Design of linear and cyclic peptide binders of different lengths only from a protein target sequence

Qiuzhen Li[1,2], Efstathios Nikolaos Vlachos[1,2] and Patrick Bryant[1,2*]

1. The Department of Molecular Biosciences, The Wenner-Gren Institute, Stockholm University, Svante Arrhenius väg 20C, 114 18 Stockholm, Sweden
2. Science for Life Laboratory, 172 21, Solna, Sweden

***Corresponding author, email: patrick.bryant@scilifelab.se**

Abstract

Structure prediction technology has revolutionised the field of protein design, but key questions such as how to design new functions remain. Many proteins exert their functions through interactions with other proteins, and a significant challenge is designing these interactions effectively. While most efforts have focused on larger, more stable proteins, shorter peptides offer advantages such as lower manufacturing costs, reduced steric hindrance, and the ability to traverse cell membranes when cyclized.

Here, we present an AI method to design novel linear and cyclic peptide binders of varying lengths based solely on a protein target sequence. Our approach does not specify a binding site or the length of the binder, making the procedure completely blind. We demonstrate that high-affinity binders can be selected directly from predicted confidence metrics, and adversarial designs can be avoided through orthogonal *in silico* evaluation, tripling the success rate.

We selected a single designed linear sequence for lengths ranging from 8 to 20 residues and evaluated the affinity using surface plasmon resonance (SPR). Of the sequences tested, 6 out of 13 (46%) displayed affinity with dissociation constants (Kd) in the micromolar range; the strongest binder had a Kd of 19 nM, and the weakest had a Kd of 7.9 μM. Our protocol, *EvoBind2* (https://github.com/patrickbryant1/EvoBind), enables binder design based solely on a protein target sequence, suggesting the potential for a rapid increase in the number of proteins that can be targeted for various biotechnological applications.

Introduction

Protein design is in its golden age due to the advent of accurate structure prediction technology [1–3]. While single protein structures can now be designed with high accuracy [4,5], designing new functions, such as binding, without relying on structural information remains challenging [6,7]. Additionally, there are numerous issues to address before designed proteins can benefit society, including scalable production. However, shorter proteins, such as peptides, can bind with high affinity to a diverse set of targets [8,9] and are cost-effective to produce via conventional chemical synthesis.

The success rate in predicting the structure of protein complexes using AlphaFold2 (AF) and AlphaFold-Multimer (AFM) is approximately 60% [2,3]. However, for protein-peptide complexes, AF reports a high false negative rate (84/96) [10,11]. While AF can identify true binders with high confidence (high PPV), it misses most binders [12]. Therefore, the challenge in binder design is to find a sequence likely to bind that AF can also accurately predict. The subset of such sequences among the 20^L possibilities may be very small (or even absent), necessitating a highly efficient and robust search process.

Currently, most protein design workflows consist of two separate steps. Methods like RFdiffusion [5] and backpropagation through AF [13] have been used to design backbone scaffolds for both single and multi-chain proteins. The sequences for these scaffolds are then redesigned using ProteinMPNN [14]. Alternatively, a joint search over a relaxed sequence-structure space [15,16] can be performed, allowing for simultaneous design of sequence and structure. This approach may increase the likelihood of success as the joint sequence-structure space encompasses all possible sequences and structures.

AlphaFold2 relies on coevolution to produce accurate predictions and has been trained to be robust to minor changes in a multiple sequence alignment (MSA), thereby learning the redundancy between sequence and structure [17]. For design tasks, where no MSA or coevolution exists and only a single sequence is used [14,18], AF must infer the structure from the single sequence. Small changes such as single mutations can now have a significant impact on the outcome and it has been shown that AF can distinguish such changes in the binding residues of peptides [10,16].

Based on these principles, we developed a method to generate peptide binders for a target protein using only the protein target sequence. We allow AF to select a suitable binding site, binding sequence, and structure without providing any predefined information. To avoid adversarial designs and increase the success rate, we incorporate a modified version of AFM [2] for evaluation. Additionally, we implement a cyclic offset [19], [20], demonstrating that cyclic peptide binders can be designed to match the efficacy of linear designs using the same input: a single protein target sequence.

Results

Linear and Cyclic peptide binder design

Many proteins exert their functions through interactions. The ability to target specific proteins through binding is essential for numerous biotechnological applications such as fluorescent labelling [21], payload delivery [22], targeted degradation [23] and receptor activation/inhibition [24,25]. For novel targets, information about **1)** target structure, **2)** target site (residues), or **3)** binder size is often unknown. Therefore, we set out to design novel peptide binders for a target protein using only the target protein sequence.

To design binders, we create a framework based on protein structure prediction [1], **EvoBind2,** for *in silico* directed evolution [15], relying solely on the protein target sequence (Figure 1a). EvoBind2 iteratively mutates a peptide binder sequence to **1)** minimise the distance from the predicted positions of the peptide atoms towards any receptor protein atom (interface distances) and **2)** maximise the confidence in the prediction (peptide plDDT [1], Loss 1, equation 1). This approach allows complete freedom in generating a sequence-structure combination for a peptide binder where EvoBind2 deems appropriate. This lack of constraint removes any procedural bias, as preconceived notions about what constitutes a good binding structure, sequence, or site may be incorrect.

Using EvoBind2, we designed both linear and cyclic peptide binders of 8-20 residues targeting the semi-synthetic Ribonuclease (PDB ID 1SSC: https://www.rcsb.org/structure/1ssc). We expressed and purified the protein and measured the binding affinity of the designed linear peptides using surface plasmon resonance (SPR) (Methods), selecting the top sequence for each length (n=13 lengths). We found that 6 out of 13 (46%) of the designed linear peptide binders displayed affinity, with the weakest binder having a dissociation constant (Kd) of 7.93 μM and the strongest a Kd of 19 nM (Figure 1b). Compared to the positive control (Kd = 35 nM, Methods), the strongest binder exhibited almost twice the affinity from a single selected sequence. Additionally, the peptides displayed a diverse set of sequences and predicted binding modalities, highlighting EvoBind2's capability to generate varied solutions.

The loss function used for the optimisation is defined as:

$$Loss\ 1 \ = \ peptide\ plDDT^{-1} \cdot \left(\frac{1}{n} \sum_{j=1}^{n} d_j \right) \qquad (1)$$

Where the peptide plDDT is the average plDDT over the peptide and dj is the shortest distance between all atoms n in the peptide and any atom in the target receptor.

