Computational design of soluble analogues of integral membrane protein structures

Casper A. Goverde[1]†, Martin Pacesa[1]†, Lars J. Dornfeld[1], Sandrine Georgeon[1], Stéphane Rosset[1], Justas Dauparas[2,3], Christian Schellhaas[1], Simon Kozlov[5], David Baker[2,3,4], Sergey Ovchinnikov[5], Bruno E. Correia[1]*

[1] Laboratory of Protein Design and Immunoengineering, École Polytechnique Fédérale de Lausanne and Swiss Institute of Bioinformatics; Lausanne, Switzerland

[2] Department of Biochemistry, University of Washington; Seattle, WA, USA

[3] Institute for Protein Design, University of Washington; Seattle, WA, USA

[4] Howard Hughes Medical Institute, University of Washington; Seattle, WA, USA

[5] John Harvard Distinguished Science Fellowship Program, Harvard University; Cambridge, MA, 02138 USA

† These authors contributed equally

* Corresponding author. Email: bruno.correia@epfl.ch

Abstract:

D*e novo* design of protein folds with complex topologies and intricate structural features using solely computational means remains a significant challenge. Here, we use a robust deep learning pipeline to design soluble analogues of integral membrane proteins. Some unique protein topologies, such as the GPCR fold, are not found in the soluble proteome and we demonstrate that their structural features can be recapitulated in soluble analogues. Biophysical analyses reveal high thermal stability of the designs and experimental structures show remarkable accuracy. Our workflow enables the design of complex protein topologies with high experimental success rates and low sequence similarity to natural proteins, leading to a *de facto* expansion of the soluble fold space. We anticipate such design strategies will give us access to new functionalities such as novel small molecule binders, enzymes and potentially a new approach to drug discovery studies.

Main Text:

Introduction

Protein design enables the expansion of nature's molecular machinery, creating synthetic proteins with novel functionalities. Traditionally, protein design has been dominated by physics-based approaches, such as Rosetta[1]. However, these methods require empirical restraints to guide the design process, and often extensive experimental screening and optimization. This proves problematic for the design of functional proteins with complex structural topologies. Recently, structure prediction pipelines, such as AlphaFold2 (AF2)[2] and RoseTTAfold (RF)[3] achieved unprecedented accuracy in predicting protein structure given the amino acid sequence. With the rise of deep learning-based methods, exploring novel sequence space has become increasingly feasible, allowing the discovery of proteins with stable topologies and new functions. Deep learning powered methods have been impactful in a number of tasks that include the generation of novel designable backbones[4–8], oligomeric protein assemblies[9,10], proteins with embedded functional motifs[11], novel protein structural descriptors[12], the sequence design problem[13–15], and more recently the generation of a diverse range of protein topologies using diffusion models[6,10,16–19]. Additionally, it was found that structure prediction networks can be inverted and used for protein design, resulting in the generation of plausible protein backbones[5,8,20].

Nevertheless, designing protein folds with complex structures, including non-local topologies and large sizes, remains challenging yet essential for creating new protein functions. In addition to design proficiency, many questions about the fundamental determinants of protein structure and folding remain elusive, particularly regarding the generalizability of deep learning methods beyond natural protein structures and sequences. To probe some of these questions, we analyzed the protein fold space in the SCOP database[21] and observed a segregation at the structural level between proteins in the soluble proteome and in the cell membrane environments (Fig. 1a). We observed that 1075 membrane proteins exhibit unique topologies that are not found in soluble form, with only 189 folds being present in both soluble and membrane environments. This raises the question whether integral membrane protein topologies have some fundamental structural features that preclude them from existing in the soluble fold space. Consequently, we investigated if membrane folds could be designed as soluble analogs,

thus achieving a *de facto* fold expansion of the soluble proteome and creating opportunities for designing novel functions using these previously inaccessible protein folds.

To this end, we developed a computational pipeline for robust *de novo* protein design based on the inversion of the AF2 network[5] coupled with sequence design using ProteinMPNN[14] (Fig. 1b). Our approach allowed us to design highly stable folds that were previously very challenging to obtain by computational design (Ig-like folds, β-barrels, TIM-barrels), as well as soluble analogues of integral membrane protein folds (claudin, rhomboid protease, G-protein coupled receptor (GPCR)) without the need for experimental optimization or empirical restraints. Our findings showcase the remarkable success and accuracy of deep learning-based methods in protein design, paving the way for exploring new protein topologies and sequences for improved functional design strategies.

Results

Structure-driven sequence generation using deep learning

AF2-based design approaches have been shown to generate plausible protein backbones[5,22,23], however, their performance in sequence design was suboptimal as evidenced by the low experimental success rates[5,9,24].

Wicky and coworkers[9] have demonstrated the efficiency of using ProteinMPNN on AF2-generated structures to enhance their expression and solubility, but it remained unclear whether it could be successfully applied to explore the sequence space of complex protein folds with intricate topological features including those only found in membrane environments (Fig 1a). To address this challenge, we integrated a previously developed AF2-based design approach ($AF2_{seq}$)[5] with the ProteinMPNN framework (Fig 1b).

In this pipeline we used $AF2_{seq}$ to generate sequences that adopt a desired target fold. $AF2_{seq}$ optimizes a sequence based on a loss function that comprises both structural and structural confidence loss components (see Methods). The loss function combined with a gradient descent algorithm was used to update a Position Specific Scoring Matrix (PSSM), which tracks the sequence probabilities of the input sequence. This step is repeated until a sequence is found that folds to the desired topology. We then apply ProteinMPNN sequence generation to the designed structure, which given a backbone

generates the most likely sequences for this fold. Finally, all designed sequences are repredicted with AF2 and RF, and filtered based on their structural similarity to the target (TM-score > 0.8), confidence scores (pLDDT > 80), and sequence novelty relative to naturally occurring sequences (e-value > 0.1).

In silico assessment showed that, despite the restricted structural diversity (Supp Fig. S1), AF2-designed backbones enable ProteinMPNN to generate a much larger protein sequence diversity for a desired fold than classical backbone sampling methods, such as Rosetta Backrub[25] or molecular dynamics (MD) simulations (Fig. 1c,d). This was especially striking since the MD simulations reached an $RMSD_{C\alpha}$ of 4.10 Å whilst AF2-designed backbones exhibited a lower $RMSD_{C\alpha}$ of 1.37 Å on average (Supp Fig. S1). Additionally, we found that for some more stringent design examples, the target structure could not be predicted in single sequence mode by AF2 upon ProteinMPNN redesign. However, when using a combination of $AF2_{seq}$ and ProteinMPNN ($AF2_{seq}$-MPNN), we found the input sequence to result in accurate structural predictions of the target folds (Supp Fig. S2). Therefore, we sought to test whether our design strategy would be successful in designing protein folds that have thus far been challenging to other approaches.

