Mining the endogenous peptidome for peptide binders with deep learning-driven optimization and molecular simulations

Erik Hartman[1] Firdaus Samsudin[2] Peter J. Bond[2,3]
Artur Schmidtchen[†,4] Johan Malmström[†,1]

Abstract

The endogenous peptidome is the complete set of naturally occurring peptides, some of which are involved in processes like intercellular communication, and in innate immunity. The peptidome is also a potential source of uncharacterized peptide-based therapeutics, which has remained largely unexplored due to challenges in identifying bioactive peptides in vast peptidomic datasets. Here, we present a generalized computational pipeline that mines datasets for peptides with high affinity binding capacities. The approach employs protein-peptide docking, binding interface scoring and a deep ensemble model to dynamically learn the mapping between peptide embeddings and binding capacity. We demonstrate the utility of the pipeline by identifying endogenous peptide binders for the LPS-binding site of CD14 using experimentally defined wound fluid peptidomes. Promising candidates underwent conformational refinement and validation through all-atom molecular dynamics simulations. This pipeline offers a systematic way to uncover novel peptide binders within large peptidomes, paving the way for accelerated discovery of peptide-based therapeutics.

[†] These authors contributed equally to this work.

[1] Division of Infection Medicine, Faculty of Medicine, Lund University, Sweden
[2] Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), 30 Biopolis Street, #07-01 Matrix, Singapore 138671, Republic of Singapore
[3] Department of Biological Sciences, National University of Singapore, Singapore 117543, Singapore
[4] Division of Dermatology and Venereology, Department of Clinical Sciences, Lund University, Sweden

Introduction

The identification and design of peptide binders with high affinity for specific protein targets is a fundamental challenge in drug discovery, therapeutic development, and biomolecular engineering [1–3]. The large peptide sequence space, combined with the complexity of protein-peptide interactions, makes the discovery of optimal binders a challenging task [4]. Traditional experimental approaches, such as high-throughput screening and combinatorial libraries, are resource-intensive and often inefficient, particularly when navigating the vast landscape of potential peptide candidates. As a result, computational methods have emerged as promising tools to accelerate and streamline the peptide discovery process.

Machine learning techniques, particularly protein language models such as Evolutionary Scale Modeling (ESM) [5], and structural prediction frameworks such as AlphaFold [6] and RoseTTAFold [7] leverage large-scale data to predict protein and peptide sequences as well as their structures and molecular interactions. During training, these models learn a latent representation of amino acid sequences that capture their biological and chemical properties. When combined with search algorithms and generative techniques, like diffusion methods or Monte Carlo-based searches, these tools provide sophisticated approaches for exploring the vast landscape of sequences. As a result of these developments, methods

like PepMLM [8], EvoBind [9], BindCraft [10] and others [11,12], have emerged as potent tools for designing peptide binders.

Parallel to these developments, there has been growing interest in identifying encrypted peptides with desirable properties from naturally occurring sequences in organisms, including endogenous peptidomes [13], the genomes of organisms, and extinct species [14]. Part of the underlying rationale is that evolution has effectively "sampled" the peptidomic landscape to yield bioactive peptides that have functional roles. By interrogating these peptides, nature's pre-existing optimization can be harnessed to discover novel potential therapeutics. However, current approaches often focus on identifying peptides with broad, low-dimensional properties, such as antimicrobial activity. In contrast, therapeutic peptides are characterized by highly specificity and potent binding capabilities. While several tools exist to evaluate and predict protein-peptide interactions [15–18], typical peptidomic datasets contain tens of thousands of peptides, currently making it computationally unfeasible to assess binding affinity for all peptides and targets. Currently, no method effectively combines structural docking techniques with large-scale endogenous peptidomic analysis.

Previous studies have demonstrated that encrypted peptides derived from the C-terminal part of thrombin (TCPs) are present in wound fluids and interact with the LPS-binding site of CD14, desensitizing cells to LPS-induced responses [19]. Such encrypted peptide binders, which modulate immune responses by targeting CD14, hold biological and clinical value [20,21] and demonstrate that endogenous peptidomes contain peptides with immunomodulatory properties.

Motivated to identify additional immunomodulatory peptides, we developed a general methodology that systematically searches for peptide binders in large peptidomic datasets using Bayesian optimization. The approach subsamples peptidomic datasets of tens of thousands of peptides with the help of a deep ensemble model that dynamically learns the mapping between peptide embeddings and binding capacity. The method utilizes the ESM2 protein language model to generate peptide embeddings [5], AlphaFold2 multimer v3 for rapid docking [22,23] and tools in the Rosetta suite for interface scoring [24].

We apply the approach to discover novel peptide binders to the LPS-binding site of CD14 within deep mass spectrometric wound fluid peptidomic datasets. This was followed by conformational refinement using FlexPepDock [25] for promising candidates. Finally, top-scoring peptides were subjected to all-atom molecular dynamics simulations to accurately sample protein-peptide dynamics and calculate binding energies.

Results

The primary objective of this work was to develop an efficient method for identifying peptide binders in large-scale peptidomic datasets. While current state-of-the-art tools provide potent ways to predict aspects of protein-peptide interactions, the high computational demands make them impractical for exhaustive screening of large-scale data. To this end, we developed a generalized multistep computational pipeline that leverages Bayesian Optimization (BO) to efficiently subsample the peptide space, minimizing the amount of necessary compute. The optimization process is guided by an ensemble of multilayer perceptrons (MLPs) that are trained to predict a multi-objective binding score that captures various aspects of binding quality. After the optimization process is completed, the top candidates are subjected to conformational searches using FlexPepDock and then validation with all-atom molecular dynamics (MD) simulations.

Our second objective was to investigate whether the infected and inflammatory peptidome contains endogenous peptide binders for immunomodulatory targets such as CD14. To do this, we assembled a concatenated peptidomic dataset by integrating previously acquired tandem mass spectrometry data from three clinical conditions: non-healing leg ulcers (N=18) [26], infected wounds (N=3), and acute

surgical wound fluid (N=5) [27]. After rudimentary filtering, this yielded 35,863 unique peptide sequences, which we then subjected to our computational pipeline, aiming to uncover potential binders that might play a role in modulating inflammation.

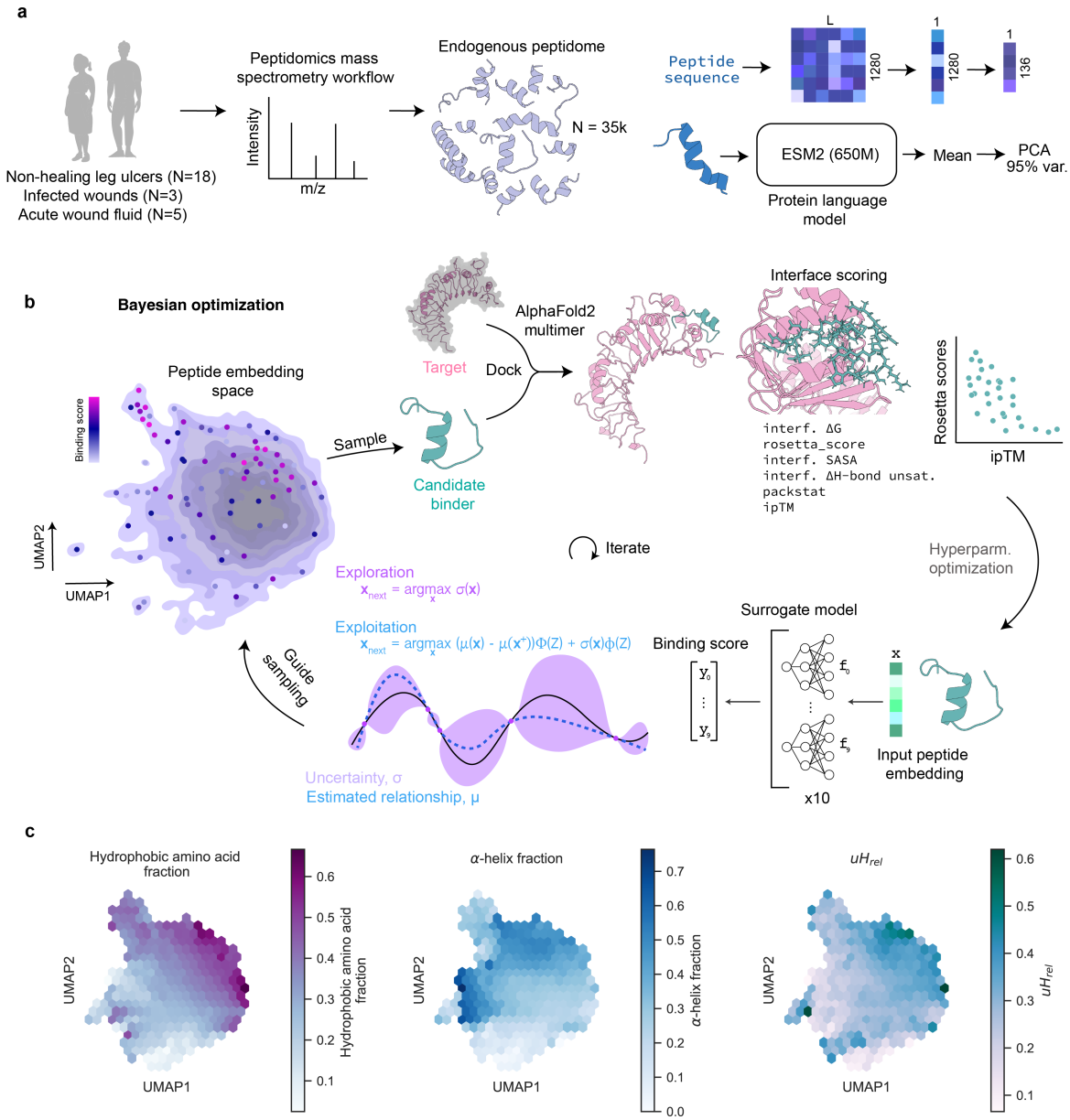


Fig. 1 Guided peptide sampling using Bayesian optimization. a Tandem mass spectrometry datasets from clinical wound fluid samples including non-healing leg ulcers (N=18), infected wounds (N=3), and acute wound fluid (N=5) were curated. The compiled final curated peptidomics dataset contained 35k endogenous peptides. The peptides were embedded into a latent space using the protein language model ESM2 and subjected to dimensionality reduction via PCA (explaining 95% of the variance). The PCA reduced the dimensionality from 1,280 to 136 peptide features. **b** The Bayesian optimization process starts with initial sampling of peptides from the latent space. Peptides are then iteratively sampled based on acquisition functions tailored for exploration or exploitation. The selected peptides are docked using AF2m against the target protein (CD14, in this case, PDB: 4GLP), and interfaces are scored using metrics from the Rosetta toolkit and from AlphaFold. Binding

