Computational design of membrane proteins using RosettaMembrane

Amanda M. Duran [1,2] and Jens Meiler[1,2]*

[1]Department of Chemistry, Vanderbilt University, Nashville, Tennessee 37235
[2]Center for Structural Biology, Vanderbilt University, Nashville, Tennessee 37240

Abstract: Computational membrane protein design is challenging due to the small number of high-resolution structures available to elucidate the physical basis of membrane protein structure, multiple functionally important conformational states, and a limited number of high-throughput biophysical assays to monitor function. However, structural determination of membrane proteins has made tremendous progress in the past years. Concurrently the field of soluble computational design has made impressive inroads. These developments allow us to tackle the formidable challenge of designing functional membrane proteins. Herein, Rosetta is benchmarked for membrane protein design. We evaluate strategies to cope with the often reduced quality of experimental membrane protein structures. Further, we test the usage of symmetry in design protocols, which is particularly important as many membrane proteins exist as homo-oligomers. We compare a soluble scoring function with a scoring function optimized for membrane proteins, RosettaMembrane. Both scoring functions recovered around half of the native sequence when completely redesigning membrane proteins. However, RosettaMembrane recovered the most native-like amino acid property composition. While leucine was overrepresented in the inner and outer-hydrophobic regions of RosettaMembrane designs, it resulted in a native-like surface hydrophobicity indicating that it is currently the best option for designing membrane proteins with Rosetta.

Keywords: Rosetta; RosettaMembrane; computational design; membrane proteins; membrane protein design; membrane protein engineering

Introduction

Membrane proteins comprise approximately 30% of all open reading frames of known genomes.[1] However, in the Protein Data Bank (PDB)[2] membrane proteins continue to be underrepresented. Membrane proteins, many of which are alpha-helical, include classes of proteins that are responsible for functions such as channel and transporter proteins, or signal transduction in receptors. Additionally, more than 60% of drugs target membrane proteins,[3] therefore insight to the structure and function of membrane proteins is valuable for the development

of treatment strategies for diseases such as cancer,[4,5] cardiac arrhythmia,[6,7] schizophrenia,[8,9] and many more.

Membrane proteins are difficult to structurally characterize because over-expression of the protein is typically toxic to bacterial cells,[3,10] resulting in low protein yields. Additionally, membrane proteins must be reconstituted into micelles, bicelles, nanodisks, or liposomes to provide a native-like environment. Often an extensive screening for the optimal detergents and lipids is needed for maximal solubility and stability.[3] However, membrane mimetics can have a destabilizing effect on the structure of the membrane protein. Finally, membrane proteins have inherent conformational dynamics,[11] which often requires engineering of a thermodynamically stabilized mutant for structural studies.

Challenges in membrane protein structure determination has resulted in limited available structural information for membrane proteins. In the PDB less than 3% of structures are membrane proteins. Approximately 700 unique membrane proteins structures have been deposited in the PDB[2,12] to date, which is a vast improvement to the structural information that was available nearly a decade ago, but far away from complete coverage of membrane protein folds. Computational modeling by de novo and comparative modeling can provide structural insights to membrane proteins without experimentally determined structures. However, in order to obtain more accurate models of membrane proteins, more high-resolution structures are needed to understand the physical basis of membrane protein folding and derive more accurate scoring functions.

The PDB is a depository of structure files which provides the knowledge-base for proteins of known structure to drive the development of accurate scoring functions and for rigorous testing of newly developed computational methods. As a result, methods for computational membrane protein structure prediction lag behind considerably, and computational design of function—an area of great success for soluble proteins in the past ten years—is largely absent for membrane proteins. However, the structures of many important membrane proteins have been determined at a stunning rate over the past 10 years[13–17] increasing the knowledge-base for scoring function development, providing higher-resolution structures for benchmarking, and yielding templates of important membrane protein classes to begin engineering.

Computational protein design is a difficult problem due to the large number of possible sequences for a particular protein backbone. Computational design tools aim to rapidly evaluate possible interactions between side-chains to determine likely sequences of low-energy. Some methods have an emphasis on calculations that evaluate electrostatics and solvation of a side-chain in its environment.[18–20] However the environment for membrane proteins is complicated and consideration for differences in membrane protein folding should be taken into account.[21] Additionally, these methods fail to consider features that many membrane proteins have that are important for function and membrane solubility.[22] Tools have been developed empirically to overcome the shortcomings of these calculations for membrane proteins. Walters and Degrado[23] developed idealized geometries and position-specific sequence propensities for helix-packing motifs most commonly seen in membrane proteins. Senes *et al.*[24] developed a potential based on the membrane depth dependent propensities of amino acids to predict if sequences would insert in the membrane.

The Rosetta software suite for biomolecular modeling and design has an impressive track record in the design of soluble proteins including the design of a de novo protein fold,[25] enzymes,[26–29] protein–protein interactions,[30–33] protein–small molecule interfaces,[34] and self-assembling materials.[35–38] The Monte Carlo search strategy that allows changes to amino acid identities during sampling combined with a multiscale knowledge-based scoring function that is optimized to capture structural features at the protein fold level as well as at atomic detail create a unique ability to engineer proteins that set Rosetta apart from other computational strategies. The scoring function and sampling methods used by Rosetta, however, are tailored for the needs of soluble-protein modelers; despite some progress in adapting it for membrane proteins, modeling abilities in membrane proteins lag behind those of soluble proteins.

Rosetta's knowledge-base has been derived in large part using statistical analysis of geometric arrangements within structures reported in the PDB. For protocols involving minimization, backbone torsion angles are randomly perturbed and rotational side-chain conformations are optimized for interactions including van der Waals, electrostatics, and hydrogen-bonding.[39,40] Interactions with the solvent are modeled implicitly by determining the likelihood of a certain amino acid type being in a particular burial state. Monte Carlo sampling combined with knowledge-based scoring functions are parameterized so that resulting models exhibit properties of proteins of known structure.[41] The membrane protein scoring function, RosettaMembrane, additionally considers the likelihood of an amino acid being in a particular membrane environment and burial state.[42,43]

Previously, Rosetta was used to completely redesign 108 soluble proteins. Designs recovered 51% of the native sequence in the protein core. The terms involving the Lennard–Jones potential and Lazaridis solvation drove the scoring function to design sequences that were native-like.[44] In the current study,

complete redesign of membrane proteins was benchmarked using RosettaMembrane,[42,43] and for comparison, the Rosetta scoring function for soluble proteins "Talaris."[45,46] Many membrane proteins like channels and transporters are functional homo-oligomers. In order to model membrane proteins in their native states and obtain correct representation of the surfaces and interfaces, one must consider how such a protein might symmetrically assemble. Therefore, homo-oligomeric membrane proteins were modeled with RosettaSymmetry[47] which is able to sample and rapidly score these larger assemblies while considering interface interactions between subunits.

One important application of membrane protein design is thermostabilization to facilitate structural characterization. Membrane proteins often require flexibility in order to perform their function.[11,48] By stabilizing a single conformation, one can reduce the flexibility, thus yielding a more ideal protein for experimental structure determination. Computational methods like RosettaDesign can propose an optimal sequence for a particular conformation by using information from known membrane protein structures. The proposed mutations in the optimized sequence could presumably lead to a thermostabilized membrane protein.

This study evaluates how well Rosetta recovers native sequences for membrane proteins when fully redesigned. We find that the methods for minimizing the structure prior to design play a role in native sequence recovery. Additionally, total sequence recovery was similar among different scoring functions; however, unsurprisingly, RosettaMembrane performed best in designing membrane proteins with native-like properties.

Results and Discussion

Initial energy minimization improves membrane protein design for low-resolution experimental structures

When benchmarking protein design algorithms, the question arises whether or not to minimize the starting experimental structure with the respective scoring function. The argument against minimization is that adjustment of backbone and side-chain coordinates to minimize energy will imprint a "memory" for the correct amino acid into the backbone coordinates. The native amino acid will score better as the backbone is positioned in such a way that the native amino acid can be placed in an energy minimum for the scoring function used. As a result, artificially inflated sequence recovery values might be reported. The counter argument is that energetic frustrations such as clashes in the starting structures that could be relieved with energy minimization might cause the design algorithm to prefer smaller, non-native amino acids in these locations. This is a particular

concern for membrane proteins where many structures of reduced resolution are deposited in the PDB. For soluble proteins the latter problem can be easily circumvented by benchmarking only on highest-quality protein structures with resolutions better than 2 Å.[44] However, the sparseness of membrane proteins in the PDB requires usage of lower-quality structures. Accordingly, we developed a protocol that applies an initial moderate energy minimization to resolve frustrations but avoids an aggressive optimization that might result in inflated sequence recovery values.

Without initial energy minimization, the sequence recovery of fully redesigned membrane proteins correlates with the resolution of the input structure such that low-resolution structures tend to have reduced sequence recovery (Fig. 1). For monomeric membrane proteins, the Pearson's correlation coefficient is strongly negative at -0.75 ($R^2 = 0.56$). For homo-oligomeric membrane proteins, the Pearson's correlation coefficient is -0.47 ($R^2 = 0.22$). When extrapolated, sequence recovery for a structure with 0 Å resolution is approximately 57 and 45% for monomeric and homo-oligomeric membrane proteins, respectively. Upon energy minimization, the correlation is absent independent of the Rosetta minimization protocol employed (Fig. 1). At the same time, we observe that average sequence recovery for monomeric membrane proteins improves from 31% without backbone energy minimization to 38, 49, 48, and 54% with the four Rosetta minimization protocols minimization with constraints (MWC), constrained to the start coordinate (CSC) relax, FastRelax, and Dualspace, respectively. For homo-oligomeric membrane proteins, average sequence recovery starts at 36% and results in 35, 48, 48, and 55%, respectively.