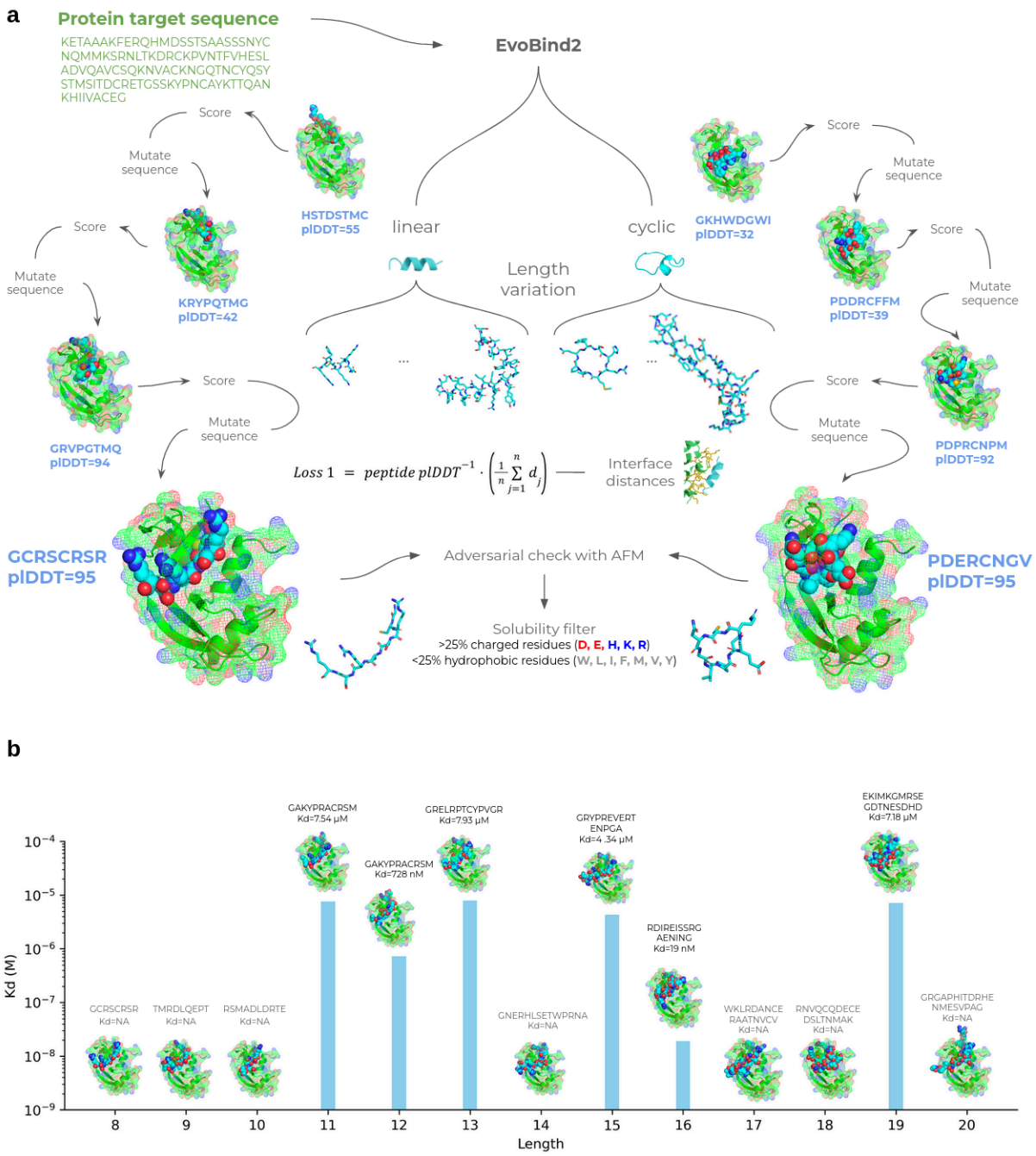

Figure 1. a) Procedure for the design of peptide binders with EvoBind2. The target protein sequence is the only input to the design pipeline. No information about the binding site or peptide length is provided. Both linear and cyclic (head-to-tail) peptides are then designed of various lengths (8-20 amino acids). A random starting peptide sequence is initialised and updated by introducing point mutations (Methods). The predicted structures are shown with the target protein in green and the designed peptide in cyan with coloured oxygen(red) and nitrogen (blue) atoms. Mutations that improve the loss function (equation 1, methods) are selected for. To avoid adversarial designs (see below), AlphaFold-multimer (AFM) is run on the designs and the predictions from both methods are compared (equation 2, Methods). The designs with the best losses from both the design process and the adversarial comparison are selected for experimental validation. **b)** Affinity (Kd) vs sequence length for the top selected linear designs measured by surface plasmon resonance (SPR, methods).

The predicted structures are shown with the target protein in green and the designed peptide in cyan with coloured oxygen(red) and nitrogen (blue) atoms. For 6/13 lengths, μM affinity is measured from a single selected sequence. For the designs with undetectable affinity, no bar is shown.

Adversarial designs

When designing binders to proteins with AI, such as structure prediction technology, it is possible to obtain designs that exhibit high *in silico* confidence (low loss, equation 1) without being true binders. These sequences that “trick” EvoBind2 are deemed adversarial [26]. To avoid adversarial designs, we modified AlphaFold-multimer [2] (AFM) to predict protein-ligand complexes using only MSA information for the receptor protein and a single sequence for the peptide. This additional validation increases the likelihood of success, as it is unlikely that two different structure prediction networks, trained on different data, would agree while both are incorrect.

To analyse the usefulness of this additional adversarial check, we selected sequences from the linear design where EvoBind2 and AFM disagree - specifically, where the AFM loss (equation 2) is high but the EvoBind2 loss (equation 1) is low (Figure 2). We selected the lowest EvoBind2 loss for each length when the AFM loss was >1, obtaining EvoBind2 losses in the range of 0.04-0.1. We synthesised the peptides and evaluated their binding affinity against the target protein using surface plasmon resonance (SPR).

We found that only two of the sequences in the adversarial selection displayed detectable affinity compared to six when both losses (equations 1 and 2) were low. This suggests that adversarial selection helps distinguish true binders from false ones, increasing the success rate threefold. Conversely, some sequences seem adversarial to AFM as 2/13 sequences did display affinity (Figure 2c).

The loss for the adversarial selection is defined as:

$$Loss\,2 \;=\; peptide\,plDDT^{-1} \cdot \left(\frac{1}{m}\sum_{i=1}^{m} d_i + \frac{1}{n}\sum_{j=1}^{n} d_j\right) \cdot \frac{1}{2} \cdot \;\Delta COM \qquad (2)$$

Where the peptide plDDT is the average plDDT over the peptide, di is the shortest distance between all atoms in the receptor target atoms (Cβs within 8 Å between the peptide and the receptor protein as predicted by EvoBind2) m and any atom in the peptide, dj is the shortest distance between all atoms n in the peptide and any atom in the receptor target atoms and ΔCOM is the CA centre of mass distance between the predicted peptides from the design and validation procedures, respectively.

