Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.07.22.000293. This version posted July 22, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

ASPIRE improves protein binder prioritization through cross-target priors and sparse preference feedback

Active Selection with Prior-Informed Ranking for Experiments (ASPIRE)

Mingqian Chen[1,2,#], Liping Huang[3,#,*], Yuxue Guo[3], Yue Peng[3], Yihui Yang[2], Yuhong Li[2], Wen Li[2], Zhonghua Liu[1,*], Gang Logan Liu[2,4,*], Wenjun Hu[2,*]

[1] The National & Local Joint Engineering Laboratory of Animal Peptide Drug Development, College of Life Sciences, Hunan Normal University, Changsha, Hunan, PR China.

[2] State Key Laboratory for Diagnosis and Treatment of Severe Zoonotic Infectious Diseases, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, PR China.

[3] School of Food Science and Pharmaceutical Engineering, Nanjing Normal University, Nanjing, Jiangsu, PR China.

[4] Liangzhun Life Science Company, Wuhan, Hubei, PR China.

# These authors contributed equally. * Correspondence: Liping Huang(lphuang@nnu.edu.cn), Zhonghua Liu(liuzh@hunnu.edu.cn), Gang Logan Liu(loganliu@hust.edu.cn) and Wenjun Hu(hu_wenjun@hust.edu.cn).

Abstract

Generative protein design can produce thousands to hundreds of thousands of candidates for a single target, making accurate experimental prioritization a central bottleneck. Existing protein Bayesian optimization and active-learning workflows generally require initial measurements from the current target and therefore retain a cold-start problem. Here, we introduce Active Selection with Prior-Informed Ranking for Experiments (ASPIRE), which combines a transferable cross-target ranking prior with sparse within-target preference updates.

ASPIRE learns target-binder ranking from multi-target interaction data and assigns candidate priorities before target-specific measurements are available. Experimental outcomes are converted into preference relations within the same target and assay method, enabling subsequent ranking updates. Across four parallel protein landscapes containing 2,000-149,361 candidates, first-round ASPIRE selection identified candidates within the top 0.055%-3.54% of the true landscape using batches of four or ten. Iterative updates further improved candidate ranks and outperformed random-initialized Gaussian-process, prior-initialized Gaussian-process, and random-initialized Bradley-Terry/Laplace controls.

In a prospective ASPIRE-guided VISTA-binding peptide campaign, four of eight synthesized candidates produced clear binding signals. A ninth lead advanced to kinetic validation yielded an affinity range of 1.62-2.31 pM. ASPIRE therefore improves protein-binder prioritization from the first experimental round, with reduced experimental burden emerging as a practical consequence of more accurate selection.

Introduction

RFdiffusion, ProteinMPNN, and related generative and sequence-design methods have greatly increased the scale and structural success of protein-binder generation [1,2]. Target-structure-conditioned design, learned surface fingerprints, and one-shot binder design further expand the types and numbers of candidates that can be produced [3–5], and de novo design has yielded picomolar miniprotein binders in selected systems [6]. These methods can rapidly generate large structure and sequence pools and filter them using confidence, interface geometry, or energy-based criteria. However, generating many computationally plausible candidates does not establish which candidates should enter experiments first. Scores from different generators lack a common experimental scale, and performance on one

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.07.22.000293. This version posted July 22, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0
https://creativecommons.org/licenses/by-nc-nd/4.0

target does not imply transferable ranking on another. As candidate libraries grow from tens to thousands or hundreds of thousands of sequences, the gap between library size and experimental budget becomes a primary bottleneck in binder discovery.

Experiments provide more direct binding evidence, but labels from different platforms are not naturally commensurate. Surface plasmon resonance (SPR), WeSPR, and biolayer interferometry record association and dissociation behavior [7,8]; ELISA and other fixed-concentration plate assays provide relative binding signals [9]; display, enrichment, and deep mutational scanning provide population enrichment or variant effects [10]; and early screens may only distinguish positive from negative candidates. Meta-SPR and WeSPR platforms developed by Liangzhun Life Science reduce the cost of molecular-interaction measurement and support multi-region and kinetic analysis through imaging-based readout [11–13]. Although these measurements use different units, each can provide a relative order among candidates measured for the same target under the same assay method.

Gaussian processes, Bayesian optimization, machine-learning-assisted directed evolution, and low-N learning have shown that limited experiments can guide subsequent protein selection [14–20]. Self-driving laboratories and active-learning-assisted directed evolution further connect prediction, synthesis, measurement, and retraining into closed loops [21–23]. Yet these optimizations usually begin within the current target: an initial batch is measured, a target-specific surrogate is fitted, and the next candidates are selected within the same sequence landscape or experimental system. If a new target has no target-specific labels and the first-round budget is only four to ten candidates, this target-internal paradigm retains a decisive cold-start problem.

ASPIRE addresses this gap by connecting cross-target learning to within-target optimization. A transferable target-binder representation and ranking prior are first learned from external interaction data, allowing a new target to receive first-round candidate priorities without experimental labels. Quantitative, ordinal, or binary outcomes are then converted into preference relations only within that target and assay method, and a Bradley-Terry/Laplace update refines the next batch [24,25]. Compared with target-internal optimization that begins from random measurements on each new target, ASPIRE moves the available assay budget toward more valuable candidates from the first round while retaining the ability to correct the prior with target-specific evidence.

We constructed and evaluated this screening paradigm in three stages. We first established ESMC encoding, low-rank bilinear target-binder fusion, and an updatable ranking layer, and compared pair representations and training objectives. We then evaluated zero-label cold starts, iterative updates, library size, and batch size on GB1 and three BPTI-protease landscapes, using a 2 × 2 factorial design to separate cross-target prior and within-target update effects. Finally, we used ASPIRE to select two small experimental batches in a prospective VISTA binder campaign and advanced a lead to WeSPR Imager kinetic validation. ASPIRE thus converts post-generation screening from repeated random starts into a process that inherits cross-target experience and adapts with very limited target-specific experiments.

