction is central to understanding biological networks and de novo protein design. However, the systematic exploration remains impeded by the astronomic dimensionality of sequence space. Here, we present a platform that synergizes deep experimental screening with deep learning to decode interaction plasticity. By developing a ribosome display stripped of all known ribosome termination and rescue functions, we produce a comprehensive dataset comprising 47.8 million unique peptides spanning a broad spectrum of Streptactin-binding activity. A deep learning architecture, systematically trained on sequence context, enrichment dynamics, and subsequence abundance, achieves high accuracy (Pearson's $r = 0.902$) on predicting Streptactin-binding activity. Through sequence dimensionality reduction, exhaustive subsequence elucidation, and enriched motif elicitation, we identify 799 strong-binding sequences containing a canonical motif and 219 sequences harboring novel motifs with divergent docking conformations. These findings reveal an unanticipated depth and breadth in protein-binding plasticity. We propose that this integrated experimental-AI framework will facilitate the systematic exploration of protein interactions and enable the data-driven design of synthetic peptides.

Protein interactions commonly exhibit promiscuity, a notable trait that empowers proteins to engage with a wide-ranging array of partners[1]. This inherent ability underscores the adaptability of proteins, allowing them to fulfill a wide variety of functions within complex biological systems. The plasticity in the binding of proteins such as calmodulin[2], β-catenin[3], CDKs[4], and p53[5] forms a crucial mechanistic foundation for complex biological processes. It enables these proteins to participate in cellular signaling, metabolic regulation, and other essential functions, ensuring the proper functioning and homeostasis of biological systems. In essence, the plasticity of protein interactions is central to the complex and dynamic nature of life, offering valuable insights into the underlying mechanisms of biological phenomena. Beyond fundamental biology, decoding the interaction plasticity holds transformative potential for biotechnology. It informs the rational design of synthetic proteins with programmable binding specificities, aids in targeting elusive interfaces in drug discovery, and underpins the engineering of adaptive biomaterials. Thus, dissecting the molecular logic of plasticity bridges the gap between understanding native protein networks and innovating next-generation therapeutic and biotechnological tools.

Systematic mapping of protein-protein interactions (PPIs) relies on scalable experimental and computational approaches to decipher complex biological networks. Established techniques such as yeast two-hybrid (Y2H) assays[6], affinity purification-mass spectrometry (AP-MS)[7], and protein microarrays are increasingly integrated with automated workflows and bioinformatic pipelines[8]. Among these, display technologies offer distinct advantages for high-throughput interaction analysis, combining genetic encoding with selective pressure to interrogate binding events. Phage display, a pioneering method, exploits bacteriophages to surface-present protein libraries, enabling affinity-driven selection through iterative panning and bacterial amplification, a process central to antibody engineering[9–11]. Microbial display systems further expand this paradigm: *E. coli* provides rapid prokaryotic expression for screening[12], while yeast systems accommodate eukaryotic post-translational modifications and folding requirements[13]. Recent advances have incorporated deep learning to extract deeper insights from these in vivo display platforms, yet current datasets remain constrained to a size of ~10^4[14], limiting the scope of predictive modeling. Concurrently, advanced protein structure prediction algorithms,

[1]Frontiers Science Center for Synthetic Biology (Ministry of Education), School of Synthetic Biology and Biomanufacturing, Tianjin University, Tianjin, China. [2]State Key Laboratory of Synthetic Biology, Tianjin University, Tianjin, China. [3]School of Chemical Engineering and Technology, Tianjin University, Tianjin, China. [4]School of Cyber Science and Engineering, Wuhan University, Wuhan, China. [5]These authors contributed equally: Mengtong Tang, Jiawei Li. [6]These authors jointly supervised this work: Jingsong Cui, Hao Qi. ✉e-mail: jscui@whu.edu.cn; haoq@tju.edu.cn

including AlphaFold[15,16] and RoseTTAFold[17] have been adapted to facilitate the computational prediction of PPIs.

In contrast to viral and cellular in vivo systems, ribosome display (RD)[18,19] exploits cell-free mRNA-ribosome-protein complexes to bypass inherent biological constraints. This in vitro approach not only accommodates highly diverse libraries but also enables stringent selection under non-physiological conditions, thereby facilitating the evolution of high-affinity interactions. Yet despite these advantages, deep learning integration with RD remains underdeveloped, which is a notable limitation given the technology's potential. Consequently, conventional methods relying on static structural models or sparse sampling of sequence-function space prove inadequate for deciphering the complex epistatic networks underlying interaction plasticity.

In this study, we developed a reliable, high-throughput experimental-computational pipeline that integrates RD, next-generation sequencing (NGS), and deep learning to model sequence-function landscapes at scale, and demonstrated how this model can be used to explore the vast space for protein engineering and binding specificity prediction (Supplementary Fig. 1). We engineered a deep RD system by leveraging *mf*-Lon-mediated orthogonal protein degradation to precisely and thoroughly eliminate all four translation termination factors and five ribosomal dissociation related rescue proteins from *E. coli* lysates, stabilizing translation complexes and enhancing display efficiency. This reprogrammed RD platform generated a dataset of 47.8 million unique Streptactin-binding peptide variants with experimentally resolved binding activities. Iterative binding enrichment refined the dataset, highlighting the system's capacity to capture expansive sequence-activity landscapes (Supplementary Fig. 1c). A deep learning framework systematically trained on sequence context, enrichment dynamics, and subsequence abundance achieved strong predictive accuracy (Pearson's $r = 0.927$ on training set, 0.902 on testing set) on Streptactin-binding activity (Supplementary Fig. 1e), establishing the first framework to decode high-dimensional RD data. Exhaustive analysis of the 1.28×10^9 7-mer subsequence space, combined with dimensionality reduction, mapped sequence determinants of binding plasticity. We identified 799 high-affinity sequences within canonical Streptactin motifs and 219 novel motifs adopting distinct docking orientations, revealing unprecedented structural adaptability (Supplementary Fig. 1f). By integrating experimental and computational methods, this study demonstrates a powerful approach for elucidating interaction adaptability and establishing a generalizable framework to investigate protein plasticity, thereby accelerating progress in artificial protein design.

Results

Ribosome display in reprogrammed cell extract

RD is an in vitro technique that uses specific regulatory sequences to stall translation[20–22], forming a nascent peptide–ribosome–mRNA ternary complex that directly links phenotype to genotype, enabling the screening of large libraries (10^7–10^{15}) without cell transformation[23,24]. However, conventional RD relies on multiple affinity enrichment cycles, limiting datasets to ~10^4 variants[24], and the complexes are unstable in cell lysates, where endogenous factors rapidly dismantle them, restricting high-throughput analysis. Advancements in complex stability and data processing are essential to fully realize RD's potential for generating deep, quantitative protein-binding datasets.

In cellular systems, translation termination is mediated by Class I (RF1, RF2) and Class II (RF3, RRF) release factors, which disassemble ribosomes upon stop codon recognition, while rescue factors resolve aberrant ribosomal complexes stalled during elongation (Fig. 1a). These endogenous machineries inherently destabilize the ternary ribosomal complex (nascent peptide-ribosome-mRNA) in RD. However, as these factors are essential for translation fidelity and cell viability, their removal from conventional lysates is incompatible with cellular survival. To address this, we engineered a cell-free system[25] by eliminating all 4 release factors and 3 rescue factors (ArfA, ArfB and Pth) through insertion of a synthetic pdt3 degradation tag at the C-terminus of essential genes on the genome (Fig. 1b), enabling normal cellular proliferation under standard growth conditions (Supplementary Fig. 2). Two nonessential rescue factor genes (*smrB* and *ssrA*) were directly deleted from BL21 genome. Finally, we constructed two engineered *E. coli* strains: BL21-T5, in which the four release factors (RF1, RF2, RF3, RRF) and *ssrA* were genomically edited, and BL21-TR9, in which additionally four ribosome rescue factors (ArfA, ArfB, Pth, and SmrB) were edited for deletion (Supplementary Table 1). During S30 lysate preparation, the pdt3 degradation tag directs the exogenous *mf*-Lon protease to selectively degrade these tagged essential proteins. Western blot analysis confirmed the complete removal of the Class I release factor RF1 and RF2 (Fig. 1b), validating the efficacy of the orthogonal degradation system. Consequently, this approach produced a translationally active lysate devoid of both translation termination and ribosome rescue machinery. The observed reduction in superfolder GFP (*sf*GFP) expression (Fig. 1c) in the BL21-TR9 lysate compared to the standard BL21 lysate aligns with the expected decrease in protein synthesis efficiency caused by termination deficiency, consistent with prior studies[25].

To enhance the expression efficiency of saturated NNK mutagenesis libraries, we optimized the suppression of the remaining TAG stop codon inherent to the NNK codon scheme. Building on prior work, we engineered a suppressor $tRNA^{leu}$ (derived from *E. coli* MG1655 gene 945662) to decode the TAG codon. Total RNA, enriched with the suppressor $tRNA^{leu}$ through overexpression in *E. coli* BL21(DE3) (see Methods), was purified (Fig. 1d) and supplemented into the cell lysate to enable the synthesis of proteins encoded by NNK-mutated sequences. Using wild-type *sf*GFP and an *sf*GFP variant containing the N149TAG mutation as reporter proteins, we systematically titrated the total RNA concentration to balance protein yield and decoding fidelity. At a total RNA concentration of 1.0 A_{260}, the system achieved optimal performance, with a decoding efficiency of 97.73% for the TAG codon (Fig. 1e), demonstrating robust suppression while maintaining high translational activity.

