Evaluating zero-shot prediction of protein design success by AlphaFold, ESMFold, and ProteinMPNN

Mario Garcia[1]; Sugyan M. Dixit[1]; Gabriel J. Rocklin[1,2*]

1. Department of Pharmacology & Center for Synthetic Biology, Northwestern University Feinberg School of Medicine, Chicago, IL, USA
2. Robert H. Lurie Comprehensive Cancer Center, Northwestern University Feinberg School of Medicine, Chicago, IL, USA*

*To whom correspondence should be addressed:

Gabriel J. Rocklin

303 E Superior St

Simpson Querrey Building 11-517

Chicago, IL 60611

grocklin@gmail.com

Abstract

De novo protein design has enabled the creation of proteins with diverse functionalities that are not found in nature. Despite recent advances, experimental success rates remain inconsistent and context-dependent, posing a bottleneck for broader applications of *de novo* design. To overcome this, structure and sequence prediction models have been applied to assess design quality prior to experimental testing to save time and resources. In this study, we aimed to determine the extent to which AlphaFold2, Protein MPNN, and ESMFold can discriminate between experimentally successful and unsuccessful designs. For this, we curated a benchmark dataset of 614 experimentally characterized *de novo* designed monomers from 11 different design studies between 2012 and 2021. All predictive models demonstrated moderate ability to discriminate experimental successes (expressed, soluble, monomeric, and fold into the intended design structure) from failures, with many failed designs having better confidence metrics than successful designs. Among all computational models evaluated, ESMFold average pLDDT yielded the best individual performance at distinguishing between successful and unsuccessful designs. A logistic regression model combining all confidence metrics provided only modest improvement over ESMFold pLDDT alone. Overall, these results show that these models can serve as an initial filtering strategy prior to experimental validation; however, their utility at accurately predicting experimental successful designs remains limited without task-specific training.

Introduction

De novo protein design permits the exploration of protein sequence space beyond the constraints imposed by natural evolution (1). This approach allows for the design of novel protein folds with precise control over the designed structure to achieve specific functions, representing promise for developing novel proteins with applications in medicine and biotechnology (2). Despite recent breakthroughs in AI-guided structure prediction and design strategies, experimental success rates of designs remain inconsistent with rates varying based on the complexity of the designed topology and design objective (3). For example, α-helical protein designs tend to have higher experimental success rates compared to designs rich in β-sheet content or designs attempting to incorporate functionality (3, 4). These inconsistencies emphasize the need for improved computational models to better predict which designs will likely succeed experimentally.

Protein design can fail at many experimental stages, including through poor expression, low solubility, aggregation, folding to an undesired or non-functional conformation, or low overall folding stability. Implementing computational strategies to filter out bad designs prior to characterization can reduce the experimental resources and cost associated with validating designs (5, 6).

Over the last several years, deep learning models have increasingly been applied for this purpose. Structure-based prediction models such as AlphaFold, inverse folding prediction models such as Protein MPNN, and protein language models such as ESMFold have been integrated in design strategies as final validation steps to remove poor designs from experimental characterization (7-12). Although these computational models have shown great capabilities at sequence and structure prediction, they were not trained to predict protein biophysical properties such as expression, aggregation, or stability. As a result, many *de novo* designed proteins achieve highly confident model metrics even when these designs fail to express or fold under typical experimental conditions (13-15). In one large-scale test on *de novo* designed αββα miniproteins, AlphaFold pLDDT metric showed minimal discrimination between experimentally stable and unstable designs (16). These examples highlight the importance of assessing the performance of sequence and structure prediction models when tasked to discriminate between experimental successes and failures in *de novo* design. Furthermore, design and prediction are separate

challenges. Even if a protein design method arose with near-perfect success at a certain design task, it would remain important to accurately predict whether diverse sequences generated by other approaches could also achieve the design objective.

In this study, we aim to quantify the extent to which existing deep learning models developed for structure and sequence prediction can distinguish between experimentally successful and unsuccessful *de novo* protein designs. To accomplish this, we compiled a dataset of 614 experimentally characterized *de novo* designed proteins covering many different designed topologies from the past decade. All proteins were designed using Rosetta-based methods to fold into soluble monomers without deep learning, indicating that these designs all had strongly favorable energies according to a classical protein design energy function (17). Since these models have never been trained on experimental outcomes such as protein expression, solubility, or folding stability, asking them to predict design success is a "zero-shot" evaluation of these models. This analysis provides a baseline for their performance in predicting design success. We also use our benchmark dataset containing many unique folds to examine the influence of protein topology on model confidence and accuracy.