Our analysis indicates that both initial concerns have merit. A clear correlation between model resolution and sequence recovery is observed. Upon energy minimization this correlation vanishes. However, aggressive minimization protocols such as Dualspace, may inflate sequence recovery beyond what would be expected from the extrapolation to a membrane protein model with 0 Å resolution. To measure movement of the minimized protein from the original structure, the root-mean-square deviation (RMSD) was calculated. FastRelax and Dualspace move the protein beyond 1 Å $RMSD_{100}$,[49] whereas CSC attains similar average sequence recovery rates despite movement of less than 1 Å $RMSD_{100}$ during minimization [Fig. 1(E,F)]. We conclude that CSC, the limited energy minimization with a constraint to starting coordinates, is a good compromise to avoid over- and under-reporting algorithm accuracy.

Interestingly, for the highest resolution monomer, PDBID, the pack-only preparation resulted in an average sequence recovery of 42%, while MWC was 46%. Using the recommended CSC protocol, the average
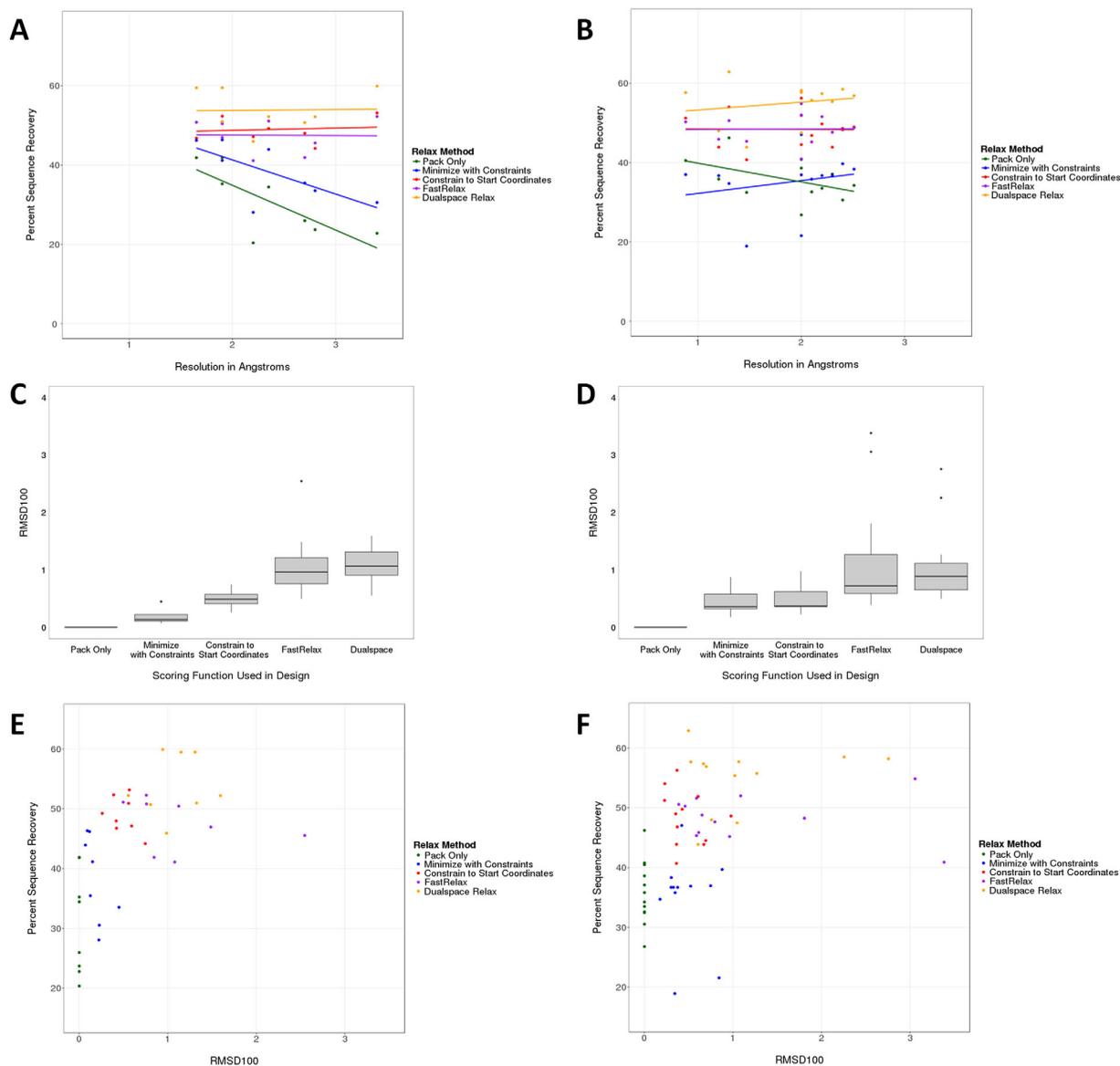
Figure 1. Sequence recovery for monomeric (A,C,E) and homo-oligomeric (B,D,F) sets. Various minimization methods were used to prepare crystal structures as input for Rosetta. When considering sequence recovery by resolution (A,B), pack-only and less stringent minimization (MWC) result in a correlation. CSC, FastRelax, and Dualspace minimization resulted in a consistently high sequence recovery independent of the initial structure resolution. The normalized, average movement of minimized structures for each minimization protocol (C,D) showed that FastRelax and Dualspace tend to move the protein further away from the starting structure. When examining sequence recovery by average movement (E,F), we find that pack-only and MWC had a larger range over low sequence recovery whereas protocols that allowed more movement during minimization, CSC, FastRelax, and Dualspace, yielded more consistently high sequence recovery rates. FastRelax and Dualspace in some cases moved the backbone further than 1 Å.

sequence recovery is 47% [Fig. S1(A), Supporting Information]. This indicates that any major clashes that typically lessen sequence recovery were resolved prior to minimization. Additionally, for the lowest resolution monomer, PDBID, the pack-only and MWC preparations resulted in sequence recoveries of 23 and 31%. However, after more flexible minimization strategies, CSC, FastRelax, and Dualspace, sequence recoveries increased to 53, 52, and 60%, respectively, indicating that perhaps major clashes were resolved once more flexibility was introduced.

For homo-oligomers, this analysis had a different finding. While most of the homo-oligomeric structures were of high-resolution more stringent minimization—CSC, FastRelax, or Dualspace—was required in order to achieve higher sequence recovery percentages [Fig. S1(B)]. This is likely due to an option used during symmetric relax which enables rigid body movement (see protocol capture in Supporting Information). Whereas the pack-only preparation would only move side-chains while MWC might constrain
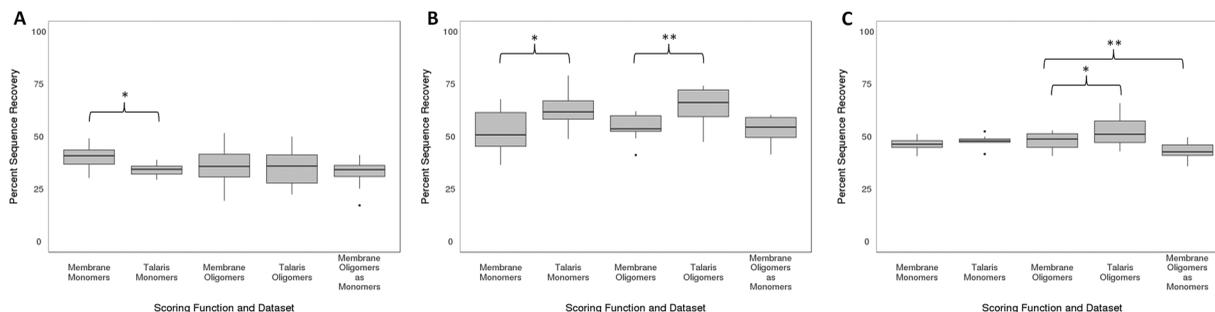
Figure 2. Percent of native sequence recovery for design of membrane proteins using various scoring functions. Boxplots show recovery of native sequence on the surface (A) and core (B) of the protein. RosettaMembrane (Membrane) designed monomeric proteins have a higher average surface recovery than Talaris. The total sequence recovery (C) shows that both scoring functions evaluated appear to have similar native sequence recovery percentages; however, core recovery is higher in Talaris which likely contributes to the total sequence recovery. When homo-oligomers were modeled as monomers, the total average sequence recovery rate was approximately 5% lower than the sequence recovery rate for design considering homo-oligomeric interfaces.

the minimization without considering the placement of the rigid bodies with respect to each other.

### Sequence recovery is highest in the core of the protein

To evaluate the performance of RosettaMembrane[42,43] redesigning membrane proteins, we compared the performance of the soluble scoring function Talaris.[45,46] The largest differences in score terms between RosettaMembrane and Talaris are the membrane-related terms that describe the membrane-specific environment (including burial state) and differences in solubility. We used Talaris to test how well Rosetta can design native-like membrane proteins in the absence of these membrane protein specific terms.