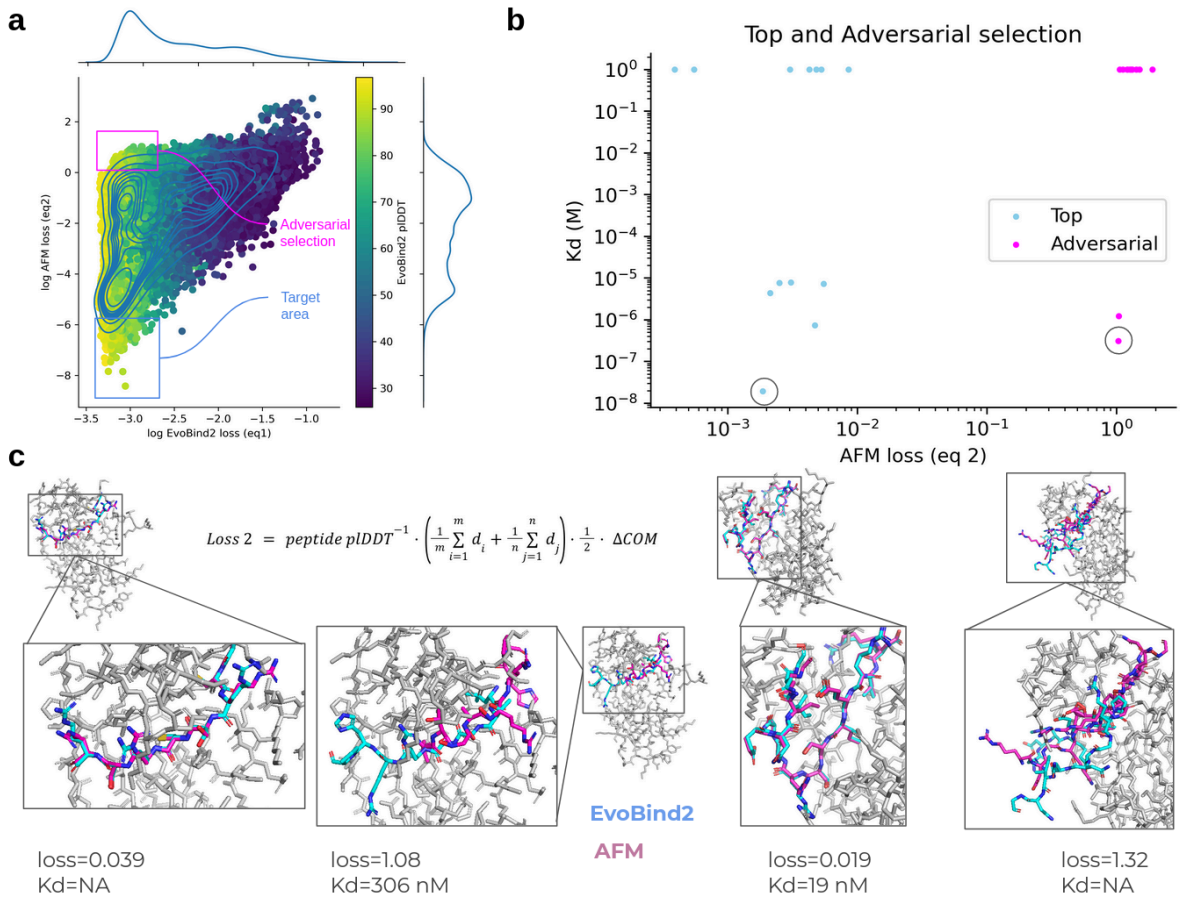

Figure 2. a) EvoBind2 loss (equation 1) vs AlphaFold-multimer (AFM) loss (equation 2) on log scale for all linear design runs (n=64935). The points are coloured by the pIDDT from EvoBind2 and the density is represented with blue lines. The target area represents cases where both losses are low, and the adversarial selection cases where the EvoBind2 loss is low but the AFM loss is high. The discrepancy between the losses in the adversarial area signifies that those sequences may trick EvoBind2. **b)** Affinity (Kd) vs AFM loss (eq2) for both the top selection (in the target area, Figure 2a) and the adversarial selection (AFM loss>1) for the linear design (n=13). The peptides with no detectable affinity are set to 1 M. The peptides with low AFM loss do have lower Kd overall (6/13 with μM affinity) compared to sequences with AFM loss>1 (2/13 with μM affinity). The circled points mark examples shown in c. **c)** Examples of adversarial selections. The predicted structure is shown in stick format with the receptor structures in grey and the predicted peptide structures in cyan/magenta for EvoBind2/AlphaFold-multimer (AFM), respectively. Left; an example adversarial to AFM (length=8) where the Kd is 306 nM although the AFM loss is high and undetectable at a low AFM loss. Right; an example where the adversarial prediction of AFM is accurate (length=16) where the Kd is 19 nM with a low AFM loss and undetectable at a high AFM loss.

Affinity and *in silico* metrics

The loss functions (equations 1 and 2) used in this study produce high-affinity binders, but not all selected binders display measurable affinity. To determine whether we can distinguish between true and false binders or even predict affinity, we analysed various in silico metrics (Figure 3). One of the most important metrics used in the loss functions is the predicted local distance difference test (pLDDT) [27], which evaluates the local confidence of predicted protein structures. Higher pLDDT scores indicate higher confidence, but we observe no meaningful correlation between the dissociation constant (Kd) and EvoBind2 or AFM pLDDT scores (Figure 3b, 3f). This observation holds for the loss functions as well as for distinguishing true binders from false ones.

In addition to generating accurate binders, these peptides must be functional under biological conditions to be useful in biotechnological applications. To ensure this, we applied a simple solubility filter requiring >25% of peptide residues to be charged and <25% to be hydrophobic (Methods). These metrics, however, do not correlate with success in the binder design process (Figure 6c, 6d) and cannot be used to select for affinity.

The length of a peptide can influence its function, as longer peptides offer greater complexity and potential for diverse interactions. However, we find no relationship between peptide length and Kd (Figure 3c), suggesting that the binding residues themselves are what matter most. A comparison of all metrics with the presence or absence of measured μM affinity is displayed in Figure 3i, highlighting that *in silico* metrics fail to effectively distinguish between true and false binders. This underscores the complexity of predicting peptide interactions.

