

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

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

Mingqian Chen^{1,2,3,#}, Liping Huang^{4,#,*}, Yuxue Guo⁴, Yue Peng⁴, Yihui Yang², Yuhong Li², Wen Li², Zhonghua Liu^{1,3,*}, Gang Logan Liu^{2,5,*}, 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 Hunan Research Center of the Basic Discipline Developmental Biology, College of Life Sciences, Hunan Normal University, Changsha 410081, China.

⁴ School of Food Science and Pharmaceutical Engineering, Nanjing Normal University, Nanjing, Jiangsu, PR China.

⁵ Liangzhun Life Science Company, Wuhan, Hubei, PR China.

These authors contributed equally. * Correspondence: Liping Huang(lp Huang@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 three benchmark tasks, first-round ASPIRE selection identified candidates within the top 0.225%–3.542% of the measured landscape using batches of four or ten. Additional GB1 evaluations covered libraries containing 2,000–149,361 candidates. Iterative updates further improved candidate ranks and outperformed random-initialized Gaussian-process, prior-initialized Gaussian-process, and random-initialized Bradley–Terry/Laplace controls.

In an experimental evaluation of ASPIRE-selected VISTA-binding peptides, 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 initial screening stage, 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 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 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 initial selection without target-specific experimental labels, feedback-driven iterative selection, library size, and batch size on GB1 and the BT/BPTI and MT/BPTI interaction landscapes. Combinations of random or prior-guided initialization with Gaussian-process regression or Bradley–Terry/Laplace preference updates were used to assess the contributions of initialization strategy and update method. VISTA, the V-domain immunoglobulin suppressor of T cell activation, is an immune-checkpoint protein involved in T-cell suppression in the tumor microenvironment [29]. We used VISTA as an application target to evaluate ASPIRE-selected peptides through structure prediction, molecular dynamics, binding measurements, and WeSPR Imager kinetic characterization. Together, these analyses connect cross-target candidate prioritization with feedback-driven optimization and experimental binder screening.

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 $A > B$. These relations update the candidate score

distribution before the next batch is selected. ASPIRE provides target-conditioned candidate priorities and refines them as experimental preference evidence accumulates.

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]:

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

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. Test-set comparisons and paired architecture differences are shown in Supplementary Figure S1A–C.

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

We evaluated three benchmark tasks: the GB1 combinatorial fitness landscape, bovine trypsin (BT) against BPTI variants, and human trypsin-3 (PRSS3; mesotrypsin, MT) against BPTI variants. GB1 covers joint variants at four positions in the Protein G B1 domain [27]. The two BPTI libraries describe interface-variant interaction landscapes against the respective 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 0.225% to 3.542% across the three tasks and two batch sizes (Fig. 3A). Increasing the batch from four to ten improved the best selected rank in all three tasks. Relative to random selection, the prior improved the best selected rank across the three tasks (Fig. 3B). In GB1, this advantage persisted across library sizes of 2,000, 10,000, 50,000, and 149,361 candidates (Fig. 3C). Across these GB1 configurations, initial-batch top-1% enrichment averaged 12.7–25.0-fold over random expectation (Supplementary Figure S2A,B).

We compared the ASPIRE prior with static ESMC masked-marginal likelihood-ratio (LLR) ranking to examine the contribution of target-binder ranking training (Fig. 3D). Values above one favor the ASPIRE prior and values below one favor static LLR. ASPIRE strongly outperformed static LLR on GB1, while LLR was informative on the BPTI mutation landscapes. These task-dependent comparisons characterize the relative performance of the learned target-binder prior and generic sequence likelihood, and support ranking training as an effective route from protein representations to experimental candidate priorities. In a complementary comparison within the same 200-candidate pool for each task, the ASPIRE prior selected candidates with higher rank-enrichment scores than OpenDDE-v2 and Protenix-v2 in all three tasks (Fig. 3E).

Prior initialization and preference updates jointly improve closed-loop selection

We compared four combinations of initialization strategy and update method: random + GP, prior + GP, random + Bradley–Terry/Laplace, and prior + Bradley–Terry/Laplace (ASPIRE). This comparison evaluates the benefit of prior-guided initialization and the relative performance of regression-based and preference-based updates under matched experimental budgets. 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 reached 0.0355%, compared with 0.203% for prior + GP, 0.225% for random + Bradley–Terry/Laplace, and 0.775% for random + GP. Random initialization followed by preference updates showed rapid recovery after the first round, while prior initialization retained a substantial advantage under the same total budget.

Across GB1, BT/BPTI, and MT/BPTI at batch size four, ASPIRE produced the lowest mean final rank percentile in each task (Fig. 4B). Across four GB1 library sizes, prior initialization with Bradley–Terry/Laplace updates also 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 favor ASPIRE. Stars report one-sided paired Wilcoxon tests on 100 matched simulation seeds. Effect sizes and paired distributions jointly summarize the selection advantage.

These results show complementary contributions from prior-guided initialization and preference updating. The transferable prior moves the initial batch into a higher-ranking region of the landscape. Bradley–Terry/Laplace updates then use sparse relative feedback to refine candidate priorities. Their combination retained the strongest mean final selection performance across the evaluated tasks and library sizes.

Mechanism analyses reveal target-dependent encoder priors and GP instability

The full-GB1 selection trajectories summarize the four initialization–update combinations (Fig. 5A). Encoder-score feedback and GP-collapse controls then examine the contributions of sequence-based initialization and target-dependent update behavior (Fig. 5B,C).

Static ESMC-LLR showed target-dependent selection performance (Fig. 5B). In the batch-four comparison, it selected a final best rank percentile of 19.0% in GB1; adding Bradley–Terry/Laplace feedback improved this to 0.225%. For BT/BPTI, static LLR was already informative and the same update did not improve it; MT/BPTI was essentially unchanged. Preference feedback therefore substantially improved a weak GB1 initialization, with its additional benefit varying across target landscapes.

GP behavior also depended on the target and initialization strategy (Fig. 5C). Under the predefined diagnostic, random-initialized GP collapse rates were 3%, 54%, and 21% for GB1, BT/BPTI, and MT/BPTI, respectively; prior-initialized rates were 0%, 17%, and 12%. The diagnostic considered final selection quality and posterior numerical stability, as specified in Methods. Prior initialization reduced the observed collapse rate in all three tasks, with the largest reduction in BT/BPTI.

Structural and molecular dynamics evaluation of ASPIRE-selected VISTA peptides

BoltzGen [33] was used to generate a library of 5,316 VISTA peptide sequences without specifying an epitope. We evaluated VISTA candidates using structure-prediction confidence, joint structural criteria, and MD. Representative peptide poses are displayed against a shared VISTA reference surface (Fig. 6A). The ASPIRE prior Top200 set had a median candidate-mean ipTM of 0.779, compared with 0.730 for ESMC masked-PLL and 0.754 for Random (Fig. 6B). Joint pass rates were 48.7%, 25.4%, and 37.84% under loose criteria, and 38.4%, 6.7%, and 21.16% under strict criteria, respectively (Fig. 6C). Additional analyses show complex confidence (Supplementary Fig. S3A), threshold sensitivity (S3B), random-batch variation (S3C), candidate lengths (S3D), and prediction-seed consistency (S3E,F).