Methods

Dataset Construction

The design studies selected for this analysis focused on the *de novo* design of monomeric proteins (18-28). These studies did not include AlphaFold2, Protein MPNN, or ESMFold during the design process. All monomeric designs in the dataset were experimentally characterized using expression in E. coli, protein purification, size exclusion chromatography combined with multi-angle light scattering (SEC-MALS), and secondary structure assessment by circular dichroism (CD). We defined a design to be experimentally successful if it was expressed, soluble, monomeric, and had a CD spectrum consistent with the intended design fold as reported in each study (Fig. 1A). Protein designs that did not report experimental success or designs needing to be stabilized as dimers were excluded from the final analysis.

AlphaFold Modeling

AlphaFold2 (AlphaFold 2.1.1 version) was used to generate predicted structures for all monomeric *de novo* designs in the dataset (7). For each sequence, five predicted protein structures were generated using the AlphaFold monomer model. The confidence for each structure was determined by the pLDDT metric. The structure with the highest average pLDDT value was selected as the best-ranked AlphaFold predicted structure. AlphaFold structures were generated using both the default protocol of 3 recycles and an extended protocol with 25 recycle steps (7). In addition, AlphaFold PTM model was used to compute the average PAE of each protein design. The average AlphaFold pLDDT and PAE of each design were used as confidence metrics to assess model performance when predicting experimental success. A higher AlphaFold pLDDT and a lower PAE indicate better model predictive performance.

ProteinMPNN Modeling

We used the highest-ranked AlphaFold predicted structures with the default recycle parameter as the input backbone for Protein MPNN (9). The computed Protein MPNN score for the original sequence was used to assess the model's performance in predicting experimental success, with a lower Protein MPNN score indicating better performance.

ESMFold Modeling

To assess the ability of protein language models at predicting experimental success, ESMFold was applied to all *de novo* designed proteins in the dataset (8). ESMFold (esmfold_v1) was used to predict the structures for all designs. ESMFold performance was assessed using the average

pLDDT of the predicted structures. A higher average pLDDT indicates better model performance for ESMFold.

Logistic Regression Modeling

Regression models and ROC curves were computed using scikit-learn (29).

Results

Roughly half of the designs across 11 studies met all experimental success criteria

We compiled 11 *de novo* protein design studies that collectively included 614 individual monomeric *de novo* designs. Each study typically focused on several different design "topologies" specifying the lengths and connectivity of secondary structural elements in each design. Overall, 269/614 (43%) were experimentally successful (soluble, monomeric, and with the correct secondary structure). Success rates varied from study to study and topology to topology (Fig. 1B). For example, whereas most (88%) of the curved β-sheet proteins in Marcos et al. (22) met all experimental success criteria, a much smaller fraction (21%) of the β-barrel designs from Dou et al. (24) were experimentally successful. Many of the lowest success rates came from efforts to design β-sheet rich proteins, such as β-barrel and jellyroll designs (28). Most failed designs were attributed to insolubility and aggregation (65% of failures) (Fig. S1 & S2), although many soluble, monomeric designs failed to show the intended secondary structure via circular dichroism (59/345). Figure 2 displays examples of soluble designs predicted to be well-folded with high confidence by AlphaFold but failed to adopt the designed secondary structure. These designs indicate that soluble, monomeric expression is not sufficient to establish proper folding.

Structure prediction confidence metrics moderately discriminate between good and bad designs

Overall, experimentally successful designs had better confidence metric distributions compared to unsuccessful designs across all computational models (Fig. 3A). Still, discrimination was far from perfect, with ROC AUC scores ranging from 0.66-0.73 for the different confidence metrics (Fig. 3B). ESMFold's average pLDDT had the highest AUC out of the different confidence metrics (0.73±0.04, mean±std from bootstrapping) whereas MPNN score had the lowest AUC (0.66±0.05). AlphaFold pLDDT values generated with different numbers of recycles were highly correlated (Spearman = 0.94). A total of 31/614 designs showed an increase in pLDDT > 5 using 25 recycles, with three designs increasing their pLDDT > 15. These 31 designs were primarily FoldIt designs (23/31) from Koepnick et al. (Fig. S3). The different confidence metrics captured different information, with Spearman correlations between metrics ranging from 0.52-0.83 (not including the two AlphaFold protocols, Fig. 3C). Still, despite differences between these methods, we found that combining these metrics in a logistic regression model (ROC AUC of 0.74 ± 0.04) showed minimal improvement over the best individual method (ESMFold: ROC AUC of 0.73 ± 0.04) (Fig. 4).