scores are predicted by an ensemble model of 10 MLPs trained on reduced peptide embeddings. The surrogate model guides the next round of peptide selection. Hyperparameter optimization is performed to adapt the networks to the changing dataset. **c** Hexplot UMAP projections depicting the discrete peptide latent space, colored based on hydrophobic amino acid fraction, α-helix fraction and the relative hydrophobic moment. The color of a hexagon is computed by taking the mean of the respective metrics for all peptides inside the bin.

Deep learning-based search for peptide binders

To address the challenge of identifying peptide binders within large peptidomic datasets, we implemented a generalized BO pipeline that aims to minimize computational demands by guiding a sampling process towards promising peptide candidates. Before performing optimization, the peptide sequences need to be represented numerically. We therefore embed them into a latent vector-space using the protein language model ESM2 (650M parameter version). The embedding is followed by averaging and dimensionality reduction via principal component analysis (PCA), retaining 95% of the variance in the data to produce 136-dimensional vector representations (Fig. 1a and Methods). The latent space effectively captures several aspects of peptide properties that are relevant to binding capacity, including amino acid composition, secondary structure fractions, and more complex measures like the relative hydrophobic moment, uH_{rel} (Fig. 1a and Extended Data Fig. 1).

Peptide binding positions and poses are predicted by co-folding each candidate peptide with the target protein using AF2m. To capture multiple aspects of the binding quality and to minimize score-specific biases, we combine several Rosetta-based metrics into a single binding score (B_s). These metrics include ipTM (which indicates AF2m's certainty in relative chain positioning), change in interface Gibbs free energy (ΔG), the overall Rosetta score (R_{score}), solvent-accessible surface area change (ΔSASA), change in unsatisfied hydrogen bonds (ΔH_{unsat}), and the packing statistic (P_{stat}), which gauges interface packing quality. We compute B_s by taking a weighted linear sum of these individual score-terms. Peptides that are not predicted to bind within 5 Å of residues in a pre-determined binding site are penalized by assigning them a score of zero. All of these scores exhibit moderate to high correlations which reflect their commonality in assessing the binding quality (Extended Data Fig. 2).

The binding score serves as the objective for a surrogate model composed of an ensemble of MLPs, which takes the peptide embeddings as input. By approximating the relationship between peptide features and binding capacity, this model enables the identification of promising candidates without exhaustively testing every peptide. The hyperparameters of the ensemble are optimized with another BO process every 50th iteration, allowing the networks to adapt to the growing dataset as more peptides are being sampled. The performance of each hyperparameter configuration is evaluated through cross-validation, and the mean validation R^2 score was used as the objective function.

The BO operates in two distinct phases with two different acquisition functions defining the phase-specific goal. In the initial exploration phase, peptides are selected based on high uncertainty (σ) to maximize the diversity of sampled candidates. Here, uncertainty, $\sigma(\mathbf{x})$, is defined as the intra-ensemble standard deviation of the predicted binding score of each peptide in the dataset which has not been docked. This phase facilitates broad exploration of the peptide space, ensuring that underrepresented or highly variable regions are sufficiently sampled.

In the subsequent exploitation phase, the focus is shifted towards selecting peptides that maximize the expected improvement (EI) over the best candidate discovered so far. The expected improvement is calculated as:

$$\text{EI}(\mathbf{x}) = (\mu(\mathbf{x}) - \mu(\mathbf{x}^+))\Phi(Z) + \sigma(\mathbf{x})\phi(Z),$$

where $\mu(\mathbf{x})$ represents the average predicted binding score at $\mathbf{x}$, $\mu(\mathbf{x}^+)$ is the highest binding score

obtained so far, $Z = \frac{\mu(\mathbf{x})-\mu(\mathbf{x}^+)}{\sigma(\mathbf{x})}$, and $\Phi(Z)$ and $\phi(Z)$ are the cumulative distribution function and probability density function of the normal distribution, respectively. This strategy prioritizes peptides likely to have a high binding score to select the most promising candidates in the peptidome while also taking uncertainty into account.

The BO methodology is generalized, and an implementation is available at GitHub/ErikHartman/bopep.

Mining for endogenous peptide binders to CD14

After developing the pipeline, we applied it to identify potential peptide binders to the LPS binding site on CD14 within wound fluid peptidomic datasets, addressing the second objective. For this search, the BO process was initialized with 300 dockings, selecting peptides as centroids of clusters generated through minibatch k-means. Thereafter, 200 iterations of exploration was followed by 600 iterations of exploitation. With a batch size of 4 this resulted in 3500 dockings corresponding to <10% of the dataset. Hyperparameter optimization was performed every 50th iteration, which resulted in increasingly more complex models during the course of the optimization (Extended Data Fig. 3). Over time, the surrogate model gradually learned a mapping between the peptide embeddings and the binding score in the defined LPS-binding pocket. This can be seen in the increase of the validation R^2 which reached a peak at iteration 520 of $R^2 = 0.44 \pm 0.058$ (Fig. 2a-b).

As the phases transitioned from exploration to exploitation, the binding scores of the selected peptides exhibits a shift to more potent binders. This trend can be identified in the individual score components as well as the cumulative binding score (Fig. 2c-d). Additionally, the fraction of peptides predicted to bind in proximity to the site increases from 0.75 during initialization and exploration to 0.92 during exploitation. This shift in peptide properties is also apparent in their distribution within the embedding space (Fig. 2e), which, as previously noted, captures biophysical peptide characteristics. A similar trend emerges when analyzing the biophysical properties of peptides and correlating these with the individual binding score terms (Extended Data Fig. 4).

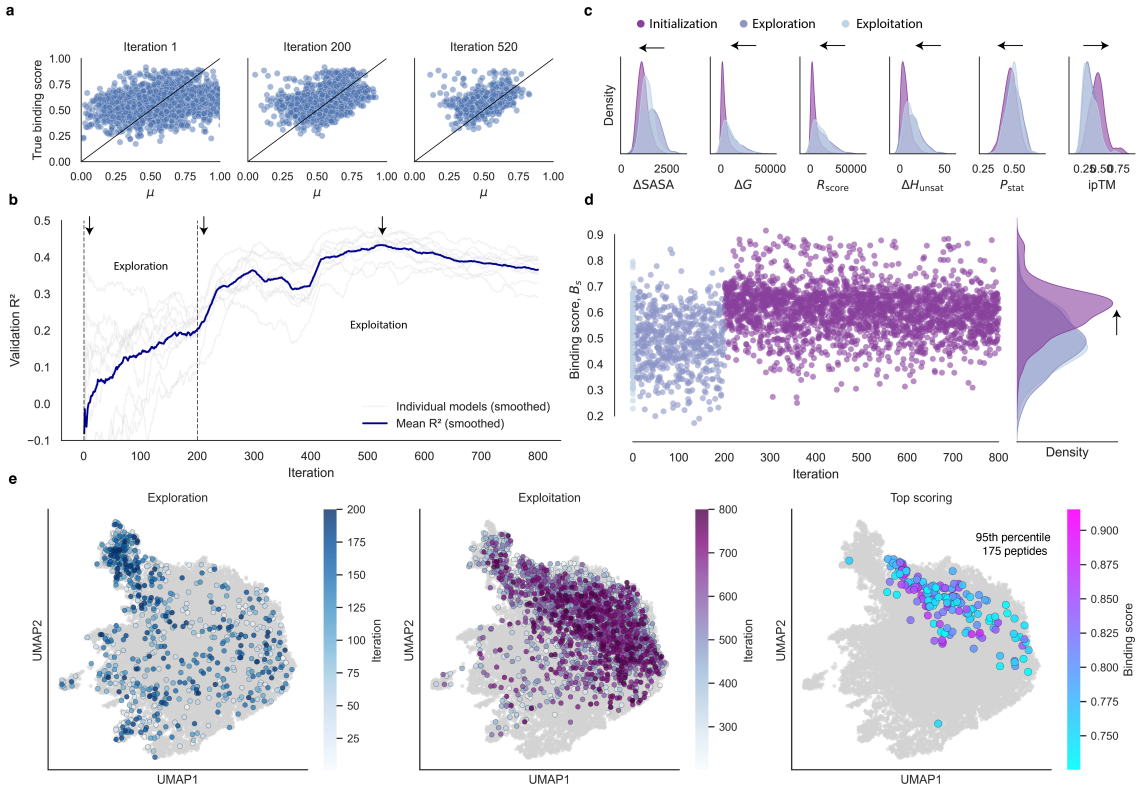


Fig. 2 The surrogate model learns a representation of specific binding activity. **a** Correlation between the true and estimated binding score by the MLP ensemble (μ) at iteration 1, 200 and 500 in the optimization process. The peptides included in each plot are the peptides that are docked after the selected iteration, and are therefore not a part of the training data. **b** Validation R^2 of the MLPs in the ensemble model during the optimization process. The gray lines are the values for each model and the dark blue line is the mean (μ). **c** Density plots of the individual score components that contribute to the binding score separated on optimization phase. The arrow depicts which direction is better. **d** The binding score the peptides sampled during the optimization process. **e** Projections of the latent peptide space where the colored scatters highlight which peptides were selected during exploration and exploitation, and which of the selected peptides exhibited the highest binding scores.

