Expanding the Soluble β-Barrel Protein Space: Computational Design and Characterization of OmpW Analogues from *Escherichia coli and Vibrio cholerae*

By: Anastasia Ganshof van der Meersch

Supervisor: Ekaterina Pyatova

Professor: Bruno Correia

Bachelor semester 5 (2024-2025)

Laboratory of Protein Design and Immunoengineering (LPDI)

Swiss Federal Technology Institute of Lausanne (EPFL)

Table of Contents

Introduction

Proteins are biopolymers composed of amino acids that mediate the fundamental processes of life. Their unique amino acid sequences dictate the three-dimensional structures that govern their specific functions. Among the many protein folds, β-barrels stand out due to their distinctive cylindrical architecture. They form a closed barrel-shaped structure, stabilized by hydrogen bonding between β-strands. This fold exists in two biological contexts: soluble β-barrels and membrane-embedded β-barrels.

β-Barrel Proteins: Structure, Function, and Immunological Relevance

Soluble proteins adopt stable conformations in aqueous environments, whereas membrane proteins are structured to function within the lipid bilayer. Thus, membrane and soluble β-barrels differ in structure, function, and localization due to their environments. Membrane β-barrels are wider (8–36 strands) than soluble β-barrels (6–12 strands). To minimize the exposure of polar backbone groups to the hydrophobic lipid bilayer, membrane β-barrels maximize intra-strand hydrogen bonding and thus have a more vertical β-strand arrangement than soluble β-barrels. In lipid bilayers, β-barrels span the membrane, requiring greater height and increased amino acid content per strand. Unlike soluble β-barrels, which bury hydrophobic residues inside and present hydrophilic residues on the outside, membrane β-barrels interact with the lipid bilayer through hydrophobic residues on its exterior while keeping hydrophilic residues inward. Their amino acid composition also differs. Membrane β-barrels are rich in glycine and alanine for flexibility, while soluble β-barrels contain more lysine and glutamate to interact with water [2].

These structural differences directly influence function. Membrane β-barrels serve in small molecule transport, structural support, surface binding, and enzymatic catalysis. Their increased strand count often allows them to function as pores for polar solutes. Soluble β-barrels perform diverse tasks such as light emission, ligand binding, and enzymatic catalysis, as seen in proteins like green fluorescent protein, lipocalins, avidins, and allene oxide cyclase [2].

Beyond structure and function, localization further defines the roles of β-barrels. Soluble β-barrels are typically found in the cytoplasm or extracellular space, while membrane β-barrels are found in the outer membranes of Gram-negative bacteria, mitochondria, and chloroplasts[2]. In bacteria, membrane β-barrel proteins often facilitate infection and trigger an immune response. Exposed loops on the bacterial surface can serve as epitopes recognized by the host immune system.

When antigen receptors on B-cells and T-cells bind these epitopes, they trigger an immune response, leading to antibody production and pathogen elimination, thereby protecting the host from infection [3]. One membrane β-barrel that draws attention as a

potential target for vaccine development is OmpW, an 8-strand protein that is highly immunogenic and found in all known *Vibrio cholerae (V. cholerae)* strains and *Escherichia coli* (*E. coli*).

Computational Tools for Protein Engineering: From Structure to Sequence

De novo protein design focuses on creating new proteins with sequences unrelated to natural ones. While the design space is immense, it contains few functional sequences[4], posing a challenge in efficiently navigating the fitness landscape. Approaches to tackle this include combinatorial libraries and computationally intensive tools like Rosetta[5]. Recently, machine learning has transformed the field, making the design of functional *de novo* proteins more efficient and user-friendly[6,7].

Two major advances have shaped the current state of the protein design field. In 2020, AlphaFold2 (AF2) solved the sequence-to-structure problem, enabling accurate prediction of protein structures from amino acid sequences [8]. Conversely, the structure-to-sequence challenge—designing sequences that fold into a desired structure—has been effectively addressed by tools like Protein Message Passing Neural Network (pMPNN)[9]. Leveraging the rules and physical principles governing protein folding captured by these AI tools has enabled the successful design of *de novo* proteins [10].

AF2 predicts protein structures from amino acid sequences by leveraging evolutionary data and structural templates. It constructs multiple sequence alignments (MSAs) to identify conserved regions and co-evolutionary patterns among homologous proteins. These MSAs, along with structural templates when available, are processed by AF2's neural network to predict inter-residue distances and angles. This information is then used to assemble a highly accurate 3D model of the protein's structure [8].

While AF2 solves the forward problem of predicting structure from sequence, designing sequences for a target structure requires a different approach. In inverted AF2 version, AF2seq, the input is a target 3D structure. AF2seq initiates with a random amino acid sequence, uses AF2 to predict its structure, and then compares this prediction to the target structure. It iteratively adjusts the amino acid sequence and re-predicts its structure with AF2, refining the sequence until the predicted structure matches the target 3D structure[10].

PMPNN is a deep learning model designed to create protein sequences that will fold into a given 3D structure. It takes the 3D coordinates of the protein backbone, excluding side chains, as an input. The backbone encoder then processes input features to generate node (atom) and edge (bond) features. These encodings are then fed into a sequence decoder that generates amino acid sequence in autoregressive manner.[9]

Study Aim: Expanding the Soluble Protein Space

Membrane β-barrels are naturally hydrophobic on their exterior, making them difficult to work with outside their native membrane environment. Their poor solubility in water often complicates structure determination and binding screens. Common methods to solubilize membrane proteins—such as detergent micelles, proteoliposomes, nanodisks, and virus-like particles—are typically time-consuming and expensive. Grafting a protein's extracellular regions onto soluble scaffolds can simplify their use while preserving functionality, such as small molecule or antibody binding and enzymatic activity. In some cases, grafting can also enhance antigen accessibility, improving immune recognition and response [11]. Preserving the native fold of the scaffold [12] might be crucial for epitope integrity of complex structures like β-barrels.

This study employs a computational pipeline [12] *(Fig.1)* combining AF2seq and pMPNN to graft the extracellular region of the immunogenic membrane protein OmpW onto a soluble scaffold. The goal of this grafting is to maintain the exact β-barrel fold as in the wild type (WT) while redesigning the sequence to ensure solubility. The AF2seq pipeline generates amino acid sequences optimized for structural compatibility and predicts their 3D structures using AF2. During this process, a loss function assesses structural similarity to the target fold and guides iterative sequence adjustments until an optimal match is achieved. PMPNN then generates 'soluble' sequences by analyzing the atomic coordinates and dihedral angles, ensuring they align with the target fold. This approach has previously enabled the design of soluble analogs of helical membrane proteins while preserving the structure of their extracellular epitopes, thereby expanding the accessible protein fold space [12].

