RECEPTOR.AI

Receptor.AI Peptide Discovery Platform: End-to-End Computational Design of Peptides

Abstract

Therapeutic peptide design is a multi-property optimization problem across large and chemically diverse search spaces, where changes intended to improve affinity, permeability, stability, or other properties can also alter peptide conformation and target binding. Here we present the Receptor.AI Peptide Platform, an end-to-end framework for orchestrating computational peptide design within iterative design–make–test–analyze programs, supporting both *de novo* hit discovery and lead optimization.

The platform combines peptide generation, ForceFold protein–peptide complex modeling, search over canonical and noncanonical chemistry, and QuorumMap multi-property optimization using relative affinity ranking and passive-permeability prediction. Alternative structural hypotheses can be retained when binding modes remain uncertain and used to define distinct design spaces, with experimental measurements from each round informing subsequent candidate selection and model adaptation.

Across method-specific benchmarks and experimental programs, ForceFold achieved CAPRI Acceptable-or-better accuracy for 74.4% of top-ranked predictions on 172 protein–peptide complexes. In a *de novo* GPCR program, 50 of 100 synthesized peptides were active *in vitro*, with a best IC_{50} of 90 nM; optimization of an existing interleukin-binding peptide produced 9 sub-100 nM binders among 24 synthesized designs; and in a Vps29 program, two experimental rounds improved PAMPA logPerm from −8.0 to −5.63 while improving affinity from 780 nM to 270 nM.

1 Introduction

Peptide discovery requires decisions about target binding, chemical modification, and the experiments needed to test each design. The platform brings these decisions into a common workflow, from the choice of starting peptide to the analysis of measured results.

1.1 The peptide-design problem

Therapeutic peptide discovery requires searching large sequence and chemical spaces for candidates that meet several property objectives. This involves choosing which structural hypotheses to consider, which chemistry to allow, which candidates warrant more expensive evaluation, and which designs to test experimentally.

Peptides can bind extended or shallow protein surfaces that are difficult to target with conventional small molecules [1–4]. Their size allows them to form contacts across a larger interface, but their flexibility makes the bound conformation difficult to predict. Changes intended to improve affinity, stability, or permeability can also alter this conformation and the interactions with the target. These properties therefore need to be considered together during design.

Peptide optimization includes changes to both sequence and chemical structure. In addition to canonical amino-acid substitutions, it can involve D-amino acids, N-methylation, C-α-methylation, beta- and gamma-amino acids, as well as stapling and different cyclization chemistries. Structural models must represent these modifications explicitly to assess their effects on peptide conformation and target binding.

A central task is to identify plausible binding modes and use them to guide the choice of modifications. When several binding modes are consistent with the available evidence, optimization must account for these alternatives. Experimental measurements are then needed to test the predictions and guide further changes.

1.2 Design–make–test–analyze orchestration

The Receptor.AI Peptide Platform supports hit discovery and lead optimization within iterative design–make–test–analyze (DMTA) programs. StratAI is the platform's agentic DMTA orchestrator, coordinating computational design and candidate selection within the project's synthesis and testing cycle. A project can start from a target structure alone, a target with one or more starting peptides, or an existing protein–peptide complex. Starting peptides can come from known ligands, mRNA-, phage-, or DNA-encoded library (DEL) screening, or previous design rounds. When no starting peptide is available, the platform generates candidates based on the target structure.

StratAI coordinates computational tasks, maintains project state, selects the appropriate modeling and optimization strategies, and incorporates experimental results into subsequent design rounds. Within a design round it delegates the optimization itself to QuorumMap, which in turn calls the structural, affinity and permeability models.

The cycle has four stages (Figure 1):

- Design: Form structural hypotheses, define the allowed chemistry, and optimize peptides against the project objectives to select candidates for experimental testing.
- Make: Transfer selected designs and synthesis specifications to the project's experimental team, including the required noncanonical residues, chemical modifications, and cyclization patterns.
- Test: The experimental team measures project-relevant endpoints such as binding, functional activity, stability, and permeability.
- Analyze: Compare predictions with measurements, identify how chemical changes affect the measured properties, and use these results to guide candidate selection and model adaptation in the next design round.

All starting points enter the same downstream lab-in-the-loop system. Each peptide is represented as an explicit atomic structure, including its chemical modifications and topology. Predicted protein–peptide complexes provide structural hypotheses for defining the allowed modifications and evaluating candidates. Experimental results guide candidate selection and model adaptation in subsequent rounds.

This white paper focuses on the computational decision layer within the cycle: forming structural hypotheses, defining and searching design spaces, evaluating properties, selecting candidates, and incorporating experimental feedback.

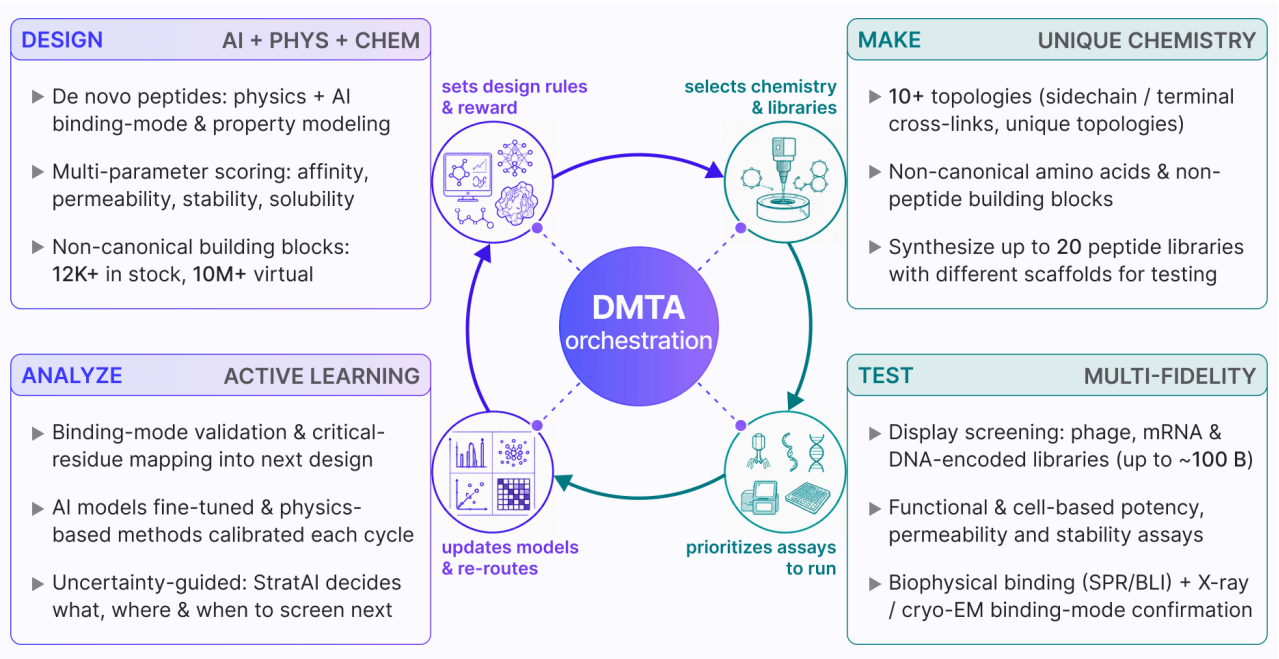

Figure 1. The design–make–test–analyze cycle for peptide discovery. Computational design produces candidates for synthesis and testing. Analysis of the experimental results guides the next design round.

1.3 Lab-in-the-loop optimization system

Within the DMTA cycle that StratAI orchestrates, QuorumMap is the platform's peptide optimization and candidate-selection engine. It makes the computational decisions inside a design round: which structural hypotheses to carry forward, how the design space is defined, which property models evaluate which candidates, which candidates are selected for experimental testing, and how returned measurements enter the next round. Each round produces selected candidates together with a design record containing:

- Composition and topology: The sequence and atomic structure, including noncanonical amino acids, other building blocks, cyclization, and backbone constraints.
- Structural hypotheses: The preferred predicted protein–peptide complex together with relevant alternative binding modes when structural ambiguity remains, including the confidence or disagreement signals used during selection.
- Ranking results: The physical and model-based scores used to compare the candidate with other designs, together with the structural hypothesis or hypotheses on which those scores were based.
- Design history: The starting peptide, the structural hypothesis and allowed design space used for optimization, and the modifications that produced the candidate.

The record is linked to synthesis and assay results as they become available. QuorumMap uses the returned measurements to guide next candidate selection and, where applicable, project-specific model adaptation. This allows each design decision to be reviewed against its structural assumptions, scores, and experimental outcome.

2 Platform workflow

The platform workflow implements the computational part of the QuorumMap lab-in-the-loop system. It converts project objectives, target information, and starting peptides into candidates for experimental testing. Target preparation and complex prediction establish structural hypotheses. Each retained hypothesis defines a design space for QuorumMap optimization. Affinity, permeability, and more expensive structural assessments provide evaluation signals used by QuorumMap to search these spaces and select candidates for testing.

2.1 Target preparation and binding-site analysis

The workflow begins by defining the target state and the regions where peptide binding will be evaluated. Target preparation establishes the biological and chemical state used in subsequent calculations. It includes selecting the biological assembly, repairing missing local structure where possible, assigning protonation and covalent states, and treating metals, cofactors, structural waters, and crystallographic artifacts. Known functional states of the target are specified before conformational sampling.

An experimentally established binding site defines the region to be evaluated. When the site is unknown or depends on the target state, the platform combines cavity detection, energetic affinity maps, fragment hotspots, and docking across the protein surface. Regions that repeatedly support favorable poses across receptor states, peptide conformations, and docking engines are retained as candidate sites. These are alternative site hypotheses, and several can proceed to structural modeling when the evidence does not uniquely support one region.

2.2 Starting peptides

With the target prepared, starting peptides provide a second source of information for forming binding-mode hypotheses. Known ligands, hits from mRNA display, phage display, and DNA-encoded library screening [5–7], experimentally characterized peptides, and candidates from earlier design rounds constrain the binding models considered. When no suitable starting peptide is available, *de novo* generation proposes backbones and sequences based on the geometry and interaction hotspots of the target surface.

Generated candidates are checked for chemical and geometric validity before structural modeling. The remaining candidates undergo the same ForceFold prediction and scoring procedures as externally supplied peptides. Generation therefore provides starting designs; their binding properties must still be assessed computationally and tested experimentally.

2.3 ForceFold protein–peptide complex modeling

ForceFold is the platform's protein–peptide complex modeling system used to establish structural hypotheses for design. First, a physics-based search samples receptor states, peptide conformations, candidate sites, and docking poses, with ArtiDock-P used as a peptide docking model alongside other docking engines. Filtering, clustering, local minimization, and scoring reduce this pose pool to a set of distinct binding modes. Second, proprietary conditioning adaptors encode information from those modes for an all-atom co-folding model, which produces the final complex structures with the most correct binding mode.

If several binding modes remain comparably supported after refinement, they are retained as alternative structural hypotheses for downstream optimization. Figure 2 shows example predictions across four target classes — an oncogenic GTPase, a complement protein, a peptide-MHC, and an apoptosis regulator — with experimental structures released after the co-folding model's training cutoff.

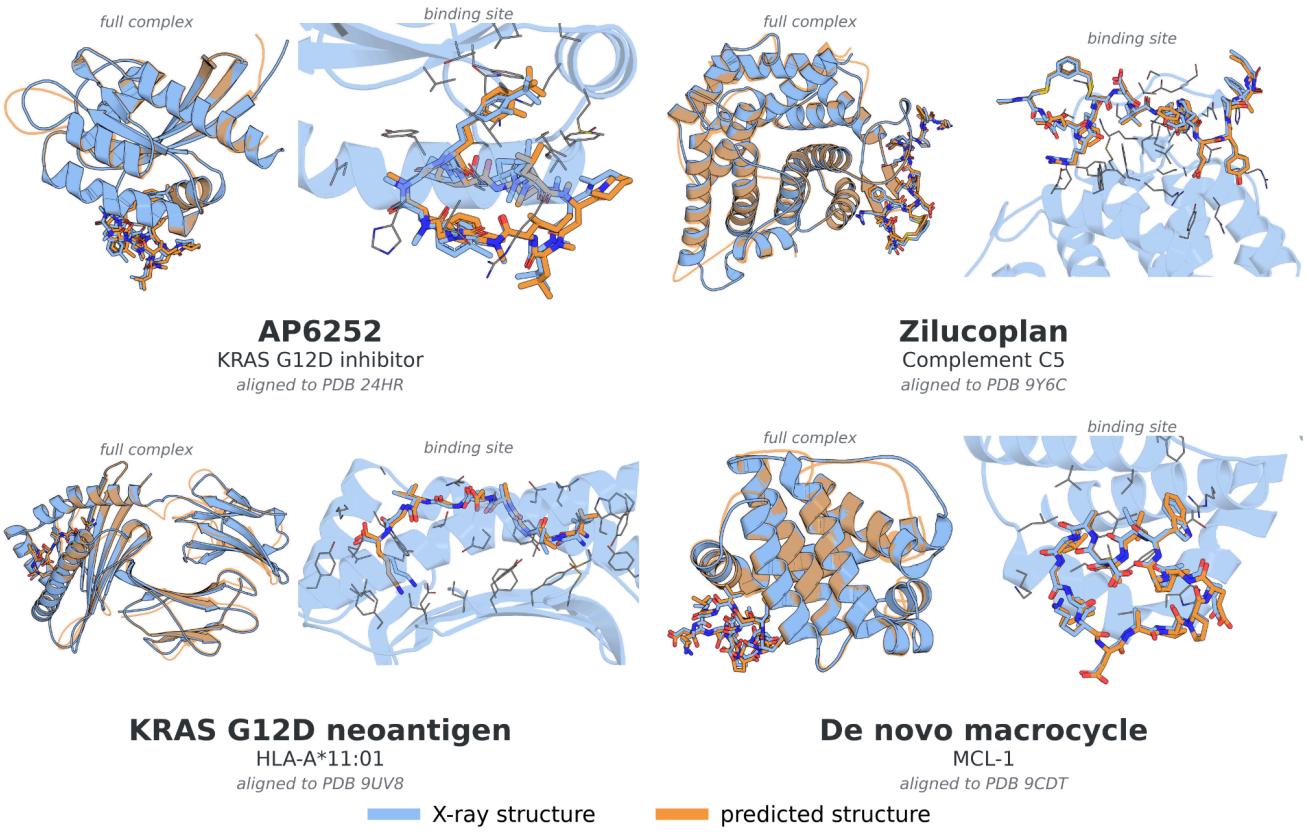

Figure 2. ForceFold predictions for four protein–peptide complexes deposited after the Boltz-2 training cutoff (1 June 2023). Predicted structures (orange) are overlaid with experimental reference structures (blue): AP6252-KRAS G12D (24HR), zilucoplan-class inhibitor-complement C5 (9Y6C), KRAS G12D neoantigen-HLA-A*11:01 (9UV8), and a de novo macrocycle-MCL-1 (9CDT).

2.4 Sequence and chemical optimization

For each retained structural hypothesis, a design space specifies the variable and protected positions, allowed canonical and noncanonical building blocks, and topology and backbone constraints. The project objectives determine

which properties are optimized and which constraints candidates must satisfy. Each variable position has a list of permitted building blocks, called its substitution alphabet.

Substitution alphabets are constructed either by manual specification or with Auto-Alphabet. Auto-Alphabet removes building blocks that do not fit the local structure and ranks the remaining blocks with a fast physical interaction score. Users can edit the proposed alphabets before sequence optimization.

The QuorumMap Optimization Engine selects the search and evaluation strategy for each design space according to its size and the evaluation budget. Small spaces can be evaluated exhaustively. For larger spaces, QuorumMap uses Bayesian optimization or hierarchical search to select candidates within the defined space. Affinity and other property evaluations guide which designs receive further computational assessment and which are selected for experimental testing.

2.5 Relative affinity ranking

Relative affinity ranking provides one of the evaluation signals used by QuorumMap during optimization and candidate selection. Related peptides are compared using a proprietary empirical multi-component scoring function. The scoring function is designed for high-throughput evaluation, typically screening on the order of 10^5 peptide candidates within a single computational optimization cycle. Its output is used for relative prioritization within an optimization series and should not be interpreted as an absolute binding constant.

2.6 Complex stability assessment

Before experimental selection, high-ranking designs undergo an additional check of their predicted complexes. Short molecular dynamics (MD) simulations examine changes in peptide conformation and the persistence of receptor–peptide contacts under thermal motion. Loss of the predicted interaction pattern during a short trajectory is treated as a warning that the structural hypothesis requires further investigation. Conversely, persistence of the complex is used as a gross-stability quality check and is not interpreted as independent evidence that the predicted binding mode or affinity is correct.

Because MD is more expensive than initial docking and scoring, this assessment is applied to a small final set of candidates. The resulting stability assessment is returned to QuorumMap as an additional signal for selection for experimental testing.

2.7 Passive permeability assessment

When passive membrane permeability is a project objective, its predicted value is used alongside affinity to guide optimization and candidate selection. A fast, conformation-sensitive permeability AI model evaluates the candidate set. Candidates for which AI confidence or chemistry/topology coverage is insufficient can be assessed with two mechanistically distinct structure-based physics tiers, termed “Light Physics” and “Heavy Physics.” These methods use the same peptide conformational ensembles as the AI model but differ in computational cost and physical description (representative sampled conformations shown on Figure 3). Their outputs are returned to QuorumMap as evaluation signals for optimization and candidate selection.

The ensembles sample aqueous-like, membrane-interface-like, and membrane-interior-like conditions to represent environment-dependent changes in conformation, exposed polarity, and intramolecular hydrogen bonding (Figure 3). These implicit environments are screening approximations to relevant dielectric and solvation conditions rather than atomistic representations of a biological membrane, and they do not represent an explicit kinetic trajectory through the membrane. Permeability predictions are evaluated against the parallel artificial membrane permeation assay (PAMPA). Validation is strongest for cyclic peptides; other peptide classes require separate evaluation.

3 Methods

The workflow depends on methods that operate on the same atomic structures but answer different questions about binding, chemical modification, and permeability. Here we describe how these methods work and the evaluations used to assess them.