For both monomeric and homo-oligomeric sets, average core sequence recovery was higher with the Talaris scoring function when compared to RosettaMembrane [Fig. 2(B)]. Talaris had an average core sequence recovery of 63 and 65% for monomeric and homo-oligomeric datasets, respectively, compared to RosettaMembrane with 52 and 55%. A Wilcoxon signed rank test determined that the difference in percent core sequence recovery between Rosetta-Membrane and Talaris was significant for both monomers and homo-oligomers ($z = 2.49$, $P = 0.013$; $z = 3.04$, $P = 0.002$). Residues in the core are less influenced by the membrane environment than surface residues that are likely interacting with the lipid bilayer. Therefore, sampling and scoring in the core is driven by van der Waals packing interactions that are similar for membrane and soluble proteins. RosettaMembrane was derived from score 12, the scoring function that preceded Talaris. Membrane specific scoring terms were added. Meanwhile, score 12 evolved to Talaris through improvement of the electrostatic term, hydrogen bond terms, and reference energies.[45,46] These changes give rise to the improved core sequence recovery observed with the

Talaris energy function (Fig. 2) as amino acid interactions are modeled more precisely.

Surface sequence recovery for monomers improved in designs using RosettaMembrane (40%) when compared with Talaris [34%, Fig. 2(A)]. However for homo-oligomers, the average surface sequence recovery was 35% for both RosettaMembrane and Talaris. A Wilcoxon signed rank test determined that the difference in percent surface sequence recovery between RosettaMembrane and Talaris was significant for monomers ($z = 2$, $P = 0.046$), and not significant for homo-oligomers ($z = 0.69$, $P = 0.492$). RosettaMembrane models a membrane of fixed thickness implicitly. The higher surface sequence recovery observed with Rosetta-Membrane is attributed to the membrane-specific score terms that adjust the polarity of the environment (Fig. 2). However, the improvement in sequence recovery on the surface within Rosetta-Membrane when compared to Talaris is only moderate. We attribute this to the absence of specific interactions on the surface of the proteins that allow for the presence of only one specific amino acid. A more pronounced improvement is observed when comparing amino acid property composition between RosettaMembrane and Talaris (Fig. 3).

Finally, when evaluating the total sequence recovery in monomers, RosettaMembrane had an average of 46% while Talaris had an average of 48%. In homo-oligomers, the average total sequence recovery was calculated to be 48% for RosettaMembrane and 53% for Talaris. A Wilcoxon signed rank test revealed that the difference in percent total sequence recovery between RosettaMembrane and Talaris was not significant for monomers ($z = 0.81$, $P = 0.421$) while it was significant for homo-oligomers ($z = 2.1$, $P = 0.036$). When homo-oligomers were designed as monomers, the average percent native sequence recovery for surface [Fig. 2(A)] and core [Fig. 2(B)] were similar to that of homo-oligomers designed in a
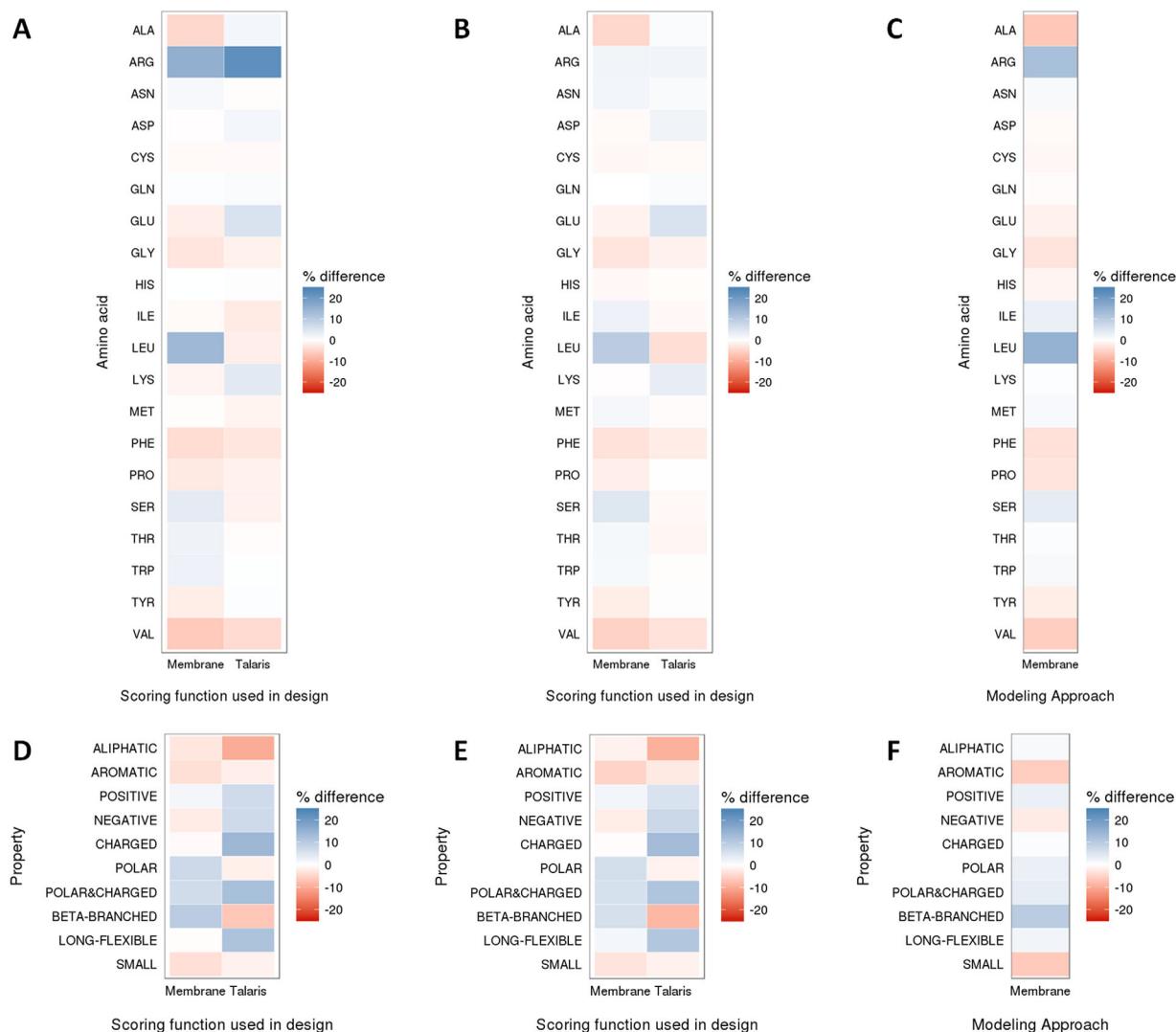
Figure 3. Heatmaps for composition of sequence (A–C) and amino acid properties (D–F) by percent difference of wild-type from design. Datasets evaluated were monomers (A,D), homo-oligomers (B,E), and homo-oligomers as monomers (C,F). Both RosettaMembrane (Membrane) and Talaris scoring functions have strong and weak amino acid recovery for different amino acids in the monomeric set (A,D). The homo-oligomeric set (B,E) performs similarly to the monomeric set for each respective scoring function. Finally, when homo-oligomers are designed as monomers using RosettaMembrane (C,F), the design is less native-like, but has a similar sequence composition as the homo-oligomeric design.

homo-oligomeric state. A Wilcoxon signed rank test confirmed there was no significant difference ($z = 1.24$, $P = 0.217$; $z = 0.33$, $P = 0.739$). However, the difference in percent total sequence recovery was found to be significant ($z = 2.77$, $P = 0.006$). This is likely due to a subset of residues not classified as either surface (less than or equal to 16 neighbors within a c-beta (C-β) distance of 10 Å) or core residues (more than 24 neighbors within a C-β distance of 10 Å) contributing to the difference in percent total sequence recovery differences.

We selected top models as representatives to better understand which residues were designed by mapping those residues on the structure. For both scoring functions, designed residues tended to be on the surface where residues would be lipid-exposed

(Fig. S3), in monomers (Fig. S4), and homo-oligomers (Fig. S5). Residues at the interface of sub-units [Figs. S3(C,E) and S5(C,F)] appear to be designed less frequently and result in core-like recovery indicating that design considers neighboring residues from different chains when using RosettaSymmetry.

Amino acid properties are most native-like in proteins designed using RosettaMembrane

Sequence recovery is a limited metric for design in that it only reports how much of the sequence changes from the native sequence. The percent difference in sequence composition (design percent composition—native percent composition) was calculated to further detail how design sequences differed