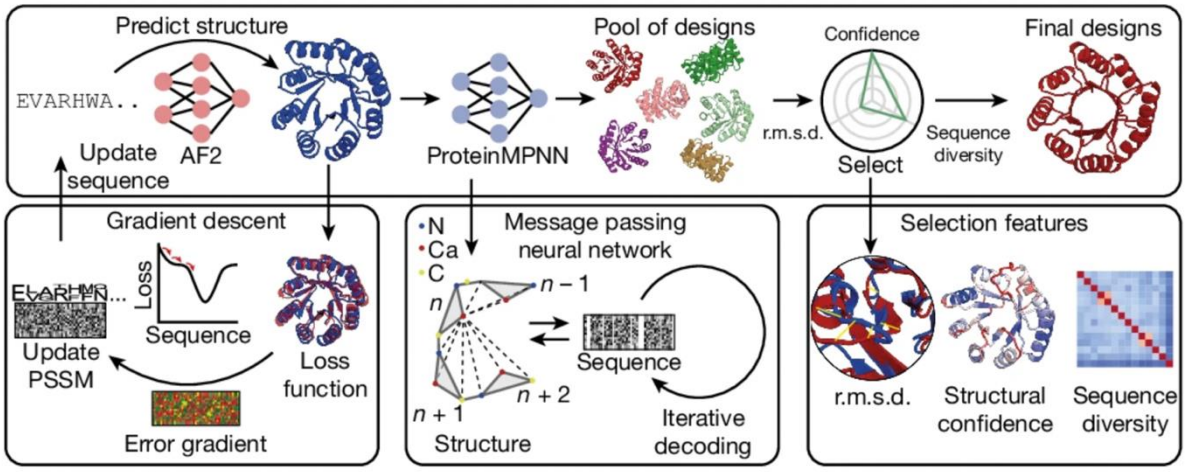

Figure 1: ***Schematic representation shows a two-step "backbone and sequence" design workflow.*** *First, an initial protein sequence is optimized against a target backbone using AlphaFold2 in a gradient-descent loop: the loss function tracks how well the predicted 3D structure matches the target. Once a suitable initial sequence is identified, pMPNN is used to sample and generate a diverse set of new sequences that preserve the target conformation. Finally, those pMPNN-generated designs are screened by checking how closely their structures match the original target, the model's confidence in those structures, and the resulting sequence diversity to select a final set of high-quality protein designs.* [12]

There are two key questions to address: Can a membrane β-barrel fold exist within the soluble protein space? And can we achieve the precision required to preserve the structural integrity and antigenic properties of the β-barrel's extracellular region? As discussed above, there are several essential differences between membrane and soluble β-barrel folds, such as β-strand length, number of hydrogen bonds, and average tilt angle per protein. While computational solubilization of alpha-helical membrane proteins has been successfully demonstrated [12][13][14] to our knowledge, this approach has not yet been tested on all-β membrane proteins.

De novo β-barrel design presents unique challenges due to the structural and chemical constraints of β-strands. A major issue is the risk of aggregation into amyloid-like structures if strand alignment is disrupted. The natural curvature of β-barrels also introduces strain, increasing instability. Excessive symmetry can further hinder proper folding by disrupting hydrogen bonding patterns. Overcoming these challenges requires precise sequence adjustments to balance stability, solubility, and function [1].

This study focuses on OmpW, a membrane β-barrel from *V. cholerae* and *E. coli*, with potential vaccine applications. We first computationally solubilized *E. coli* OmpW, leveraging its available X-ray structure. Then, we experimentally assessed the stability, oligomeric state, and secondary structure of *E. coli* OmpW soluble analogues. Then, this framework guided the design of soluble analogues of *V. cholerae* OmpW, which currently lacks structural data. Our computational approach assessed the AF2-foldability, structural congruency, and sequence similarity of the designs to the WT protein.

Methods

Protein expression and Purification

Plasmid digestion

The genes encoding the proteins of interest included an N-terminal His-tag for purification and 30-50 bp overhangs for subsequent cloning into a plasmid. We used the pET-11a plasmid[15] (Novagen), which was digested with NdeI and BlpI for our cloning experiments.

Gibson Assembly

Gibson Assembly for genes of proteins KP8, KP10, KP15 and KP21 was performed in the following way. 1 µl of plasmid DNA (concentration 30-50 ng/µl), 1.5 µl of the DNA insert and 3.5 µl of Gibson Assembly Master Mix were combined and incubated at 50°C for at least 1 hour in a thermocycler. Gibson Assembly works by using three enzymes contained in the master mix: exonuclease to chew back the 5' ends of the insert and plasmid and expose the complementary sequence for annealing, a DNA polymerase to fill in gaps where the fragments overlap, and finally a DNA ligase to seal the nicks, creating a continuous DNA strand.

Transformation with Heat Shock

The four recombinant plasmid samples (KP8, KP10, KP15 and KP21) obtained through Gibson assembly were transformed into *E.coli* strain HB101 using the heat shock method. *E.coli* cells were pretreated with calcium chloride to make the cell membranes more permeable and thereby competent. Then, competent cells were premixed with Gibson assembly mix and placed on ice for 5 mins allowing the DNA to adhere to the cell membrane. The samples were then subject to heat shock in which they were placed at 42°C for 30-45 seconds which causes a temporary destabilization of the bacterial cell membrane, creating pores through which the plasmid DNA could enter the cell. After the heat shock the samples were put back on ice to allow the cell membrane to reseal trapping the plasmids inside the cell. The samples were plated on LB agar plates containing Ampicillin and incubated overnight at 37°C. This allowed us to pick the colonies that had been successfully transformed with a plasmid containing an ampicillin resistance gene. [16,17]

MiniPrep

To isolate and purify plasmid DNA, we began by selecting two single colonies from each of four different plates to amplify the plasmid for further use. Using a sterile tip, the selected colonies were transferred into tubes containing 5 mL of LB medium supplemented with ampicillin. These cultures were incubated at 37°C for 12-16 hours in a shaking incubator set to 200-250 rpm to allow for bacterial growth.

To begin, the plasmid DNA isolation, the cultures were transferred to microcentrifuge tubes and centrifuged at 8000 rpm (6800 × g) for 2 minutes to pellet the bacterial cells. The supernatant was removed, leaving the bacterial pellet, which was then resuspended in 250 µL of chilled Resuspension Buffer. This step involved vortexing and pipetting to ensure complete resuspension without any clumps.

Next, 250 µL of Lysis Buffer was added to the resuspended cells to lyse the bacterial membranes and release the cellular contents, including the plasmid DNA. The tubes were gently inverted 4-6 times and incubated for 5 minutes to ensure effective lysis.

To neutralize the lysate and precipitate cell debris, proteins, and chromosomal DNA, 350 µL of Neutralization Buffer was added. This buffer brings the pH back to neutral, allowing the plasmid DNA to renature and remain in solution. The mixture was then centrifuged at maximum speed (>12,000 × g) for 5 minutes to separate the precipitate from the supernatant. The clear supernatant, which contains the plasmid DNA, was carefully transferred to a new tube without disturbing the pellet.

The supernatant was transferred to a spin column, and centrifuged for 1 minute, allowing the plasmid DNA to bind to the silica membrane within the column while other soluble components were washed away. The flow-through was discarded, and to further remove contaminants, 750 µL of Wash Buffer was added to the column and centrifuged at maximum speed for 15 seconds. This wash step was repeated with an additional 500 µL of Wash Buffer, followed by a 1-minute centrifugation to ensure complete removal of any residual buffer. The spin column was then put into a clean microcentrifuge tube.

Finally, to elute the purified plasmid DNA, 30 µL of nuclease-free water was added to the spin column, which was then allowed to sit for 3 minutes to facilitate DNA dissolution into the elution buffer and then centrifuged for 1 min. The eluted solution was then recuperated in the tube and stored at -20°C [18].

Sequencing

The concentrations of plasmids were measured using the Quibit and subsequently diluted to a concentration of 50-100 ng/µL — 12 µL of each plasmid were sent for sequencing. Sequencing was performed to verify that the plasmids contained the correctly inserted gene without any mutations before transforming them for protein expression.

Protein Expression and Purification

The verified plasmids with the genes of proteins KP8, KP10, KP15 and KP21 were used to transform into *E. coli* cells strain T7 for expression. The transformation proceeded the same way as the heat shock transformation [see Transformation with Heat Shock above].