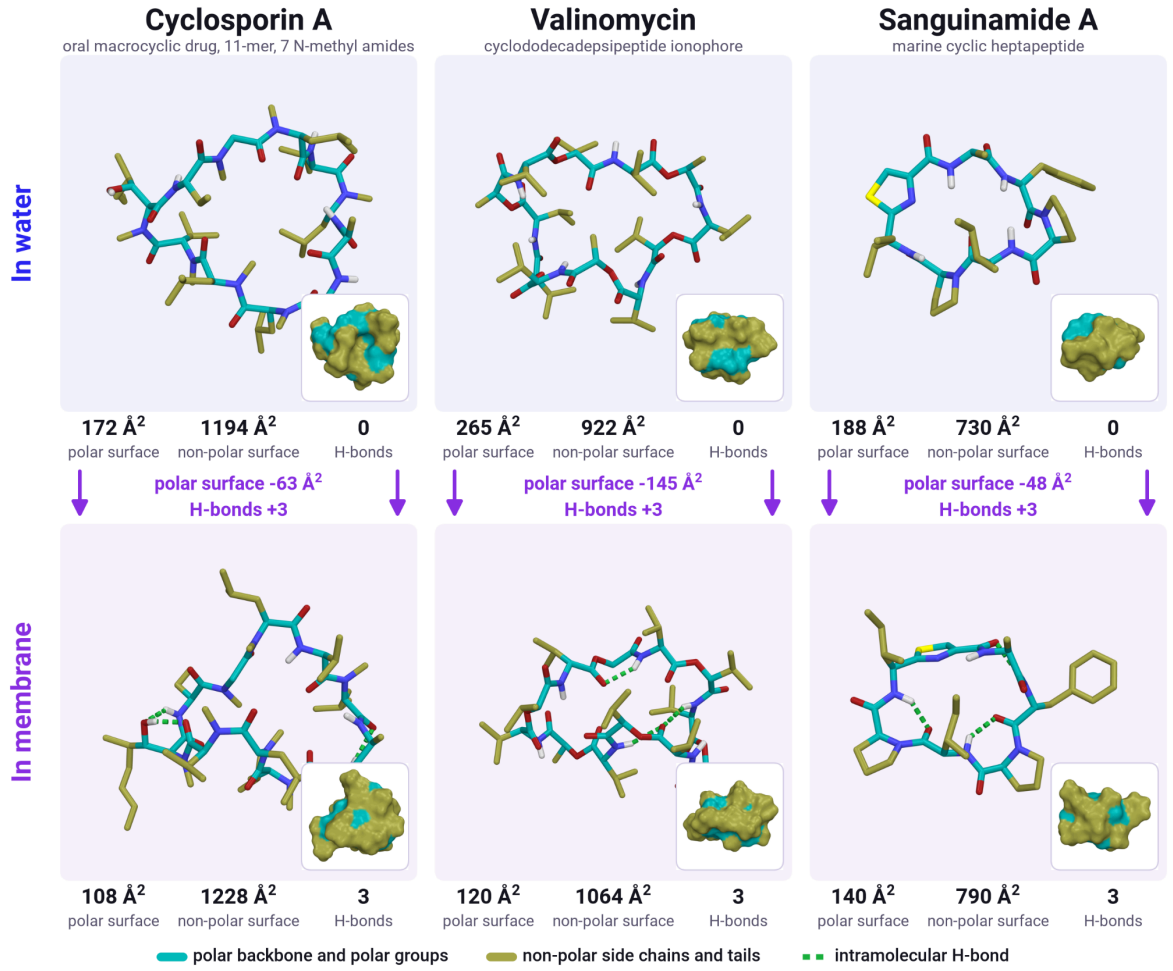

Figure 3. Representative conformations sampled for cyclosporin A, valinomycin, and sanguinamide A in aqueous and membrane environments. The examples illustrate phase-dependent changes in exposed polar and non-polar surface areas and intramolecular hydrogen bonding.

3.1 *De novo* peptide generation

For projects without a suitable starting peptide, the first computational task is to propose candidates that can be evaluated against the target. The generator uses the three-dimensional environment around selected target hotspots to propose peptide backbone conformations and sequences. It produces a diverse candidate set to sample alternative starting structures and chemistries.

Each candidate is checked for correct stereochemistry, severe internal or receptor clashes, and the integrity of specified cyclic or constrained structures [8]. The generator is used to propose structurally and chemically diverse starting peptides rather than to assign final confidence in target binding. Generated candidates therefore pass through ForceFold for binding mode identification, physical ranking, and experimental selection steps as externally supplied peptides.

3.2 ArtiDock-P peptide docking model

Both generated and supplied peptides require candidate binding poses for structural assessment. Peptide docking must handle a flexible ligand with many rotatable bonds and a frequently shallow interface [9, 10]. ArtiDock-P is the proprietary peptide docking model used for this purpose within ForceFold's physical-search stage. It contributes poses alongside other docking engines, all docking the same receptor and peptide conformational ensembles. All poses enter the pose-pool reduction procedure. The model retains the atom-level graph architecture of ArtiDock (Figure 4).

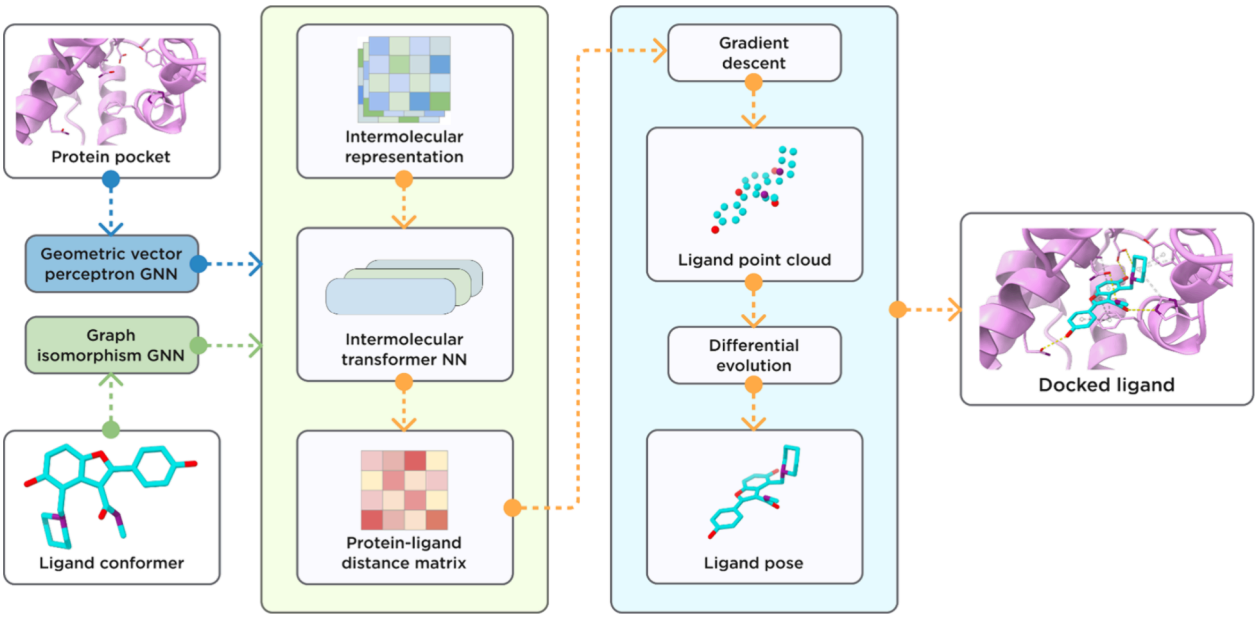

Figure 4. ArtiDock/ArtiDock-P architecture. Separate graph neural networks encode the protein pocket and ligand. Their combined representations predict interatomic distances, which are used to reconstruct the ligand pose. ArtiDock-P retains the published ArtiDock architecture [11] and extends training to peptide ligands.

A graph isomorphism neural network encodes the ligand, a geometric vector perceptron graph neural network encodes protein-pocket heavy atoms, and an intermolecular transformer predicts distances between pocket and ligand atoms:

$$D = \left[D_{ij}\right] \in R^{p \times c}, \qquad D_{ij} \approx \| p_i - l_j \|_2 \tag{1}$$

Here p and c are the numbers of pocket and ligand atoms, p_i and l_j are their coordinates, and D_{ij} is the predicted distance between the atom pair. The ligand pose is reconstructed by minimizing the difference between predicted and realized distances:

$$pose = \arg\min_{l_j} \sum_{i,j} \left(\| p_i - l_j \|_2 - D_{ij} \right)^2 \tag{2}$$

The published ArtiDock reconstruction procedure uses gradient descent and differential evolution to optimize the ligand coordinates and pose [11].

The base small-molecule model was trained on the PLINDER training split [12]: 227,066 protein–ligand pairs from 206,555 systems and 67,888 PDB entries, grouped into 40,148 redundancy-controlled clusters. ArtiDock-P extends this training set with curated protein–peptide complexes and synthetic examples containing noncanonical amino acids. Canonical and modified peptides use the same atom-level representation. The synthetic examples are used to broaden chemistry/topology coverage rather than presented as experimentally validated binding geometries; training-source breadth should therefore not be interpreted as equal structural certainty across all examples. Additional peptide-specific affinity scoring function is used as described in Section 3.5.

Docking985 [13] is a Receptor.AI-curated standalone peptide self/re-docking benchmark compiled from seven published protein–peptide structural benchmark collections. After harmonization and quality control, it contains 985 complexes representing 955 PDB entries and 473 receptor UniProt accessions, with peptide lengths from 2 to 20

residues; 973 structures are X-ray and 12 are NMR. Bound receptor structures are used, so the benchmark measures peptide pose recovery with the receptor already in the experimentally observed bound geometry rather than receptor-state prediction. Quality control includes structure quality, native-contact, interface-coverage, and steric-clash criteria. Pose quality is assessed with fraction of native contacts (fNat), interface RMSD, DockQ, and CAPRI quality classes. Because Docking985 evaluates ArtiDock-P alone, its results are reported separately from the end-to-end ForceFold benchmark.

3.3 ForceFold protein–peptide complex modeling

ForceFold is the protein–peptide complex modeling system that combines physical sampling with peptide-specific docking and further all-atom co-folding [14]. Its proprietary components include ArtiDock-P, peptide conformational sampling, pose-pool reduction, physical scoring, and conditioning adaptors. Boltz-2 [15] is used as the underlying all-atom co-folding model and constructs the final complex. Generalization of co-folding methods to unseen systems remains an active area of assessment [16].

The physical search generates diverse plausible complexes from receptor states, peptide conformations, candidate sites, and multiple docking engines. Pose-pool reduction retains distinct binding modes. Conditioning adaptors then encode information from these modes into the co-folding model, which can adjust both receptor and peptide coordinates (Figure 5). Experimentally established binding sites are used where available; otherwise, the site analysis described in Section 2.1 supplies candidate regions.

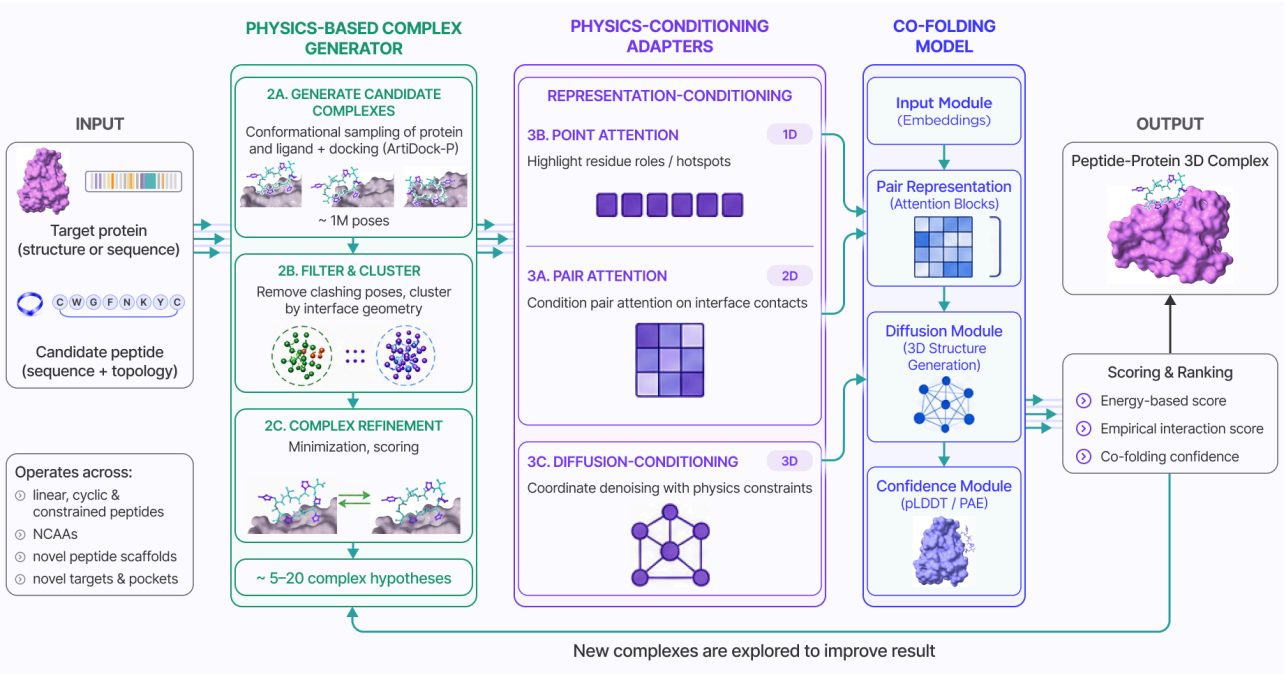

Figure 5. ForceFold architecture. Physical search and pose-pool reduction produce candidate binding modes. Point, pair, and diffusion-conditioning adaptors supply this information to the co-folding model. Confidence, interaction, and energy scores support final selection; persistent disagreement prompts further sampling.

3.3.1 Receptor and peptide conformational sampling

A useful binding-mode search must account for conformations of both the receptor and the peptide. After target preparation, MD samples conformations that capture protein and pocket flexibility, following ensemble docking [17]. Trajectories are clustered by local binding-site geometry, and representative states are selected for docking. AI-guided conformational sampling can propose additional conformations or substitute MD to reduce computational cost depending on target size.

Peptides are represented with explicit atom identities, connectivity, stereochemistry, cyclization, and chemical modifications. A fast proprietary physics-based generator samples conformations consistent with the correct secondary structure while exploring a broad range of backbone torsion angles. Enhanced-sampling MD of the free peptide is available when equilibrium populations or slow transitions are relevant, but it was not used in the ForceFold benchmark.

Target protein and peptide ensembles are designed to cover possible structures, not to estimate equilibrium populations. They are the input to the docking stage described next.

3.3.2 Docking and pose-pool reduction

The receptor and peptide conformational ensembles are docked with several engines, including ArtiDock-P. Chemically invalid docking poses and severe clashes are removed first. The remaining poses are deduplicated and clustered by binding mode and interaction pattern. Cluster representatives undergo local minimization and physical scoring.

Energy, geometry, recurrence across docking engines, and pose origin are retained as separate information. The set of retained binding modes has no fixed size. A pose can be retained if it represents a distinct binding mode, even when that mode occurs infrequently in the initial search.

3.3.3 Conditioning adaptors and co-folding

The retained binding modes must be converted into information that the co-folding model can use. Three adaptors provide this connection. The point adaptor represents local atom and pocket features. The pair adaptor represents receptor–peptide distances, contacts, orientations, and recurring interaction geometry. The diffusion-conditioning adaptor applies a soft bias during coordinate generation.

Repeatedly supported binding modes can receive greater weight, but the adaptors do not fix peptide coordinates or select the final complex independently. Co-folding can reject a docking hypothesis or apply ligand-conditioned local structural adjustments to the receptor–peptide complex.

Generated complexes are assessed using co-folding confidence, an empirical interaction score, and an energy-based estimate. Agreement supports selection of a complex family. Persistent disagreement prompts re-evaluation of existing samples or additional receptor and peptide sampling. If disagreement remains, alternative structures are retained for subsequent assessment.

3.3.4 Adaptor training

Because pose pools contain incorrect and incomplete binding-mode hypotheses, adaptor training must reflect that uncertainty. The adaptors are trained on pose pools generated without using the native coordinates as input. For each experimental protein–ligand complex, the physical-search workflow produces a pose pool. The binding modes retained from it condition co-folding, and the difference between predicted and experimental structures provides the training signal. The loss terms, weights, and optimization procedure are proprietary.

Training varies random seeds and retained pose subsets and includes incorrect binding modes, incomplete pose pools, and non-native receptor starting states. This exposes the adaptors to the imperfect structural information encountered during inference. Closely related receptor and ligand systems are separated between adaptor training and test sets to reduce similarity-based bias. This separation applies to the proprietary adaptor training. Benchmark interpretation also considers the training history of the underlying co-folding model through the post-cutoff and structural-similarity analyses described in Section 4.3; these analyses probe, but do not by themselves eliminate, possible effects of prior exposure to related structures.

3.4 QuorumMap Optimization Engine

3.4.1 Search space and surrogate model

QuorumMap is the peptide optimization engine within the platform's lab-in-the-loop system. It receives a structural hypothesis from upstream modeling and a design space that specifies the allowed peptide modifications and project objectives. A peptide optimization campaign is often limited to only a few hundred expensive experimental

evaluations, while the design space may contain trillions of possible sequences. QuorumMap allocates the available evaluation budget to candidates expected to improve the property profile or resolve uncertainty about where to search next.

Each variable position has a substitution alphabet containing permitted canonical residues, noncanonical residues, D-isomers, beta- or gamma-amino acids or other building blocks. These alphabets are constructed either by manual specification or with Auto-Alphabet (Section 3.4.5). Certain positions can be completely fixed, protected from selected substitution classes, or restricted to predefined building blocks. Cyclization anchors can be specified as constraints. Small spaces are enumerated exhaustively, while larger spaces use budget-limited search. Depending on the evaluation cost (both computational and experimental), computational campaigns can assess approximately 10^5 candidates from spaces of millions or trillions. A predefined evaluation limit controls computational cost. The search can stop early when scores stop improving or proposed variants remain close to the best candidates already identified.

QuorumMap searches for peptide variants with improved binding affinity or other specified properties. In each iteration, it selects a set of variants defined by the allowed residue substitutions and chemical modifications. These variants are evaluated in parallel, and their scores guide the selection of variants for the next iteration (Figure 6). Experimental measurements, when available, can also guide this selection. The output is a ranked list of peptide variants with predicted properties and associated uncertainty estimates. Depending on the size and structure of the search space, QuorumMap uses a Tree-structured Parzen Estimator (TPE) [18], a Gaussian-process model [19], or hierarchical search.

The Gaussian-process model uses chemical similarity between building blocks to predict the properties of untested peptide variants. At each variable position, the allowed building blocks are grouped hierarchically according to charge, hydrophobicity, size, and aromaticity. A tree-based kernel uses these groupings to calculate similarity between peptide variants:

$$K(x, x') = \frac{\sum_p \alpha_p S_p(x_p, x'_p) + \sum_{p<q} \rho_{pq} S_p(x_p, x'_p) S_q(x_q, x'_q)}{Z} \tag{3}$$

Here, x and x' are two peptide variants, and $K(x, x')$ is the kernel value describing their similarity for the Gaussian-process model. The indices p and q denote variable peptide positions; x_p and x'_p are the building blocks at position p in the two variants. $S_p(x_p, x'_p)$ measures the chemical similarity of these building blocks using the hierarchical grouping defined for position p. Building blocks grouped closely together, such as leucine and isoleucine, have higher similarity than chemically distinct building blocks, such as aspartate and lysine. The first sum runs over all variable positions. The coefficient α_p weights the contribution of position p. The second sum runs over all distinct position pairs, with $p < q$ ensuring that each pair is counted once. The product $S_p(x_p, x'_p)S_q(x_q, x'_q)$ describes similarity at both positions, and ρ_{pq} weights this joint contribution. These pair terms allow the model to represent effects that depend on combinations of substitutions. Z is a normalization factor that sets the scale of the kernel. The positional weights, pair weights, and parameters defining similarity within each hierarchy are fitted using evaluated peptide variants and their corresponding objective values.