Results

ASPIRE reframes binder discovery as low-budget experimental prioritization

Candidate generation and experimental validation operate at different scales (Fig. 1A). The generated or combinatorial libraries studied here contained 10^3–10^5 sequences, whereas one experimental round may measure only 10^0–10^2 candidates. High-fitness variants occupied a narrow region of the landscape: in GB1, only 3.3% of measured variants exceeded 10% of the maximum observed fitness (Fig. 1B). A small random batch therefore cannot reliably cover the high-value region.

ASPIRE inserts an updatable selection loop between the candidate library and the assay (Fig. 1C). For a fixed target, $s_i^{(0)}$ first ranks the full library. A small batch is synthesized or measured, and quantitative, ordinal, or binary outcomes are converted into preference relations such as $A \succ B$. These relations update the candidate score distribution before the next batch is selected. ASPIRE outputs the candidates most worth testing next rather than claiming a universally calibrated absolute affinity for every sequence.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.07.22.000293. This version posted July 22, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

A target-binder ranking model supplies a transferable prior and an inexpensive update state

ASPIRE represents the target and candidate binders using ESMC mean-pooled embeddings (Fig. 2A). Independent target and binder projectors produce compact vectors. A low-rank bilinear interaction, the projected vectors, their absolute difference, and their signed difference are processed by an MLP to produce a 64-dimensional pair representation φ_i, denoted h_pair in the implementation diagram. The final Bayesian linear ranking layer defines

$$s_i = \varphi_i{}^T w, \quad \sigma_i{}^2 = \varphi_i{}^T \Lambda^{-1} \varphi_i$$

where w is the posterior mean of the final-layer weights and Λ is a diagonal posterior precision matrix. Candidate i is preferred to candidate j according to a Bradley-Terry model with a Laplace posterior approximation [24,25]:

$$P(i \succ j) = sigmoid(s_i - s_j)$$

The network that produces φ_i remains frozen during a benchmark or experimental campaign. Preference feedback updates w and Λ, thereby changing candidate scores and posterior uncertainty without end-to-end fine-tuning of the ESMC encoder or pair representation.

To harmonize heterogeneous training records, comparisons were generated only among candidates sharing a target and an assay or measurement method (Fig. 2B). Kinetic measurements, plate-based binding signals, enrichment scores, and binary calls can therefore contribute ordering evidence without assuming that their raw values occupy a common absolute scale. The first trainable data freeze contained 4,728 primary SKEMPI interaction pairs [26] and controlled antibody-antigen auxiliary subsets sampled from 75,209 embedding-ready AbBiBench pairs [30]. Missing mean-pooled embeddings were completed for 100,846 antibody binders and two antigens, with no failed or over-length sequences.

Architecture ablations supported mean pooling with low-rank bilinear fusion over the earlier mean-pool baseline, token-level attention, and a VBLL-lite alternative (Fig. 2C). In three-seed objective comparisons, ranking-only training achieved the strongest held-out discrimination among the tested objectives (Fig. 2D). The model reached validation AUROC 0.889 and test AUROC 0.766 on SKEMPI, while avoiding cross-database preference pairs and target-specific heads.

The learned prior enriches high-ranking candidates before target-specific feedback

We evaluated four parallel tasks: the GB1 combinatorial fitness landscape; bovine trypsin (BT) against BPTI variants; chymotrypsin (ChT) against BPTI variants; and human trypsin-3 (PRSS3; historically mesotrypsin, MT) against BPTI variants. GB1 covers joint variants at four positions in the Protein G B1 domain [27]. The three BPTI libraries derive from interface-variant interaction landscapes against distinct protease targets [28]. Each target-library pair was treated as an independent ranking task.

At round 0, the deterministic ASPIRE prior selected candidates with best true rank percentiles ranging from 0.055% to 3.54% across targets and batch sizes (Fig. 3A). Increasing the batch from four to ten improved the best selected rank in all four tasks, as expected. Relative to random selection, the prior improved the best selected rank across the four targets (Fig. 3B). In GB1, this advantage persisted across library sizes of 2,000, 10,000, 50,000, and 149,361 candidates (Fig. 3C), showing that the prior benefit was not restricted to a small subsample.

The ASPIRE prior was not equivalent to a raw protein-language-model score. We compared it with static ESMC masked-marginal likelihood-ratio (LLR) ranking (Fig. 3D). Values above one favor the ASPIRE prior and values below one favor static LLR. ASPIRE strongly outperformed static LLR on GB1, whereas LLR itself was informative on several BPTI mutation landscapes. Together, these comparisons show that the ASPIRE prior is distinct from static ESMC-LLR. These results support target–binder ranking training as an effective strategy for transforming general protein representations into experimentally actionable candidate priorities.

Prior initialization and preference updates jointly improve closed-loop selection

We next separated initialization from update mechanism in a 2 x 2 design: random or ASPIRE-prior initialization, followed by GP regression or Bradley-Terry/Laplace (BT) updates. In the largest GB1 library ($N=149,361$, batch size four), random initialization began at an average best rank percentile of approximately 17.6% (Fig. 4A). Prior initialization began at 0.802%. After four rounds, ASPIRE (prior + BT) reached 0.0355%, compared with 0.203% for

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.07.22.000293. This version posted July 22, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0
https://creativecommons.org/licenses/by-nc-nd/4.0

prior + GP, 0.225% for random + BT, and 0.775% for random + GP. Random + BT showed rapid recovery after the first round, but the prior preserved a substantial advantage under the same total budget.

Across GB1, BT, ChT, and MT at batch size four, ASPIRE produced the lowest mean final rank percentile in each task, with a tie between ASPIRE and prior + GP for ChT (Fig. 4B). The same pattern was not attributable only to the largest GB1 library: across four GB1 library sizes, prior + BT retained the best mean final rank (Fig. 4C). Seed-paired comparisons quantified the magnitude and consistency of these differences (Fig. 4D). Heatmap color represents the mean difference in final rank percentage points, calculated as ASPIRE minus the comparator; negative values therefore favor ASPIRE. Stars report one-sided paired Wilcoxon tests on 100 matched simulation seeds. Effect sizes and paired distributions, rather than star counts alone, are the primary evidence.

These results support a two-component interpretation. The transferable prior moves the initial batch into a better region of the landscape. BT/Laplace updates then use sparse relative feedback to refine priorities without fitting an unrestricted target-specific regression model. Neither component alone reproduced the full behavior in every task.