To express the protein of interest we inoculated 5 ml of the overnight culture with 500 ml LBAmp at 37˚C ~3h until OD600 0.6-0.8. Once achieved, the cultures were put to rest at

4°C for 15 minutes. To induce the expression of our protein, 250 µl of 1M IPTG was added to the cultures and they were then incubated at 18°C overnight for optimal expression. The following day the cultures were spun down 4000 xg, for 20 min, and then the supernatant was removed. The bacterial pellet of each sample was then resuspended in 25 mL of Lysis Buffer. The following was added to each of the 4 samples: 400 µL of PMSF (a protease inhibitor), 400 µL of lysozyme (to break down the bacterial cell wall), 800 µL of DNase (to degrade DNA) and 2 mL of CellLytic reagent (to facilitate cell lysis). The pellet was resuspended by pipetting up and down. Each sample was transferred into separate 50 ml falcon tubes and incubated on a rotating wheel for 60 minutes at 4°C. The tubes were then centrifuged at 4000 xg for 20 minutes at 4°C to pellet cellular debris. The supernatant was subsequently filtered through a 0.2 µM filter. To initiate purification, 1.5 ml of 50:50 Agar-His beads were added to each tube and incubated for 60 minutes at 4°C to bind the His-tagged proteins. The contents of each tube were then applied to bench-top columns, which were washed with 10 column volumes (CV) of His-Wash buffer to remove non-specifically bound particles. His-Elution buffer (2 mL) was added to each column and incubated for 20 minutes to facilitate the release of bound proteins. The eluates containing the proteins of interest were collected.

For further purification, the eluate was subjected to Size Exclusion Chromatography (SEC) using a Superdex 75 16/600 column (S75), which separates proteins ranging from 3,000 to 70,000 Da. SEC was performed overnight in PBS buffer. The fractions containing the protein of interest were pooled and concentrated using an Amicon 15 ml centrifugal filter with a 3 kDa cutoff. Protein concentration was measured and adjusted to >1 mg/mL, after which the proteins were snap-frozen in liquid nitrogen and stored at -80°C for long-term preservation.

Circular Dichroism (CD)

Circular Dichroism (CD) spectroscopy is a technique used to obtain structural information about proteins, particularly their secondary structure. This method works by shining circularly polarized light through a protein solution and measuring the difference in absorption between left- and right-circularly polarized light, which produces characteristic signals based on the protein's structure[19]. The measurements were performed using a Chirascan V100 CD spectrometer. To ensure accuracy, the system was purged with nitrogen gas to eliminate residual oxygen, which can absorb in the UV range (200-250 nm) and interfere with the CD measurements. For our experiments, proteins were diluted to 0.3 mg/mL in PBS buffer, and 300 µL of the solution was placed into a quartz cuvette for each measurement. The CD spectra were recorded between 190-250 nm.

MALS

Size Exclusion Chromatography coupled with Multi-Angle Light Scattering (SEC-MALS) is a technique used to accurately determine the molecular weight of macromolecules in solution. In this method, proteins are first separated by size using Size Exclusion Chromatography (SEC). As the proteins elute from the SEC column, they pass through a Multi-Angle Light Scattering (MALS) detector, which measures the intensity of light scattered by the proteins at multiple angles. This scattering data is then used to calculate the absolute molecular weight of the proteins in each fraction [20]. To analyze the molecular weight and oligomeric state of our protein samples, they were concentrated to 1 mg/mL in PBS. The samples were then subjected to SEC-MALS, with the experiment conducted in PBS at a flow rate of 0.5 mL/min using a Superdex 75 10/300 size-exclusion chromatography column on an UltiMate 3000 system. The molecular mass was determined using Astra software, and UV absorbance was measured with the DAWN detector's integrated sensor.

Computational Methods

To achieve effective membrane protein redesign and solubilization while preserving the epitope, we employed a computational pipeline that integrates AF2seq and pMPNN methodologies. The design target was the *V. cholerae* OmpW protein, for which no PDB structure was available. To address this, we employed AlphaFold3 (AF3)[21] to predict its structure using the protein sequence obtained from UniProt. AF3 generates structural predictions using five different models, each providing a variation within the AF2 framework. For our design process, we selected the structure predicted by model 0.

AF2seq

AF2seq leverages AF2's structure prediction model to design protein sequences by iteratively modifying them to achieve the desired 3D structure, guided by a loss function that aligns the sequence with the target fold.

In our pipeline, AF2seq was used to design sequences converging to the target structure via gradient descent optimization. We applied two different weights to the loss function to balance structural fidelity and diversity. The target structure, derived from the AF3 model 0 prediction, had fixed positions to preserve the epitope conformation.

To maintain epitope integrity while promoting structural diversity, the loss function weights were adjusted: the distogram weight on the epitope (dgram_cce) was set higher to preserve its structure, while the full distogram weight (full_dgram_loss) was lower to encourage diverse overall structures. These variations defined two versions of the design: version 1 (dgram_cce = 0.5, full_dgram_loss = 1.0) and version 2 (dgram_cce = 0.1, full_dgram_loss = 1.0).

Each design started with an initial sequence based on the predicted secondary structure, using alanines for helices, valines for β-sheets, and glycines for loops. This sequence was processed through inverted AF2 to generate 43 predicted structures, with a composite loss function assessing alignment with the target structure. Errors at each position were backpropagated to update the sequence via a Position-Specific Scoring Matrix (PSSM) and Adam optimizer, guiding amino acid mutations to reduce the error.

The PSSM was transformed into a probability distribution through softmax, selecting the most probable amino acid for each position. This process was repeated iteratively, refining the sequence until minimal error and high alignment were achieved. For each trajectory, 500 rounds of gradient descent optimization were run, and after convergence, three additional rounds of AF2 recycling were performed, followed by relaxation in an AMBER force field [22] to simulate physical interactions and stabilize the structure.

PMPNN

The predicted backbones obtained through AF2seq and relaxation were processed by pMPNN to generate five optimized sequences per design. For each input structure, three versions were generated using model types designed for soluble proteins: v_48_010, v_48_020, and v_48_030. These models correspond to different levels of Gaussian noise (0.10 Å, 0.20 Å, and 0.30 Å) applied during training, which influences the model's exploration. Higher noise introduces more variation, allowing broader sequence exploration, while lower noise helps the model stay closer to the original structure. This controlled noise ensures robustness and flexibility in the resulting sequences, accommodating natural protein conformation fluctuations.

Additionally, sampling temperature was used to control sequence diversity. A high temperature (close to 1) makes the model more random and exploratory, while a low temperature focuses on specific outcomes. We set the temperature to 0.2, favoring more deterministic sequences. This pipeline generated 1,290 sequences, which were then processed through AF2 in single-sequence mode with three recycles to predict their final amino acid sequences.

The pool of designed proteins was further filtered based on criteria such as pLDDT > 70, epitope RMSD < 3 Å, sequence similarity (original vs. output) < 0.5, and apolar surface fraction < 0.3. A manual selection process was also conducted to identify unfavorable interactions, including side chain clashes, same-charge interactions, buried unsatisfied polar side chains (BUNS), and core packing issues, with a focus on maintaining a hydrophobic core.

Results

Experimental characterization of *E.coli* OmpW soluble analogues

Soluble analogues of *E. coli* OmpW (KP8, KP10, KP15, and KP21) were designed in silico using the computational pipeline shown in *Fig. 1*. These theoretical sequences were then expressed and purified experimentally, yielding the respective proteins KP8, KP10, KP15, and KP21. Size Exclusion Chromatography (SEC) purification showed that the major peaks for all four proteins appeared at similar retention times, between 60 and 80 mL. Circular Dichroism (CD) spectra revealed that the proteins KP8, KP10, KP15, and KP21 exhibited peaks around 215–220 nm at 20°C. However, at 90°C, the CD curves of KP10 and KP21 shifted further to the left, with KP10 showing a peak at approximately 210–215 nm and KP21 at 205–210 nm, suggesting a transition to a random coil conformation. Multi-Angle Light Scattering (MALS) analysis indicated that all four proteins eluted as sharp peaks, with experimentally determined molecular weights closely matching the theoretical value of approximately 23 kDa.