For the four ASPIRE candidates selected in this study, mean MD contact retention was 0.663, compared with 0.379, 0.423, and 0.442 for OpenDDE, Protenix-v2, and ESMC masked-PLL. Mean peptide pose RMSD was 3.77 Å for ASPIRE and 8.51, 5.16, and 5.52 Å for the respective comparators (Fig. 6D,E). The ASPIRE-selected group therefore retained more of its initial interface and showed less pose displacement in these simulations. Independent trajectories are presented in Supplementary Fig. S4A,B, and analysis-window comparisons in S4C,D.

Experimental validation of ASPIRE-selected VISTA-binding peptides

We next evaluated ASPIRE-selected VISTA peptides using binding measurements. The BoltzGen-generated library was prioritized using the ASPIRE prior within the workflow illustrated in Fig. 7A. Eight candidates were synthesized in two batches of four, forming the initial experimental screening cohort (Round 0). Binding signals were measured in triplicate at a final peptide concentration of 10 μ M in PBS at room temperature using a Liangzhun WeSPR300 instrument, and PBS blank responses were subtracted. Four of the eight candidates produced clear binding signals (Fig. 7B). These measurements provide experimental feedback for subsequent preference-based updating.

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. 7C). 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. Together, binding screening and kinetic characterization demonstrate the application of ASPIRE-based candidate prioritization to VISTA-binding peptide discovery.

Discussion

Large-scale generative models have greatly increased the supply of binder candidates, and candidate selection determines how efficiently these libraries yield experimental hits. Protein active learning and Bayesian optimization established that a small number of target-specific measurements can guide later experiments [14,15]. ASPIRE extends this principle to the start of a new target by transferring ranking information from cross-target interaction data. This prior concentrates the initial assay budget on higher-value candidates. Subsequent preference feedback enables target-specific adaptation within the same selection framework.

The connection between cross-target prioritization and within-target optimization is central to ASPIRE. GB1 and the two BPTI-protease landscapes showed enrichment of high-ranking candidates before target-specific labels were observed. Comparisons of initialization strategies and update methods further characterized the benefit of prior-guided initial selection and the relative performance of Bradley–Terry/Laplace and GP updates. Random initialization followed by preference updating recovered rapidly as measurements accumulated, while prior-guided initialization retained an advantage under the same small total budget. First-round ranking is therefore a determining component of low-budget screening.

ASPIRE combines target-conditioned representations with within-target preferences. Static ESMC likelihood-ratio rankings showed task-dependent performance [31,32]. Target-binder ranking training extracts transferable information from protein representations, and preference feedback allows evidence from a new target to refine candidate ordering. GP behavior also varied with the task and initialization strategy under the small-batch protocol. These comparisons support a screening framework that uses cross-target information for initial prioritization and experimentally derived relative ordering for subsequent adaptation.

The VISTA evaluation extends ASPIRE-based prioritization to a generated peptide library. Structural and molecular dynamics analyses characterized the selected candidates, and binding measurements identified multiple active candidates within two small experimental batches. Kinetic characterization of an advanced lead yielded picomolar-range affinity under free- R_{max} fitting. Together with the benchmark selection trajectories, these results support the use of ASPIRE for low-budget binder prioritization and provide a basis for subsequent experimental preference updates.

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,

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

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 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.

Held-out objective and architecture comparisons

Objective comparisons used three training seeds and report individual test AUROCs with their mean and sample SD. Architecture comparisons report the recorded test-set runs. Paired test-AUROC differences from the mean-pooling baseline used bootstrap 95% confidence intervals across 48 groups, with 3,000 resamples for the bilinear variants and 2,000 for token attention (Supplementary Figure S1).

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 $P_t = i > j$, the posterior mode was updated by maximizing the sum of the Gaussian log prior and Bradley-Terry log likelihoods,

$$\log p(w | P_t) \propto -\frac{1}{2}(w - w_0)^T \Lambda_0 (w - w_0) + \sum_{(i,j) \in P_t} \log \text{sigmoid}(\varphi_i^T w - \varphi_j^T w)$$

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 $\sigma_i = (\varphi_i^T \Lambda^{-1} \varphi_i)^{1/2}$.

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 and MT/BPTI benchmarks were drawn from the protein-protein interaction landscapes reported by Heyne and colleagues [28]. Their effective libraries contained 13,341 and 13,607 candidates, respectively. Each target-library pair was treated as an independent ranking task.

Baselines and combinations of initialization and update methods

Random selection sampled candidates without a prior or update model. GP controls used the same 64-dimensional pair representation (GPphi64) to reduce representation confounding. We compared random and prior-guided initialization, each combined with GP regression or Bradley–Terry/Laplace preference updates. This yielded four protocols: random + GP, prior + GP, random + Bradley–Terry/Laplace, and ASPIRE. Within each evaluation configuration, all protocols used the same library, seed schedule, number of rounds, and batch budget.

Static ESMC-LLR controls ranked candidates using masked-marginal likelihood ratios. ESMC-LLR + Bradley–Terry/Laplace used the same initial LLR ordering followed by the ASPIRE last-layer preference update. This control evaluates the contribution of preference feedback when initial candidate ordering is supplied by a generic sequence score. For the external-method comparison, ASPIRE, OpenDDE-v2, and Protenix-v2 ranked the same 200-candidate comparison pool within each task. Native method scores were averaged over five seeds before selection, and the selected candidates were evaluated using rank enrichment (Fig. 3E).

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. Figure 4B,C reports standard errors across 100 seeds; other simulation error bars show standard deviations unless specified in the figure legend. 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.

Cold-start top-1% enrichment was calculated as the fraction of selected candidates in the true top 1% divided by 0.01. Initial batches of four and ten were evaluated in GB1 libraries of 2,000, 10,000, 50,000, and 149,361 candidates, with 100 recorded evaluation seeds per configuration. Summaries use the mean and sample SD; repeated deterministic selections in the complete library yield zero dispersion (Supplementary Figure S2).

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 structural evaluation

The ASPIRE prior and ESMC masked-PLL Top200 sets and five Random sets of 200 candidates were evaluated with five OpenDDE seeds per candidate. Confidence distributions use candidate-mean ipTM and pLDDT, with pLDDT expressed on a 0–100 scale. Box plots show medians, interquartile ranges, and whiskers extending to 1.5 times the interquartile range; outlier markers are omitted. Representative peptides are shown in orange against a shared VISTA reference surface with matched orientation and scale.

Loose structural criteria required ipTM > 0.75, pLDDT > 85, and binder-only aligned C α RMSD < 2 Å; strict criteria required ipTM > 0.80, pLDDT > 87.5, and the same RMSD cutoff. Pass rates use candidate-seed records. Candidate-bootstrap 95% confidence intervals retain all five seeds together. The pooled Random estimate is the mean of five equal-sized batches; its bootstrap resamples unique sequence clusters while retaining repeated batch appearances and all seeds together. These intervals use 10,000 resamples.

Threshold sensitivity varies the ipTM cutoff while holding pLDDT > 87.5 and binder-only aligned C α RMSD < 2 Å fixed. The Random curve shows the five-batch mean and its shading spans the batch minimum and maximum.