Mechanism analyses reveal target-dependent encoder priors and GP instability

Figure 5 uses control experiments to compare generic encoder priors with the target-conditioned prior and to examine GP behavior across tasks and initialization conditions.

Raw ESMC-LLR was a failed-sanity negative control on GB1 but not on all targets (Fig. 5B). In the batch-four comparison, static LLR selected a final best rank percentile of 19.0% in GB1, whereas adding BT/Laplace feedback improved it to 0.225%. For BT and ChT, static LLR was already informative and the same update did not improve it; MT was essentially unchanged. This result makes two points. First, ASPIRE's trained prior cannot be reduced to static language-model likelihood. Second, ranking feedback can rescue a poor initialization, but it is not guaranteed to improve a prior that is already aligned with the target landscape.

GP behavior was also target dependent (Fig. 5C). Under the predefined diagnostic, random-initialized GP collapse rates were 3%, 54%, 35%, and 21% for GB1, BT, ChT, and MT, respectively; prior-initialized GP rates were 0%, 17%, 99%, and 12%. The diagnostic combined poor final selection (best rank fraction above 5%) with non-finite or highly unstable posterior uncertainty. Prior initialization therefore did not universally stabilize GP regression and was associated with nearly complete collapse on ChT. It shows that GP regression is sensitive to target and initialization, whereas the ranking-only update provides a more controlled interface to ordinal evidence.

A prospective VISTA campaign links ASPIRE to experimental binder selection

We applied ASPIRE prospectively to select VISTA-binding peptides. VISTA is an immune-checkpoint protein associated with tumor immunosuppression and microenvironment-dependent ligand recognition [29]. Candidates were supplied by an external ODesign generation workflow [33], whereas ASPIRE performed first-round ranking and selected the next batch after experimental feedback (Fig. 6A). Four candidates were synthesized in each of two rounds. Binding signals were measured in triplicate at a final peptide concentration of 10 μM in PBS at room temperature using a Liangzhun WeSPR300 instrument, and PBS blank responses were subtracted. Four of the eight round-specific candidates produced clear binding signals (Fig. 6B). Within-round experimental orders were converted into preference relations for the next ASPIRE update.

A ninth candidate was advanced as the lead and characterized using a Liangzhun WeSPR Imager and a carboxylated LifeDisc MetaSPR chip. VISTA was immobilized using the same method as in the fixed-concentration screen, and peptide was delivered in the flow phase. Sensorgrams showed concentration-dependent association followed by dissociation after approximately 1,000 s (Fig. 6C). Kinetic analysis used the free-R_{max} method described for imaging-based Meta-SPR [13], yielding a fitted affinity range of 1.62–2.31 pM. This prospective campaign translated cross-target cold-start ranking and small-batch adaptation into experimental selection of a picomolar-range VISTA binder.

Discussion

Large-scale generative models have greatly increased the supply of binder candidates, but selection still determines whether these candidates can be converted efficiently into experimental hits. Protein active learning and Bayesian optimization established that a small number of target-specific measurements can guide later experiments

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.07.22.000293. This version posted July 22, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0
https://creativecommons.org/licenses/by-nc-nd/4.0

[14,15]. ASPIRE extends this principle to the start of a new target. Rather than spending the first batch on an uninformed target-specific sample, it inherits ranking ability from cross-target interaction data and concentrates the limited budget on more valuable candidates from the first experiment. Once target-specific feedback is available, the selector transitions from cross-target transfer to target-specific adaptation.

This continuous transition from cross-target learning to within-target optimization is the central innovation of ASPIRE. GB1 and the three BPTI-protease landscapes showed that the learned prior enriched high-ranking candidates before any target-specific label was observed. The 2 × 2 factorial analysis further separated the cold-start contribution of prior initialization from the target-adaptation contribution of the Bradley-Terry/Laplace update. Random initialization followed by preference updating recovered rapidly as measurements accumulated, but within the same severe total budget it did not fully erase the loss incurred by the first random batch. First-round ranking is therefore a determining component of low-budget screening rather than a preliminary detail before closed-loop optimization.

ASPIRE also avoids dependence on a single generic score by combining target-conditioned representation with within-target preferences. Static ESMC likelihood-ratio rankings changed direction across tasks, demonstrating that generic protein-language-model scores are strongly landscape dependent [31,32]. ASPIRE extracts transferable information through target-binder ranking training and then permits evidence from the new target to correct the ordering. GP behavior was likewise sensitive to task and initialization under the current small-batch protocol. Ordinal updates instead use the relative order that experiments can provide consistently across heterogeneous readouts. The method changes the screening paradigm rather than simply exchanging one regressor or acquisition function for another.

The prospective VISTA campaign demonstrates that the same framework can act on a real generated candidate pool. With two rounds of four peptides, ASPIRE advanced multiple binding candidates and identified a lead with picomolar-range affinity under free-R_max fitting. Cross-target cold starts in computational landscapes, sparse within-target adaptation, and prospective experimental advancement together establish ASPIRE as a low-budget binder-screening framework.

Methods

Problem formulation

For a target T, a candidate library B, and a batch budget b, the objective is to select candidates whose unknown target-specific assay utility ranks near the top of the full library. Performance was evaluated with the best rank fraction after each round,

$$r_t = rank(best\ candidate\ observed\ by\ round\ t) / N$$

where rank one is best and lower values are better. Figures report $100r\_t$ as rank percentile unless otherwise stated. The protocol used round 0 plus four update rounds and batch sizes of four or ten.

Training data and frozen data release

The primary interaction-training source was SKEMPI 2.0 [26]. Records were grouped by target complex, and pairwise preferences were generated only among related candidates. AbBiBench served as a controlled antibody-antigen auxiliary source [30]. Its embedding-ready pool contained 75,209 pairwise records and 1,039 unique binder hashes; subsets of 500, 1,000, and 1,200 pairs were prepared during data freezing. FLIP GB1 single-sequence pairs were retained in a separate auxiliary table.

All protein sequences were normalized and indexed by sequence hashes. Missing ESMC mean-pooled embeddings were computed for 100,846 antibody binders and two target antigens (13,106,705 residues in total). No sequences failed and none exceeded the configured length gate. The resulting unified embedding index and training tables formed the first data freeze marked ready for model training.