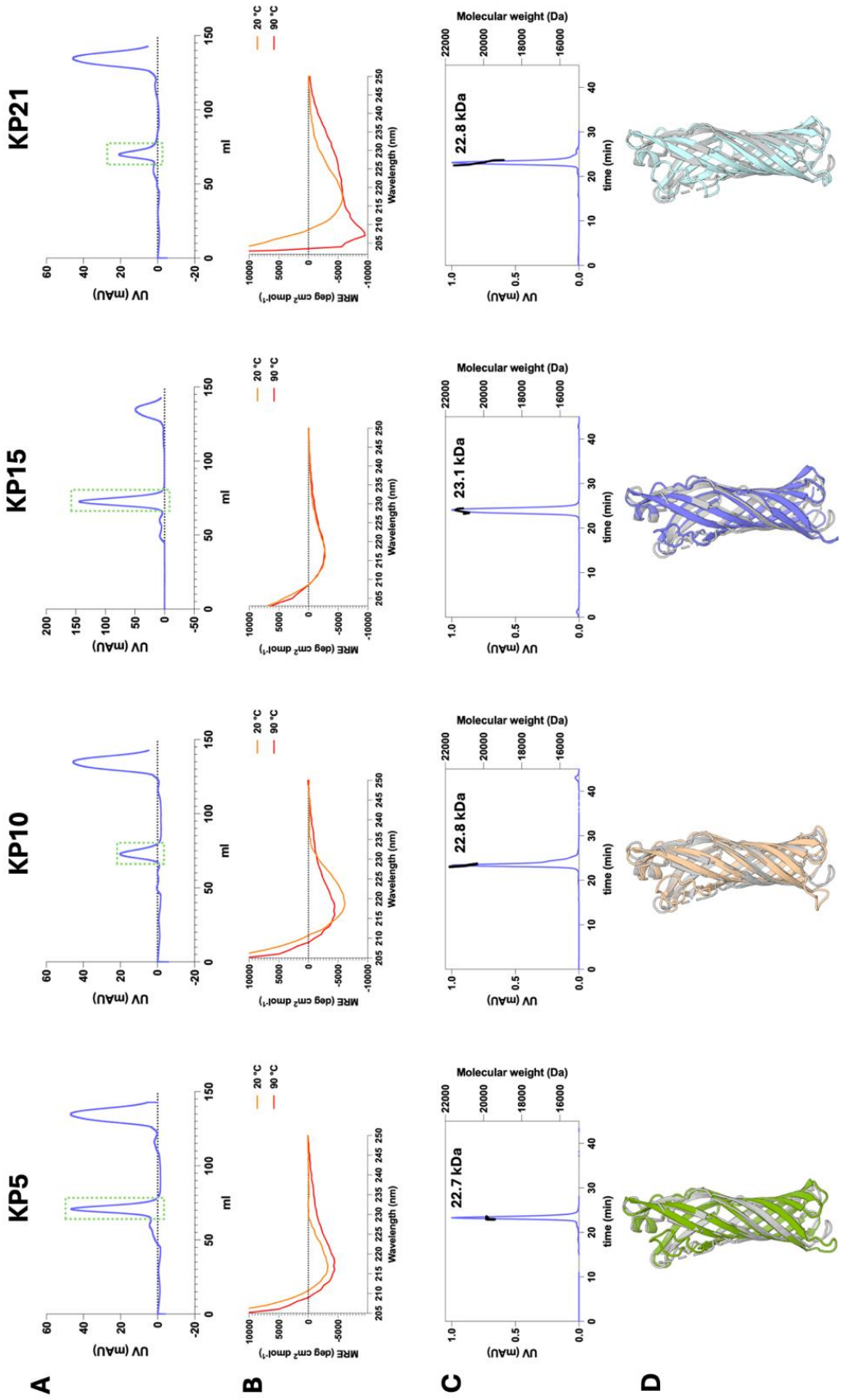
KP21
KP15
KP10
KP5
A
B
C
D
UV (mAU)
ml
MRE (deg cm² dmol⁻¹)
Wavelength (nm)
20 °C
90 °C
Molecular weight (Da)
22.8 kDa
23.1 kDa
22.8 kDa
22.7 kDa
time (min)

Figure 5: In Vitro characterization of four E. coli OmpW soluble analogues. ***(A)*** *Size Exclusion Chromatograms following Ni-NTA chromatography, with the green square indicating the collected peak used for further tests.* ***(B)*** *Circular Dichroism (CD) spectra at 20°C and 90°C, normalized to protein concentration and length.* ***(C)*** *Multi-Angle Light Scattering (MALS) chromatograms, with the left y-axis showing UV absorbance (blue line) and the right y-axis representing molecular weight (MW) determined by MALS (black line).* ***(D)*** *AF2 predictions of the selected E. coli OmpW analogues. The grey structure represents the WT E. coli OmpW PDB structure (PDB 2f1v), and the colored structures show AF2-predicted models for KP8, KP10, KP15, and KP21 in single-sequence mode.*

Computational design and analysis of *V.cholerae* OmpW soluble analogues

This section presents the results from a computational pipeline designed to solubilize OmpW from *V. cholerae* while preserving its epitope structure. In the absence of an available resolved structure, we used the AF3 model 0 prediction for subsequent experiments. The pLDDT scores were lower for the epitope residues (*Fig.2*), with an average of 54.4 for the epitope and 81.7 for the β-barrel.

Starting from an AF3-predicted structure, AF2seq analysis was conducted using two sets of weights—version 1 (more stringent) and version 2 (less stringent)—resulting in a total of 43 structures. Upon relaxation, these structures were then processed through different versions of pMPNN (010, 020, and 030), which varied in the level of Gaussian noise applied during design, ultimately generating 1,290 unique sequences. To track progress, the key scores were monitored throughout the design process (*Fig.3*).

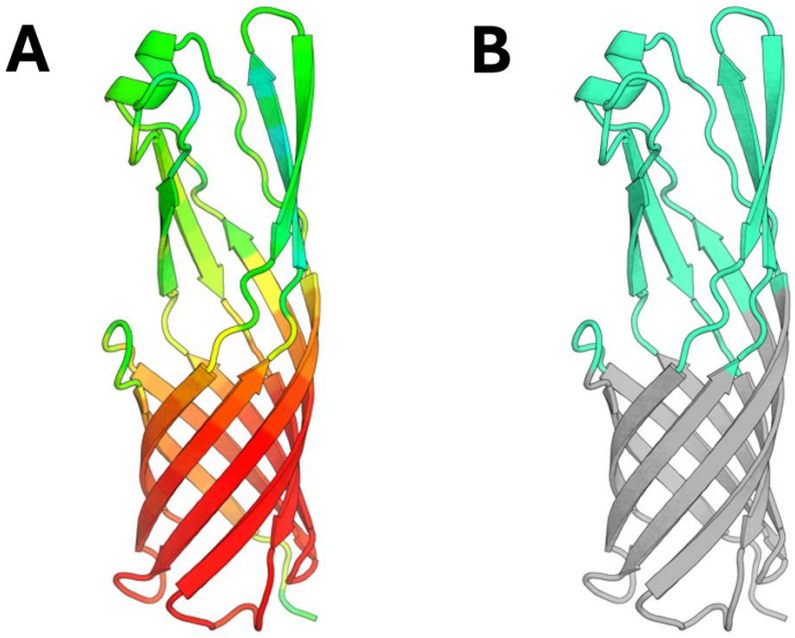

Figure 2: AF3-based structural models of Vibrio cholerae OmpW. ***(A)*** *Ribbon representation colored by pLDDT confidence scores, transitioning from high-confidence regions (red: 98.24) to lower-confidence*

regions (green: 37.86). ***(B)*** *The same predicted structure highlighting the putative epitope (blue) and the membrane-embedded region (grey).*