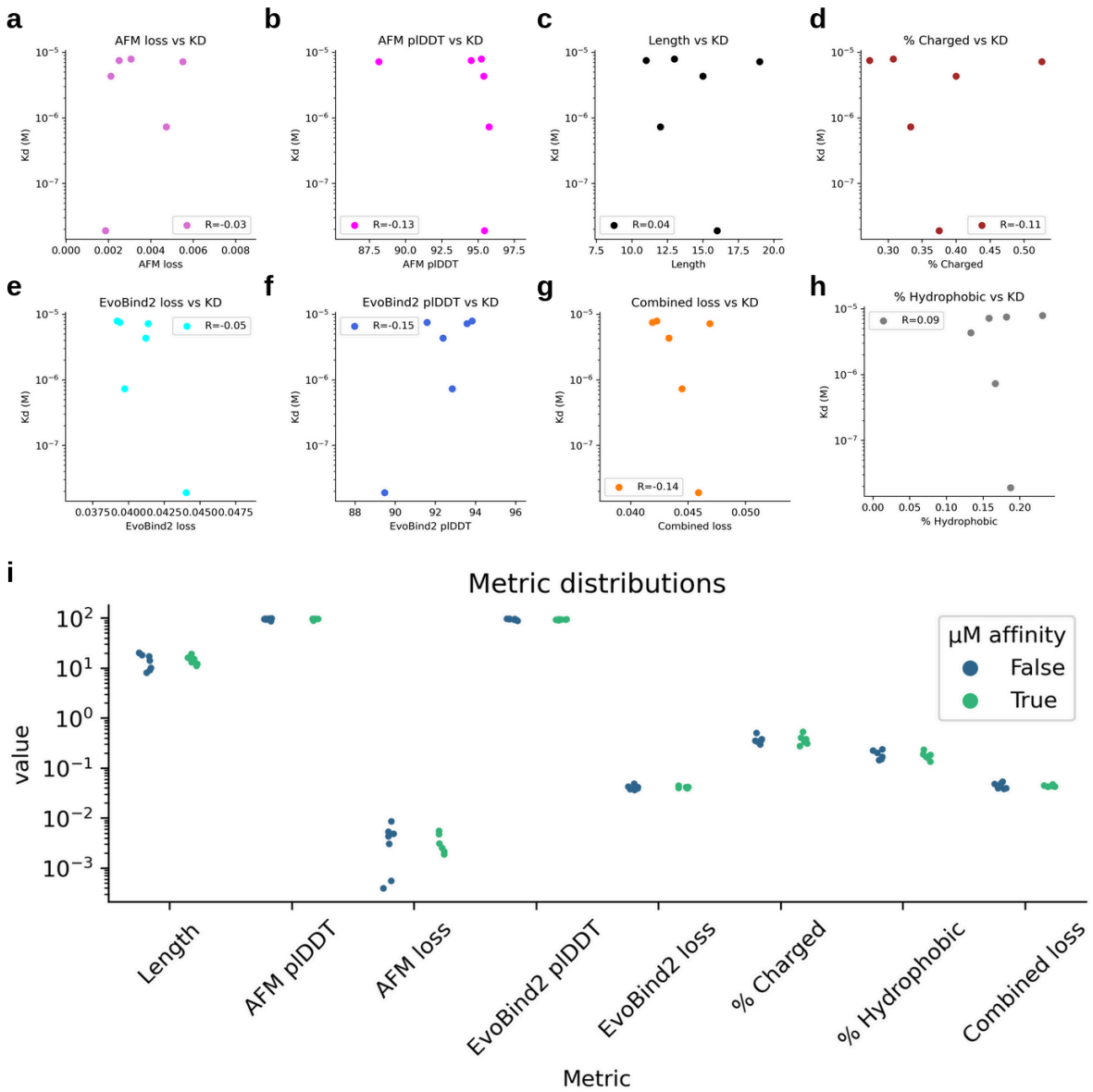

Figure 3. a-h Scatter plots showing the affinities of linear peptides (Kd) vs *in silico* metrics for successful designs from the top selection (n=6). The Spearman correlation coefficient is annotated in each subfigure. None of the metrics display a meaningful correlation with the measured Kd. **a)** AlphaFold-multimer loss (equation 2) **b)** pIDDT from AlphaFold-multimer **c)** the length of designed peptide **d)** the fraction of charged amino acids in the sequence **e)** pIDDT from EvoBind2 **f)** the loss from EvoBind2 (equation 1) **g)** the combined loss from EvoBind2 and AlphaFold-multimer (equations 1 and 2) **h)** the fraction of hydrophobic amino acids in the sequence. **i)** Strip plot of the categories in a-h divided by if μM affinity could be measured/not (n=13). No *in silico* metric separates true from false binders.

Discussion

AI-based peptide binder design offers the potential to rapidly expand the repertoire of proteins that can be targeted for biotechnological applications. The EvoBind2 framework developed here designs peptide binders to a protein target using only its amino acid sequence. This approach does not require knowledge of binding sites, template structures, or binder sizes. We observed binding affinities in the range of Kd = 7.9 μM to 19 nM from a single sequence selection for linear peptides of lengths between 8 and 20 residues, achieving a 46% success rate (6 out of 13 lengths).

Unlike protein structure prediction, which relies on coevolutionary information from multiple sequence alignments (MSAs), novel peptide design lacks this evolutionary context. This absence of coevolutionary information presents a significant challenge when designing new peptides, as the structure must be predicted without these constraints. The finding that successful peptide design is still possible [15] suggests that important protein sequence-structure relationships have been learned independently of coevolution.

One of the key challenges in designing peptide binders is avoiding sequences that have high in silico confidence (high pLDDT) but do not bind to the target protein—these are known as adversarial designs. To address this issue, we adapted AlphaFold-multimer (AFM) [2] to predict protein-ligand complexes using MSA information for the receptor protein and a single sequence for the peptide. This additional validation step, which filters out sequences that appear promising according to EvoBind2 but do not function as true binders, enhances the likelihood of successful binder design threefold. The rationale is that it is unlikely for two distinct neural networks, trained on different data, to agree on an incorrect prediction. Previous design methods have not included adversarial checks, which may explain lower success rates observed with other strategies [6,7,16].

The protein targeted in this study, a semi-synthetic Ribonuclease (PDB ID 1SSC: https://www.rcsb.org/structure/1ssc) has a known peptide binder of length 11. To ensure the expressed protein is functional for affinity measurements, we also synthesised this sequence and measured its affinity as a positive control with a resulting Kd of 35 nM. Interestingly, one of the binders designed here outperforms this control with a Kd of 19 nM, suggesting that EvoBind2 can also be used to achieve improved affinities for targets with known binders.