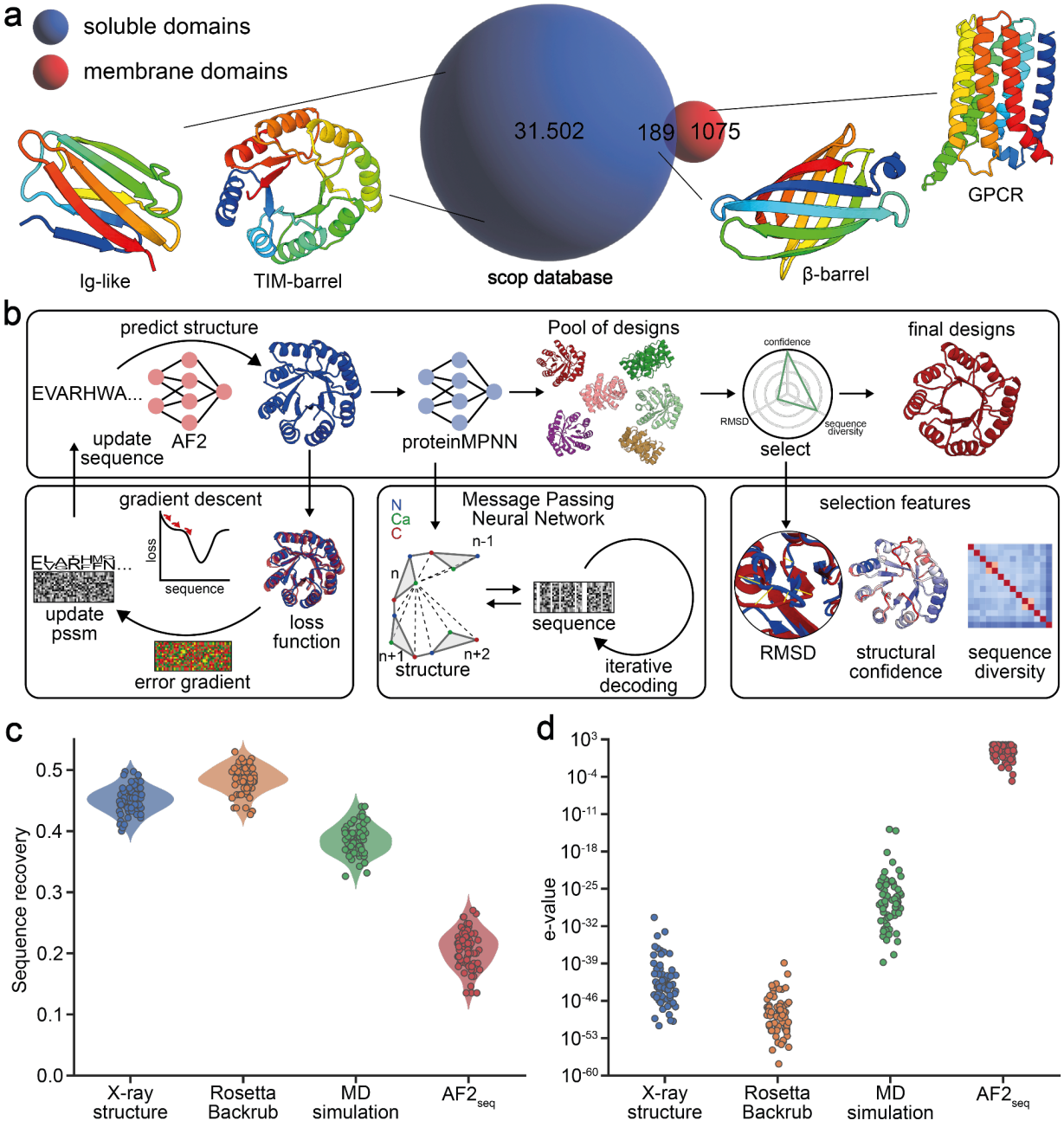

Figure 1 | Overview of the fold space across different environments and computational design approach. a, Overview of the occurrence of soluble and membrane folds in the SCOP structural database with depictions of selected representatives. **b,** Schematic representation of the integrated design pipeline for backbone and sequence generation. Given a target structure, an initial sequence is generated using AF2 through loss function optimization. The resulting structure is then passed to ProteinMPNN to sample novel sequences for a given fold. ProteinMPNN designs are filtered based on structural similarity to the target, confidence and sequence diversity. **c,** Sequence recovery of TIM-barrel ProteinMPNN designs generated based on the reference crystal X-ray structure (5BVL), Rosetta-perturbed backbones, MD simulation trajectories, or $AF2_{seq}$ generated structures. **d,** Novelty of generated sequences resulting from different backbone sampling methods evaluated by e-values relative to the non-redundant protein sequence database.

Computational design of topologically complex soluble folds

Topological and structural complexity of protein folds is difficult to quantify, nevertheless we used the metrics of protein length and sequence contact order to provide an assessment on topological complexity of protein structures (Supp Fig. S3, Methods). Our rationale for these metrics derives from the intuition that folds with non-local contacts may have a more restricted set of viable sequences and protein length inevitably leads to an explosion of the sequence space. Based on this assessment and on how challenging some folds have been for computational design, we selected three folds to test our approach: Ig-like fold (IGF), β-barrel fold (BBF), and TIM-barrel fold (TBF) (Fig. 2a). All of these folds have long-range contacts, especially when compared to other computationally designed proteins of the same size (Supp Fig. S3). Our initial focus was on one of the most prevalent folds in nature: the Ig-like fold, which is an essential building block for immunological effectors such as antibodies[26] or cell surface receptors[27,28]. The development of IGFs for specific antigen binding has given rise to novel therapeutics, with demonstrated potential for increasing manufacturability and extending the spectrum of therapeutic action[29,30]. The IGF consists of two stacked β-sheets, presenting a significant design challenge due to the non-local nature of its interactions and the edge β-strands that render these proteins susceptible to aggregation[31,32]. This has been well exemplified by previous design efforts of this fold that required a fine hand-crafting of structural features and strict empirical restraints[33,34]. We used the $AF2_{seq}$-MPNN protocol (Fig 1b) to design novel IGFs and observed they were confidently predicted to fold into the target structure (Supp Fig. S4). The majority of the designed sequences were significantly distant from natural protein sequences (Fig. 2b). Upon filtering based on predicted AF2 confidence scores and sequence novelty, we selected 19 highly diverse designs (24% average sequence identity, with some as low as 8%) for experimental characterization (Supp Fig. S5). Seven designs were soluble and purifiable, and 4 exhibited monodisperse peaks in solution according to size exclusion chromatography coupled to multi-angle light scattering (SEC-MALS) (Supp Fig. S6). Exemplified by the design IGF_10 (Fig. 2c), the designed IGFs exhibited the typical β-sheet rich secondary structure profile according to circular dichroism (CD) spectroscopy, along with unusually high thermostability (Fig. 2c, Supp Fig. S6)[35].

Next, we attempted to design a *de novo* β-barrel (BBF), a fold which can be found both in the soluble as well as in the membrane proteome. Design of novel BBFs can enable the creation of new small molecule binders, transporters, and sensors[36–40]. The BBF consists of eight antiparallel β-strands with precise hydrogen bonding patterns[36], which makes the design of β-barrels extremely challenging. Dou et al. were the first to overcome this challenge using a set of design principles involving the introduction of glycine kinks, β-bulges, and tryptophan corners to alleviate backbone strain and allow for continuous hydrogen bonding connectivity[36]. We investigated whether our pipeline could successfully generate novel BBFs without explicitly defining these restraints. We selected 25 designs for experimental characterization, out of which six were folded according to CD spectroscopy and were monomeric in solution as assessed by SEC-MALS (Fig. 2d). Many of the designed proteins also showed extremely high thermal stability (Fig. 2d, Supp Fig. S7). Upon closer inspection of the sequences, we found that our folded designs had high glycine residue recovery at the glycine kink positions of the design target, as also observed by Verkuil et al[41]. However, some designs showed that different residues could be utilized without compromising protein folding (Supp Fig. S8). Additionally, the tryptophan corners do not appear to be essential since many of the designs have different residues at these positions. This demonstrated that not all empirically-derived design features were necessary to obtain folded β-barrel designs and successfully explored uncharted sequence space.

Finally, we attempted to design the TIM-barrel fold (TBF), also a challenging protein topology, which has a critical importance in biology, as its structure is highly proficient in supporting enzymatic active sites, making it an ideal candidate for the design of enzymes with novel catalytic functions[42–44]. The TIM-barrel fold comprises eight parallel-paired β-strands, each separated by an alpha helix, resulting in long-range interactions between the β-strands[45]. The TBF has been a long-standing challenge in protein design[45–47], and only very recently, several studies successfully designed this fold[10,48,49]. Previous TBF design strategies imposed symmetry and other empirical constraints both at the structural and sequence level[48,49]. Removing the symmetry constraints increases the possible sequence space that is explored by many orders of magnitude, 10^{59} with symmetry to 10^{239} with no symmetry. When designing the TBF with our pipeline, we did not impose any constraints besides the target fold backbone, thus allowing for a much larger

structural and sequence diversity in the designs Supp Fig. S9). We experimentally assessed 25 designs, CD spectroscopy confirmed that five of the designs were folded and highly thermostable, and SEC-MALS analysis determined them to be monomeric in solution (Fig. 2e, Supp Fig. S10).

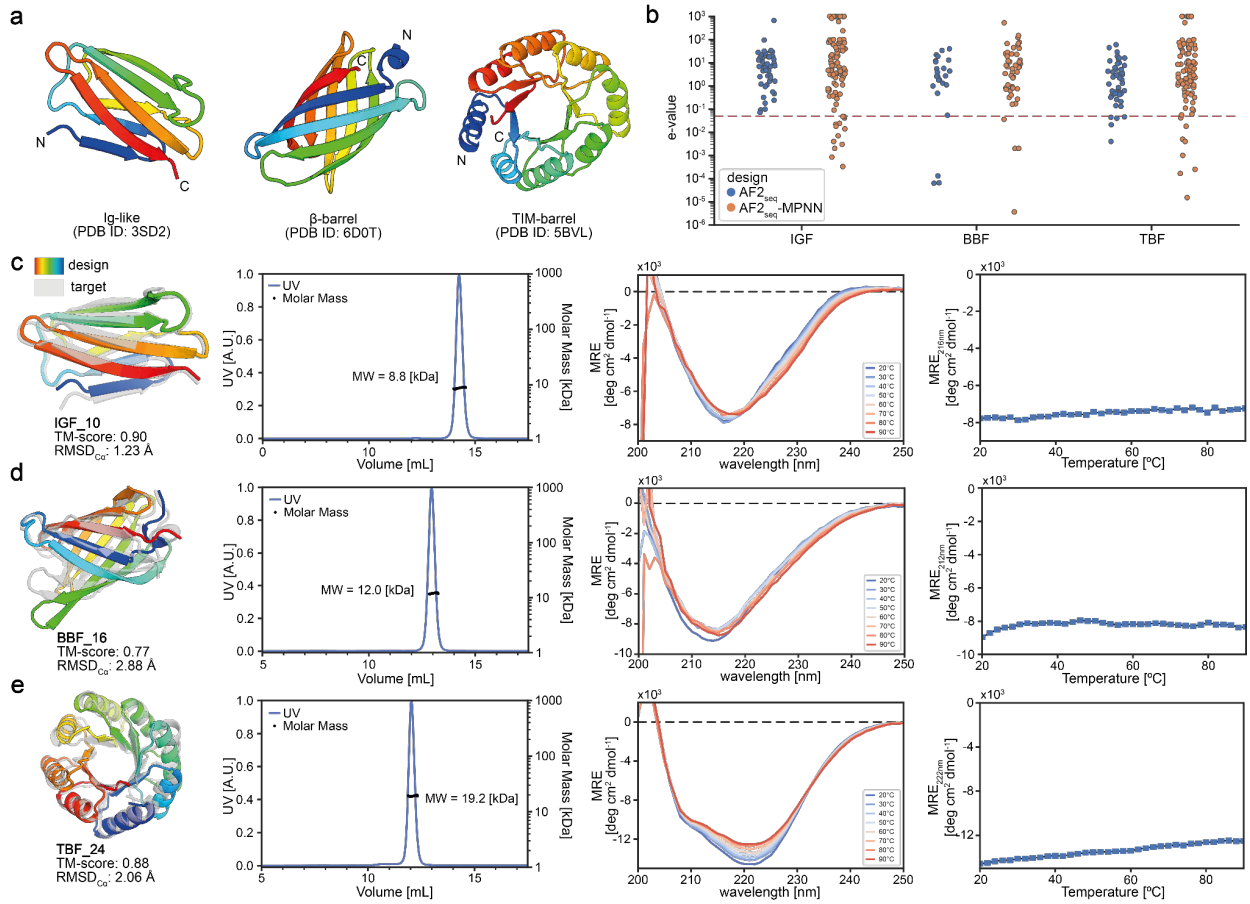

Figure 2 | Experimental characterization of designed soluble topologies. a, Cartoon depiction of three protein topologies which have been challenging for computational design: Ig-like Fold (IGF), β-barrel Fold (BBF), and TIM-barrel Fold (TBF). **b,** Closest e-value hit of the $AF2_{seq}$ and $AF2_{seq}$-MPNN designs when searching a non-redundant protein sequence database. Significance threshold of 0.05 is highlighted, indicating little sequence homology to natural sequences. Characterization of design **c,** IGF_10, **d,** BBF_16, and **e,** TBF_24, respectively. Superposition of the design (color) and the target fold (gray), the corresponding SEC-MALS measurement, CD-spectra at different incubation temperatures, and the CD melting curve.