Computational Design using RosettaMembrane
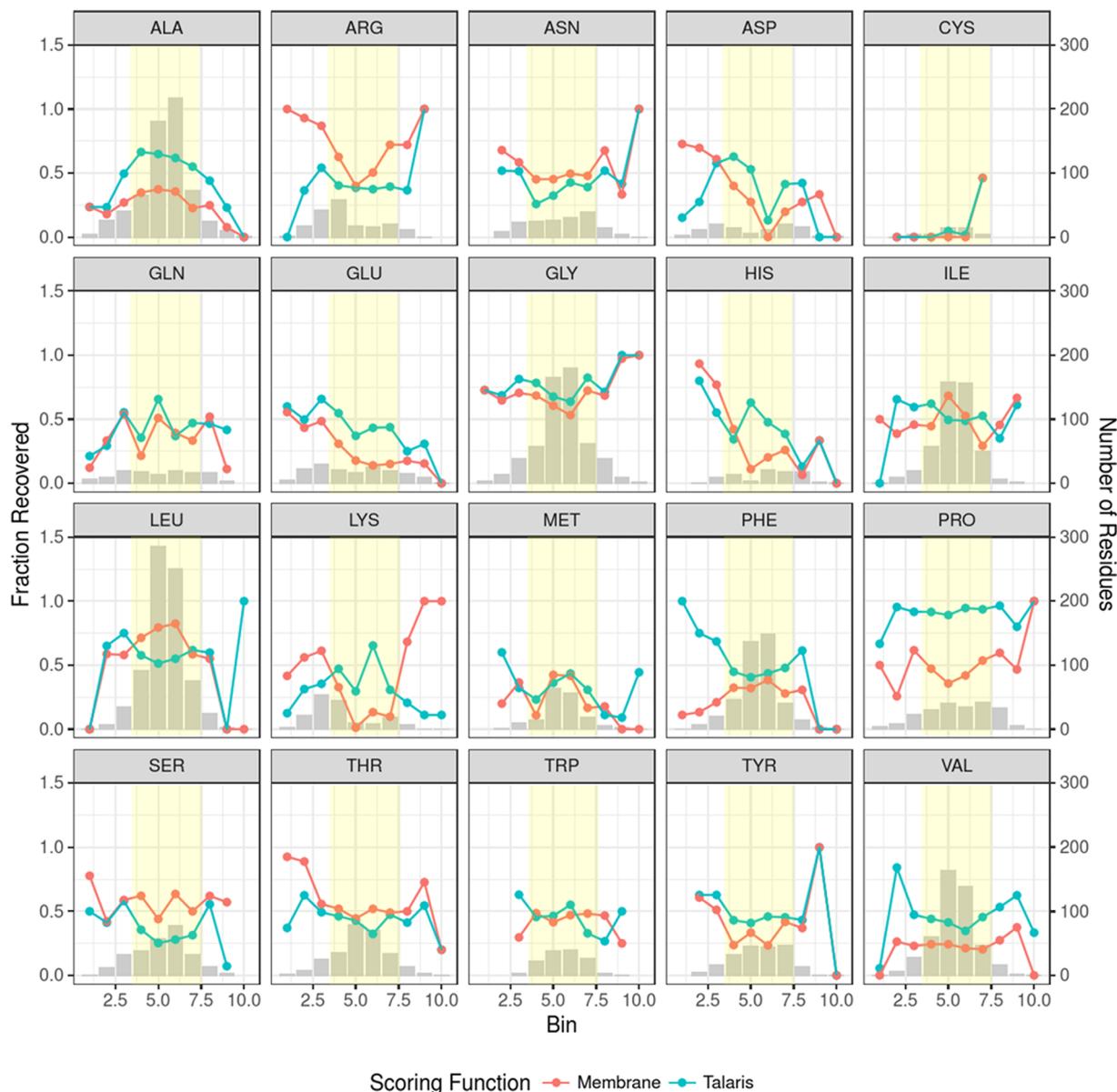
Figure 4. Fraction of sequence recovery for each amino acid with respect to distance from the membrane center. Bars indicate the raw number of residues observed at a particular distance bin (see Table I) from the membrane center. Dots indicate the fraction recovered at that particular distance bin and lines are not to infer a continuous dataset. The distance bins are discrete and the lines are only to aid the eye in following the trend between layers. The yellow box overlays bins of distance that would contain the inner and outer hydrophobic layers of the protein.

from native (Fig. 3). A negative percent difference (red) indicates that Rosetta introduces that particular amino acid less frequently than is observed in the native proteins in our dataset, while a positive percent difference (blue) indicates Rosetta introduces it more frequently. The average absolute deviation from native sequence composition for monomers was ±3.4% for RosettaMembrane, and ±2.8% for Talaris. For homo-oligomers, a similar trend was seen with ±2.5% for RosettaMembrane, ±1.6% for and Talaris.

Arginine was found more frequently in designs than in native membrane proteins. To visualize where arginines are found in native proteins compared to designs, we have plotted the fraction recovered

(Fig. 4) and number of occurrences (Fig. 5) of arginines in all native, best-scoring RosettaMembrane designs, and best-scoring Talaris designs with respect to their position in the membrane layer. This representation can also be seen broken down by monomeric (Figs. S6 and S8) and homo-oligomeric datasets (Figs. S7 and S9). In Figure 4, the fraction recovered drops in the inner hydrophobic layer for RosettaMembrane designs. In Figure 5, it is clear that Talaris is solubilizing the designs as an increase in occurrence of arginine is seen in the inner and outer hydrophobic regions.

However, for RosettaMembrane, only the outer hydrophobic and interface regions have an increase
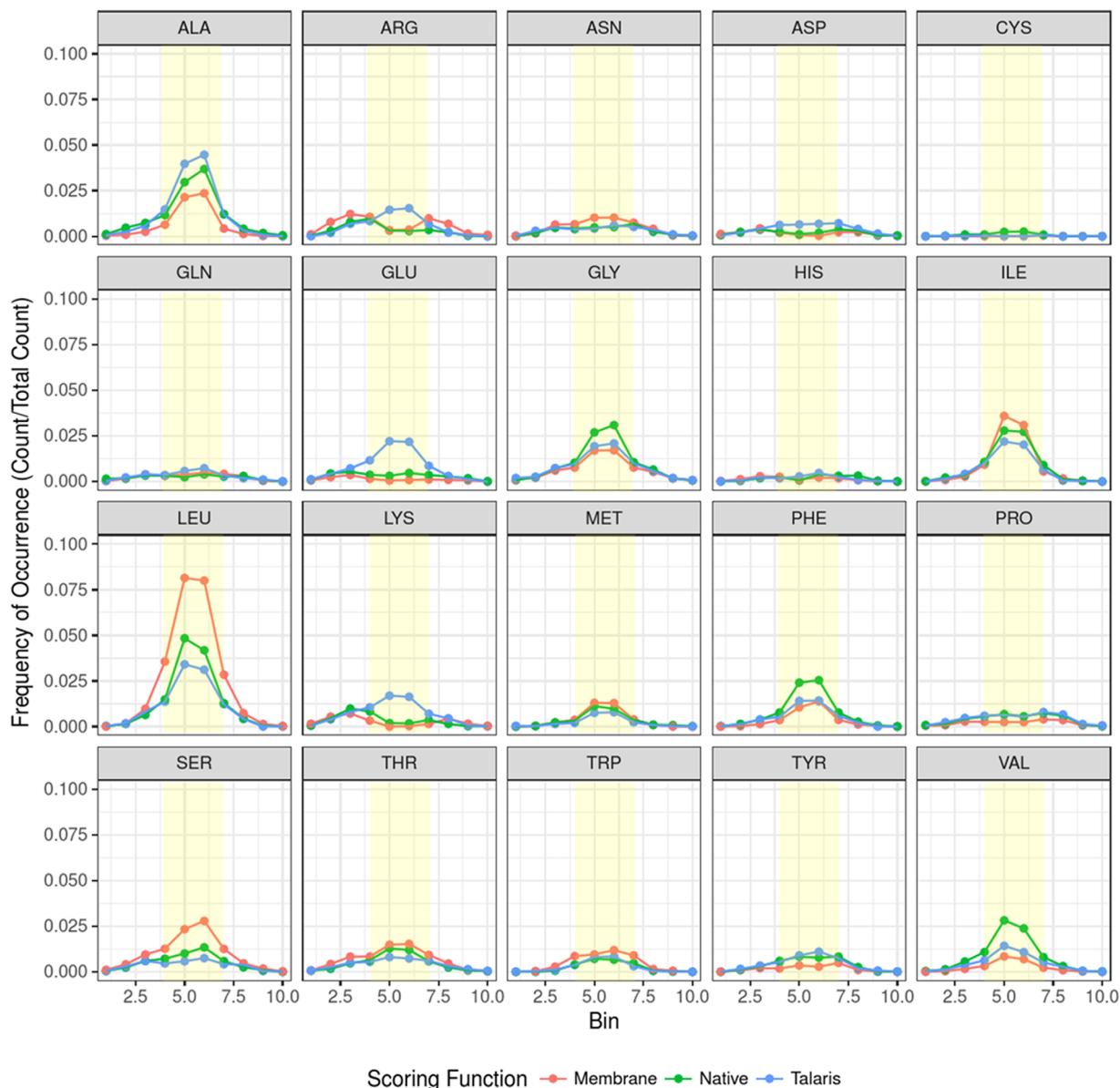
Figure 5. Frequency of occurrence for each amino acid by membrane layer. Bins are a range of distances from the membrane center (see Table I). Dots indicate the frequency of occurrence of an amino acid seen at a particular distance from the membrane center, and lines are not to infer a continuous dataset. The distance bins are discrete and the lines are only to aid the eye in following the trend between layers. The yellow box overlays bins of distance that would contain the inner and outer hydrophobic layers of the protein.

of occurrence. Additionally, this is more pronounced in the monomeric dataset (Fig. S8), perhaps indicating that there is an additional cost of designing in a bulky residue at a protein-protein interface region (Fig. S9). Talaris adds charged residues such as arginine, aspartate, glutamate, and lysine on the surface and in the inner and outer hydrophobic regions, as expected, to solubilize the protein.