To assess the stability of ternary ribosomal complexes in lysates from engineered strains, stalled ribosomal complexes were generated in S30 extracts derived from BL21-T5 (termination factor-deficient) and BL21-TR9 (termination + rescue factor-deficient) strains and isolated via affinity pulldown. For quantification, FLAG epitope tags were inserted into the C-termini of ribosomal proteins RplQ (50S subunit) and RpsB (30S subunit) by precise genome editing (Supplementary Fig. 3a). Ribosome stalling was induced by translating a linear DNA target template encoding an N-terminal standard Strep-tag (WSHPQFEK), a central GeneIII spacer sequence spanning the ribosomal tunnel, and a C-terminal SecM stalling motif (lacking a stop codon) (Supplementary Fig. 4a). The DNA template (Supplementary Table 2), generated by PCR, was incubated for 2 hours in S30 lysates from standard BL21, BL21-T5, and BL21-TR9 strains respectively. During translation, the SecM motif stalled ribosomes after displaying the Strep-tag, while the GeneIII sequence occupied the ribosomal tunnel. Post-translation, Strep-tag-bearing complexes were affinity-captured using Streptactin-coated magnetic beads (ST-MB), co-pulldown stalled ribosomes (see Methods). Western blot analysis of pulldown fractions (Fig. 1f) and total lysate ribosomes (Supplementary Fig. 3b) quantified FLAG-tagged RplQ (50S) and RpsB (30S). Compared to standard BL21 lysates, the engineered strains exhibited markedly enhanced ribosomal recovery. BL21-T5 lysates achieved a 14.79-fold increase in ribosomal subunit recovery, while BL21-TR9 further elevated intact 70S complex retention by 1.90-fold over BL21-T5, a result attributable to the special suppression of 70S ribosome dissociation. These results demonstrate that eliminating termination factors (BL21-T5) substantially stabilizes ternary complexes, while additional removal of rescue factors (BL21-TR9) preferentially enhances 30S subunit retention, likely by preventing rescue factor-mediated complex disassembly during translation stalling.

The integrity of genotype-phenotype linkage in RD relies on the preservation of the intact 70S ribosomal complex. Consistent with this principle, lysates from the BL21-TR9 strain demonstrated superior capacity to generate stable, stalled ternary ribosomal complexes. To rigorously evaluate the system's selection fidelity, a competition assay was

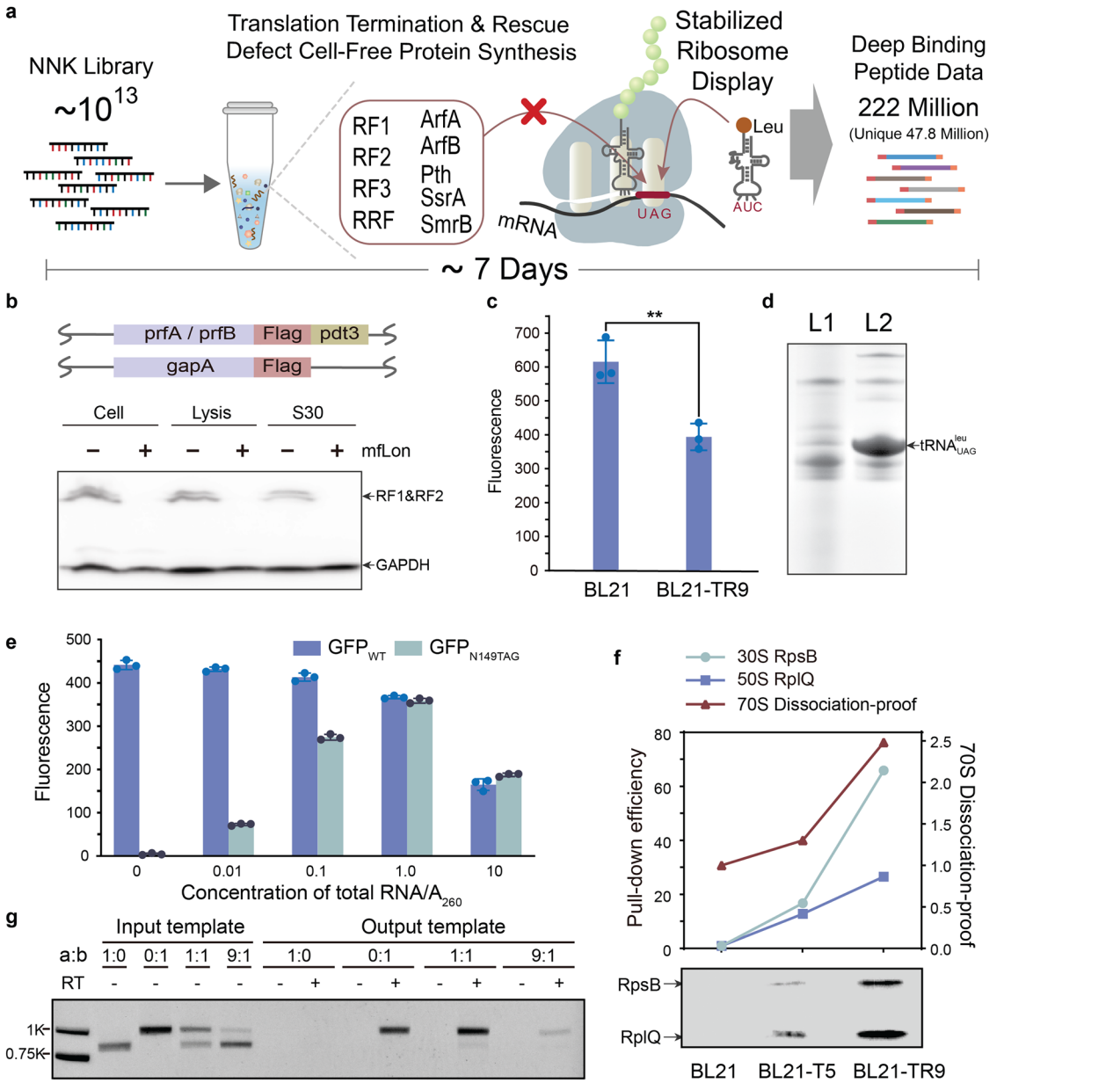


Fig. 1 | RD in Reprogrammed Cell Extract. a Schematic diagram outlining the steps involved in collecting protein interaction data. **b** Western blot analysis assessing degradation efficiency of termination factors RF1 (MW. 40.5 kDa), RF2 (MW. 41.2 kDa), and GAPDH (MW. 36.9 kDa), each tagged with a C-terminal FLAG-pdt3. Samples were collected from whole cells, post-physical homogenization lysates, and S30 extracts probed with anti-FLAG antibodies. **c** Comparison of *sf*GFP synthesis capabilities between BL21 and BL21-TR9, measured by averages of three independent reactions ($n = 3$ biological replicates, bars represent the mean of all replicates with ±SD; $p = 0.0067$, two-tailed t test). **d** PAGE analysis of total RNA extracted from bacterial cells: Lane 1 contains bacteria with an empty vector plasmid, while Lane 2 contains bacteria carrying a plasmid expressing $tRNA^{Leu}_{TAG}$. **e** Assessment of the CFPS system's capability to translate $sfGFP_{WT}$ and $sfGFP_{N149TAG}$ upon varying total RNA concentrations, each conducted in triplicate. Fluorescence measurements indicate a decrease in of $sfGFP_{WT}$(blue) synthesis with increasing total RNA level, while $sfGFP_{N149TAG}$(green) translation shows initial enhancement of TAG decoding at $A_{260} < 1.0$ total tRNA levels followed by suppression at higher concentrations ($n = 3$ biological replicates, bars represent the mean of all replicates with ±SD). **f** Evaluation of ribosomal subunit dissociation during RD. The pulldown efficiencies of 30S (RpsB, green) and 50S (RplQ, blue) subunits were quantified by the ratio of pulled-down ribosomal protein (Western blot intensity) to total protein in S30 extract. Dissociation-proof efficiency (red) was measured similarly using RD pulldown. All data were normalized to standard Strep-tag controls. Western blot depicts RplQ and RpsB levels in pulldown fractions. **g** RT-PCR analysis of cDNA recovery in the first cycle of RD with different mixing ratios of positive and negative templates. The larger molecular weight band represents the positive template.

performed using mixed a target template encoding the standard Strep-tag and a reference template encoding a non-binding control peptide (GGGSGGGS) (Supplementary Fig. 4b). The templates were combined at ratios of 1:0, 0:1, 1:1, and 9:1 (Reference: Target). The translation efficiency of the two linear templates showed no significant differences (Supplementary Fig. 4c). Following a single-round RD workflow optimized for BL21-TR9 lysates (Supplementary Fig. 5), the target template was selectively enriched across all ratios, achieving robust recovery even at a 9:1 starting ratio (Fig. 1g). These results underscore the engineered lysate's capacity to stabilize ternary ribosomal complexes while enabling stringent affinity-based selection. By systematically eliminating endogenous all translation termination and ribosome rescue machineries, this cell-free system circumvents the inherent instability of ternary complexes in conventional RD. Furthermore, the integration of orthogonal TAG stop

codon suppression ensures full decoding of all codons within saturated NNK mutagenesis libraries.

Large scale of Streptactin-binding peptides from deep ribosome display

Using Streptactin as a binding model, we designed and constructed a linear DNA library for deep RD selection. This library includes the following elements: a T7 promoter and a 25nt UTR sequence for initiating in vitro transcription and translation; a Myc tag for stabilizing protein expression; a 12-amino-acid-long NNK mutational region to generate sequence diversity; a GeneIII spacer sequence to ensure the nascent peptide extends properly from the ribosome tunnel; and a SecM stalling sequence to halt ribosome translation (Fig. 2a). The library was constructed via a one-step in-fusion PCR process (Supplementary Fig. 6a). Specifically, a 137-nt primer containing T7, UTR and 12×NNK degenerate codons prepared using uniform chemical synthesis with balanced nucleotide distribution at each position were fused with downstream fragment containing GeneIII and SecM parts by a one-step PCR (see Methods) to construct the linear DNA library.

In a modified procedure of RD (Supplementary Fig. 5), the linear DNA library was translated in BL21-TR9 lysate at 30 °C for 2 hours to form a ternary ribosomal complex pool. ST-MB were used to isolate displayed peptides with binding activity. After the washing step, the mRNA was eluted from the magnetic pull-down beads using Biotin buffer and reverse-transcribed into a new DNA library and subjected to the next round of RD screening or subjected to sequencing to determine the genotypes of selected binding peptides. The selection was repeated for five rounds (see Methods). To enable high-resolution tracking of selection dynamics, we performed deep sequencing on both the initial library and on each of the selection rounds. Across six datasets (round 0 to round 5), we obtained 243.25 million reads (73.42GBase) in total. After standard processing to extract the binding peptide region (Supplementary Fig. 7a), 221.99 million perfectly matched reads were retained, achieving a sequence retention efficiency of 91.26% (Fig. 2b; Supplementary Fig. 7b).