Conformational sampling of the top candidate binders selects the most promising binders

To narrow the evaluation to the most promising candidates, the top 100 peptides with the highest binding score were subjected to further examination (Fig. 3a). Several of these peptides were subsequences of other peptides included amongst the top binders and were likely generated through exoprotease activity. Inspired by a previous implementation of peptide networks for dimensionality reduction of peptidomic data [26], we generated peptide networks based on indel (insertion and deletion) distances, where peptides were connected if their indel distance was below 5. To reduce the redundancy amongst the top binders, the top scorer in each connected component in the network was selected as a representative peptide, narrowing the dataset down to 51 candidates. We henceforth call these peptides encrypted peptides (EPs, Fig. 3b). The largest cluster of peptides contained 18 peptides derived from a region in Apolipoprotein A1, while the other clusters contained 1-4 members from proteins like hemoglobin B and DNA-PKcs.

While AF2m is useful for rough docking and scoring, it often fails to identify the correct binding pose. We therefore applied conformational sampling using FlexPepDock to refine the binding poses. Each peptide was subject to 100 separate conformational searches for the most optimal binding pose. As expected, the interface ΔG decreased when applying the conformational search and all top 50 peptides achieved an interface ΔG of < -30 Rosetta energy units (REU) amongst the top 10 scoring poses (Fig. 3c). Visual inspection of peptide binders demonstrated that most peptides had α-helical structures mixed with unstructured loop regions and that they tended to fold into the pocket located in the binding site (Fig. 3d).

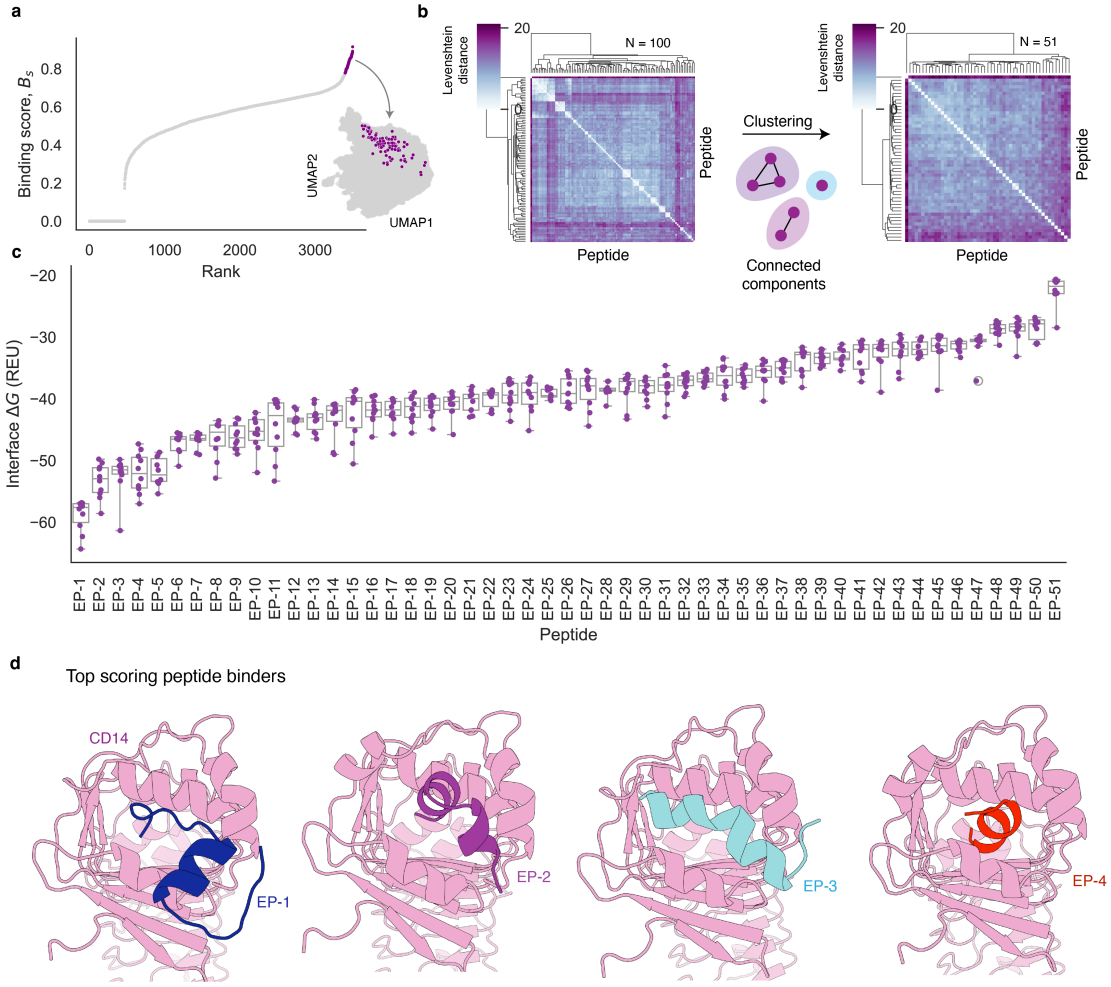


Fig. 3 Peptide selection and refinement of top binders with conformational searches. **a** The binding score of all docked peptides ranked by interface ΔG. The purple scatters in the UMAP highlight the top 100 scoring peptides. **b** Clustered Levenshtein distance matrices for the top 100 peptides (left) and the remaining peptides after removing subpeptide cluster (right). **c** Energy (Rosetta energy units, REU) of the peptides (EPs) after performing conformational searches using FlexPepDock, sorted based on the mean interface ΔG of the 10 best poses. **d** Top scoring pose for EP-1 to EP-4.

All-atom molecular dynamics simulations validates the binders

The top 15 candidates based on ΔG after conformational optimization were selected for validation through MD simulations. Given that the focus of this study was the N-terminal LPS-binding site of

CD14, residues beyond position 130 were removed prior to modeling to create a truncated version of the protein (Fig. 4a). Restraints were applied to residues 120-130 to ensure structural integrity. Stability analyses comparing the truncated region of the full-length and truncated CD14 proteins demonstrated similar stability profiles. Furthermore, protein-peptide contact fingerprint comparisons between the original and truncated CD14 with the known binder sHVF18 showed largely consistent behavior (Extended Data Fig. 5). These findings supported the decision to use the truncated CD14 in subsequent MD simulations.

The top scoring peptide-CD14 complexes were simulated for 200 nanoseconds, during which 2000 snapshots were extracted and analyzed from the trajectories. The first 50 ns were discarded to allow the peptides to settle into more stable conformations. While minor shifts in binding poses were observed for some top peptides, all peptides demonstrated overall stability compared to 3 control peptides which tended to unfold and/or drift in the pocket (Fig. 4b). The control peptides were the 3 worst scorers that were still predicted to bind in the site. This was confirmed by visual inspection of the trajectories and evaluation of peptide RMSD (Fig. 4b, c). To estimate the dissociation constant (K_d) for each peptide, the 500 snapshots with the lowest interface ΔG were selected for robust estimations (Fig. 4c). The K_d was calculated using the equation:

$$K_d = \frac{e^{\Delta G}}{c},$$

where c is a calibration constant derived from previous experimental binding studies of the linear peptide HVF18 ($K_d = 4.7, \mu$M) [20]. Of the 15 peptides, 7 had an estimated $K_d < 4.7, \mu$M, and one peptide was predicted to bind with nanomolar affinity (Fig. 4d). The contacting residues on CD14 align closely with previous analyses of the interacting residues in the CD14-LPS complex, including residues such as D44, S46, F49, and T85 [19]. Notably, many of these peptides also formed interactions deeper within the binding site (residues 100–120) which likely serve to anchor the peptides and contribute to their strong affinity (Fig. 4e-f).