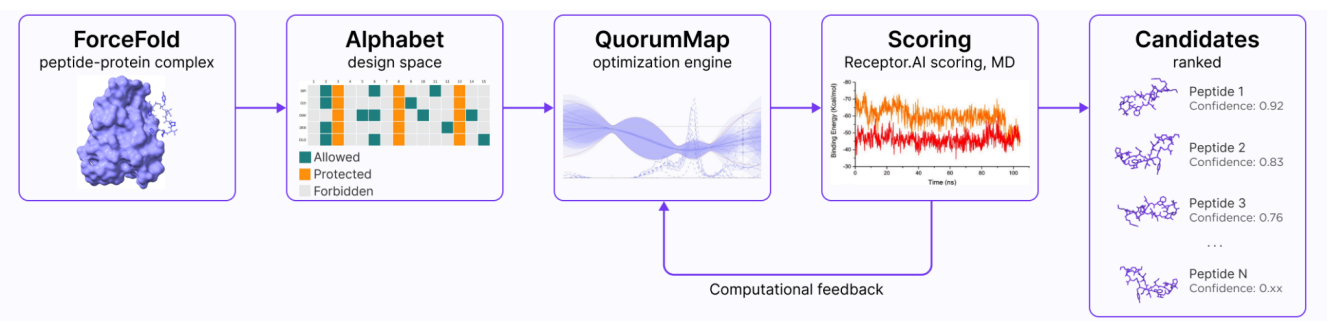

Figure 6. Computational peptide optimization workflow. For each retained structural hypothesis from ForceFold, position-specific alphabets and protected or forbidden design constraints define the admissible design space. QuorumMap proposes candidates and our scoring support ranking. MD stability assessment can provide additional evidence. Evaluation results guide the next computational iteration.

3.4.2 Avoiding trapping in local minima

An optimization campaign may fall into a local minimum and concentrate early on one scaffold family without testing whether a better family exists. QuorumMap manages this risk when selecting individual peptides, assembling sets of peptides for evaluation, choosing which regions of sequence space to explore, and selecting the search strategy.

Peptide selection considers both predicted performance and prediction uncertainty. An exploration coefficient controls the contribution of uncertainty to the selection score. When this coefficient is zero, candidates are ranked by predicted performance alone. Increasing the coefficient gives greater priority to candidates with uncertain predictions, allowing QuorumMap to explore substitutions whose effects are not yet well established.

Selecting peptides solely by their individual scores can produce a batch of closely related sequences. To increase chemical diversity, QuorumMap reduces the selection priority of candidates similar to peptides already included in the batch. The similarity threshold is adjusted according to the diversity of the candidates available in that iteration.

QuorumMap divides the sequence space into regions, each containing peptides that share defined structural or chemical features, such as residue classes at specified positions, a common scaffold, or a cyclization topology. It allocates candidate evaluations among these regions using the following priority score:

$$priority = promise + \lambda_u uncertainty + \lambda_n novelty + \lambda_d diameter \quad (4)$$

Here, *promise* is a conservative estimate of predicted peptide quality within a region, designed to limit the influence of a small number of favorable results. *Uncertainty* reflects limited evaluation of a region and decreases as more results become available. *Novelty* measures how different a region is from those already explored. *Diameter* measures the chemical variation within a region, giving greater priority to regions that contain a broader range of candidates. The coefficients λ_u, λ_n, and λ_d control the contributions of uncertainty, novelty, and diameter relative to promise.

Regions with higher priority scores receive more evaluations. Once a promising region has been sufficiently evaluated, it can be divided into smaller groups according to a chemical feature that distinguishes their members. Further evaluations can then focus on the more promising groups. Regions with consistently poor results may be excluded from further evaluation, while a minimum number of regions remain active to maintain coverage of different peptide families.

When multiple search strategies run in parallel, QuorumMap allocates evaluations according to their observed performance. It also reserves evaluations for less-tested strategies so that their potential can be assessed. These controls address different aspects of exploration: uncertainty guides the selection of individual candidates, similarity penalties maintain diversity within each batch, and allocation across regions and search strategies broadens coverage of the search space.

3.4.3 Optimization of multiple properties

Peptide optimization often requires simultaneous improvement of properties that can conflict. For example, a substitution that improves binding affinity may reduce passive permeability. The relative importance of these properties depends on the intended route of administration and formulation. QuorumMap therefore does not collapse the objectives into a single fixed score but retains candidates that represent different trade-offs between them.

Candidates are compared using Pareto dominance. A candidate is dominated if another candidate performs at least as well on every objective and better on at least one. Candidates that are not dominated form the Pareto set P. The hypervolume indicator measures the volume of objective space dominated by this set and bounded by a reference point r:

$$HV(P \mid r) = \lambda\left(\bigcup_{y \in P}[r,\ y]\right) \quad (5)$$

Here, [r, y] is the hyper-rectangle spanned by the reference point r and the objective vector y of a member of the Pareto set, and λ denotes volume (the Lebesgue measure). The hypervolume increases whenever a candidate that is not dominated by the current set is added. It is the only known unary quality indicator with this strict Pareto-compliance property [20, 21], and it therefore measures both the quality and the range of the retained trade-offs.

For two to four objectives, QuorumMap selects candidates with the parallel noisy expected hypervolume improvement (qNEHVI) acquisition function [22]. qNEHVI extends expected hypervolume improvement to batches of candidates evaluated in parallel and to noisy measurements [23]. It scores a proposed batch X of q candidates by the hypervolume gain that the batch is expected to add to the current Pareto set:

$$\alpha_{qNEHVI}(X) = \mathbb{E}_f\left[HV\left(P_f\left(X_n \cup X\right) \mid r\right) - HV\left(P_f\left(X_n\right) \mid r\right)\right] \tag{6}$$

Here, f is a function drawn from the surrogate posterior, X⃗ is the set of peptides already evaluated, Pf(·) is the Pareto set of a group of peptides under f, and $\mathbb{E}$f is the expectation over the posterior. Because the Pareto set of the evaluated peptides is also computed under f, the current front is treated as uncertain rather than fixed at noisy measured values. The expectation is estimated with quasi-Monte Carlo samples from the posterior, and the hypervolume is computed with a box decomposition of the non-dominated region [23, 24]. The batch is scored jointly, so candidates that would improve the same part of the front are not rewarded twice, which favours diverse batches. Uncertainty enters through the posterior samples, so exploration and exploitation are balanced without the exploration coefficient described in Section 3.4.2. Maximizing the expected hypervolume improvement is the optimal one-step decision for the hypervolume criterion under the surrogate model, and qNEHVI has shown state-of-the-art sample efficiency on noisy, batched multi-objective benchmarks [22].

The cost of computing the hypervolume and its box decomposition grows rapidly with the number of objectives: the number of boxes grows polynomially with the size of the Pareto set, with an exponent that increases with the number of objectives [24]. As the number of objectives increases, a growing fraction of candidates also becomes mutually non-dominated, which weakens the hypervolume signal. For five or more objectives, QuorumMap therefore uses a lighter random-scalarization upper-confidence-bound acquisition [25] with an augmented Chebyshev scalarization [26]. The posterior mean $\mu_i(x)$ and standard deviation $\sigma_i(x)$ of each objective i are scaled to comparable ranges and oriented so that larger values indicate better performance, and are combined into an optimistic estimate $u_i(x) = \mu_i(x) + \beta\sigma_i(x)$. For each position in the batch, a weight vector w is drawn from the probability simplex, and candidates are scored by:

$$s_w(x) = \min_i \; w_i\left(u_i(x) - z_i\right) + \rho \sum_{i=1}^{m} w_i\left(u_i(x) - z_i\right) \tag{7}$$

Here, x is a candidate peptide, m is the number of objectives, β controls the contribution of uncertainty, w_i is the weight of objective i, and z_i is a reference value below the observed values of that objective. The first term rewards the objective with the smallest weighted margin above the reference, so a candidate scores highly only when it performs well on all weighted objectives at once. The coefficient ρ is a small positive value that adds the total weighted margin and separates candidates that share the same smallest margin. The score increases with every objective, so its maximizer is Pareto-optimal with respect to the optimistic estimates, and, for small ρ, essentially every point of the front, including points on nonconvex parts, is the maximizer for some choice of weights. The candidate with the highest score fills that position in the batch, and the next position is filled with a newly drawn weight vector from the remaining candidates. Different positions therefore target different trade-offs, and repeated iterations spread the evaluations along the whole front. The cost of this score grows only linearly with the number of objectives, so it remains fast for large candidate pools. Hypervolume is still used to report and compare the retained Pareto sets.

Feasibility constraints are applied separately from the optimization objectives. These constraints can specify molecular-weight limits, minimum solubility, prohibited sequence motifs, or structural requirements. A property can therefore serve either as an objective to improve or as a requirement to satisfy. For example, when permeability is specified only as a minimum requirement, candidates below that threshold are excluded, but higher permeability does not receive additional weight at the expense of affinity.

3.4.4 Candidate and evaluation selection

QuorumMap allocates computational and experimental resources by screening candidates with inexpensive methods and selecting a subset for more costly evaluation. This approach limits the cost of assessing large peptide libraries while allowing selected candidates to undergo more detailed analysis.

The available evaluation methods depend on the property of interest. For affinity, available methods include surrogate scoring, Receptor.AI affinity scoring, molecular dynamics, free-energy perturbation, and SPR or ITC measurements. Passive permeability can be assessed using molecular descriptors, machine-learning models based on conformational ensembles, simulations in membrane-like environments, parallel artificial membrane permeability

assays (PAMPA) or artificial membrane assays (BLM). Caco-2 and MDCK assays could be used to measure transport across cell monolayers and include biological contributions beyond passive membrane permeation.

These methods differ in their physical assumptions, uncertainty, and cost. A more expensive method is not necessarily more accurate for every peptide or property. Within a given computational method, additional resources can support longer simulations or more independent replicates, and each such setting is treated as a separate evaluation level of that method. For each objective, the reference level is the measurement that the project treats as the ground truth, such as SPR or ITC for affinity or PAMPA for passive permeability.

In its Gaussian-process formulation, QuorumMap models the evaluation levels of an objective jointly rather than passing candidates through a fixed sequence of filters. For each objective j, a multi-output Gaussian process treats every evaluation level s as a separate output of the same underlying peptide property [27, 28]. Its covariance combines the peptide kernel of Equation (3) with a learned matrix of correlations between levels:

$$k_j((x,s),(x',s')) = B_j(s,s')\,K(x,x') \tag{8}$$

Here, x and x' are peptide variants, s and s' are evaluation levels, K(x, x') is the peptide kernel of Equation (3), and B⍰ is a positive semidefinite matrix whose entries describe the variance of each level and the covariance between levels for objective j. The off-diagonal entries of B⍰ are fitted from peptides evaluated at more than one level, so the model learns from the data how closely each inexpensive method tracks the reference level, including when a method is uninformative. A result from an inexpensive method therefore changes the predicted reference-level value of that peptide, and of chemically similar peptides, in proportion to the learned correlation. Because the correlations between all levels are estimated jointly, two inexpensive methods with similar biases are not counted as independent evidence. The autoregressive multi-fidelity model, in which each level is a scaled version of the level below plus an independent correction [27], is closely related to this structure.

Because each level is a separate output with its own mean and scale, raw scores from different physical models are never compared directly or pooled into one cutoff. Receptor.AI scoring and free-energy perturbation, for example, differ in physical approximation and scale; an ML permeability score and a PAMPA measurement likewise cannot share a raw-value cutoff. The model relates them only through their learned correlations. Each objective has its own ladder of levels and its own model, because the methods available for affinity and for permeability differ and are applied to different candidates. The Pareto set of Section 3.4.3 is always defined on the reference-level predictions of each objective, so inexpensive evaluations contribute only by improving these predictions.

For the hypervolume-based selection described in Section 3.4.3, the allocation objective is to choose the candidate and evaluation level that provide the greatest expected improvement of the reference-level Pareto set per unit cost [29, 30]:

$$\left(x^*, s^*\right) = arg\,max_{x,s} \frac{\mathbb{E}_y\left[HV\left(P^*(D\cup\{x,s,y\})\right) - HV\left(P^*(D)\right)\right]}{c(s)} \tag{9}$$

Here, x denotes a peptide candidate, and s specifies the objective, evaluation method, and resource setting. D is the set of results obtained so far, y is the result that the evaluation would return, drawn from the model's current prediction for level s, and P*(·) is the Pareto set of reference-level predictions after the model is updated with a given set of results. HV is the hypervolume of Equation (5), and c(s) is the cost of the evaluation. The numerator is the expected hypervolume gain that the evaluation would produce at the reference level, even when s is an inexpensive method. This gain is small when the method is weakly correlated with the reference level or when the candidate is already well characterized. The operator argmax selects the candidate and level with the highest expected gain per unit cost, denoted x* and s*. A more expensive evaluation is therefore preferred only when its expected value per unit cost exceeds that of the alternatives, and the level is chosen for each candidate rather than by a fixed schedule. Batches are assembled sequentially: after each selection, the model is updated with a simulated result for the chosen evaluation, so that later selections account for the information that it is expected to provide.

Estimating agreement with an experimental reference requires peptides evaluated by both the computational method and the reference assay. Experimental campaigns can begin with a small set of paired evaluations, typically 10–20 peptides, or with user-supplied prior estimates of agreement, which are updated as paired results accumulate. The simulated benchmarks in Section 4.4 assess a different setting: the reference landscape is used to measure optimization performance, but exact reference scores are not available to the optimizers. Table 1 illustrates the resulting allocation across surrogate scoring, Receptor.AI affinity scoring, and molecular dynamics.

The procedure also has limitations. The correlations between levels are learned from peptides evaluated at several levels, so early estimates are uncertain; user-supplied prior estimates of agreement can guide allocation at this stage, but inaccurate estimates may reduce efficiency. Each correlation is assumed to be constant across the design space, so a method that is reliable for some scaffolds and unreliable for others is represented by its average agreement. Peptides are usually evaluated at the reference level because they appeared promising at an inexpensive level, and this selection can inflate the apparent agreement between methods. Correlations between objectives are not modelled, because each objective has a separate model.

Table 1. Allocation of evaluations across methods in an affinity campaign. Evaluations is the number of candidate evaluations the method performs over a campaign; the same candidate may be evaluated by more than one method. Cost per evaluation is the resource consumed by one evaluation of one candidate, expressed in units of a single surrogate evaluation, so the column is a ratio between methods rather than an absolute price. Total cost is the product of the two, that is, the share of the campaign budget the method consumes. The allocation between methods is decided during the run by the rule in Equation (9) rather than fixed in advance, and raw scores from different methods are never compared directly. Because the cost per evaluation rises about as fast as the number of evaluations falls, no single method dominates the budget; the question is which candidates justify the more expensive methods, not how many candidates survive a stage. The counts and costs shown are illustrative of the intended allocation pattern, not production throughput or validated cost-saving estimates.

Evaluation method	Evaluations	Cost per evaluation	Total cost
Surrogate scoring	> 1,000,000	1	$\approx 10^6$
Receptor.AI scoring	> 100,000	10	$\approx 10^6$
Molecular dynamics	100–1,000	10^3	10^5–10^6

Joint affinity and permeability optimization has not been calibrated on an independent benchmark containing matched affinity and permeability measurements for the same peptides. The Vps29 program in Section 4.7.3 provides a prospective series with both endpoints, but it is a single campaign rather than a calibration benchmark. Peptide affinity ranking is evaluated in Section 4.5. Search efficiency is assessed using the methodology described in Section 3.4.7 and the benchmark results are reported in Section 4.4. Auto-Alphabet, which defines the initial substitution options, is described in Section 3.4.5 and evaluated separately in Section 4.4.

3.4.5 Auto-Alphabet

The choice of allowed building blocks determines the size and composition of the design space. Auto-Alphabet defines the position-specific substitution alphabets by selecting candidate building blocks before sequence optimization. In the benchmark described here, it screens 345 residues or other building blocks in two stages. This screening panel is distinct from the much larger parameterized chemistry library described in Section 3.4.6; the latter defines broader chemistry availability and is not exhaustively screened at every position. A voxel-based steric filter removes substitutions that do not fit the local environment. A compact physical score then ranks the remaining substitutions:

$$E_{AA} = E_{LJ} + E_{Coulomb} + E_{H-bond} + E_{desolv} + S_{torsion} \tag{10}$$

Here, E_{LJ} is the Lennard-Jones term for steric and van der Waals complementarity between the substituted building block and the surrounding structure, $E_{Coulomb}$ is the electrostatic interaction between their partial charges, E_{H-bond} is a directional hydrogen-bonding term, E_{desolv} is the desolvation penalty for burying polar and charged atoms, and $S_{torsion}$ is a torsional strain term for the conformation the building block must adopt. The highest-ranked building blocks form the proposed alphabet. Users can inspect the scores and edit the alphabet before optimization.

For evaluation, every candidate substitution at each test position is scored with the full physical workflow, providing a higher-cost computational reference ranking. Auto-Alphabet performance is measured by recall@10: the fraction of the ten highest-ranked substitutions from this reference recovered in the Auto-Alphabet top ten. On the six-target benchmark, a baseline ranks substitutions by fingerprint similarity to the original residue. A separate IL-11 test and repeated full-workflow reference calculations assess additional performance and reference reproducibility (Section 4.4).

3.4.6 Noncanonical chemistry

The substitution alphabets draw on a shared library that makes modified chemistry available to the full workflow. This parameterized library contains more than 12,000 in-stock building blocks and more than 10 million virtual on-demand building blocks. It includes D-isomers, N-methylated residues, noncanonical side chains, beta- and gamma-amino acids, and cross-linker attachment groups.

Entries include the atom types, charges, connectivity, and other parameters required for use across the computational workflow. This common representation allows previously parameterized noncanonical chemistry to enter conformational sampling, docking, scoring, and MD without project-by-project setup for every building block. Predictive accuracy can nevertheless vary for unfamiliar chemotypes or topologies, and computational parameterization should not be interpreted as evidence of equal synthetic accessibility or experimental tractability.

3.4.7 Benchmarks of search and candidate selection

The benchmark evaluates two distinct capabilities: finding sequences with favorable properties and selecting those sequences for experimental validation. These capabilities require separate assessment because an optimizer may find a useful candidate but fail to recommend it if the scoring method ranks it incorrectly.

For a real peptide optimization problem, the best possible sequence is generally unknown. Deep mutational scanning can provide experimental measurements for all combinations of substitutions at a small number of positions [31], but such datasets do not cover the number of variable positions and allowed building blocks considered here. The benchmark therefore uses mathematical sequence–property models with exactly known optima. These controlled models test search and selection under defined conditions, following an approach also used in biophysical sequence optimization [32]. They do not validate predictions of experimental peptide properties.

Each sequence x is assigned a reference objective value:

$$X(x) = X_0 + \sum_i h_i(x_i) + \sum_{(i,j)\in E} J_{ij}(x_i, x_j) \tag{11}$$

Here, $X(x)$ is the reference value for sequence x, and X_0 is a constant offset. The index i denotes a variable position, and x_i is the building block at that position. The function $h_i(x_i)$ describes its individual contribution. The set E contains the interacting position pairs, and $J_{ij}(x_i, x_j)$ describes the contribution of the building-block combination at positions i and j beyond their individual effects. This pairwise contribution represents epistasis [33]. For example, two substitutions that are favorable individually may be unfavorable when combined.

The reference model contains both individual-position and pairwise terms, allowing the benchmark to examine how the corresponding components of the Gaussian-process kernel affect optimization. The model is linear in its coefficients, which supports decomposition of the variance and calibration to specified measurements or summary statistics. This linearity does not, by itself, make the optimal sequence easy to determine.

To keep the optimum exactly computable, the benchmark restricts the connectivity of the interacting position pairs. The computational cost of exact optimization scales as $O(L \cdot q^{(w+1)})$, where L is the number of variable positions, q is the maximum number of allowed building blocks per position, and w is the treewidth of the interaction graph. Treewidth describes how the graph can be decomposed into overlapping groups of interacting positions; smaller values permit more efficient exact optimization. Restricting treewidth allows the optimum to be calculated even when the sequence space contains hundreds of trillions of variants. The restriction therefore concerns the structure of the interactions, rather than the total number of sequences.