Individual Random batches retain their candidate-bootstrap intervals. Seed-consistency distributions count zero to five passing predictions per candidate under each criterion and retain each Random batch occurrence (Supplementary Figure S3).

VISTA molecular dynamics analysis

MD analyses used the four ASPIRE candidates selected in this study and four candidates from each comparator. Each candidate had three independent 30-ns trajectories with fixed pH 7.4 protonation assignments. Contact retention is the fraction of initial heavy-atom residue contacts maintained using a 4.5 Å cutoff. Peptide pose RMSD uses C α coordinates after target alignment. Primary summaries average each trajectory over 20–30 ns and then average the three trajectories per candidate; method-level summary marks and error bars show the mean and sample SD across four candidates.

The reproducibility analysis displays all three trajectory means for each candidate, with identical candidate order across the contact-retention and RMSD panels. Sensitivity analyses repeat the candidate-level aggregation over 15–30 ns and compare it with the primary 20–30 ns window, using equality lines as references (Supplementary Figure S4).

VISTA peptide generation and screening

BoltzGen [33] generated a VISTA peptide library containing 5,316 sequences without a specified epitope. The sequences were then prioritized using the ASPIRE prior. Eight peptides were synthesized in two batches of four and analyzed as the initial screening cohort (Round 0). Binding signals were measured in triplicate using a Liangzhun WeSPR300 instrument at a final peptide concentration of 10 μ M in PBS at room temperature. VISTA was immobilized on a carboxylated LifeDisc, and PBS blank responses were subtracted. The measured responses were retained as experimental feedback for subsequent preference-pair construction and model updating.

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 μ M. 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); the Major Program (JD) of Hubei Province (grant no. 2023BAA011); and the Hunan Province Major Basic Research Projects (Grant No. 2026JC0008). 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).
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).
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. Stark, H. et al. BoltzGen: Toward Universal Binder Design. *bioRxiv* (2025) doi:10.1101/2025.11.20.689494.

Figures and legends

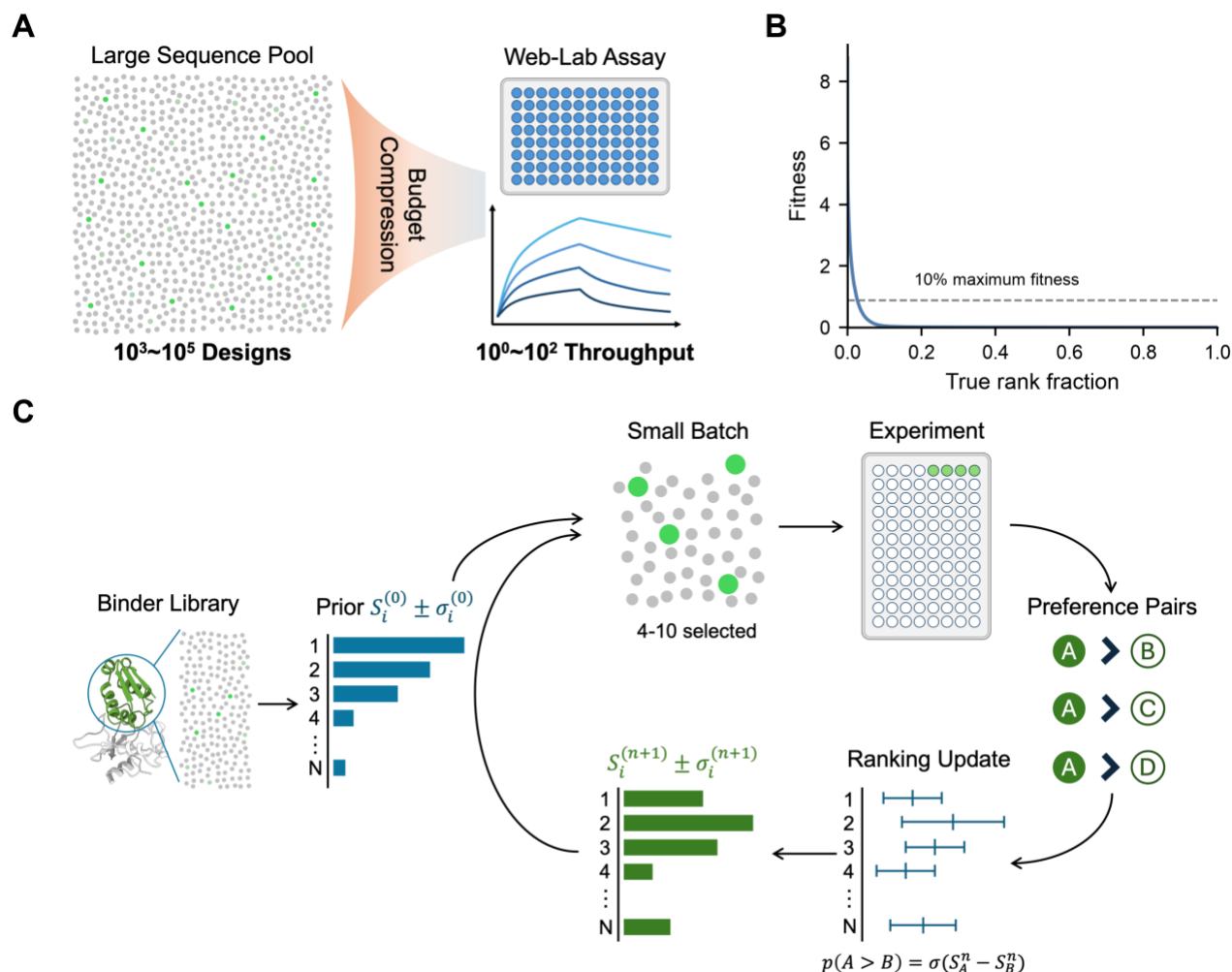


Figure 1 | ASPIRE connects large candidate libraries to small-batch experimental decisions. (A) Orders-of-magnitude gap between candidate-library size and assay throughput. (B) GB1 fitness ordered by true rank fraction; the dashed line marks 10% of maximum fitness. (C) Prior scoring, small-batch measurement,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