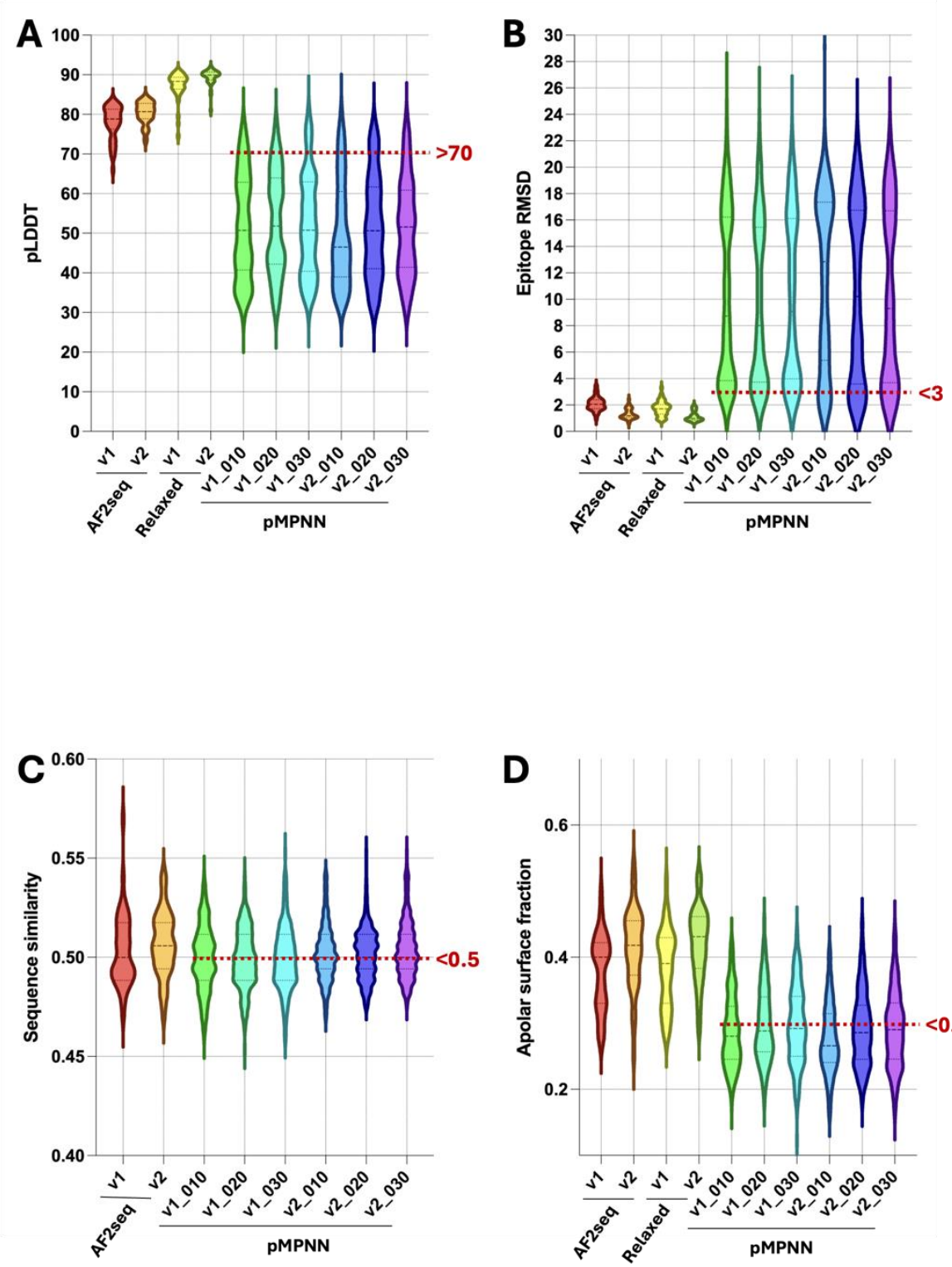

Figure 3: Evaluating and selecting V. cholerae OmpW soluble analogues across AF2seq-pMPNN pipeline.*The red line represents computational selection criteria. These designations—v1_010, v1_020, v1_030, v2_010, v2_020, v2_030—represent distinct model versions. "v1" and "v2" indicate the loss*

function used at the AF2seq step, while "010," "020," and "030" refer to sequences generated by pMPNN models trained with varying levels of Gaussian noise. This naming convention provides clarity in differentiating the various stages or configurations of the models. ***(A)*** *pLDDT (predicted Local Distance Difference Test) scores, illustrating the confidence in the predicted structures.* ***(B)*** *Epitope RMSD (Root-Mean-Square Deviation) values, reflecting the structural deviations of the epitopes of the designed proteins from the target V. cholerae OmpW structure.* ***(C)*** *Sequence similarity to the original OmpW of V. cholerae sequence.* ***(D)*** *Apolar surface fraction, with lower values indicating improved solubility.*

Protein designs were filtered based on apolar surface fraction (<0.3), sequence similarity (<0.5), epitope RMSD (<3 Å), and pLDDT (>70), yielding 10 sequences for v1_010, 9 for v1_020, 11 for v1_030, 20 for v2_010, 16 for v2_020, and 14 for v2_030. Manual selection followed, filtering for unfavorable interactions like side chain clashes, same-charge interactions, buried unsatisfied sidechains (BUNS), and water cavities, resulting in 6 optimal designs. Version 2 had more selected sequences per noise level than version 1 (*Table 1*).

Design Version	V1_010	V1_020	V1_030	V2_010	V2_020	V2_030
Computational selection criteria	10	9	11	20	16	14
Manual Selection	2	0	1	0	1	2

Table 1. Selection of soluble V. cholerae OmpW soluble analogues based on computational and manual criteria. *V1 and V2 are AF2seq outputs with different weight sets, while 010, 020, and 030 are pMPNN models. Stage 1 (row 1) involved computational filtering based on apolar surface fraction (<0.3), sequence similarity (<0.5), epitope RMSD (<3 Å), and pLDDT (>70). Stage 2 (row 2) involved manual curation, assessing side chain clashes, charge interactions, polar cores, and water cavities for structural integrity.*

The six selected designs (*Table 1*) correspond to structures generated by different pMPNN models and AF2seq weights: two from v1_010, two from v2_030, one from v1_030, and one from v2_020.