The most striking difference for RosettaMembrane designs when compared with native membrane protein sequences was that the amino acid composition is shifted toward leucine residues (Fig. 4) while other hydrophobic amino acids such as phenylalanine, valine, and alanine, have a lower than

native probability. This indicates that RosettaMembrane has a bias toward leucine at the cost of other hydrophobic amino acids. The fraction recovered for leucine in the inner and outer hydrophobic regions ranged from 58 to 82% while valine and alanine had recoveries in the ranges of 20–24 and 23–37%, respectively (Fig. 4). When the number of occurrences of leucine in native proteins and designed proteins was plotted with respect to their position in the membrane layer, leucine was found to be overrepresented by 1.9-fold in the inner and outer hydrophobic regions for RosettaMembrane designs (Fig. 5). An increase is also seen in both datasets with a 2.2-fold increase for monomers (Fig. S8), and

Computational Design using RosettaMembrane
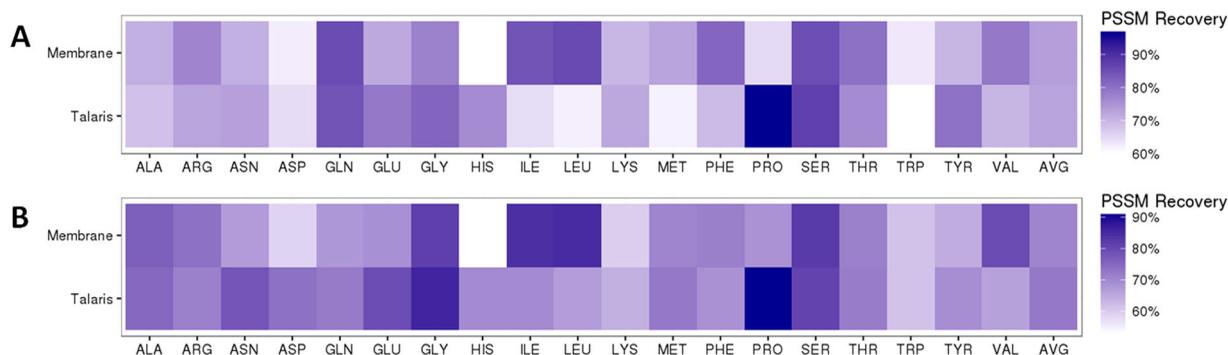
Figure 6. Heatmaps for PSSM recovery for the monomeric set (A) and homo-oligomeric set (B). The PSSM recovery for each scoring function is similar when comparing the monomeric set to the homo-oligomeric set. RosettaMembrane (Membrane) has limitations for recovering histidine and proline, but shows improved recovery for isoleucine, leucine, valine, and phenylalanine.

a 1.6-fold increase for homo-oligomers (Fig. S9). Additionally, RosettaMembrane designs valine and alanine less frequently than what is seen in native proteins in the inner and outer hydrophobic regions by 3.4- and 1.6-fold, respectively. This further supports that in the hydrophobic regions, valine and alanine are replaced by leucine in RosettaMembrane designs.

Sequence recovery may be too crude of an analysis to determine the extent of which designed proteins have changed. In addition to calculating recovery of native amino acid identities, we calculated the percent difference in the composition of amino acids grouped by properties such as polarity and charge (design percent composition—native percent composition). Here, the average absolute deviation from native amino acid property composition in monomers was 3.9% for RosettaMembrane, and 7.4% for Talaris, while in homo-oligomers, it was 3.4% for RosettaMembrane, and 7.3% for Talaris. When considering the composition of all amino acid properties, RosettaMembrane resulted in proteins with more native-like properties in both monomeric and homo-oligomeric sets [Fig. 3(D,E)]. The differences in sequence composition between native and designed proteins are primarily caused by mutations on the protein surface as core sequence recovery is high for both, Talaris and RosettaMembrane. Recall that surface sequence recovery rates of monomers averaged at 40% for RosettaMembrane designs, whereas Talaris had lower averages of 34 and 38%, respectively [Fig. 2(A)]. However, when comparing the difference in amino acids that are aliphatic [Fig. 3(D,E)], RosettaMembrane is near native with a percent difference of nearly −3% in monomers and −1% in homo-oligomers whereas Talaris had a percent difference near −10% for both monomers and homo-oligomers.

To further investigate which amino acid mutations would be tolerated by evolution, position specific scoring matrix (PSSM) recovery[50] was calculated using the uniref50membrane database. Because PSSM recovery is considering all tolerated amino acids that have been seen in known sequences, PSSM recovery will be higher than sequence recovery alone.[51] In monomers, RosettaMembrane had an average PSSM recovery of 73% while Talaris had a recovery of 72% [Fig. 6(A)]. In homo-oligomers, RosettaMembrane had an average PSSM recovery of 69% while Talaris was at 70% [Fig. 6(B)]. Despite using a membrane specific database, the PSSM recovery did not favor RosettaMembrane designs.

RosettaMembrane designs a native-like hydrophobicity gradient and predicted $\Delta G_{transfer}$

The HotPatch server[52] was used to visualize the relative hydrophobicity on the surface of proteins (Fig. S10). For Talaris, despite having a similar sequence composition as native structures [Fig. 3(A,B)], the resulting designs had a noticeably different surface composition. This is supported by the sequence recovery analysis where core sequence recovery is typically much higher than the surface sequence recovery [Fig. 2(A,B)]. Representative design models selected for monomers show that both scoring functions resulted in a large amount of surface residues being redesigned [Fig. S1(A,B)]. Design models of assembled homo-oligomers highlight a similar feature; however, design at the interface of subunits is typically more restricted and thus more core-like [Fig. S1(C–F)]. For Talaris, the surfaces of the majority of the protein designs were covered in hydrophilic residues (Fig. S10) as the scoring function attempted to solubilize the surface of the protein. However, RosettaMembrane resulted in a designed protein with a native-like hydrophobicity gradient on the surface. These models had more strongly hydrophobic and hydrophilic areas whereas native surfaces had moderate hydrophobic and hydrophilic regions.

The positioning of proteins in membrane (PPM) server[53,54] was used to predict the $\Delta G_{transfer}$ for both monomeric and homo-oligomeric sets (Fig. 7). The server tends to predict that integral membrane proteins and peptides have a $\Delta G_{transfer}$ between −400
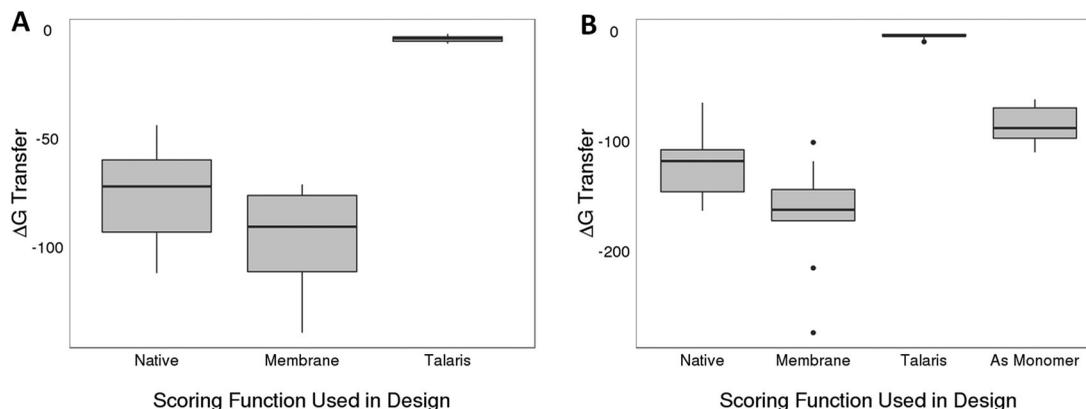
Figure 7. Predicted $\Delta G_{transfer}$ for designs from RosettaMembrane (membrane) and Talaris. For both monomeric (A) and homo-oligomeric (B) sets, the membrane scoring function resulted in more native-like $\Delta G_{transfer}$ values in comparison to Talaris. For the soluble scoring function Talaris, the value was nearly zero indicating it would likely not partition into the membrane. Finally, the homo-oligomeric design took into account surfaces when assembled as an homo-oligomer, resulting in more native-like values.

and −10 kcal/mol.[54] For our datasets, the native proteins were in the range of −44 to −164. Designs by the RosettaMembrane scoring function were near and above native in a range of −71 to −275 whereas designs by Talaris were near zero indicating that the designed protein would not be membrane soluble.

### *RosettaMembrane replaces other hydrophobics with leucine*

RosettaMembrane chooses leucine over other hydrophobic amino acids. Although leucine may be ideal for the particular membrane environment modeled in Rosetta, this may not be ideal biologically as it does not account for asymmetry and heterogeneity of the membrane. A previous study showed leucine to be the most frequent amino acid in the inner hydrophobic and outer hydrophobic layers of the membrane.[42] Because leucine has such a high frequency compared to other amino acids, it scores quite favorably in RosettaMembrane and is overrepresented in designs often replacing native, hydrophobic amino acids [Figs. 3(A) and 5].

To further investigate how leucine might replace hydrophobic amino acids such as alanine, valine, and phenylalanine, we mapped their occurrences onto the structures to understand where each scoring function would typically place them compared to where they are found on the native membrane protein. For both monomers and homo-oligomers, native membrane proteins have alanine in the core as well as on the surface (Fig. S11). Both scoring functions typically placed alanine in the core of the protein and RosettaMembrane had a lower alanine sequence composition than native membrane proteins. In homo-oligomers, very few alanine occur on the surface of the protein that would be lipid-exposed, and very few are seen in the interface between subunits, likely due to alanine's small size.

Designs from both scoring functions resulted in fewer valine and phenylalanine. Both residues are hydrophobic and, in the case of RosettaMembrane, were likely replaced by leucine. Valine was typically designed in the core of the protein regardless of scoring function; however, in homo-oligomers, Talaris does place valine in the core-like interface between subunits more frequently than RosettaMembrane (Fig. S12). Despite phenylalanine typically occurring in the interface and inner and outer hydrophobic layers, fewer phenylalanines are seen on the surface of designs from both scoring functions (Fig. S13). This suggests that leucine's abundance in these layers overshadows the presence of phenylalanine in native membrane proteins. As a comparison, arginine, was also highlighted onto structures (Fig. S14). Although the percent difference in composition was like that of leucine, the number of occurrences (Fig. 4) was much lower, so the effect was pronounced.