We solved a crystal structure of TBF_24 at 1.34 Å resolution (Fig. 3). Our design showed noticeable structural deviations from the initial template (Fig. 3a), with an overall backbone $RMSD_{C\alpha}$ of 2.06 Å, with variations at the residue level as large as 5.53 Å (Supp Fig. S11). When comparing the x-ray structure to the designed model, the Cα and full-atom $RMSD_{fa}$ are 0.80 Å and 2.12 Å, respectively (Fig. 3b). These structural comparisons

demonstrate the remarkable accuracy of our design approach, further underlined by the almost identical sidechain placement within both the core and peripheral regions of the protein (Fig. 3c). Given the encouraging results from our initial designs, we wondered whether our approach would allow us to probe the sequence space of topologies not present in the soluble proteome, such as those of integral membrane proteins.

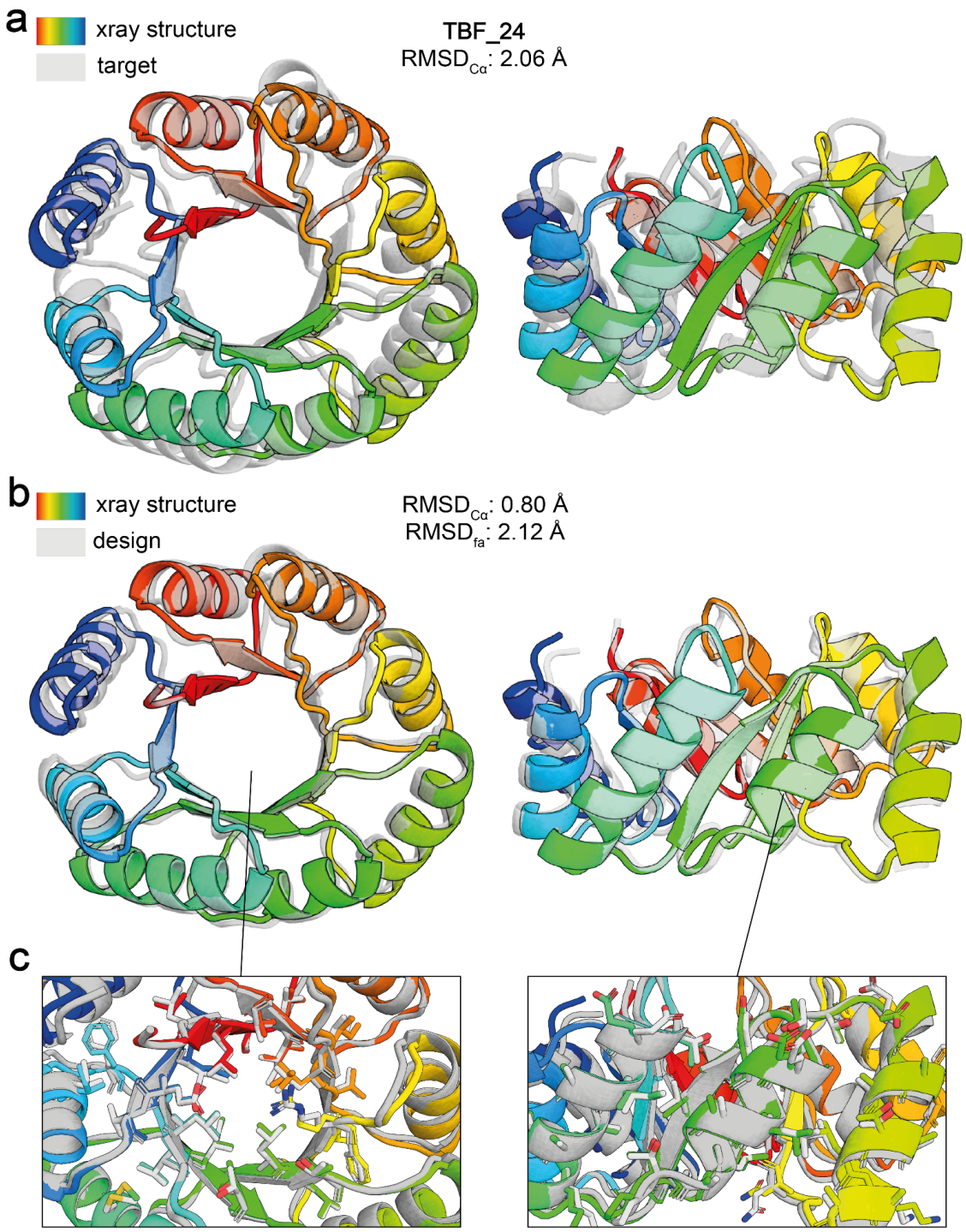

Figure 3 | X-ray structure of TIM-barrel design TBF_24. a, X-ray structure (colored) superimposed on the target fold (gray). **b,** X-ray structure (colored) superimposed on the design model (gray). **c,** Comparison of the sidechains in the core and on the surface of TBF_24 versus the X-ray structure. $RMSD_{C\alpha}$ - root mean square deviation computed over the Cα atoms of the backbone. $RMSD_{fa}$ - root mean square deviation computed over all the atoms in the structure.

Expanding the soluble fold space with membrane protein topologies

In a domain analysis performed over the SCOP database, we observed that both the soluble and membrane proteome each encompass a group of unique structural protein topologies, with only a narrow overlap between the two (Fig. 1a, Fig. 4a). This prompted us to ask whether it is possible to design soluble analogues of such membrane-only folds or whether they contain intrinsic structural features that preclude them from existing in soluble form. Previous studies have demonstrated that simply substituting exposed hydrophobic residues for polar or charged amino acids might not be sufficient to solubilize them, since the interactions between the surface residues have to be carefully considered[50–53]. To address this question, we set out to design soluble analogues of membrane proteins using the $AF2_{seq}$-MPNN pipeline with several methodological variations (see Methods). We selected three membrane folds to test the design strategy: the claudin fold[54], the rhomboid protease fold[55], and the GPCR fold[56] (Fig. 4b).

Initial designs by $AF2_{seq}$ exhibited a high sequence novelty when compared to natural proteins (Fig. 4c) and a low fraction of surface hydrophobics (Fig. 4d), however, none could be expressed in soluble form (Fig. 4e). We attempted to optimize the sequences using the standard ProteinMPNN model, but the resulting sequences consistently recovered the surface hydrophobics, most likely due to the topology's similarity to membrane proteins encountered during training (Fig. 4d). Hence, we tried biasing the amino acid sampling towards hydrophilic amino acids ($AF2_{seq}$-$MPNN_{bias}$), which improved the solubility of the designs slightly (Fig. 4d), however, the number of successful designs generated was low. Therefore, we retrained the ProteinMPNN network using a dataset of only soluble proteins, which we named soluble MPNN ($MPNN_{sol}$) (Fig. 4c-4e). $AF2_{seq}$-$MPNN_{sol}$ was able to produce novel sequences with high confidence scores predicted by AF2 and low fraction of surface hydrophobics (Fig. 4d, Supp Fig. S4). As a result, we were able to generate high confidence designs of membrane protein topologies that do not exist in the soluble proteome.

We started by designing soluble analogues of the claudin fold, a class of proteins involved in the formation of tight junctions, which are critical in controlling the flow of molecules between layers of epithelial and endothelial cells[57]. Claudin folds are composed of an alpha/beta mixed secondary structure where there are four transmembrane alpha helices and an extracellular β-sheet[57]. The composition of the β-sheet determines the type of tight

junction between cells that is being formed, resulting in highly selective permeability of ions and solutes[57]. Claudin-targeting therapies hold great promise as novel cancer therapies and soluble claudin analogues could introduce a novel route to screen for claudin binders[58]. We tested 13 designs for the claudin-like fold (CLF) out of which 10 were found to be expressed in soluble form (Fig. 4e). Five designs were further biochemically characterized, and three were monomers in solution according to SEC-MALS and were folded, with two showing a T_m above 90 °C (Fig. 4f, Supp Fig. S12). The CLF designs showed a sequence identity below 13% relative to the native fold and nearest e-values to natural sequences below 0.063 (Fig. 4c, Supp Fig. S4). AF2 predicted structures in comparison with the designed models exhibited $RMSDs_{C\alpha}$ ranging from 2.84 - 4.03 Å (Supp Fig. S12). The CLF design series showed that our approach could successfully design soluble analogues of membrane proteins with rather simple topologies such as 4-helix bundles.

Next, we attempted to design a topologically more complex fold, the rhomboid protease fold (RPF). The RPF comprises six transmembrane alpha helical domains, with many intricate loops and long-range contacts[59] (Fig. 4b). Additionally, it harbors a serine-histidine catalytic dyad buried in the cell membrane, allowing it to cleave transmembrane protein domains, which plays an important role in cell signaling, making it a therapeutically interesting target[60,61]. We selected 15 designs for protein expression, out of which 13 were found to be soluble (Fig. 4e) and five were selected for further experimental characterization. Three of the five designs showed a single monomeric species in solution and the expected helical secondary structure as assessed by CD (Fig. 4g, Supp Fig. S13). All of the three monomeric species of RPFs exhibited high thermal stability, with T_m above 90 °C (Fig. 4g, Supp Fig. S13). Interestingly, the AF2 structure predictions of the designs were less accurate than those of the CLF designs, with the $RMSD_{C\alpha}$s between models and predictions ranging between 3.34 and 5.57 Å (Fig. 4g, Supp Fig. S13). Overall, the high $RMSD_{C\alpha}$s between design models and AF2 predictions highlight the inherent difficulty in designing folds with such structural complexity.