The predicted structures of the designed peptides show diverse binding modalities (Figure 1). However, it is not clear what specific factors contribute to one peptide binding with higher affinity than another (Figure 3). While sequences with a high AFM loss (equation 2) have a lower chance of binding, for sequences with low AFM loss, it remains impossible to distinguish between nM and μM affinity based on predicted confidence metrics alone.

Cyclic peptides offer advantages over linear peptides in terms of stability and potential membrane traversal [28]. Several drugs on the market today are cyclic peptides [29]. Interestingly, our results indicate that cyclic peptides can be designed to be comparable to linear ones using the same input information. However, this finding still needs to be validated in the laboratory.

In conclusion, our study demonstrates the potential of EvoBind2 for designing novel peptide binders with high affinity using only a target protein sequence. The lack of a requirement for defined binding sites or binder sizes is crucial for targeting novel proteins and suggests a potential for a rapid increase in the number of proteins that can be targeted for various biotechnological applications.

Availability

Data: https://zenodo.org/uploads/11543503
Code: https://github.com/patrickbryant1/EvoBind

Contributions

PB designed the studies. PB developed and QL executed the binder design procedure. QL performed all laboratory analyses and analysed the results in consultation with PB and ENV. PB wrote the first draft of the manuscript which was later improved further by all authors. PB obtained funding.

Acknowledgements

This study was supported by the SciLifeLab & Wallenberg Data Driven Life Science Program (grant: KAW 2020.0239, P.B). Computational resources were obtained from LiU with project id Berzelius-2024-78 (P.B).

The protein purification was facilitated by the Protein Science Facility at Karolinska Institutet, Stockholm, and we would like to thank Dr Henry Ampah-Korsah, Dr Henrik Spåhr, and Dr Tomas Nyman for their assistance.

References

1. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596: 583–589.

2. Evans R, O'Neill M, Pritzel A, Antropova N, Senior A, Green T, et al. Protein complex prediction with AlphaFold-Multimer. bioRxiv. 2022. p. 2021.10.04.463034. doi:10.1101/2021.10.04.463034

3. Bryant P, Pozzati G, Elofsson A. Improved prediction of protein-protein interactions using AlphaFold2. Nat Commun. 2022;13: 1–11.

4. Ingraham JB, Baranov M, Costello Z, Barber KW, Wang W, Ismail A, et al. Illuminating protein space with a programmable generative model. Nature. 2023; 1–9.

5. Watson JL, Juergens D, Bennett NR, Trippe BL, Yim J, Eisenach HE, et al. De novo design of protein structure and function with RFdiffusion. Nature. 2023;620: 1089–1100.

6. Cao L, Coventry B, Goreshnik I, Huang B, Sheffler W, Park JS, et al. Design of protein-binding proteins from the target structure alone. Nature. 2022;605: 551–560.

7. Bennett NR, Coventry B, Goreshnik I, Huang B, Allen A, Vafeados D, et al. Improving de novo protein binder design with deep learning. Nat Commun. 2023;14: 1–9.

8. Yu L, Wang R, Wen T, Liu L, Wang T, Liu S, et al. Peptide Binder with High-Affinity for the SARS-CoV-2 Spike Receptor-Binding Domain. ACS Appl Mater Interfaces. 2022 [cited 18 Jun 2024]. doi:10.1021/acsami.2c03707

9. Quartararo AJ, Gates ZP, Somsen BA, Hartrampf N, Ye X, Shimada A, et al. Ultra-large chemical libraries for the discovery of high-affinity peptide binders. Nat Commun. 2020;11: 1–11.

10. Bryant P, Elofsson A. Peptide binder design with inverse folding and protein structure prediction. Communications Chemistry. 2023;6: 1–10.

11. Tsaban T, Varga JK, Avraham O, Ben-Aharon Z, Khramushin A, Schueler-Furman O. Harnessing protein folding neural networks for peptide–protein docking. Nat Commun. 2022;13: 1–12.

12. Johansson-Åkhe I, Wallner B. Improving peptide-protein docking with AlphaFold-Multimer using forced sampling. Front Bioinform. 2022;2: 959160.

13. Frank C, Khoshouei A, de Stigter Y, Schiewitz D, Feng S, Ovchinnikov S, et al. Efficient and scalable de novo protein design using a relaxed sequence space. bioRxiv. 2023. p. 2023.02.24.529906. doi:10.1101/2023.02.24.529906

14. Dauparas J, Anishchenko I, Bennett N, Bai H, Ragotte RJ, Milles LF, et al. Robust deep learning-based protein sequence design using ProteinMPNN. Science. 2022;378: 49–56.

15. Bryant P, Elofsson A. EvoBind: in silico directed evolution of peptide binders with AlphaFold. bioRxiv. 2022. p. 2022.07.23.501214. doi:10.1101/2022.07.23.501214

16. Wang Y, Tang H, Huang L, Pan L, Yang L, Yang H, et al. Self-play reinforcement learning guides protein engineering. Nature Machine Intelligence. 2023;5: 845–860.

17. Hopf TA, Schärfe CPI, Rodrigues JPGL, Green AG, Kohlbacher O, Sander C, et al. Sequence co-evolution gives 3D contacts and structures of protein complexes. 2014 [cited 18 Jun 2024]. doi:10.7554/eLife.03430

18. Roney JP, Ovchinnikov S. State-of-the-Art Estimation of Protein Model Accuracy Using AlphaFold. Phys Rev Lett. 2022;129: 238101.

19. Rettie SA, Campbell KV, Bera AK, Kang A, Kozlov S, De La Cruz J, et al. Cyclic peptide structure prediction and design using AlphaFold. bioRxiv. 2023. p. 2023.02.25.529956.

doi:10.1101/2023.02.25.529956

20. Kosugi T, Ohue M. Design of Cyclic Peptides Targeting Protein-Protein Interactions Using AlphaFold. Int J Mol Sci. 2023;24. doi:10.3390/ijms241713257

21. Tillberg PW, Chen F, Piatkevich KD, Zhao Y, Yu C-C (jay), English BP, et al. Protein-retention expansion microscopy of cells and tissues labeled using standard fluorescent proteins and antibodies. Nat Biotechnol. 2016;34: 987–992.

22. Fani M, Maecke HR, Okarvi SM. Radiolabeled Peptides: Valuable Tools for the Detection and Treatment of Cancer. Theranostics. 2012;2: 481.

23. Békés M, Langley DR, Crews CM. PROTAC targeted protein degraders: the past is prologue. Nat Rev Drug Discov. 2022;21: 181–200.

