RareFold: Structure prediction and design of proteins with noncanonical amino acids

Qiuzhen Li[1,2], Diandra Daumiller[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

Protein structure prediction and design have traditionally been limited to the 20 canonical amino acids. Expanding this space to include noncanonical amino acids (NCAAs) offers new opportunities for probing novel interactions and engineering proteins with enhanced or entirely new functions. Some NCAAs also offer practical advantages, such as increased proteolytic stability and reduced immunogenicity, as they are rarely encountered by the human immune system. Here, we present **RareFold**, a deep learning model capable of accurate structure prediction for proteins containing both the 20 canonical amino acids and an additional 29 NCAAs. By treating each amino acid as a distinct token, RareFold learns residue-specific atomic interaction patterns, enabling precise modelling of chemically diverse sequences. This tokenised representation also supports sequence-structure co-optimisation, allowing efficient inverse design. We leverage this capability in **EvoBindRare**, a design framework for generating linear and cyclic peptide binders that incorporate NCAAs. Applying EvoBindRare, we design binders targeting a ribonuclease and experimentally validate these, obtaining μM affinity in both the linear and cyclic cases. RareFold thus enables binder design with an expanded chemical vocabulary, opening the door to next-generation peptide therapeutics with improved stability, specificity, and immune evasion. RareFold is available at: https://github.com/patrickbryant1/RareFold

Introduction

Structure prediction with AlphaFold2 (AF2) has revolutionised structural biology by achieving near-experimental accuracy [1], enabling new possibilities in rational protein engineering. Beyond predicting natural protein structures, structure prediction now plays a central role in *inverse design*, where new sequences are generated to adopt desired structures and functions [2]. Most successful design pipelines rely on AF2 or similar models to evaluate whether candidate sequences fold as intended [3–5], making forward prediction a key step in modern protein design.

However, these approaches are limited by the chemistry of the 20 canonical amino acids shared across life, restricting the range of possible interactions, stability, and functionality. While cyclic peptide design can improve stability [6–8], expanding to noncanonical amino acids (NCAAs) offers far greater potential. Over 300 amino acid types are found in the PDB, and more than 500 occur in nature [9], with 140 NCAAs known to be naturally incorporated into proteins [10]. Despite this, NCAAs remain largely unexplored in design. They offer unique benefits, such as protease resistance and potential immune evasion, since many are not recognised as self by the human immune system, an advantage in therapeutic settings.

There are two main challenges in predicting structures with NCAAs. First, standard multiple sequence alignments (MSAs) cannot distinguish these residues, as most appear as “X” in sequence databases, masking their identity. Nonetheless, the presence of “X” still conveys useful information about variability and alignment context. Second, while recent frameworks like AlphaFold3 (AF3) [11] can model modified residues atom by atom, specifying each modified atom is impractical for design workflows due to the vast search space and changing input/output dimensions. In addition, it is not known how accurate AF3 is in predicting different types of modified amino acids, especially on unseen proteins, as it has been trained on almost all proteins, making evaluation difficult.

To overcome these limitations, we introduce RareFold, a structure prediction network trained on the standard 20 amino acids plus 29 additional NCAAs. By representing each residue, canonical or modified, as a unique token, RareFold learns coevolutionary and structural relationships in a compact, tractable manner. This representation enables both accurate structure prediction and joint sequence-structure optimisation. We further demonstrate its power by inverting the network into a design framework, EvoBindRare, which successfully designs linear and cyclic peptide binders incorporating NCAAs.

Results

RareFold

The genetic code comprises 64 codons, but due to redundancy, it typically encodes only 20 standard amino acids in proteins. In contrast, nature employs a much wider chemical repertoire: over 500 amino acid types have been identified [9], and 140 of these noncanonical amino acids (NCAAs) are known to be incorporated into proteins [10], offering an expanded landscape for molecular interactions and functionality. The Protein Data Bank

(PDB) contains 331 distinct amino acid types in single-chain proteins, underscoring the untapped potential of NCAAs in protein engineering.

To enable accurate structure prediction across this chemically diverse space, we extended the EvoFormer architecture [1] to include 29 of the most common NCAAs found in the PDB. Unlike previous efforts [11,12], which operate at the atomic level, our approach treats each amino acid, canonical or noncanonical, as a unique token. This design choice improves computational efficiency and allows the model to learn coevolutionary relationships that influence how NCAAs are incorporated into protein structures (Figure 1).

We evaluate our model, RareFold, on a held-out set of structures with less than 20% sequence identity to the training data and compare its predictions to those of AlphaFold3 (AF3). RareFold accurately predicts both the overall protein fold and the positions of diverse noncanonical amino acids (NCAAs) (Figure 1b). In Figure 1c, we report the side chain (sc) RMSD for 24 different NCAA types across 743 validation structures (AF3 not applicable; see *Methods*), using varying numbers of recycles. The RMSD values are consistently low and comparable to those observed for standard amino acids (see *Comparison with proteinogenic amino acids*), indicating that RareFold generalises well across chemically diverse residue types. Notably, increasing the number of recycles does not lead to substantial improvements, suggesting that accurate predictions can be achieved efficiently.

Figures 1d-f compare RareFold and AF3 on a test set of 174 structures (171 for AF3 due to memory limitations), covering 13 NCAAs remaining after excluding AF3's training data. RareFold performs comparably to AF3, with side chain RMSDs differing by only 1-2 Å for most amino acids. RareFold outperforms AF3 on SEP, SAH, and PCA, while AF3 shows slightly better accuracy for others. Two cases stand out: MSE, which is frequently mispredicted by AF3 due to intra-residue clashes (Figure 1f), while RareFold produces physically realistic structures without such clashes (Figure 2a); and SAH, which is often not covalently linked in the PDB, making its prediction more akin to small molecule docking [13]. AF3 treats SAH as covalently bound, potentially reducing accuracy, while RareFold remains agnostic to covalent linkage.

Figure 1e shows the Cα lDDT scores grouped by the type of NCAA present in each protein. AF3 outperforms RareFold on this metric, indicating stronger global structure prediction and likely explaining its slightly lower side chain RMSD in some cases. This is expected, as AF3 has been trained on a larger number of structures, including similar ones to those in the test set via template-based self-distillation, and uses longer input crops (768 residues vs 254 for RareFold), which may also contribute to the lower lDDT scores observed in RareFold (Figures 2b, c).

Despite this, RareFold still achieves high structural accuracy overall. On the validation set, it reaches a median Cα lDDT of 0.87 (Figure 3). Additionally, RareFold is more computationally efficient due to its token-level representation. For example, two test set structures could not be processed by AF3 even with 80 GB of RAM, while RareFold predicted all structures using only 40 GB. This efficiency is critical in protein design settings, where large-scale searches and evaluations of candidate sequences are required.

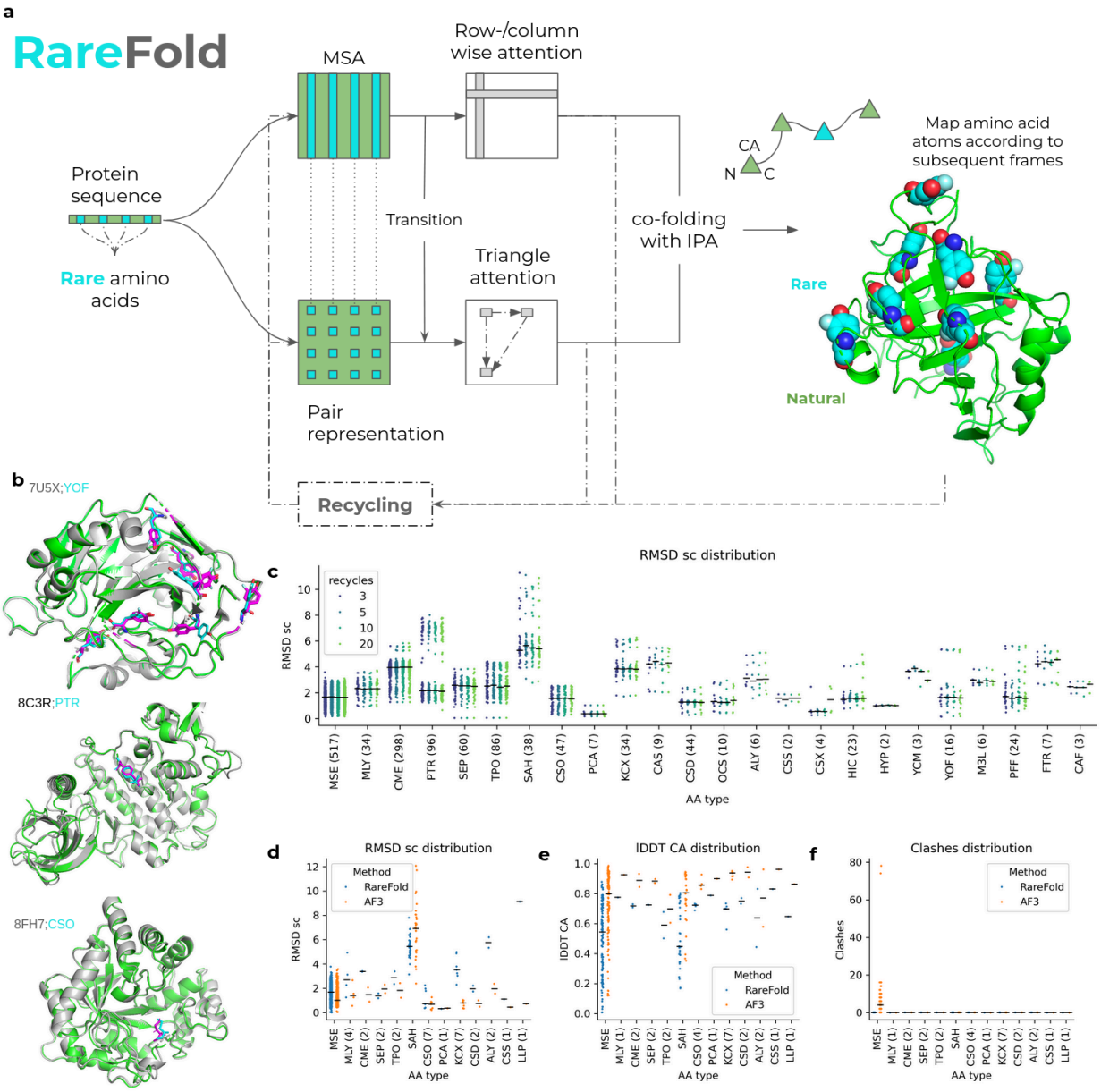


Figure 1. RareFold network architecture and evaluation.
a) RareFold is a structure prediction network based on the EvoFormer architecture from AlphaFold2, trained to model both the 20 canonical amino acids and 29 additional noncanonical amino acids (NCAAs). Each amino acid, whether canonical or modified, is treated as a distinct token, enabling the model to learn residue-specific structural and coevolutionary patterns. This tokenised representation is essential for supporting iterative sequence optimisation and forward prediction in design tasks. For each input sequence, a multiple sequence alignment (MSA) is generated without templates; positions corresponding to NCAAs are marked as "X" due to missing homology information, shown in cyan. The MSA is processed using row and column attention, while pairwise residue interactions are captured using triangle attention mechanisms. Final residue representations are mapped to atomic coordinate frames, with each amino acid assigned a specific set of local frames, resulting in 49 distinct frame types in total. This architecture enables accurate structural modelling across chemically diverse protein sequences. **b)** Examples of accurate RareFold predictions across diverse folds and noncanonical amino acids (NCAAs). Superpositions of predicted (green cartoon) and native (grey cartoon) single-chain protein structures

containing the NCAAs YOF (PDBID 7U5X), PTR (PDBID 8C3R), and CSO (PDBID 8FH7). The NCAA side chains are shown as sticks: RareFold predictions in cyan, and native PDB structures in magenta. These examples illustrate RareFold's ability to accurately model structurally diverse proteins incorporating chemically distinct NCAAs. **c)** Distribution of aligned side chain RMSD on the validation set divided by NCAA type and number of recycles used at inference (3, 5, 10 and 20). The number of NCAAs evaluated is provided in the parentheses next to each AA type. **d)** Distribution of aligned test set side chain RMSD per NCAA type for RareFold and AlphaFold3 (AF3). The number of NCAAs evaluated is provided in the parentheses next to each AA type. There are two discrepancies since AF3 was OOM for some structures; MSE 118 vs 120 and SAH 31 vs 32, for AF3 and RareFold, respectively. **e)** Test set Cα lDDT per protein divided by the NCAA type the protein contains for RareFold and AlphaFold3 (AF3). The number of NCAAs evaluated is provided in the parentheses next to each AA type. There are two discrepancies since AF3 was OOM for some structures; MSE 118 vs 120 and SAH 31 vs 32, for AF3 and RareFold, respectively. **f)** Clash distribution within each NCAA summed per target on the test set. Clashes are defined as atoms within each NCAA closer than 1 Å apart. Only AF3 exhibits clashes for MSE. This is analysed in more detail in Figure 2. The number of NCAAs evaluated is provided in the parentheses next to each AA type. There are two discrepancies since AF3 was OOM for some structures; MSE 118 vs 120 and SAH 31 vs 32, for AF3 and RareFold, respectively.