A closer look at trends seen in designs

Core residues have a better chance of recovering the native amino acid. For example, the core of PDBID has several residues surrounding asparagine 64 that remain the same for both scoring functions [Fig. 8(A–C)]. The native core is likely well-packed with favorable hydrophobicity. The largest differences among designs are expected at the surface of the protein. While RosettaMembrane is designing toward an optimal hydrophobicity gradient so that the protein can partition in the membrane, Talaris is designing toward a soluble protein [Fig. 8(D–F)]. For this reason, many of the surface residues that were designed by Talaris are charged when the native protein would likely not tolerate multiple charged residues embedded in the membrane. As previously noted, an interesting finding was the abundance of leucine on the surface of proteins designed using RosettaMembrane. In many cases,

Computational Design using RosettaMembrane
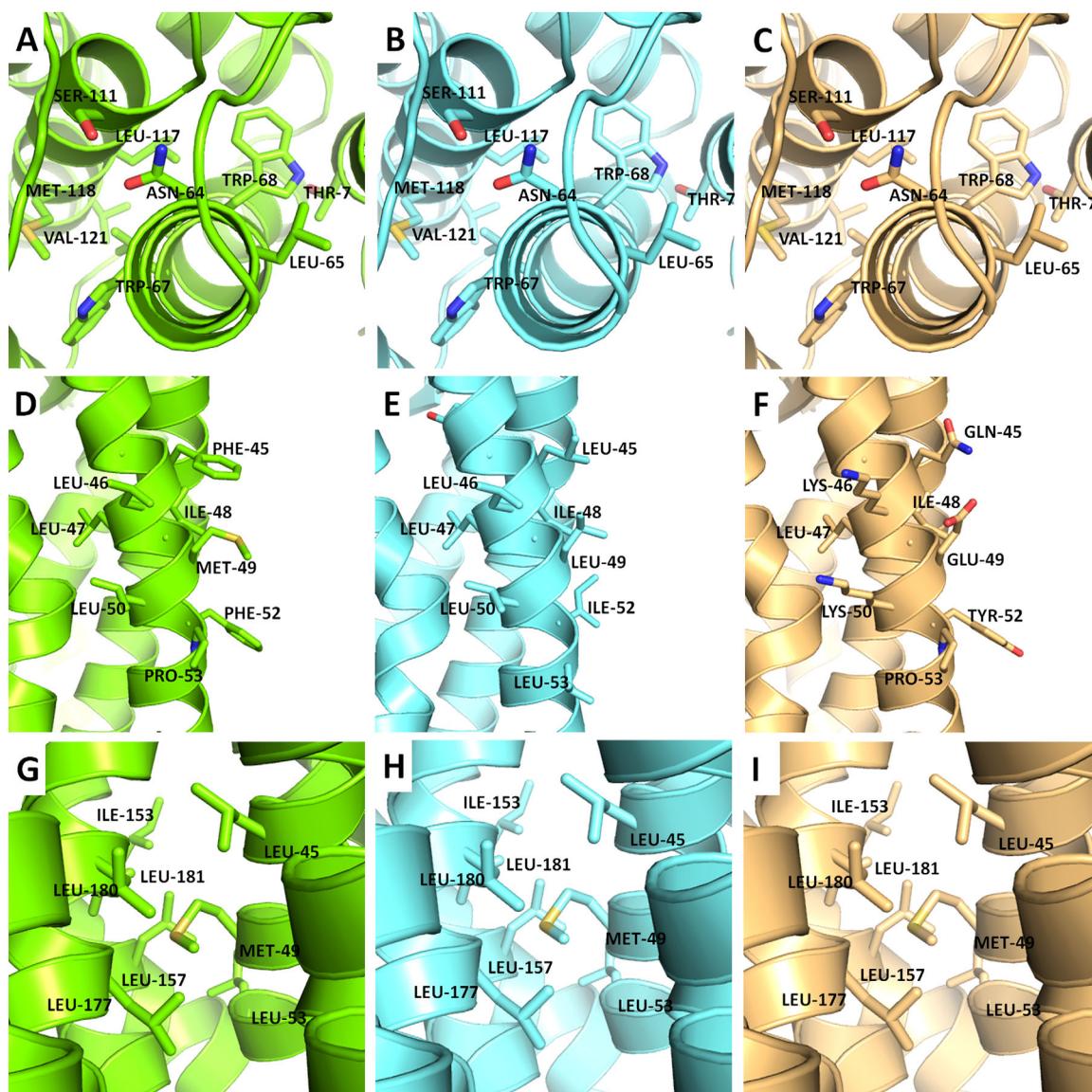
Figure 8. Atomic detail of designs compared to wild-type. A closer look at typical interactions at the core (A–C), surface (D–F), and homo-oligomeric interface (G–I). Representative cases were selected from PDBID (A–C), PDBID (D–F), and PDBID (G–I). Green represents respective minimized native, aquamarine is RosettaMembrane, and light orange is Talaris.

native hydrophobic residues, such as phenylalanine at position 45 and methionine 49 [Fig. 8(D–F)], were replaced by leucine.

In homo-oligomers, the surface and core are similar to that in monomers; however, the homo-oligomers have interface regions between the subunits. The interface regions should be designed similarly to the core in that they are surrounded by neighboring residues, provided that distance is close enough to be considered buried, despite those residues residing on a different chain. As expected, these regions, when well packed, will remain the native amino acid for both scoring functions [Fig. 8(G–I)].

RosettaMembrane designs membrane proteins that capture native-like properties. We have reported *in silico* sequence redesign experiments using two different Rosetta scoring functions. Despite having similar

sequence recoveries (Fig. 2), Talaris did not, as expected, appropriately design the surface. RosettaMembrane was developed to implicitly model an appropriate hydrophobic gradient that is often seen in native membrane proteins.[43] RosettaMembrane designed a hydrophobic gradient that was native-like (Fig. S10). However, an artifact of designing in RosettaMembrane was the overuse of leucine because of their high frequency at various layers in the membrane (Figs. 5 and 9).

Also indicative of a native-like surface, the $\Delta G_{transfer}$ was above or near native for RosettaMembrane designs, whereas Talaris designs were near zero (Figs. 6 and 7). Interestingly, although both scoring functions resulted in a similar amino acid composition [Fig. 3(A,B)], the difference in composition of amino acid properties made it evident that RosettaMembrane designed in amino acids that
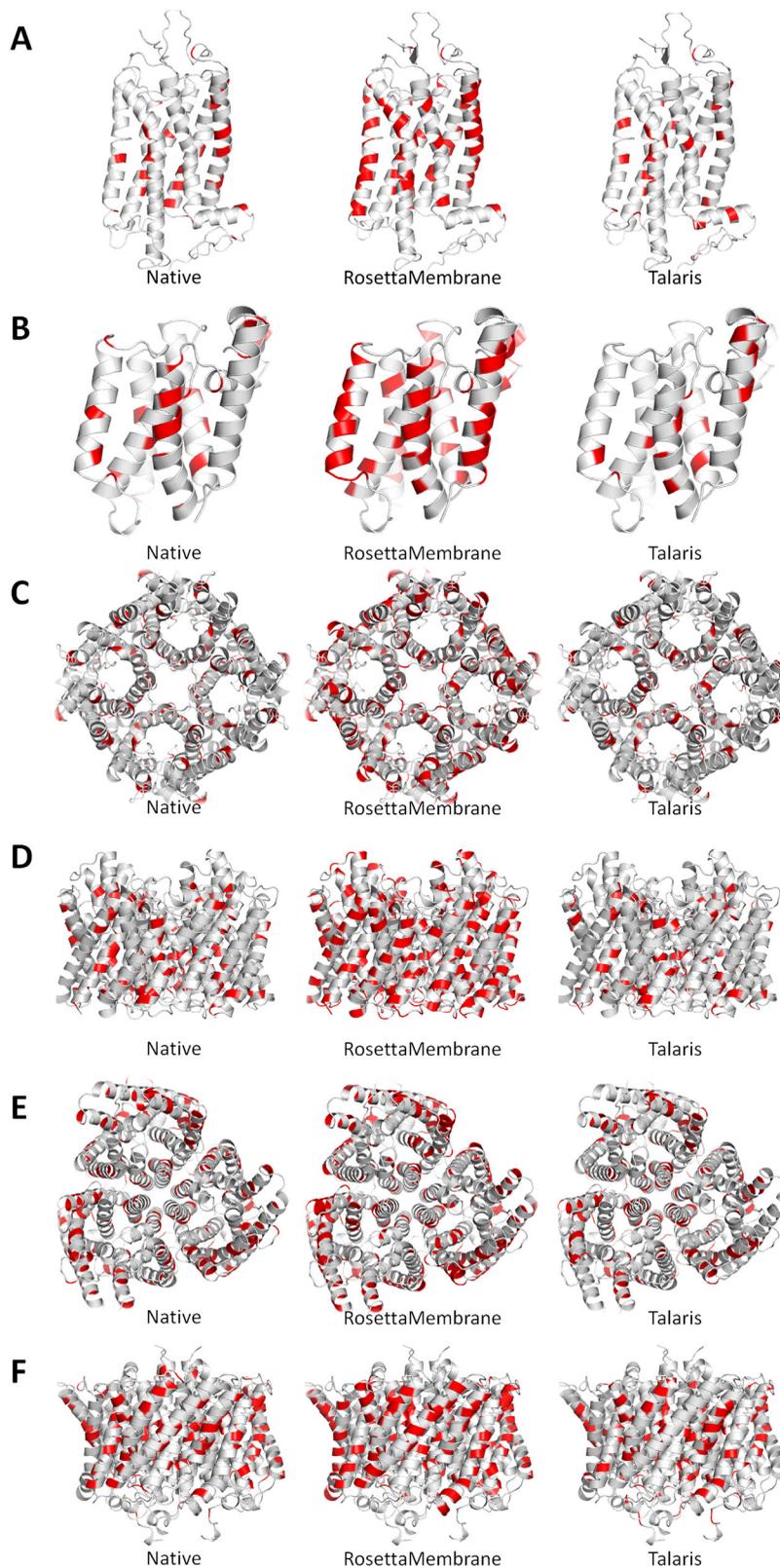
Figure 9. Visualization of leucine on models. Top models were selected to visualize where leucines occur in monomers PDBID (A), PDBID (B) and in homo-oligomers PDBID (C, top; D, down) and PDBID (E, top; F, down). Native structures (left) were compared to representative models of proteins designed using RosettaMembrane and Talaris. RosettaMembrane designs proteins with an abundance of leucine at multiple layers of the membrane and surface residues. In homo-oligomers, leucine is also seen in regions that are buried at the interface between subunits and in the core of the protein.