24. Donnelly D. The structure and function of the glucagon-like peptide-1 receptor and its ligands. Br J Pharmacol. 2012;166: 27.

25. Portha B, Tourrel-Cuzin C, Movassat J. Activation of the GLP-1 Receptor Signalling Pathway: A Relevant Strategy to Repair a Deficient Beta-Cell Mass. Exp Diabetes Res. 2011;2011. doi:10.1155/2011/376509

26. Yuan X, He P, Zhu Q, Li X. Adversarial Examples: Attacks and Defenses for Deep Learning. 2017. Available: http://arxiv.org/abs/1712.07107

27. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics. 2013;29: 2722.

28. Ji X, Nielsen AL, Heinis C. Cyclic Peptides for Drug Development. Angew Chem Int Ed. 2024;63: e202308251.

29. Joo SH. Cyclic Peptides as Therapeutic Agents and Biochemical Tools. Biomol Ther . 2012;20: 19.

30. Mirdita M, von den Driesch L, Galiez C, Martin MJ, Söding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. 2017;45: D170–D176.

31. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics. 2019;20: 473.

32. Hsu C, Verkuil R, Liu J, Lin Z, Hie B, Sercu T, et al. Learning inverse folding from millions of predicted structures. bioRxiv. 2022. p. 2022.04.10.487779. doi:10.1101/2022.04.10.487779

33. Goudy OJ, Nallathambi A, Kinjo T, Randolph NZ, Kuhlman B. In silico evolution of autoinhibitory domains for a PD-L1 antagonist using deep learning models. Proceedings of the National Academy of Sciences. 2023;120: e2307371120.

34. Simulated single-cycle kinetics improves the design of surface plasmon resonance assays. Talanta. 2013;114: 211–216.

Methods

Design protocol

EvoBind2

For designing peptide binders, we use a modified version of EvoBind [15] (Figure 4) that does not require a defined target area. This is the first method developed that can generate both sequence and structure from a single protein target sequence (to our knowledge). We design over a relaxed sequence-structure space that is searched by mutating one residue randomly at a time to minimise the loss function:

$$Loss\,1 \;=\; peptide\,plDDT^{-1} \cdot \left(\frac{1}{n}\sum_{j=1}^{n} d_j\right) \qquad (1)$$

Where the peptide plDDT is the average plDDT over the peptide and dj is the shortest distance between all atoms n in the peptide and any atom in the target receptor.

The receptor is represented by a multiple sequence alignment that is constructed from searching uniclust30_2018_08[30] with HHblits[31] version 3.1.0:

```
hhblits -E 0.001 -all -oa3m -n 2
```

The peptide is represented by a single sequence (initialised with a Gumbel distribution over all standard 20 amino acids). The single chain AlphaFold2 folding pipeline [1] with model_1, one ensemble and 8 recycles was used. The optimisation is run for 1000 iterations and repeated five times from random starting points (sequences).

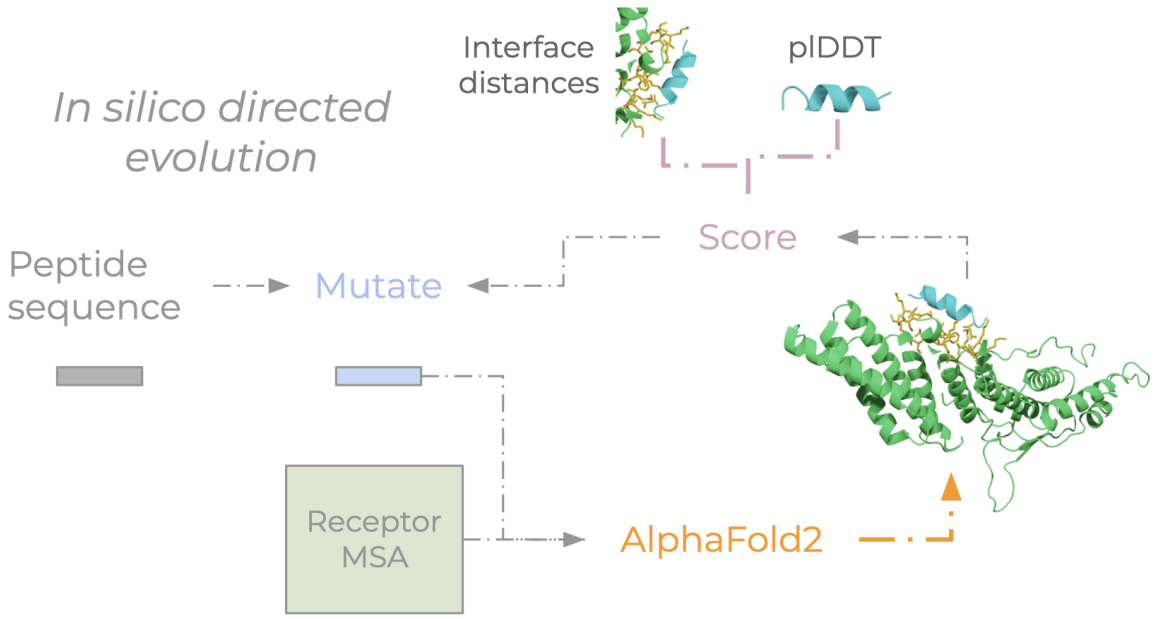

Figure 4. *In silico* directed evolution procedure towards an unspecified binding site. By not specifying a binding site, EvoBind2 is free to generate a structure-sequence combination to

any part of the target protein as long as 1. the confidence (plDDT) of the predicted peptide is high and 2. the predicted peptide atom positions are close to the receptor surface (interface distances).

Other binder sequence design methods

MC and MCTS with AlphaFold2 modifications
There are other methods[16] for designing binders using the loss function from the previously developed EvoBind [15], but no binder design method is entirely untargeted (i.e. a target site is required) and purely sequence-based (i.e. a target scaffold is needed).

Design with inverse folding
Inverse folding has been applied to the problem of binder design [14,32]**.** However, these methods require scaffolds that are not available for novel targets and have only been applied to design larger protein structures (not peptides) [6,7,33]. Other methods have recently been developed to design peptide structures, but how to generate sequences for these is not clear. In addition, sequences for only 4-7% of structural seeds can be redesigned into sequences that structure prediction networks understand using inverse folding [10]. Relaxing the entire sequence-structure space will allow for more possibilities, increasing the probability that a structure prediction network can understand a designed sequence-structure combination.