Artefacts from diffusion for MSE in AF3

One limitation of AF3 is the presence of unnatural atomic representations introduced by its diffusion module, leading to overlapping atoms, particularly in selenomethionine (MSE), the most common NCAA in the PDB. While AF3 appears to perform well on MSE based on RMSD metrics, a closer examination reveals systematic structural artefacts. MSE residues frequently exhibit steric clashes (atoms closer than 1 Å apart) due to improper placement of the selenium atom, leading to artificially low RMSD values that do not reflect realistic atomic configurations (Figure 2a). The reliance on RMSD as a sole metric for evaluating the performance of NCAAs is problematic. In the AF3 paper [11], a threshold of 2 Å is used to assess prediction accuracy; however, this does not capture the overall structural quality of NCAAs as the median scRMSD is close to 1 for MSE (Figure 1d).

Similar issues were observed with secondary structure predictions, where AF3 tends to overproduce ordered regions and underrepresent disordered loops due to biases in its diffusion-based training process [11]. To mitigate such problems, AF3 incorporates fine-tuning on AF2 predictions. However, this approach is unlikely to be generalizable to NCAAs due to the limited availability of training data. As the diffusion-based model primarily minimises RMSD, it is possible to find suboptimal solutions. This issue is inherent to diffusion-based approaches, as the stochastic nature of denoising makes it difficult to direct the model toward uncommon data regimes such as NCAAs.

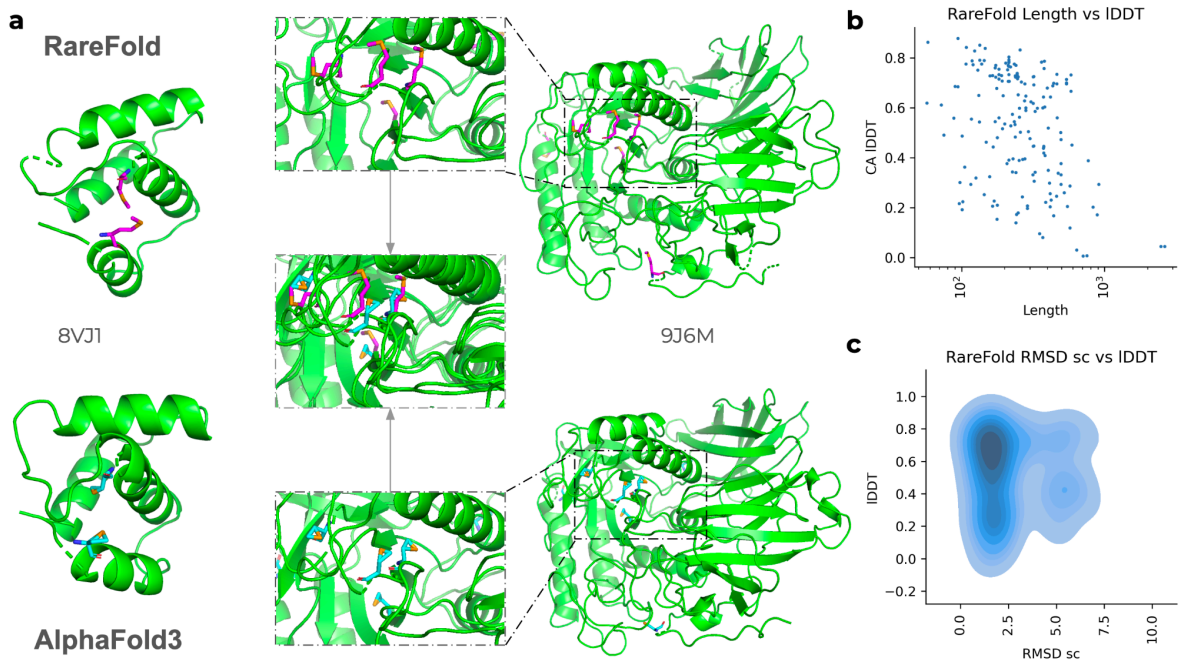


Figure 2. Limitations of AlphaFold3 and RareFold in Predicting Modified Amino Acids **a)** Comparison of MSE side chain predictions for two examples (PDB IDs 8VJ1 and 9J6M). AF3 frequently produces unnatural atomic configurations for MSE (cyan sticks for AF3, magenta for RareFold) due to limitations in its diffusion module, even though the overall structural accuracy (green cartoon) is high. These include steric clashes (atoms <1 Å apart) caused by misplacement of the selenium atom. Although AF3 reports good performance on MSE using RMSD, structural inspection reveals unfolded side chains and systematic artefacts not captured by this metric. The over-reliance on RMSD, including a 2 Å success threshold, fails to reflect true accuracy; median side chain RMSD for MSE is close to 1 Å (Figure 1d). **b)** Length vs Cα lDDT on the test set (n=174). Cα-lDDT scores plotted against sequence length for each test protein. RareFold shows a mild decline in global accuracy with increasing sequence length, which may be attributed to the use of shorter input crops (254 residues) compared to AF3 (768 residues). Despite this, RareFold maintains high overall accuracy across a broad range of sequence lengths, demonstrating robust generalisation even without extensive self-distillation or template-based training data. **c)** Median side chain RMSD (scRMSD) for NCAAs versus average Cα-lDDT over the full protein on the test set (n=174). The side chain accuracy for NCAAs is higher when the global structure is accurately predicted. This highlights the influence of overall structural accuracy on NCAA modelling performance.

Amino acid accuracy and confidence

NCAAs are harder to predict simply because the network does not have as much data to learn from compared to the more common proteinogenic ones. However, even if an NCAA can't be predicted with high accuracy, the prediction may still be useful. A key aspect of structure prediction, and design, is the ability to distinguish accurate from inaccurate predictions. The predicted lDDT score [14] (plDDT) provides a confidence measure for individual residues and the overall structure. As shown in Figure 3a, the plDDT (all-atom,

averaged per residue) correlates with the NCAA RMSD after global structural alignment and NCAAs with plDDT 0-30 have median RMSD above 10 compared to around 2 for those with plDDT 90-100.

While not all NCAAs are predicted with high accuracy, we can reliably identify those that are, along with the overall structural accuracy, using plDDT (Figure 3b, Spearman R = 0.87, median Cα lDDT=87). Notably, the vast majority of predictions achieve lDDT scores above 80, indicating high structural quality. Figure 3c highlights examples of predictions with high and low lDDT. For many NCAA types, the overall structure and atomic side chain details are both correctly predicted and identified by RareFold. At lDDT scores above 0.8, the predicted structures are almost identical to the native ones.

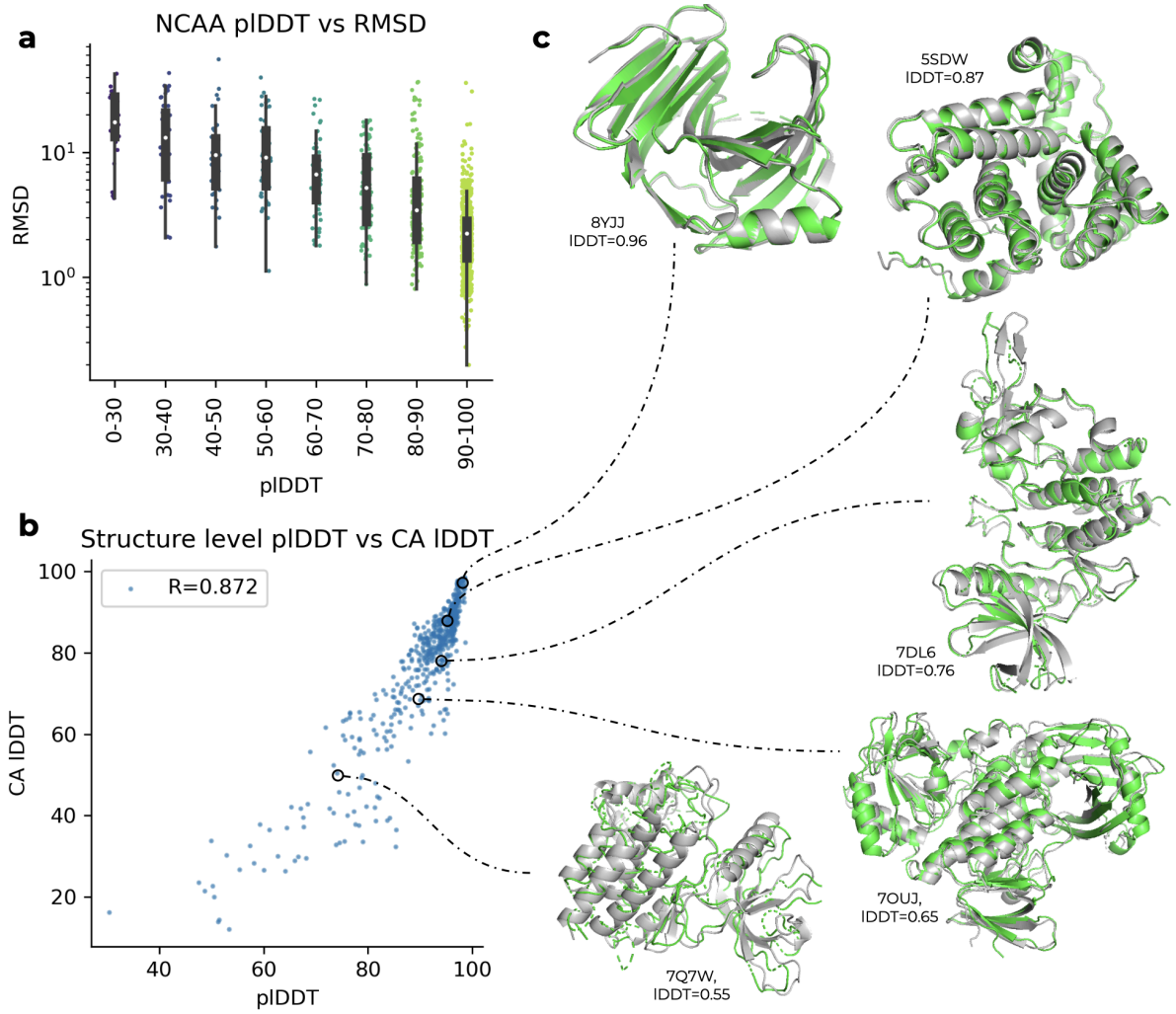


Figure 3. Relationship between confidence metrics and structural accuracy on the validation set. a) Relationship between predicted lDDT (plDDT) and RMSD after global alignment for NCAAs in the validation set. High plDDT scores generally correspond to lower RMSD values, indicating that RareFold can distinguish accurate predictions. Most residues fall in the 90-100 plDDT range (n = 896), supporting reliable confidence estimation. The number of amino acids per predicted lDDT bin are as follows: 896 in 90-100, 186 in 80-90, 78 in 70-80, 53 in 60-70, 55 in 50-60, 36 in 40-50, 45 in 30-40, and 27 in 0-30.**b)** Predicted lDDT (plDDT) vs actual Cα lDDT (scaled by 100) across full protein structures in the

validation set (n = 731). The strong correlation (Spearman R = 0.87) indicates that RareFold can reliably identify high-accuracy predictions. Most structures achieve Cα lDDT > 80, reflecting high global structural quality. **c)** Examples at different levels of accuracy (CA lDDT). At lDDT scores above 0.8, the structures are almost identical.