Of these, the initial library yielded 34,002,357 reads. After processing with a standard sequence extraction pipeline (see Methods), we identified 26,154,209 unique amino acid sequences with an average sequencing depth of 1.30, confirming the high diversity of the library (Fig. 2c). Based on total DNA template molecules added in the first round of selection, the theoretical sequence space of the library was estimated to be approximately 3.6×10^{13}. Additionally, sequence logo analysis showed uniform distribution without significant base bias (Supplementary Fig. 6b). Serine (S), arginine (R), and leucine (L) showed slightly higher entropy due to broader codon usage in the NNK scheme. Through five iterative rounds of screening, we identified a total of 21,685,678 unique amino acid sequences. Strikingly, only 593 of these sequences (0.0027%) were identified in the initial library, demonstrating its exceptionally high diversity that exceeded the limits of the sequencing depth. We quantified screening dynamics by analyzing the proportion of sequences unique to each round (round-specific unique sequences), calculated as the number of sequences appearing exclusively in a given round divided by the total number of unique sequences detected across five rounds. This analysis revealed a progressive decline from 16,128,508 (33.7%) in round 1 to 506,712 (1.1%) in round 5 (Fig. 2d), indicating effective enrichment of functional peptides and progressive library convergence. The systematic decline in round-specific unique sequences reflects enrichment of selected peptides and validates this deep screening protocol's capacity to resolve high-performance candidates within the vast diversity space in the library. Analysis of peptide sequence depth distribution revealed a clear enrichment pattern during selection from round 0 to round 5 (Fig. 2e). Sequencing depths were categorized into bins, with the y-axis representing the frequency (percentage) of sequences in each depth bin. The depth frequency histogram exhibited a rightward progression of the distribution curve, reaching a maximum observed depth of 68,653.

Through five consecutive rounds of screening, we systematically analyzed sequence depth correlations (Fig. 2f). Early rounds showed near-zero correlations with other rounds (e.g. R1 vs R5, r = -0.00) due to insufficient overlapping sequences; From round 2 onward, correlations increased significantly (R2 vs R5, r = 0.48; R2 vs R4, r = 0.59), reaching high consistency in later stages (R4 vs R5, r = 0.91) (Supplementary Fig. 8). The data indicate that this Streptactin-binding screening of engineered RD has a good reproducibility and enrichment progress plateaued by round 2, with sequence diversity contracting sharply as low-activity peptides were competitively eliminated during screening. This stabilized phase retained a strong correlation with subsequent rounds (R2 vs R3, r = 0.74) while providing richer sequence variations. These metrics underline the validity of our RD experiment and the high-throughput sequencing data, paving the way for further computational analysis.

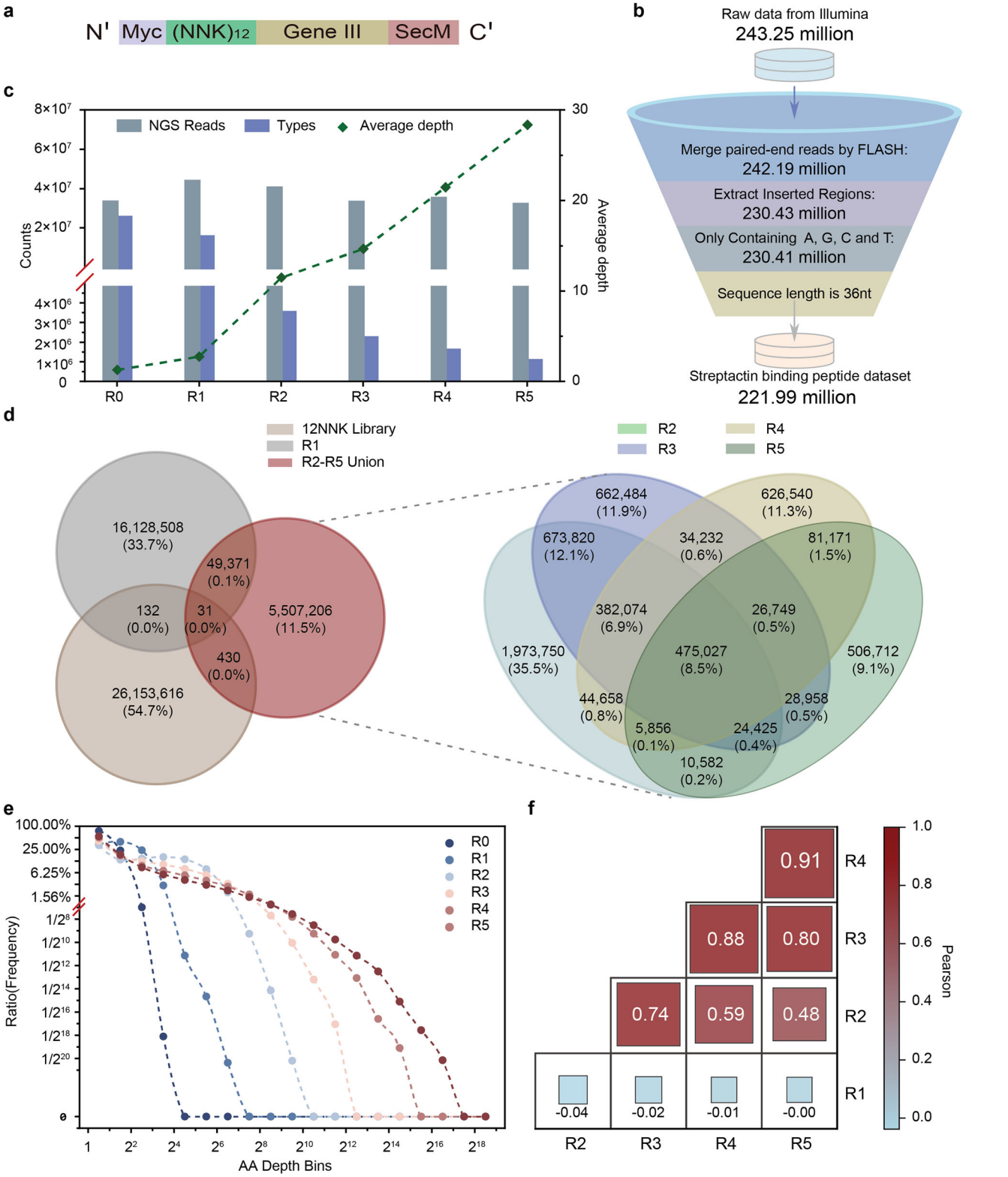
Deep learning model for characterizing streptactin-binding peptides

To resolve the complexity in this vast binding peptide dataset, we adopted a deep learning framework combining a recurrent neural network (RNN) with a multilayer perceptron (MLP) for regression analysis, aiming to calculate the binding activity based on the sequence context and sequencing metrics. The RNN, considered effective for tasks on short sequences[26–29], serves as an encoder to extract features from Onehot-encoded input sequences, followed by a fully connected network for activity prediction (Fig. 3a). This architecture enables multi-level peptide feature extraction, enhancing prediction accuracy. The long-tailed distribution of sequence depths (Fig. 2e) was alleviated through logarithmic conversion, generating an estimated continuous value reflecting the binding potential.

Understanding the interaction between sequencing depth and activity is crucial for effective modeling and analysis. While these two factors are generally positively correlated, this relationship exhibits nonlinear and context-dependent behavior in practice due to stochastic sampling biases and activity-dependent selection pressures. In conventional RD approaches, high-activity sequences are typically determined by screening dozens to hundreds or thousands of individual colonies. This approach is fundamentally based on the probabilistic assumption that sequences with higher activity will demonstrate proportionate enrichment after screening. However, this correlation holds reliably only for the most abundant top sequences (Supplementary Fig. 8), as stochastic sampling noise and activity-independent biases disproportionately distort the abundance-activity relationship for mid- to low-frequency sequences.

Furthermore, given that the essential binding region may represent only a subsequence of the 12-amino-acids random sequence in this huge dataset, particularly when sequencing depth is insufficient to fully cover the entire sequence space, we observed that sequencing depth fails to maintain a strong positive correlation with actual activity for the vast majority of sequenced variants. Therefore, we systematically evaluated alternative metrics to derive quantitative activity values from sequencing data for optimal model training. There are two key issues that need to be addressed. First, the total sequence space of a 12-amino-acid peptide is extremely large (20^{12}), far exceeding the number of sequencing reads that can be obtained (approximately 10^8 to 10^9). As a result, many sequences with low abundance may not necessarily have significantly weaker activity than those with higher depth; rather, their lower representation could simply be due to stochastic under-sampling, experimental bias or statistical randomness. Second, the biophysical characteristics of Streptactin[30–32] indicated that the effective binding interface is actually shorter than the full 12-amino-acids peptide. To address this, we introduced a subsequence diversity metric, measured by their occurrence within the full-length deduplicated dataset, to quantify the binding activity.