In silico evaluation with AlphaFold-multimer to avoid adversarial designs

In the EvoBind2 protocol developed here, no target site is used as this information is not available for novel target proteins. To ensure the design outcomes from the untargeted protocol are viable we use AlphaFold-multimer (AFM, Figure 5) as an additional check to avoid adversarial designs. This allows us to utilise the original EvoBind loss function (equation 12) which relies on a specified target site and a known peptide binder. In this case, we compare the loss between the two predictions using the EvoBind2 designs as “pseudo-native” structures.

Previous benchmarking has found that the best option for protein-peptide complex prediction is AFM v2.1.0 (params_model_1_multimer_v2) with 3 recycles and dropout everywhere but in the structural model (highest correlation with the ranking confidence from AFM+median DockQ) [12]. To account for that some examples improve with more recycles, we set the number of recycles to 20 and include ‘early stopping’ (no more recycles are included if the confidence score is not improved).

Calculating the loss using EvoBind2 vs. AFM predictions circumvents the need to know the native structure or CA centre of mass to use the EvoBind loss function. The EvoBind loss function, which has demonstrated accuracy both *in silico* and *in vitro [15,16]*, can thus be used to evaluate the agreement between different neural networks (AF2 and AFM).

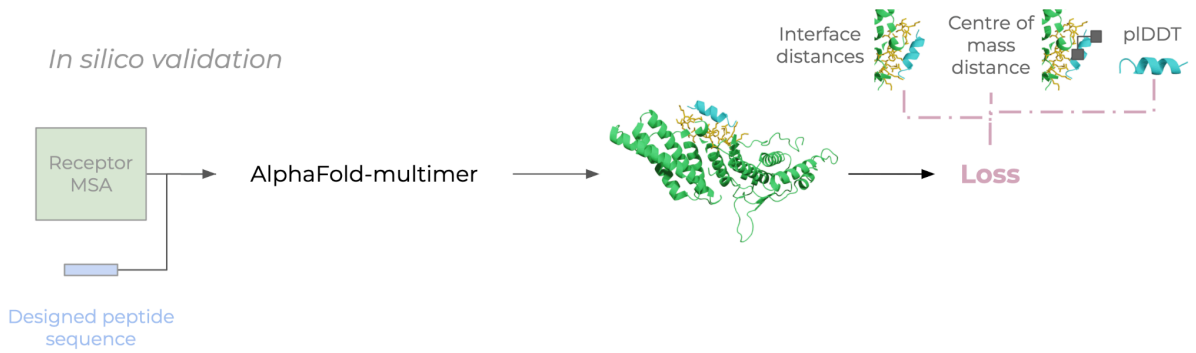

Figure 5. AFM validation. Given a receptor MSA and a single sequence for the designed peptide, AlphaFold-multimer is used to predict a protein-peptide complex. Based on the similarity of the predicted structure to that predicted with EvoBind2 and the confidence (pIDDT) of the prediction a loss is calculated (equation 2).

$$Loss\,2 \;=\; peptide\,plDDT^{-1} \cdot \left(\frac{1}{m} \sum_{i=1}^{m} d_i + \frac{1}{n} \sum_{j=1}^{n} d_j \right) \cdot \frac{1}{2} \cdot \;\Delta COM \qquad (2)$$

Where the peptide pIDDT is the average pIDDT from AFM over the peptide, di is the shortest distance between all atoms in the receptor target atoms (Cβs within 8 Å between the peptide and the receptor protein as predicted by EvoBind2) m and any atom in the peptide, dj is the shortest distance between all atoms n in the peptide and any atom in the receptor target atoms and ΔCOM is the CA centre of mass distance between the predicted peptides from the design and validation procedures, respectively.

Binder design selection

We select the designs with the best combined loss calculated as the sum of equations 1 and 2. For the adversarial selection, we select designs with an AFM loss >1 according to equation 2 and the lowest possible loss according to equation 1, obtaining losses in the range 0.04-0.1 with equation 1 for the linear design task.

Cyclic offset

To design cyclic binders, we implement a cyclic offset [19] informing the structure prediction network to connect the peptide amino acids in a continuous cycle. This feature is called the relative positional encoding in the AlphaFold2 [1] and AlphaFold-multimer [2] networks.

Positive control

To ensure that the expressed protein (semi-synthetic Ribonuclease, PDB ID 1SSC: https://www.rcsb.org/structure/1ssc) is in a structural state where it can interact with designed binders we analysed the affinity of the known peptide binder with sequence: PYVPVHFDASV. Using single-cycle kinetics with surface plasmon resonance (SPR) we obtain a Kd of 35 nM (Supplementary Figure 6). At the same time, the adversarial designs

report no response suggesting that the protein is in a selective structural state where only specific peptide sequences will bind with high affinity.

Length and initialisation

Figure 6 shows the best EvoBind2 loss (equation 1) vs the binder length for linear and cyclic design across five different runs (initialisations). For the linear design, the loss increases with the length, but no such relationship is apparent for the cyclic design. The losses vary substantially between runs (Supplementary Figures 1 and 2), but the results show that given five tries comparable results can be obtained in terms of affinity for linear designs. The finding that longer lengths display higher best losses and larger variability may be related to the higher number of possibilities accompanied by a longer sequence length making it less likely to end up in a favourable design.