The benchmark varies both the strength of pairwise interactions and the distribution of local optima. Epistatic strength is specified by the fraction of objective variance attributable to pairwise interactions. When this fraction is zero, positions contribute independently, and the optimum can be obtained by choosing the best building block at each position. The optimum is unique if every position has a unique best building block. Stronger interactions can introduce local optima, where individual substitutions do not improve the objective even though better sequences exist elsewhere. Landscape families additionally vary the number and depth of traps.

The benchmark also represents systematic errors in candidate evaluation. Repeated evaluations can reduce random noise, but they cannot remove a persistent bias associated with a particular chemical class. Each simulated evaluation method therefore observes a distorted version of the reference objective:

$$Z_{method}(x) = r\,Z_X(x) + \sqrt{1 - r^2}\,\varepsilon(x) \tag{12}$$

Here, $Z_{method}(x)$ is the standardized score assigned to sequence x by the simulated method, and $Z_{X(x)}$ is its standardized reference value. The function $\varepsilon(x)$ specifies a fixed error across sequence space, with similar sequences receiving related errors. It is not resampled at each evaluation. The mixing coefficient r is calibrated to a specified rank correlation between the method and the reference objective. No docking, empirical scoring or free-energy-perturbation calculation is performed in the benchmark. The evaluation methods are analogues, defined only by their cost and their rank correlation with the reference objective; where they are named, the names indicate the class of evaluation each level stands for and carry no implication that the corresponding calculation was run. This construction tests the consequences of persistent ranking errors and distinguishes the value of additional, independent evidence from repeated evaluation using the same biased method.

Table 2 summarizes the benchmark conditions and the aspect of optimizer performance assessed by each.

Performance is reported as the reference value of the best candidate among the k candidates the optimizer recommends, which is the quantity a synthesis decision acts on rather than the best candidate the search happened to evaluate. Recommendations are made using the information available to the optimizer, without access to the reference values. X* denotes the global minimum and X_{med} the median reference value over the landscape; reported values are relative to X_{med}, which is set to zero.

Table 2. Synthetic benchmark conditions and their diagnostic purposes.

Benchmark conditions	Controlled feature of the landscape model	Diagnostic question
Pairwise epistatic landscape model	Fraction of variance attributable to pairwise interactions	How do interactions between substitutions affect search efficiency?
Landscape model with Multiple basins	Number and depth of local optima	Does exploration maintain access to alternative sequence families?
Variable rank correlation evaluation methods	Persistent evaluation bias and specified rank correlation with the reference objective	How far does the rank correlation of the evaluation method limit the result that can be reached?

These benchmarks assess performance within the specified class of pairwise interaction models with restricted treewidth. They do not establish performance across all peptide sequences–property relationships. They also do not test whether the reference objective predicts experimental success. An optimizer may identify the best sequences under a computational objective while those sequences perform poorly in the laboratory. Agreement between computational objectives and experimental outcomes requires separate validation.

3.5 Physics-based affinity scoring

Structural prediction provides candidate complexes, but optimization also requires a way to compare related peptides. The proprietary scoring supplies a relative affinity score by estimating binding free energy from the complex and isolated receptor and ligand states:

$$\Delta G_{bind} = G_{complex} - \left(G_{receptor} + G_{ligand}\right) \tag{13}$$

$$G = E_{MM} + G_{solv} - TS \tag{14}$$

$$E_{MM} = E_{bond} + E_{vdW} + E_{elec} \tag{15}$$

$$G_{solv} = G_{GB} + G_{SA} \tag{16}$$

E_{MM} contains bonded, van der Waals, and electrostatic terms. G_{solv} contains polar generalized-Born and nonpolar surface-area contributions. T is temperature, and S is entropy, which is treated approximately.

The resulting score is used to rank related peptides within a series. It is not an absolute dissociation constant or a rigorous alchemical free-energy calculation. Within ForceFold, it contributes to structural assessment alongside other scores and does not independently select among distinct binding modes. In the affinity-ranking benchmark, the unbound receptor is evaluated in the bound geometry, while the isolated peptide is minimized separately rather than sampled as an independent free-state ensemble.

The affinity-ranking calculations start from crystallographic/reference protein–peptide complex structures rather than ForceFold-predicted geometries. The benchmark [34] contains 191 peptides in ten series across eight targets: MCL1, MDM2, MDMX, FcRn, KEAP1, Calmodulin, Src SH2, and uPA (Table 3).

Spearman rank correlation is calculated within each series and summarized across the ten series, so differences in affinity scale between unrelated targets do not dominate the result. Scores are summarized using the best-conformer scheme used in the final workflow; uncertainty in the summary result is reported as the standard error across the ten series.

The same peptide sets are evaluated with Schrödinger Prime MM-GBSA and Glide SP. The comparator results are treated as practical ranking-pipeline comparisons rather than as isolated tests of the scoring functions.

Table 3. Composition of the peptide affinity-ranking benchmark.

Target	Series	Peptide type	Peptides, n
MCL1	1	Linear	12
MDM2	2	Linear	11 + 27
MDMX	1	Linear	11
FcRn	1	Linear	72
KEAP1	2	Cyclic + linear	21 total
Calmodulin	1	Linear	17
Src SH2	1	Linear	8
uPA	1	Constrained	12
Total: 8 targets	10	Linear and cyclic	191

3.6 Passive permeability prediction

Chemical changes selected for binding can also affect membrane permeation. Affinity and passive permeability are therefore modeled separately so that both can be considered during optimization. For cyclic peptides, environment-dependent conformation can alter exposed polarity, intramolecular hydrogen bonding, charge shielding, and compactness [35–37]. The permeability models therefore include conformational information as well as molecular connectivity.

3.6.1 Data and conformational representation

Permeability modeling requires both measured transport data and a representation of peptide conformations in different environments. PAMPA provides the experimental reference by measuring passive diffusion across an artificial lipid-containing barrier [38]. PAMPA is treated here as an operational assay of apparent passive permeability rather than a direct measurement of an intrinsic phospholipid-bilayer permeability coefficient; it does not directly measure cellular uptake or oral bioavailability. The public permeability benchmark [39] is constructed from the

CycPeptMPDB [40] PAMPA export used for this analysis, which spans 42 source-publication identifiers and contains 7,298 records. Excluding 271 assay-floor records reported at logPerm = -10 leaves 7,027 PAMPA measurements. Source-level quality control then excludes all 1,772 PAMPA rows from 13 publications that failed the predefined source-level consistency checks. Eleven additional publications contributed too few measurements to be evaluated at source level and were retained.

After source exclusion and Sequence deduplication, 5,231 distinct public PAMPA measurements remain for model development and retrospective evaluation. The source filter is applied before construction of the training, validation, and test sets and before the retrospective benchmarks in Figures 18–20. Measurements from different source publications can still reflect assay-specific variability, which limits direct comparability across sources. These public data are supplemented by approximately 3,000 proprietary Receptor.AI PAMPA records. Across the broader public and proprietary collections, 176 distinct noncanonical amino acids occur, and about 70% of peptides contain at least one NCAA.

The filtered public set remains concentrated in six- and seven-monomer Circle peptides and ten-monomer Lariat peptides, with much lower representation elsewhere (Figure 7).

After source-level quality filtering and Sequence deduplication, the 5,231 public measurements are partitioned by Bemis–Murcko scaffold [41] into 70% training (n = 3,662), 10% validation (n = 523), and 20% held-out test (n = 1,046). Unless otherwise stated, the AI benchmark, AI–physics consistency analysis, and confidence-routing analysis below use this same held-out test fold.

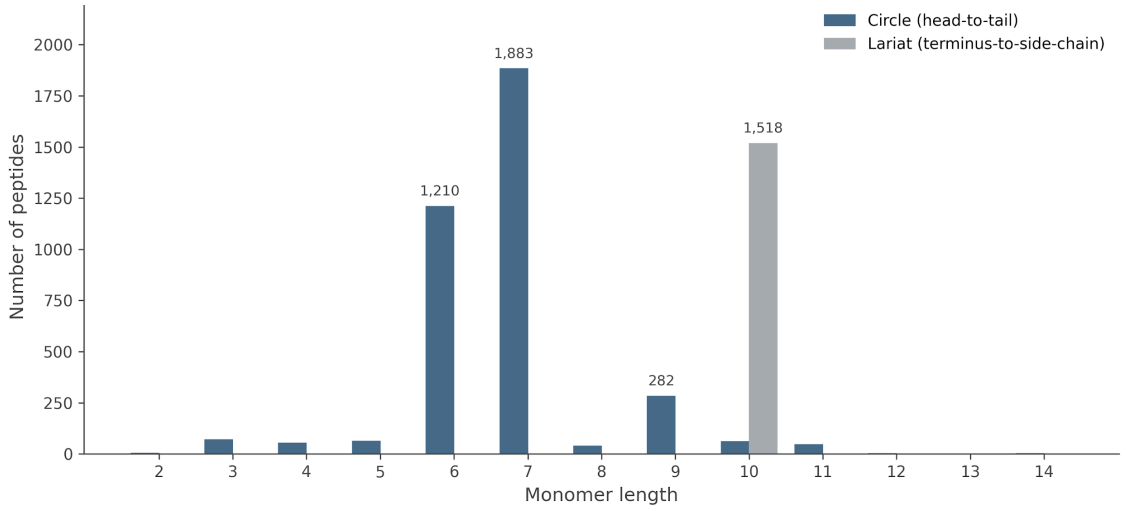

Figure 7. Coverage of the source-QC-filtered public CycPeptMPDB PAMPA collection by monomer length and topology (n = 5,231).

Each peptide is represented by a fast physics-based conformational generator in aqueous-like, membrane-interface-like, and membrane-interior-like implicit solvents. The objective is broad, reproducible coverage of physically reasonable conformations that expose permeability-relevant structural variation; sampling is diversity-driven rather than exhaustive, and the resulting conformers are not treated as Boltzmann-weighted equilibrium populations. Redundant conformations are removed before use.

The same conformational representation is used during AI training and inference and by both physics-based methods. Changing the evaluation method therefore changes the scoring procedure, not the underlying conformer set. Higher-fidelity physics can improve the energetic description of sampled structures, but it cannot recover an important conformation that was not sampled.

3.6.2 AI and physics-based methods

The conformational ensembles are evaluated at different computational costs according to the candidate and the reliability of the AI prediction. The permeability AI model is a graph neural network that combines two-dimensional molecular information with three-dimensional descriptors from the environment-specific ensembles. It predicts PAMPA logPerm and a confidence estimate. Published AI approaches to cyclic-peptide permeability have been

benchmarked systematically elsewhere [42]. The confidence estimate serves as an operational indicator of model applicability and is complemented by explicit checks for sparsely represented chemistry, peptide length, and topology. It is not a calibrated probability of correctness or a complete out-of-distribution detector [43].

The full passive membrane permeability prediction and optimization pipeline is shown in Figure 8. Lower-confidence candidates or candidates with weak chemistry/topology coverage can be evaluated by two mechanistically distinct structure-based physics tiers.

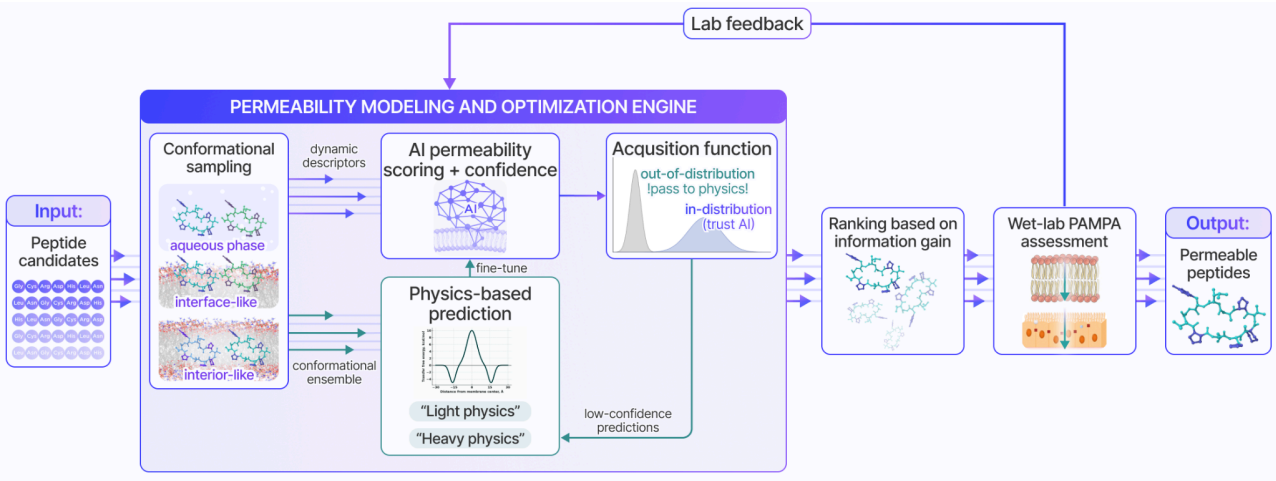

Figure 8. Multi-fidelity architecture for passive permeability assessment. Each peptide candidate undergoes conformational sampling in aqueous-like, membrane-interface-like, and membrane-interior-like conditions. The AI model scores candidates for permeability and confidence, while "Light Physics" and "Heavy Physics" are applied selectively as additional structure-based evidence. The available signals are combined by the QuorumMap acquisition function, followed by PAMPA measurement and project-specific model adaptation.

"Light Physics" uses a continuum-electrostatics membrane model with a force-field description of the peptide and typically requires seconds per peptide. "Heavy Physics" uses the same conformations and membrane representation with a quantum-level description of surface charge and polarity and typically requires one to several hours per peptide [44].

Both tiers return aggregate relative membrane-transfer scores from the same pre-generated conformational ensemble. These scores primarily capture the transfer-thermodynamic contribution to passive permeation; they are not absolute permeability coefficients, are not necessarily a single free-energy barrier in physical units, and do not explicitly model position-dependent diffusivity or full permeation kinetics. Raw physics scores are not intrinsically on the PAMPA logPerm scale. When an absolute PAMPA-scale estimate is required, the raw physics score s from each tier is mapped to a calibrated PAMPA logPerm prediction using a two-parameter affine transformation:

$$\widehat{logPerm} = a + bs \tag{17}$$

Here, the left-hand side is the calibrated predicted PAMPA logPerm; s is the raw physics score; a is the fitted intercept; and b is the fitted slope. The parameters a and b are fitted on the validation partition only, frozen before evaluation, and then applied unchanged to held-out peptides. Raw membrane-transfer scores and calibrated PAMPA-scale predictions are reported separately.

Operational routing uses the normalized AI confidence score c together with explicit chemistry/topology coverage checks; the confidence cutoffs are operational defaults rather than calibrated probabilities. For $c \geq 0.80$ in well-covered chemistry, the AI result can be used without additional physics. For $0.50 \leq c < 0.80$, or for weakly represented or new chemistry or topology, "Light Physics" is added. For $c < 0.50$, "Light Physics" is run first; "Heavy Physics" is used by default for $c < 0.30$ and selectively for $0.30 \leq c < 0.50$ when the local chemistry remains insufficiently resolved after "Light Physics," when disagreement between the AI model and "Light Physics" is large enough to change candidate ranking or advancement decisions, or for selected anchors, checks, and model-improvement points. Available AI, "Light Physics," and "Heavy Physics" signals are returned to the QuorumMap acquisition function, which combines them with model confidence and local chemistry coverage to rank candidates.

3.6.3 Evaluation

Evaluation examines how well the models rank measured permeability and how performance changes under distribution shift. AI performance is assessed primarily by Spearman correlation on scaffold-held-out and distribution-shift tests. “Light Physics” and “Heavy Physics” scores are compared with PAMPA using Pearson correlation on the same source-QC-filtered, Sequence-deduplicated public corpus: 3,529 peptides with 4–9 monomers and 1,629 peptides with 10–15 monomers; 73 two- and three-monomer records lie outside these displayed bins. The 4–9 group contains 3,528 Circle and one Lariat peptide, while the 10–15 group contains 111 Circle and 1,518 Lariat peptides. Because a within-subset correlation cannot be estimated from a single peptide, the 4–9-monomer Lariat subgroup is omitted from the topology-stratified analysis in Figure 18B. These subsets therefore differ strongly in topology as well as size. Pearson and Spearman values measure different aspects of agreement and are reported separately.

Validation is strongest for cyclic peptides. Linear peptides and new or multicyclic topologies require separate evaluation before equivalent performance can be claimed.

4 Results

We evaluated the computational methods separately to distinguish complex-prediction accuracy, affinity ranking, search efficiency, and permeability prediction. We then examined their combined use in experimental peptide programs. This order connects the performance of individual methods with the outcomes of the complete workflow.

4.1 ArtiDock-P docking performance

We first examine docking as a component of the structural workflow. The base ArtiDock small-molecule model was evaluated before integration into ForceFold. On the PLINDER holo benchmark, 48% of its poses were within 2 Å RMSD of the experimental pose, compared with 39% for Glide, 38% for AutoDock Vina, and 35% for AutoDock. On the PLINDER MLSB apo/predicted-receptor benchmark, success was 23% for ArtiDock and 8–13% for the classical engines (Figure 9). In the internal PoseBusters evaluation, the raw ArtiDock distance output achieved 71.9% within 2 Å RMSD; UFF refinement increased physical validity to 97.4%, while 68.8% of the refined poses remained within 2 Å.

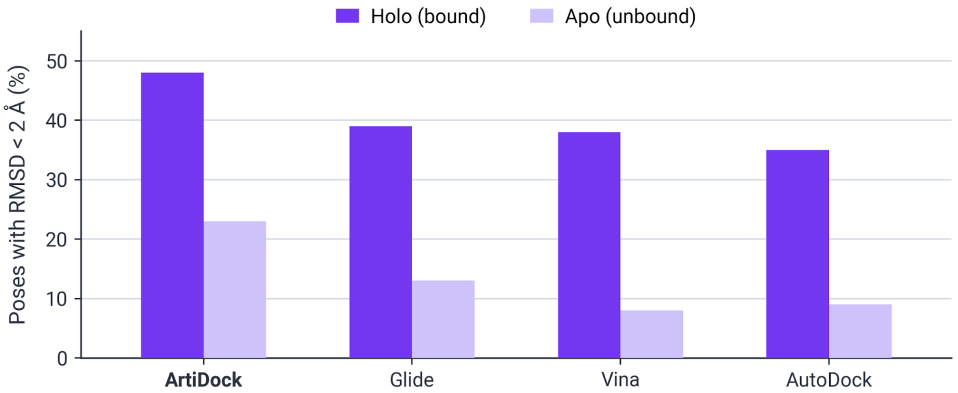

Figure 9. ArtiDock small-molecule docking performance on PLINDER holo and MLSB apo/predicted-receptor benchmarks. Bars show the fraction of predicted poses within 2 Å RMSD of the experimental pose.

On Docking985 [13], ArtiDock-P recovered a CAPRI Acceptable-or-better binding mode in 73% of complexes. The distributional summary in Figure 10 shows the accompanying whole-peptide RMSD and fraction of native contacts (fNat) distributions overall and after stratification by peptide topology and length. In the Vina X-ray-seeded condition, Vina is initialized from the experimentally observed crystallographic peptide geometry; in the Vina generated condition, Vina instead starts from a computationally generated peptide conformer. ArtiDock-P + UFF denotes the ArtiDock-P output after UFF relaxation. Because fNat is bounded between 0 and 1 and only five-number summary statistics are available for these runs, the fNat panels show the median and interquartile range rather than the full reported whisker range. These distributions complement, rather than replace, the headline CAPRI and interface-RMSD success rates.