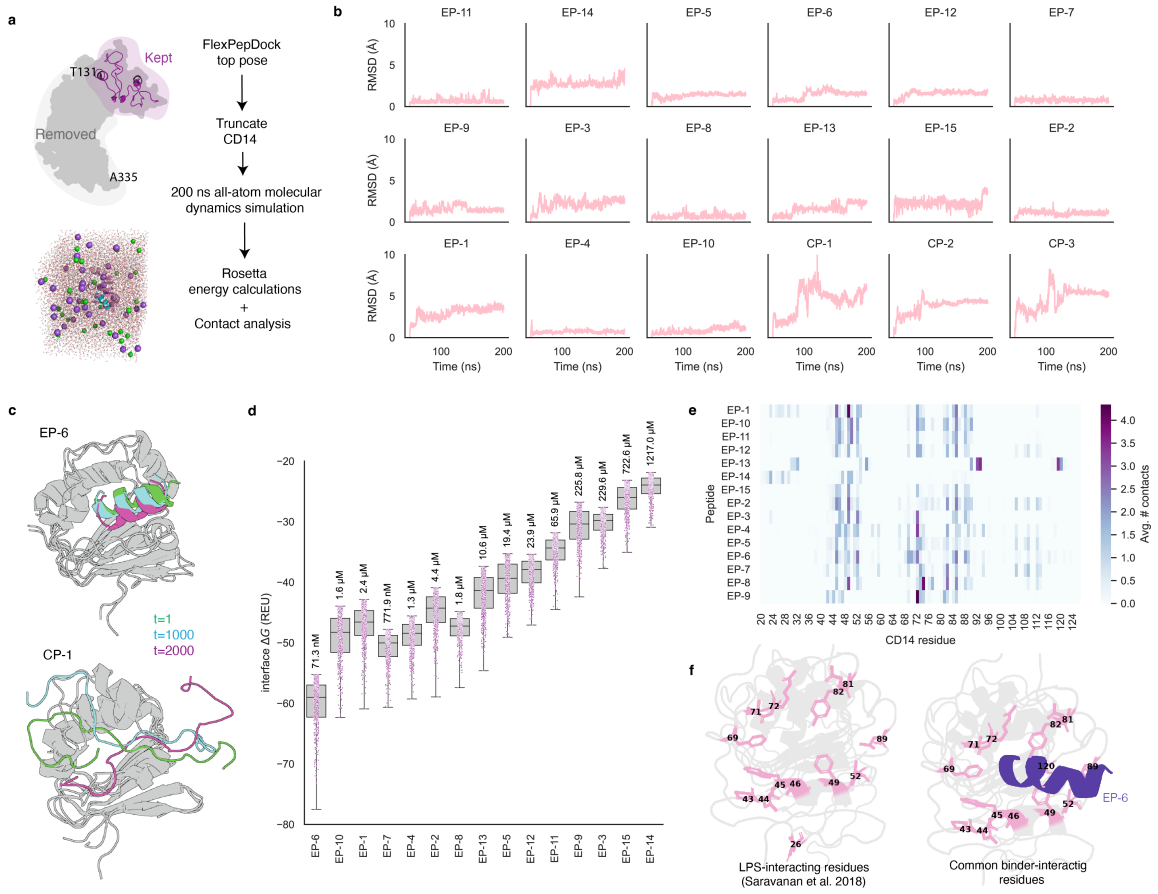


Fig. 4. Molecular dynamics simulations validate peptide binders. **a** The truncated N-terminal domain of CD14 was simulated in complex with each peptide binder for 200 ns using all-atom molecular dynamics simulations. **b.** Root-mean-square deviation (RMSD, Å) of the protein backbone over the simulation trajectories (50-200 ns). CP-(1,2,3) are control peptides, and were selected from the worst binders based on B_s that were still predicted to bind in the pocket. **c** Three superimposed snapshots from the trajectories (frames 1,1000 and 2000) of EP-6 and CP-1. **d** Interface binding energy (ΔG, REU) calculated for the top 500 poses sampled along the simulation trajectories. Peptides are annotated with their estimated dissociation constant (K_d). **e** Average number of contacts for each residue on CD14 highlighting interaction patterns with peptides. A contact was defined by heavy atom distances < 4 Å between the protein and peptide residues. **f** Visualization of LPS-interacting residues on CD14 and residues commonly interacting with peptides, defined as being within 4 Å of the peptide.

Discussion

Identifying peptides that bind with high affinity to specific protein sites in endogenous peptidomic data enables the investigation of selective protein degradation for active encrypted peptides. It also allows for the utilization of the *a priori* optimization conducted by evolution to identify potential therapeutics. Here, we present a method that effectively mines large peptidomic datasets for binders to specific targets by combining optimization with established structural proteomics tools. We apply this method to search for binders to the LPS-binding domain of CD14 from peptidomics data. This leads us to identify several endogenous peptides which are predicted to bind strongly to this domain and interact with LPS-binding residues, and that could provide an immunomodulatory environment

in infectious and inflamed wounds. These findings support the hypothesis that peptide degradation is selective and utilized to generate bioactive peptides which have immunomodulatory effects.

Interestingly, co-folding several thousands of peptides with CD14 led us to identify that most peptides favored positioning in proximity to either the LPS-binding site at the N-terminal or to the C-terminal. These binding hotspots suggest that the protein plays an important role in guiding peptide interactions by providing specific regions that are more favorable for binding. The high number of potential binders also suggest that potential immunomodulatory effects could be mediated by a vast number of peptides with transient interactions, leading to a cumulative effect.

The method we present is generalized and broadly applicable for identifying peptide binders to any target from experimental or theoretical peptidomic datasets. Its modular design also paves the way for future enhancements, such as adopting alternative peptide embedding strategies, refining the cumulative binding score, or swapping out the surrogate model, to further extend its utility.

Our method also has several limitations. A major limitation is the large amount of required compute. With access to 4 GPUs, many hyperparameter choices were constrained by computational feasibility. However, this limitation presents an opportunity for expansion, as increased computational resources could enhance binding predictions by e.g. enabling a greater number of AF models or by incorporating FlexPepDock directly into the Bayesian optimization process. Further refinement could also include adopting Bayesian neural networks or other Bayesian techniques and leveraging hypervolume optimization to better handle multi-objective scoring rather than collapsing individual terms into a single score.

We also suspect that a bias in AlphaFold may impact our findings, as it tends to favor short peptides with well-defined secondary structures (Extended Data Fig. 4) [28]. This and other possible biases could result in false negatives, potentially missing binders that either lack defined secondary structures or cannot be accurately predicted by AlphaFold. Addressing this limitation might be possible by increasing computational resources, as AlphaFold's accuracy has been shown to improve with more seeds and recycles [29], which could help mitigate structural prediction biases.

Despite these limitations, the framework we present is scalable, adaptable, and effective for identifying binders in large endogenous peptidomic datasets. It provides a powerful tool for advancing our understanding of degradomics, the role of encrypted peptides in peptide-protein interactions and in discovering therapeutic peptides.

Method

Peptidomic dataset of infected and inflamed conditions

The peptidomic dataset was curated by integrating data from three clinical sources. These include the wound fluid peptidomes from: non-healing leg ulcers (N=18, 4 patients, PXD: PXD048892), infected wounds (N=3, datamatrices available in [27]), and acute surgical drainage (N=5, datamatrices available in [27]). The peptide GKY25 and its subpeptide HVF18 are generated in an inflammatory setting [30], and are known to bind to the LPS binding pocket of CD14 [19] and were also included, along with homologues and variants of these, as benchmarks. These were included as a benchmark, and because of their previously documented generation in an inflammation setting. The list of peptides was filtered to retain only those with lengths between 5 and 30 amino acids. Peptides exhibiting repeating patterns or containing over 75% of a single amino acid were excluded, as they introduced excessive variability and were unlikely to be binding candidates.

The structure of CD14 used here is based on the crystal structure deposited to the Protein Data Bank (PDB: 4GLP). Similarly to in Petruk et al. [20], 5 residues were added to the N-terminal (residues 20-25) alongside a disulfide bridge (C25-C36) since these residues are proximate to the binding site

and may affect predictions.

When developing the pipeline, a dataset consistent of the porcine wound fluid peptidome (PXD: PXD048892) was used to prevent overfitting during technological development. The human dataset described above was first introduced for the production run.

A filtering step was performed to remove peptides with repetitive sequences or other characteristics that were unlikely to result in strong binding interactions and that introduced artificially high variations in the data, potentially misguiding the optimization process. Peptides with overly repetitive sequences, such as those consisting of a single amino acid repeated multiple times (e.g., `"AAAAAAAA"`), were excluded from further consideration. Specifically, the exclusion criteria included peptides with a single amino acid frequency exceeding 75% or a repeat length of 5 or more. Additionally, peptides with post-translational modifications, non-standard residues or sequences longer than 30 amino acids were removed. The final dataset consisted of 35,863 peptide sequences.

Peptide embedding and featurization

Peptide sequences were transformed into high-dimensional embedding vectors using the pre-trained protein language model, ESM2 (650M). This model generates residue-level embedding matrices, where each peptide is represented as a matrix of dimensions $l \times e_i$, where l is the length of the peptide, and $e_i = 1280$ corresponds to the embedding dimension for each residue. To generate fixed-length embeddings for peptides of varying lengths, residue-level embeddings were averaged, yielding a vector $\in \mathbb{R}^{1280}$ for each peptide.

To improve computational efficiency and reduce the noise in the embedding space, principal component analysis (PCA) was applied to reduce the dimensionality of the embeddings while retaining 95% of the variance. This way, the final dataset was reduced to 136 dimensions. Since we suspected that peptide length is an important feature, we appended this explicitly to the feature vector.

Bayesian optimization for identifying peptide binders

A Bayesian optimization (BO) framework was used to systematically explore the peptidomic space and identify peptides with high binding affinity for a target receptor. This approach integrates predictive modeling, docking simulations, and iterative selection strategies to explore and traverse the peptide landscape. The goal is to efficiently sample diverse peptide sequences and identify those with the highest potential for receptor binding whilst sampling a small fraction of the peptide landscape.

The general BO goes as follows: 1. Map the peptide sequences to a latent vector space. 2. Select candidate peptides. 3. Dock and score the candidates. 4. Train a Bayesian surrogate model to estimate the binding capacity of the candidates. 5. Use the model to select the next batch of peptide candidates using an acquisition function. 6. Iterate 1-5 for a fixed number of iterations or until a stopping criteria is reached.

Initialization

The optimization process commences by selecting an initial set of peptides from the embedding space to ensure diversity in the sampled peptides. To achieve this, batch k-means clustering was applied to group peptides based on their embeddings. The peptide closest to the centroid of each cluster was selected to represent that cluster, ensuring that the initial peptide set spanned different regions of the embedding space.