We first trained various lengths of the peptide subsequences to determine the optimal window size that retained the most informative features while discarding irrelevant or noisy residues. The subsequence metric derives from an evolutionary selection hypothesis: functional motifs critical for binding are retained under positive selection pressure, resulting in conservation of these subsequences across the dataset. This conservation manifests as high diversity of flanking sequence contexts around stable core



motifs, while non-functional residues accumulate neutral mutations that contribute primarily to stochastic sequence noise. Adopting this strategy brings various benefits. First, it transforms the intractable ~20^{12} combinatorial sequence space into analyzable motif-occurrence distributions, mitigating undersampling. Second, the use of deduplicated occurrences alleviates the impact of sampling noise. Third, it prevents distortion by hyperabundant non-functional sequences to ensure balanced inclusion of distinct sequences. Next, we trained sequence data from different screening rounds, with sequences from early-stage screening rounds exhibiting a broader, less refined sequence distribution, and late-stage exhibiting a more enriched set of deep sequence binders. Therefore, we investigated different metrics for the activity of a subsequence in a deep learning pipeline. Specifically, we considered three distinct metrics: sequencing depth (reflecting the enrichment dynamics across rounds), the number of unique sequence

Fig. 2 | Large Scale of Streptactin-binding Peptides from Deep RD. a Protein structure in RD screening. Myc tag equalizes translation levels; 12×NNK random peptides; Gene3 spacer; SecM sequence enables ribosome stalling and stable ternary complex formation. **b** Statistics at each step of the standard sequence extraction pipeline. 243.25 million NGS reads containing 73.42GBase were processed to extract the inserted NNK region, obtaining 221.99 million high-quality reads for further analysis. **c** Sequencing data statistics after five rounds of RD screening. Total sequencing reads and unique amino acid sequence variants after each screening round. While read counts remained similar across rounds, type number of variants decreased significantly with screening progression, indicating sequence enrichment. **d** Venn diagram illustrating the distribution of sequence sets across screening stages: the gray and brown circles represent the initial library and the sequences after the first round of screening, respectively; the red area denotes the union of sequences from the second to fifth rounds of screening. The right subpanel shows an independent Venn diagram of the results from rounds 2, 3, 4, and 5, with overlapping regions representing shared sequences between rounds. **e** Distribution of sequencing depth after five rounds of RD screening. Sequence depth (x-axis) versus cumulative frequency (y-axis). The proportion of high-depth sequences increases with screening rounds (0~5, color gradient from blue to red). Both axes are scaled logarithmically. **f** Correlation analysis of sequence depth across five rounds of RD screening. Heatmap of Pearson's correlation coefficients between sequencing depths in rounds 1~5. Round 1 showed low correlation with other rounds ($r \approx 0.00$), while rounds 2~5 exhibited progressively stronger correlations ($r = 0.48$~0.91). Color gradient indicates correlation strength (blue: low; red: high).

variants (reflecting the retained diversity), and a composite metric defined as the product of sequencing depth and diversity. Subsequently, we conducted an exhaustive search over all possible combinations of these hyperparameters to identify the optimal configuration. Comprehensive training experiments and cross-validation revealed that the best performance was achieved with a 7-amino-acid subsequence, data sourced from R2, and diversity as the primary label (Fig. 3b, c**;** Supplementary Fig. 9**;** Supplementary Fig. 10). The Pearson's correlation coefficient (r) of the resulting model achieved 0.927 on the training set and 0.902 on the testing set (Supplementary Figs. 9a**;** 11). This result indicates that the model has been highly successful in both training and prediction, and its predictions align closely with the observed binding affinity of input peptide sequences.

To validate the efficacy of this trained Streptactin-binding potential (SBP) model trained on our high-throughput sequencing data, we independently screened candidate variants using two distinct computational frameworks: a full-length sequence model analyzing complete 12-amino-acid sequences and a model trained on 7-amino-acid subsequences targeting a refined motif previously identified as the core determinant of functional activity. From each of the two models, 200 variants predicted to outperform the standard Strep-tag sequence and not in the training set were validated using the independent round 3 RD dataset. The model trained on 7-amino-acid subsequences demonstrated robust accuracy, with its activity predictions for high-performance candidates exhibiting a statistically significant correlation to experimental measurements (Pearson's $r = 0.6835$) (Fig. 3d). This strong agreement suggests that the 7-amino-acid motif captures sufficient information to compensate for the omission of flanking regions, effectively filtering nonessential sequence noise while preserving functional patterns. Concurrently, the full-length sequence model also demonstrated its utility, albeit through a different lens. When screening for high-activity variants, the full-length model selected sequences that, upon experimental validation, exhibited significantly enhanced binding activity relative to those derived from random sampling (Fig. 3e). Such findings underscore that even when the entire sequence is taken into account—despite the potential introduction of extraneous or less relevant information—the full-length model is capable of identifying high-activity variants with clear advantages over non-directed selection methods. This performance reinforces the idea that comprehensive sequence data, when integrated into a robust deep learning framework, can yield meaningful and actionable predictions in protein engineering.

Decoding Streptactin-binding plasticity

While Streptactin is a gold-standard binding protein, the plasticity of its binding pocket remains underexplored. Leveraging the trained SBP model, we conducted an in-depth exploration of the global features of Streptactin-binding peptides to demonstrate the diversity of sequences being able to well accommodate into a protein binding pocket, with potential surpassing current benchmarks through distinct structural epistasis.

We exhaustively evaluated the entire 7-amino-acid sequence space (1.28×10^9 total sequences) using the SBP model and identified 1,018 strong sequences with SBP values exceeding that of the standard Strep-tag sequence "WSHPQFEK" (evaluated as 8.338). We applied t-SNE to visualize latent representations of these sequences in a reduced-dimensional space. Notably, the full dataset of screened sequences in R2 exhibited diffused, overlapping clusters (Fig. 4a). In contrast, these strong sequences clearly resolved into discrete clusters (Fig. 4b). Several other dimensionality reduction methods were also applied, followed by clustering algorithms on these sequences. Although the results vary slightly depending on parameters, the overall conclusions are consistent: the reduced set of 1,018 high-activity sequences can be grouped into three major classes, with HP sequences further subdivided into smaller clusters (see Methods; Supplementary Fig. 12 and Supplementary Fig. 13). On the reduced set, we also performed motif detection using STREME, which produced similar results (Supplementary Fig. 14). These results suggest the existence of 6 distinct sequence motifs. Four motifs, "H-P-Q-G", "H-P-M-G", "H-P-Q-F", and "H-P-M-F", converge into a large "H-P" centered supercluster, comprising 799 strong sequences (Fig. 4c). The canonical Strep-tag sequence localizes to this supercluster, and this large number of sequences aligning with this supercluster underscores the depth of plasticity inherent to the binding pocket. Beyond the supercluster, novel motifs "E-x-W-L" (198 sequences) and "P-x-W-W-x-x-L" (21 sequences) were discovered (Fig. 4c, d). These newly identified motifs further illustrate the unexpected breadth of binding pocket plasticity. The "H-P" cluster exhibited a higher average SBP value compared to the "E-x-W-L" and "P-x-W-W-x-x-L" clusters. Three representative peptides, HPMGERS (SBP = 10.9) from the "H-P" supercluster, SGLELWL (SBP = 10.2) from the "E-x-W-L" motif and PSWWYSL (SBP = 9.2) from the "P-x-W-W-x-x-L" motif, were cloned as the N-terminal fusions to *sf*GFP (Fig. 4e) and expressed in *E. coli* BL21(DE3) cells for further experimental validation. These purified fusion *sf*GFP were subjected to a pulldown assay with Streptactin magnetic beads. The robust binding was confirmed for all three peptides from these distinct motifs using western blot (Fig. 4e). Precise fluorescence quantification further revealed that all these peptides exhibited significantly higher efficiencies than the standard Strep-tag peptide (SBP = 8.338), consistent with their SBP rankings (Fig. 4f). Notably, the SBP model validation experiment (Fig. 4e, f) confirmed the binding function of sequences encompassing all these identified motifs (Supplementary Fig. 15a, b).

In addition to the binding assays, we conducted molecular docking computation on the three selected peptide sequences, aiming to elucidate the detailed molecular interactions underlying the wide binding plasticity. The most stable binding modes of the three peptides with Streptactin is displayed, revealing that all three peptides can be successfully embedded within the active site of Streptactin (Fig. 4g–i). Meanwhile, the peptides form complexes with specific amino acid residues of Streptactin through non-covalent interactions. 2D interaction diagrams generated using LigPlot illustrate different binding patterns (Supplementary Fig. 16). Among them, HPMGERS primarily forms hydrogen bonds with Ser76 and Thr78, while engaging in hydrophobic interactions with Leu13, Gly33, Trp42, Trp67, Leu92, and Leu98. SGLELWL exhibits a more extensive hydrogen-bonding network, forming five hydrogen bonds with Asn11, Ser15, Tyr31, Ser40, and Asp116, along with hydrophobic contacts involving Leu13, Arg35, Trp42, Trp67, Arg72, Ser76, Trp80, Trp96, and