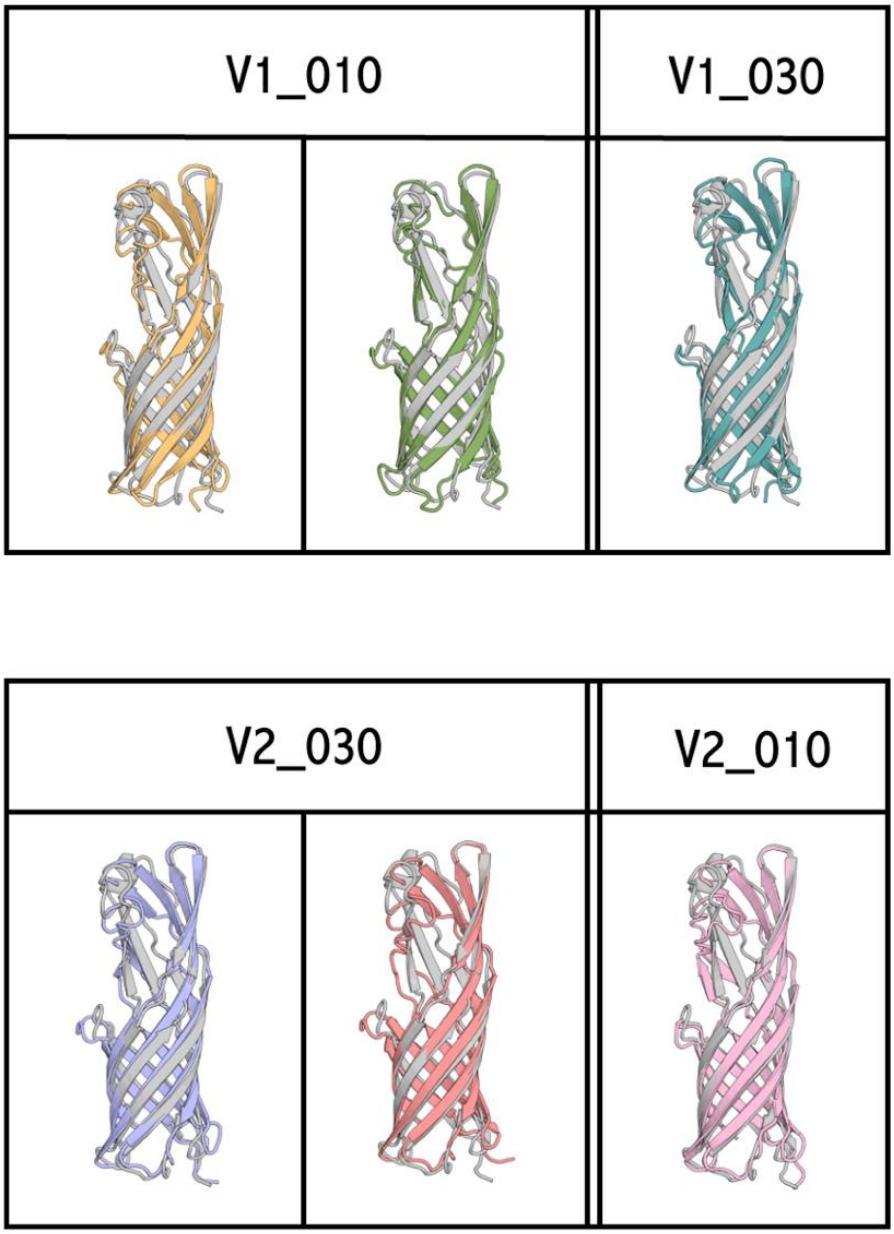

Figure 4: AF2 structures of the top selected V. cholerae OmpW soluble analogues. *The grey structure represents the reference V. cholerae WT AF3 prediction, while the colored structures correspond to AF2 models of the soluble analogues.*

Discussion

All-β proteins pose a significant design challenge due to their dependence on long-range hydrogen bonding and a high propensity for aggregation. In this study, our goal was to computationally 'solubilize' the eight-stranded immunogenic membrane β-barrel protein OmpW while preserving the sequence and structural integrity of its epitope. Soluble analogues of immunogenic proteins facilitate binder library screening and serve as promising vaccine candidates, in contrast to their wild-type counterparts, which are often poorly expressed and require cumbersome handling. Our research was conducted

in two phases. First, we expressed, purified, and characterized the oligomerization status and secondary structure of *E. coli* OmpW soluble analogues (KP8, KP10, KP15 and KP21) designed using the computational pipeline *(Fig.1)*. Next, employed the same computational pipeline as *E.coli* to generate soluble analogues of *V. cholerae* OmpW. Given the close homology between *E. coli* and *V. cholerae* OmpW, both hold therapeutic potential as vaccine candidates.

Protein expression and characterization of *E. coli* OmpW soluble analogues

In vitro characterization of solubilized OmpW from *E. coli* offered valuable insights into both the strengths and limitations of our computational pipeline for solubilizing membrane proteins. While the AF2seq-pMPNN approach has been primarily applied to α-helical membrane proteins, this study extended its application to the challenging design of soluble analogues for membrane all-β proteins [23]. Soluble analogues of *E. coli* OmpW were generated similarly to those of *V. cholerae* OmpW, with slightly different selection criteria, including a stricter pLDDT threshold. Unlike *V. cholerae* OmpW, which required AF3 prediction due to the lack of a resolved structure, the *E. coli* OmpW process leveraged its available PDB structure (2f1v) in the initial stage. Access to a real structure likely contributed to the higher model confidence observed for the *E. coli* OmpW soluble analogues.

Circular Dichroism (CD) spectroscopy confirmed that the designed proteins retained their β-barrel secondary structures. CD spectra recorded at 20°C showed that KP8, KP10, KP15, and KP21 maintained their β-barrel conformations, indicating successful solubilization *(Fig.5B)*. At 90°C, KP8 and KP15 remained structured, while KP10 and KP21 began to unfold *(Fig.5B)*. This high-temperature stability differs from natural proteins, which typically have melting temperatures (Tm) of 40–60°C in mesophilic organisms like *E. coli* and *V. cholerae* [24]. The enhanced stability at 90°C is likely due to the sequence design by pMPNN, which optimizes for highly stable proteins[25].

Multi-Angle Light Scattering (MALS) analysis further validated the designed proteins. The measured molecular weights matched theoretical predictions, and all four proteins (KP8, KP10, KP15, and KP21) were monodisperse with no signs of oligomerization *(Fig.5C)*. Their sharp elution peaks indicated the absence of multiple populations or aggregation *(Fig.5C)*. The agreement between experimental data and theoretical expectations confirms that the proteins were successfully solubilized and remained stable in solution. The next steps include using these solubilized proteins in binding assays to determine if antibodies specific to native *E.coli* OmpW recognize the soluble analogues. Additionally, we could perform immunization studies in animals to assess whether these proteins elicit a strong immune response when introduced alongside *E. coli*.

Computational Design of *V. cholerae* OmpW Soluble Analogues

To generate *V. cholerae* OmpW soluble analogues, we first predicted the template structure from sequence using AF3, as no resolved structure is available for *V. cholerae* OmpW. The resulting model revealed a region of low pLDDT scores corresponding to the proposed epitope (*Fig.2*). During AF2 backpropagation, the model shifts logits to concentrate probability mass toward the correct distance bins, thereby increasing confidence (pLDDT). However, high pLDDT does not necessarily equate to true foldability or solubility. In the relaxation step, the structure is repredicted without a template and optimized using the AMBER force field, reducing steric clashes and strain, which further aligns the structure with AF2's priors and increases pLDDT. After processing the relaxed structures through pMPNN and repredicting in single-sequence mode, pLDDT drops—indicating that pMPNN may mitigate AF2's inherent biases, leading to more diverse sequences that better reflect genuine foldability. We selected designs with pLDDT above 70, as these are considered confident with generally accurate backbone predictions.

Using single-sequence mode during relaxation ensures that the epitope's stability is predicted solely from its sequence (*Fig.3*). In contrast, after applying pMPNN, most designs fail to fold into a similar conformation, correlating with a drop in pLDDT. Selecting structures with an epitope RMSD below 3 Å allows to select variants with only minor local adjustments.