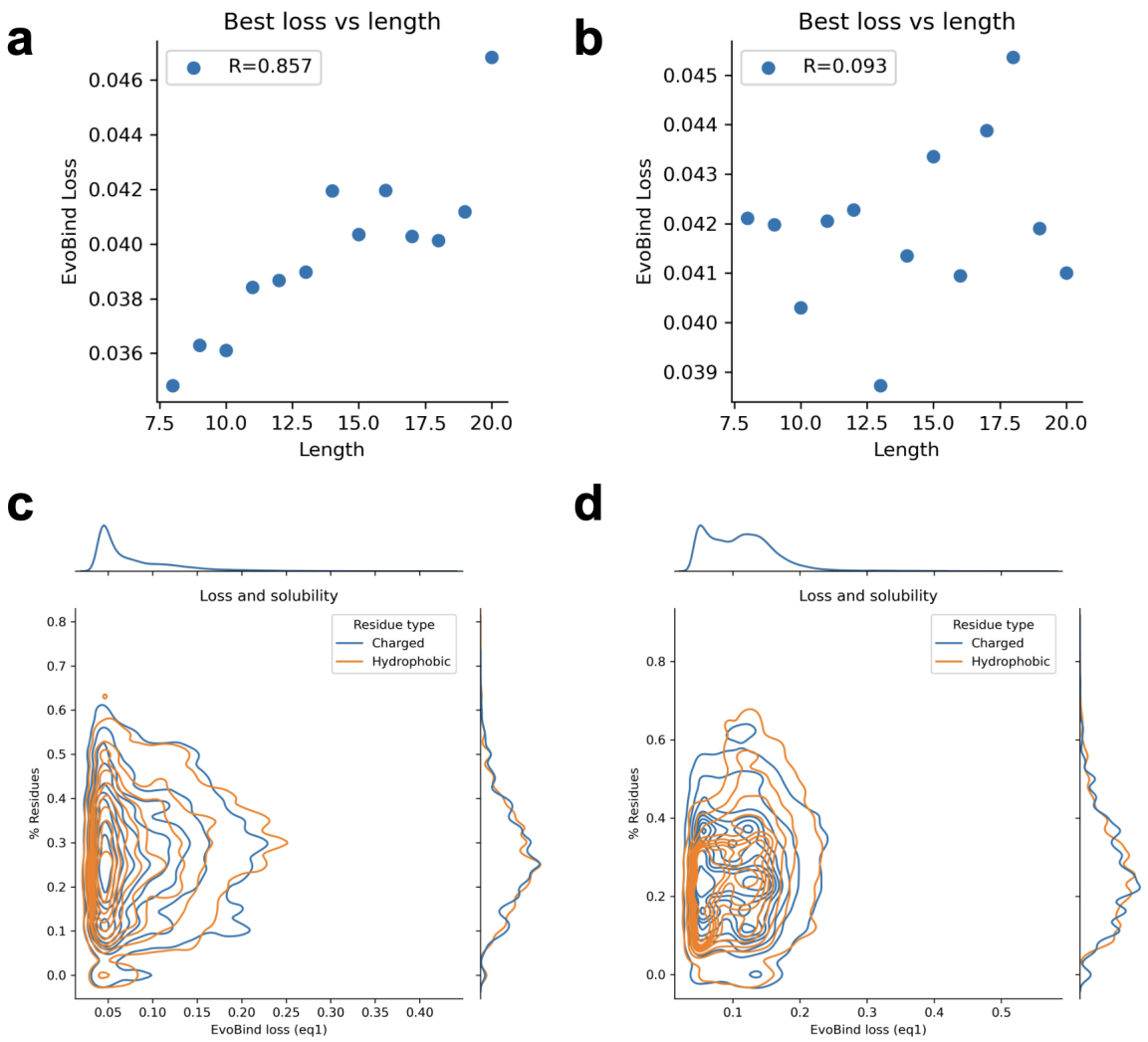

Figure 6. Best EvoBind2 loss (equation 1) vs length for linear **(a)** and cyclic **(b)** design. For the linear design, the loss increases with the length (Spearman R=0.857), but for the cyclic, such a relationship is not as apparent (Spearman R=0.093). **c, d** Fraction of charged (D, K, R, H and E) and hydrophobic residues (W, L, I, F, M, V, Y) vs EvoBind2 loss (equation 1) for linear **(a)** and cyclic **(b)** design. Different fractions of charged/hydrophobic residues are

observed for a variety of losses suggesting that the solubility is not important for EvoBind2 when generating designs.

Solubility

To ensure water solubility we select designed binders that have an average plDDT>90, >25% charged residues (D, K, R, H and E) and <25% hydrophobic residues (W, L, I, F, M, V, Y). Figures 6c and d show different fractions of charged/hydrophobic residues are observed for a variety of losses suggesting that solubility is not important for EvoBind2 when generating designs. All synthesised linear peptides were soluble in the SPR running buffer (see Affinity measurement with SPR).

Protein expression and purification

The psfRNAseA-c001 construct was transformed into E. coli BL21 (DE3) T1R cells and cultivated in Terrific Broth (TB) medium supplemented with glycerol, ampicillin, and chloramphenicol. The cultures were grown at 37°C, and protein expression was induced with IPTG at an OD of approximately 3, after which the temperature was reduced to 18°C for overnight expression. The cells were harvested, resuspended in an IMAC lysis buffer, and disrupted by sonication. The soluble fractions were purified using IMAC and size exclusion chromatography, and the target protein was pooled, concentrated, and stored at -80°C. The psfRNAseA protein eluted together with some contaminants or possibly proteolytic fragments, and the full-length psfRNAseA protein with the MBP-tag intact was identified by SDS-PAGE analysis.

Peptide synthesis

Lyophilized powders of affinity peptides by EvoBind2's design were synthesised and purified by JPT Peptide Technologies GmbH (linear ones), and GenScript (cyclic ones). The quality of the peptide (purity above 90%) was verified by high-performance liquid chromatography and mass spectroscopy. Peptides were resuspended in HBS-P+ Buffer (0.01 M HEPES, 0.15 M NaCl and 0.05% v/v Surfactant P20) to achieve a stock concentration of 10 mM.

Affinity measurement with SPR

To directly measure the binding affinity of designed peptides toward 1SSC we used surface plasmon resonance (SPR). An experimental method that measures real-time binding interactions between biomolecules by detecting changes in refractive index near a sensor surface. All interaction analysis was performed at 25 °C in a running buffer of HBS-P+ (10 mM HEPES, pH 7.4, 150 mM NaCl, 0.05% Tween-20) for Biacore 8K. MBP-8xHis-1SCC was immobilised to about 12000 RU using standard amine coupling reagents 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC) and N-hydroxysuccinimide (NHS) onto a Series S Sensor Chip CM5. The designed peptides were captured onto the prepared surface to generate an interaction with a maximum response (R_{max}) from 20 RU to 300 RU.

We applied the single-cycle kinetics method, in which each cycle consists of a series of association and dissociation phases. A series of peptide concentrations were prepared at

dilutions of 3.125 μM, 6.25 μM, 12.5 μM, 25 μM, 50 μM. We injected each peptide concentration sequentially without a regeneration step in between, and after the end of each dissociation phase, we injected the subsequent higher concentration of peptide. The final dissociation phase was extended to ensure complete dissociation of the peptide from the protein. The sensorgrams obtained from the single-cycle kinetics experiment were analysed using the Biacore 8K Evaluation software available from Cytiva [34].

To calculate the binding affinity the data were fit to a 1:1 binding model to obtain the association rate constant (k_a), dissociation rate constant (k_d), and equilibrium dissociation constant (Kd). See Supplementary Figures 4-6 for all sensorgrams.

$$A + B \leftrightarrow AB \tag{3}$$

$$Kd = \frac{[A]\cdot[B]}{[AB]} = \frac{kd}{ka} \tag{4}$$

Where A is the ligand that was immobilised on the sensor chip, and B is the analyte (peptide). The equilibrium dissociation constant Kd is calculated as the ratio of kd to ka, where ka represents the association rate constant for complex formation (A+B →AB, $M^{-1}s^{-1}$), and kd the dissociation rate constant (AB →A+B, s^{-1}).