Finally, we tested our design approach in one of the most prevalent membrane folds in nature: the GPCR fold. GPCRs are the largest and most diverse family of membrane receptors in eukaryotes, playing important roles in signaling pathways[56]. About 34% of all US Food and Drug Administration (FDA) approved drugs target GPCRs and they remain

the most studied drug target[62,63]. The core topology of GPCRs comprises seven transmembrane helices that facilitate numerous non-local interactions, enabling them to bind a variety of ligands, including photoreceptors, odors, pheromones, hormones, and neurotransmitters[56]. De novo design of GPCR-like folds (GLF) provides the potential of creating new small molecule receptors and protein scaffolds with functional sites of the GPCRs. We tested 56 designs out of which 36 were expressed to be soluble, and we selected the 10 most highly expressed designs for further biochemical characterization. From the ten designs, nine were monodisperse monomers in solution, and all showed the characteristic CD signature of alpha-helical rich proteins (Fig. 4h, Supp Fig. S14). All ten designs also showed high thermal stabilities (T_m > 90 °C) (Fig. 4h, Supp Fig. S14).

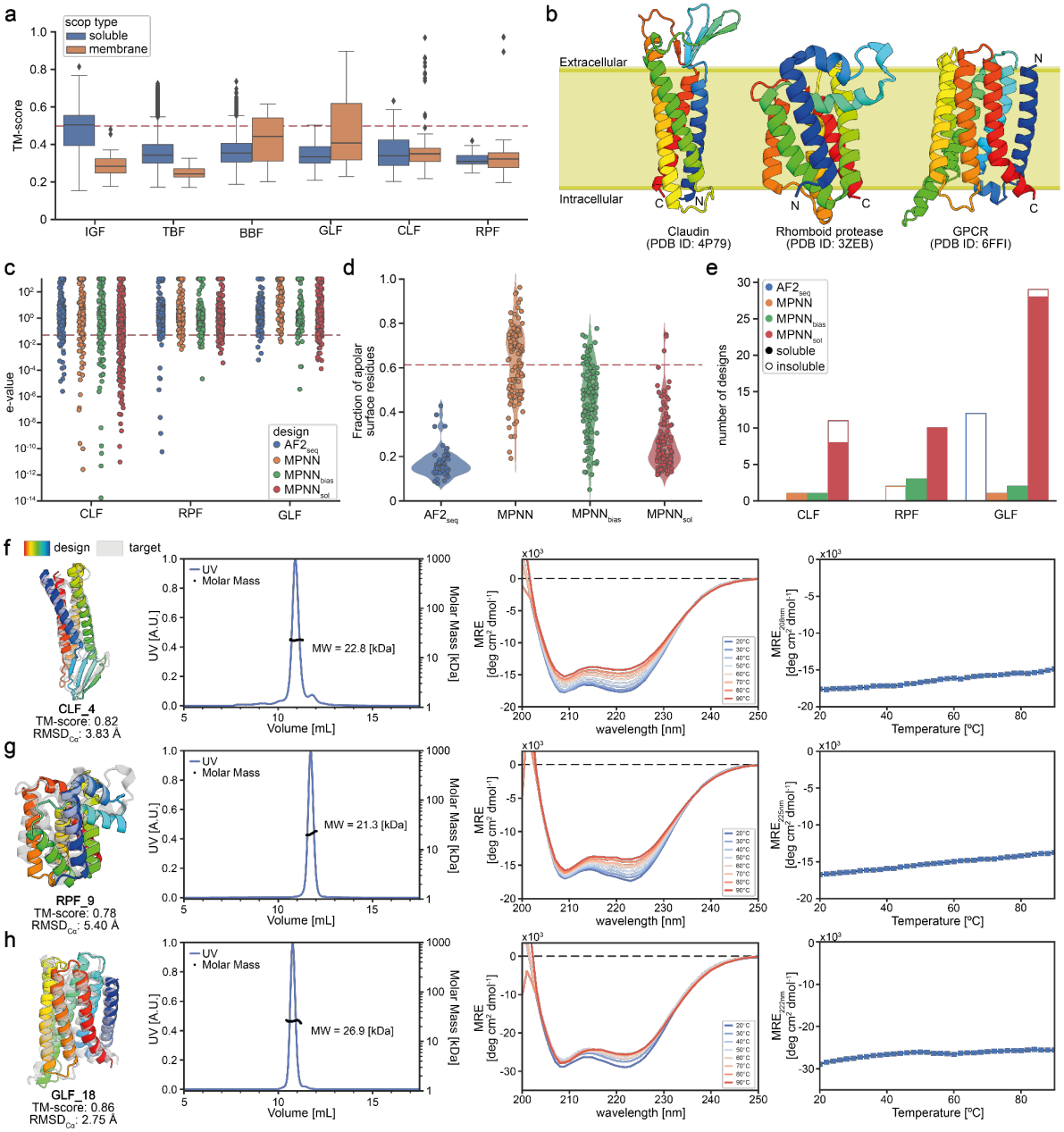

Figure 4 | Experimental characterization of soluble analogues of membrane proteins. a, Structural similarity for each of the target folds against the SCOP database. TM-score cut-off of 0.5 is highlighted, denoting significant structural similarity as compared to the reference fold. **b,** Cartoon representation of three transmembrane topologies chosen to be redesigned as soluble folds: Claudin-Like Fold (CLF), Rhomboid Protease Fold (RPF) and GPCR-Like Fold (GLF). **c,** Closest e-value hits of the solubilized CLF, RPF and GLF against a non-redundant protein sequence database. The vast majority of the designed sequences differ substantially from natural sequences as indicated by e-values higher than the significance threshold of 0.05 (red line). **d,** Fraction of hydrophobic residues found on the surface of the GLF designs using different sequence-generation methods following $AF2_{seq}$ backbone generation. The fraction of surface hydrophobics of the native GLF is 0.61 (red line) **e,** Number of designs resulting in soluble expression of the designed soluble membrane protein analogues. Experimental characterization of **f,** CLF_4, **g,** RPF_9 and **h,** GLF_18. Comparison between the design (color) and target fold (gray), solution behavior by SEC-MALS, CD spectra at different incubation temperatures and melting temperature profiles by CD.

Next, we attempted to crystallize the soluble analogues of membrane topologies and obtained high resolution X-ray structures for one claudin, one rhomboid protease, and two GPCR designed folds (Fig. 5, Supp Fig. S15). The CLF_4 structure exhibited exceptional design precision in both backbone and sidechains, as indicated by the backbone $RMSD_{C\alpha}$ of 0.73 Å and full-atom $RMSD_{fa}$ of 1.28 Å (Fig. 5a). The comparison between CLF_4 and the native claudin revealed accurate secondary structural element positioning, with an $RMSD_{C\alpha}$ of 3.63 Å, where most of the deviation arises from the β-sheet region (Fig. 5b, Supp Fig. S16). Additionally, the four helices are mostly hydrophilic, as evidenced by the low lipophilicity potential on the surface (Fig. 5c). In the case of the rhomboid protease design RPF_9, we observed high accuracy between the X-ray structure and the design model, with a backbone $RMSD_{C\alpha}$ of 0.97 Å and full-atom $RMSD_{fa}$ of 1.83 Å. However, the structural similarity was significantly lower when compared to the target fold, as indicated by the backbone $RMSD_{C\alpha}$ of 5.67 Å (Fig. 5d, 5e). Specifically, large structural deviations in the first extracellular loop (Supp Fig. S17) were observed, which could be due to its native positioning in the water/membrane interface[60]. The designed RPF_9 fold showed significantly increased hydrophobicity on the transmembrane surface when compared to the native RPF fold (Fig. 5f). Lastly, the designed GLFs preserved the canonical seven-helical bundle characteristic of native GPCRs. In terms of structural accuracy, the crystal structures were in very good agreement with the design models, 1.05 Å and 0.88 Å of $RMSD_{C\alpha}$ for GLF_18 and GLF_32, respectively (Supp Fig. S15, Fig. 5g, 5h). This accuracy further extended to the side-chain level, where comparisons of crystal structures vs the design models for GLF_18 and GLF_32 showed 1.54 and 1.40 Å full-atom $RMSD_{fa}$s, respectively. Comparing the structures of the soluble analogues to that of the reference native GPCR, the overall backbone $RMSD_{C\alpha}$s were 3.08 and 3.51 Å for GLF_18 and GLF_32, respectively (Supp Fig. S15, Fig. 5g, 5h). By analyzing the lipophilicity potential at the surface[60] of the designed proteins we can observe a clear transition from an initially hydrophobic surface to a hydrophilic one (Fig. 5i, Supp Fig. S15). Additionally, much of the sequence signatures of the GPCR fold were removed, such as the evolutionary conserved DRY motif in the first intracellular loop[64], the (N/D)PxxY motif in the seventh helix[65] and the transmembrane proline rich domains[66,67] (Supp Fig. S18). This demonstrates that by design one can explore very diverse sequence spaces while

removing potential evolutionary biases. Interestingly, at the structural level we observe that the irregular local structure of the terminal helix is preserved, while the intracellular segment of the GPCR fold exhibits notable structural deviation in GLF_18 (Supp Fig. S19). Our results demonstrate that integral membrane folds can be successfully designed in soluble form, hinting that these topologies share many of the same designability principles and constraints of folds present in the soluble proteome.