Pair representation and prior training

Target and binder sequences were encoded using ESM Cambrian and mean pooled over residues [31,32]. Independent learned projectors produced r_T and $r_{B,i}$. Pair features comprised the projected vectors, a low-rank bilinear

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.07.22.000293. This version posted July 22, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

interaction, their absolute difference, and their signed difference. An MLP produced the 64-dimensional representation φ_i.

The model was trained with within-target Bradley-Terry ranking losses. The selected ranking-only configuration used 100 epochs, Adam optimization with learning rate 10^{-4}, weight decay 10^{-4}, SKEMPI and antibody batch sizes of 256 and 128, and random seed 42. Temperature parameters were 0.2 for SKEMPI and 0.3 for antibody data. Model selection used validation AUROC. Alternative objectives included ΔK_D regression, weighted pairwise training, group-centered K_D regression, and combined pairwise plus group-K_D training.

Bayesian ranking layer and Laplace update

For a frozen candidate representation φ_i, the final ranking score was linear in weights w. The prior checkpoint initialized a diagonal Gaussian posterior over w, represented by a posterior mean and diagonal precision Λ. Given cumulative preferences $P_t = i \succ j$, the posterior mode was updated by maximizing the sum of the Gaussian log prior and Bradley-Terry log likelihoods,

$$\log p(w \mid P_t) \propto -\tfrac{1}{2}(w - w_0)^T \Lambda_0 (w - w_0) + \Sigma_{(i,j)\in P_t} \log \mathrm{sigmoid}(\varphi_i^T w - \varphi_j^T w)$$

A diagonal Laplace approximation at the mode produced the updated precision. Candidate means and standard deviations were then computed as $s_i = \varphi_i^T w$ and $\sigma_i = \left(\varphi_i^T \Lambda^{-1} \varphi_i\right)^{1/2}$.

Batch selection

Candidates were ranked by an upper-confidence acquisition score $a_{i,t} = s_{i,t} + \beta \sigma_{i,t}$. The benchmark scanned $\beta \in \{0, 0.5, 1, 2\}$ for non-random method families and reported the pre-specified best summarized setting for each cell. The main ASPIRE protocol used continuous prior scores without forced diversity bins. Additional ablations retained only prior rank, applied diversity bins in the first round, or applied them in every round.

Benchmark landscapes

The GB1 benchmark derived from a complete combinatorial landscape of four jointly varied residues in the Protein G B1 domain [27]. After quality filtering, the largest library contained 149,361 variants; nested libraries of 2,000, 10,000, and 50,000 candidates were used to evaluate library-size effects. Fitness values served only as hidden benchmark truth for simulation and evaluation.

The BT/BPTI, ChT/BPTI, and trypsin-3/BPTI benchmarks derived from three homologous protein-protein interaction landscapes reported by Heyne and colleagues [28]. The effective libraries contained 13,341, 12,765, and 13,607 candidates, respectively. Each target-library pair was treated as an independent ranking task.

Baselines and factorial protocols

Random selection sampled candidates without a prior or update model. GP controls used the same 64-dimensional pair representation (GP_{phi64}) to reduce representation confounding. The factorial analysis compared random + GP, prior + GP, random + BT, and prior + BT (ASPIRE). All protocols used the same library, seed schedule, number of rounds, and batch budget within a cell.

Static ESMC-LLR controls ranked candidates with masked-marginal likelihood ratios and did not use target-binder ranking training. ESMC-LLR + BT used the same initial LLR ordering followed by the ASPIRE last-layer preference update. This control tests whether ranking feedback can recover from a poor generic encoder prior; it does not represent another trained ASPIRE model.

Metrics and statistical analysis

The primary metric was final best rank fraction or percentile; lower is better. Cold-start comparisons used the best true rank among round-0 selections and ratios of comparator rank to ASPIRE-prior rank; ratios above one favor ASPIRE. Simulation summaries used 100 matched seeds unless a panel explicitly showed deterministic prior values. Error bars are standard deviations across seeds. Seed-paired differences were tested with one-sided Wilcoxon signed-rank tests using the alternative that ASPIRE had a lower final rank fraction. Heatmap symbols denote $**** p < 10^{-4}$,$*** p < 10^{-3}$,$** p < 10^{-2}$,$* p < 0.05$, and ns otherwise. No multiplicity correction was applied to the displayed exploratory comparison grid.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.07.22.000293. This version posted July 22, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

GP collapse was defined by a pre-specified compound diagnostic: final best rank fraction above 0.05, non-finite posterior moments, or unstable posterior uncertainty, including normalized variance above 0.1. Collapse rates were calculated across 100 seeds. This metric describes protocol failure in the current benchmark and is not a general property of GP models.

VISTA peptide generation and screening

Candidate peptides were produced by an external ODesign workflow [33], and ASPIRE performed initial ranking and iterative selection. Four peptides were synthesized in round 1 and four in round 2. Binding signals were measured in triplicate using a Liangzhun WeSPR300 instrument at a final peptide concentration of 10 μM in PBS at room temperature. VISTA was immobilized on a carboxylated LifeDisc, and PBS blank responses were subtracted. Within each round, candidate pairs separated by more than 2 RU were converted into preference relations for the next update.

Lead kinetics were measured using a Liangzhun WeSPR Imager. VISTA was immobilized on a carboxylated LifeDisc using the same method, and peptide was supplied in the flow phase at 0, 10, 25, 40, 55, 70, 85, and 100 μM. Association was monitored for approximately 1,000 s and followed by a dissociation phase. Kinetics were fitted with the free-R_{max} method, which allows concentration-specific R_max values to reduce bias from unknown ligand density and mass-transport limitation in high-affinity interactions [13]. The fitted affinity range was 1.62–2.31 pM.

Data and code availability

Data and source code supporting this study will be released with the preprint.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (grant nos. 32471472 and 82072735); the Fundamental Research Funds for the Central Universities (grant nos. 2020kfyXJJS113 and YCJJ20251406); the State Key Laboratory for the Diagnosis and Treatment of Severe Zoonotic Infectious Diseases (grant nos. 2024ZZ10007 and 2024ZZ00019); and the Major Program (JD) of Hubei Province (grant no. 2023BAA011). The authors thank contributors from the Gang Logan Liu laboratory and Liangzhun Life Science for support with Meta-SPR/WeSPR measurements.