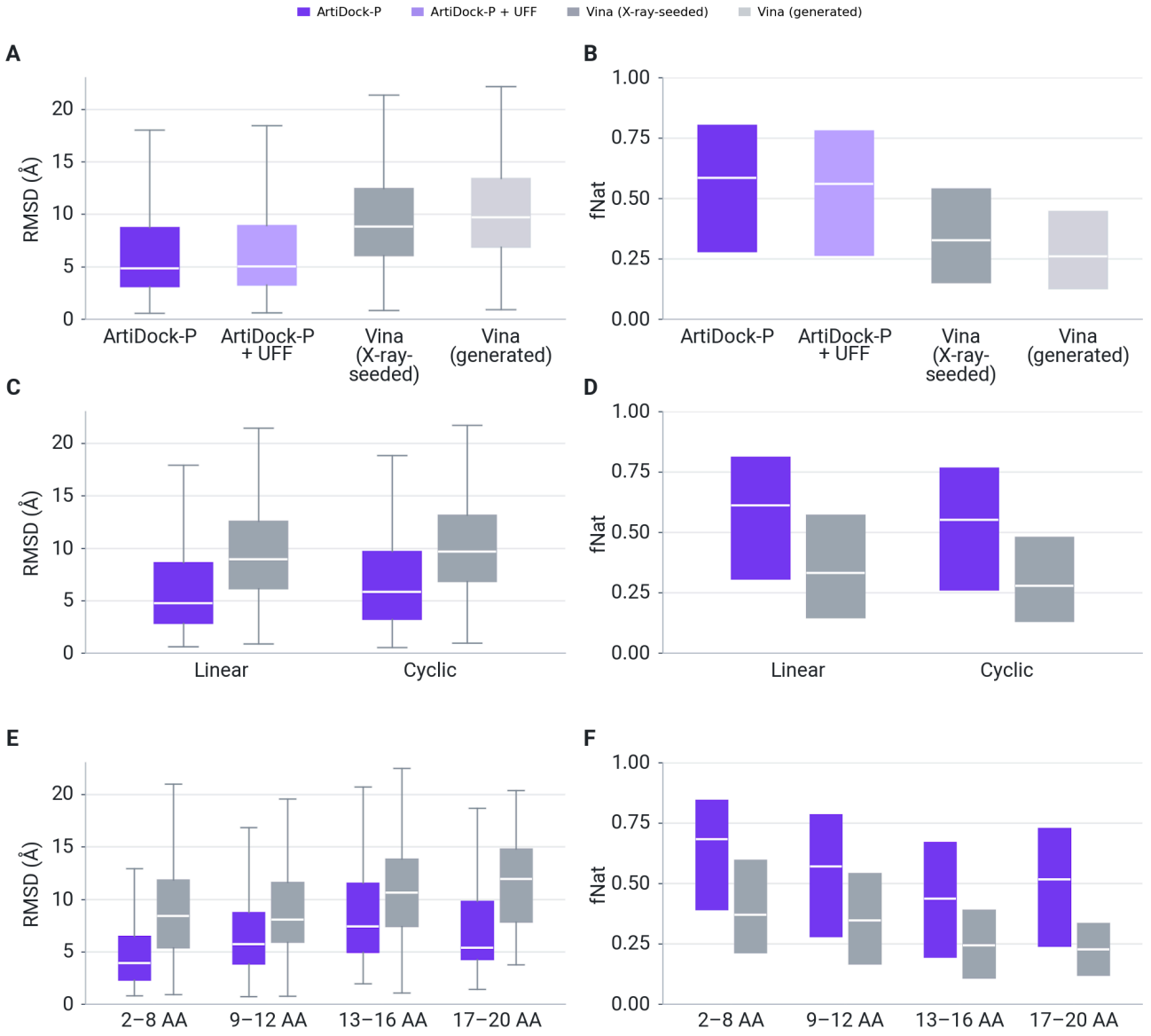

Figure 10. Standalone ArtiDock-P peptide docking benchmark (Docking985). (A) Overall whole-peptide RMSD for ArtiDock-P, ArtiDock-P + UFF, Vina (X-ray-seeded), and Vina (generated). (B) Overall fNat for the same four conditions. (C) Whole-peptide RMSD stratified by peptide topology (linear versus cyclic) for ArtiDock-P and Vina. (D) fNat stratified by peptide topology for the same methods. (E) Whole-peptide RMSD stratified by peptide length. (F) fNat stratified by the same length bins. RMSD panels (A, C, E) use box-and-whisker summaries: boxes span the interquartile range (Q1–Q3), the horizontal white line denotes the median, and whiskers show the reported low and high whiskers. fNat panels (B, D, F) use the same box representation for Q1–Q3 with a horizontal white median line, but omit whiskers because fNat is bounded to [0,1]. "X-ray-seeded" denotes Vina initialized from the experimentally observed crystallographic peptide geometry, whereas "generated" denotes Vina initialized from a computationally generated peptide conformer; "ArtiDock-P + UFF" denotes UFF relaxation of the ArtiDock-P output.

4.2 ForceFold complex-prediction accuracy

The BMI-200 benchmark source [45] contains 202 complexes. After excluding antibody/designed-binder cases (n = 24) and MHC cases (n = 6), 172 protein–peptide complexes remained, spanning 120 distinct receptor targets. Four configurations were evaluated on these same 172 complexes with matched random seeds: unconditioned co-folding, physical search without co-folding, co-folding supplied with physical contacts through its native restraint input, and complete ForceFold. Each system's top-ranked prediction was assessed against the experimental structure using CAPRI-style peptide-interface criteria [46]; dedicated protein-peptide complex-prediction benchmarks follow comparable conventions [47].

Top-1 Acceptable-or-better accuracy was 56.8% for unconditioned co-folding, 59.5% for physical search, 66.3% for native-restraint conditioning, and 74.4% for ForceFold (Figure 11). The comparison with native-restraint conditioning tests two ways of supplying the same underlying physical-search evidence to the same co-folding model. The conditioning interfaces differ in how that evidence is represented and therefore should not be interpreted as an isolated test of a single encoding variable. ForceFold improved accuracy by 8.1 percentage points, with a 95% confidence interval of 3.3–12.8 points.

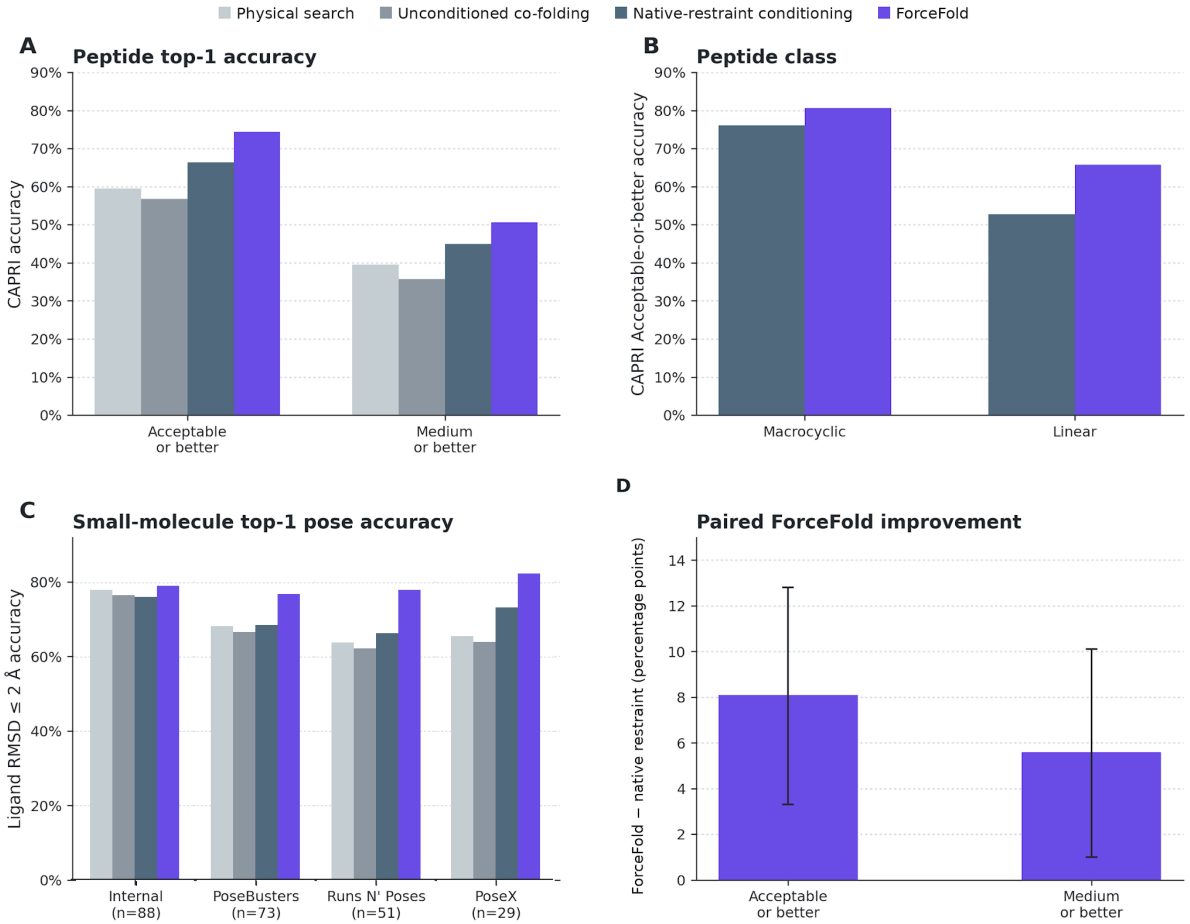

Figure 11. ForceFold benchmark comparisons. (A) Peptide-complex prediction accuracy for the four configurations at the CAPRI Acceptable-or-better and Medium-or-better thresholds (n = 172). (B) Peptide-complex prediction accuracy at the Acceptable-or-better threshold, split into macrocyclic and linear peptides, for ForceFold and the built-in restraint route. (C) Small-molecule pose-prediction accuracy at ligand RMSD ≤ 2 Å on the internal collection and three post-cutoff benchmarks, for physical search, co-folding alone, the built-in restraint route, and ForceFold. Whiskers in panels B and C denote confidence intervals. (D) Paired ForceFold advantage over the built-in restraint route at the two peptide thresholds; whiskers show 95% CIs for the paired difference.

At the Medium-or-better threshold, ForceFold reached 50.6%, compared with 45.0% for native-restraint conditioning. The difference was 5.6 percentage points, with a 95% confidence interval of 1.0–10.1 points.

For linear peptides, Acceptable-or-better accuracy was 65.7% for ForceFold and 52.8% for native-restraint conditioning. For macrocycles, the corresponding values were 80.7% and 76.0%. The larger gain for linear peptides is consistent with a greater benefit in the more conformationally heterogeneous subset, although this benchmark does not isolate ligand flexibility from peptide length, chemistry, or target composition.

4.3 Effect of structural information on ForceFold

To understand where ForceFold gains accuracy, we examined the available structural information and tracked binding modes through the workflow. ForceFold's improvement over native-restraint conditioning increased with the difference between the supplied receptor and its experimentally observed bound geometry. For post-cutoff peptides, the

improvement also increased as the receptor interface became less similar to structures available before the co-folding model's training cutoff. These are associations across difficulty groups, not controlled tests of individual causes. The largest receptor-mismatch groups also contain few systems.

Tracking binding modes through the workflow identifies where accurate structures are lost (Figure 12). An Acceptable-or-better mode was present in 95.3% of complete search pools, retained after pose-pool reduction in 90.5%, and selected as the final top-ranked prediction in 74.4% of systems. At the Medium-or-better threshold, the corresponding values were 84.1%, 75.0%, and 50.6%.

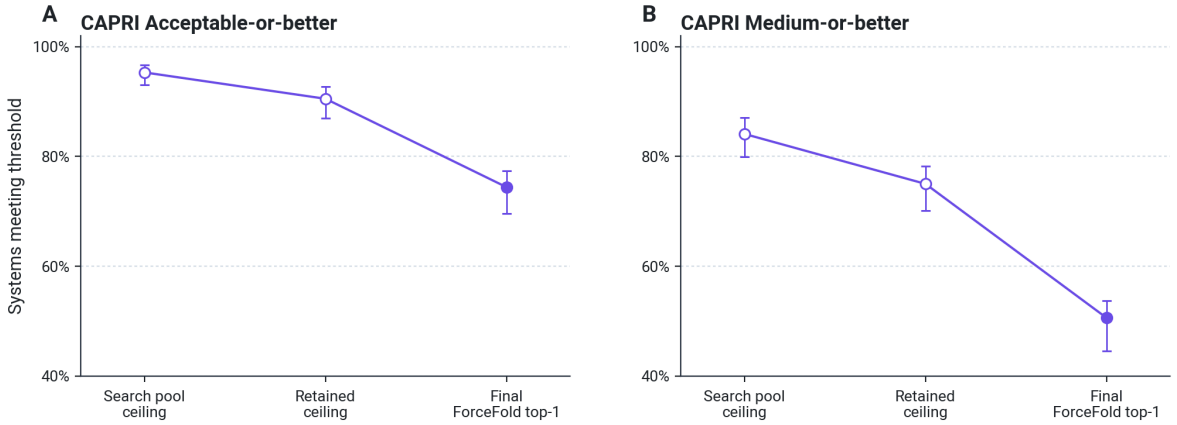

Figure 12. Search, retention and final-selection diagnostics on the peptide benchmark (n = 172). Search-pool and retained-mode values are retrospective diagnostic ceilings and are not prospective ranking outputs; the final value is the ForceFold top-1 prediction.

The search-pool and retained-mode values are retrospective measures: they show that a qualifying mode was available, not that the method selected it. The largest decrease in recovery occurs after mode retention, across conditioning, co-folding, and final selection. This diagnostic does not isolate final ranking as the downstream failure mechanism; it points instead to representation, reconstruction, and discrimination of retained binding-mode families.

Replacing the retained binding modes with those from a different complex reduced Acceptable-or-better accuracy to 57.0%, close to the 56.8% of unconditioned co-folding. The improvement thus depended on system-specific structural information. Adaptor ablation showed the largest separately resolved contribution from the pair adaptor, a smaller contribution from the point adaptor, and no clearly resolved effect from diffusion conditioning in that experiment.

4.4 Search efficiency and Auto-Alphabet

4.4.1 QuorumMap search and candidate-selection efficiency

We evaluated search and candidate selection using three controlled sequence–property landscapes constructed as described in Section 3.4.7. The first benchmark assessed whether the optimizer could identify distinct favorable sequence families within a predominantly inactive library. The second assessed optimization within an established favorable sequence family. The third used a chemical space derived from a published FcRn peptide series, with experimental measurements used to constrain the reference values. These benchmarks assess optimization separately from the accuracy of structural, affinity, and permeability predictions.

The hit discovery benchmark contained five variable positions with 100 building blocks per position, giving 10^{10} possible sequences. Building blocks were sampled independently from the internal library of noncanonical amino acids to create a chemically diverse search space. A common inactive reference value was assigned to 95% of the landscape. The remaining space contained six separated favorable regions, each with a local optimum of a different reference value.

The lead optimization benchmark contained four variable positions with 64 building blocks per position. One favorable region occupied 99.5% of the sequence space and contained additional local optima. Construction of the FcRn landscape is described below.

Figure 13 shows each landscape as a two dimensional map and a three dimensional surface. These visualizations represent the arrangement of favorable regions and local optima rather than a projection of every sequence. They preserve the measured numbers and depths of optima and traps, the sizes of their associated regions, and the chemical distances between them; their placement in the plane is fitted. Reference values are expressed relative to the median of each library, which is set to zero. Lower values indicate better performance. Values can be compared within each landscape but do not represent experimental binding free energies.

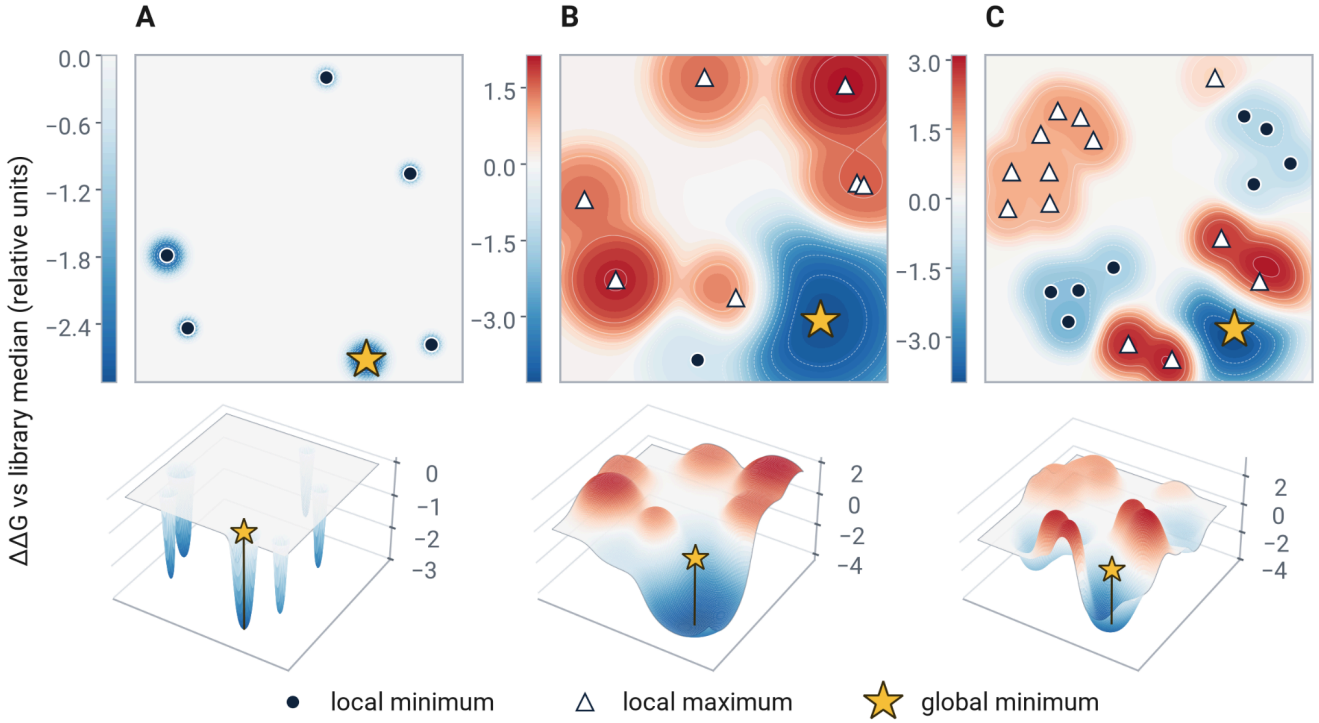

Figure 13. Controlled sequence–property landscapes used in the optimization benchmark, each shown as a two-dimensional map (top) above the same landscape as a three-dimensional surface (bottom). (A) Hit-discovery landscape containing several separated favorable basins in a predominantly inactive sequence space. (B) Lead-optimization landscape dominated by one broad chemical basin with internal local structure. (C) FcRn-derived lead-optimization landscape. These are representations of the measured landscape topology, not projections of the underlying sequence spaces: the number of optima and traps, their depths, the size of the basin draining to each, and the chemical distances between them are measured, and only their placement in the plane is fitted. On the map, filled circles denote local minima, open triangles local maxima, and the star the global minimum, which is also pinned on the surface below. Colour on the map and height on the surface carry the same reference value relative to the library median; each column has its own scale, read from the bar beside the map or the z axis beside the surface.