Peptide docking with AlphaFold2 multimer v3

Initial docking position and potential were carried out using AlphaFold2 multimer v3 (AF2m) as implemented in LocalColabFold (v1.5.5) [23]. Prediction of peptide-receptor complexes were carried

out in single-sequence mode. Additionally, a PDB template was provided for the receptor, avoiding the need for multiple sequence alignments. The structural refinement of each complex was performed over nine recycles, with a positional tolerance of 0.3 Å. For each peptide, all 5 different AF2m models were used, and dockings were parallelized across four Tesla T4 GPUs to further accelerate the process.

Constructing the binding score

To obtain a reliable estimate of binding affinity, we developed a multi-objective scoring function that integrates several metrics. The rationale being that by themselves, these scores represent noisy estimates of different aspects of the true binding affinity, and that the score-dependent error and bias may be reduced when combining them. Additionally, an optimal peptide binder should score well amongst all metrics. These scores include the inter-protein Template Modeling score (ipTM) from AF2m, which reflects the certainty in the relative positioning of protein chains. It also includes the predicted change in interface Gibbs free energy (ΔG) upon binding, calculated using the Rosetta suite. Additionally, the Rosetta score (R_{score}), which represents the overall energy of the system, was incorporated, along with the change in solvent-accessible surface area (ΔSASA), indicating the extent of interface formation during peptide binding. We also considered the change in unsatisfied hydrogen bonds (ΔH_{unsat}), a metric that can affect both binding stability and specificity, and the packing statistic, P_{stat}, which evaluates the quality of side-chain packing in the interface, contributing to the structural stability of the complex. Scoring was done in full-atom mode using the scoring matrix `ref2015`.

In each iteration of the optimization process, these metrics were scaled to a common range between 0.1 and 1, ensuring comparability across different units and magnitudes. The binding site was defined by key residues which contact LPS during interaction, specifically residue numbers 44, 49, 74, 82, 89, 105. Peptide candidates that were not localized within 5 Å of any of these residues, or those with fewer than two models in agreement with the highest scoring model (from multiple structural predictions), were assigned a binding score of 0 to penalize them.

The scaled metrics were subsequently combined into a single binding score, B_s, by taking their weighted sum. A weight of 0.5 was assigned to the ipTM score. The remaining 0.5 was equally distributed amongst the interface-based metrics derived from the Rosetta suite, including ΔG, R_{score}, ΔSASA, ΔH_{unsat}, and P_{stat}. The complete expression is therefore:

$$B_s = \mathbf{w}^T \mathbf{m},$$

where $\mathbf{m}$ is the vector of scaled metrics:

$$\mathbf{m} = [\text{ipTM}, \Delta G, R_{\text{score}}, \Delta\text{SASA}, \Delta H_{\text{unsat}}, P_{\text{stat}}],$$

and $\mathbf{w}$ is the vector of corresponding weights:

$$\mathbf{w} = [0.5, 0.1, 0.1, 0.1, 0.1, 0.1].$$

Multi-objective optimization with a deep ensemble

The surrogate model in the BO is an ensemble consisting of 10 multi-layer perceptrons (MLPs). Each MLP in the ensemble is trained to predict the binding score (B_s) from the peptide embeddings for the peptides from previous docking experiments by minimizing the $R^2 \in (-\infty, 1]$ score, which is defined as:

$$R^2 = (1 - \frac{u}{v})$$

where u is the residual sum of squares and v the total sum of squares. To monitor the performance of the MLPs during the optimization process, they are evaluated on a 80/20 train/validation split. This split is also used to determine the number of training epochs using early stopping. After validation, the models are retrained on the complete dataset, using the number of epochs identified with early stopping.

Since the dataset evolves during the optimization process - both in size and complexity as new peptides are docked, the hyperparameters of the MLPs are adjusted dynamically to maintain model performance. Therefore, hyperparameter optimization is conducted every 50th iteration of the optimization process. For this, we employed Optuna using the Tree-structured Parzen Estimator (TPE) as the sampler. The optimized parameters are the number of layers, the size of the initial layer, the alpha regularization parameter and the initial learning rate. Other parameters such as the choice of activation function (ReLU) and the solver (ADAM) are fixed. The neurons in each layer after the initial layer is halved in size. These are kept constant to restrict the search space and reduce computational demands.

Each network in the ensemble produces a prediction for the binding score, and the final output prediction is the mean, $\mu(\mathbf{x})$, of these predictions, while $\sigma(\mathbf{x})$ reflects the model's epistemic uncertainty, i.e. the uncertainty due to limited knowledge or insufficient data in certain regions of the search space.

Exploring and exploiting to identify the best candidates

The BO operates in two modes: exploration and exploitation.

During the exploration phase, peptides are selected with the aim to reduce the uncertainty of the surrogate model by guiding the selection to regions of the search space where the model has high uncertainty. During this phase, the next batch of peptides is chosen based on the standard deviation of the ensemble's predictions, $\sigma(\mathbf{x})$. By sampling in these uncertain regions, the model gradually reduces its epistemic uncertainty, improving its overall knowledge of the search space. Mathematically, the selection criterion for exploration is simply:

$$\mathbf{x}_{\text{next}} = \arg\max_{\mathbf{x}} \sigma(\mathbf{x}),$$

During the exploitation phase we shift the focus to selecting peptides that are likely to yield high binding scores based on the model's current predictions. There are many acquisition functions that accomplish this, such as simply selecting the peptides with the highest predicted mean binding score, or by taking uncertainty into account using metrics like the upper confidence bound (UCB) or the expected improvement (EI). Here, we use EI.

The EI for a candidate peptide with embedding $\mathbf{x}$ is defined as:

$$\text{EI}(\mathbf{x}) = \mathbb{E}\left[\max(\mu(\mathbf{x}) - \mu(\mathbf{x}^+), 0)\right],$$

where $\mu(\mathbf{x}^+)$ is the highest observed binding score so far. The expectation, $\mathbb{E}$, is taken over the predictive distribution of $\mu(\mathbf{x})$. Assuming that this predictive distribution is normal with mean $\mu(\mathbf{x})$ and standard deviation $\sigma(\mathbf{x})$, EI can be expressed as:

$$\text{EI}(\mathbf{x}) = (\mu(\mathbf{x}) - \mu(\mathbf{x}^+))\Phi(Z) + \sigma(\mathbf{x})\phi(Z),$$

where $Z = \frac{\mu(\mathbf{x})-\mu(\mathbf{x}^+)}{\sigma(\mathbf{x})}$, and $\Phi(Z)$ and $\phi(Z)$ are the cumulative distribution function and probability density function of the standard normal distribution, respectively.

To ensure that the selected peptides are diverse in each batch we employ a k-center greedy selection approach. This weighs the expected improvement of each peptide with its distance from other peptides

in the batch, encouraging the selection of peptides that are both promising and dissimilar. Mathematically, given a set of already selected peptides $P = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{k-1}\}$, the next peptide $\mathbf{x}_k$ is chosen as:

$$\mathbf{x}_k = \arg\max_{\mathbf{x} \in X} \min_{\mathbf{x}_i \in P} \left(\mathrm{dist}(\mathbf{x}, \mathbf{x}_i) \cdot \mathrm{EI}(\mathbf{x}) \right),$$

where $\mathrm{dist}(\mathbf{x}, \mathbf{x}_i)$ is the euclidean distance between peptide embeddings $\mathbf{x}$ and $\mathbf{x}_i$. This approach ensures that we don't waste compute on similar peptides.

Conformational searches using FlexPepDock

The most promising candidates based on the binding score were further evaluated with FlexPepDock as implemented in PyRosetta. In FlexPepDock, their pose is relaxed and refined using a high-resolution Monte Carlo-based energy minimization protocol that optimizes peptide backbone and side-chain conformations, as well as peptide-receptor interactions. This step allows for fine-tuning of the peptide's binding orientation and improves the accuracy of the predicted pose as well as interaction quality.

The top 100 peptides by binding score were selected for FlexPepDock searches. Several of these peptides came from the same peptide cluster. To reduce the redundancy in the dataset, peptide networks were created by connecting peptides with an indel distance < 5. Isolated connected components were defined as a cluster, and was represented by its top binder. This reduced the dataset to 64 peptides. The FlexPepDock protocol was applied 100 times on these 64 peptides. The top candidates after relaxation were subject to further binding energy evaluation and ranking.

Molecular dynamics simulations of top-scoring peptides

To further assess the binding capacity of candidate peptides, we selected 15 peptides with the highest FlexPepDock interface ΔG score (Figure 3C) and subjected them to atomic resolution molecular dynamics (MD) simulations. The initial coordinates of CD14-peptide complexes were obtained from FlexPepDock. To reduce computational resources required to run these simulations, we truncated residue 131-335 in CD14 as extensive microsecond timescale simulations with sHVF18 showed similar peptide binding profile and stability to the full-length protein simulations performed in our previous study (Extended Data Fig. 3) [20].