Table I. *Layers of the Membrane Represented by Bins. Calculated Distances from the Membrane Center have been Binned to Aid in Visualization of Data. Bins have been Defined by the Layers Described by Yarov Yarovoy et al.*[42]

Bin number	Distance (Å) from membrane center	Membrane layer
1	−40 to −30	Water
2	−30 to −24	Polar
3	−24 to −18	Interface
4	−18 to −12	Outer hydrophobic
5	−12 to 0	Inner hydrophobic
6	0 to 12	Inner hydrophobic
7	12 to 18	Outer hydrophobic
8	18 to 24	Interface
9	24 to 30	Polar
10	30 to 40	Water

were aliphatic, charged, or long and flexible more realistically [Fig. 3(D–F)]. Additionally, when evaluating PSSM recovery, RosettaMembrane's strength was recovering hydrophobic residues such as isoleucine, leucine, valine, and phenylalanine (Fig. 6). Despite both of the scoring functions resulting in similar amino acid composition, design using RosettaMembrane results in membrane protein designs with more native-like properties.

RosettaMembrane and symmetry can be used in conjunction to model obligate homo-oligomeric membrane proteins

Because many membrane proteins are functional as homo-oligomers, it is important the RosettaDesign algorithm works well with RosettaSymmetry so that both the internal energy of all subunits and interface interactions are taken into account during the design process. RosettaSymmetry is ideal for larger, symmetric systems because the subunits in homo-oligomers are moved in the same way, which enables the sampling process to rapidly occur. The homo-oligomeric set performed similarly to the monomeric set in amino acid composition and slightly better in recovering native-like properties. To ensure this comparison was not an artifact of the sets of proteins, the homo-oligomeric set was modeled as monomers in a separate design experiment. This revealed that although the patterns for amino acid composition were similar, the monomeric representation deviated further from the native [Fig. 3(B,C)] indicating that homo-oligomeric modeling result in more native-like designs.

Conclusion

This study illustrates that with minimized structures, membrane proteins have core sequence recovery rates of 52–63% for monomeric membrane proteins and 53–65% for homo-oligomeric membrane proteins. These rates are similar to the 51% core sequence recovery rates calculated from a large soluble protein set.[44] The chance of designing a position

with the correct amino acid identity is roughly 5% (selecting the correct amino acid out of 20), so a recovery of approximately 50% indicates the algorithm is working well. Increasing sequence recovery even further would involve extensive backbone minimization and/or an improved scoring function. We find that PSSM recovery (here averaging around 70%) is a more reliable metric because the recovery tolerates mutations that have been seen in evolution. Additionally, to avoid minimizing structures that imprint the native sequence, we recommend using CSC to prepare structures for design as this reduces backbone RMSD from native during minimization and still achieves moderately high sequence recovery for a range of starting resolutions.

While RosettaMembrane designs native-like surface hydrophobicity, it is important to note that RosettaMembrane has a tendency to favor leucine over other hydrophobic residues at these positions. This may be due to high occurrence of leucine for proteins in the original training set. An updated RosettaMembrane scoring function with a larger, more diverse, and higher resolution membrane protein knowledge-base may help dampen this bias. Finally, as membrane protein structures have varying membrane thicknesses, an accurate depiction of the hydrophobicity gradient during modeling and design of membrane proteins in Rosetta could improve the quality of native-like designs even further.

Methods

A set of 20 membrane proteins with resolutions ranging from 0.88 to 3.4 Å was compiled. Twelve of these membrane proteins are modeled as homo-oligomers (Table S1). All of the coordinates were obtained from the PDB. Solvent and ions were excluded for the duration of this study. Span files that specify the trans-membrane spanning region were created using information obtained from the protein data bank of transmembrane proteins (PDBTM).[55] The symmetry definition files were created using the noncrystallographic symmetry mode in the make_symmdef_file.pl script provided in Rosetta. This mode calculates the point symmetries using the homo-oligomers present in the PDB file, or from symmetry mates generated in Pymol from the original PDB file. The RosettaScripts eXtensible markup language (XML) scripting language framework[33] from the Rosetta week 52 build was used for all of the protocols tested. The Rosetta software suite is publically accessible and free for noncommercial use.

Preminimization trials

Five minimization protocols were tested on this benchmark set: pack-only where the backbone is not perturbed and only the side-chains conformations are optimized; minimize with constraints (MWC) where harmonic constraints are used to minimize both the backbone and side-chains to within 0.5 Å of the starting position[56] (used to prepare structures for

thermostability calculations[57]); FastRelax with an added CSC which is similar to MWC, but ramps the weight of the repulsive term to allow for more flexibility. FastRelax, the standard minimization protocol; and DualSpace relax[58] which uses a combination of internal and Cartesian minimization. Three of these protocols, CSC, FastRelax, and Dualspace, were set up using the FastRelax mover in Rosetta Scripts and can also be set up using the relax application by including command-line options appropriate for each protocol. For pack-only and MWC, the appropriate applications and options were used (please see a complete, detailed protocol capture in Supporting Information, parts 1a, 1b).

Full redesign to assess preminimized structures

Full redesign, where all canonical amino acids identities are allowed to be sampled at each position, was performed on the preminimized membrane protein sets. For each minimization protocol, two to three top models by score and RMSD for each membrane protein were chosen as the input models for full redesign to introduce backbone diversity. Full redesign was set up using PackRotamersMover and the SymPackRotamersMover, where appropriate, to generate design models of each minimized model. The top ten percent models by score were chosen for sequence recovery analysis (protocol capture, Supporting Information, parts 1a, 1b).

Full redesign using various scoring functions

Full redesign was performed on the top three models by score and RMSD from the CSC protocol. The scoring functions tested were the RosettaMembrane full atom smoothed potential (membrane_highres_-Menv_smooth.wts) and Talaris (talaris2013.wts). Full redesign was set up using PackRotamersMover and SymPackRotamersMover, where appropriate, to generate design models from each selected minimized model. The top scoring ten percent models were used to calculate sequence recovery of the native protein sequence (protocol capture, Supporting Information, parts 2a, 2b).

Sequence analysis of redesigned proteins

The top 10% of designs by score were analyzed. Native sequence recovery was calculated for the full protein, core residues (a residue with at least 24 contacts within a C-β distance of 10 Å), and surface residues (a residue with at most 16 contacts within a C-β distance of 10 Å) using the Sequence Recovery application in Rosetta. Additionally, we determined whether the scoring functions reproduced native-like amino acid composition.