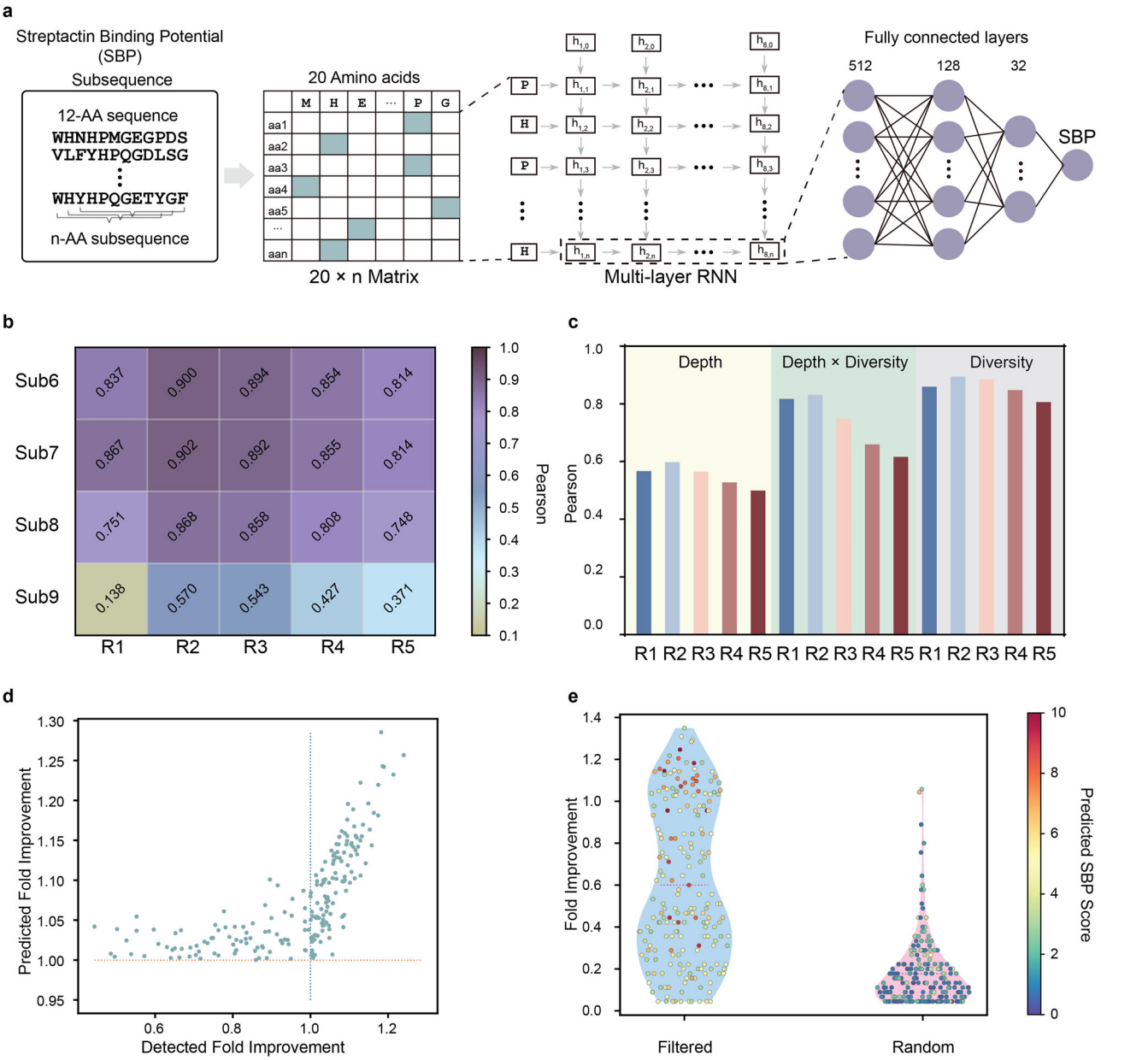


Fig. 3 | Deep Learning Model for Characterizing Streptactin-Binding Peptides. **a** A schematic overview of the SBP model architecture. The peptide subsequences are first one-hot encoded, and then passed through a multi-layer RNN. The embedded feature is then passed through an MLP to compute a final prediction value on the activity of the given sequence. **b** Heatmap visualization comparing the Pearson's correlations on testing set when training on different truncation lengths and data sources, using diversity as true label. Truncation lengths greater than 8 exhibits a substantial drop. **c** Bar plot comparing the Pearson's correlations on testing set across different rounds and three different training metrics (depth, diversity, depth×diversity). **d** Scatter plot comparing predicted fold improvement (x-axis) with experimentally measured fold improvement (y-axis) in binding affinity across 200 selected 7-amino-acid sequences. **e** Vertical violin plots comparing experimentally measured fold improvement distributions between 200 variants filtered using trained model (left) and 200 randomly selected counterparts (right). Each sample is visualized as a colored dot (warm: high, cool: low) reflecting its predicted activity and positioned at the appropriate value within the distribution for each group.

Leu98. PSWWYSL interacts via hydrogen bonds with Tyr31 and Ala34, while its hydrophobic interactions involve Leu13, Ser15, Trp67, Ala74, Ser76, Leu98, and Leu112. Generally speaking, a more negative binding free energy (ΔG) indicates stronger and more stable binding. The predicted binding energies for HPMGERS, SGLELWL, and PSWWYSL were -9.054, -8.071, and -6.225 kcal/mol, respectively. These values align with their SBP rankings. This suggests that HPMGERS binds most readily to Streptactin, followed by SGLELWL and then PSWWYSL. These results are in strong agreement with both the pulldown experiments and the computational predictions. Furthermore, predictions on complexes of Streptactin with the three short peptides using AlphaFold3 and calculation of binding free energy using PRODIGY are generally consistent in trend with the docking results from AutoDock (Supplementary Fig. 17).

Overall, the integration of dimensionality reduction, exhaustive subsequence elucidation, enriched motif elicitation, and molecular docking computation with the deep learning framework demonstrates the ability to efficiently and systematically map protein-binding plasticity while identifying high-binding potential sequences with novel, unexpected features. This consistency across independent methods confirms the robustness of our approach and its potential for rational protein design.

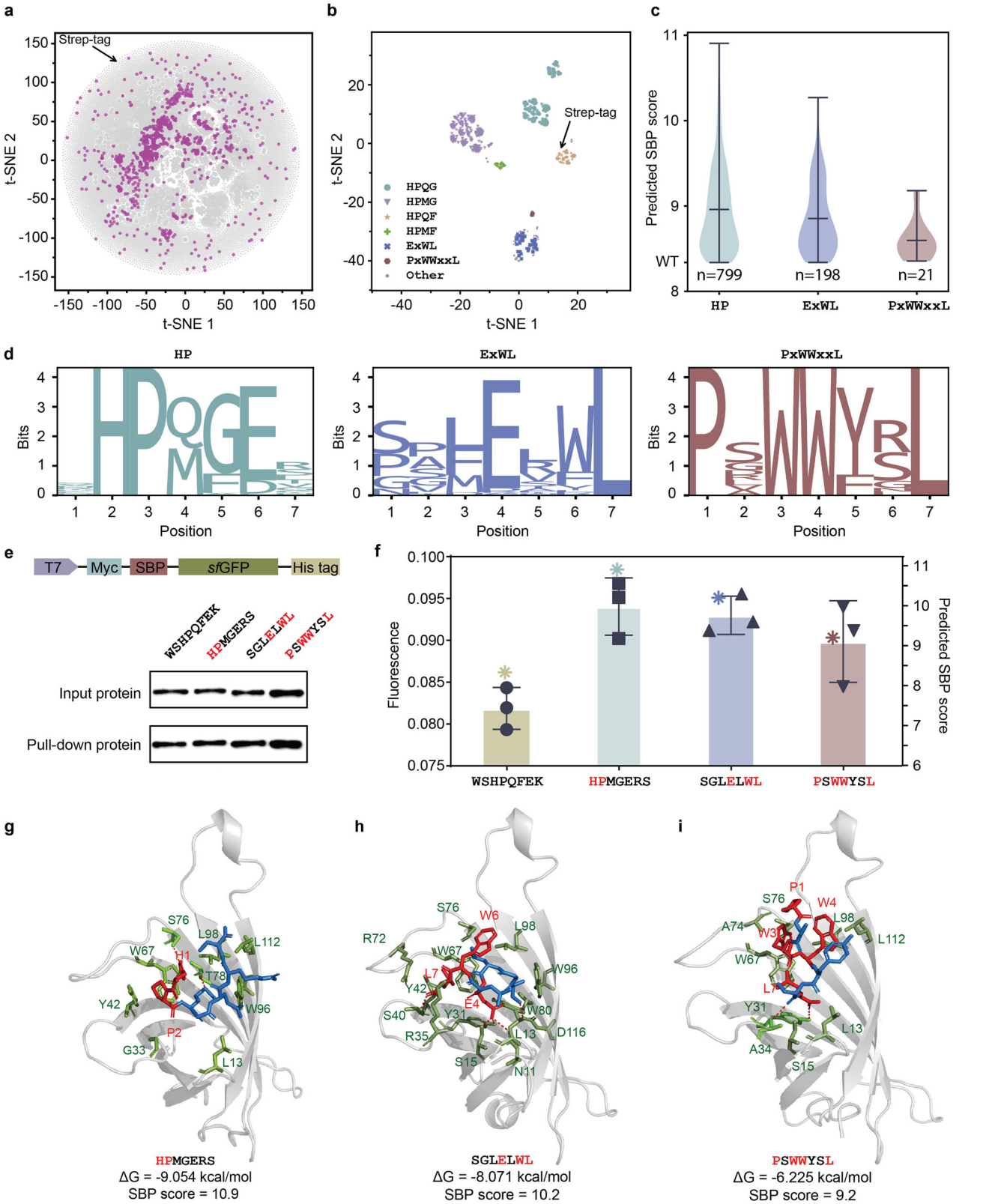


Fig. 4 | Decoding Streptactin-binding plasticity. Dimensionality reduction using t-SNE on the latent representations of all available 7-amino-acid sequences in R2 (**a**) and those with predicted activities over standard Strep-tag (**b**). **c** Distribution of predicted values and of the three clusters "H-P", "E-x-W-L", and "P-x-W-W-x-x-L". **d** Aligned sequence logos of the sequences of three clusters. **e** Western blot analysis of the pull-down ability of three representative 7-amino-acid peptides, compared to that of the standard Strep-tag peptide. **f** Fluorescence measurement (left) of the pull-down ability and predicted SBP score of three representative 7-amino-acid peptides, compared to those of standard Strep-tag. ($n = 3$ biological replicates, bars represent the mean of all replicates with ±SD). 3D Binding diagram of Streptactin with short peptides, HPMGERS (**g**), SGLELWL(**h**) and PSWWYSL (**i**), generated using PyMOL.

Discussion

This deep RD approach, enabled by an engineered cell lysate lacking ribosome rescue and termination machinery, enables high-throughput generation of interaction datasets across a wide range of affinities. While the PURE system (Protein Synthesis Using Recombinant Elements)[20,33], a reconstituted cell-free synthesis platform comprising purified translational components, can similarly exclude termination factors. However, its preparation involves labor-intensive purification steps that contrast with the streamlined production of crude cell lysates. The complexity of the PURE system assembly renders it time- and resource-intensive compared to the single-step lysate preparation methodology described here. Furthermore, the engineered lysate offers modular versatility for functional customization, such as orthogonal tRNA integration or protease-mediated degradation, and can be readily adapted to other bacterial strain backgrounds, enabling broader applications in synthetic biology and directed evolution workflows.

The comprehensive exploration of a target protein's binding plasticity necessitates large-scale datasets to encapsulate its diverse conformational states and interaction profiles. Proteins often adopt distinct structural conformations to engage with multiple binding partners, modulating affinities in response to environmental or biochemical stimuli. This adaptability is evolutionarily optimized. To systematically study such plasticity, the engineered cell-free RD platform offers dual advantages: (1) compatibility with expansive DNA libraries (>10^{13} variants) and (2) enhanced stabilization of ribosomal complexes, ensuring robust genotype-phenotype linkage for high-throughput data generation (>10^{8} variants). The platform's capacity to produce datasets spanning millions of unique sequences provides an unprecedented resource for probing binding landscape heterogeneity. However, the inherent high dimensionality of such datasets presents computational challenges.