At the sequence level, similarity to the wild-type drops significantly to around 60% after AF2 backpropagation and is further refined by pMPNN (*Fig.3*). Notably, similarity remains above 40% because 40% of the designed protein comprises the epitope. Regarding surface polarity, AF2 backpropagation produces sequences with a high fraction of apolar residues (approximately 40–50%), but pMPNN reduces this to around 30%, in line with our selection criteria (*Fig. 3*). This ensures that for every apolar residue on the surface, there are at least two polar residues, thereby minimizing hydrophobic exposure, preventing aggregation, and enhancing stability in aqueous environments.

Following computational selection, which resulted in 80 sequences, further manual curation filtered out designs with BUNS, unfavorable interactions, water cavities, and side-chain clashes *(Table. 1)*. These structural flaws could destabilize the protein and promote unfolding, ultimately leaving us with six selected sequences (*Fig. 4*).

Outlook

Of the 2,100 computationally designed *E. coli* OmpW soluble analogues, 11 were soluble and expressed well in bacteria, yielding a 5% success rate. Despite a sequence identity of 57% between *E. coli* and *V. cholerae* OmpW, pLDDT values were higher for *E. coli* designs, likely due to the availability of a resolved structure (*PDB 2F1V*), unlike *V. cholerae* OmpW. Given this limitation, a larger computational dataset may be necessary for *V.*

cholerae OmpW, as only six sequences were selected from the 1,200 generated. A higher experimental screening effort may also be required to ensure not only solubility but also epitope structural integrity.

For the E. coli OmpW analogues, the successful solubilization *(Fig. 5)* underscores that membrane β-barrels can indeed be redesigned to adopt a stable, water-soluble form. The monomeric, β-stranded structures now pave the way for binding assays with anti-OmpW antibodies and the protein's natural ligand, colicin S4, to confirm whether the epitope is preserved sufficiently for high-affinity recognition. Additionally, techniques such as Surface Plasmon Resonance (SPR) or Isothermal Titration Calorimetry (ITC) can quantify binding affinities and compare them with theoretical predictions—providing critical insight into how closely the soluble analogues recapitulate the native epitope surface. If these tests show robust binding, X-ray crystallography would be the next step to visualize the fine structural details and confirm the accuracy of the epitope.

Applications

This study tested a deep learning-driven method for solubilizing native beta-barrel membrane proteins while preserving their extracellular sequences and structures. This approach has wide potential for supporting the biochemical and structural analysis of integral membrane proteins, facilitating therapeutic discovery through screening of purified soluble targets, and creating antigenically intact molecules for vaccine development.

Sources

(1) Dou, J.; Vorobieva, A. A.; Sheffler, W.; Doyle, L. A.; Park, H.; Bick, M. J.; Mao, B.; Foight, G. W.; Lee, M. Y.; Gagnon, L. A.; Carter, L.; Sankaran, B.; Ovchinnikov, S.; Marcos, E.; Huang, P.-S.; Vaughan, J. C.; Stoddard, B. L.; Baker, D. De Novo Design of a Fluorescence-Activating β-Barrel. *Nature* **2018**, *561* (7724), 485–491. https://doi.org/10.1038/s41586-018-0509-0.

(2) Dhar, R.; Feehan, R.; Slusky, J. S. G. Membrane Barrels Are Taller, Fatter, Inside-Out Soluble Barrels. *J Phys Chem B* **2021**, *125* (14), 3622–3628. https://doi.org/10.1021/acs.jpcb.1c00878.

(3) *Epitope - an overview | ScienceDirect Topics*. https://www.sciencedirect.com/topics/medicine-and-dentistry/epitope (accessed 2024-09-10).

(4) Axe, D. D. Estimating the Prevalence of Protein Sequences Adopting Functional Enzyme Folds. *Journal of Molecular Biology* **2004**, *341* (5), 1295–1315. https://doi.org/10.1016/j.jmb.2004.06.058.

(5) Leman, J. K.; Weitzner, B. D.; Lewis, S. M.; Adolf-Bryfogle, J.; Alam, N.; Alford, R. F.; Aprahamian, M.; Baker, D.; Barlow, K. A.; Barth, P.; Basanta, B.; Bender, B. J.; Blacklock, K.; Bonet, J.; Boyken, S. E.; Bradley, P.; Bystroff, C.; Conway, P.; Cooper, S.; Correia, B. E.; Coventry, B.; Das, R.; De Jong, R. M.; DiMaio, F.; Dsilva, L.; Dunbrack, R.; Ford, A. S.; Frenz, B.; Fu, D. Y.; Geniesse, C.; Goldschmidt, L.; Gowthaman, R.; Gray, J. J.; Gront, D.; Guffy, S.; Horowitz, S.; Huang, P.-S.; Huber, T.;

Jacobs, T. M.; Jeliazkov, J. R.; Johnson, D. K.; Kappel, K.; Karanicolas, J.; Khakzad, H.; Khar, K. R.; Khare, S. D.; Khatib, F.; Khramushin, A.; King, I. C.; Kleffner, R.; Koepnick, B.; Kortemme, T.; Kuenze, G.; Kuhlman, B.; Kuroda, D.; Labonte, J. W.; Lai, J. K.; Lapidoth, G.; Leaver-Fay, A.; Lindert, S.; Linsky, T.; London, N.; Lubin, J. H.; Lyskov, S.; Maguire, J.; Malmström, L.; Marcos, E.; Marcu, O.; Marze, N. A.; Meiler, J.; Moretti, R.; Mulligan, V. K.; Nerli, S.; Norn, C.; Ó'Conchúir, S.; Ollikainen, N.; Ovchinnikov, S.; Pacella, M. S.; Pan, X.; Park, H.; Pavlovicz, R. E.; Pethe, M.; Pierce, B. G.; Pilla, K. B.; Raveh, B.; Renfrew, P. D.; Burman, S. S. R.; Rubenstein, A.; Sauer, M. F.; Scheck, A.; Schief, W.; Schueler-Furman, O.; Sedan, Y.; Sevy, A. M.; Sgourakis, N. G.; Shi, L.; Siegel, J. B.; Silva, D.-A.; Smith, S.; Song, Y.; Stein, A.; Szegedy, M.; Teets, F. D.; Thyme, S. B.; Wang, R. Y.-R.; Watkins, A.; Zimmerman, L.; Bonneau, R. Macromolecular Modeling and Design in Rosetta: Recent Methods and Frameworks. *Nat Methods* **2020**, *17* (7), 665–680. https://doi.org/10.1038/s41592-020-0848-2.

(6) Watson, J. L.; Juergens, D.; Bennett, N. R.; Trippe, B. L.; Yim, J.; Eisenach, H. E.; Ahern, W.; Borst, A. J.; Ragotte, R. J.; Milles, L. F.; Wicky, B. I. M.; Hanikel, N.; Pellock, S. J.; Courbet, A.; Sheffler, W.; Wang, J.; Venkatesh, P.; Sappington, I.; Torres, S. V.; Lauko, A.; De Bortoli, V.; Mathieu, E.; Ovchinnikov, S.; Barzilay, R.; Jaakkola, T. S.; DiMaio, F.; Baek, M.; Baker, D. De Novo Design of Protein Structure and Function with RFdiffusion. *Nature* **2023**, *620* (7976), 1089–1100. https://doi.org/10.1038/s41586-023-06415-8.