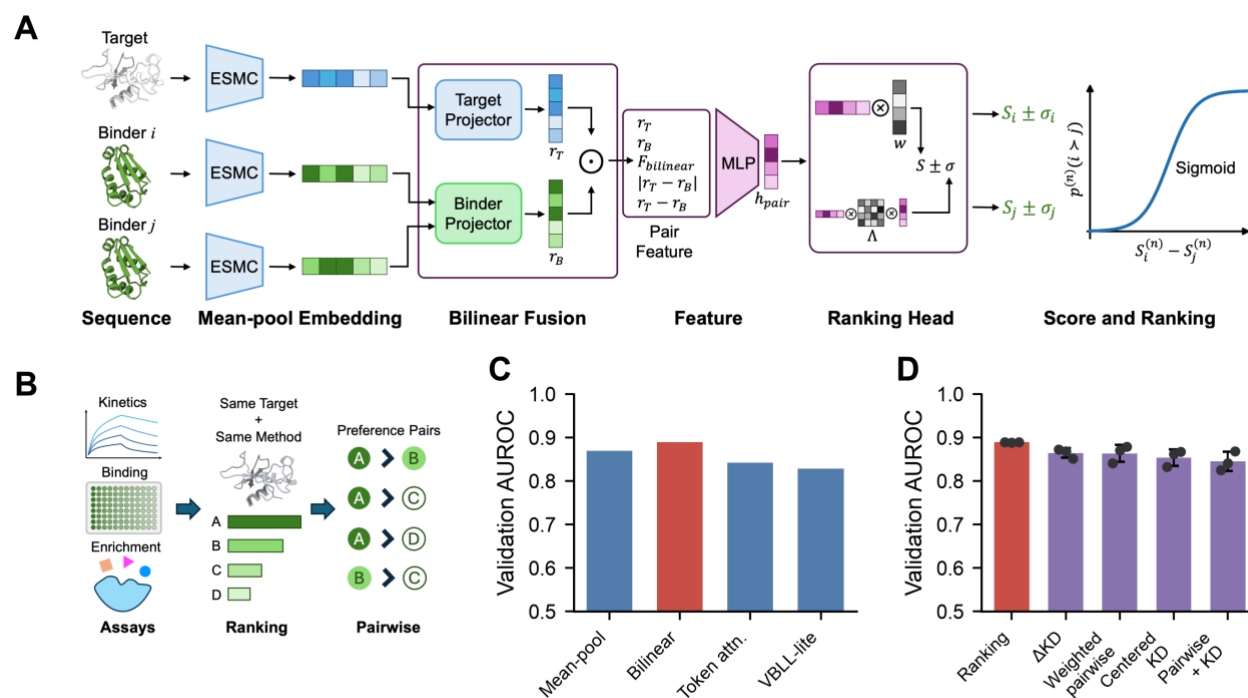


Figure 2 | Target-binder representation, preference harmonization, and model selection. (A) ESMC mean pooling, low-rank bilinear fusion, pair-feature encoding, and the Bayesian ranking head. (B) Preference pairs generated within the same target and assay method. (C) Pair-representation route screen.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

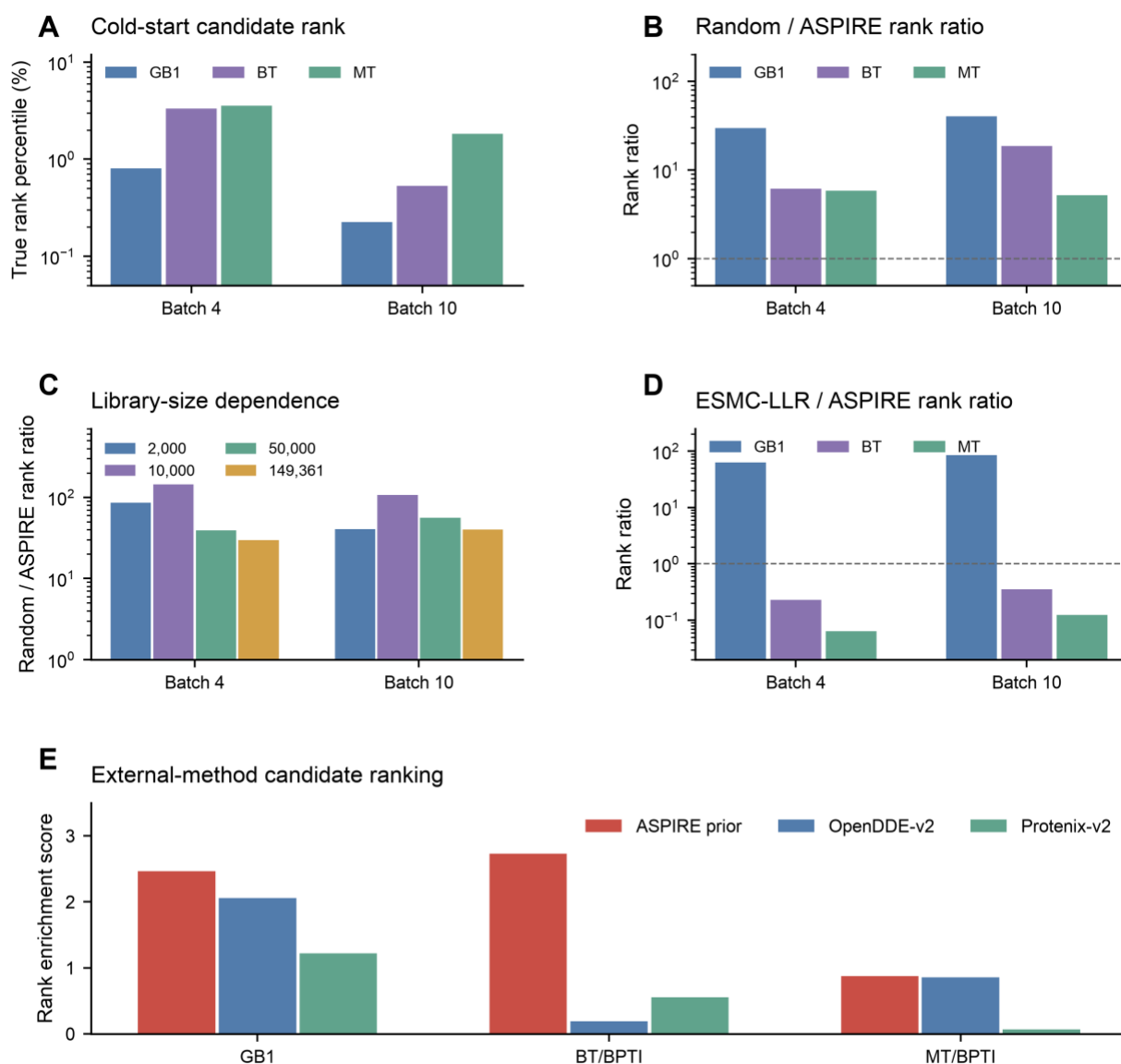


Figure 3 | Target-conditioned cold-start ranking. (A) Round-0 best true rank percentile across three tasks and two batch sizes. (B) Random-to-prior best-rank ratio across tasks. (C) Random-to-prior best-rank ratio across four GB1 library sizes. (D) Static ESMC-LLR-to-ASPIRE-prior best-rank ratio. (E) Best-ranked candidates selected by ASPIRE, OpenDDE-v2, and Protenix-v2 from the same 200-candidate comparison pool for each task. Rank enrichment is $-\log_{10}(\text{rank percentile}/100)$, with larger values indicating higher positions in the measured landscape. Native method scores were averaged over five seeds before candidate selection.