Faced with a dataset of exceptional richness, multiple analytical paths are feasible. We demonstrate that subsequence-focused deep learning strategy, combined with dimensionality reduction and clustering, efficiently extracts information and identifies core motifs. Alternative methods, including manual inspection, rule-based pattern extraction, and conventional motif detection tools, were attempted but proved to be inefficient and unscalable for processing this bulk data. In contrast, a well-trained deep learning model with high predictive fidelity effectively compresses the entire high-dimensional dataset into a compact functional representation. This model enables efficient prioritization of sequences and interpretation of an otherwise intractable high-dimensional sequence space. This approach makes the rules underlying Streptactin plasticity more explicit and computationally accessible, while the training process inherently performs pattern extraction, noise reduction, and data smoothing.

Nevertheless, we recognize several limitations in our current strategy. Our choice of an RNN architecture was initially informed by the nature of protein translation and protein folding. We have also evaluated other structures (Supplementary Fig. 18). We acknowledge that modern transformer-based models, such as those employed in large language models, are powerful options for such modeling; their application, however, requires careful consideration of the performance-compute trade-off. Another limitation stems from the fact that our activity labels are round-specific, necessitating evaluation via within-round splits. Ideally, a smaller set of data points from high-precision assays would be collected for validation. However, such low-throughput validation methods, common in protein engineering workflows, incur substantial costs in time and resources while yielding far fewer data points, especially when compared to our deep RD system. We also consider data leakage as an important issue. Analysis of sequence similarity between splits confirms that while a minor fraction of sequences is similar, the majority are different (Supplementary Fig. 19). While random stratified splitting of the dataset is viable, analysis and careful control of data leakage is essential. We are actively planning to address these aspects in future investigations.

By integrating the complementary strengths of RD technology, NGS, and deep learning frameworks, we establish a unified platform that addresses critical limitations of traditional peptide screening methods. RD enables in vitro selection from large libraries without cellular transformation constraints. Deep learning algorithms further distill these datasets to identify latent patterns, enabling predictive modeling of peptide binding landscapes. This synergistic approach allows for exhaustive exploration of sequence space, uncovering both dominant and rare functional conformations that underlie protein binding plasticity. The platform's capacity to resolve high-dimensional sequence-activity relationships offers a transformative methodology for the rational design and optimization of binding peptides. We anticipate its application will accelerate therapeutic peptide discovery, particularly for targets with complex binding dynamics, such as intrinsically disordered proteins or allosteric sites, and advance protein engineering efforts by linking sequence diversity to functional outcomes. Furthermore, the integration of deep learning with experimental workflows enables data-driven design of synthetic peptides, bypassing heuristic optimization and directly informing iterative library refinement.

Methods

All unique materials are available from the corresponding authors.

Materials

The sequences and functional details of the oligonucleotides employed in this study are provided in Supplementary Table 3. All oligonucleotides and genes were synthesized by AZENTA. Degenerate primers were synthesized and purified using denaturing high-performance liquid chromatography (DHPLC) to ensure uniform quality. DNA purification was performed using the SPARKeasy Gel/PCR Purification Kit (Shandong Sparkjade Biotechnology Co., Ltd.). Cloning enzymes and reverse transcriptase were obtained from Vazyme, while DNA polymerases and RNA extraction/purification reagents were sourced from TransGen Biotech. Plasmid extractions were carried out using the TIANpure Mini Plasmid Kit (TIANGEN). All antibodies used in the experiments were supplied by ABclonal Technology Co., Ltd. (Wuhan, China).

Strains and plasmids

The *E. coli* BL21 Star (DE3) strain, obtained from Addgene, was employed as the wild-type host for both protein expression and the preparation of S30 extracts. The DH5α strain, also sourced from Addgene, was used for cloning purposes. Genome editing across all strains was conducted using λ-Red-mediated homologous recombination, and the pTKRED plasmid used in this process was purchased from Addgene (#41062). The sequences of the altered regions on the genome are shown in Supplementary Table 4. Cloning of plasmids was facilitated by the ClonExpress II One Step Cloning Kit (Vazyme). Plasmid pZA16-*mf*-Lon for expression of *mf*-Lon protease was purchased from Addgene (#75439). The tRNA expression plasmid was derived from the pET28a vector, incorporating the pBR322 origin of replication (which allows for high copy number), the T7 promoter, and the lac operator to regulate the expression of the tRNA of interest, along with the *rrnC* terminator located downstream of the tRNA gene. Detailed information about the full sequence of this transcription unit can be found in Supplementary Table 5.

Cell-free protein synthesis

The genetically modified *E. coli* chassis strains employed in cell-free protein synthesis (CFPS) are detailed in Supplementary Table 1. The protocol for preparing the S30 cell extract utilized in CFPS has been described previously[25]. During the resuspension of the cell pellet, 1 mL of S30 buffer was added per 1.5 g of cell mass. A standard CFPS reaction mixture typically consists of the following components: 40 mM HEPES, 130 mM potassium glutamate (K-Glu), 18 mM magnesium acetate ($Mg(CH_3COO)_2$), 2 mM dithiothreitol (DTT), 1 mM putrescine, 0.34 mM nicotinamide adenine dinucleotide (NAD), 1.5 mM spermidine, 1 mM of each nucleotide (ATP, UTP, CTP, GTP), 0.3 mM coenzyme A, 170 μg/mL *E. coli* tRNA mix, 34 μg/mL calcium folinate, 33.33 mM phosphoenolpyruvate (PEP), 0.5 mM of each of the 20 standard amino acids, 35% v/v cell extract, and linear DNA

templates. Variations in these components were made depending on the specific requirements of the experiment.

Total RNA extraction

The protocol for total tRNA extraction was adapted from a previously established method. tRNA expression plasmids were introduced into *E. coli* BL21 (DE3) cells through transformation. A single colony was inoculated into rich medium (2×YT) and incubated at 37 °C with shaking at 220 rpm for 12 hours. Following the incubation, 100 μL of preculture was transferred to 5 mL of prewarmed 2×YT, and the culture was further incubated at 37 °C with shaking at 220 rpm for 1-2 hours. When the optical density at 600 nm (OD_{600}) reached 0.4~0.5, protein expression was induced by adding 1 mM IPTG. The cells were then pelleted by centrifugation (2 min, >13,000 RCF, 4 °C). The resulting cell pellet was resuspended in 1 mL of TRIzol® reagent and incubated at 25 °C for 5 minutes. Subsequently, 0.2 mL of chloroform was added, followed by vertexing for 15 seconds and a 5-minute incubation at 25 °C. Afterward, the lysed cells were centrifuged (15 min, >13,000 RCF, 4 °C) and the aqueous phase, approximately 500 μL containing the RNA, was transferred to a clean tube. To precipitate the RNA, 500 μL of isopropanol was added, and the solution was mixed gently before being incubated for 5 minutes at 25 °C. The RNA was then pelleted by centrifugation (10 min, >13,000 RCF, 4 °C). The pellet was washed with 1 mL of 75% ice-cold ethanol and centrifuged again (5 min, >15,000 RCF, 4 °C). Finally, the RNA pellet was resuspended in 20~50 μL of RNase-free water. This solution was either immediately used in the CFPS reaction or stored at −80 °C for future use.

12NNK library construction

The DNA library was designed with a range of functional elements: an upstream T7 promoter for transcription initiation, a ribosomal binding site, a Myc tag, a random sequence for screening Streptactin-binding proteins, a spacer to facilitate the extension of the nascent peptide chain within the ribosomal tunnel, a stalling sequence to induce ribosome stalling and form a stable ternary complex, and a reverse transcription sequence. The sequence information of the successfully constructed library is detailed in Supplementary Table 2. Libraries exhibiting substantial diversity were constructed through a one-step PCR amplification process, using primers specified in Supplementary Table 3. High-fidelity DNA polymerase (TransStart® FastPfu DNA Polymerase, TransGen Biotech) was employed for the PCR, and a total of 25 amplification cycles were performed. NGS of the resulting libraries confirmed uniform sequence coverage under the specified amplification conditions. Finally, a ~445 bp fragment was purified by DNA agarose gel extraction to generate the DNA library, which is referred to as the input library.

Ribosome display

The standard CFPS reaction system for RD was set up in a total volume of 100 μL. The purified PCR product was used as the template at a concentration of 30 ng/μL, and total RNA was supplemented to a concentration corresponding to 1.0 A_{260} unit. The reaction was incubated at 30 °C for 2 hours, after which it was transferred to ice to halt the reaction. To terminate the reaction, 1.9 mL of ice-cold binding buffer (BB) was added, which contained 50 mM Tris-acetate (pH 7.5), 150 mM NaCl, 50 mM $Mg(CH_3COO)_2$, and 0.1% (v/v) Tween-20. This RD reaction mixture was subsequently used for binding to immobilized protein targets in the context of in vitro selection. When varying volumes of CFPS reactions were required, the volume of the stop buffer was scaled accordingly. During each selection round, the reaction was prepared in multiples of the standard 1× preparation: the first round used a 5× concentration, the second round used 4×, the third round used 3×, the fourth round used 2×, and the fifth round used the standard 1× preparation.

In vitro selection

ST-MB was used to isolate peptide-ribosome-mRNA ternary complexes that could bind to it. After the recovery of mRNA with multiple washes and reverse transcription to DNA, the sequences of the candidate Streptactin-binding peptides were determined by NGS. The experimental procedure is described as follows: RD solutions were first pre-screened using empty magnetic beads not coated with Streptactin, and after incubation at 30°C for 30 min to remove components that might bind the beads non-specifically, the supernatant was collected for subsequent formal screening. Each standard reaction used 300 μl of ST-MB, and the beads were first washed three times with BB. RD reaction mixture was incubated with ST-MB at 30 °C for 30 minutes, followed by five 5-minute washes with washing buffer (WB) (50 mM Tris-acetate, pH 7.5, 300 mM NaCl, 50 mM Mg $(CH_3COO)_2$, 0.1% (v/v) Tween-20) at 30 °C. After washing, mRNA was isolated from the beads by incubation in 100 μl elution buffer (EB) (50 mM Tris-acetate, pH=7.5, 150 mM NaCl, 1 mM biotin) at 30 °C for 30 min. To eliminate any residual DNA templates, the eluted mRNA was treated with DNase I for 1 hour at 37 °C. The mRNA was purified using an EasyPure RNA Purification Kit (TransGen Biotech) according to supplier's procedure. Reverse transcription of purified RNA was performed using HiScript II 1st Strand cDNA Synthesis Kit (Vazyme) and primer as described in Supplementary Table 3. The reverse transcription product was directly used as a template for PCR amplification. The number of PCR cycles, typically between 5 and 15, was carefully selected to avoid over-amplification. The amplification products were divided into two: samples for NGS and DNA template input for the RD reaction to perform additional rounds of in vitro selection, each requiring specific amplification primers (Supplementary Table 3).