References

1. Tan S, Hwee TT, Chung MCM (2008) Membrane proteins and membrane proteomics. Proteomics 8:3924–3932.
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242.
3. Arinaminpathy Y, Khurana E, Engelman DM, Gerstein MB (2009) Computational analysis of membrane proteins: the largest class of drug targets. Drug Discov Today 14:1130–1135.
4. Lemmon MA, Schlessinger J (2010) Cell signaling by receptor tyrosine kinases. Cell 141:1117–1134.
5. Jura N, Endres NF, Engel K, Deindl S, Das R, Lamers MH, Wemmer DE, Zhang X, Kuriyan J (2009) Mechanism for activation of the EGF receptor catalytic domain by the juxtamembrane segment. Cell 137:1293–1307.
6. Moss AJ, Kass RS (2005) Long QT syndrome: from channels to cardiac arrhythmias. J Clin Invest 115: 2018–2024.
7. Wang Q, Curran ME, Splawski I, Burn TC, Millholland JM, VanRaay TJ, Shen J, Timothy KW, Vincent GM, de Jager T, Schwartz PJ, Towbin JA, Moss AJ, Atkinson DL, Landes GM, Connors TD, Keating MT (1996) Positional cloning of a novel potassium channel gene: KVLQT1 mutations cause cardiac arrhythmias. Nat Genet 12:17–23.
8. Meisenzahl EM, Schmitt GJ, Scheuerecker J, Möller H (2007) The role of dopamine for the pathophysiology of schizophrenia. Int Rev Psychiatry 19:337–345.
9. Conn PJ, Lindsley CW, Jones CK (2009) Activation of metabotropic glutamate receptors as a novel approach for the treatment of schizophrenia. Trends Pharmacol Sci 30:25–31.
10. Wagner S, Bader ML, Drew D, de Gier J-W (2006) Rationalizing membrane protein overexpression. Trends Biotechnol 24:364–371.
11. Bowie JU (2001) Stabilizing membrane proteins. Curr Opin Struct Biol 11:397–402.
12. White SH. Membrane Proteins of Known 3D Structure. Available at: http://blanco.biomol.uci.edu/mpstruc/.
13. Sanders CR, Sönnichsen F (2006) Solution NMR of membrane proteins: practice and challenges. Magn Reson Chem 44:24–40.
14. Wiener MC (2004) A pedestrian guide to membrane protein crystallization. Methods 34:364–372.
15. Loll PJ (2003) Membrane protein structural biology: The high throughput challenge. J Struct Biol 142:144–153.
16. White SH (2004) The progress of membrane protein structure determination. Protein Sci 13:1948–1949.
17. Wu H, Wang C, Gregory KJ, Han GW, Cho HP, Xia Y, Niswender CM, Katritch V, Meiler J, Cherezov V, Conn PJ, Stevens RC (2014) Structure of a class C GPCR metabotropic glutamate receptor 1 bound to an allosteric modulator. Science 344:58–65.
18. Pokala N, Handel TM (2004) Energy functions for protein design I: Efficient and accurate continuum electrostatics and solvation. Protein Sci 13:925–936.
19. Marshall SA, Vizcarra CL, Mayo SL (2005) One- and two-body decomposable Poisson-Boltzmann methods for protein design calculations. Protein Sci 14:1293–1304.
20. Vizcarra CL, Zhang N, Marshall SA, Wingreen NEDS, Zeng C, Mayo SL (2008) An improved pairwise decomposable finite-difference Poisson–Boltzmann method for computational protein design. J Comput Chem 29: 1153–1162.
21. Senes A (2011) Computational design of membrane proteins. Curr Opin Struct Biol 21:460–466.
22. Perez-Aguilar JM, Saven JG (2012) Computational design of membrane proteins. Struct Des 20:5–14.
23. Walters RFS, Degrado WF (2006) Helix-packing motifs in membrane proteins. Proc Natl Acad Sci USA 103: 13658–13663.

24. Senes A, Chadi DC, Law PB, Walters RFS, Nanda V, Degrado WF (2007) Ez, a depth-dependent potential for assessing the energies of insertion of amino acid side-chains into membranes: derivation and applications to determining the orientation of transmembrane and interfacial helices. J Mol Biol 366:436–448.

25. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, Baker D (2003) Design of a novel globular protein fold with atomic-level accuracy. Science 302:1364–1368.

26. Röthlisberger D, Khersonsky O, Wollacott AM, Jiang L, DeChancie J, Betker J, Gallaher JL, Althoff EA, Zanghellini A, Dym O, Albeck S, Houk KN, Tawfik DS, Baker D (2008) Kemp elimination catalysts by computational enzyme design. Nature 453:190–195.

27. Jiang L, Althoff EA, Clemente FR, Doyle L, Rothlisberger D, Zanghellini A, Gallaher JL, Betker JL, Tanaka F, Barbas CF, Hilvert D, Houk KN, Stoddard BL, Baker D (2008) De novo computational design of retro-aldol enzymes. Science 319:1387–1391.

28. Korkegian A, Black M, Baker D, Stoddard BL (2005) Computational thermostabilization of an enzyme. Science 308:857–860.

29. Siegel JB, Zanghellini A, Lovick HM, Kiss G, Lambert AR, Clair JLS, Gallaher JL, Hilvert D, Gelb MH, Stoddard BL, Houk KN, Michael FE, Baker D (2010) Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction. Science 329:309–313.

30. Joachimiak LA, Kortemme T, Stoddard BL, Baker D (2006) Computational design of a new hydrogen bond network and at least a 300-fold specificity switch at a protein-protein interface. J Mol Biol 361:195–208.

31. Kortemme T, Joachimiak LA, Bullock AN, Schuler AD, Stoddard BL, Baker D (2004) Computational redesign of protein-protein interaction specificity. Nat Struct Mol Biol 11:371–379.

32. Strauch E-M, Fleishman SJ, Baker D (2014) Computational design of a pH-sensitive IgG binding protein. Proc Natl Acad Sci USA 111:675–680.

33. Fleishman SJ, Leaver-Fay A, Corn JE, Strauch E-M, Khare SD, Koga N, Ashworth J, Murphy P, Richter F, Lemmon G, Meiler J, Baker D (2011) RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite. PLoS One 6:e20161.

34. Tinberg CE, Khare SD, Dou J, Doyle L, Nelson JW, Schena A, Jankowski W, Kalodimos CG, Johnsson K, Stoddard BL, Baker D (2013) Computational design of ligand-binding proteins with high affinity and selectivity. Nature 501:212–216.

35. King NP, Sheffler W, Sawaya MR, Vollmar BS, Sumida JP, André I, Gonen T, Yeates TO, Baker D (2012) Computational design of self-assembling protein nanomaterials with atomic level accuracy. Science 336:1171–1174.

36. King NP, Bale JB, Sheffler W, McNamara DE, Gonen S, Gonen T, Yeates TO, Baker D (2014) Accurate design of co-assembling multi-component protein nanomaterials. Nature 510:103–108.

37. Fortenberry C, Bowman EA, Proffitt W, Dorr B, Combs S, Harp J, Mizoue L, Meiler J (2011) Exploring symmetry as an avenue to the computational design of large protein domains. J Am Chem Soc 133:18026–18029.

38. Eisenbeis S, Proffitt W, Coles M, Truffault V, Shanmugaratnam S, Meiler J, Höcker B (2012) Potential of fragment recombination for rational design of proteins. J Am Chem Soc 134:4019–4022.

39. Rohl CA, Strauss CEM, Misura KMS, Baker D (2004) Protein structure prediction using Rosetta. Methods Enzymol 383:66–93.

40. Schueler-Furman O (2005) Progress in modeling of protein structures and interactions. Science 638:638–642.

41. Kaufmann KW, Lemmon GH, Deluca SL, Sheehan JH, Meiler J (2010) Practically useful: what the Rosetta protein modeling suite can do for you. Biochemistry 49:2987–2998.

42. Yarov-Yarovoy V, Schonbrun J, Baker D (2006) Multipass membrane protein structure prediction using Rosetta. Proteins 62:1010–1025.

43. Barth P, Schonbrun J, Baker D (2007) Toward high-resolution prediction and design of transmembrane helical protein structures. Proc Natl Acad Sci USA 104:15682–15687.

44. Kuhlman B, Baker D (2000) Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci USA 97:10383–10388.

45. Leaver-Fay A, O'Meara MJ, Tyka M, Jacak R, Song Y, Kellogg EH, Thompson J, Davis IW, Pache RA, Lyskov S, Gray JJ, Kortemme T, Richardson JS, Havranek JJ, Snoeyink J, Baker D, Kuhlman B (2013) Scientific benchmarks for guiding macromolecular energy function improvement. Methods Enzymol 523:[PAGE #S].

46. O'Meara MJ, Leaver-Fay A, Tyka MD, Stein A, Houlihan K, Dimaio F, Bradley P, Kortemme T, Baker D, Snoeyink J, Kuhlman B (2015) Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. J Chem Theory Comput 11:609–622.

47. Dimaio F, Leaver-Fay A, Bradley P, Baker D, Andre I (2011) Modeling symmetric macromolecular structures. PLoS One 6:e20450.

48. Chen K-YM, Zhou F, Fryszczyn BG, Barth P (2012) Naturally evolved G protein-coupled receptors adopt metastable conformations. Proc Natl Acad Sci USA 109:13284–13289.

49. Carugo O, Pongor S (2008) A normalized root-mean-spuare distance for comparing protein three-dimensional structures. Protein Sci 10:1470–1473.

50. Deluca S, Dorr B, Meiler J (2011) Design of native-like proteins through an exposure-dependent environment potential. Biochemistry 50:8521–8528.

51. Allison B, Combs S, Deluca S, Lemmon G, Mizoue L, Meiler J (2014) Computational design of protein-small molecule interfaces. J Struct Biol 185:193–202.

52. Pettit FK, Bare E, Tsai A, Bowie JU (2007) HotPatch: a statistical approach to finding biologically relevant features on protein surfaces. J Mol Biol 369:863–879.

53. Lomize AL, Pogozheva ID, Lomize MA, Mosberg HI (2006) Positioning of proteins in membranes: a computational approach. Protein Sci 15:1318–1333.

54. Lomize MA, Pogozheva ID, Joo H, Mosberg HI, Lomize AL (2012) OPM database and PPM web server: resources for positioning of proteins in membranes. Nucleic Acids Res 40:D370–D376.

55. Kozma D, Simon I, Tusnády GE (2013) PDBTM: Protein Data Bank of transmembrane proteins after 8 years. Nucleic Acids Res 41:D524–D529.

56. ddg_monomer application. Available at: https://www.rosettacommons.org/docs/latest/application_documentation/analysis/ddg-monomer.

57. Kellogg EH, Leaver-Fay A, Baker D (2011) Role of conformational sampling in computing mutation-induced changes in protein structure and stability. Proteins Struct Funct Bioinforma 79:830–838.

58. Conway P, Tyka MD, DiMaio F, Konerding DE, Baker D (2014) Relaxation of backbone bond geometry improves protein energy landscape modeling. Protein Sci 23:47–55.