Comparison with proteinogenic amino acids

Figure 4a shows the side chain (sc) RMSD distribution on the validation set for all AAs present. As can be seen, the RMSD for both proteinogenic and NCAA varies, although the values are similar. Take e.g. ARG and compare it to CME which is of similar size. In both of these AAs, the sc RMSD varies between 1-5 Å. This showcases the structural uncertainty overall, which is a part of protein structure prediction as it should not be possible to always obtain predictions with RMSD <1 Å given that the resolution of the structures used for training and evaluation do not have such high resolution.

Many of the NCAAs do display higher median values than the proteinogenic ones, but this can be attributed to the big difference in their occurrence (e.g. CME occurs 298 times and ARG 10627). The median global average Cα lDDT scores are also consistently higher for the proteinogenic AAs (Figure 4b). Again, this is related to their occurrence as the proteinogenic AAs are present in all structures, resulting in the overall median of 0.87, while the NCAAs are not. Still, many of the structures with NCAAs are accurately predicted and contain some examples with Cα lDDT >0.9 (all except MLY, SAH, CSS, CSX, HYP, YCM and M3L).

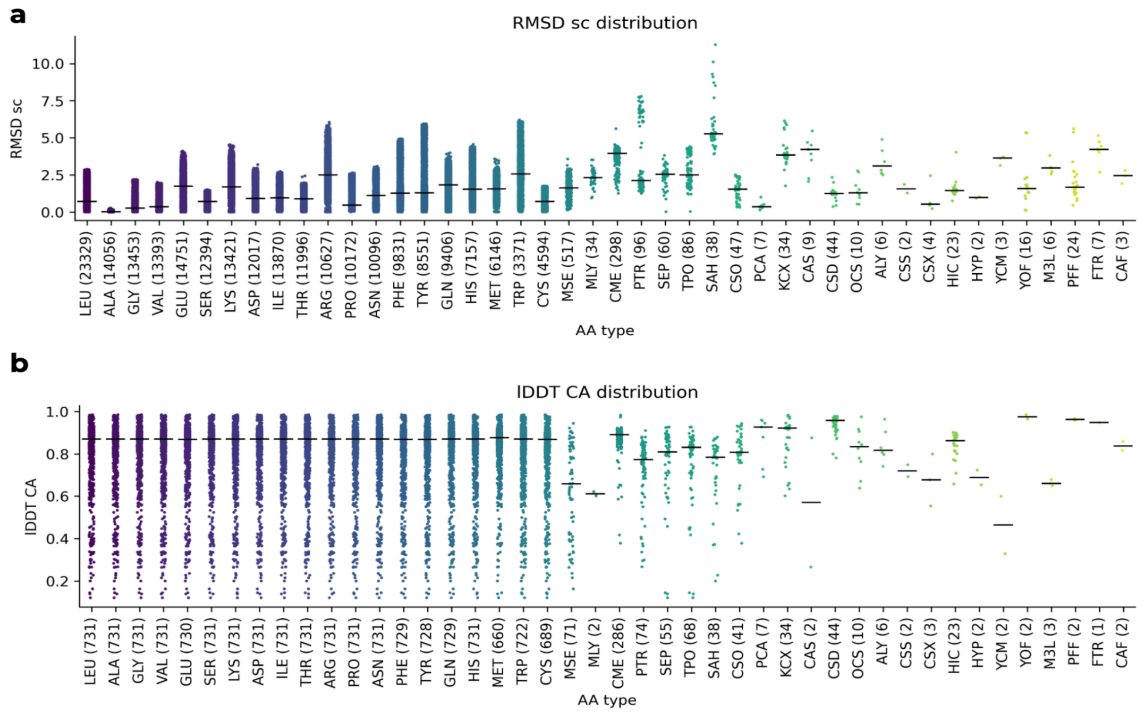


Figure 4. Comparison between proteinogenic and NCAAs on the validation set.
a) Distribution of aligned side chain RMSD values for all amino acid types in the validation set (n = 743). The occurrence of each amino acid is noted in parentheses after the three-letter codes. While the RMSD varies across both canonical and noncanonical amino acids (NCAAs), the ranges are comparable (e.g. ARG vs CME), reflecting inherent structural

uncertainty. Higher median RMSDs for some NCAAs are explained by their lower representation in the dataset. **b)** Median global Cα lDDT scores for structures containing each amino acid type. Canonical amino acids show consistently higher median scores due to their ubiquity across the dataset. Several NCAAs still achieve high accuracy (Cα lDDT > 0.9), with a few exceptions (MLY, SAH, CSS, CSX, HYP, YCM, M3L).

Peptide binder design with noncanonical amino acids

The best utility of protein structure prediction is arguably protein design, as made accurate by AlphaFold2 [1]. It is possible to design single-chain proteins by inverting structure prediction networks, however, this does not necessarily result in functional outcomes. A more interesting, albeit difficult, problem is protein binder design - especially that of small binders, as these have the potential to traverse cell membranes and be administered orally. To test the ability of RareFold, *which has not seen any protein complexes during training*, for design, we invert the network and adapt it to design linear and cyclic peptide binders from sequence information, as in our previous work with the standard 20 amino acids (EvoBind [8]). We incorporate the 29 NCAA of RareFold into a new framework called EvoBindRare (EBR, Figure 5). Twelve different NCAAs are available from our supplier (MSE, MLY, PTR, SEP, TPO, MLZ, ALY, HIC, HYP, M3L, PFF and MHO), and we design both linear and cyclic binders using these and the canonical 20 AAs (32 AAs in total).

Figure 5a shows the design procedure with EBR towards a Ribonuclease (PDB ID 1SSC). Starting from the protein target sequence, we generate an MSA with HHblits [15]**,** and the binding sequence is randomly initiated. The MSA and the binder sequence are then input to a version of RareFold modified for protein-peptide structure prediction. We use the model from step 20000, fine-tuned for 5000 additional training steps to deter inter-residue clashes (Methods). We run 1000 steps of mutation and design binders of different lengths (10-20 residues).

Figure 5b shows the top 10% of designs ranked by loss and their corresponding plDDT scores across peptide lengths, for both linear and cyclic cases. While plDDT scores generally increase with length in the linear designs, this trend is absent in the cyclic ones, likely due to the structural constraints imposed by cyclisation and the resulting variability in binding modes. Horizontal lines at plDDT=85 and loss=0.05 indicate the thresholds for high-confidence structure and low design loss, respectively. These benchmarks reflect the true objective: to obtain at least one high-quality binder per length. The distributions themselves quantify the likelihood of achieving such a design. We selected designs with plDDT>85 and chose the lowest-loss sequence at each length for synthesis, resulting in 7 linear and 6 cyclic peptides. Only one of the cyclic peptides could be synthesised at high purity.

The binding affinities of the synthesised peptides were evaluated using surface plasmon resonance (SPR). Of the linear designs, one out of seven (14%) showed measurable binding in the micromolar range, while the single tested cyclic peptide (1/1) also exhibited micromolar affinity. The linear and cyclic binders had affinities (Kd) of 2.13 μM and 8.77 μM, respectively, compared to 1.81 μM for the known wild-type (WT) binder (see Supplementary Figure 7 for sensorgrams). The successful linear binder contains three types of noncanonical

amino acids (NCAAs), all predicted to interact with the target interface, while the cyclic binder includes one NCAA, also located at the predicted interface (**Figure 5c**). These results demonstrate that RareFold-enabled design can incorporate NCAAs into functional interfaces, yielding binders with experimentally confirmed affinity. Notably, the predicted binding modes differ from that of the WT, indicating that EBR can generate novel interfaces with comparable affinity, without prior exposure to any protein complex (Figure 5d).

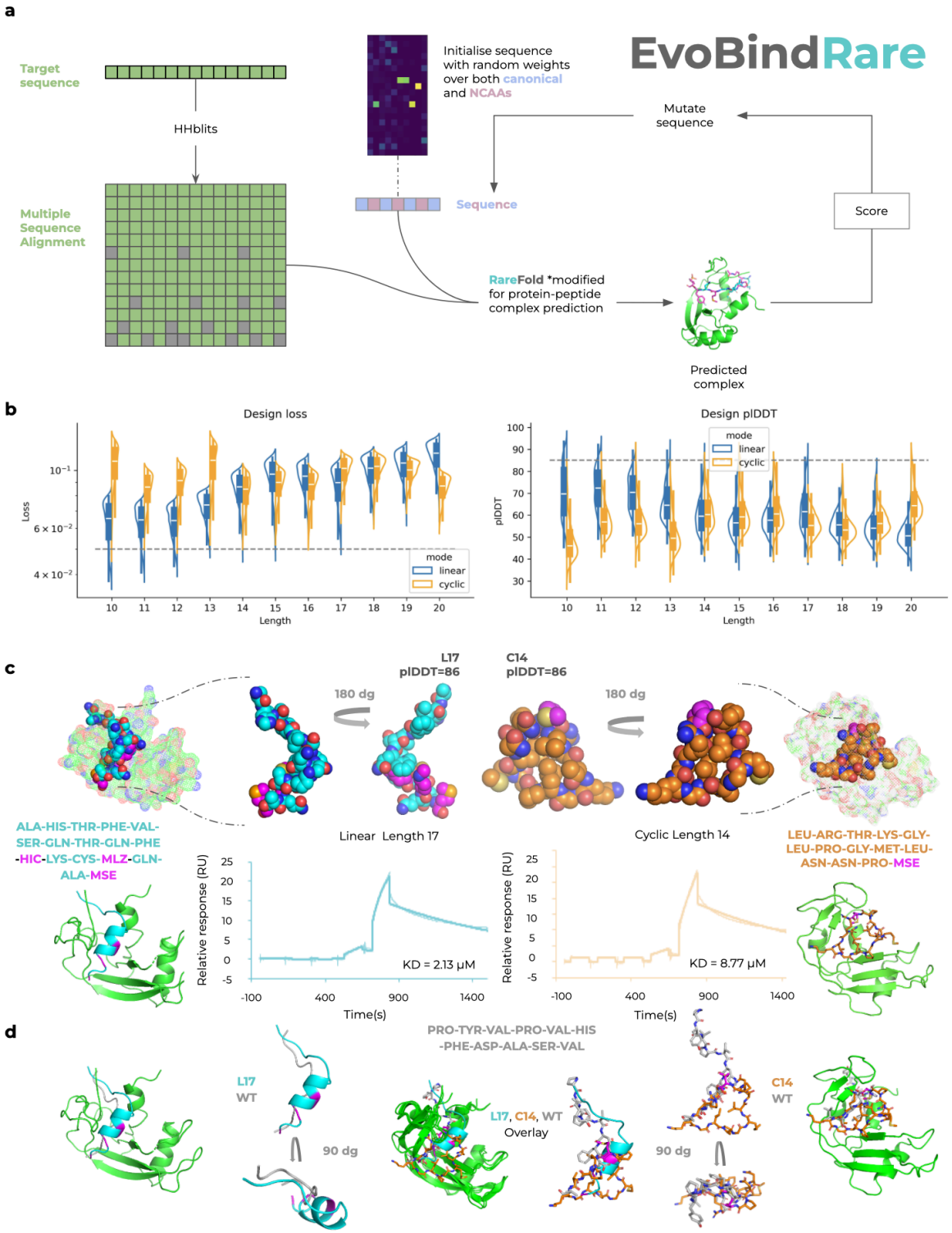


Figure 5. a) Design of peptide binders using EvoBindRare (EBR). Starting from the sequence of a protein target, an MSA is generated with HHblits and a random peptide