Confidence metrics were not generalizable across design topologies, preventing a reliable threshold from being established for experimental success

We then analyzed confidence metrics across different design topologies, focusing initially on AlphaFold average pLDDT values (Fig. 5, Fig. S4-S6). Importantly, we observed that the distribution of AlphaFold pLDDT values varied across different design topologies (Fig. 5A). This topology-dependence indicates that there is no clear "threshold" in AlphaFold pLDDT (or ESMFold pLDDT, Fig. S5) for experimental success. Failed designs from "easily predictable" topologies on the left of Fig. 5 often had higher pLDDT scores than successful designs from "difficult to predict" topologies on the right on Fig. 5. Although AlphaFold (and ESMFold) pLDDT

scores remain useful ranking metrics within an individual topology, this illustrates the difficulty of comparing designs between topologies and using pLDDT as a proxy for experimental success.

Discussion

In this study, we found that sequence and structure prediction models offer a moderate ability to distinguish between experimentally successful and unsuccessful *de novo* designs, with ESMFold providing the highest discrimination. Integrating confidence metrics from sequence and structure prediction models into a logistic regression model resulted in marginal improvements compared to ESMFold pLDDT alone. Notably, many unsuccessful designs received better confidence metrics than successful designs across all models, and confidence metrics were correlated with the design topology.

A caveat to this analysis is that many of the successful designs with crystal structures were likely implemented in AlphaFold's training set. This would potentially bias the models to more confidently predict these structures and artificially inflate discriminatory power. However, AlphaFold's confidence metrics struggled to consistently distinguish experimentally successful designs from failures. Examining the predicted structures of these models can potentially overcome these inconsistencies. For this analysis, we did not have access to the intended designed backbones; however, RMSD calculations between predicted backbone structures and the intended backbone designs can be implemented in future work to identify large deviations from the intended backbone design to further filter poor designs (6). Moreover, computational models are continuously being developed to predict experimental properties such as solubility, conformational dynamics, and stability (30-33). Incorporating these models into design strategies can further improve the experimental success rates. As large-scale experimental datasets become available, we anticipate that training and fine-tuning these models on this data will help improve the inconsistencies in experimental success rates of *de novo* designs. This work provides a baseline analysis of experimental success rates for *de novo* designs.