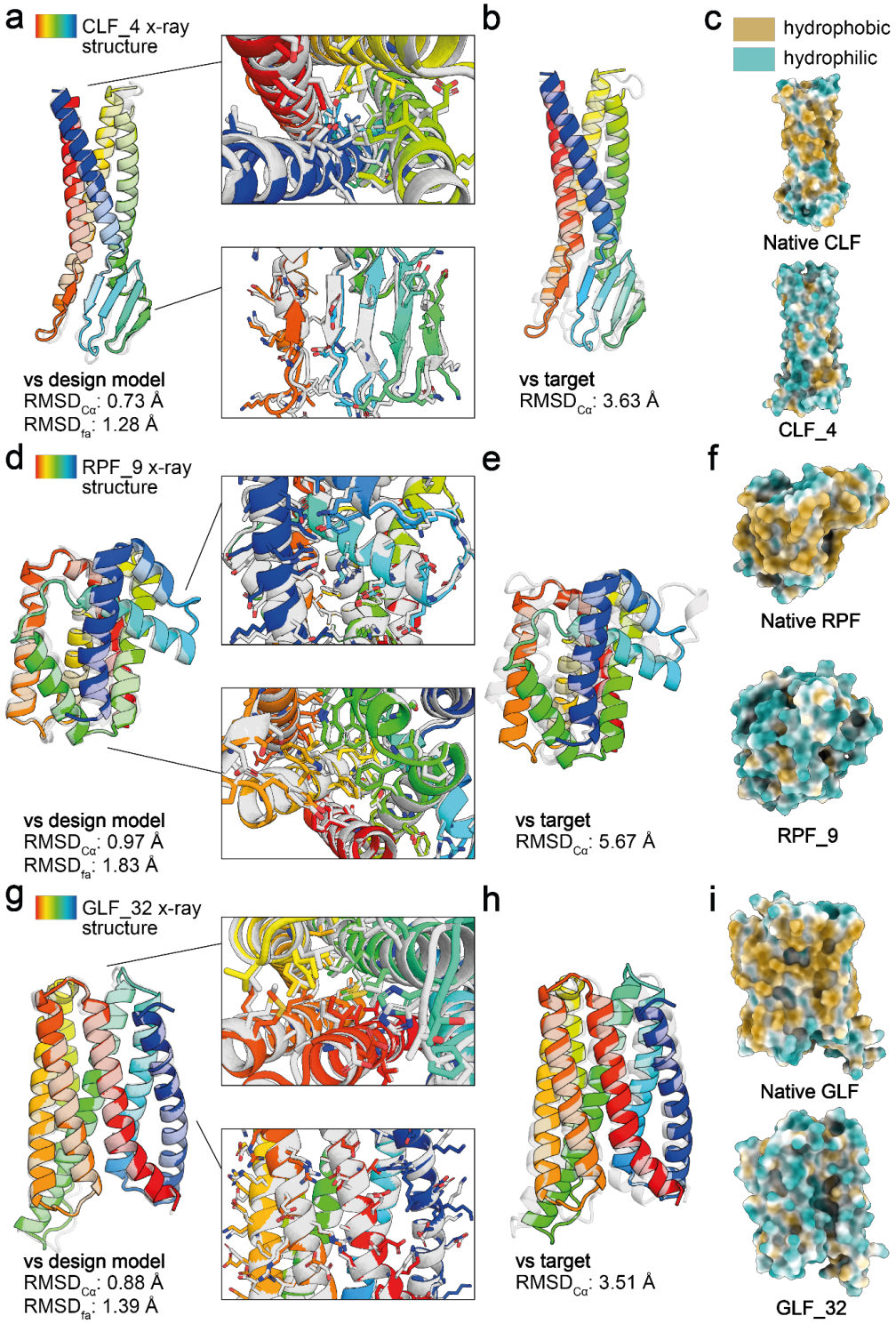

Fig. 5 | Soluble analogues of membrane proteins solved by X-ray crystallography. a, X-ray structure of CLF_4 (colored) superimposed on the design model (gray). **b,** X-ray structure of CLF_4 (colored) superimposed on the design model (gray). **c,** Molecular lipophilicity potential of the surface of the claudin design target and the soluble design CLF_4. **d,** X-ray structure of RPF_9 (colored) superimposed on the design model (gray). **e,** X-ray structure of RPF_9 (colored) superimposed on the design model (gray). **f,** Molecular lipophilicity potential of the surface of the rhomboid protease design target and the soluble design RPF_9. **g,** X-ray structure of GLF_32 (colored) superimposed on the design model (gray). **h,** X-ray structure

of GLF_32 (colored) superimposed on the design model (gray). **i,** Molecular lipophilicity potential of the surface of the GPCR design target and the soluble GLF_32 design. After redesign of the original membrane folds with $MPNN_{sol}$, the hydrophobicity (yellow) of the surface is significantly reduced, and polarity increased (blue). $RMSD_{C\alpha}$ - root mean square deviation computed over the Cα atoms of the backbone. $RMSD_{fa}$ - root mean square deviation computed over all the atoms in the structure.

Conclusions

The robust computational design of novel protein folds remains a difficult endeavor. Here, we present a computational approach based on deep learning that through the generation of high-quality protein backbones enables the efficient search of non-natural sequences for a variety of protein topologies. The computational framework based on $AF2_{seq}$-MPNN is flexible and generalizable, avoiding the need to perform fold-specific re-training or providing tedious empirical restraints for fold-conditioning. We designed and characterized several folds that have been very challenging to engineer with previous methods, achieving high experimental success rates in terms of soluble and folded designs. Structural characterization of the designs showed an exquisite accuracy of the computational models, both in terms of overall fold, as well as in the fine details of the side chain conformations, which are critical for the design of function. Additionally, we aimed to test the capability of the computational approach to provide the means to expand the soluble fold space and enable the design of analogues of protein topologies only found in membrane environments. By allowing full sequence design, we designed three different membrane fold analogs, including two with highly elaborate helical topologies (rhomboid protease and GPCR) and showed that such designs were folded and monomeric in solution. The experimental structures showed once again that the design method was very accurate, and that we recreate soluble analogues both for the rhomboid protease fold, as well as for the canonical 7 helical GPCR fold, which are not present in the soluble fold space. By doing so, we showed that membrane protein folds generally follow the same design principles as soluble protein folds, and that many of such folds can be readily designed in soluble form.

Moreover, we hypothesize that this could promote the designability of functional proteins by accessing a new plethora of folds that are not present in the soluble fold space, as well as by exporting certain functions exclusive to membrane proteins into the soluble proteome, such as ligand binding recognition and conformational signal transduction. Another exciting perspective is the creation of soluble analogues of membrane proteins

that retain many of the native features of the original membrane proteins, such as enzymatic or transport functions, which could greatly accelerate the study of their function in more biochemically accessible soluble formats. Similarly, this would be critical to facilitate the development of novel drugs and therapies that target this challenging class of proteins, which remain one of the most important drug targets. In summary, we present a deep learning approach for computational protein design that demonstrates the usefulness of high-quality structure representations to enable the effective exploration of new sequence spaces that can yield viable proteins and contribute to the expansion of the designable fold space that can ultimately impact our ability to design functional proteins.

Materials and Methods

AF2seq design protocol

Design target preparation. The design target structures were sourced from the Protein Data Bank (PDB) and included the following protein folds: Ig-like fold (3SD2), β-barrel fold (6D0T)[36], TIM-barrel (5BVL)[69], claudin (4P79)[48], rhomboid protease (3ZEB)[54], and GPCR (6FFI). Due to missing residue positions in the TIM-barrel, claudin, and GPCR X-ray structures, we used AlphaFold2 to predict the protein structure using the X-ray structure as a template. The claudin target contained a disordered region (residues 34 to 40) that we replaced with a three glycine linker and the GPCR target contained a large disordered region (residues 875 to 896) which we replaced with a five glycine linker. Additionally, the GPCR X-ray target contained an endolysin domain that was inserted into one of the loops to enhance protein solubility for crystallization (residues 679 to 838). Therefore, we predicted the GPCR sequence using the X-ray target as a template.

Loss function. For the computation of error gradients, a composite loss function was used:

$$loss = W_{fape}L_{fape} + W_{dist}L_{dist} + W_{pLDDT}L_{pLDDT} + W_{pTM}L_{pTM}$$

The loss function is represented as a combination of L, which denotes the value of the loss, and W, which denotes the weight of the loss. The Frame Aligned Point Error (FAPE) loss quantifies the L2 norm between the predicted C_α atoms and the target structure. The distogram (dist) loss is the cross entropy over the C_β distogram for non-glycine residues and the C_α distance in the case of glycine. The model confidence (pLDDT) loss of the C_α positions is computed by taking 1 – pLDDT, penalizing low confidence. And finally, the pTM-score (pTM) loss, which is a prediction confidence metric focused on global structural similarity. In this paper, the designs were tested with varying weights of the losses, encompassing a range of W_{fape} = 0-1.0, W_{dist} = 0-1.0, W_{pLDDT} = 0-0.3 and W_{pTM} = 0-0.3.

Gradient descent. As previously described by Goverde et al.[70], the amino acid sequences were initialized based on the secondary structure of the target fold. The secondary structure assignments were then encoded in sequences using alanines for helix, valines

for β-sheet and glycines for loop residues. This introduces a bias towards the correct local structure, aiding faster convergence of the design trajectories. In order to diversify the generated designs, 10% of the amino acids are randomly mutated in the initial sequence of each design trajectory. Subsequently, the sequence was passed through the AF2 networks, which generated five structures. These structures were then utilized to calculate the loss using the previously defined loss function. The error gradient was obtained by backpropagating the errors to the one-hot-encoded input, resulting in a 5x20xN error gradient, where N represents the sequence length. We then take the average of the five matrices to obtain the mean error gradient (20xN) which is used for gradient descent. A Position Specific Scoring Matrix (PSSM) of 20xN is updated using the ADAM optimizer[5] with the normalized error gradient. Following the update, the PSSM undergoes a softmax function that transforms the matrix into a probability distribution of the amino acid identity for each position. The argmax function is subsequently used to determine the most probable amino acid identities per position, which are then utilized to construct the new input sequence for the next iteration. The cysteine residues in the PSSM are masked so the designed sequences do not contain any cysteines.