sequence is initialised. Both are input to a modified RareFold model for protein-peptide complex prediction. The predicted complex is scored (Methods), a mutation is introduced, and the process is repeated. EBR performs joint structure and sequence optimisation over 1000 mutation steps to design both linear or cyclic binders of 10-20 residues using up to 49 amino acid types. In this study, 32 amino acids (20 canonical + 12 NCAAs) were used due to synthesis constraints. **b)** Optimisation results using EvoBindRare. The top 10% of designs, ranked by loss (left) and plDDT (right), are plotted by peptide length (n = 4804 per length), with the hues indicating linear or cyclic topology. The dashed grey lines represent loss and plDDT cutoffs at 0.05 and 85, respectively. These cutoffs reflect the true objective: to obtain at least one high-quality binder per length. The distributions themselves quantify the likelihood of achieving such a design. **c)** Predicted structures and single-cycle kinetics experiments obtained from surface plasmon resonance (SPR). The sensorgrams show 5 injections from peptide concentrations 2 nM, 20 nM, 200 nM, 2,000 nM, and 20 μM. Both the linear (l17, cyan, Kd=2.13 μM) and cyclic (c14, orange, Kd=8.77 μM) designed peptides are predicted to bind to the same site on the target (green) with NCAAs (magenta) in the interfaces - although in different ways. The affinity measurements reveal that our design strategy successfully creates binders in the same micromolar affinity range as the WT peptide (Kd=1.81 μM) towards the 1SSC target protein. **d)** Structural overlay the predicted designs and the crystal of the known binder (WT). Left: linear binder L17. Right: cyclic binder C14. Middle: combined overlay of all structures. The alignment illustrates the overall agreement between predicted and observed structures and highlights the diversity of binding modes generated by EBR.

Discussion

The results show that RareFold enables accurate structure prediction of proteins incorporating a diverse set (29) of noncanonical amino acids (NCAAs), achieving performance comparable to AlphaFold3 (AF3) on global structure and often exceeding it on local side chain accuracy for key residues. Notably, RareFold produces more physically plausible structures for certain NCAAs such as MSE, avoiding intra-residue atomic clashes observed in AF3. While AF3 yields higher lDDT scores on average, this is expected due to its training on a larger set of structures (including test-similar examples via self-distillation) and longer crop sizes. RareFold, in contrast, maintains high predictive accuracy using a leaner architecture with token-level representations throughout, allowing efficient inference even with limited computational resources. This efficiency is essential for practical protein design, where forward models are called repeatedly to explore large sequence-structure spaces.

Few existing forward models support design with NCAAs, and none, to our knowledge, have been systematically validated with both large-scale benchmarks and experimental tests. HighFold2, a recent transfer-learning extension of AlphaFold-multimer (AFM)[16], has attempted to include modified amino acids for cyclic peptide prediction (not design) [17], but it is evaluated without separating structures seen during AFM pretraining, raising concerns about generalisability. Moreover, performance on liganded structures remains a major challenge, as evidenced by AF3's underwhelming results on novel interfaces [13]. These findings underscore the need for joint progress in computational methods and rigorous

experimental evaluation. RareFold addresses this by offering a scalable and generalisable foundation that we have validated through experimental binder design - specifically, through the EvoBindRare (EBR) platform, which we believe is the first successful method for designing linear and cyclic peptide binders with NCAAs from scratch.

The ability to model and design with NCAAs opens up new avenues in therapeutic development, including targeting previously inaccessible interfaces, improving proteolytic stability, and evading immune recognition. RareFold's token-based architecture is particularly well suited for this, allowing rapid evaluation, adaptability to new target functions, and compatibility with sequence-only design pipelines. Inverse design approaches [8] such as those used in EBR, makes it possible to extend protein design into novel chemical and structural regimes. This is exemplified in our recent HIV-1 work [18], where we successfully designed a potent peptide inhibitor against a challenging viral interface. By uniting accurate NCAA structure prediction with efficient design strategies such as EBR, RareFold lays the groundwork for a new generation of peptide therapeutics that combine structural novelty with functional precision.

Methods

Data

All monomeric protein structures from the PDB were selected on 2024-12-10 determined by X-RAY diffraction or Electron Microscopy (EM) with a resolution ≤5 Å (n=75'232 structures). We extracted the first protein chain in each PDB file and the corresponding sequences with less than 80% non-regular amino acids and more than 50 residues (n=74882/75232 proteins fulfilled these criteria, 99.5%). We clustered the sequences at 20% identity using MMseqs2 (version f5f780acd64482cd59b46eae0a107f763cd17b4d) [19] with the command:

```
mmseqs easy-cluster examples/DB.fasta clusterRes tmp --min-seq-id
0.2 -c 0.8 --cov-mode 1
```

This analysis resulted in 9,031 clusters. We also examined the distribution of amino acid types across the 74,882 structures, which contained a total of 331 unique amino acids. However, the frequencies of noncanonical amino acids (NCAAs) are much lower compared to the standard 20 amino acids commonly found in humans. Among the NCAAs, MSE is the most frequent, largely due to its use in promoting crystallisation, while the others occur with frequencies ranging from 10^1-10^3. To address the limited availability of data for NCAAs, we focus on learning from the top 50 amino acids by frequency, which includes 30 NCAAs. The rarest amino acid in this top 50 is MHO, appearing only 21 times.

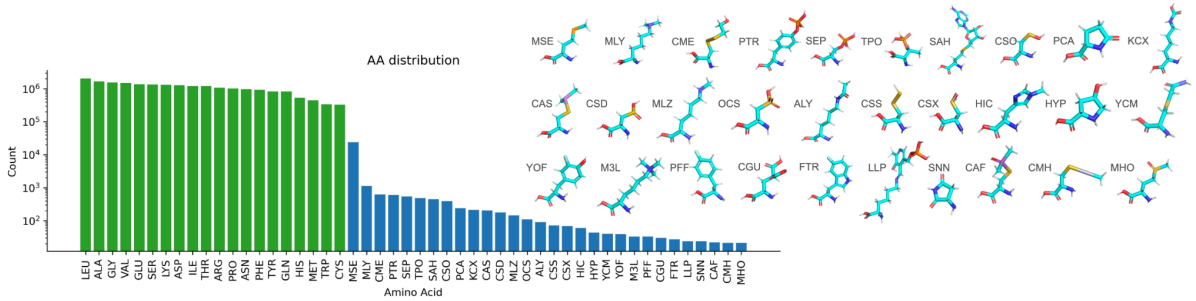


Figure 6. Distribution of amino acids in single-chain protein structures in the PDB. The distribution shows both proteinogenic (green) and noncanonical amino acids (NCAAs, blue) across single-chain protein structures. The 30 most common NCAAs (blue) were included in the training of RareFold. The minimum count in the top 50 most frequent amino acids is 21, represented by MHO. See Supplementary Figure 1 for the distribution of all amino acids present in the PDB.

Noncanonical amino acid descriptions

To predict the coordinates of the additional 29 NCAAs (SNN was excluded since continuous peptide bonds can't be formed with this amino acid) included in this study, we apply the following modifications to the frames available in AlphaFold2 for the standard 20 amino acids. New frames were implemented for atoms extending beyond the standard 20 amino acids structural frames available in AlphaFold2 (see the code, code availability section, for a detailed implementation).

'MSE':,https://www.rcsb.org/ligand/MSE, S → Se in MET
'MLY', https://www.rcsb.org/ligand/MLY, H→ CH3 in LYS
'CME', https://www.rcsb.org/ligand/CME, H → S-C-C-OH in CYS
'PTR', https://www.rcsb.org/ligand/PTR, OH → PO4 in TYR
'SEP', https://www.rcsb.org/ligand/SEP, OH → PO4 in SER
'TPO', https://www.rcsb.org/ligand/TPO, OH → PO4 in THR
'SAH', https://www.rcsb.org/ligand/SAH, add the large group to MET instead of the CH3 group as rigid
'CSO', https://www.rcsb.org/ligand/CSO, H →OH in CYS
'PCA', https://www.rcsb.org/ligand/PCA, add O to PRO
'KCX', https://www.rcsb.org/ligand/KCX, add COO to LYS
'CAS', https://www.rcsb.org/ligand/CAS, H → AS-C-C in CYS
'CSD', https://www.rcsb.org/ligand/CSD - O2 to S rigid in CYS
'MLZ', https://www.rcsb.org/ligand/MLZ, H3 → CH3 in LYS
'OCS', https://www.rcsb.org/ligand/OCS, O3 to S in CYS
'ALY', https://www.rcsb.org/ligand/ALY, CCO to N in LYS
'CSS', https://www.rcsb.org/ligand/CSS, SH to S in CYS
'CSX', https://www.rcsb.org/ligand/CSX, O to S in CYS
'HIC', https://www.rcsb.org/ligand/HIC, add CH3 to HIS
'HYP', https://www.rcsb.org/ligand/HYP, add OH to PRO
'YCM', https://www.rcsb.org/ligand/YCM, add C-CONH2 to CYS
'YOF', https://www.rcsb.org/ligand/YOF, add F to TYR
'M3L', https://www.rcsb.org/ligand/M3L, add 3 methyl groups to LYS

'PFF', https://www.rcsb.org/ligand/PFF, replace OH with F in TYR or add F to PHE
'CGU', https://www.rcsb.org/ligand/CGU, add COO to second carbon in GLU
'FTR', https://www.rcsb.org/ligand/FTR, add F to TRP
'LLP', https://www.rcsb.org/ligand/LLP, added the complex group to LYS
'SNN', https://www.rcsb.org/ligand/SNN, can't form continuing peptide bonds
'CAF', https://www.rcsb.org/ligand/CAF, add DIMETHYLARSINOYL to CYS
'CMH', https://www.rcsb.org/ligand/CMH, add Hg-met to CYS
'MHO', https://www.rcsb.org/ligand/MHO, add O to S in MET

Multiple sequence alignments

There is no one-letter code for these noncanonical amino acids, which means that alignments can't be made. The only ones that can be mapped are:
SEC':'U', 'PYL':'O', 'GLX':'X',. However, since GLX maps to X, it will look like "UNK" (unknown), which also maps to "X". In addition, SEC and PYL are not even among the top 50 AAs. Therefore, we substitute all non-std AAs with X for the MSA search. AlphaFold2 (and later versions) generates three different MSAs. This process constitutes the main bottleneck for the predictions as very large databases such as the Big Fantastic Database [20,21] are searched, which is very time-consuming [22]. To simplify this process, we instead search only uniclust30_2018_08 [23] with HHblits (from HH-suite [15] version 3.1.0):

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

Training and evaluation sets

To compare the method developed here (RareFold) with AlphaFold3 [11] (AF3) which can handle amino acid modifications, although these have to be specified on the atomic level, we analyse what structures in the PDB have less than 20% sequence identity to the AF3 training set (all data in the PDB up to 2021-09-30) and contain NCAAs. Before 2021-09-30, there are 62530 examples, and after 12352, belonging to 8125 and 2360 clusters, respectively, of which 906 are only found after the date cutoff. In these 906 clusters, there are 1654 unique structures, and 177 of these have modified amino acids. However, only 59 are non-MSE, leaving relatively little evaluation data for the other amino acid types.

In total, there are 7599 structures (10.1%) with modified amino acids, making the 177 not in the AF3 training set only 2.3% of the total set. To obtain a valid comparison, we choose to continue with this set and train on all examples before 2021-09-30, test on the 177, and validate on the remainder (Table 1). Supplementary Figure 2a displays the length distribution for the different partitions and Supplementary Figure 2b the number of NCAA in each protein. For the test set, most proteins are of shorter lengths, while the training and validation sets correspond well to each other.

Table 1. Training, validation and test partitions.