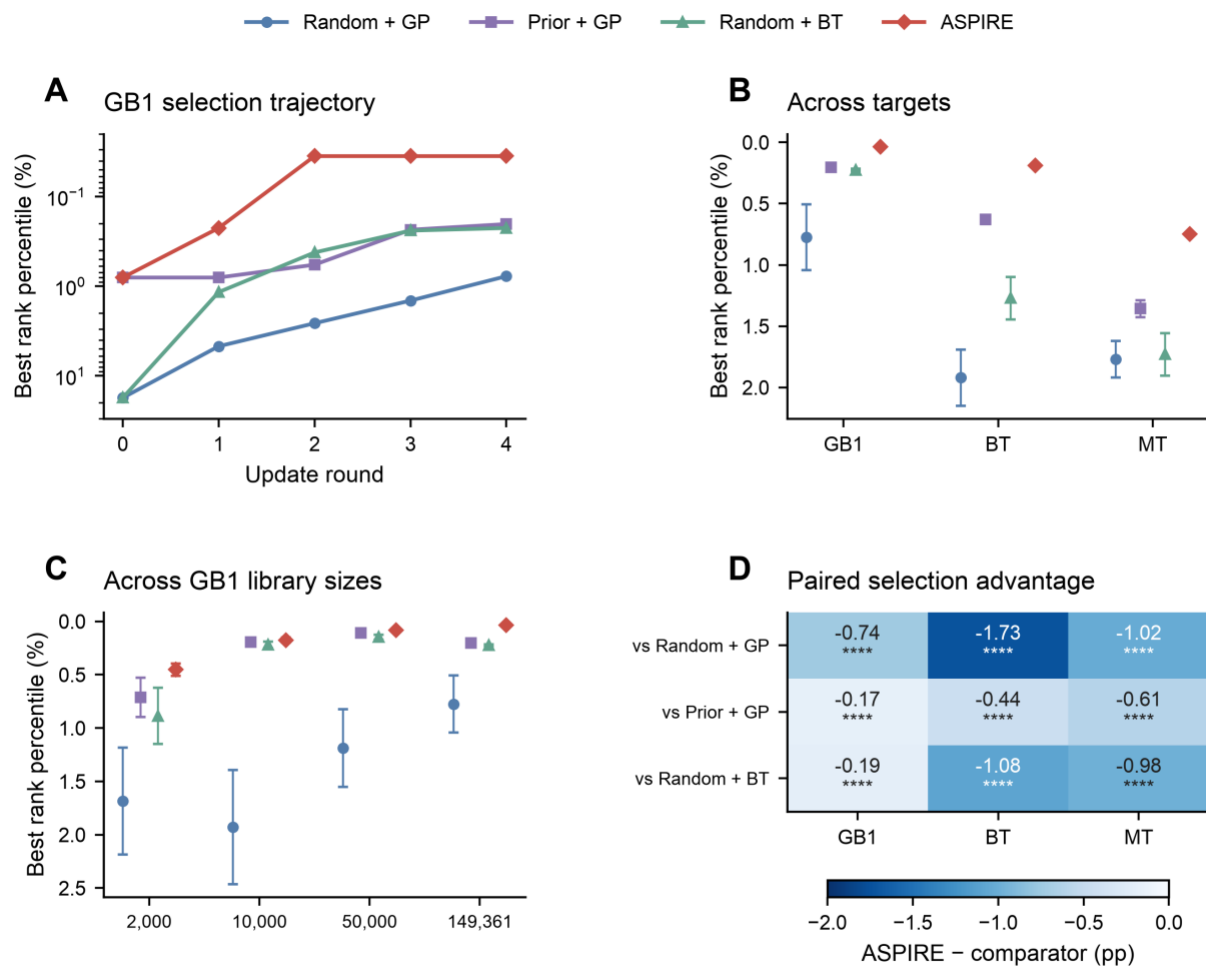


Figure 4 | Joint effects of prior initialization and ranking update. (A) Four-round trajectories in the full GB1 library at batch size four. (B) Final best rank percentile across three tasks. (C) Final best rank percentile across four GB1 library sizes. Points and error bars in (B, C) show mean $\pm$ standard error across 100 seeds. (D) Seed-paired ASPIRE-comparator effects and one-sided Wilcoxon tests.

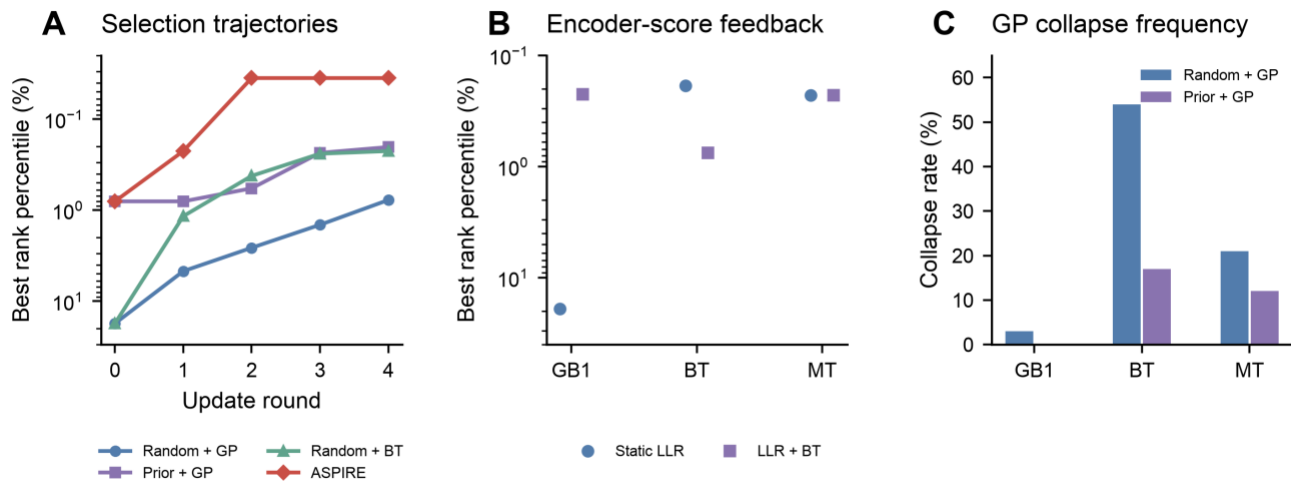


Figure 5 | Target-dependent encoder priors and GP behavior. (A) Four initialization-update trajectories in the full GB1 library. (B) Final best rank percentile for static ESMC-LLR and ESMC-LLR followed by BT/Laplace updating. (C) GP collapse rates under random and prior initialization across three tasks.