(7) Pacesa, M.; Nickel, L.; Schellhaas, C.; Schmidt, J.; Pyatova, E.; Kissling, L.; Barendse, P.; Choudhury, J.; Kapoor, S.; Alcaraz-Serna, A.; Cho, Y.; Ghamary, K. H.; Vinué, L.; Yachnin, B. J.; Wollacott, A. M.; Buckley, S.; Westphal, A. H.; Lindhoud, S.; Georgeon, S.; Goverde, C. A.; Hatzopoulos, G. N.; Gönczy, P.; Muller, Y. D.; Schwank, G.; Swarts, D. C.; Vecchio, A. J.; Schneider, B. L.; Ovchinnikov, S.; Correia, B. E. BindCraft: One-Shot Design of Functional Protein Binders. bioRxiv December 7, 2024, p 2024.09.30.615802. https://doi.org/10.1101/2024.09.30.615802.

(8) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(9) Dauparas, J.; Anishchenko, I.; Bennett, N.; Bai, H.; Ragotte, R. J.; Milles, L. F.; Wicky, B. I. M.; Courbet, A.; de Haas, R. J.; Bethel, N.; Leung, P. J. Y.; Huddy, T. F.; Pellock, S.; Tischer, D.; Chan, F.; Koepnick, B.; Nguyen, H.; Kang, A.; Sankaran, B.; Bera, A. K.; King, N. P.; Baker, D. Robust Deep Learning–Based Protein Sequence Design Using ProteinMPNN. *Science* **2022**, *378* (6615), 49–56. https://doi.org/10.1126/science.add2187.

(10) Goverde, C. A.; Wolf, B.; Khakzad, H.; Rosset, S.; Correia, B. E. De Novo Protein Design by Inversion of the AlphaFold Structure Prediction Network. *Protein Sci* **2023**, *32* (6), e4653. https://doi.org/10.1002/pro.4653.

(11) Correia, B. E.; Bates, J. T.; Loomis, R. J.; Baneyx, G.; Carrico, C.; Jardine, J. G.; Rupert, P.; Correnti, C.; Kalyuzhniy, O.; Vittal, V.; Connell, M. J.; Stevens, E.; Schroeter, A.; Chen, M.; MacPherson, S.; Serra, A. M.; Adachi, Y.; Holmes, M. A.; Li, Y.; Klevit, R. E.; Graham, B. S.; Wyatt, R. T.; Baker, D.; Strong, R. K.; Crowe, J. E.; Johnson, P. R.; Schief, W. R. Proof of Principle for Epitope-Focused Vaccine Design. *Nature* **2014**, *507* (7491), 201–206. https://doi.org/10.1038/nature12966.
(12) Goverde, C. A.; Pacesa, M.; Goldbach, N.; Dornfeld, L. J.; Balbi, P. E. M.; Georgeon, S.; Rosset, S.; Kapoor, S.; Choudhury, J.; Dauparas, J.; Schellhaas, C.; Kozlov, S.; Baker, D.; Ovchinnikov, S.; Vecchio, A. J.; Correia, B. E. Computational Design of Soluble and Functional Membrane Protein Analogues. *Nature* **2024**, *631* (8020), 449–458. https://doi.org/10.1038/s41586-024-07601-y.
(13) Yao, Z.; Kuhlman, B. Design of a Water-Soluble CD20 Antigen with Computational Epitope Scaffolding. bioRxiv December 6, 2024, p 2024.12.05.627087. https://doi.org/10.1101/2024.12.05.627087.
(14) Nikolaev, A.; Orlov, Y.; Tsybrov, F.; Kuznetsova, E.; Shishkin, P.; Kuzmin, A.; Mikhailov, A.; Galochkina, Y. S.; Anuchina, A.; Chizhov, I.; Semenov, O.; Kapranov, I.; Borshchevskiy, V.; Remeeva, A.; Gushchin, I. Engineering of Soluble Bacteriorhodopsin. bioRxiv November 21, 2024, p 2024.11.20.624543. https://doi.org/10.1101/2024.11.20.624543.
(15) LLC, G. B. *pET-11a Sequence and Map*. https://www.snapgene.com/plasmids/pet_and_duet_vectors_(novagen)/pET-11a (accessed 2025-02-25).
(16) *Bacterial Transformation Workflow - ES*. https://www.thermofisher.com/es/es/home/life-science/cloning/cloning-learning-center/invitrogen-school-of-molecular-biology/molecular-cloning/transformation/bacterial-transformation-workflow.html (accessed 2024-08-20).
(17) Asif, A.; Mohsin, H.; Tanvir, R.; Rehman, Y. Revisiting the Mechanisms Involved in Calcium Chloride Induced Bacterial Transformation. *Front Microbiol* **2017**, *8*, 2169. https://doi.org/10.3389/fmicb.2017.02169.
(18) QIAprep Miniprep Handbook.
(19) Miles, A. J.; Janes, R. W.; Wallace, B. A. Tools and Methods for Circular Dichroism Spectroscopy of Proteins: A Tutorial Review. *Chem Soc Rev 50* (15), 8400–8413. https://doi.org/10.1039/d0cs00558d.
(20) *Size-Exclusion Chromatography with Multi-Angle Light Scattering (SEC-MALS)*. https://cmi.hms.harvard.edu/SEC-MALS (accessed 2024-08-24).
(21) Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; Bodenstein, S. W.; Evans, D. A.; Hung, C.-C.; O'Neill, M.; Reiman, D.; Tunyasuvunakool, K.; Wu, Z.; Žemgulytė, A.; Arvaniti, E.; Beattie, C.; Bertolli, O.; Bridgland, A.; Cherepanov, A.; Congreve, M.; Cowen-Rivers, A. I.; Cowie, A.; Figurnov, M.; Fuchs, F. B.; Gladman, H.; Jain, R.; Khan, Y. A.; Low, C. M. R.; Perlin, K.; Potapenko, A.; Savy, P.; Singh, S.; Stecula, A.; Thillaisundaram, A.; Tong, C.; Yakneen, S.; Zhong, E. D.; Zielinski, M.; Žídek, A.; Bapst, V.; Kohli, P.; Jaderberg, M.; Hassabis, D.; Jumper, J. M. Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3. *Nature* **2024**, *630* (8016), 493–500. https://doi.org/10.1038/s41586-024-07487-w.

(22) Case, D. A.; Cheatham III, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz Jr., K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber Biomolecular Simulation Programs. *Journal of Computational Chemistry* **2005**, *26* (16), 1668–1688. https://doi.org/10.1002/jcc.20290.
(23) Hong, H.; Patel, D. R.; Tamm, L. K.; van den Berg, B. The Outer Membrane Protein OmpW Forms an Eight-Stranded Beta-Barrel with a Hydrophobic Channel. *J Biol Chem* **2006**, *281* (11), 7568–7577. https://doi.org/10.1074/jbc.M512365200.
(24) *Mesophile - an overview | ScienceDirect Topics*. https://www.sciencedirect.com/topics/immunology-and-microbiology/mesophile# (accessed 2024-10-06).
(25) Sumida, K. H.; Núñez-Franco, R.; Kalvet, I.; Pellock, S. J.; Wicky, B. I. M.; Milles, L. F.; Dauparas, J.; Wang, J.; Kipnis, Y.; Jameson, N.; Kang, A.; De La Cruz, J.; Sankaran, B.; Bera, A. K.; Jiménez-Osés, G.; Baker, D. Improving Protein Expression, Stability, and Function with ProteinMPNN. *J. Am. Chem. Soc.* **2024**, *146* (3), 2054–2061. https://doi.org/10.1021/jacs.3c10941.
(26) Graham, B. S.; Gilman, M. S. A.; McLellan, J. S. Structure-Based Vaccine Antigen Design. *Annual Review of Medicine* **2019**, *70* (Volume 70, 2019), 91–104. https://doi.org/10.1146/annurev-med-121217-094234.