Hit discovery performance was measured by the number of distinct favorable regions identified as the number of inexpensive evaluations increased (Figure 14).

QuorumMap, which uses a statistical surrogate model to guide search, recovered four of the six optima after approximately 1.3×10^4 evaluations and all six within 10^6 evaluations. At the full budget, genetic algorithms and simulated annealing each recovered four optima, simple Bayesian optimization recovered three, random sampling recovered two, and greedy search recovered one.

Methods that recovered fewer optima preferentially sampled broader favorable regions, which contained more sequences. Because all six favorable regions are known in this benchmark, their recovery directly measures coverage of the defined favorable sequence families. This coverage cannot generally be measured in a prospective experimental screen, where the complete set of favorable families is unknown.

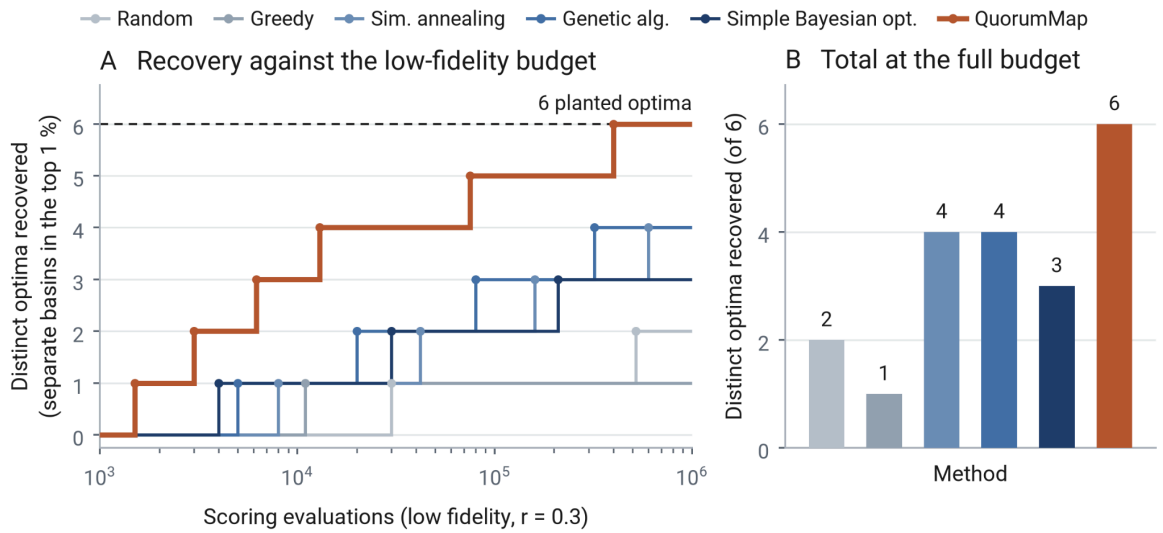

Figure 14. Hit-discovery benchmark on the landscape of Figure 13A. (A) Number of distinct favorable basins recovered as a function of low-fidelity evaluations. Six separated optima are present in the landscape. (B) Total number recovered by each search method at the full 10^6-evaluation budget.

Lead optimization performance was measured using the reference value of the best sequence among the 25 candidates recommended by each method at a given budget (Figure 15). Three simulated evaluation methods were used, with relative costs of 1, 10, and 100 and rank correlations of 0.3, 0.5, and 0.7 with the reference objective, respectively. Budgets are reported relative to a campaign cost of 10,000 units. The figure labels docking, Receptor.AI affinity scoring, and FEP identify the evaluation classes represented by these simulated methods. No docking, scoring, or free energy perturbation calculations were performed.

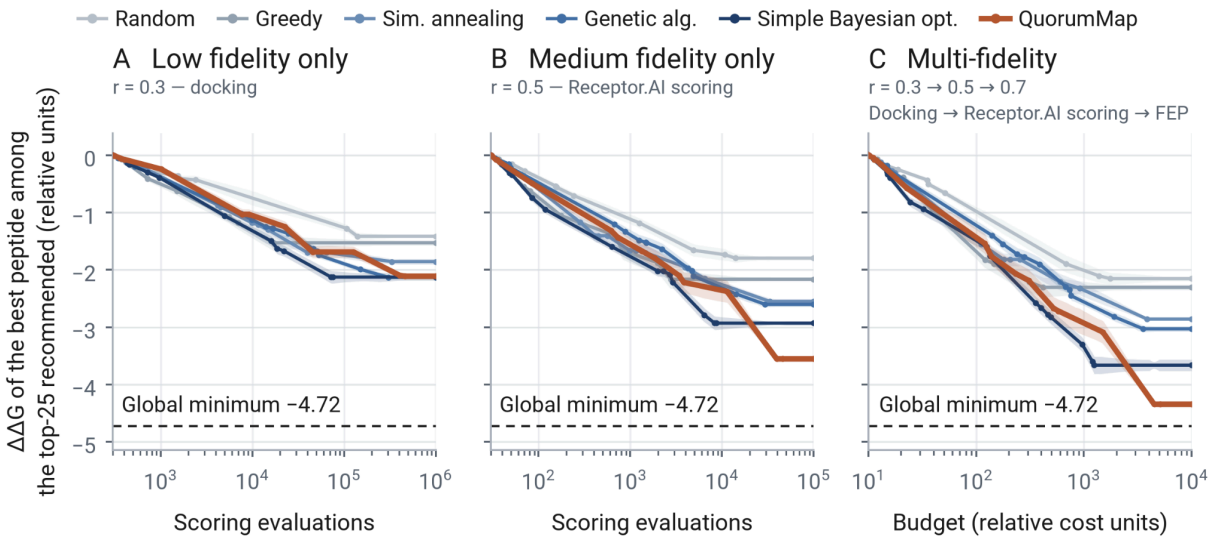

Figure 15. Lead-optimization benchmark on the landscape of Figure 13B. Reference value of the best sequence among the 25 candidates recommended by each method. (A) Low-fidelity evaluation, rank correlation 0.3. (B) Medium-fidelity evaluation, rank correlation 0.5, at matched evaluation cost. (C) Multi-fidelity search using three evaluation levels with rank correlations of 0.3, 0.5 and 0.7. The panels label these levels docking, Receptor.AI scoring and FEP to indicate the class of evaluation each represents; no such calculation was run. The dashed line denotes the global minimum; bands show variation between repeated runs.

When only the least expensive and least accurate evaluator was used, all search methods reached plateaus well above the global minimum. At the same total evaluation cost, the intermediate evaluator improved recommendation quality. QuorumMap reached a reference value of −3.6, compared with the global minimum of −4.72. Performance curves

crossed repeatedly when the least accurate evaluator was used, indicating that comparisons at a single intermediate budget could give a different method ranking from comparisons across the full budget range.

Using all three evaluation methods improved the QuorumMap result to −4.3. The comparator methods used a fixed schedule to advance candidates between evaluation levels, whereas QuorumMap selected the evaluation level separately for individual candidates. No method reached the global minimum of −4.72. The most accurate simulated evaluator still contained systematic ranking error, with a rank correlation of 0.7 against the reference objective. None of the optimizers therefore had access to exact reference scores.

The third benchmark used a published series of FcRn binding peptides cyclized through a disulfide bond. Seventy-two experimentally measured analogues were aligned, and four positions present in every analogue were selected as variables. The initial set of allowed building blocks at each position included only those observed in analogues with $IC_{50} \leq 5$ μM. These sets were then expanded with chemically related building blocks from the parameterized library.

The resulting space contained 64 building blocks per position and approximately 1.7×10^7 sequences. Selecting building blocks from active analogues enriched the space for potential binders, making this a lead optimization benchmark. The reference landscape was constrained to reproduce the measured values associated with the 22 distinct sequence combinations represented at the four variable positions.

Using all three simulated evaluation methods, QuorumMap reached a reference value of −4.0, compared with a global minimum of −4.47. Comparator methods reached values between −1.9 and −3.4 (Figure 16).

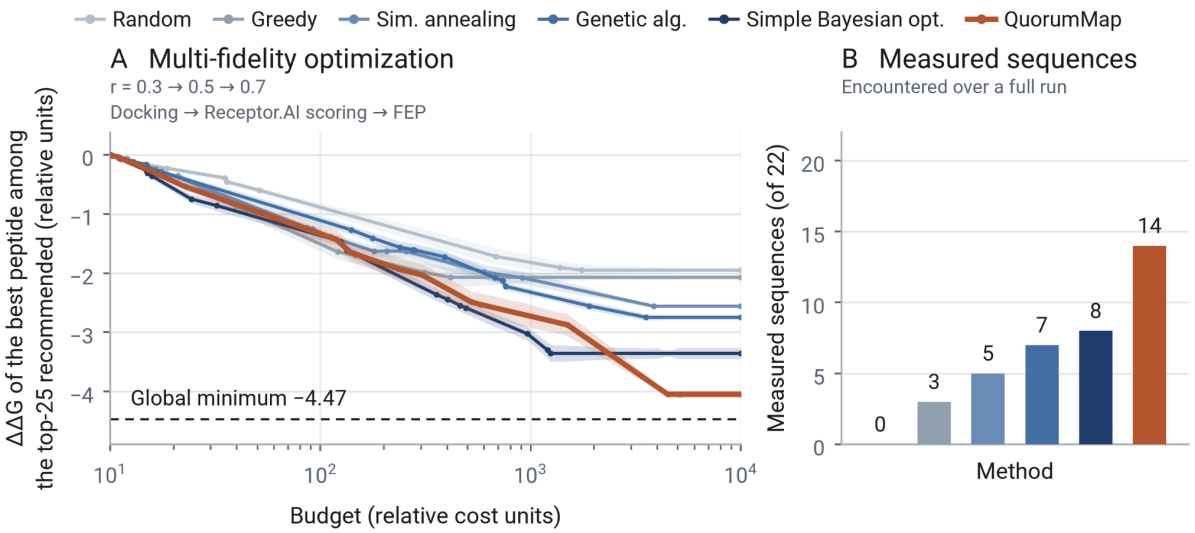

Figure 16. FcRn-anchored lead-optimization benchmark on the landscape of Figure 13C. (A) Multi-fidelity optimization as a function of cumulative evaluation budget. The dashed line denotes the global minimum of the constructed reference landscape. (B) Number of the 22 experimentally measured sequence combinations encountered during a complete run by each search method.

During a complete run, QuorumMap evaluated 14 of the 22 experimentally measured sequence combinations embedded in the landscape. Simple Bayesian optimization evaluated eight, the genetic algorithm seven, simulated annealing five, greedy search three, and random search none. This measure complements the reference score comparison by showing whether the search reached experimentally characterized sequences, in addition to favorable sequences assigned values by the computational model.

These results assess search and candidate selection within the controlled model family described in Section 3.4.7. The reference landscapes contain contributions from individual positions and interacting position pairs, with restrictions on the interaction graph. The surrogate kernel was designed to represent these same contributions. The benchmarks therefore assess optimization when the surrogate structure is compatible with the reference model. They do not establish robustness when that structure is incorrect or performance across arbitrary peptide sequence–property relationships.

For the FcRn benchmark, the substitution sets and 22 reference measurements are based on experimental data. Reference values for the remaining sequences in the approximately 1.7×10^7 sequence space were generated computationally. The benchmark therefore assesses optimization in a chemical space constrained by experimental observations; it does not reproduce the experimental outcome of the original FcRn campaign.

4.4.2 Evaluation of Auto-Alphabet substitution ranking

Auto-Alphabet was evaluated for its ability to identify favorable substitutions before sequence optimization. Candidate substitutions were ranked using the simplified Auto-Alphabet scoring function and the full, more computationally expensive scoring function. The latter provided the reference ranking. Agreement was measured by recall@10, defined as the number of substitutions shared by the two normalized top ten lists.

To determine whether Auto-Alphabet improves on selection by chemical similarity alone, the same substitutions were also ranked by molecular fingerprint similarity to the original residue. Across six targets, mean recall@10 was 0.261 for Auto-Alphabet and 0.106 for similarity ranking. Agreement with the reference was therefore 2.5 times higher for Auto-Alphabet. In a separate IL-11 test, Auto-Alphabet achieved recall@10 of 0.338. A corresponding similarity comparison was not reported for this test.

The reproducibility of the reference ranking was assessed by repeating the reference calculations. Spearman correlations between repeated rankings ranged from 0.21 to 0.65, and the corresponding top ten lists shared no more than three substitutions. Thus, the reference rankings themselves varied substantially between calculations.

These results measure agreement with a variable computational reference, not with experimentally determined substitution preferences. Auto-Alphabet provides an inexpensive way to construct the initial position-specific substitution alphabets. Selected peptide candidates still require subsequent structural assessment and scoring with the full scoring function.

4.5 Affinity ranking within peptide series

The benchmark asks whether the scoring implementation preserves affinity ordering once a crystallographic/reference protein–peptide complex geometry is available. Across 191 peptides in ten series spanning eight targets, the Receptor.AI scoring achieved a mean within-series Spearman correlation of 0.524 ± 0.168. The corresponding values were 0.432 ± 0.240 for Schrödinger Prime MM-GBSA and 0.322 ± 0.290 for Glide SP. Uncertainty is reported as the standard error between the ten series (Figure 17).

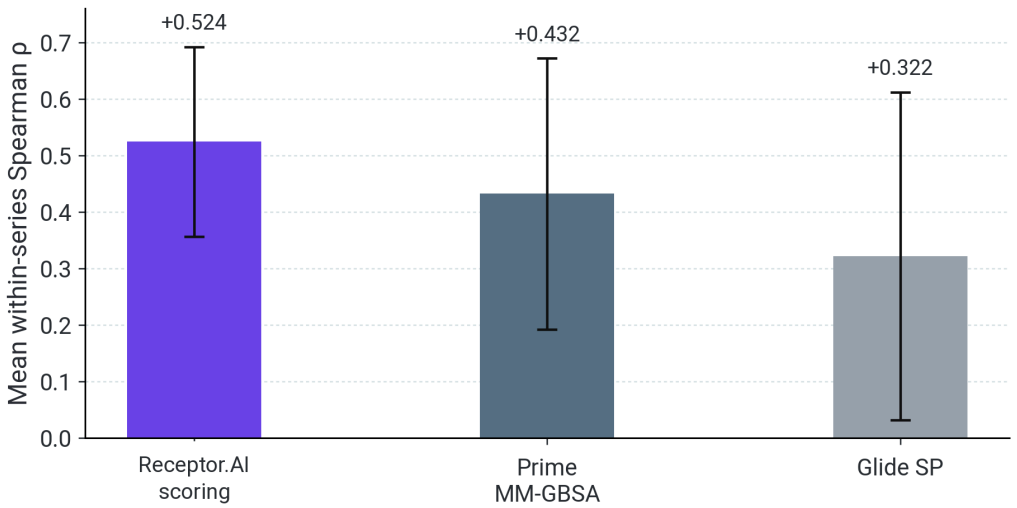

Figure 17. Peptide affinity-ranking benchmark on crystallographic protein–peptide complex structures (191 peptides across n = 10 series across 8 targets). Best-conformer results using all five repeats are shown. Bars report the mean within-series Spearman ρ across the ten peptide series. Whiskers show the standard deviation between series.

4.6 Passive permeability prediction

Permeability evaluation proceeds from AI benchmark performance to conformational and physics-based validation, cross-tier consistency, and confidence-based routing.

4.6.1 Scaffold-held-out and distribution-shift tests

We first tested the AI model on peptides with scaffolds excluded from training, then examined changes in the peptide-length distribution. On the held-out test set (n = 1,046), the AI model achieved Spearman ρ = 0.71, mean absolute error (MAE) = 0.52 log units, and mean squared error (MSE) = 0.42. The earlier chemistry-stratified NCAA/canonical subgroup counts are not carried forward after source-level filtering.

Performance was also assessed across peptide-length distributions. Spearman correlation decreased from 0.71 on held-out peptides from length classes represented in training to 0.64 on held-out 8–9-mers and 0.59 on 11–12-mers beyond the training pool. The DMPNN values of 0.68, 0.46, and 0.33 are taken from the published systematic benchmark of AI methods for cyclic peptide permeability [42]; the baseline was not retrained by us. Our model was evaluated under the corresponding length-defined training and evaluation conditions (Figure 18).

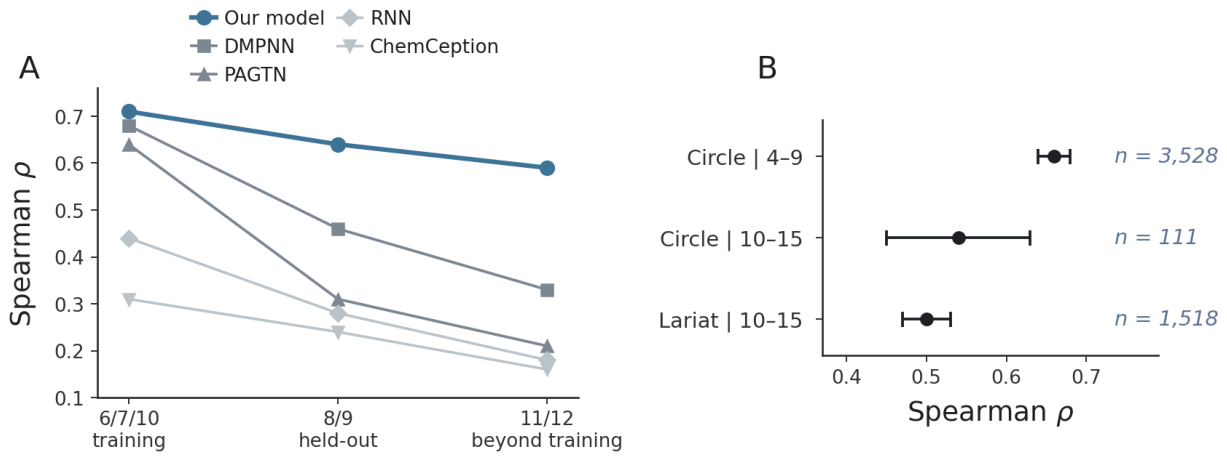

Figure 18. Generalization across peptide-length strata and topology-stratified performance after source-level PAMPA quality filtering. (A) Spearman ρ for our model and published CycPeptMPDB reference models across held-out peptides from training-represented length classes, held-out 8–9-mers, and 11–12-mers beyond training. (B) Spearman ρ with 95% confidence intervals for topology–length subsets with sufficient sample size after filtering.

4.6.2 Conformational descriptors and physical scores

To assess the physical information underlying the permeability models, we compared conformational descriptors with MD and physical scores with PAMPA measurements. Six production conformational descriptors were compared with explicit-solvent all-atom MD for approximately 3,000 cyclic peptides. The descriptors were aqueous and membrane-like polar surface area, environment-dependent change in polar surface area, intramolecular hydrogen-bond occupancy, polarity shielding, and radius of gyration. Pearson correlations ranged from 0.59 to 0.70. MD is used here as a higher-fidelity offline method-development reference rather than as experimental ground truth. MD-derived descriptor values are not supplied to the permeability AI model, “Light Physics,” or “Heavy Physics”. The comparison assesses consistency with the MD-derived descriptor ensemble; it does not establish that either conformational ensemble is an exact equilibrium representation.