The peptide-truncated CD14 complexes were parametrized using the CHARMM36m forcefield [31]. The complexes were solvated with TIP3P water molecules and 0.15 M NaCl salt was subsequently added to neutralize the system. Steepest descent energy minimization was then performed to remove any atomic clashes. A short 125 ps equilibration was performed with positional restraints (force constant, Fc = 400 kJ $\mathrm{mol}^{-1}\mathrm{nm}^{-2}$) applied to all backbone atoms of the protein and peptide. Afterwards a 200 ns production MD simulations was performed for each complex using GROMACS 2024 [32]. To maintain the stability of the truncated CD14 during the production runs, we applied positional restraints (Fc = 500 kJ $\mathrm{mol}^{-1}\mathrm{nm}^{-2}$) to the backbone atoms of residues 120-130. The temperature of the system was maintained at 310 K using the Nosé-Hoover thermostat [33,34], while the pressure was maintained at 1 atm by an isotropic coupling to a Parrinello-Rahman barostat [35]. Electrostatic interactions were computed using the particle mesh Ewald (PME) method [36], while the van der Waals interactions were truncated at 1.2 nm with a force switch function applied between 1.0-1.2 nm. A 2-fs integration timestep was used whereby all covalent bonds involving hydrogen atoms were constrained using the LINCS algorithm [37]. Snapshots of the MD trajectories were saved at 0.1 ns intervals to capture the conformational dynamics. These snapshots were then used to calculate interface characteristics using PyRosetta.

To translate ΔG expressed in Rosetta energy units (REU) to the dissociation constant, K_d, we utilize the relationship:

$$\Delta G = RT \log K_d.$$

Since the simulations are equilibrated with the same parameters, RT is constant. Thus, we can express ΔG as proportional to $\log K_d$:

$$\Delta G = c \log K_d.$$

Here, c is a proportionality constant that can be determined using experimental data. Previous binding experiments determined the dissociation constant for the linear peptide HVF18 to be $K_d = 4.7\,\mu$M. Using this value and ΔG from molecular dynamics (MD) simulations, we calculate c.

Once c is known, the dissociation constant for any peptide p can be computed using the formula:

$$K_{d,p} = \frac{e^{\Delta G_p}}{c}.$$

To quantify protein-peptide interactions, MDAnalysis was used [38]. Pairwise distances between all non-hydrogen atoms of CD14 and each peptide were computed from their MD trajectories. A contact was defined as any residue pair having at least one atom-atom distance within a 4.0 Å cutoff.

Computing peptide properties

All peptide properties except the relative hydrophobic moment, uH_{rel}, was calculated using Biopython [39]. These include molecular weight, aromaticity, instability index, isoelectric point, grand average of hydropathy (GRAVY), and secondary structure fractions (helix, turn, sheet). Amino acid compositions were determined for hydrophobic, polar, positively charged, and negatively charged residues, and the net charge was estimated by subtracting the fraction of negatively charged residues from the fraction of positively charged residues. To assess the uH_{rel}, which quantifies the distribution of hydrophobic and hydrophilic residues along the peptide sequence, assuming an α-helical conformation [40]. The hydrophobic moment (μ_H) was calculated using the Kyte-Doolittle hydrophobicity scale (Extended Data Table 1) with the following equations:

$$\mu_H = \sqrt{\left(\sum_{i=1}^{N} H_i \cos\theta_i\right)^2 + \left(\sum_{i=1}^{N} H_i \sin\theta_i\right)^2}$$

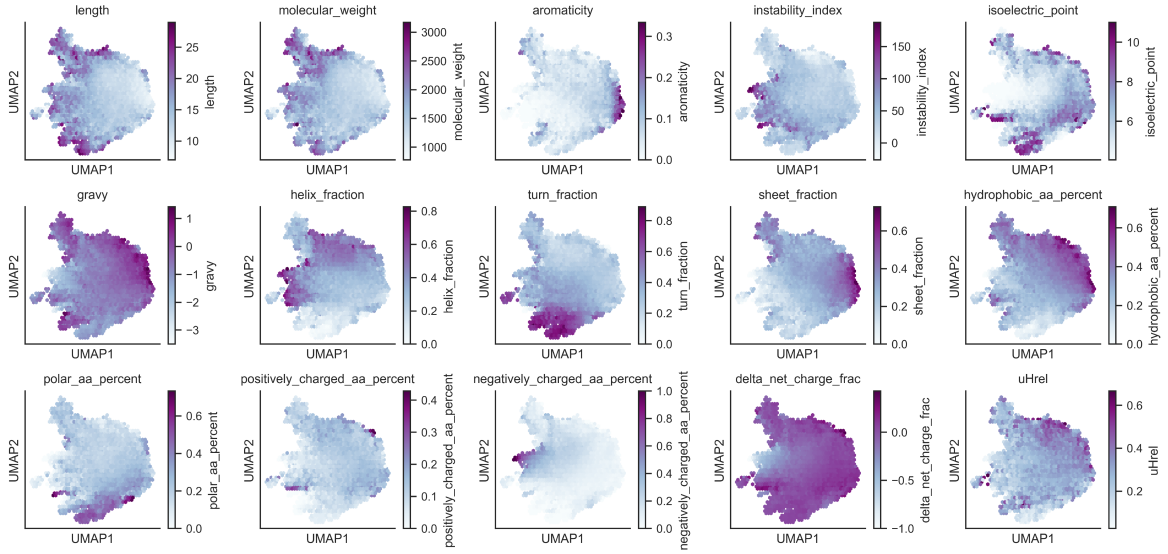
where H_i is the hydrophobicity of the i-th amino acid, $\theta_i = i \times \frac{100^\circ \times \pi}{180^\circ}$, and N is the total number of amino acids in the peptide. The relative hydrophobic moment is then calculated by normalizing μ_H by the sum of the absolute hydrophobicities:

$$uH_{\text{rel}} = \frac{\mu_H}{\sum_{i=1}^{N} |H_i|}$$

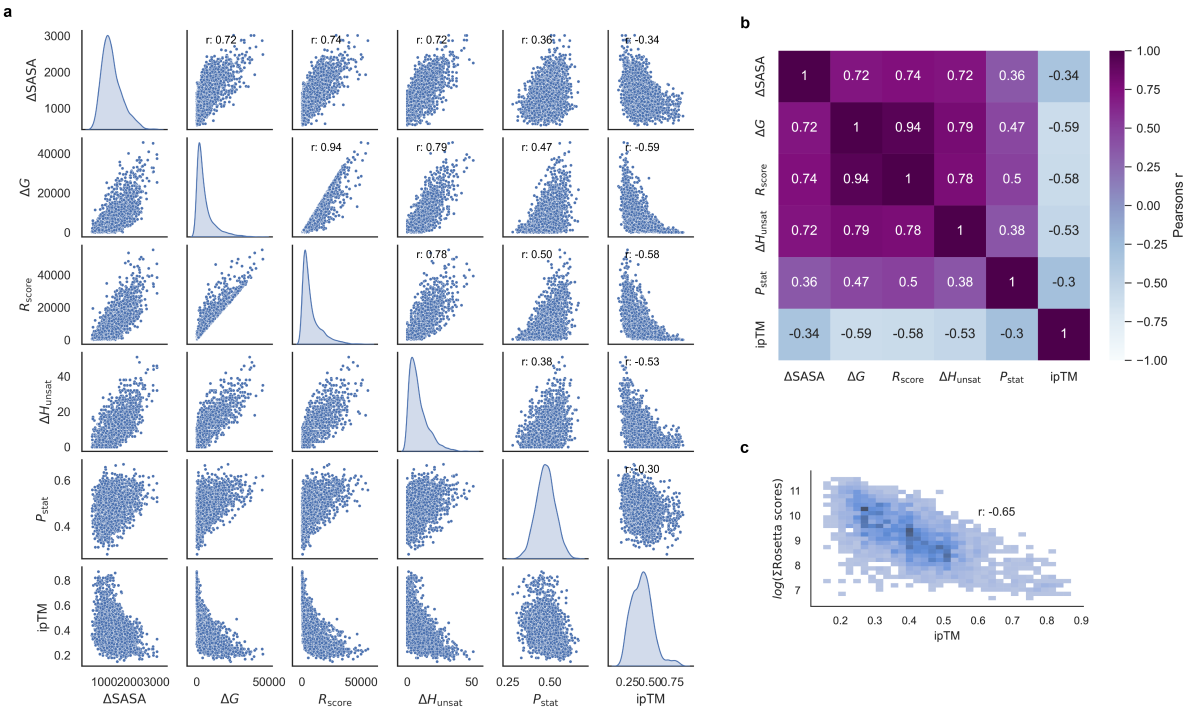
Extended Data

Extended Data Table 1 The hydrophbocity constants used for calculating uH_{rel} (Kyte-Doolittle scale).