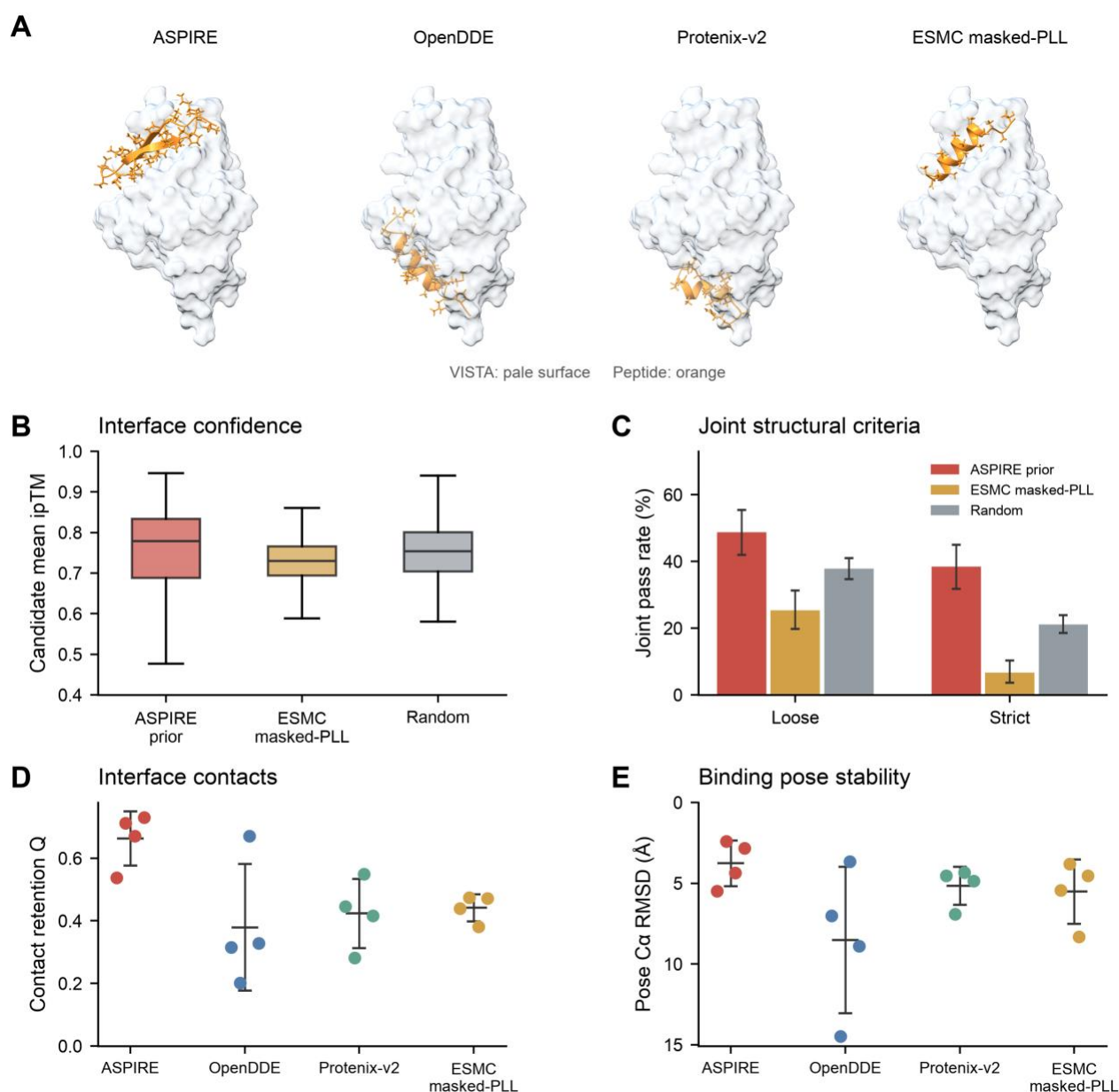


Figure 6 | Structural evaluation and molecular dynamics of ASPIRE-selected VISTA peptides. (A) Representative peptides on a shared VISTA surface with matched orientation and scale. (B) Candidate-mean ipTM for ASPIRE prior and ESMC Top200 sets and five Random sets of 200 candidates; boxes show medians and interquartile ranges with 1.5-IQR whiskers. (C) Joint pass rates under loose and strict criteria, with candidate-bootstrap 95% confidence intervals. (D, E) Contact retention and target-aligned peptide Ca RMSD for four candidates per method. Points show candidate means from three 30-ns trajectories at pH 7.4; summary marks show mean $\pm$ SD across candidates.