NGS

To prepare sequencing samples, reverse-transcribed cDNA was amplified using the primers T7-FL and NGS-R. The amplification process employed high-fidelity DNA polymerase (TransStart® FastPfu DNA Polymerase, TransGen Biotech). The resulting PCR products were purified through DNA agarose gel extraction and quantified using a Qubit 3 Fluorometer. Subsequently, the ~250 bp amplicon libraries were sequenced by Novogene using an Illumina NovaSeq6000 platform, with 150 bp paired-end reads.

Protein purification

All target proteins were expressed as His-tagged fusions in *E. coli* BL21(DE3). Protein expression was induced with 1 mM IPTG during the logarithmic growth phase in 1 L cultures. After 7 hours of induction, cells were harvested and lysed by sonication (SCIENTZ-IID). For purification, the lysate supernatant was loaded onto an ÄKTA pure 25 system (GE Healthcare) equipped with a 5 mL HisTrap HP Ni-Sepharose column. The column was pre-equilibrated with buffer A (50 mM HEPES-KOH pH 7.6, 1 M NH_4Cl, 10 mM $MgCl_2$, 20 mM imidazole, 7 mM β-mercaptoethanol). Bound proteins were eluted using a linear gradient from 0% to 100% buffer B (50 mM HEPES-KOH pH 7.6, 100 mM KCl, 10 mM $MgCl_2$, 400 mM imidazole, 7 mM β-mercaptoethanol) over 30 min at 1 mL/min, with UV monitoring at 280 nm. The eluted protein was subsequently desalted into storage buffer (50 mM HEPES-KOH, pH 7.6, 100 mM KCl, 30% glycerol, 7 mM β-mercaptoethanol) using a HiTrap Desalting column (5 mL, GE Healthcare). Purified proteins were aliquoted and stored at −80 °C until further use.

Ribosome pull-down assay

Ribosome: template (Supplementary Fig. 4a) was transcribed and translated in 100 μL CFPS reaction, incubated at 30 °C for 2 h. For target isolation, 800 μL of ST-MB were washed twice with 1.6 mL BB and then resuspended to 0.5× concentration, and divided into four aliquots. Reactions were terminated on ice, and 100 μL of each CFPS product was diluted into 1.9 mL BB, followed by incubation with beads at 30 °C for 30 min with rotation. Beads were subsequently washed with 1 mL WB (5 min under rotation). For elution, bound complexes were incubated with 100 μL EB at 30 °C for 30 min in the dark. Eluted samples were mixed with 4× sample buffer (250 mM Tris-HCl, pH 6.8, 10% (w/v) SDS, 8‰ (w/v) bromophenol blue, 40% (v/v) glycerol, 2.86 M β-Mercaptoethanol), denatured at 98 °C for

10 min, and analyzed by SDS-PAGE. A Western blot was performed using an anti-FLAG primary antibody.

Streptactin-binding peptide pull-down assay

For the Streptactin-binding peptide pull-down assay, purified proteins (equal amounts) were incubated with streptavidin-coated magnetic beads in BB The initial fluorescence intensity of the protein-bead mixture was measured (ex/em: 485 nm/ 535 nm) before incubation at 30 °C for 30 min with rotation. Unbound proteins were removed by washing the beads five times with 1 mL WB. Bound complexes were eluted with 100 μL biotin-containing EB at 30 °C for 30 min (protected from light), and the fluorescence of the eluate was measured. The elution efficiency was calculated as (Fluorescence elution / Fluorescence initial). For Western blot analysis, the eluate was mixed with 4× sample buffer, denatured (98 °C, 10 min), resolved by SDS-PAGE, and probed with anti-His antibody.

Western blot analysis

Protein samples were mixed with 4× sample buffer at a 3:1 ratio and denatured at 98 °C for 5 min. Proteins were separated by SDS-PAGE using discontinuous gels consisting of a 5% acrylamide stacking gel and 10% acrylamide resolving gel. Electrophoresis was performed at 80 V through the stacking gel and 120 V through the resolving gel in running buffer (25 mM Tris base, 200 mM glycine, 0.1% SDS). Proteins were transferred to PVDF membranes (10 cm × 7.5 cm) using a wet transfer system at 0.12 A for 1.5 h in transfer buffer (25 mM Tris-HCl pH 8.3, 192 mM glycine, 10% methanol). Membranes were blocked overnight at 4 °C in blocking buffer (5% skim milk, 10 mM Tris-HCl pH 7.4, 157.5 mM NaCl, 0.02% Tween 20). Immunoblotting was performed by incubating with primary antibody (1:1000 dilution in blocking buffer) for 1 h at room temperature, followed by four washes with blocking buffer. Membranes were then incubated with HRP-conjugated secondary antibody (1:5000 in blocking buffer) for 1 h at room temperature. After four washes with stripping buffer (10 mM Tris-HCl pH 7.4, 157.5 mM NaCl, 0.02% Tween 20), protein bands were visualized using enhanced chemiluminescence substrate.

NNK mutant region extraction from NGS raw data

In the first step, FLASH[34] was used for read merging, retaining only the successfully merged reads. In the second step, sequence slicing was performed using the fixed sequences at both the N-terminus and C-terminus of the NNK region, with only the reads that contained both the N-terminal and C-terminal sequences retained. In the third step, sequences containing the degenerate base "N" were identified and excluded, retaining only the reads without "N". The final step involved assessing the length of the mutated region, where only reads of exactly 36nt in length were retained. The total number of unique DNA sequences was determined, representing the DNA types, while the number of reads for each identical sequence was considered its corresponding depth value. Upon translating, due to codon degeneracy, different DNA sequences may translate to the same amino acid sequence. The total number of unique amino acid sequences was counted (total types of amino acid sequences), while the sum of the depth of all DNA sequences for identical amino acid sequences was considered their corresponding depth values.

Machine learning model architecture and training

To model the relationship between protein sequences and their binding activities, we implemented a recurrent neural network (RNN) with a fully connected output layer. Each sequence is first Onehot-encoded. Each amino acid is represented as a binary value among the $H_{in} = 21$ possible symbols (including "*" for stop codon); on every given position in the peptide sequence, the value for the observed amino acid (or stop codon) is set to 1, and all other values on that position are set to 0. Next, the embedded sequence, each element of which is now an H_{in}-dimensional vector, is sent into a classic recurrent neural network. Specifically, an RNN with $N_{layer} = 8$ hidden layers and hidden feature size $H_{out} = 64$ was employed. The hidden layers update whenever a new embedded amino acid is input. The hidden state of the final time step served as a compressed representation of the entire protein sequence, with a total dimension of $N_{layer} * H_{out} = 512$. Following the RNN encoder, a fully connected feedforward neural network was employed to map the extracted sequence features to the predicted activity value. This network consists of two dense layers, with sizes 128 and 32, respectively, each followed by ELU (Exponential Linear Unit) activation, and an output layer with a linear activation function to predict a continuous activity value. Other model structures are also tested (Supplementary Fig. 18).

The model was trained on a training set consisting of 80% sequences chosen using random stratified sampling and their corresponding labels. Hyperparameters were tuned by further splitting the valid sets from the training set and applying k-fold cross-validation ($K = 5$) together with grid search. The remaining 20% sequences served as a testing set and did not participate in the training stage. Huber loss was used as the loss function, and a classic stochastic gradient descent as the optimizer. Dropout layers ($p = 0.1$) were applied after each hidden layer in the MLP during training to prevent overfitting, and batch normalization was employed to alleviate the problems of gradient vanishing and gradient exploding. The learning rate was reduced by multiplying by 0.75 every 80 epochs. The training was conducted using NVIDIA T1000 and NVIDIA RTX 2080 Ti GPU.

Evaluation of sequences

The exhaustive prediction of all possible 7-amino-acid sequences was conducted on an NVIDIA RTX 2080 Ti GPU. Utilizing a model trained on second-round NGS data, we identified 1,018 sequences with predicted activity values > 8.338 from the complete combinatorial space (1.28×10^9). Distances among sequences were defined as the Levenshtein distances[35]. The resultant distance matrix underwent nonlinear dimensionality reduction via t-SNE[36] (t-distributed stochastic neighbor embedding) to generate an interpretable 2D visualization (Fig. 4a, b). To evaluate longer sequences such as the Strep-tag, we employed a sliding window approach to extract all possible contiguous n-mer subsequences, where n is the target length of a model. Each of these n-mer fragments was scored individually using the model, and the maximum score among them was taken as the representative score for the full-length peptide. This approach ensures that the highest-affinity segment within the longer sequence is captured.

Dimensionality reduction and clustering

For dimensionality reduction, three approaches were employed: t-distributed stochastic neighbor embedding (t-SNE), multidimensional scaling (MDS)[37], and isometric mapping (Isomap)[38]. All three methods were implemented in Python using the scikit-learn library with default parameters unless otherwise specified. Following dimensionality reduction, clustering analyses were performed using three algorithms, namely K-means clustering, BIRCH clustering[39], and density-based spatial clustering of applications with noise (DBSCAN)[40]. For K-means and BIRCH, the number of clusters was set as a variable parameter; for DBSCAN, the neighborhood radius (epsilon) and minimum number of samples were adjusted to evaluate cluster stability. In addition, motif detection was conducted using the STREME tool (from the MEME Suite)[41]. To facilitate visualization, scatter plots from dimensionality reduction were colored according to the presence of specific motifs (i.e., HPQG, HPMG, HPQF, HPME, ExWL, and PxWWxxL), with sequences containing the same motif assigned the same color.