After source-level quality filtering, the 4–9-monomer physics benchmark comprises 3,529 peptides (3,528 Circle, 1 Lariat), with Pearson correlation to PAMPA of 0.53 for “Light Physics” and 0.72 for “Heavy Physics”. The 10–15-monomer benchmark comprises 1,629 peptides (111 Circle, 1,518 Lariat), with corresponding values of 0.44 and 0.59 (Figure 19). “Heavy Physics” shows stronger linear association with the measurements in both groups. The groups differ strongly in topology as well as size, and these Pearson correlations should not be compared numerically with the AI model's Spearman correlations. The displayed physics outputs are raw, uncalibrated membrane-transfer scores; PAMPA-scale calibration is a separate validation-fitted mapping.

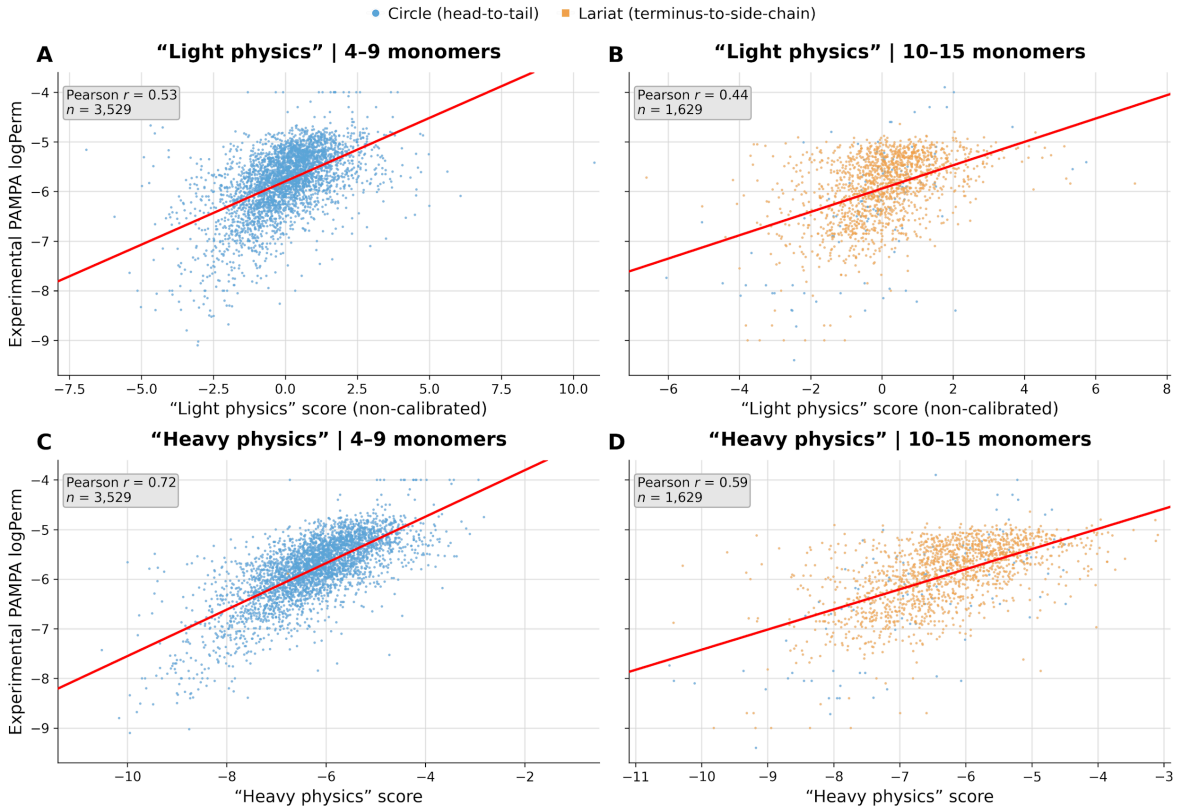

Figure 19. Performance of raw "Light Physics" and "Heavy Physics" membrane-transfer scores versus experimental PAMPA logPerm after source-level PAMPA quality filtering. (A,C) 4–9-monomer subset (n = 3,529; Circle n = 3,528, Lariat n = 1). (B,D) 10–15-monomer subset (n = 1,629; Circle n = 111, Lariat n = 1,518). Circle and Lariat peptides use the same topology encoding as in the other permeability figures. Solid black lines are fitted linear trends.

4.6.3 Consistency between AI and physics-based tiers

On the same test fold, the frozen AI predictions showed a strong positive rank correlation with both physics tiers: Spearman ρ = 0.69 for "Light Physics" (95% CI) and ρ = 0.79 for "Heavy Physics" (95% CI) (Figure 20).

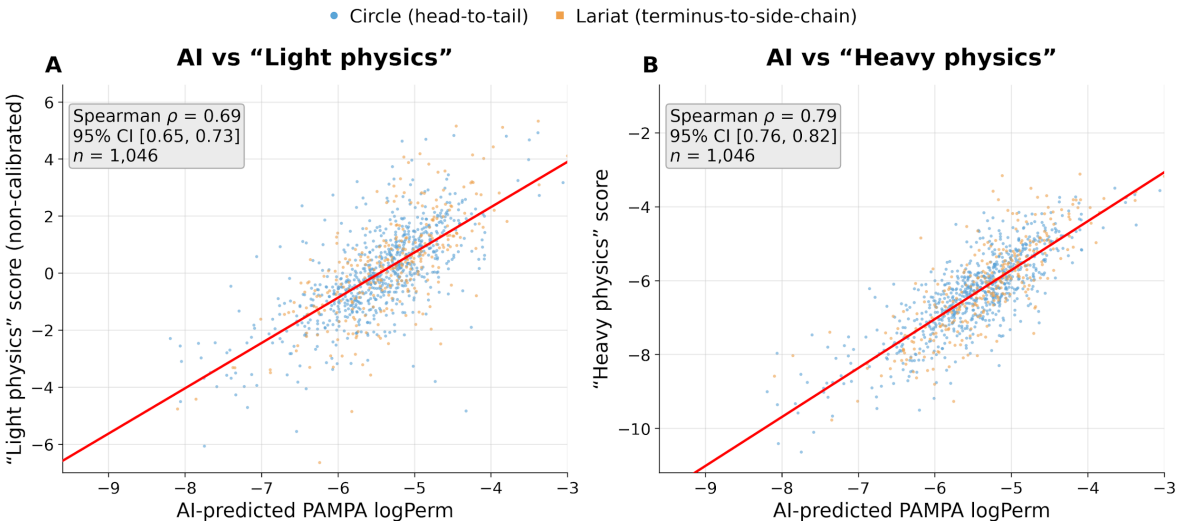

Figure 20. Internal cross-tier consistency between the conformation-sensitive AI predictor and physics-based permeability scores. (A) AI-predicted PAMPA logPerm versus raw "Light Physics" membrane-transfer score. (B) AI-predicted PAMPA logPerm versus raw "Heavy Physics" membrane-transfer score. The held-out test fold contains n = 1,046 peptides (Circle n = 742, Lariat n = 304). Insets report Spearman ρ, 95% bootstrap confidence intervals, and n. Raw physics scores are not calibrated to the PAMPA scale; therefore no identity line is shown.

Because the AI and physics tiers use the same production conformational representation, this analysis is interpreted as an internal cross-tier consistency check rather than an independent validation of predictive accuracy. It shows that the tiers recover a similar monotonic ordering on this held-out cohort, but it does not by itself establish out-of-distribution accuracy.

4.6.4 Confidence-based selection of the evaluation method

The remaining question is whether AI confidence can identify candidates that benefit from physical evaluation. Within this fold, the 210 lowest-confidence peptides, 20% of the cohort, are routed to the physics-based workflow, while the remaining 836 form the higher-confidence group (Table 4). AI predictions have MAE of 0.39 log units for the higher-confidence candidates and 1.06 log units for the lower-confidence candidates. For the same 210 lower-confidence candidates, the calibrated physics-derived estimate has MAE of 0.55 log units.

For this confidence-routing MAE comparison only, raw physics scores are mapped to the PAMPA logPerm scale using the two-parameter affine calibration fitted on validation data and frozen before evaluation; no held-out labels are used to refit the mapping. The confidence grouping is an operational routing decision rather than a calibrated probabilistic partition.

Table 4. Confidence-based routing within the held-out test fold (n = 1,046).

Evaluation set	Estimator	MAE (log units)	n
Higher-confidence candidates	AI	0.39	836
Lower-confidence candidates	AI	1.06	210
Same lower-confidence candidates	Calibrated physics-derived logPerm (“Light Physics” and “Heavy Physics”)	0.55	210

4.7 Experimental peptide programs

The component benchmarks do not show whether a complete design round produces useful peptides. The following programs address that question through synthesis and testing, starting from an existing binder, a target structure alone, or a peptide that requires joint affinity and permeability optimization.

4.7.1 Optimization of an interleukin binder

The first program applied the workflow to affinity optimization of an existing peptide rather than *de novo* discovery. The starting point was a single 26-residue β-strand peptide constrained by an organic linker, with micromolar affinity and no established structure–activity relationship. Because an experimental protein–peptide complex was not available, optimization was performed from a computationally modeled structural hypothesis. Position-specific substitution alphabets, including noncanonical amino acids, were defined around this structure and explored in a single computational design pass.

Twenty-four selected candidates were synthesized. Nine had measured affinities below 100 nM, and the best reached 97 nM, approximately tenfold better than the starting peptide. In the structural models, the optimized designs retained the principal hydrogen-bond and ionic interaction pattern of the starting structural hypothesis while introducing additional hydrophobic contacts and an intrapeptide π-cation interaction predicted to stabilize the bound conformation (Figure 21). These interaction changes provide a structural rationale for the selected substitutions but are computational predictions rather than experimentally determined contacts. The program therefore evaluates the combined use of structural modeling, substitution search, physical ranking, and candidate selection, rather than the performance of any individual component.

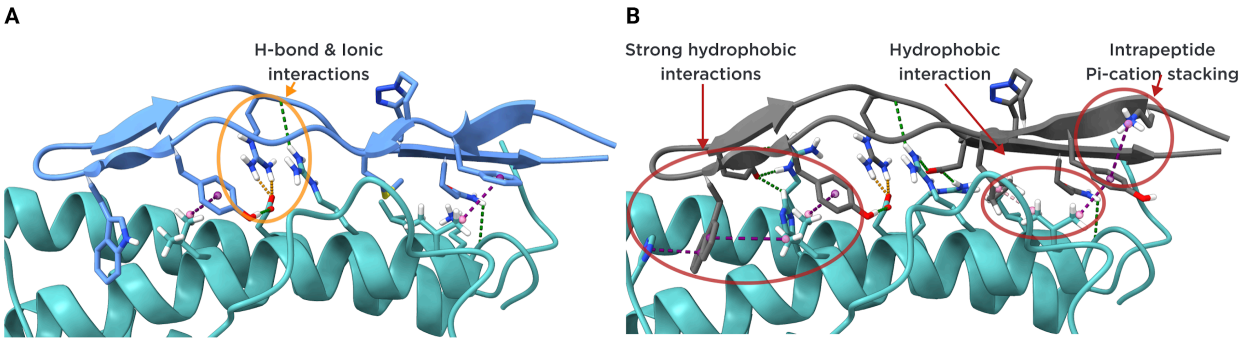

Figure 21. Modeled interleukin-binding complexes before and after affinity optimization. (A) Starting modeled complex, showing the principal hydrogen-bond and ionic interactions in the initial structural hypothesis. (B) Optimized modeled complex, retaining this interaction pattern while introducing additional hydrophobic contacts and an intrapeptide π-cation interaction predicted to stabilize the bound conformation.

4.7.2 *De novo* GPCR peptide design

The second program tested *de novo* peptide design from a target structure without a known peptide binder. The target was a GPCR involved in protease-induced itch and inflammation, and the design brief required linear 4–10-residue antagonists compatible with an INCI-compliant amino-acid alphabet, nanomolar activity, topical delivery, and low *in vitro* toxicity. ~10,000 peptides were generated. Structural and multiparameter evaluation reduced this set to ~2,000 candidates for final ranking by predicted activity, binding-mode stability, and skin permeability, from which 100 peptides were selected for synthesis and wet-lab validation.

50 of the 100 synthesized peptides showed antagonist activity *in vitro* (Figure 22). Five candidates had IC_{50} < 0.3 μM, nine were in the 0.3–1 μM range, 21 in the 1–10 μM range, and 15 in the 10–30 μM range; the best measured potency was IC_{50} = 90 nM. Eight active peptides also showed measurable skin permeation in a Franz diffusion cell assay using a Strat-M membrane, with the best candidate reaching Q_{24h} = 1.8 μg cm^{-2}. In a HaCaT keratinocyte MTT assay, the top tested peptides maintained >90% viability at 100 μM. The program therefore demonstrates *de novo* generation followed by experimental selection against activity and developability criteria.

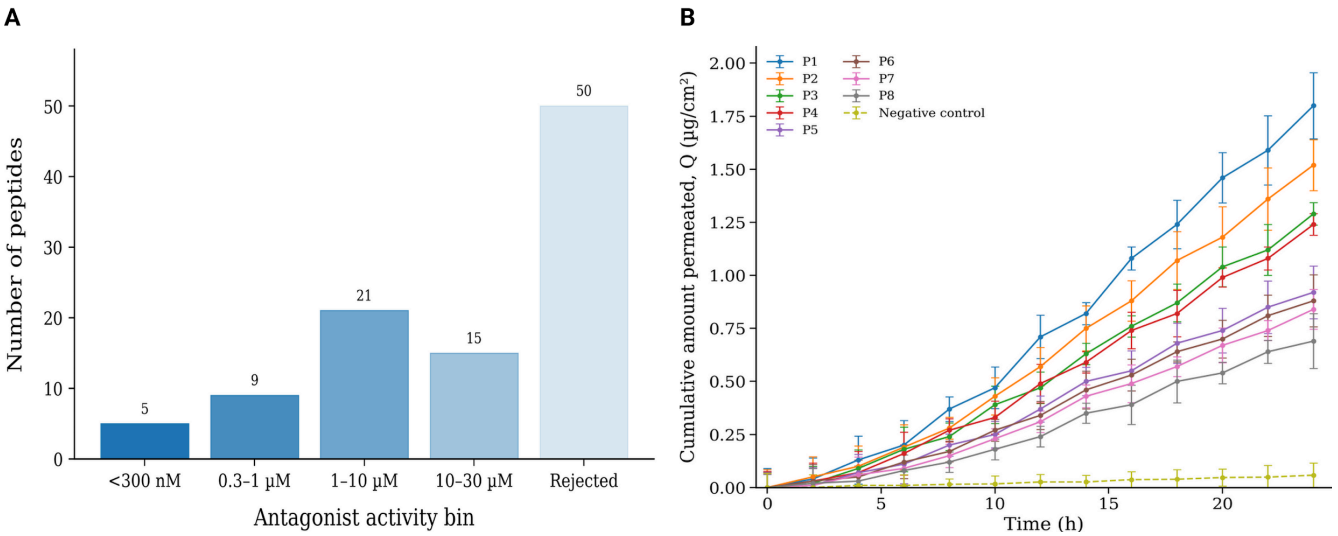

Figure 22. Experimental outcome of the *de novo* GPCR peptide program. (A) Distribution of measured antagonist activity across the 100 synthesized candidates; 50 were active *in vitro*, with a best-measured IC_{50} of 90 nM. (B) Skin permeation of eight active peptides in a Franz diffusion cell assay using a Strat-M membrane; the highest 24-h cumulative permeation was Q_{24h} = 1.8 μg cm^{-2}.

4.7.3 Vps29 affinity and permeability optimization

The third program combined affinity and permeability optimization over two experimental rounds. The starting peptide was a 16-residue disulfide cycle targeting Vps29, with K_d = 780 nM and PAMPA logPerm = −8.0 at the assay detection floor. Fourteen positions could vary, including substitutions with noncanonical amino acids. Candidates were ranked jointly on affinity and permeability. Because the peptide was longer than the AI training range, physics-based scores supplied additional evidence.

In the first round, 20 peptides were synthesized. Nine improved both affinity and permeability, and five reached logPerm ≥ −6.5. The best affinity was 250 nM, while the highest permeability was −5.92.

The second round introduced selected backbone N-alkylations to reduce exposed hydrogen-bond donor character and to favor permeability-compatible conformational state. The physics-based scores and first-round PAMPA data were used for AI model fine-tuning and adaptation to the specific series. Six candidates were synthesized. The best second-round peptide reached logPerm = −5.63 with an affinity of 270 nM (Figure 23).

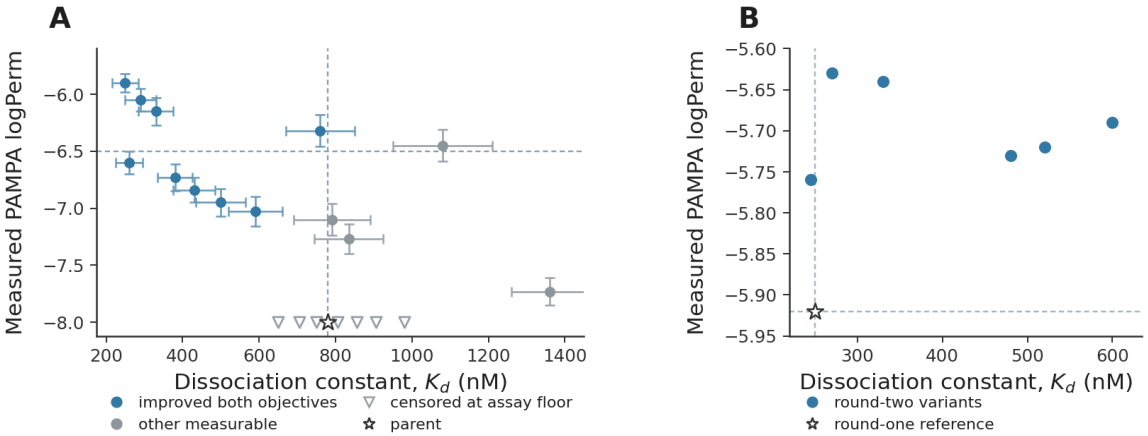

Figure 23. Optimization results in the affinity–permeability plane. (A) The twenty first-round designs with measurement error bars, the parent affinity reference, the permeability threshold, and censored values at the assay floor. (B) The best six second-round N-alkylation variants relative to the best first-round peptide.

Table 5 summarizes all shared peptide programs:

Table 5. Summary of the experimental peptide programs.

Program	Starting point	Experimental selection	Best result
Interleukin binder	Constrained binder; μM affinity	24 synthesized; 9 below 100 nM	97 nM; tenfold affinity improvement
De novo GPCR	Target structure only	10,000 generated; 2,000 selected; 100 synthesized; 50 active *in vitro*	IC_{50} 90 nM; 8 showed measurable skin permeation
Vps29	16-residue disulfide cycle; K_d 780 nM; logPerm −8.0	20 first-round designs and 6 N-alkylated variants synthesized	logPerm −5.63 at 270 nM affinity

5 Discussion

The results address successive decisions in peptide design. Structural hypotheses determine the interactions and constraints considered during optimization. Design spaces specify which chemical changes can be explored, and QuorumMap uses property evaluations to select candidates for testing. Experimental measurements then inform the next design round.