Partition	Examples	Modified examples	20% clusters

Train	62530	6679 (10 SNN)	8125
Valid	12175	743	2265
Test	177	177	140
Total	74882	7599	10530

Training

To bias the learning to noncanonical amino acids, we ensure that we always include examples with noncanonical amino acids in half of the batch, sampled according to the inverse frequency of the noncanonical amino acids. The other half of the batch is sampled according to the occurrence of the sequences in the 20% sequence identity clusters to enable the learning of diverse protein structures. We use a batch size of 24 and train across 8 A100 GPUs with 80GB of RAM each. **Figure 7** shows the training curve for individual losses and metrics defined as in AlphaFold2 [1]. We use the same loss function as in AlphaFold2:

$$Loss = 0.5 \cdot FAPE + 0.5 \cdot AUX + 0.3 \cdot Distance + 0.2 \cdot MSA + 0.01 \cdot Confidence \quad (1)$$

Where *FAPE* is the frame aligned point error, *AUX* a combination of the *FAPE* and angular losses, *Distance* a pairwise distance loss, *MSA* a loss over predicting masked out MSA positions and *Confidence* the difference between true and predicted lDDT scores. These losses are defined exactly as in AlphaFold2 and we refer to the description there [1]. From step 20000, the model is finetuned further for 5000 steps by including additional losses (see Fine-tuning).

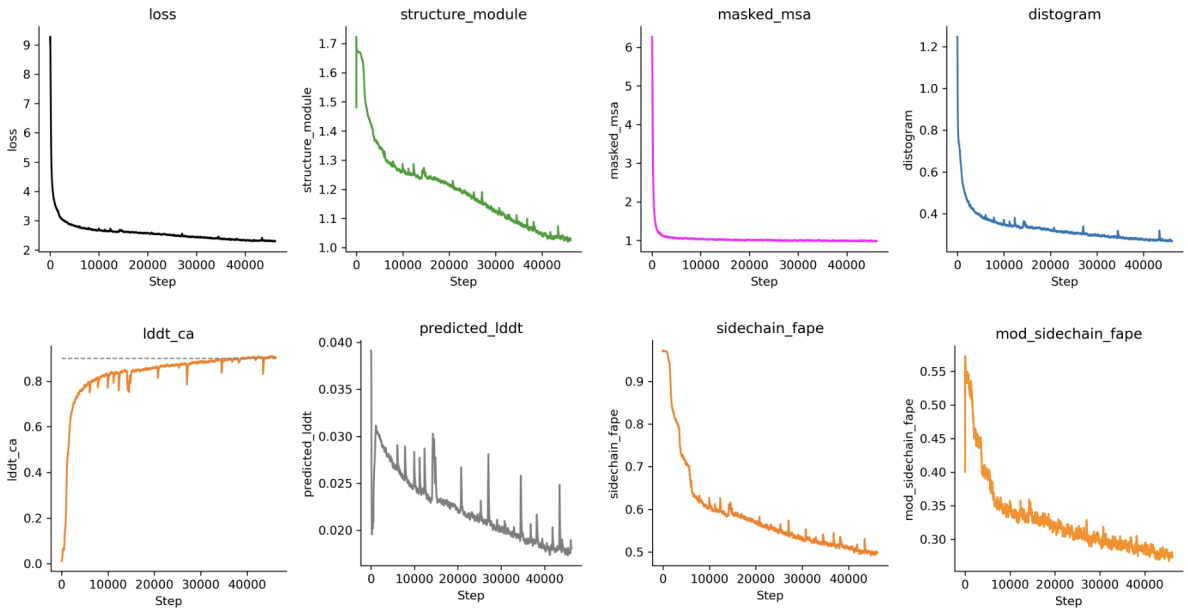


Figure 7. Training curves for the RareFold structure prediction model.

Training losses and metrics over approximately 40000 steps are shown with a smoothing window of 500. The losses include: the combined loss function (equation 1), AUX (structure_module, a combination of FAPE and angular losses), distogram (pairwise distance loss), FAPE (frame aligned point error) for regular amino acid side chains and modified (mod_sidechain_fape), and predicted_lddt (difference between true and predicted lDDT scores), all defined as in AlphaFold2 [1]. The Cα lDDT is also monitored. From step 20000, the model undergoes fine-tuning for 5000 steps with additional losses applied to reduce residue bond violations, clashes, and extreme Cα–Cα distances (see Fine-tuning), resulting in improved structural quality without the need for relaxation.

Fine-tuning

To resolve clashes that may appear between residues, we apply the same fine-tuning losses with the same loss weights as used for AlphaFold2 [1]: between residue bond violations, clashes and extreme CA-CA distances. We do not fine-tune any intra-residue violations. We fine-tuned for 5000 steps, reaching a total of 25000 training steps (600000 seen training examples, Figure 8). The fine-tuning was performed mainly to obtain a model which does not require relaxation (as this is equivalent, see below). This model is useful for design purposes, where the extra relaxation step can become costly if performed at each iteration.

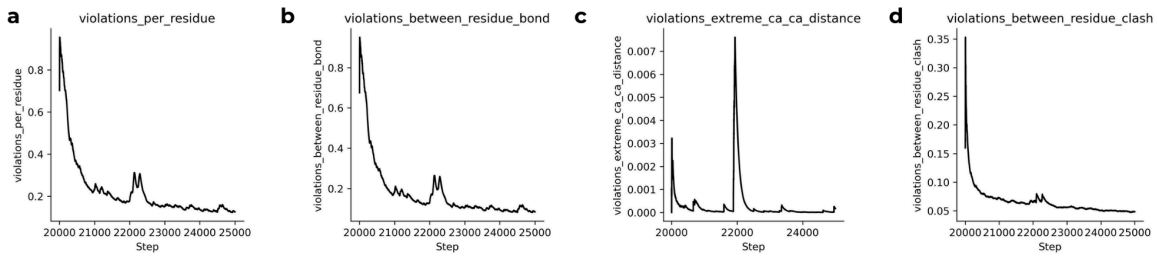


Figure 8. Fine-tuning improves structural violation metrics.
Fine-tuning RareFold for 5,000 additional steps (total 25,000 training steps) using loss functions and weights adapted from AlphaFold2 [1] reduces residue **(a)** and bond violations **(b)**, extreme Cα–Cα distances **(c)** and clashes **(d)**. No intra-residue violations were fine-tuned. This process minimises structural clashes, producing models that do not require computationally expensive relaxation steps, enhancing efficiency for iterative design workflows.

Relaxation

To resolve inter-residue clashes and improve local geometry, we apply molecular dynamics (MD) relaxation as a final step in structure prediction. We use OpenMM [24] with the CHARMM36 force field [25] and the Langevin middle integrator. Since NCAAs are not natively supported, we temporarily replace them with their closest canonical analogues during relaxation and substitute them back afterwards. While relaxation does not substantially alter the predicted structures, it slightly worsens most evaluation metrics (Figure 9).

Validation

We validated every 10000 steps, starting at step 20000, using the 743 examples with modified amino acids from Table 1. For each example, one structure was predicted using the previously generated HHblits MSAs. Of these, 731 (98.3%) were successfully scored, with 12 excluded due to structural inconsistencies. Figure 9 shows that relaxation, performed at step 20,000, does not significantly impact accuracy. However, not all residues could be included in the relaxation due to missing representations in OpenMM, resulting in 679 (93%) structures being relaxed.

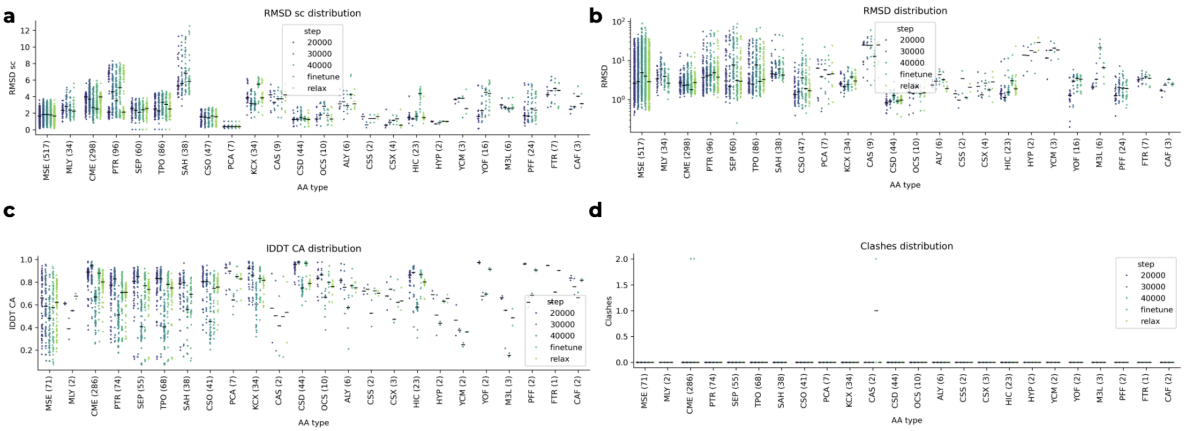


Figure 9. Validation metrics. The validation was performed at steps 20000, 30000 and 40000 with additional relaxation at step 20000 and finetuning 5000 steps from step 20000. Distributions are shown per NCAA for **a)** RMSD of side chains after side chain alignment, **b)** RMSD of side chains after global alignment, **c)** Overall Cα lDDT scores and **d)** intra-residue clashes. Internal clashes are only observed for finetuning and at 40000 steps for CME and CAS.

Scoring

For the evaluation, we align the Cα atoms of the predicted structures and calculate the RMSD of modified atoms for each modified amino acid (mod AA), considering all atoms in that AA (RMSD). The overall Cα lDDT score is computed for global structure accuracy [14] (lDDT CA). For each modified amino acid, the backbone atoms (N, Cα, C) are also aligned, and the side chain RMSD is calculated for the remaining atoms(RMSD sc). The test set consists of 174 structures for RareFold and 171 for AF3 (due to three being OOM using 80 GB of RAM), and the validation set of 731 structures for RareFold.

Recycling

We analysed the effect of varying the number of recycles using 3, 5, 10 and 20 recycling iterations. Interestingly, the number of recycles does not significantly improve the structure prediction of the NCAAs (Supplementary Figure 3). No internal AA clashes were observed at any of the investigated number of recycles except for in one case for TPO at 20 recycles.

AlphaFold3

AlphaFold3 (AF3) [11] can be used to introduce modifications to amino acids and thereby generate NCAAs by the addition of covalent modifications. In the AF3 publication, only 40 examples were evaluated and these were from proteins that had less than 40% sequence similarity with their training set. The predictions were evaluated based on if the aligned RMSD of the NCAAs were <2 Å. Although only around half of these examples were predicted with an RMSD < 2 Å, the evaluation is not rigorous due to the small number of examples, high sequence overlap and disregard of the type of NCAA.

We evaluate AF3 on the test set outlined above, which contains 14 different NCAA types (Figure 1). We use HHblits MSAs and no templates (the same conditions as for RareFold). We take the first prediction using the default settings of 5 predicted models; modelSeeds": [1], "dialect": "alphafold3", "version": 1. We introduce modifications as explained in the AF3 github: https://github.com/google-deepmind/alphafold3/blob/main/docs/input.md. In total, 175/177 proteins could be predicted and 171 scored. Two failed due to being out of memory using NVIDIA A100 GPUs with 80 GB RAM. The remainder failed for unknown reasons, but will not significantly impact the results.

Binder design by inversion of RareFold

Building on our previous work with EvoBind [8] we inverted RareFold to create EvoBindRare (Figure 5a). We constrain the modified amino acids to a set that is readily available for synthesis from our supplier GenScript (Table 2). We initialise a random sequence using the standard 20 AAs and these 12 NCAAs and update and score each sequence using the loss function:

$$Loss\,1 \;=\; peptide\,plDDT^{-1} \cdot \left(\frac{1}{n}\sum_{j=1}^{n} d_j\right) \;+\; \%clashes \qquad (2)$$

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 %clashes is the number of protein-peptide atoms closer than 1.5 Å, divided by the number of peptide atoms. We run the design process for 1000 iterations using 24 different initialisations, lengths 10-20 and both linear and cyclic residue offsets [6] for the peptides. For all design runs, three recycles were used. Due to a rewriting of the prediction with RareFold, we design in batches, resulting in the possibility to design 24 different binders simultaneously on a single NVIDIA A100 GPU with 40 GB of RAM. The 1000 iterations are completed within 24 hours.