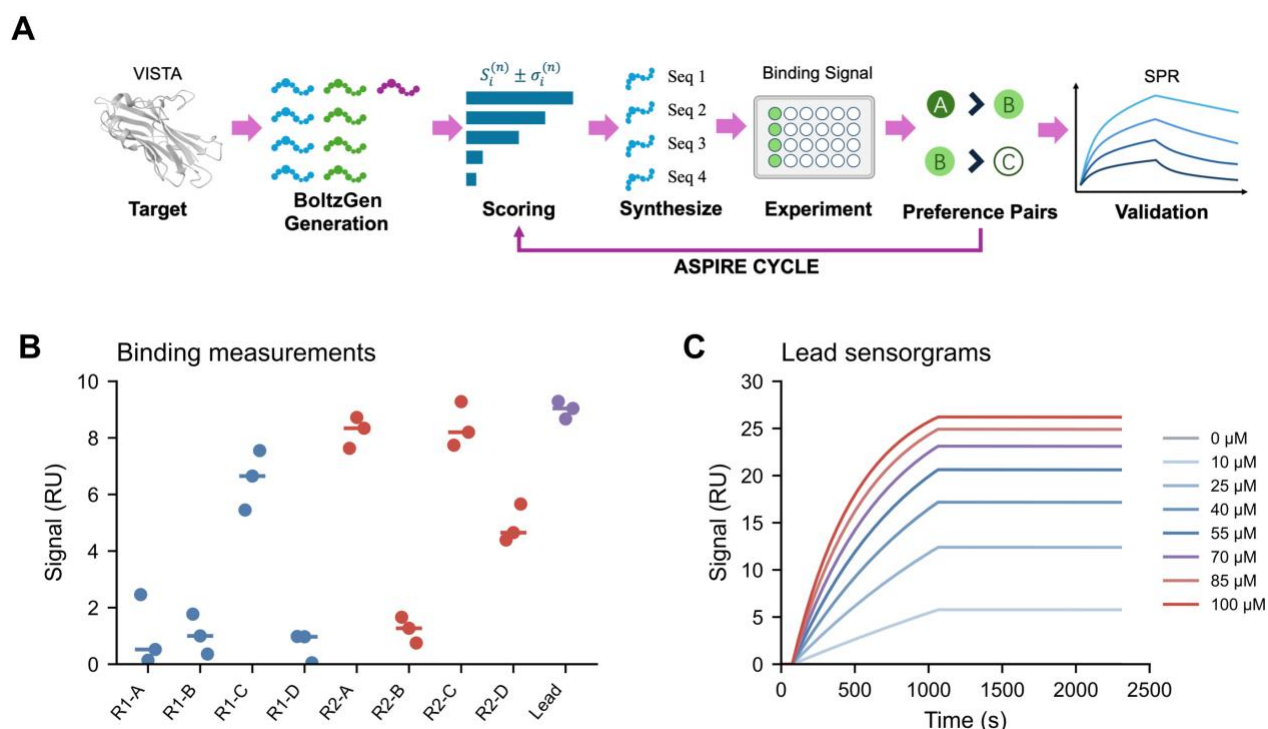
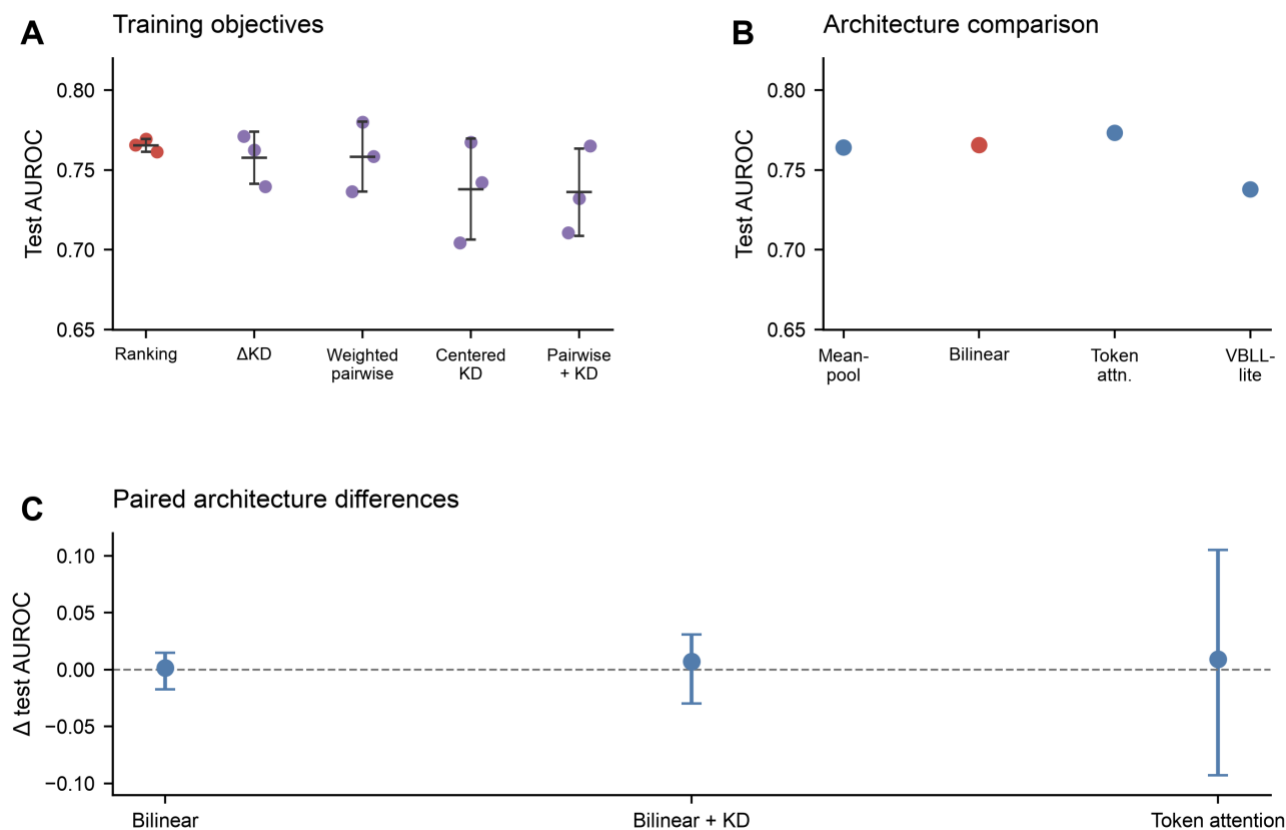
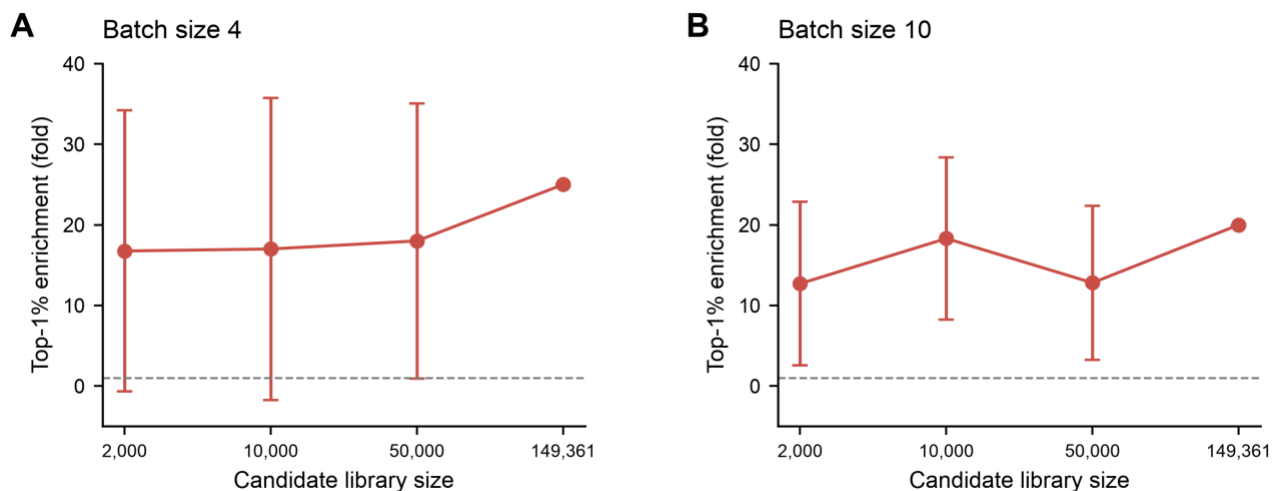


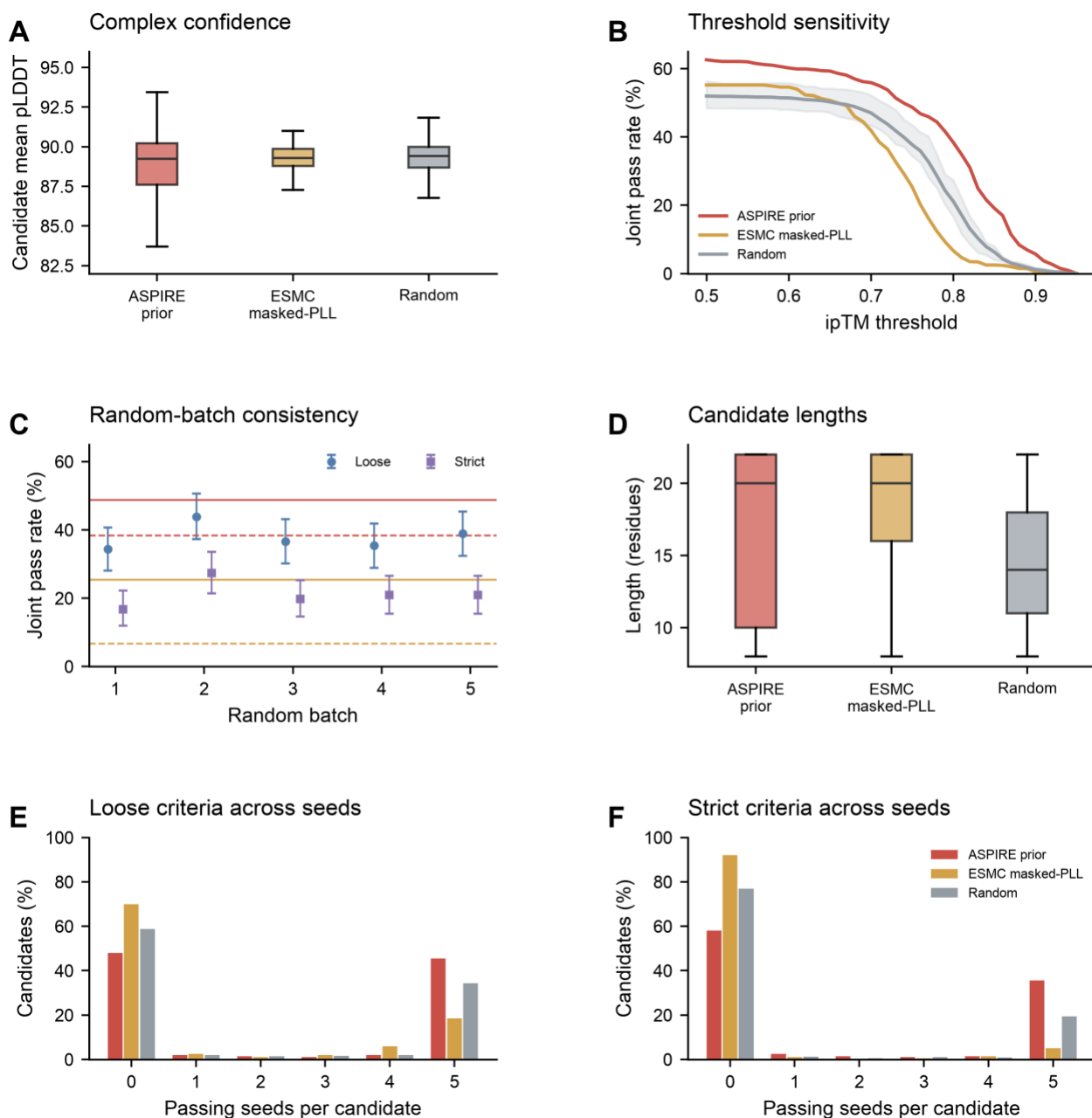
Figure 7 | Experimental validation of ASPIRE-selected VISTA-binding peptides. (A) ASPIRE workflow connecting BoltzGen candidate generation, ranking, binding measurements, preference updating, and kinetic characterization. (B) Triplicate 10 μM binding signals for two four-peptide experimental batches (R1 and R2), together forming the initial screening cohort, and the advanced lead; responses are PBS-blank-subtracted and horizontal marks show medians. (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.