Model settings. AF2 was run in single sequence mode using the network configurations of 'model_5_ptm' with MSAs and templates disabled. For the design trajectories, we used zero recycles, meaning that each AF2 network was only executed once. All design runs were executed on a single Nvidia Tesla V100 (32GB) GPU.

Computational design protocol. In each AF2-sequence design trajectory 500 rounds of gradient descent optimization were performed (https://github.com/bene837/af2seq). Not all design trajectories of the claudin, rhomboid protease and GPCR converged. Hence, we sampled sequences from successful trajectories and introduced mutations. These sequences were then used as starting points for new design trajectories, we named soft-starts, which resulted in a higher convergence rate. All generated sequences were then predicted using AF2 with three recycles, followed by relaxation in an AMBER force-field[71]. This results in high quality structures which were used as an input for proteinMPNN sequence generation.

Training of SolubleMPNN ($MPNN_{sol}$)

The $MPNN_{sol}$ model was trained on protein assemblies in the PDB (as of Aug 02, 2021) determined by X-ray crystallography or cryoEM to better than 3.5 Å resolution and with less than 10,000 residues. We followed training as described in the work by Daupauras et al.[72] with the only modification of excluding annotated transmembrane PDB codes. The list of excluded PDBs and $MPNN_{sol}$ model weights are available here: https://github.com/dauparas/ProteinMPNN/tree/main/soluble_model_weights.

ProteinMPNN sequence redesign

The generated backbones by $AF2_{seq}$ were used as input to proteinMPNN. For the vanilla proteinMPNN we used the provided model weights trained on a dataset with 0.1 Å gaussian noise[24]. For the biased proteinMPNN (referred in the main text as $MPNN_{bias}$) we used a modified version of the script 'submit_example_8.sh' as provided on the proteinMPNN github mentioned above. We found the best results by giving a positive sampling bias to the polar amino-acids whilst giving a negative sampling bias to alanine. For soluble proteinMPNN ($MPNN_{sol}$) we generated sequences with two different models which had different levels of noise during training (0.1 Å and 0.2 Å). For all proteinMPNN models we generated two sequences per $AF2_{seq}$ designed backbone. No gaussian noise was added to the input backbone and cysteine residues were masked during the decoding process.

Structural similarity calculations

The C_α atoms of the structures were aligned using the super imposer from the Biopython package[24]. The Root Mean Square Deviation of the C_α atoms ($RMSD_{C\alpha}$) is calculated as the mean Euclidean distance between predicted and target C_α atom coordinates. The full atom RMSD ($RMSD_{fa}$) is calculated by first aligning all atoms with the superimposer after which the mean Euclidean distance between atoms is computed. The Template Modeling (TM) scores were determined using TM-align[73].

Sequence diversity analysis

Sequence recovery is quantified as the number of positions at which the corresponding residue matches the residue in the target fold divided by the total number of residues in

the sequence multiplied by 100%. The e-values were obtained through a protein BLAST search of the NCBI RefSeq database of 10/1/2022 with a maximum hit value of 1000.

Surface hydrophobicity calculations.

The fraction of surface hydrophobics were calculated using Rosetta[74]. First, all surface residues were identified using the layer selector; these residues are defined as residues with Solvent Accessible Surface Area (SASA) > 40 $Å^2$. Of these surface residues we count the number of apolar amino acids (defined as "GPAVILMFYW") and divide it by the total number of surface residues.

Design filtering and selection.

All generated sequences were predicted with AF2 using three recycles and a relaxation step in an AMBER force-field. Next the sequences were filtered using the following criteria: I) TM-score > 0.80 for all designs except the rhomboid protease. The rhomboid protease yielded slightly lower TM-scores in the design trajectory hence we chose a cutoff value 0.75 instead; II) pLDDT > 80 for all designs except the rhomboid protease (pLDDT > 75); III) an e-value threshold > 0.1 for sequence novelty.

Structural fold similarity search

The fold similarity search was performed using FoldSeek[1] on the SCOP-database[75] (downloaded March 2023). For each of the design target folds, an exhaustive search based on TM-score alignment was performed. The SCOP-database contains globular and membrane domain annotations which were used for the hit classification.

Fold complexity calculations

Relative Contact Order (RCO) is calculated at the secondary structure level by computing the residue distance in the sequence between secondary structures for all pairs within 8 Å of each other, and then averaging these distances for all contacts that are more than 4 residues apart. To ensure consistency in secondary structure annotations across all structures, we employ DSSP for the determinations of secondary structural elements[21]. The de novo protein dataset comprises 70 helical proteins, 6 beta sheet proteins, and 42 proteins containing both alpha helices and beta sheets[41,68]. Meanwhile, the natural protein

dataset consists of 1,000 proteins, randomly selected from the entire collection of proteins in the CATH dataset (version 4.3)[76].

Molecular dynamics-based backbone perturbation

All molecular dynamics (MD) simulations were performed using the Groningen Machine for Chemical Simulations (GROMACS) 2021.4 software package[77]. For the unbiased simulations in water, the Amber ff99SB-ILDN force field for proteins[78] was used. The simulations were carried out under *NPT* conditions with a leapfrog integration scheme and a time step of 2.0 fs. Rhombic dodecahedron (triclinic) periodic boundary conditions were applied and the TIP3P[78] water model was used as solvent. Temperature coupling using stochastic velocity rescaling[79] to two separate temperature baths for the protein and for the water solvent was applied with a reference temperature of 300 K and a relaxation time of 0.1 ps. The pressure was coupled isotropically to a Parrinello-Rahman barostat[80] at 1.0 bar with a coupling constant of 2.0 ps and an isothermal compressibility of 0.45 nm^2 N^{-1}. For both the short-range electrostatic- and van der Waals interactions, a single cutoff distance of 0.9 nm was used. The long-range electrostatics were calculated by the particle mesh Ewald (PME) algorithm[81] with a Fourier spacing of 0.12 nm. The linear constraint solver (LINCS) algorithm[82] was used to impose constraints on the bond lengths with fourth order expansion. Preceding the simulations, the solvated protein structures were energy minimized with a steepest descent algorithm, until the maximum force was below 100 kJ mol^{-1} nm^{-1}. For the unbiased simulations in the POPC lipid bilayer (a mimic for a cellular membrane), the GROMOS 54A8 force field[83] was used in combination with lipid parameters from the 1-Palmitoyl-2-oleoylphosphatidylcholine (POPC) model of Marzuoli *et al.*[84]. The simulations were carried out under *NPT* conditions with a leapfrog integration scheme and a time step of 2 fs. Rectangular periodic boundary conditions were applied and the SPC water model[85] was used as solvent. Nose-Hoover temperature coupling[86,87] was applied to two separate temperature baths at 323 K, one for the protein and the POPC bilayer and one for the solvent, with a relaxation time of 0.5 ps. The pressure was coupled semi-isotropically to a Parrinello-Rahman barostat[80] at 1.0 bar with a coupling constant of 2.0 ps and an isothermal compressibility of 0.45 nm^2 N^{-1} was applied. For both the short-range electrostatic and van der Waals interactions, a single cutoff of 1.2 nm was used. The long-range electrostatics were treated using the PME

algorithm[81] with a Fourier spacing of 0.16 nm. The bond lengths were constrained with the linear constraint solver (LINCS) algorithm[82]. The simulation systems contained 512 POPC molecules in a bilayer (256 per leaflet). These lipids were packed around the embedded protein structures in the center of the simulation box (2.4963 nm x 13.0172 nm x 13.7189) using the InflateGRO methodology [KANDT2007475], as described by Lemkul[88]. Center-of-mass (COM) motion removal was applied in every simulation step to remove the motion of the bilayer and protein relative to the solvent. Preceding the simulations, the solvated simulation systems were energy minimized with a steepest descent algorithm, until the maximum force was below 1000 kJ mol^{-1} nm^{-1}. For all simulations, the energy minimization was followed by 100 ps *NVT* thermalisation and 1.0 ns *NPT* equilibration. After equilibration, initial velocities were generated using a random number generator. Three unbiased MD runs were performed for 11 ns each and the first 1.0 ns of each simulation was discarded, resulting in 30 ns of sampling for each starting structure. Trajectory frames were extracted every 100 ps.

Design solubility screen

Designs were synthesized and cloned into the pET21 vector with C-terminal His_6-tag by GenScript using the NheI and XhoI sites. Designs were transformed into *E. coli* BL21 DE3 cells and expressed overnight with 1 mM IPTG at 18 ºC in 1 ml LB medium supplemented with 100 µg/ml ampicillin in a 96-well format. After expression, cells were chemically lysed, cell debris was pelleted, and the supernatant was incubated with 50 µl of equilibrated Ni-NTA agarose (QIAGEN) beads. After incubation, unbound supernatant is discarded, beads are washed 5 times with 50 mM Tris-HCl pH 7.5, 500 mM KCl, 15 mM imidazole, and designs are eluted using buffer containing 50 mM Tris-HCl pH 7.5, 500 mM KCl, 400 mM imidazole. Elutions are analyzed using SDS-PAGE and designs are denoted soluble if the appropriate protein band is clearly visible following Coomassie staining.

Protein expression, purification and characterization

Designed proteins were expressed in *E. coli* BL21 (DE3) (Novagen) for 16 h at 18 °C. Bacterial pellets were resuspended and sonicated in 50 mM Tris-HCl pH 7.5, 500 mM KCl, 15 mM imidazole buffer supplemented with 1 mg/ml lysozyme, 1 mM PMSF, and 1

µg/ml DNase. Cell lysates were clarified using ultracentrifugation and loaded on a 10 ml Ni-NTA Superflow column (QIAGEN) and washed with 7 column volumes of 50 mM Tris-HCl pH 7.5, 500 mM NaCl, 10 mM imidazole. Designs were eluted with 10 column volumes of 50 mM Tris-HCl pH 7.5, 500 mM NaCl, 500 mM imidazole. Main protein fractions were concentrated and injected onto a Superdex 75 16/600 gel filtration column (GE Healthcare) in PBS buffer. Protein fractions were concentrated, flash frozen in liquid nitrogen and stored at -80 °C. Molar mass and homogeneity was confirmed using SEC-MALS. Folding, secondary structure content, and melting temperatures were assessed using circular dichroism in a Chirascan V100 instrument from Applied Photophysics.