Docking

Molecular docking was employed to simulate the binding interactions between Streptactin and three short peptides: HPMGERS, SGLELWL, and PSWWYSL. The crystal structure of Streptactin was retrieved from the RCSB Protein Data Bank (PDB ID: 1KL4)[30]. The three-dimensional structures of the peptides were predicted using AlphaFold3[16], with the sequences listed above.

Prior to docking, all structures were preprocessed by removing water molecules, adding hydrogens, correcting potential amino acid errors, and

adjusting partial charges. Semi-flexible molecular docking was performed using AutoDock Vina[42]. which is based on the AutoDock 4.2 framework. The docking procedure employed the Lamarckian Genetic Algorithm (LGA)[43]. to explore optimal binding conformations.

Binding energies were reported as scoring values calculated by AutoDock Vina's empirical scoring function, which is trained and optimized through machine learning approaches. The general form of this scoring function is as follows:

$$\Delta G_{binding} = \sum w_i \times f_i(interaction) + Constant$$

where w_i denotes the regression-derived weights of individual interaction terms, and f_i represents specific molecular descriptors, such as hydrogen bonding, van der Waals interactions, and electrostatic forces.

To evaluate and visualize the docking outcomes, PyMOL was used to examine the spatial features of the top-ranked binding conformations. Additionally, LigPlot was employed to analyze key molecular interactions between Streptactin and each of the three peptides, including hydrogen bonds and hydrophobic contacts.

Statistics and reproducibility

Fluorescence and growth-curve experiments were performed with n = 3 biological replicates (defined as independent culture batches), and data are reported as mean ± SD. Unless otherwise stated, all statistical tests were two-tailed: two-group comparisons used unpaired student's *t*-tests; significance thresholds were $P < 0.05$ (*) and $P < 0.01$ (**). No pre-specified exclusion criteria were applied. Visualizations were generated in GraphPad Prism v9. Figure legends provide per-experiment details (sample sizes, time points, *p*-values and other notes).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The raw sequencing data is available on Zenodo (10.5281/zenodo.17481122)[44]. Source data behind the graphs in the paper are provided as Supplementary Data. Additional information is available from the corresponding author upon reasonable request.

Code availability

Code of the model is available on Zenodo (10.5281/zenodo.17481122)[44] and github (https://github.com/iiLiniesta/StrepBindingPredict).

Received: 13 May 2025; Accepted: 30 October 2025;
Published online: 02 December 2025

References

1. Hakes, L., Pinney, J. W., Robertson, D. L. & Lovell, S. C. Protein-protein interaction networks and biology—what's the connection?. *Nat. Biotechnol.* **26**, 69–72 (2008).
2. Hoeflich, K. P. & Ikura, M. Calmodulin in action: diversity in target recognition and activation mechanisms. *Cell* **108**, 739–742 (2002).
3. Behrens, J. et al. Functional interaction of β-catenin with the transcription factor LEF-1. *Nature* **382**, 638–642 (1996).
4. Pellarin, I. et al. Cyclin-dependent protein kinases and cell cycle regulation in biology and disease. *Signal Transduct. Target. Ther.* **10**, 11 (2025).
5. Lill, N. L., Grossman, S. R., Ginsberg, D., DeCaprio, J. & Livingston, D. M. Binding and modulation of p53 by p300/CBP coactivators. *Nature* **387**, 823–827 (1997).
6. Stelzl, U. et al. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–968 (2005).
7. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).
8. Zhu, H. et al. Global analysis of protein activities using proteome chips. *Science* **293**, 2101–2105 (2001).
9. Smith, G. P. Filamentous fusion phage: novel expression vectors that display cloned antigens on the Virion surface. *Science* **228**, 1315–1317 (1985).
10. Clackson, T., Hoogenboom, H. R., Griffiths, A. D. & Winter, G. Making antibody fragments using phage display libraries. *Nature* **352**, 624–628 (1991).
11. Boder, E. T., Midelfort, K. S. & Wittrup, K. D. Directed evolution of antibody fragments with monovalent femtomolar antigen-binding affinity. *PNAS* **97**, 10701–10705 (2000).
12. Binz, H. K. et al. High-affinity binders selected from designed ankyrin repeat protein libraries. *Nat. Biotechnol.* **22**, 575–582 (2004).
13. Gan, R. & Jewett, M. C. Evolution of translation initiation sequences using in vitro yeast ribosome display. *Biotechnol. Bioeng.* **113**, 1777–1786 (2016).
14. Philpott, D. N. et al. Rapid On-Cell Selection of High-Performance Human Antibodies. *ACS Cent. Sci.* **8**, 102–109 (2022).
15. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
16. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
17. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
18. Mattheakis, L. C., Bhatt, R. R. & Dower, W. J. An in vitro polysome display system for identifying ligands from very large peptide libraries. *PNAS* **91**, 9022–9026 (1994).
19. Li, R., Kang, G., Hu, M. & Huang, H. Ribosome display: a potent display technology used for selecting and evolving specific binders with desired properties. *Mol. Biotechnol.* **61**, 60–71 (2018).
20. Ueda, T., Kanamori, T. & Ohashi, H. in Cell-Free Protein Production: Methods and Protocols. (eds. Y. Endo, K. Takai & T. Ueda) 219-225 (Humana Press, Totowa, NJ; 2010).
21. Evans, M. S., Ugrinov, K. G., Frese, M. A. & Clark, P. L. Homogeneous stalled ribosome nascent chain complexes produced in vivo or in vitro. *Nat. Methods* **2**, 757–762 (2005).
22. Contreras-Martinez, L. M. & DeLisa, M. P. Intracellular ribosome display via SecM translation arrest as a selection for antibodies with enhanced cytosolic stability. *J. Mol. Biol.* **372**, 513–524 (2007).
23. Hoffmüller, U. & Schneider-Mergener, J. In vitro evolution and selection of proteins: ribosome display for larger libraries. *Angew. Chem.* **37**, 3241–3243 (1998).
24. Chen, X., Gentili, M., Hacohen, N. & Regev, A. A cell-free nanobody engineering platform rapidly generates SARS-CoV-2 neutralizing nanobodies. *Nat. Commun.* **12**, 5506 (2021).
25. Wu, Y. et al. Efficient in vitro full-sense-codons protein synthesis. *Adv. Biol*. **6**, 10 (2022).
26. Elman, J. L. Finding structure in time. *Cogn. Sci.* **14**, 179–211 (1990).
27. Caterini, A. L., Chang, D. E., Caterini, A. L. & Chang, D. E. Recurrent neural networks. In *Deep Neural Networks in a Mathematical Framework, SpringerBriefs in Computer Science*. 59–79 (Springer, Cham, 2018).
28. Teğin, U., Dinç, N. U., Moser, C. & Psaltis, D. Reusability report: Predicting spatiotemporal nonlinear dynamics in multimode fibre optics with a recurrent neural network. *Nat. Mach. Intell.* **3**, 387–391 (2021).
29. Shen, J., Liu, F., Tu, Y. & Tang, C. Finding gene network topologies for given biological function with recurrent neural network. *Nat. Commun.* **12**, 3125 (2021).
30. Korndörfer, I. P. & Skerra, A. Improved affinity of engineered streptavidin for the Strep-tag II peptide is due to a fixed open conformation of the lid-like loop at the binding site. *Protein Sci.* **11**, 883–893 (2009).

31. Schmidt, T. G. M. et al. The role of changing loop conformations in streptavidin versions engineered for high-affinity binding of the Strep-tag II peptide. *J. Mol. Biol*. **433**, 9 (2021).
32. Erlich, K. R., Baumann, F., Pippig, D. A. & Gaub, H. E. Strep-Tag II and monovalent strep-tactin as novel handles in single-molecule cut-and-paste. *Small Methods* **1**, 1700169 (2017).
33. Shimizu, Y. et al. Cell-free translation reconstituted with purified components. *Nat. Biotechnol.* **19**, 751–755 (2001).
34. Magoč, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
35. Levenshtein, V. I. In Soviet physics doklady, Vol. 10, 707–710 (Soviet Union, 1966).
36. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
37. Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* **29**, 1–27 (1964).
38. Tenenbaum, J. B., Silva, V. D. & Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science* **290**, 2319–2323 (2000).
39. Zhang, T., Ramakrishnan, R. & Livny, M. BIRCH: an efficient data clustering method for very large databases. *SIGMOD Rec.* **25**, 103–114 (1996).
40. Schubert, E., Sander, J., Ester, M., Kriegel, H. P. & Xu, X. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Trans. Database Syst.* **42**, 19 (2017).
41. Bailey, T. L. STREME: accurate and versatile sequence motif discovery. *Bioinformatics* **37**, 2834–2840 (2021).
42. Pagadala, N. S., Syed, K. & Tuszynski, J. Software for molecular docking: a review. *Biophys. Rev.* **9**, 91–102 (2017).
43. Bitencourt-Ferreira, G., Pintro, V. O. & de Azevedo W. F. Jr. In Docking screens for drug discovery 125–148 (Springer, 2019).
44. Tang, M., Li, J. & Qi, H. Streptactin Binding Plasticity NGS from deep Ribosome Display. *Zenodo* https://doi.org/10.5281/zenodo.17481122 (2025).

Acknowledgements

We thank Jiaxi Lu and Juncheng Zhou at School of Cyber Science and Engineering, Wuhan University, for the constructive discussions.

Author contributions

Mengtong Tang designed and performed biochemical experiments and assisted with data analysis. Jiawei Li wrote the code and conducted all of the bioinformatic analyses, model training, and computational experiments. Zhixi Li performed the analyses of molecular docking. Jingsong Cui supervised the deep learning model development and computational experiments. Hao Qi supervised the whole project and led the development and designing of algorithms and experiments. Mengtong Tang, Jiawei Li and Hao Qi prepared the manuscript.

Competing interests

The authors declare the following competing interests: H.Q. and M.T. are listed as inventors of a patent application (CN202510273593) for peptide screening and cell free protein synthesis, related to work in this manuscript. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42003-025-09160-y.

Correspondence and requests for materials should be addressed to Jingsong Cui or Hao Qi.

Peer review information *Communications Biology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editors: Michal Kolar and Laura Rodríguez.