Table 2. Three-letter codes of amino acids available from GenScript **(**MSE, MLY, PTR, SEP, TPO, MLZ, ALY, HIC, HYP, M3L, PFF, MHO) and their corresponding names at GenScript.

Three letter code	**GenScript name**
MSE	Se-Met

MLY	Lys(Me2)
PTR	pTYR
SEP	pSER
TPO	pTHR
MLZ	NMe-Lys
ALY	Lys-Ac
HIC	His(1-Me)
HYP	HYP
M3L	Lys(Me3)
PFF	L-4-F-Phe
MHO	Met(O)

Binder Design Selection

From the 1000 iterations x 24 initialisations x 11 lengths (n=264'000) we select the sequences with average peptide plDDT above 85 for both the linear and cyclic case. From these we select the sequence for each length with the lowest loss (*equation 2*) for synthesis resulting in 7 peptides for the linear case and 6 for the cyclic (Tables 3 and 4, Supplementary Figures 4 and 5). All linear peptides could be synthesised at a purity above 90%, but only one of the cyclic designs (length 14).

Table 3. Linear selection (7 peptides). The table shows the iteration and run from which each peptide originated, predicted pLDDT (per-residue confidence), model loss, percentage of atomic clashes, interface distance to the binder, amino acid sequence, peptide length, and experimental purity as assessed by HPLC.

iter	run	plddt	loss	%clashes	if_dist_binder	sequence	length	Purity (%)
950	0	90.91	0.04	0.00	3.44	ASN-GLY-MET-TYR-ILE-GLY-THR-**M3L**-THR-ASN	10	95.1
497	14	88.94	0.04	0.00	3.53	THR-**MLY**-THR-PRO-THR-**MHO**-ILE-**HIC**-LYS-LYS-GLY	11	96.5
636	7	87.89	0.04	0.00	3.65	TYR-**M3L**-ILE-**HYP**-ASP-ILE-PHE-VAL-SER-ASP-**M3L**-PRO	12	90.5
855	3	87.58	0.06	0.01	4.16	ASP-GLN-ARG-VAL-**HYP-PFF**-ILE-VAL-CYS-MET-ASN-PHE-VAL	13	91.7
197	11	86.70	0.05	0.00	4.19	GLU-ALY-**HYP**-PRO-CYS-ASP-LEU-HIS-GLU-**ALY-MLZ**-GLY-GLU-**SEP**	14	96.1

949	16	85.20	0.05	0.00	4.19	PHE-GLU-VAL-HIS-**HIC**-ILE-GLN-VAL-SER-SER-PRO-LEU-GLN-LEU-ASP	15	94.1
610	21	85.76	0.05	0.00	4.62	ALA-HIS-THR-PHE-VAL-SER-GLN-THR-GLN-PHE-**HIC**-LYS-CYS-**MLZ**-GLN-ALA-**MSE**	17	97

Table 4. Cyclic selection (6 peptides). The table shows the iteration and run from which each peptide originated, predicted pLDDT (per-residue confidence), model loss, percentage of atomic clashes, interface distance to the binder, amino acid sequence, peptide length, and experimental purity as assessed by HPLC. Only length 14 could be synthesised at high purity (>90%).

iter	run	plddt	loss	%clashes	if_dist_binder	sequence	length	Purity
906	5	86.07	0.06	0.01	4.21	LEU-SER-SER-GLY-**MHO**-VAL-SER-SER-ASN-ILE-ILE	11	NA
843	3	89.65	0.05	0.00	4.57	ILE-LYS-GLY-LEU-HIS-GLU-ILE-ASP-**MLY**-MET-ILE-ARG	12	NA
895	19	85.67	0.07	0.02	4.21	LEU-ARG-THR-LYS-GLY-LEU-PRO-GLY-MET-LEU-ASN-ASN-PRO-**MSE**	14	92.6%
803	11	89.49	0.05	0.00	4.91	LEU-LYS-ASN-**MHO-SEP**-ILE-CYS-PRO-**PTR**-GLY-LEU-**MLZ-MHO**-GLN-ASP	15	NA
666	9	85.41	0.05	0.00	4.53	**SEP-PFF**-LEU-ALA-CYS-SER-ASP-PRO-**SEP-MHO-PFF**-GLN-ASP-PRO-**M3L-HIC**	16	NA
821	3	87.60	0.06	0.00	5.33	**MSE**-ARG-GLN-GLU-PRO-ILE-TYR-ILE-**PTR-HIC-MLY**-PHE-TRP-THR-**SEP**-GLN-VAL-VAL-**MLZ**-ASN	20	NA

Peptide synthesis

Lyophilised powders of peptides designed by EvoBind2 were synthesised and purified by GenScript. The quality of successfully synthesised peptides (purity >90%) was verified by high-performance liquid chromatography (HPLC) and mass spectrometry (MS). The peptides were lyophilised with TFA salt and water, resulting in a net peptide content lower than the purity achieved before lyophilisation. The net peptide content was determined by GenScript, and serial dilutions for surface plasmon resonance measurements were calculated accordingly. Peptides were resuspended in HBS-EP+ buffer (10 mM HEPES, 150 mM NaCl, 3 mM EDTA, 0.005% Tween-20) to achieve a stock concentration of 1 mM.

Protein production

The expression of psfRNAseA-c001 was described previously [8]. Here is a brief description of the protocol: Transform the psfRNaseA-c001 construct (MBP-8xHis-TEV-RNaseA in pMAL-p2E plasmid) into *E. coli* BL21 (DE3) T1R pRARE2 cells. Inoculate 15 ml/L of Terrific Broth (TB) medium supplemented with 8 g/L glycerol, 50 µg/ml ampicillin, 34 µg/ml chloramphenicol, and 0.4% glucose. Grow overnight cultures at 37°C, 175 RPM. On day 2, cultivate 9 L of culture in the LEX system at 37°C until OD600 reaches 2, then reduce temperature to 18°C. Induce protein expression at OD600 ~3 with 0.5 mM IPTG and continue overnight at 18°C. Harvest cells by centrifugation (10 min, 4500 × g), resuspend in IMAC lysis buffer (1.5 ml/g cell pellet) with Complete EDTA-free protease inhibitor and benzonase nuclease, and freeze at -80°C. In purification, Thaw cell pellets, disrupt by pulsed sonication (4 s on/off, 4 min, 80% amplitude), and centrifuge (20 min, 49000 × g). Filter the supernatant through a 0.45 µm filter and load onto a 5 ml HisTrap HP column followed by a 2 ml HisTrap HP column using ÄKTA Xpress. Wash with IMAC wash 1 (10 mM imidazole) and wash 2 (50 mM imidazole) buffers, then elute with IMAC elution buffer (500 mM imidazole). Perform size exclusion chromatography (SEC) on a HiLoad 16/60 Superdex 200 column using gel filtration buffer (20 mM HEPES, 300 mM NaCl, 10% glycerol, pH 7.5). Pool fractions containing MBP-RNaseA fusion protein (~160 mg yield). Add TEV protease to pooled fractions at a 1:15 molar ratio and incubate overnight at 4°C. Verify cleavage by SDS-PAGE (Supp. 6a). Add 20 mM imidazole to the reaction mixture and pass through two 2 ml HisTrap columns to remove the MBP-His tag. Further purify by passing over 1 ml Amylose resin. Collect the flow-through, concentrate using a Vivaspin 5 kDa MWCO filter, and exchange buffer to storage buffer (20 mM HEPES, 300 mM NaCl, 10% glycerol, pH 7.5) using a PD-10 column (Supp. 6b). Measure final concentration (8.2 mg/ml), aliquot (50 µl/tube), flash-freeze in liquid nitrogen, and store at -80 °C.

Surface plasmon resonance

To measure the binding affinity of designed peptides to 1SSC, we employed surface plasmon resonance (SPR) using a Biacore 8K system. SPR detects real-time biomolecular interactions by monitoring refractive index changes near a sensor surface. Experiments were conducted at 25°C in HBS-EP+ running buffer (10 mM HEPES, 150 mM NaCl, 3 mM EDTA, 0.005% Tween-20).

1SSC was immobilized on a Series S Sensor Chip CM5 to over 10,000 RU via amine coupling with 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide (EDC) and N-hydroxysuccinimide (NHS). Peptides were captured on the chip, yielding a maximum response (Rmax) of 25–50 RU. Single-cycle kinetics experiments used peptide concentrations of 2 nM, 20 nM, 200 nM, 2,000 nM, and 20 µM, diluted in the running buffer. Each concentration series was injected at 30 µL/min without regeneration, with 120 s association and 1,800 s dissociation phases.

Raw data were analysed using Biacore Insight Evaluation Software. Sensorgrams underwent reference subtraction (using an activated/deactivated reference flow cell) and blank subtraction (median of consecutive buffer injections) to correct for non-specific binding and bulk effects. Processed data were fitted to a 1:1 binding model [26] to derive global

kinetic parameters: association rate constant (ka), dissociation rate constant (kd), and equilibrium dissociation constant (KD).

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. Albanese KI, Barbe S, Tagami S, Woolfson DN, Schiex T. Computational protein design. Nature Reviews Methods Primers. 2025;5: 1–28.

3. 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.

4. Goverde CA, Pacesa M, Goldbach N, Dornfeld LJ, Balbi PEM, Georgeon S, et al. Computational design of soluble and functional membrane protein analogues. Nature. 2024;631: 449–458.

5. Zhu J, Liang M, Sun K, Wei Y, Guo R, Zhang L, et al. De novo design of transmembrane fluorescence-activating proteins. Nature. 2025; 1–9.

6. 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

7. Zhang C, Zhang C, Shang T, Zhu N, Wu X, Duan H. HighFold: accurately predicting structures of cyclic peptides and complexes with head-to-tail and disulfide bridge constraints. Brief Bioinform. 2024;25. doi:10.1093/bib/bbae215

8. Li Q, Vlachos EN, Bryant P. Design of linear and cyclic peptide binders of different lengths from protein sequence information. bioRxiv. 2024. p. 2024.06.20.599739. doi:10.1101/2024.06.20.599739

9. Amino acid. Wikimedia Foundation, Inc.; 28 Oct 2001 [cited 20 Jan 2025]. Available: https://en.wikipedia.org/wiki/Amino_acid

10. Non-proteinogenic amino acids. Wikimedia Foundation, Inc.; 1 Dec 2012 [cited 4 Mar 2025]. Available: https://en.wikipedia.org/wiki/Non-proteinogenic_amino_acids

11. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630: 493–500.

12. Krishna R, Wang J, Ahern W, Sturmfels P, Venkatesh P, Kalvet I, et al. Generalized biomolecular modeling and design with RoseTTAFold All-Atom. Science. 2024 [cited 20 Jan 2025]. doi:10.1126/science.adl2528

13. Škrinjar P, Eberhardt J, Durairaj J, Schwede T. Have protein-ligand co-folding methods moved beyond memorisation? bioRxiv. 2025. p. 2025.02.03.636309. doi:10.1101/2025.02.03.636309

14. 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.

15. 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: 1–15.

16. 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

17. Zhu C, Cao S, Shang T, Guo J, Su A, Li C, et al. Predicting the structures of cyclic peptides containing unnatural amino acids by HighFold2. bioRxiv. 2025. p. 2025.01.16.633493. doi:10.1101/2025.01.16.633493

18. Daumiller D, Giammarino F, Li Q, Sönnerborg A, Diez RC, Bryant P. Single-Shot Design of a Cyclic Peptide Inhibitor of HIV-1 Membrane Fusion with EvoBind. bioRxiv. 2025. p. 2025.04.30.651413. doi:10.1101/2025.04.30.651413

19. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol. 2017;35: 1026–1028.

20. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. Nat Commun. 2018;9: 1–8.

21. Steinegger M, Mirdita M, Söding J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. Nat Methods. 2019;16: 603–606.

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

23. Mirdita M, den Driesch L V, Galiez C, Martin MJ, Söding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res. 2017;45. doi:10.1093/nar/gkw1081

24. Eastman P, Swails J, Chodera JD, McGibbon RT, Zhao Y, Beauchamp KA, et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. PLOS Computational Biology. 2017;13: e1005659.

25. Huang J, MacKerell AD. CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. Journal of computational chemistry. 2013;34. doi:10.1002/jcc.23354

26. Palau W, Di Primo C. Simulated single-cycle kinetics improves the design of surface plasmon resonance assays. Talanta. 2013;114: 211–216.

Data Availability

All data presented here is available at https://zenodo.org/uploads/14892196.

Code Availability

The code for RareFold and EvoBindRare is available at:
https://github.com/patrickbryant1/RareFold

Acknowledgements

The protein purification was facilitated by the Protein Science Facility at Karolinska Institutet, Stockholm, and we would like to thank Dr Emilia Strandback, Dr Tom Reichenbach, Dr Henrik Spåhr, and Dr Tomas Nyman for assistance. PyMOL and Blender were used for structural visualisation.

Funding

This study was supported by the SciLifeLab & Wallenberg Data Driven Life Science Program (grant: KAW 2020.0239, P.B). The computing power was enabled by the Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre with project ids Berzelius-2023-267, Berzelius-2024-78, Berzelius-2024-292 and Berzelius-2025-41.

Contributions

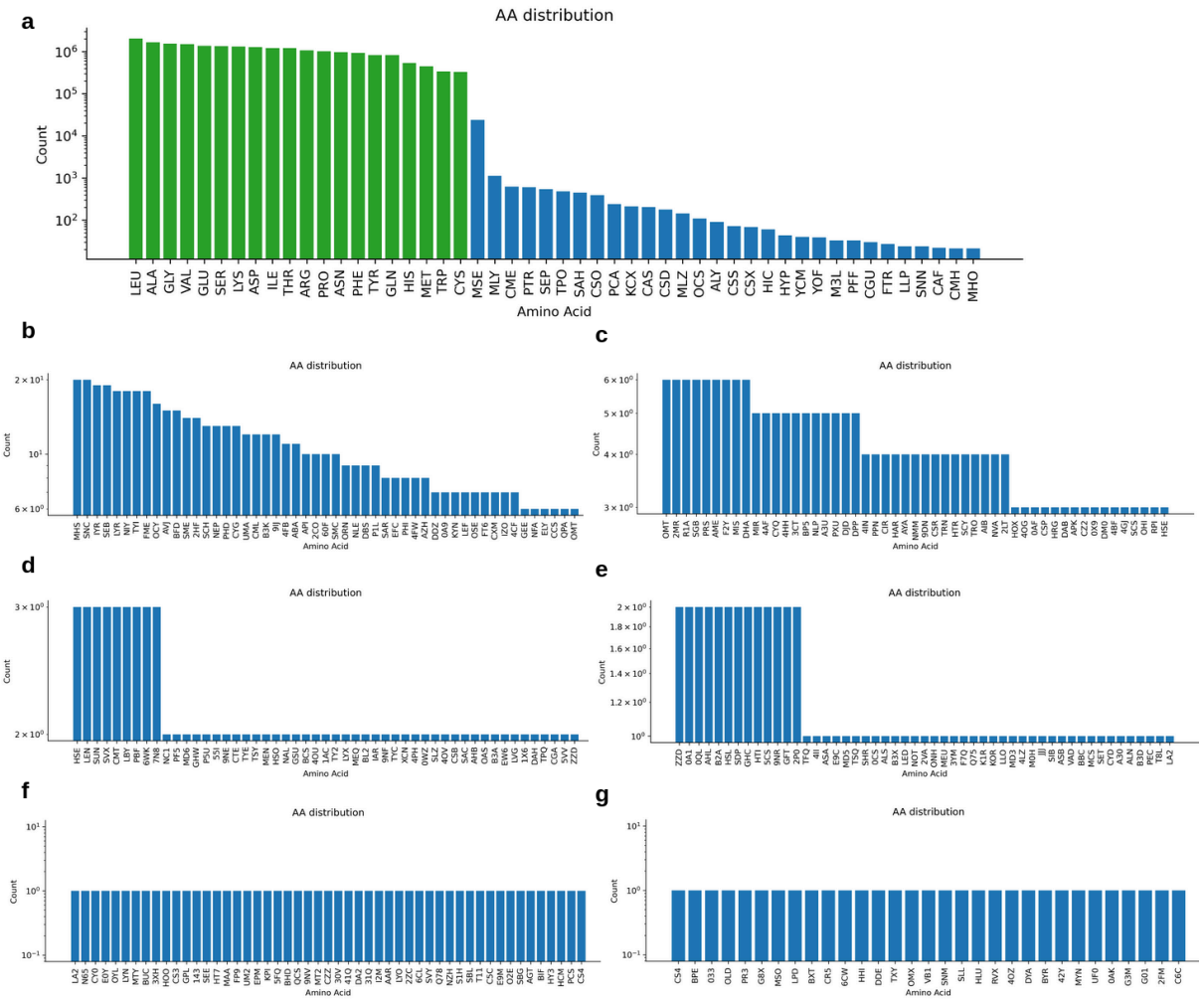
P.B. designed the study, developed RareFold and EvoBindRare, designed the binders and wrote the initial draft of the manuscript. Q.L. performed SPR affinity measurements. D.D. visualised some of the structures in coordination with P.B. and Q.L.
P.B. obtained funding.

Conflicts of interest

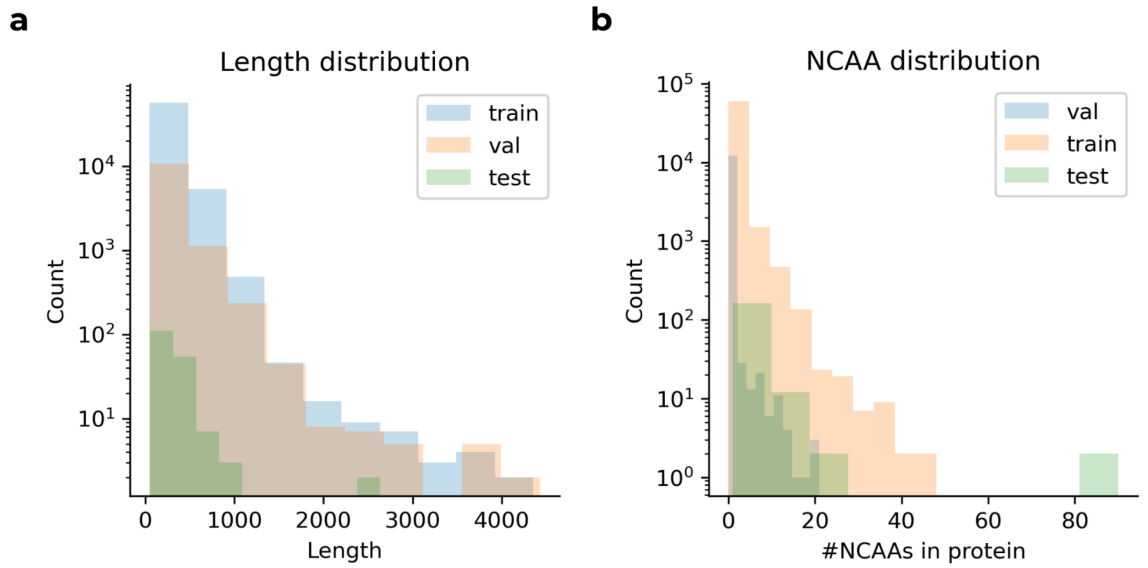
P.B. is a shareholder in Cyclic Therapeutics.

Supplementary Information

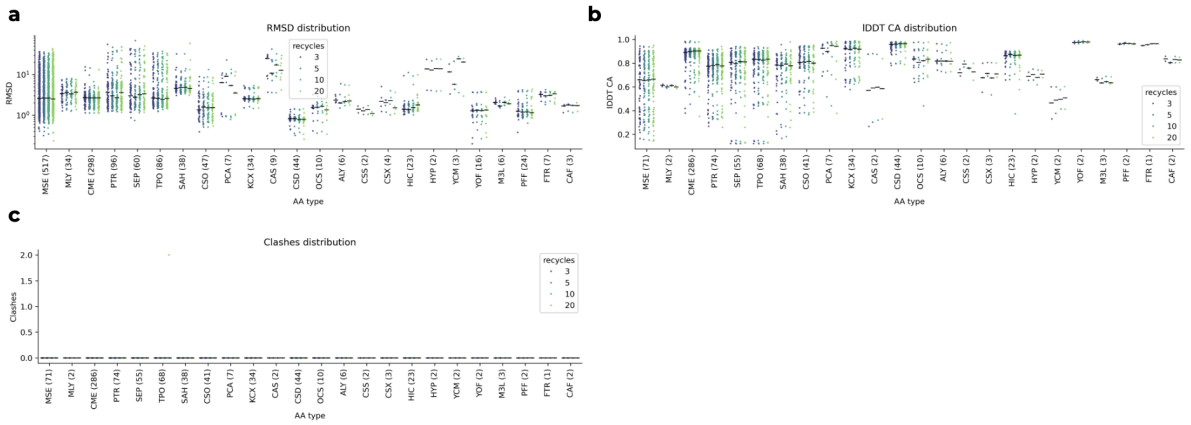
Supplementary Figures



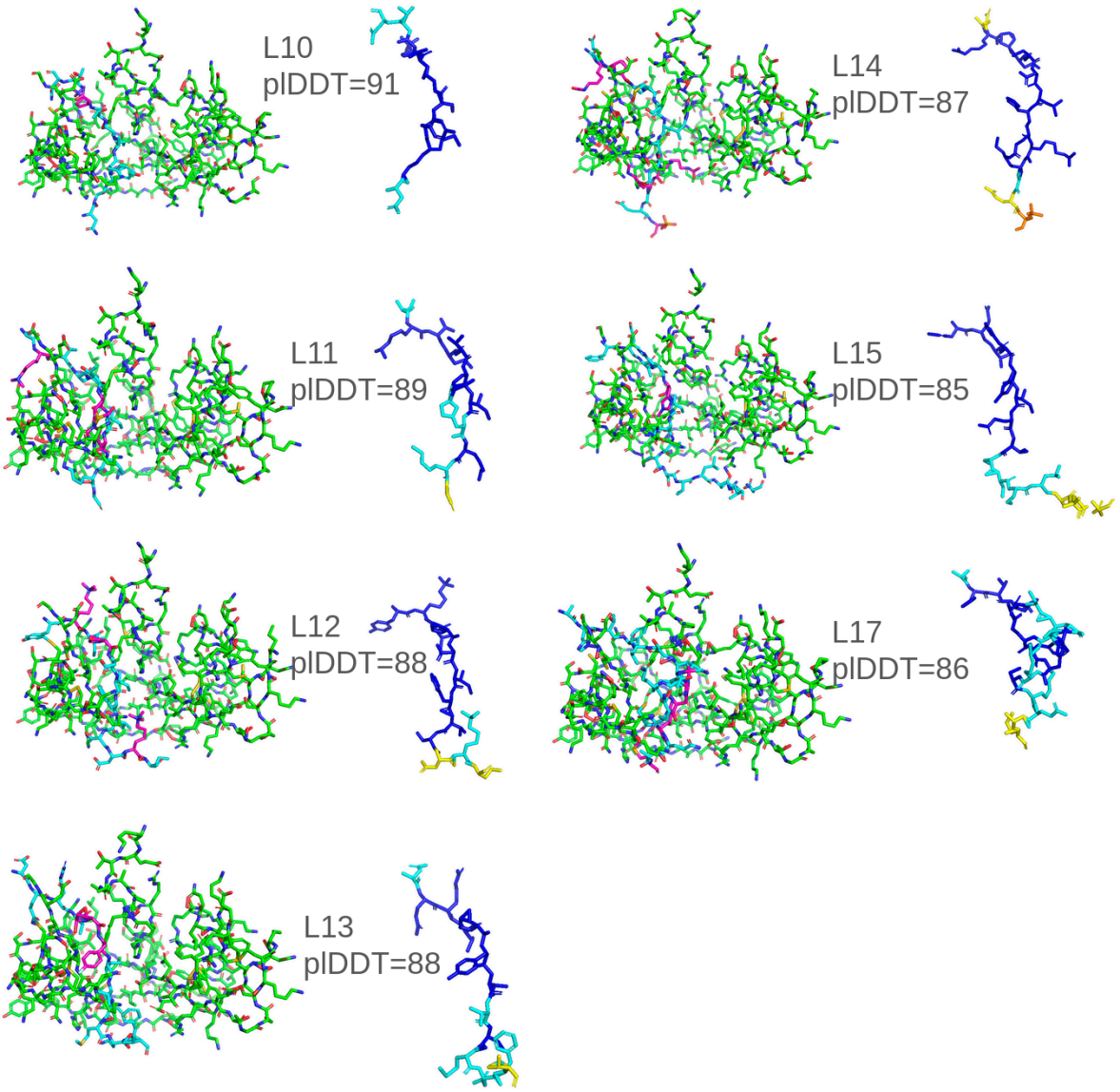
Supplementary Figure 1. Amino acid distribution among single-chain structures in the PDB. The figure shows the frequency distribution of all amino acid types observed in single-chain protein structures deposited in the Protein Data Bank (PDB). Both canonical and noncanonical amino acids are included, illustrating the chemical diversity present in experimentally determined structures. This dataset forms the basis for identifying rare amino acids (NCAAs) used in RareFold training and evaluation.