Protein crystallization and structure determination

The TIM barrel TBF_24 design was crystallized in the $P2_12_12_1$ space group using the sitting drop vapor diffusion setup at 4 °C in 0.1 M Na_3 Citrate pH 4.0, 1 M LiCl, 20% PEG 6000 buffer. The Claudin-like CLF_4 design was crystallized in the P1 space group using the sitting drop vapor diffusion method at 4 °C in 0.1 M Na_3 Citrate pH 5.0, 0.1 M Na/K phosphate pH 5.5, 0.1 M RbCl, 25% v/v PEG smear medium (BCS Screen, Molecular Dimensions). The Rhomboid protease-like RPF_9 design was crystallized in the $P2_12_12_1$ space group using the sitting drop vapor diffusion setup at 4 °C in 0.1 M HEPES pH 7.8, 0.15 M Na_3 Citrate dihydrate, 25% v/v PEG smear low (BCS Screen, Molecular Dimensions). The GPCR-like GLF_18 design was crystallized in the P1 space group using the sitting drop vapor diffusion setup at 4 °C in Na Phosphate-Citrate pH 4.2, 0.2 M LiSO4, 20% PEG 1000 buffer. The GPCR-like GLF_32 design was crystallized in the $P2_12_12_1$ space group using the sitting drop vapor diffusion setup at 4 °C in 0.1 M Na Acetate pH 5.5, 0.2 M KBr, 25% PEG MME 2000 buffer. Crystals were cryoprotected in the crystallization buffer supplemented with 20% glycerol and flash-cooled in liquid nitrogen. Diffraction data was measured at the beamline PXI (X06SA) of the Swiss Light Source (Paul Scherrer Institute, Villigen, Switzerland) and the MASSIF-1 beamline of the The European Synchrotron Radiation Facility, Grenoble, France at a temperature of 100 K. Data was processed using the autoPROC package[89], and phases were obtained by molecular replacement using the Phaser module of the Phenix package[90]. Atomic model adjustment and refinement was completed using COOT[89] and Phenix.refine[91]. The

quality of refined models was assessed using MolProbity[92]. Structural figures were generated using PyMOL (Schrödinger, LLC; https://www.pymol.org/) and ChimeraX.

Acknowledgments:

We thank SCITAS at EPFL for support in running design trajectories. We thank Beat Blattmann (Protein Crystallization Center, UZH, Zurich, Switzerland) for performing initial crystallization screens. We thank Florence Pojer, Kelvin Lau, and Amédé Larabi (Protein Production and Structure Characterization Core Facility, EPFL, Switzerland) for help with crystal optimization and data processing. We thank Takashi Tomizaki (Swiss Light Source, X06SA beamline, Paul Scherrer Institute, Villigen, Switzerland), and Didier Nurizzo (The European Synchrotron Radiation Facility, MASSIF-1 beamline, Grenoble, France) for assistance with crystallographic data collection.

Funding:

MP was supported by the Peter und Traudl Engelhorn Stiftung. BEC was supported by the Swiss National Science Foundation, the NCCR in Chemical Biology, the NCCR in Molecular Systems Engineering, and the ERC Starting grant no. 716058. JD was supported by a gift from Microsoft and DB was supported by the Howard Hughes Medical Institute. SO and SK were supported by NIH DP5OD026389, NSF MCB2032259, and Amgen.

Author contributions:

Conceptualization: CAG, MP, BEC
Methodology: CAG, MP, SK, JD, DB, SO, BEC
Investigation: CAG, MP, LJD, SG, SR, CS, SK, JD
Visualization: CAG, MP, LJD, JD, BEC
Funding acquisition: DB, SO, BEC
Project administration: CAG, MP, BEC
Supervision: DB, SO, BEC
Writing – original draft: CAG, MP, LJD, BEC
Writing – review & editing: CAG, MP, LJD, SK, JD, DB, SO, BEC

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: All data are available in the main text or as supplementary materials. Atomic coordinates and structure factors of the reported X-ray structures have been deposited in the Protein Data Bank under accession numbers 8OYS (TBF_24), 8OYV (CLF_4), 8OYW (RPF_9), 8OYX (GLF_18), and 8OYY (GLF_32). Af2seq code is available at https://github.com/bene837/af2seq. ProteinMPNN, along with soluble trained weights is available at https://github.com/dauparas/ProteinMPNN.

References

1. Leman, J. K. *et al.* Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat Methods* **17**, 665–680 (2020).
2. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
3. Baek, M. et al. Accurate prediction of protein structures and interactions using a 3-track neural network. *Sci New York N Y* **373**, 871–876 (2021).
4. Huang, B. et al. A backbone-centred energy function of neural networks for protein design. Nature **602**, 523–528 (2022).
5. Goverde, C., Wolf, B., Khakzad, H., Rosset, S. & Correia, B. E. De novo protein design by inversion of the AlphaFold structure prediction network. *Biorxiv* 2022.12.13.520346 (2022) doi:10.1101/2022.12.13.520346.
6. Lin, Y. & AlQuraishi, M. Generating Novel, Designable, and Diverse Protein Structures by Equivariantly Diffusing Oriented Residue Clouds. (2023).
7. Frank, C. et al. Efficient and scalable de novo protein design using a relaxed sequence space. *bioRxiv* (2023) doi:10.1101/2023.02.24.529906.
8. Anishchenko, I. et al. De novo protein design by deep network hallucination. Nature **600**, 547–552 (2021).
9. Wicky, B. I. M. et al. Hallucinating symmetric protein assemblies. *Science* **0**, eadd1964 (2022).
10. Watson, J. L. et al. Broadly applicable and accurate protein design by integrating structure prediction networks and diffusion generative models. bioRxiv 2022.12.09.519842 (2022) doi:10.1101/2022.12.09.519842.
11. Wang, J. et al. Scaffolding protein functional sites using deep learning. Science **377**, 387–394 (2022).
12. Gainza, P. et al. De novo design of protein interactions with learned surface fingerprints. Nature **617**, 1–9 (2023).
13. Ingraham, J., Garg, V., Barzilay, R. & Jaakkola, T. Generative models for graph-based protein design. *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)* **32**, 15741–15752.
14. Dauparas, J. et al. Robust deep learning–based protein sequence design using

ProteinMPNN. Science eadd2187 (2022) doi:10.1126/science.add2187.
15. Anand, N. et al. Protein sequence design with a learned potential. *Nat Commun* **13**, 746 (2022).
16. Yim, J. et al. SE(3) diffusion model with application to protein backbone generation. *Arxiv* (2023).
17. Wu, K. E. et al. Protein structure generation via folding diffusion. (2022).
18. Trippe, B. L. et al. Diffusion probabilistic modeling of protein backbones in 3D for the motif-scaffolding problem. *arXiv* (2022).
19. Anand, N. & Achim, T. Protein Structure and Sequence Generation with Equivariant Denoising Diffusion Probabilistic Models. arXiv (2022).
20. Norn, C. et al. Protein sequence design by conformational landscape optimization. *P Natl Acad Sci Usa* **118**, e2017228118 (2021).
21. Andreeva, A., Kulesha, E., Gough, J. & Murzin, A. G. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Res* **48**, D376–D382 (2019).
22. Moffat, L., Greener, J. G. & Jones, D. T. Using AlphaFold for Rapid and Accurate Fixed Backbone Protein Design. Biorxiv 2021.08.24.4575**49** (2021) doi:10.1101/2021.08.24.457549.
23. Jendrusch, M., Korbel, J. O. & Sadiq, S. K. AlphaDesign: A de novo protein design framework based on AlphaFold. Biorxiv 2021.10.11.463937 (2021) doi:10.1101/2021.10.11.463937.
24. Dauparas, J. et al. Robust deep learning–based protein sequence design using ProteinMPNN. Science 0, eadd2187 (2022).
25. Smith, C. A. & Kortemme, T. Backrub-Like Backbone Simulation Recapitulates Natural Protein Conformational Variability and Improves Mutant Side-Chain Prediction. *J Mol Biol* **380**, 742–756 (2008).
26. Woof, J. M. & Burton, D. R. Human antibody–Fc receptor interactions illuminated by crystal structures. *Nat Rev Immunol* **4**, 89–99 (2004).
27. Halaby, D. M., Poupon, A. & Mornon, J.-P. The immunoglobulin fold family: sequence analysis and 3D structure comparisons. *Protein Eng Des Sel* **12**, 563–571 (1999).
2**8**. Williams, A. F. & Barclay, A. N. The Immunoglobulin Superfamily—Domains for *Cell* Surface Recognition. *Annu Rev Immunol* **6**, 381–405 (1988).
29. Jost, C. & Plückthun, A. Engineered proteins with desired specificity: DARPins, other alternative scaffolds and bispecific IgGs. *Curr Opin Struc Biol* **27**, 102–112 (2014).
30. Kintzing, J. R., Interrante, M. V. F. & Cochran, J. R. Emerging Strategies for Developing Next-Generation Protein Therapeutics for Cancer Treatment. *Trends Pharmacol Sci* **37**, 993–1008 (2016).
31. Hecht, M. H. De novo design of beta-sheet proteins. *Proc National Acad Sci* **91**, 8729–8730 (1994).
32. Bork, P., Holm, L. & Sander, C. The Immunoglobulin Fold Structural Classification,