Competing interests

The authors declare no competing interests.

References

1. Watson, J. L. et al. De novo design of protein structure and function with RFdiffusion. Nature 620, 1089–1100 (2023).

2. Dauparas, J. et al. Robust deep learning–based protein sequence design using ProteinMPNN. Science 378, 49–56 (2022).

3. Cao, L. et al. Design of protein-binding proteins from the target structure alone. Nature 605, 551–560 (2022).

4. Gainza, P. et al. De novo design of protein interactions with learned surface fingerprints. Nature 617, 176–184 (2023).

5. Pacesa, M. et al. One-shot design of functional protein binders with BindCraft. Nature 646, 483–492 (2025).

6. Cao, L. et al. De novo design of picomolar SARS-CoV-2 miniprotein inhibitors. Science 370, 426–431 (2020).

7. Homola, J. Surface Plasmon Resonance Sensors for Detection of Chemical and Biological Species. Chemical Reviews 108, 462–493 (2008).

8. Abdiche, Y., Malashock, D., Pinkerton, A. & Pons, J. Determining kinetics and affinities of protein interactions using a parallel real-time label-free biosensor, the Octet. Analytical Biochemistry 377, 209–217 (2008).

9. Engvall, E. & Perlmann, P. Enzyme-linked immunosorbent assay (ELISA) quantitative assay of immunoglobulin G. Immunochemistry 8, 871–874 (1971).

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.07.22.000293. This version posted July 22, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

10. Fowler, D. M. & Fields, S. Deep mutational scanning: A new style of protein science. Nature Methods 11, 801–807 (2014).

11. Fan, H. et al. A nanoplasmonic portable molecular interaction platform for high-throughput drug screening. Advanced Functional Materials 32, 2203635 (2022).

12. Chen, M. et al. A Low-Cost Biomolecular Detection Platform Integrating Meta-SPR with Industrial Machine Vision for Affinity Analysis. ACS Sensors 10, 8628–8639 (2025).

13. Chen, M. et al. Direct Visualization and Quantitative Modeling of Mass Transport Limitation in Biomolecular Kinetics via Imaging-Based Meta-SPR. Analytical Chemistry 98, 8382–8393 (2026).

14. Romero, P. A., Krause, A. & Arnold, F. H. Navigating the Protein Fitness Landscape with Gaussian Processes. Proceedings of the National Academy of Sciences 110, E193–E201 (2013).

15. Hie, B., Bryson, B. D. & Berger, B. Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design. Cell Systems 11, 461–477.e9 (2020).

16. Cheng, L., Yang, Z., Hsieh, C., Liao, B. & Zhang, S. ODBO: Bayesian Optimization with Search Space Prescreening for Directed Protein Evolution. Arxiv https://doi.org/10.48550/arXiv.2205.09548 (2022) doi:10.48550/arXiv.2205.09548.

17. Notin, P., Weitzman, R., Marks, D. & Gal, Y. ProteinNPT: Improving Protein Property Prediction and Design with Non-Parametric Transformers. in Advances in Neural Information Processing Systems 36 (2023).

18. Wu, Z., Kan, S. B. J., Lewis, R. D., Wittmann, B. J. & Arnold, F. H. Machine learning-assisted directed protein evolution with combinatorial libraries. Proceedings of the National Academy of Sciences 116, 8852–8858 (2019).

19. Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. Nature Methods 16, 687–694 (2019).

20. Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M. & Church, G. M. Low-N protein engineering with data-efficient deep learning. Nature Methods 18, 389–396 (2021).

21. Rapp, J. T., Bremer, B. J. & Romero, P. A. Self-driving Laboratories to Autonomously Navigate the Protein Fitness Landscape. Nature Chemical Engineering 1, 97–107 (2024).

22. Yang, J. et al. Active Learning-assisted Directed Evolution. Nature Communications 16, 714 (2025).

23. Huot, M., Wang, D., Liu, J. & Shakhnovich, E. I. Predicting High-fitness Viral Protein Variants with Bayesian Active Learning and Biophysics. Proceedings of the National Academy of Sciences 122, e2503742122 (2025).

24. Bradley, R. A. & Terry, M. E. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. Biometrika 39, 324 (1952).

25. MacKay, D. J. C. A Practical Bayesian Framework for Backpropagation Networks. Neural Computation 4, 448–472 (1992).

26. Jankauskaite, J., Jimenez-Garcia, B., Dapkunas, J., Fernandez-Recio, J. & Moal, I. H. SKEMPI 2.0: An Updated Benchmark of Changes in Protein–Protein Binding Energy, Kinetics and Thermodynamics upon Mutation. Bioinformatics 35, 462–469 (2019).

27. Wu, N. C., Dai, L., Olson, C. A., Lloyd-Smith, J. O. & Sun, R. Adaptation in protein fitness landscapes is facilitated by indirect paths. eLife 5, (2016).

28. Heyne, M. et al. Climbing Up and Down Binding Landscapes through Deep Mutational Scanning of Three Homologous Protein–Protein Complexes. Journal of the American Chemical Society 143, 17261–17275 (2021).

29. Johnston, R. J. et al. VISTA is an acidic pH-selective ligand for PSGL-1. Nature 574, 565–570 (2019).

30. Zhao, X. et al. Benchmark for Antibody Binding Affinity Maturation and Design. arXiv https://arxiv.org/abs/2506.04235 (2025).