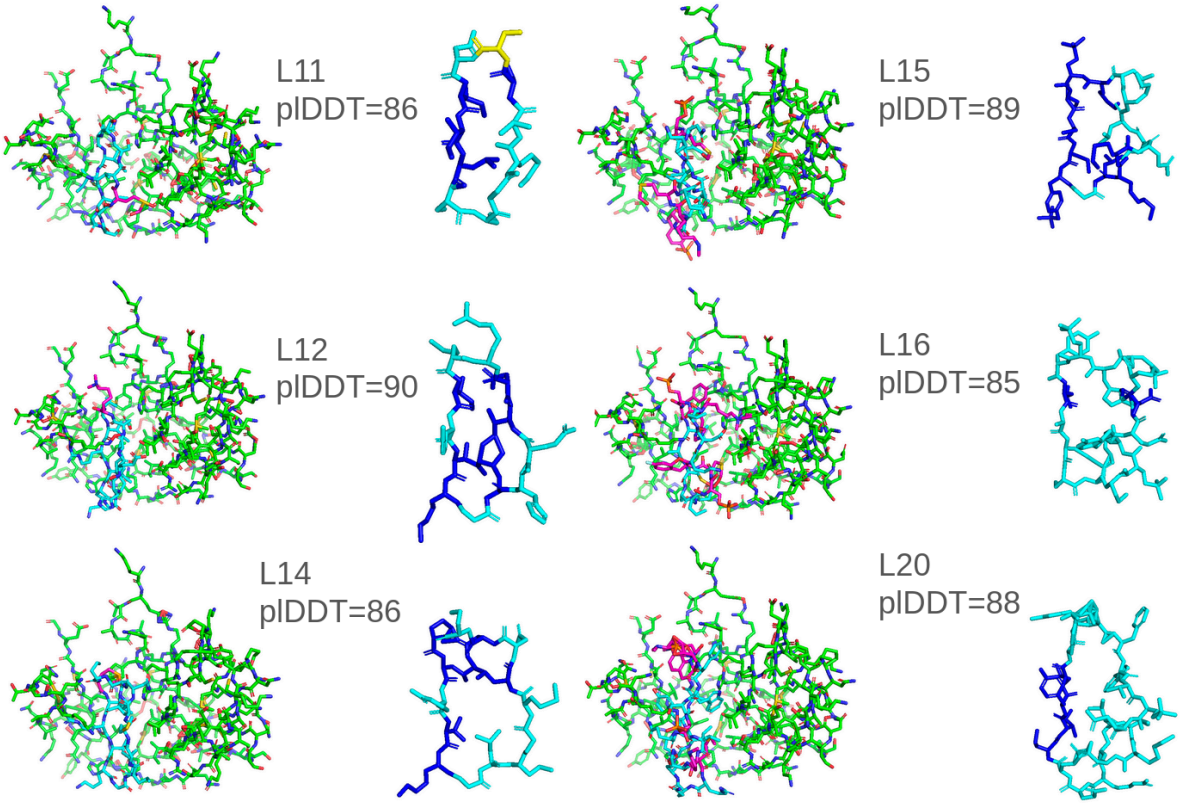
Supplementary Figure 2. Length and modification distributions per partition. **a)** Chain length distribution per partition: The distribution of protein chain lengths in the training, validation, and test sets. **b)** Number of modified amino acids per protein and partition: The number of modified amino acids in each protein across the partitions. The training set contains more proteins with modifications than the validation and test sets, with an outlier in the test set featuring more than 80 modified amino acids.



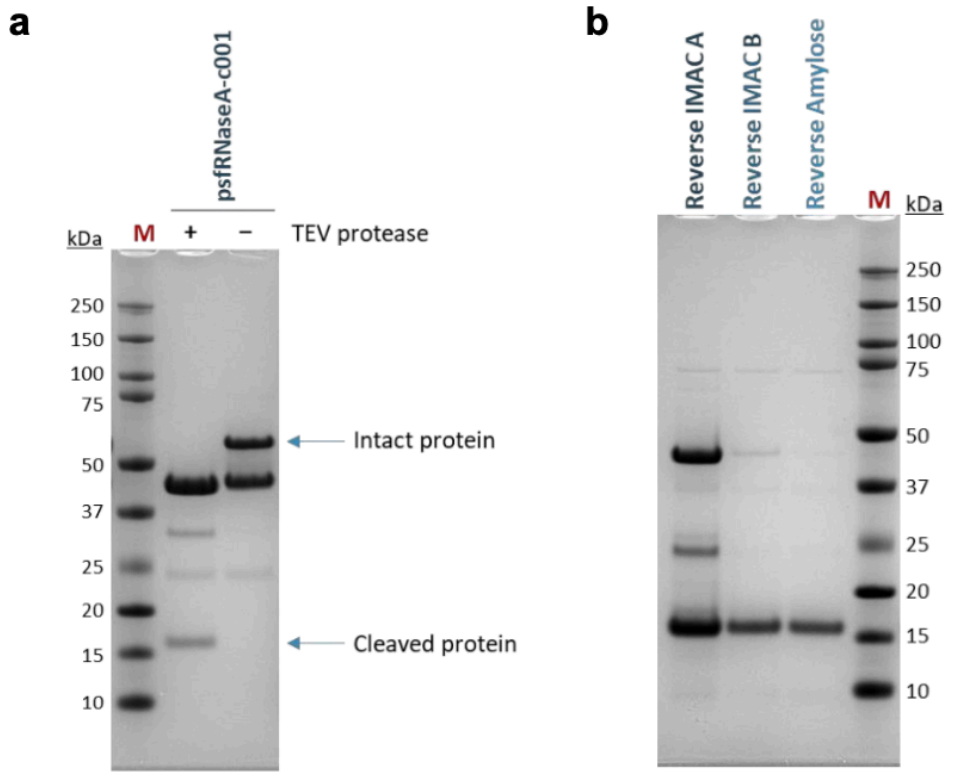
Supplementary Figure 3. Number of recycles and accuracy metrics on the validation set for NCAAs. RMSD, lDDT CA, and clash distributions are shown across different numbers of recycling steps. While the overall structural quality remains stable, additional recycling steps beyond a certain point do not lead to significant improvements in prediction accuracy. This suggests that RareFold converges early in the process, with NCAAs being predicted consistently even after fewer recycling steps.



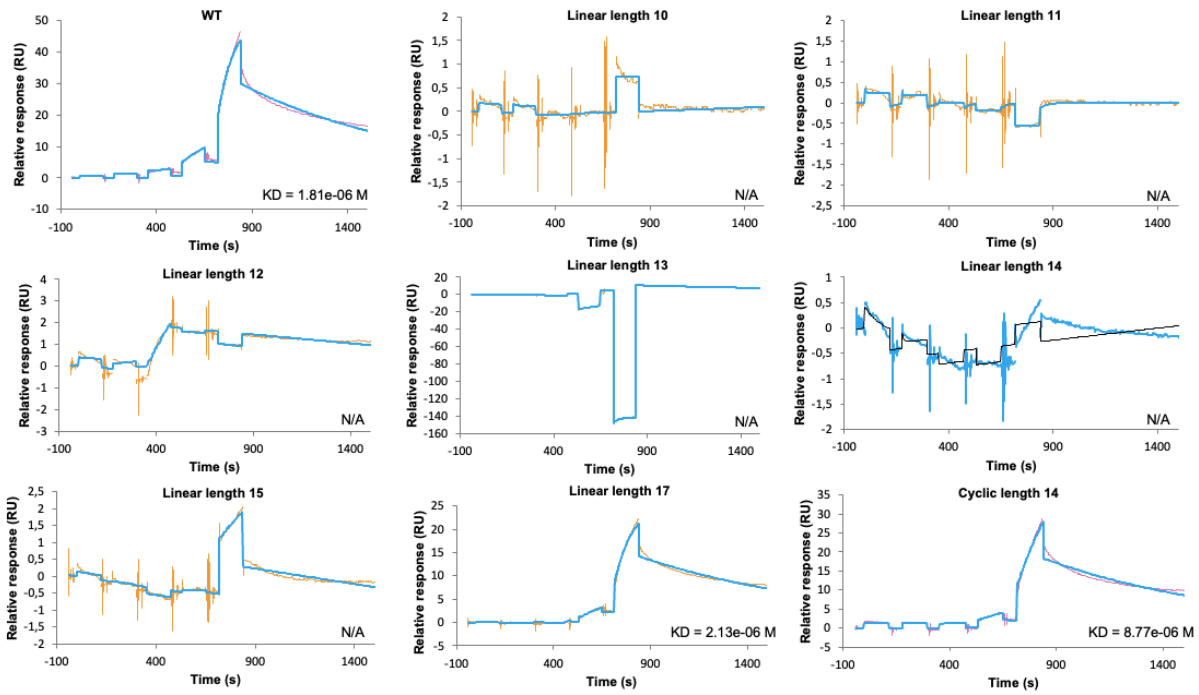
Supplementary Figure 4. Linear selection of designed peptides (unrelaxed structures). Seven peptides were selected based on structural quality and interface metrics without applying structure relaxation.



Supplementary Figure 5. Cyclic selection of designed peptides (unrelaxed structures). Six peptides were selected based on structural quality and interface metrics without applying structure relaxation.



Supplementary Figure 6. **a) SDS-PAGE analysis after tag removal with TEV protease**. TEV protease was added to the purified protein in a 1:15 molar ratio, and the reaction mixture was incubated in the cold room overnight. (+) indicates addition of TEV protease, (–) no addition of TEV protease. **b) SDS-PAGE analysis after reverse IMAC and reverse Amylose purifications.** The cleaved target protein was passed through two 2 ml HisTrap columns and one 1 ml Amylose resin to remove the MBP-His tag and TEV protease.



Supplementary Figure 7. Sensorgrams for all evaluated peptides. Surface plasmon resonance (SPR) sensorgrams were generated using a Biacore 8K system. Each graph depicts real-time binding interactions between immobilized ligands and analytes in solution. Each subfigure represents a separate binding experiment, displaying the relative units (RU) on the y-axis versus time (seconds) on the x-axis. The colored lines represent the experimental data, while the blue lines show the fitted curves based on a 1:1 binding model (Methods).