5.1 Structural models guide optimization

The structural results matter for design because a residue substitution can have different effects in different binding modes. Each retained structural hypothesis therefore provides the basis for a design space, including the positions that can vary and the interactions or conformational constraints to preserve. When several binding modes remain plausible, optimization retains these alternatives. ForceFold combines alternative structures from physical search with co-folding to improve this starting point.

The controlled benchmark shows that the conditioning adaptors improve prediction relative to the same co-folding model supplied with physical contacts through its native restraint input. The search diagnostics also show that generating an accurate mode is not sufficient: representation and final selection remain important sources of error.

ArtiDock-P contributes to this process alongside other docking engines and sampling methods. Its separate docking benchmark measures a component of the workflow, while the ForceFold benchmark measures their combined structural-prediction performance.

5.2 Chemical modifications as design variables

Within each design space, explicit atomic structures represent modifications that cannot be described by a canonical amino-acid sequence alone. Parameterized building blocks make these modifications available to conformational sampling, docking, scoring, and MD. They can therefore be evaluated during optimization alongside standard residue substitutions.

This representation expands the chemistry that can be searched. It does not establish equal predictive accuracy for every building block or topology. The relevant structural and property models must still be evaluated for unfamiliar chemistry.

5.3 Experimental feedback

Structural and property predictions guide candidate selection in QuorumMap. Experimental measurements establish whether proposed modifications improve the intended properties and are returned to QuorumMap to inform candidate selection and model adaptation in the next round. They can also help distinguish structural hypotheses, although property measurements alone may not identify a unique binding mode.

The Vps29 program illustrates this iterative use of data. First-round affinity and PAMPA measurements guided a second round of backbone modifications. First-round PAMPA data and physics-based scores were also used to adapt the AI model to the series. The outcome supports joint affinity and permeability optimization in that series. The three experimental programs do not include matched alternative workflows, so they do not quantify reductions in synthesis, design cycles, or development time.

6 Limitations

The reported benchmarks assess specific tasks on selected datasets. The main limitations are the restricted coverage of the test datasets, uncertainty in predicted structures and property scores, and limited experimental validation for unfamiliar peptide chemistry.

6.1 Scope of the benchmarks

Each benchmark measures a different aspect of platform performance. ForceFold was evaluated for protein–peptide complex prediction on 172 systems. Auto-Alphabet was evaluated for agreement with the full computational scoring function on six targets and in a separate IL-11 test. Affinity scoring was evaluated for its ability to rank peptides within ten chemical series comprising 191 peptides across eight targets. These evaluations use different datasets, reference values, and performance measures, so their results cannot be combined into a single estimate of platform accuracy.

The optimization benchmarks assess search and candidate selection using controlled sequence–property models with known optima. The FcRn benchmark incorporates experimental measurements, but values for most sequences are generated computationally. These tests establish performance under the specified model conditions. They do not establish equivalent performance with actual structural calculations or experimental assays.

The experimental peptide programs evaluate the combined workflow, from computational design to synthesis and testing. Without matched comparisons in which individual components are changed or removed, the results cannot determine how much each component contributes. Prospective experimental evaluation of permeability optimization is currently limited to the Vps29 peptide series.

6.2 Structural prediction

Structural predictions depend on how the target is prepared and which conformations are considered. Incorrect assignment of the biological assembly, protonation states, metal coordination, covalent bonds, or functional state can affect all subsequent calculations. Molecular dynamics simulations may not sample every relevant receptor conformation from the chosen starting structure.

ArtiDock-P and the complete physical search procedure were evaluated on different datasets. Their accuracy values therefore describe different tests and cannot be compared directly. Similarly, finding an acceptable binding mode among the generated or retained structures does not establish that the method can select it as the final prediction.

Analyses of receptor structural differences and similarity to training data identify conditions associated with lower prediction accuracy. They do not establish the causes of individual errors, and the groups with the lowest accuracy contain few systems. A predicted complex that remains stable during a short molecular dynamics simulation is not necessarily the experimentally observed binding mode.

The benchmarks primarily assess complex prediction from candidate structural hypotheses. They do not separately measure how accurately the platform identifies a peptide binding site when only the target structure is provided.

6.3 QuorumMap optimization and affinity ranking

The use of predicted structures for optimization introduces further uncertainty in candidate selection. Auto-Alphabet is evaluated against a computational reference whose ranking varies between repeated runs. Its recall therefore measures agreement with that reference, not experimentally established substitution preferences.

The small-molecule search-efficiency benchmark supports the optimization strategy but does not directly establish equivalent savings for peptide design. The peptide optimization benchmarks use simulated evaluators specified by rank correlation with a reference objective. No docking, Receptor.AI affinity scoring, or free-energy-perturbation calculations were performed in these tests, so their results do not establish performance with actual scoring methods.

The affinity benchmark remains limited to 191 peptides in ten related series across eight targets. Its ranking performance may not transfer to other targets or chemistries. Receptor.AI affinity scores are not absolute K_d values or rigorous alchemical free energies.

6.4 Passive permeability

The permeability models were evaluated using public PAMPA data collected under different assay conditions. These datasets mainly contain cyclic peptides within particular ranges of length and cyclization topology. Performance on linear peptides and peptides with unfamiliar or multiple ring structures requires separate validation. PAMPA measures passive permeation across an artificial membrane, so agreement with PAMPA measurements does not establish the ability to predict cellular uptake or oral bioavailability.

The AI confidence score helps identify candidates for which the model may be less reliable. It is not a calibrated probability that a prediction is correct, and it cannot identify every peptide outside the model's range of applicability.

The physical methods also require validation for unfamiliar building blocks, linkers, charge states, and mechanisms of membrane transport.

The confidence-group errors quantify routing behavior within the benchmark cohort and do not constitute validation on an additional independent dataset.

Prospective experimental evidence is limited to the Vps29 series. Some permeability measurements were at the assay detection limit, so their exact values could not be determined. The results support improved permeability in that series but do not establish predictive accuracy across other peptide families.

7 Conclusions

The Receptor.AI Peptide Platform uses project objectives, target information, and starting peptides to form structural hypotheses and define the design spaces to be searched. Within this architecture, QuorumMap serves as the peptide optimization and candidate-selection engine, using affinity and other property evaluations to search these design spaces, prioritize further assessment, and select candidates for experimental testing. Canonical and noncanonical chemistry use an explicit atomic representation throughout the workflow. Experimental measurements guide candidate selection and model adaptation in subsequent design rounds.

On the reported benchmarks, ForceFold improved complex prediction relative to native-restraint conditioning of the same co-folding model, and Receptor.AI affinity scoring provided relative affinity-ranking information within peptide series. The permeability AI model supported ranking of unseen cyclic-peptide scaffolds. The experimental programs produced active *de novo* peptides and improved existing binders, including joint gains in affinity and permeability in the Vps29 series.

Together, these results support the use of structural models and experimental feedback for peptide design. Broader prospective evaluation is needed to establish performance across additional targets, chemistries, and peptide topologies.

8 Data availability

The four benchmark datasets that support the findings of this paper are available in the Receptor.AI Peptide Benchmarks GitHub repository (https://github.com/receptor-ai/receptor-ai-peptide-benchmarks) under a Creative Commons Attribution 4.0 International licence:

- docking985 [13] — 985 curated peptide–protein X-ray and NMR complexes. Each entry holds the native receptor and peptide coordinates and a bond-order-aware peptide SDF, with the pose-scoring library used for the RMSD, DockQ and CAPRI metrics of Section 4.1. The dataset holds no predicted poses.

- BMI-200 and BMI-MODES [45] — 202 peptide–protein complexes with per-entry ring-topology, secondary-structure and crystallographic-quality metadata, and 30 mode-discrimination sites that hold 71 experimentally supported poses. The receptor-class annotations that define the 172-complex subset of Section 4.2 are in bmi200/MANIFEST.csv.

- Peptide scoring-function affinity benchmark [34] — the reference receptor and peptide structures, the analog alignments and the measured potencies for the 10 series and 191 peptides of Table 3, with the full scoring procedure. The sampled conformers (approximately 11 GB) and the per-pose scores are not included.

- Peptide passive-permeability benchmark [39] — the curated CycPeptMPDB PAMPA source data used in Section 4.6, with the source publication of every measurement. The source-level quality-control results, the filtered benchmark cohort, and the per-peptide predicted values are not included.

Coordinates and crystallographic metadata in docking985, BMI-200 and BMI-MODES come from the RCSB Protein Data Bank [48] under CC0 1.0. Receptor functional annotation in BMI-200 and BMI-MODES comes from UniProt [49] under CC BY 4.0. The permeability measurements come from CycPeptMPDB [40]. Attribute these sources when you use the datasets.

The following data are not publicly available: the proprietary part of the production permeability training corpus (approximately 3,000 records, Section 3.6.1); the per-peptide prediction caches behind the reported correlations; and the design, synthesis and assay data of the GPCR, interleukin-binding and Vps29 programs (Section 4.7). The program data are subject to partner confidentiality agreements.

9 Code availability

Code supporting public benchmark curation and structural metric calculation is released with the corresponding datasets under the MIT licence: the pose-scoring library and the entry-selection filter of docking985 [13], and the secondary-structure assignment tool of BMI-200 and BMI-MODES [45].

The platform components evaluated in this paper — QuorumMap, ForceFold, ArtiDock-P, the Receptor.AI affinity scoring, the permeability AI model, and the “Light Physics” and “Heavy Physics” permeability tiers — are proprietary software of Receptor.AI, Inc. They are not accessible to the public or the research community. They are available under a commercial licence or a collaboration agreement. Prime and Glide are commercial products of Schrödinger, Inc. and require a separate licence.

References

[1] Muttenthaler M, King GF, Adams DJ, Alewood PF. Trends in peptide drug discovery. Nat Rev Drug Discov. 2021;20:309-325.

[2] Scott DE, Bayly AR, Abell C, Skidmore J. Small molecules, big targets: drugging protein-protein interactions. Nat Rev Drug Discov. 2016;15:533-550.

[3] Vinogradov AA, Yin Y, Suga H. Macrocyclic peptides as drug candidates: recent progress and remaining challenges. J Am Chem Soc. 2019;141:4167-4181.

[4] Sawyer TK, Biswas K. Peptide Drug Discovery Raison d’Etre: Engineering Mindset, Design Rules and Screening Tools. In: Approaching the Next Inflection in Peptide Therapeutics: Attaining Cell Permeability and Oral Bioavailability. ACS Symposium Series. 2022;1417:1–25. doi:10.1021/bk-2022-1417.ch001.

[5] Huang Y, Wiedmann MM, Suga H. RNA display methods for the discovery of bioactive macrocycles. Chem Rev. 2019;119:10360-10391.

[6] Smith GP. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. Science. 1985;228(4705):1315–1317. doi:10.1126/science.4001944.

[7] Petrov D, Plais L, Schira K, et al. Flexibility-tuning of dual-display DNA-encoded chemical libraries facilitates cyclic peptide ligand discovery. Nat Commun. 2025;16:3273. doi:10.1038/s41467-025-58507-w.

[8] Buttenschoen M, Morris GM, Deane CM. PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. Chem Sci. 2024;15:3130-3139.

[9] Ling K, Li S, Zhang Z, Shin WH, Kihara D. Peptide-protein docking: from physics-based models to generative intelligence. Chem Commun. 2026;62:7235-7247.

[10] Zhao H, Zhang O, Jiang D, et al. Protein-peptide docking with a rational and accurate diffusion generative model. Nat Mach Intell. 2025;7:1308-1321.

[11] Voitsitskyi T, Koleiev I, Stratiichuk R, et al. ArtiDock: accurate machine learning approach to protein-ligand docking optimized for high-throughput virtual screening. J Chem Inf Model. 2026;66:2117-2128.

[12] Durairaj J, Adeshina Y, Cao Z, et al. PLINDER: the protein-ligand interactions dataset and evaluation resource. bioRxiv. 2024.

[13] Receptor.AI, Inc. docking985: a peptide-protein re-docking benchmark. Receptor.AI Peptide Benchmarks repository (2026). https://github.com/receptor-ai/receptor-ai-peptide-benchmarks/tree/main/docking_benchmark

[14] Abramson J, Adler J, Dunger J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630:493-500.

[15] Passaro S, Corso G, Wohlwend J, et al. Boltz-2: towards accurate and efficient binding affinity prediction. bioRxiv. 2025.06.14.659707. doi:10.1101/2025.06.14.659707.

[16] Skrinjar P, Eberhardt J, Studer G, et al. Evaluating generalization in protein-ligand cofolding methods. Nat Struct Mol Biol. 2026;33:782-794.

[17] Amaro RE, Baudry J, Chodera J, et al. Ensemble Docking in Drug Discovery. Biophys J. 2018;114:2271-2278.

[18] Bergstra J, Bardenet R, Bengio Y, Kegl B. Algorithms for hyper-parameter optimization. Adv Neural Inf Process Syst. 2011;24:2546-2554.

[19] Romero PA, Krause A, Arnold FH. Navigating the protein fitness landscape with Gaussian processes. Proc Natl Acad Sci USA. 2013;110(3):E193-E201.

[20] Zitzler E, Thiele L. Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. IEEE Trans Evol Comput. 1999;3(4):257-271.

[21] Zitzler E, Thiele L, Laumanns M, Fonseca CM, da Fonseca VG. Performance assessment of multiobjective optimizers: an analysis and review. IEEE Trans Evol Comput. 2003;7(2):117-132.

[22] Daulton S, Balandat M, Bakshy E. Parallel Bayesian optimization of multiple noisy objectives with expected hypervolume improvement. Adv Neural Inf Process Syst. 2021;34:2187-2200.

[23] Daulton S, Balandat M, Bakshy E. Differentiable expected hypervolume improvement for parallel multi-objective Bayesian optimization. Adv Neural Inf Process Syst. 2020;33:9851-9864.

[24] Yang K, Emmerich M, Deutz A, Bäck T. Efficient computation of expected hypervolume improvement using box decomposition algorithms. J Glob Optim. 2019;75:3-34.

[25] Paria B, Kandasamy K, Póczos B. A flexible framework for multi-objective Bayesian optimization using random scalarizations. Proc 35th Conf Uncertainty Artif Intell. PMLR. 2020;115:766-776.

[26] Knowles J. ParEGO: a hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. IEEE Trans Evol Comput. 2006;10(1):50-66.

[27] Kennedy MC, O'Hagan A. Predicting the output from a complex computer code when fast approximations are available. Biometrika. 2000;87(1):1-13.

[28] Bonilla EV, Chai KMA, Williams CKI. Multi-task Gaussian process prediction. Adv Neural Inf Process Syst. 2008;20:153-160.

[29] Kandasamy K, Dasarathy G, Schneider J, Poczos B. Multi-fidelity Bayesian optimisation with continuous approximations. Proc 34th Int Conf Mach Learn. 2017;70:1799-1808.

[30] Daulton S, Balandat M, Bakshy E. Hypervolume knowledge gradient: a lookahead approach for multi-objective Bayesian optimization with partial information. Proc 40th Int Conf Mach Learn. PMLR. 2023;202.

[31] Wu NC, Dai L, Olson CA, Lloyd-Smith JO, Sun R. Adaptation in protein fitness landscapes is facilitated by indirect paths. eLife. 2016;5:e16965.

[32] Stanton S, Alberstein R, Frey N, Watkins A, Cho K. Closed-form test functions for biophysical sequence optimization algorithms. arXiv:2407.00236. 2024.

[33] Hopf TA, Ingraham JB, Poelwijk FJ, et al. Mutation effects predicted from sequence co-variation. Nat Biotechnol. 2017;35(2):128-135.

[34] Receptor.AI, Inc. Peptide scoring-function affinity benchmark. Receptor.AI Peptide Benchmarks repository (2026). https://github.com/receptor-ai/receptor-ai-peptide-benchmarks/tree/main/affinity_benchmark

[35] Dougherty PG, Sahni A, Pei D. Understanding Cell Penetration of Cyclic Peptides. Chem Rev. 2019;119(17):10241-10287. doi:10.1021/acs.chemrev.9b00008.

[36] Rezai T, Bock JE, Zhou MV, Kalyanaraman C, Lokey RS, Jacobson MP. Conformational Flexibility, Internal Hydrogen Bonding, and Passive Membrane Permeability: Successful In Silico Prediction of the Relative Permeabilities of Cyclic Peptides. J Am Chem Soc. 2006;128(43):14073-14080. doi:10.1021/ja063076p.

[37] Linker SM, Schellhaas C, Kamenik AS, et al. Lessons for Oral Bioavailability: How Conformationally Flexible Cyclic Peptides Enter and Cross Lipid Membranes. J Med Chem. 2023;66(4):2773-2788. doi:10.1021/acs.jmedchem.2c01837.

[38] Kansy M, Senner F, Gubernator K. Physicochemical High Throughput Screening: Parallel Artificial Membrane Permeation Assay in the Description of Passive Absorption Processes. J Med Chem. 1998;41(7):1007-1010. doi:10.1021/jm970530e.

[39] Receptor.AI, Inc. Peptide passive-permeability benchmark on CycPeptMPDB. Receptor.AI Peptide Benchmarks repository (2026). https://github.com/receptor-ai/receptor-ai-peptide-benchmarks/tree/main/permeability_benchmark

[40] Li J, Yanagisawa K, Sugita M, Fujie T, Ohue M, Akiyama Y. CycPeptMPDB: A Comprehensive Database of Membrane Permeability of Cyclic Peptides. J Chem Inf Model. 2023;63(7):2240-2250. doi:10.1021/acs.jcim.2c01573.

[41] Bemis GW, Murcko MA. The Properties of Known Drugs. 1. Molecular Frameworks. J Med Chem. 1996;39(15):2887-2893. doi:10.1021/jm9602928.

[42] Liu W, Li J, Verma CS, Lee HK. Systematic Benchmarking of 13 AI Methods for Predicting Cyclic Peptide Membrane Permeability. J Cheminform. 2025;17:129. doi:10.1186/s13321-025-01083-4.

[43] Hirschfeld L, Swanson K, Yang K, Barzilay R, Coley CW. Uncertainty Quantification Using Neural Networks for Molecular Property Prediction. J Chem Inf Model. 2020;60(8):3770-3780. doi:10.1021/acs.jcim.0c00502.

[44] Tanizaki S, Feig M. A Generalized Born Formalism for Heterogeneous Dielectric Environments: Application to the Implicit Modeling of Biological Membranes. J Chem Phys. 2005;122(12):124706. doi:10.1063/1.1865992.

[45] Receptor.AI, Inc. BMI-200 and BMI-MODES: peptide-protein binding-mode identification benchmarks. Receptor.AI Peptide Benchmarks repository (2026). https://github.com/receptor-ai/receptor-ai-peptide-benchmarks/tree/main/bmi_benchmark

[46] Marcu O, Dodson EJ, Alam N, et al. FlexPepDock lessons from CAPRI peptide-protein rounds and suggested new criteria for assessment of model quality and utility. Proteins. 2017;85:445-462.

[47] Zhai S, Zhao H, Wang J, et al. PepPCBench is a comprehensive benchmarking framework for protein-peptide complex structure prediction. J Chem Inf Model. 2025;65:8497-8513.

[48] Burley SK, Bhatt R, Bhikadiya C, et al. Updated resources for exploring experimentally-determined PDB structures and Computed Structure Models at the RCSB Protein Data Bank. Nucleic Acids Res. 2025;53(D1):D564-D574. doi:10.1093/nar/gkae1091.

[49] UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. Nucleic Acids Res. 2025;53(D1):D609-D617. doi:10.1093/nar/gkae1010.