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.07.22.000293. This version posted July 22, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

31. Lin, Z. et al. Evolutionaryscale/esm: v3.2.3. (2025) doi:10.5281/ZENODO.14219303.

32. Hayes, T. et al. Simulating 500 million years of evolution with a language model. Science 387, 850–858 (2025).

33. Zhang, O. et al. ODesign: A World Model for Biomolecular Interaction Design. (2025) doi:10.48550/arXiv.2510.22304.

Figures and legends

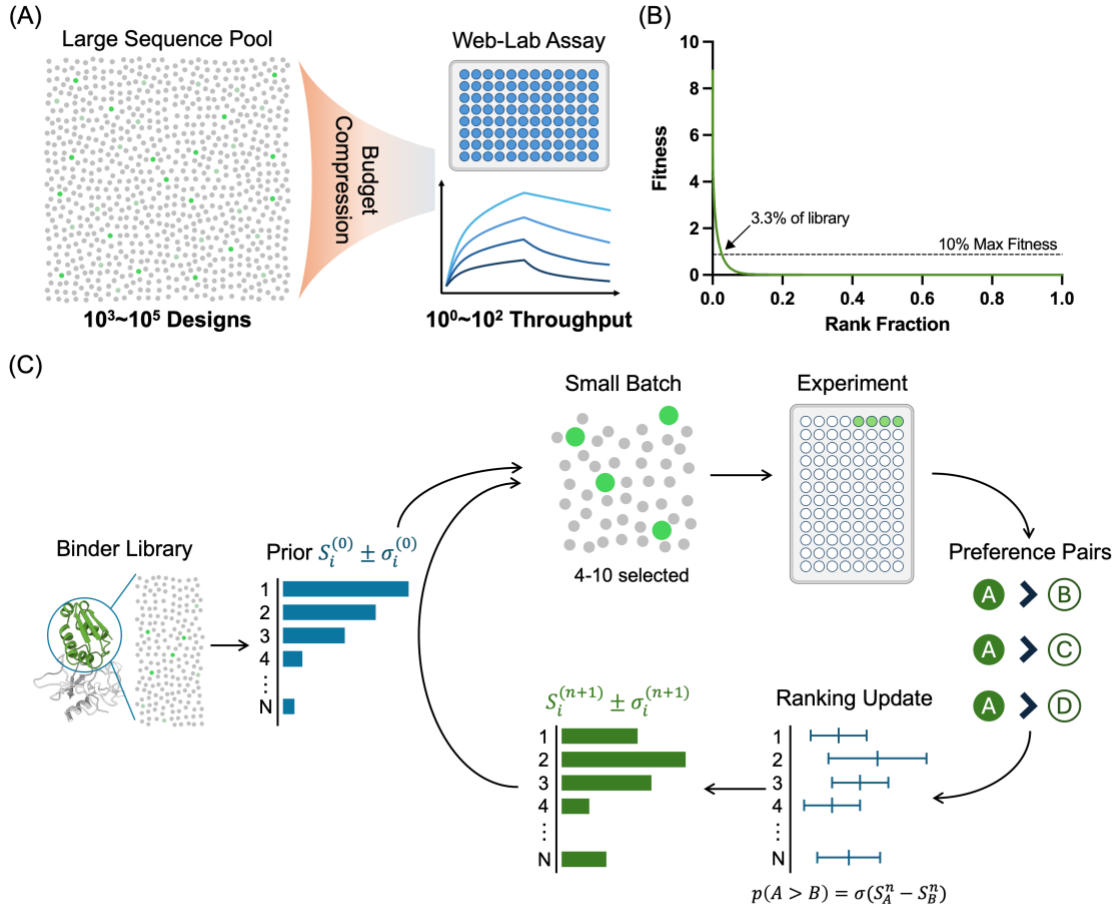

Figure 1 | ASPIRE connects large candidate libraries to small-batch experimental decisions. (A) Orders-of-magnitude gap between candidate-library size and assay throughput. (B) GB1 fitness ordered by true rank fraction; 3.3% of variants exceed 10% of maximum fitness. (C) Prior scoring, small-batch measurement, preference-pair construction, and Bradley-Terry/Laplace ranking update.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.07.22.000293. This version posted July 22, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

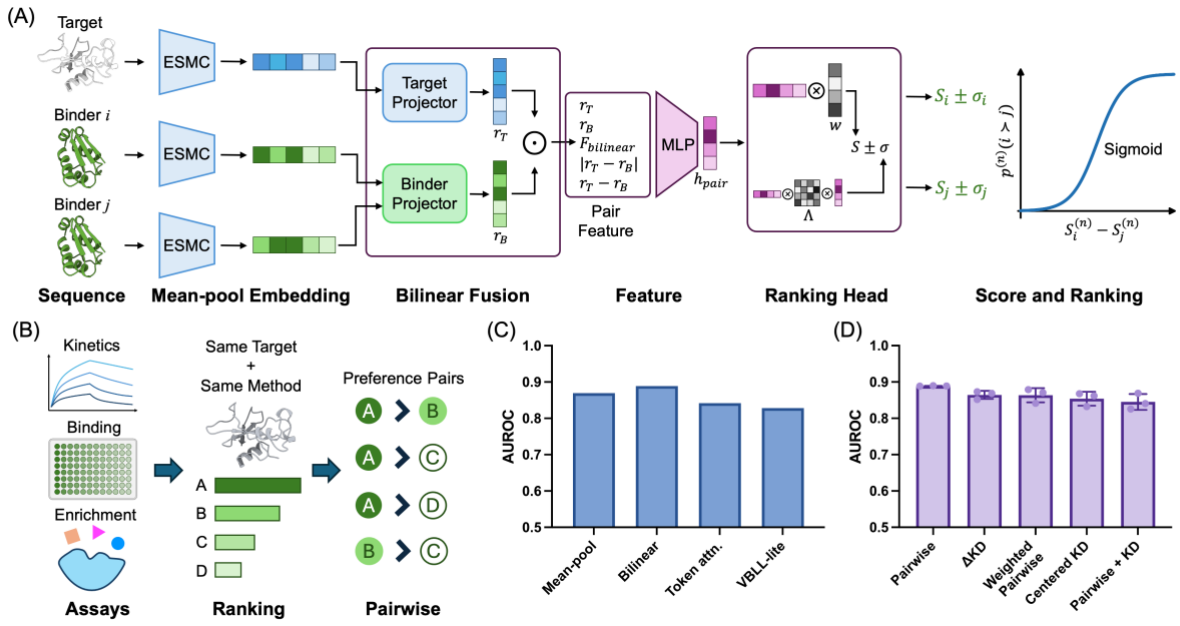

Figure 2 | Target-binder representation, preference harmonization, and model selection. (A) ESMC mean pooling, low-rank bilinear fusion, pair-feature encoding, and the Bayesian ranking head. (B) Preference pairs generated within the same target and assay method. (C) Pair-representation route screen. (D) Three-seed training-objective comparison.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.07.22.000293. This version posted July 22, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