Sequence Patterns and Common Core. J Mol Biol **242**, 309–320 (1994).
33. Marcos, E. et al. De novo design of a non-local β-sheet protein with high stability and accuracy. *Nat Struct Mol Biol* **25**, 1028–1034 (2018).
34. Chidyausiku, T. M. et al. De novo design of immunoglobulin-like domains. Nat Commun 13, 5661 (2022).
35. Weber, B. et al. A single residue switch reveals principles of antibody domain integrity. *J Biol Chem* **293**, 17**107**–17118 (2018).
36. Dou, J. et al. De novo design of a fluorescence-activating β-barrel. Nature **561**, 485–491 (2018).
37. Vorobieva, A. A. et al. De novo design of transmembrane β barrels. Science **371**, (2021).
38. Kipnis, Y. et al. Design and optimization of enzymatic activity in a de novo β-barrel scaffold. *Protein Sci* **31**, (2022).
39. Tinberg, C. E. et al. Computational design of ligand-binding proteins with high affinity and selectivity. Nature **501**, 212–216 (2013).
40. Bick, M. J. et al. Computational design of environmental sensors for the potent opioid fentanyl. *Elife* **6**, e28909 (2017).
41. Verkuil, R. et al. Language models generalize beyond natural proteins. bioRxiv 2022.12.21.52**15**21 (2022) doi:10.**1101**/2022.12.21.521521.
42. Gerlt, J. A. New wine from old barrels. *Nat Struct Biol* **7**, 171–173 (2000).
43. Sterner, R. & Höcker, B. Catalytic Versatility, Stability, and Evolution of the (βα)8-Barrel Enzyme Fold. *Chem Rev* **105**, 4038–4055 (2005).
44. Röthlisberger, D. et al. Kemp elimination catalysts by computational enzyme design. Nature **453**, 190–195 (2008).
45. Romero-Romero, S., Kordes, S., Michel, F. & Höcker, B. Evolution, folding, and design of TIM barrels and related proteins. Curr Opin Struc Biol **68**, 94–104 (2021).
46. Houbrechts, A. et al. Second-generation octarellins: two new de novo (beta/alpha)8 polypeptides designed for investigating the influence of beta-residue packing on the alpha/beta-barrel structure stability. Protein Eng 8, 249–59 (1995).
47. Nagarajan, D., Deka, G. & Rao, M. Design of symmetric TIM barrel proteins from first principles. *Bmc Biochem* **16**, 18 (2015).
48. Huang, P.-S. et al. De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. *Nat Chem Biol* **12**, 29–34 (2016).
49. Chu, A. E., Fernandez, D., Liu, J., Eguchi, R. R. & Huang, P.-S. De Novo Design of a Highly Stable Ovoid TIM Barrel: Unlocking Pocket Shape towards Functional Design. *Biodesign Res* **2022**, 1–13 (2022).
50. Slovic, A. M., Kono, H., Lear, J. D., Saven, J. G. & DeGrado, W. F. Computational design of water-soluble analogues of the potassium channel KcsA. Proc National Acad Sci 101, 1828–1833 (2004).
51. Mitra, K., Steitz, T. A. & Engelman, D. M. Rational design of `water-soluble' bacteriorhodopsin variants. Protein Eng Des Sel 15, 485–492 (2002).

52. Perez-Aguilar, J. M. et al. A Computationally Designed Water-Soluble Variant of a G-Protein-Coupled Receptor: The Human Mu Opioid Receptor. *Plos One* **8**, e66009 (2013).
53. Li, H., Cocco, M. J., Steitz, T. A. & Engelman, D. M. Conversion of Phospholamban into a Soluble Pentameric Helical Bundle †. *Biochemistry-us* **40**, 6636–6645 (2001).
54. Suzuki, H. et al. Crystal Structure of a Claudin Provides Insight into the Architecture of Tight Junctions. Science **344**, **30**4–307 (2014).
55. Tichá, A., Collis, B. & Strisovsky, K. The Rhomboid Superfamily: Structural Mechanisms and Chemical Biology Opportunities. *Trends Biochem Sci* **43**, 726–739 (2018).
56. Katritch, V., Cherezov, V. & Stevens, R. C. Structure-Function of the G Protein–Coupled Receptor Superfamily. *Annu Rev Pharmacol* **53**, 531–556 (2013).
57. Suzuki, H., Tani, K. & Fujiyoshi, Y. Crystal structures of claudins: insights into their intermolecular interactions. *Ann Ny Acad Sci* **1397**, 25–34 (2017).
58. Li, J. Targeting claudins in cancer: diagnosis, prognosis and therapy. *Am J Cancer Res* **11**, 3406–3424 (2021).
59. Vonrhein, C. et al. Data processing and analysis with the autoPROC toolbox. *Acta Crystallogr Sect D Biological Crystallogr* **67**, 293–302 (2011).
60. Wang, Y., Zhang, Y. & Ha, Y. Crystal structure of a rhomboid family intramembrane protease. Nature **444**, 179–180 (2006).
61. Urban, S., Lee, J. R. & Freeman, M. Drosophila Rhomboid-1 Defines a Family of Putative Intramembrane Serine Proteases. Cell 107, 173–182 (2001).
62. Hauser, A. S. et al. Pharmacogenomics of GPCR Drug Targets. Cell **172**, 41-54.e19 (2018).
63. Hauser, A. S., Attwood, M. M., Rask-Andersen, M., Schiöth, H. B. & Gloriam, D. E. Trends in GPCR drug discovery: new agents, targets and indications. *Nat Rev Drug Discov* **16**, 829–842 (2017).
64. LAGUERRE, M., SAUX, M., DUBOST, J. P. & CARPY, A. MLPP: A Program for the Calculation of Molecular Lipophilicity Potential in Proteins. *Pharm Pharmacol Commun* **3**, 217–222 (1997).
65. Rovati, G. E., Capra, V. & Neubig, R. R. The Highly Conserved DRY Motif of Class A G Protein-Coupled Receptors: Beyond the Ground State. *Mol Pharmacol* **71**, 959–964 (2007).
66. Konvicka, K., Guarnieri, F., Ballesteros, J. A. & Weinstein, H. A Proposed Structure for Transmembrane Segment 7 of G Protein-Coupled Receptors Incorporating an Asn-Pro/Asp-Pro Motif. *Biophys J* **75**, 601–611 (1998).
67. Srinivasan, M. & Dunker, A. K. Proline Rich Motifs as Drug Targets in Immune Mediated Disorders. *Int J Peptides* **2012**, 634769 (2012).
68. Woolfson, D. N. A Brief History of De Novo Protein Design: Minimal, Rational, and Computational. J Mol Biol **433**, 167160 (2021).
69. Christopher, J. A. et al. Structure-Based Optimization Strategies for G Protein-

Coupled Receptor (GPCR) Allosteric Modulators: A Case Study from Analyses of New Metabotropic Glutamate Receptor 5 (mGlu5) X-ray Structures. *J Med Chem* **62**, 207–222 (2019).
70. Vosyka, O. et al. Activity-based probes for rhomboid proteases discovered in a mass spectrometry-based assay. Proc National Acad Sci 110, 2472–2477 (2013).
71. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. Arxiv (2014) doi:10.48550/arxiv.1412.6980.
72. Case, D. A. et al. The Amber biomolecular simulation programs. *J. Comput. Chem.* **26**, 1668–1688 (2005).
73. Cock, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
74. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33, 2302–2309 (2005).
75. Kempen, M. van et al. Foldseek: fast and accurate protein structure search. Biorxiv 2022.02.07.479398 (2022) doi:10.1101/2022.02.07.479398.
76. Sillitoe, I. et al. CATH: increased structural coverage of functional space. Nucleic Acids Res 49, D266–D**27**3 (2020).
77. Lindahl, E., Hess, B. & Spoel, D. van der. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J Mol Model* **7**, 306–317 (2001).
78. Lindorff-Larsen, K. et al. Improved side-chain torsion potentials for the Amber ff99SB protein force field: Improved Protein Side-Chain Potentials. *Proteins Struct Funct Bioinform* **78**, 1950–1958 (2010).
79. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J Chem Phys* **126**, 014101 (2007).
80. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *J Appl Phys* **52**, 7**18**2–7190 (19**81**).
81. Essmann, U. et al. A smooth particle mesh Ewald method. J Chem Phys **103**, 8577–8593 (1995).
82. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. LINCS: A linear constraint solver for molecular simulations. J. Comput. Chem. 18, 1463–1472 (1997).
83. Reif, M. M., Winger, M. & Oostenbrink, C. Testing of the GROMOS Force-Field Parameter Set 54A8: Structural Properties of Electrolyte Solutions, Lipid Bilayers, and Proteins. *J Chem Theory Comput* **9**, 1247–1264 (2013).
84. Marzuoli, I., Margreitter, C. & Fraternali, F. Lipid Head Group Parameterization for GROMOS 54A8: A Consistent Approach with Protein Force Field Description. J Chem Theory Comput **15**, 5175–5193 (2019).
85. Berendsen, H. J. C., Postma, J. P. M., Gunsteren, W. F. van & Hermans, J. Intermolecular Forces, Proceedings of the Fourteenth Jerusalem Symposium on Quantum Chemistry and Biochemistry Held in Jerusalem, Israel, April 13–16, 1981. in 331–342 (Springer Netherlands, 1981). doi:10.1007/978-94-015-7658-1_21.
86. Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Phys*

Rev A **31**, 1695–1697 (1985).
87. Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. J Chem Phys 81, 511–519 (1984).
88. Lemkul, J. From Proteins to Perturbed Hamiltonians: A Suite of Tutorials for the GROMACS-2018 Molecular Simulation Package [Article v1.0]. *Living J Comput Mol Sci* **1**, (2019).
89. Liebschner, D. et al. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr Sect D Struct Biology* **75**, 861–877 (2019).
90. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. Acta Crystallogr Sect D Biological Crystallogr 66, 486–501 (2010).
91. Williams, C. J. et al. MolProbity: More and better reference data for improved all-atom structure validation. Protein Sci 27, 293–315 (2018).
92. Pettersen, E. F. et al. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci 30, 70–82 (2021).