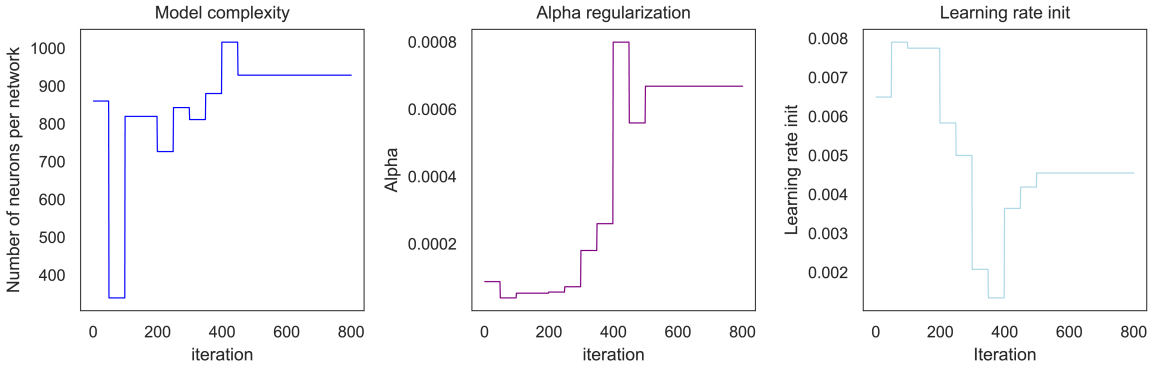
Amino acid	Hydrophobicity value
A	1.8
R	-4.5
N	-3.5
D	-3.5
C	2.5
Q	-3.5
E	-3.5
G	-0.4
H	-3.2
I	4.5
L	3.8
K	-3.9
M	1.9
F	2.8
P	-1.6
S	-0.8
T	-0.7
W	-0.9
Y	-1.3
V	4.2



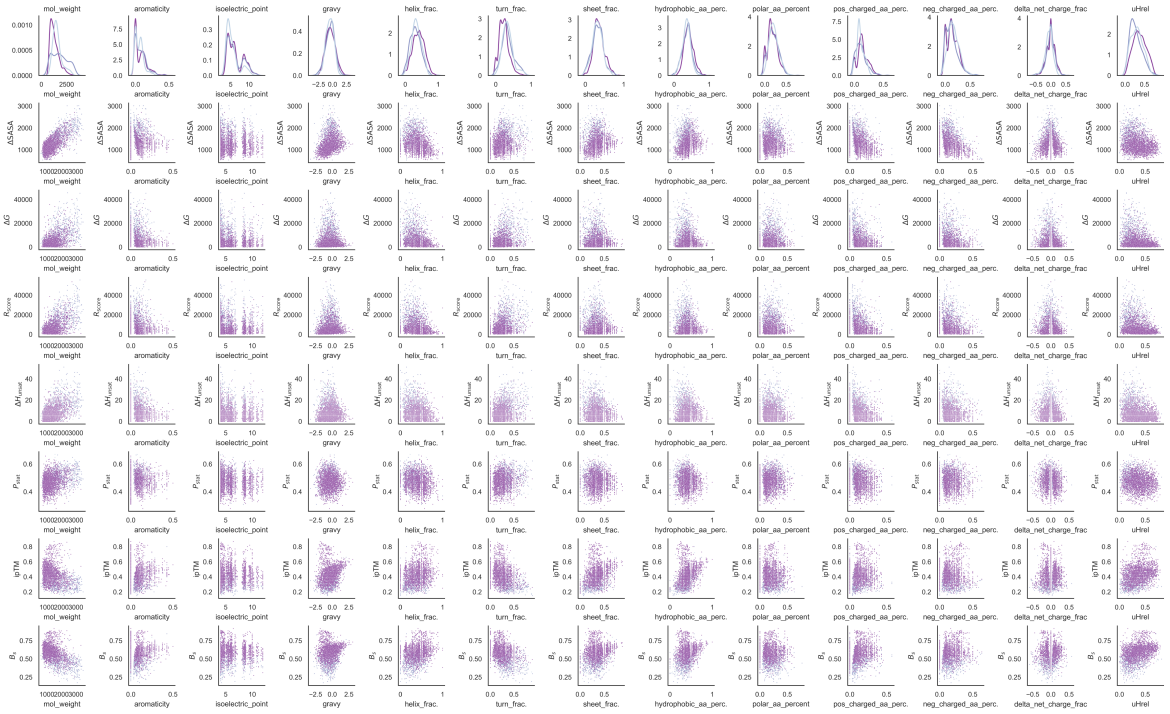
Extended Data Fig. 1 Distribution of biophysical peptide properties in the latent space. Hexbin UMAP projections of the embedded peptide space. Each hexagon aggregates the values of the peptides it covers in the embedding space, and the color indicates the average value of a given biophysical property that may influence binding capacity.



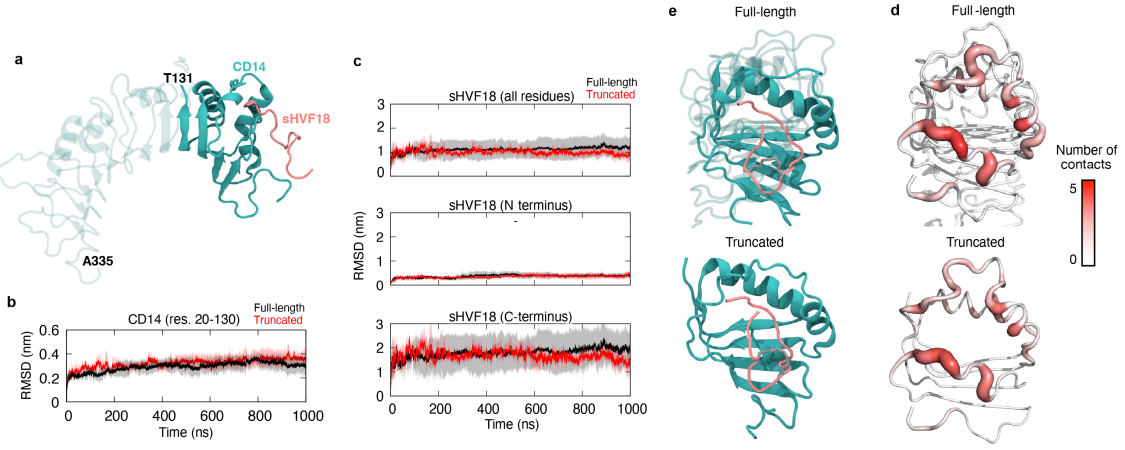
Extended Data Fig. 2 Correlation between individual interface scoring terms. a Pairwise correlation between individual score terms annotated with Pearson correlation coefficient. **b** Heatmap of Pearson correlation coefficient between individual score terms. **c** The correlation between ipTM and the scaled and weighted sum of the Rosetta-based terms.



Extended Data Fig. 3 Selected hyperparameters during optimization. Surrogate model complexity (number of nodes in the networks), alpha regularization parameter and initial learning rate over the optimization process. These are selected dynamically every 50th iteration using Bayesian hyperparameter optimization (see Methods).



Extended Data Fig. 4 Correlating the binding score terms and biophysical peptide properties. The top row shows density plots of the biophysical score terms, separated based on optimization phase. The other rows correlate the individual score terms and binding score against the biophysical properties. The correlations are only conducted for the peptides that were docked in proximity to the binding site.



Extended Data Fig. 5 Comparison of simulations of docking to full-length and truncated CD14. **a** The structure of CD14-sHVF18 complex was obtained from a previous study (a representative structure from microsecond timescale atomic simulations). Residues 131-335 on CD14 were removed to generate a truncated version. Positional restraints with Fc=500 kJ mol^{-1} nm^{-1} were applied on the backbone atoms of residues 120-130 of the truncated CD14. Three independent 1 µs simulations were performed and compared to the full-length CD14 simulations. **b** Root means square deviations (RMSD) of the backbone atoms of CD14 residues 20-130 from simulations of full-length CD14 (black) or truncated CD14 (red). Average values from three simulations are shown as thick

lines, and standard deviations are depicted as shaded areas. **c** RMSD of backbone atoms of the full-length sHVF18 (top), the first 5 residues on the N-terminus (middle), and the last 5 residues on the C-terminus (bottom). **d** Average number of contacts made by each residue of CD14 with sHVF18 projected as a heat map onto the protein structure. **e** Representative structures obtained from clustering analysis performed using an RMSD cut-off of 0.4 nm. The figure shows the central structures of the top clusters. CD14 in cyan, sHVF18 peptide in pink, and the staple shown in gray in stick representation.

Ackowledgements

EH was supported by the European Molecular Biology Organization (EMBO, Scientific Exchange Grant 11167) and the Royal Physiographic Society of Lund. FS and PJB are supported by the BII (A*STAR) core funds. MD simulations were performed on a local cluster at BII (A*STAR) and the supercomputer Fugaku provided by RIKEN through the HPCI System Research Project (Project ID: hp240347). AS are supported by the Swedish Research Council (project 2020-02016), Edvard Welanders Stiftelse and Finsenstiftelsen (Hudfonden), the Swedish Government Funds for Clinical Research (ALF). JM was supported by the Swedish Research Council (project 2018-05795, 2019-01646) and Alfred Österlund Foundation.

Supplementary

The XML-script used to run FlexPepDock in PyRosetta:

```
<ROSETTASCRIPTS>
        <TASKOPERATIONS>
        </TASKOPERATIONS>
        <SCOREFXNS>
        </SCOREFXNS>
        <FILTERS>
        </FILTERS>
        <MOVERS>
            <FlexPepDock name="prepack" ppk_only="1" />
            <FlexPepDock name="fpd"  pep_refine="1" lowres_preoptimize="1" />
        </MOVERS>
        <APPLY_TO_POSE>
        </APPLY_TO_POSE>
        <PROTOCOLS>
                <Add mover_name="prepack" />
                <Add mover_name="fpd"/>
        </PROTOCOLS>
</ROSETTASCRIPTS>
```

References

1. Muttenthaler M, King GF, Adams DJ, Alewood PF. Trends in peptide drug discovery. Nature Reviews Drug Discovery [Internet]. 2021 Feb;20[4]:309–25. Available from: http://dx.doi.org/10.1038/s41573-020-00135-8

2. Tsomaia N. Peptide therapeutics: Targeting the undruggable space. European Journal of Medicinal Chemistry [Internet]. 2015 Apr;94:459–70. Available from: http://dx.doi.org/10.1016/j.ejmech.2015.01.014

3. Fosgerau K, Hoffmann T. Peptide therapeutics: Current status and future directions. Drug Discovery Today [Internet]. 2015 Jan;20[1]:122–8. Available from: http://dx.doi.org/10.1016/j.drudis.2014.10.003

4. Wan F, Kontogiorgos-Heintz D, Fuente-Nunez C de la. Deep generative models for peptide design. Digital Discovery [Internet]. 2022;1[3]:195–208. Available from: http://dx.doi.org/10.1039/d1dd00024a

5. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. bioRxiv. 2022;

6. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature [Internet]. 2021 Jul;596[7873]:583–9. Available from: http://dx.doi.org/10.1038/s41586-021-03819-2

7. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science [Internet]. 2021 Aug;373[6557]:871–6. Available from: http://dx.doi.org/10.1126/science.abj8754

8. Chen T, Dumas M, Watson R, Vincoff S, Peng C, Zhao L, et al. PepMLM: Target sequence-conditioned generation of therapeutic peptide binders via span masked language modeling [Internet]. arXiv; 2023. Available from: https://arxiv.org/abs/2310.03842

9. Li Q, Vlachos EN, Bryant P. Design of linear and cyclic peptide binders of different lengths from protein sequence information. 2024 Jun; Available from: http://dx.doi.org/10.1101/2024.06.20.599739

10. Pacesa M, Nickel L, Schmidt J, Pyatova E, Schellhaas C, Kissling L, et al. BindCraft: One-shot design of functional protein binders. 2024 Oct; Available from: http://dx.doi.org/10.1101/2024.09.30.615802

11. Bhat S, Palepu K, Hong L, Mao J, Ye T, Iyer R, et al. De NovoDesign of peptide binders to conformationally diverse targets with contrastive language modeling. 2023 Jun; Available from: http://dx.doi.org/10.1101/2023.06.26.546591

12. Wang F, Wang Y, Feng L, Zhang C, Lai L. Target-specific de novo peptide binder design with DiffPepBuilder. Journal of Chemical Information and Modeling [Internet]. 2024 Sep; Available from: http://dx.doi.org/10.1021/acs.jcim.4c00975

13. Madsen CT, Refsgaard JC, Teufel FG, Kjærulff SK, Wang Z, Meng G, et al. Combining mass spectrometry and machine learning to discover bioactive peptides. Nature Communications [Internet]. 2022 Oct;13[1]. Available from: http://dx.doi.org/10.1038/s41467-022-34031-z

14. Wan F, Torres MDT, Peng J, Fuente-Nunez C de la. Deep-learning-enabled antibiotic discovery through molecular de-extinction. Nature Biomedical Engineering [Internet]. 2024 Jun;8[7]:854–71. Available from: http://dx.doi.org/10.1038/s41551-024-01201-x

15. Eberhardt J, Santos-Martins D, Tillack AF, Forli S. AutoDock vina 1.2.0: New docking methods, expanded force field, and python bindings. Journal of Chemical Information and Modeling [Internet]. 2021 Jul;61[8]:3891–8. Available from: http://dx.doi.org/10.1021/acs.jcim.1c00203

16. Trellet M, Melquiond ASJ, Bonvin AMJJ. A unified conformational selection and induced fit approach to protein-peptide docking. Keskin O, editor. PLoS ONE [Internet]. 2013 Mar;8[3]:e58769. Available from: http://dx.doi.org/10.1371/journal.pone.0058769

17. Kozakov D, Hall DR, Xia B, Porter KA, Padhorny D, Yueh C, et al. The ClusPro web server for protein–protein docking. Nature Protocols [Internet]. 2017 Jan;12[2]:255–78. Available from: http://dx.doi.org/10.1038/nprot.2016.169

18. Alam N, Goldstein O, Xia B, Porter KA, Kozakov D, Schueler-Furman O. High-resolution global peptide-protein docking using fragments-based PIPER-FlexPepDock. Dunbrack RL, editor. PLOS Computational Biology [Internet]. 2017 Dec;13[12]:e1005905. Available from: http://dx.doi.org/10.1371/journal.pcbi.1005905

19. Saravanan R, Holdbrook DA, Petrlova J, Singh S, Berglund NA, Choong YK, et al. Structural basis for endotoxin neutralisation and anti-inflammatory activity of thrombin-derived c-terminal peptides. Nature Communications [Internet]. 2018 Jul;9[1]. Available from: http://dx.doi.org/10.1038/s41467-018-05242-0

20. Petruk G, Puthia M, Samsudin F, Petrlova J, Olm F, Mittendorfer M, et al. Targeting toll-like receptor-driven systemic inflammation by engineering an innate structural fold into drugs. Nature Communications [Internet]. 2023 Sep;14[1]. Available from: http://dx.doi.org/10.1038/s41467-023-41702-y

21. Puthia M, Butrym M, Petrlova J, Strömdahl AC, Andersson MÅ, Kjellström S, et al. A dual-action peptide-containing hydrogel targets wound infection and inflammation. Science Translational Medicine [Internet]. 2020 Jan;12[524]. Available from: http://dx.doi.org/10.1126/scitranslmed.aax6601

22. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. Protein complex prediction with AlphaFold-multimer. 2021 Oct; Available from: http://dx.doi.org/10.1101/2021.10.04.463034

23. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: Making protein folding accessible to all. Nature Methods [Internet]. 2022 May;19[6]:679–82. Available from: http://dx.doi.org/10.1038/s41592-022-01488-1

24. Chaudhury S, Lyskov S, Gray JJ. PyRosetta: A script-based interface for implementing molecular modeling algorithms using rosetta. Bioinformatics [Internet]. 2010 Jan;26[5]:689–91. Available from: http://dx.doi.org/10.1093/bioinformatics/btq007

25. Raveh B, London N, Schueler-Furman O. Sub-angstrom modeling of complexes between flexible peptides and globular proteins: Sub-angstrom modeling of flexible peptides. Proteins: Structure, Function, and Bioinformatics [Internet]. 2010 Mar;78[9]:2029–40. Available from: http://dx.doi.org/10.1002/prot.22716

26. Hartman E, Forsberg F, Kjellström S, Petrlova J, Luo C, Scott A, et al. Peptide clustering enhances large-scale analyses and reveals proteolytic signatures in mass spectrometry data. Nature Communications [Internet]. 2024 Aug;15[1]. Available from: http://dx.doi.org/10.1038/s41467-024-51589-y

27. Hartman E, Wallblom K, Plas MJA van der, Petrlova J, Cai J, Saleh K, et al. Bioinformatic analysis of the wound peptidome reveals potential biomarkers and antimicrobial peptides. Frontiers in Immunology [Internet]. 2021 Feb;11. Available from: http://dx.doi.org/10.3389/fimmu.2020.620707

28. McDonald EF, Jones T, Plate L, Meiler J, Gulsevin A. Benchmarking AlphaFold2 on peptide structure prediction. Structure [Internet]. 2023 Jan;31[1]:111–119.e2. Available from: http://dx.doi.org/10.1016/j.str.2022.11.012

29. Raouraoua N, Mirabello C, Véry T, Blanchet C, Wallner B, Lensink MF, et al. MassiveFold: Unveiling AlphaFold's hidden potential with optimized and parallelized massive sampling. Nature Computational Science [Internet]. 2024 Nov;4[11]:824–8. Available from: http://dx.doi.org/10.1038/s43588-024-00714-4

30. Saravanan R, Adav SS, Choong YK, Plas MJA van der, Petrlova J, Kjellström S, et al. Proteolytic signatures define unique thrombin-derived peptides present in human wound fluid in vivo. Scientific Reports [Internet]. 2017 Oct;7[1]. Available from: http://dx.doi.org/10.1038/s41598-017-13197-3

31. Huang J, MacKerell AD. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. Journal of Computational Chemistry [Internet]. 2013 Jul;34[25]:2135–45. Available from: http://dx.doi.org/10.1002/jcc.23354

32. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX [Internet]. 2015 Sep;1–2:19–25. Available from: http://dx.doi.org/10.1016/j.softx.2015.06.001

33. Hoover WG. Canonical dynamics: Equilibrium phase-space distributions. Physical Review A [Internet]. 1985 Mar;31[3]:1695–7. Available from: http://dx.doi.org/10.1103/PhysRevA.31.1695

34. Nosé S. A unified formulation of the constant temperature molecular dynamics methods. The Journal of Chemical Physics [Internet]. 1984 Jul;81[1]:511–9. Available from: http://dx.doi.org/10.1063/1.447334

35. Parrinello M, Rahman A. Polymorphic transitions in single crystals: A new molecular dynamics method. Journal of Applied Physics [Internet]. 1981 Dec;52[12]:7182–90. Available from: http://dx.doi.org/10.1063/1.328693

36. Essmann U, Perera L, Berkowitz ML, Darden T, Lee H, Pedersen LG. A smooth particle mesh ewald method. The Journal of Chemical Physics [Internet]. 1995 Nov;103[19]:8577–93. Available from: http://dx.doi.org/10.1063/1.470117

37. Hess B, Bekker H, Berendsen HJC, Fraaije JGEM. LINCS: A linear constraint solver for molecular simulations. Journal of Computational Chemistry [Internet]. 1997 Sep;18[12]:1463–72. Available from: http://dx.doi.org/10.1002/(SICI)1096-987X(199709)18:12%3C1463::AID-JCC4%3E3.0.CO;2-H

38. Michaud-Agrawal N, Denning EJ, Woolf TB, Beckstein O. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. Journal of Computational Chemistry [Internet]. 2011 Apr;32[10]:2319–27. Available from: http://dx.doi.org/10.1002/jcc.21787

39. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: Freely available python tools for computational molecular biology and bioinformatics. Bioinformatics [Internet]. 2009 Mar;25[11]:1422–3. Available from: http://dx.doi.org/10.1093/bioinformatics/btp163

40. Eisenberg D, Weiss RM, Terwilliger TC. The helical hydrophobic moment: A measure of the amphiphilicity of a helix. Nature [Internet]. 1982 Sep;299[5881]:371–4. Available from: http://dx.doi.org/10.1038/299371a0