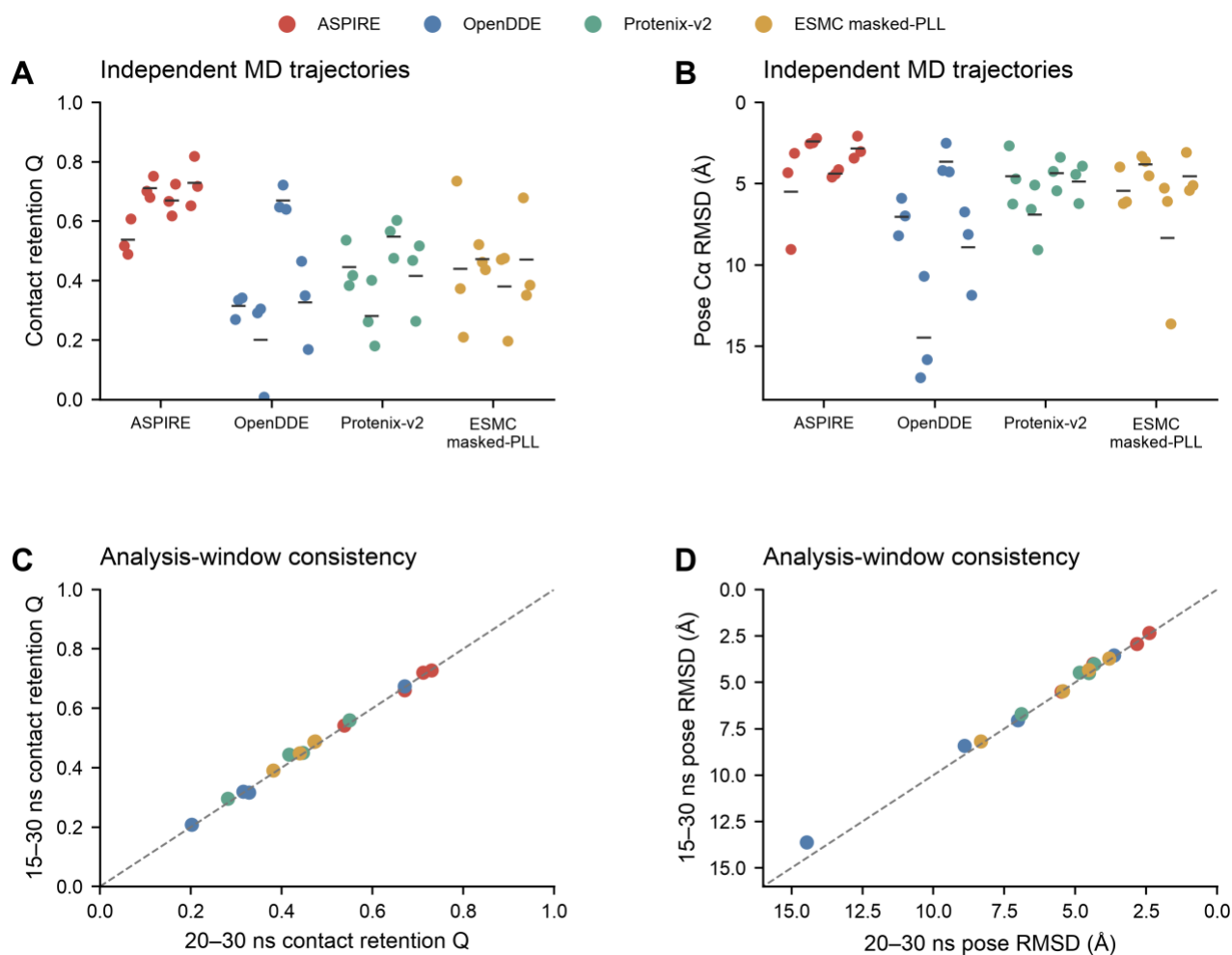
Supplementary Figure S1 | Training objectives and model architectures. (A) Test AUROC for five objectives across three training seeds; points and summary marks show individual seeds and mean $\pm$ SD. (B) Test AUROC for four architectures. (C) Paired test-AUROC differences from mean pooling, with 95% bootstrap confidence intervals across 48 groups. The dashed line denotes zero difference.



Supplementary Figure S2 | Cold-start enrichment of high-ranking GB1 candidates. (A, B) Top-1% enrichment for initial batches of four and ten across four library sizes. Points and error bars show mean $\pm$ SD across 100 evaluation seeds per configuration. Enrichment is the selected top-1% fraction divided by 0.01; the dashed line denotes random expectation.



Supplementary Figure S3 | Structural confidence and evaluation consistency. (A) Candidate-mean pLDDT. (B) ipTM-threshold sensitivity; shading spans the five Random batches. (C) Random-batch pass rates with 95% confidence intervals; red/gold reference lines denote ASPIRE/ESMC, with solid/dashed lines for loose/strict criteria. (D) Candidate lengths. (E, F) Passing-seed counts under loose and strict criteria. Boxes show medians and interquartile ranges with 1.5-IQR whiskers.



Supplementary Figure S4 | MD reproducibility and analysis-window consistency. (A, B) Contact retention and peptide pose RMSD for four candidates per method, each with three independent trajectories; horizontal marks show candidate means over 20–30 ns. (C, D) Candidate means from 20–30 ns versus 15–30 ns; dashed lines indicate equality. Candidates, their order, and simulation settings match Figure 6.