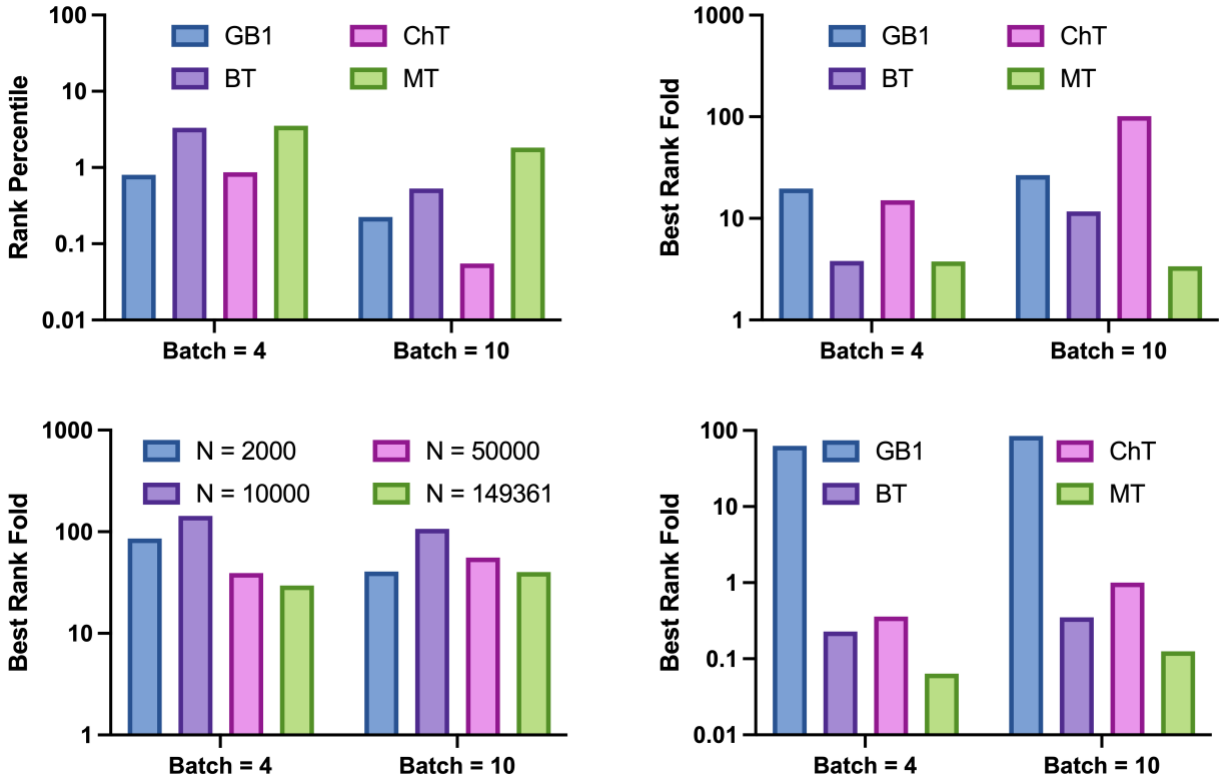

Figure 3 | Target-conditioned cold-start ranking. (A) Round-0 best true rank percentile across four tasks and two batch sizes. (B) Random-to-prior best-rank ratio across tasks. (C) Random-to-prior best-rank ratio across four GB1 library sizes. (D) Static ESMC-LLR-to-ASPIRE-prior best-rank ratio.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.07.22.000293. This version posted July 22, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

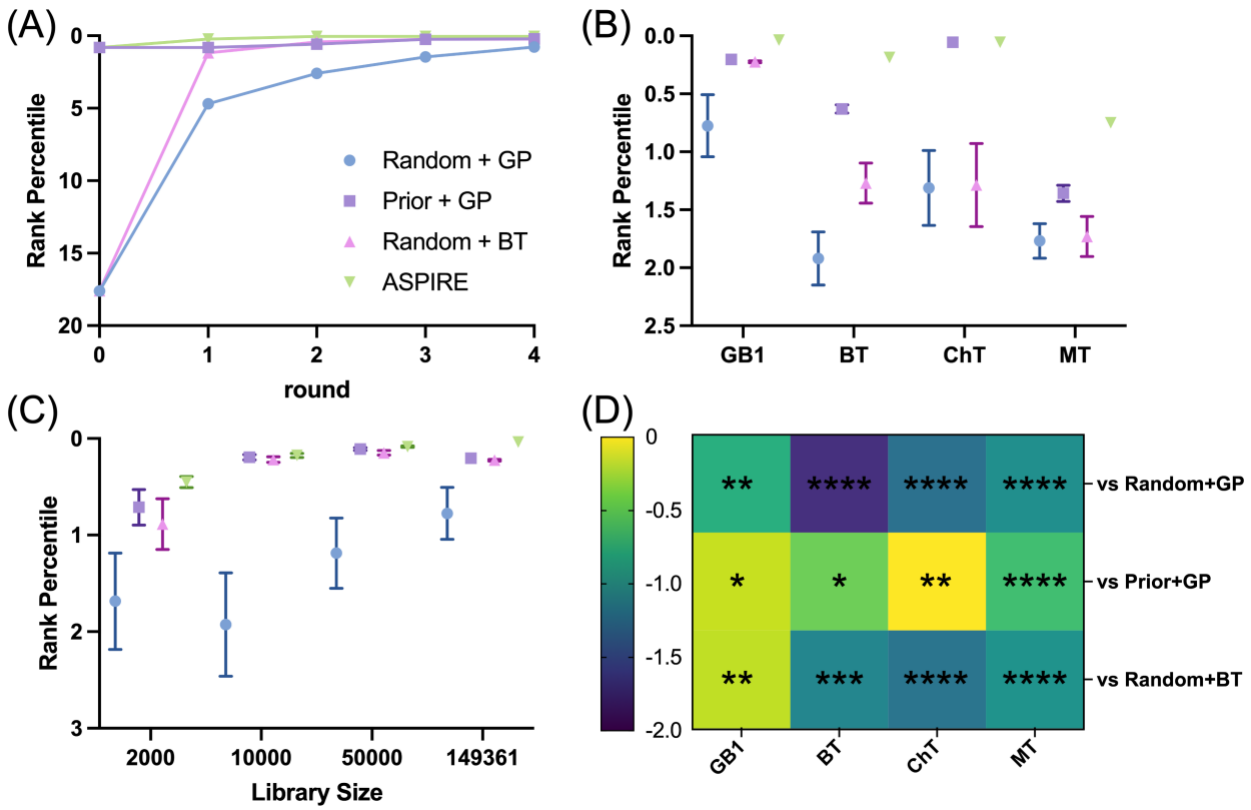

Figure 4 | Joint effects of prior initialization and ranking update. (A) Four-round trajectories in the full GB1 library at batch size four. (B) Final best rank percentile across four tasks. (C) Final best rank percentile across four GB1 library sizes. (D) Seed-paired ASPIRE-comparator effects and one-sided Wilcoxon tests.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.07.22.000293. This version posted July 22, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

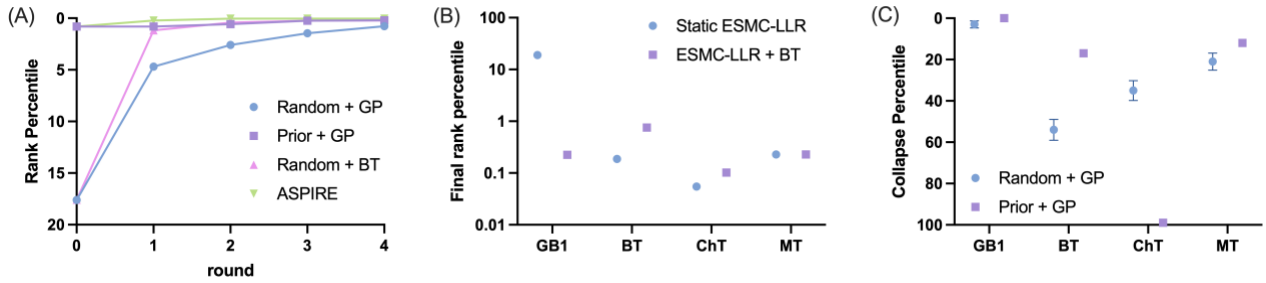

Figure 5 | Target-dependent encoder priors and GP behavior. (A) Four initialization-update trajectories in the full GB1 library. (B) Final best rank percentile for static ESMC-LLR and ESMC-LLR followed by BT/Laplace updating. (C) GP collapse rates under random and prior initialization across four tasks.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.07.22.000293. This version posted July 22, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

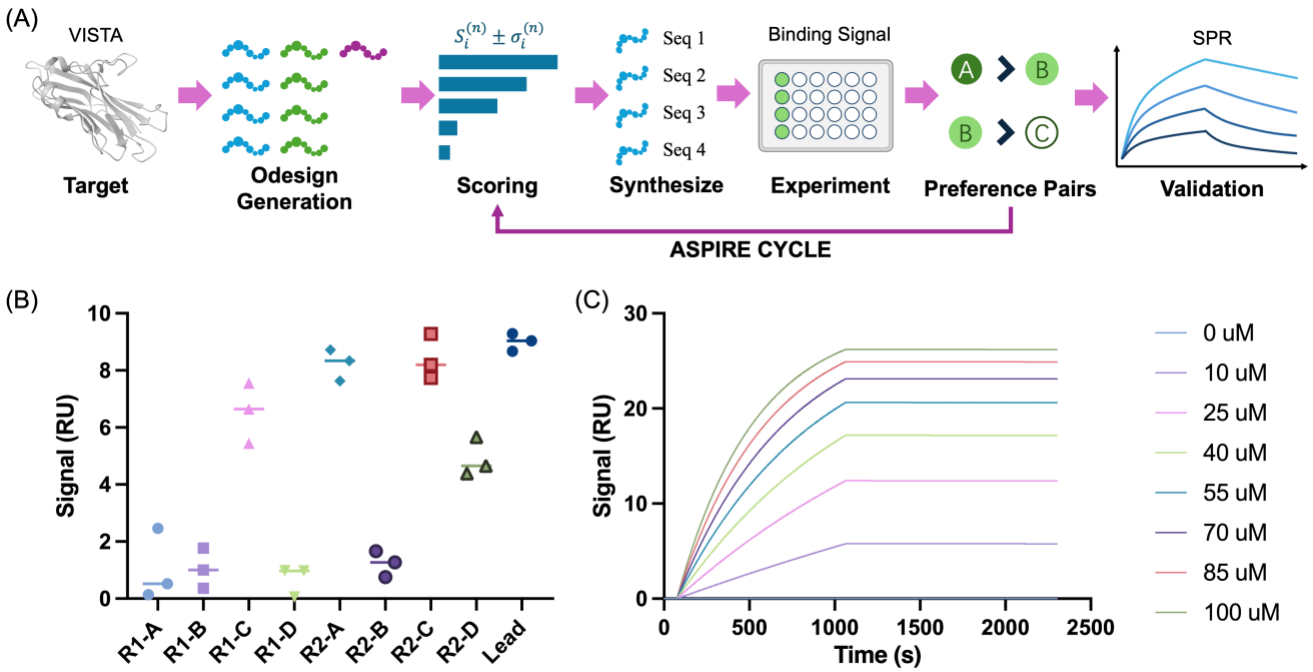

Figure 6 | Prospective ASPIRE-guided VISTA peptide selection. (A) Candidate generation, ASPIRE scoring, WeSPR300 screening, preference updating, and WeSPR Imager validation. (B) Triplicate 10 μM binding signals for two four-peptide rounds and the advanced lead after PBS blank subtraction. (C) WeSPR Imager sensorgrams with immobilized VISTA and peptide in the flow phase; dissociation begins at approximately 1,000 s, and free-R_max fitting yielded 1.62–2.31 pM.