References

1. Huang, PS., Boyken, S. & Baker, D. The coming of age of de novo protein design. Nature 537, 320–327 (2016). https://doi.org/10.1038/nature19946
2. Korendovych, I. V., & DeGrado, W. F. (2020). De novo protein design, a retrospective. Quarterly reviews of biophysics, 53, e3. https://doi.org/10.1017/S0033583519000131
3. Pan, X., & Kortemme, T. (2021). Recent advances in de novo protein design: Principles, methods, and applications. The Journal of biological chemistry, 296, 100558. https://doi.org/10.1016/j.jbc.2021.100558
4. Kortemme T. (2024). De novo protein design-From new structures to programmable functions. Cell, 187(3), 526–544. https://doi.org/10.1016/j.cell.2023.12.028
5. Peñas-Utrilla, D., & Marcos, E. (2022). Identifying well-folded de novo proteins in the new era of accurate structure prediction. Frontiers in molecular biosciences, 9, 991380. https://doi.org/10.3389/fmolb.2022.991380
6. Hermosilla, A. M., Berner, C., Ovchinnikov, S., & Vorobieva, A. A. (2024). Validation of de novo designed water-soluble and transmembrane β-barrels by in silico folding and melting. Protein science : a publication of the Protein Society, 33(7), e5033. https://doi.org/10.1002/pro.5033
7. 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., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2
8. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., Dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science (New York, N.Y.), 379(6637), 1123–1130. https://doi.org/10.1126/science.ade2574
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., … Baker, D. (2022). Robust deep learning-based protein sequence design using ProteinMPNN. Science (New York, N.Y.), 378(6615), 49–56. https://doi.org/10.1126/science.add2187
10. 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., … Baker, D. (2023). De novo design of protein structure and function with RFdiffusion. Nature, 620(7976), 1089–1100. https://doi.org/10.1038/s41586-023-06415-8
11. Wang, Jue et al. "Scaffolding protein functional sites using deep learning." *Science (New York, N.Y.)* vol. 377,6604 (2022): 387-394. doi:10.1126/science.abn2100
12. Sumida, Kiera H et al. "Improving Protein Expression, Stability, and Function with ProteinMPNN." *Journal of the American Chemical Society* vol. 146,3 (2024): 2054-2061. doi:10.1021/jacs.3c10941
13. Chu, Alexander E et al. "Sparks of function by de novo protein design." Nature biotechnology vol. 42,2 (2024): 203-215. doi:10.1038/s41587-024-02133-2
14. Wicky, B I M et al. "Hallucinating symmetric protein assemblies." *Science (New York, N.Y.)* vol. 378,6615 (2022): 56-61. doi:10.1126/science.add1964
15. Frank, Christopher et al. "Scalable protein design using optimization in a relaxed sequence space." *Science (New York, N.Y.)* vol. 386,6720 (2024): 439-445. doi:10.1126/science.adq1741
16. Kim, T. E., Tsuboyama, K., Houliston, S., Martell, C. M., Phoumyvong, C. M., Lemak, A., Haddox, H. K., Arrowsmith, C. H., & Rocklin, G. J. (2022). Dissecting the stability determinants of a challenging de novo protein fold using massively parallel design and experimentation. Proceedings of the National Academy of Sciences of the United States of America, 119(41), e2122676119. https://doi.org/10.1073/pnas.2122676119
17. Alford, R. F., Leaver-Fay, A., Jeliazkov, J. R., O'Meara, M. J., DiMaio, F. P., Park, H., Shapovalov, M. V., Renfrew, P. D., Mulligan, V. K., Kappel, K., Labonte, J. W., Pacella, M. S., Bonneau, R., Bradley, P., Dunbrack, R. L., Jr, Das, R., Baker, D., Kuhlman, B., Kortemme, T., & Gray, J. J. (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. Journal of chemical theory and computation, 13(6), 3031–3048. https://doi.org/10.1021/acs.jctc.7b00125
18. Koga, N., Tatsumi-Koga, R., Liu, G., Xiao, R., Acton, T. B., Montelione, G. T., & Baker, D. (2012). Principles for designing ideal protein structures. Nature, 491(7423), 222–227. https://doi.org/10.1038/nature11600
19. Lin, Y. R., Koga, N., Tatsumi-Koga, R., Liu, G., Clouser, A. F., Montelione, G. T., & Baker, D. (2015). Control over overall shape and size in de novo designed proteins. Proceedings of the National Academy of Sciences of the United States of America, 112(40), E5478–E5485. https://doi.org/10.1073/pnas.1509508112
20. Huang, P. S., Feldmeier, K., Parmeggiani, F., Velasco, D. A. F., Höcker, B., & Baker, D. (2016). De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. Nature chemical biology, 12(1), 29–34. https://doi.org/10.1038/nchembio.1966
21. Jacobs, T. M., Williams, B., Williams, T., Xu, X., Eletsky, A., Federizon, J. F., Szyperski, T., & Kuhlman, B. (2016). Design of structurally distinct proteins using strategies inspired by evolution. Science (New York, N.Y.), 352(6286), 687–690. https://doi.org/10.1126/science.aad8036

22. Marcos, E., Basanta, B., Chidyausiku, T. M., Tang, Y., Oberdorfer, G., Liu, G., Swapna, G. V., Guan, R., Silva, D. A., Dou, J., Pereira, J. H., Xiao, R., Sankaran, B., Zwart, P. H., Montelione, G. T., & Baker, D. (2017). Principles for designing proteins with cavities formed by curved β sheets. Science (New York, N.Y.), 355(6321), 201–206. https://doi.org/10.1126/science.aah7389
23. Marcos, E., Chidyausiku, T. M., McShan, A. C., Evangelidis, T., Nerli, S., Carter, L., Nivón, L. G., Davis, A., Oberdorfer, G., Tripsianes, K., Sgourakis, N. G., & Baker, D. (2018). De novo design of a non-local β-sheet protein with high stability and accuracy. Nature structural & molecular biology, 25(11), 1028–1034. https://doi.org/10.1038/s41594-018-0141-6
24. 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. (2018). De novo design of a fluorescence-activating β-barrel. Nature, 561(7724), 485–491. https://doi.org/10.1038/s41586-018-0509-0
25. Koga, N., Koga, R., Liu, G., Castellanos, J., Montelione, G. T., & Baker, D. (2021). Role of backbone strain in de novo design of complex α/β protein structures. Nature communications, 12(1), 3921. https://doi.org/10.1038/s41467-021-24050-7
26. Minami, S., Kobayashi, N., Sugiki, T., Nagashima, T., Fujiwara, T., Tatsumi-Koga, R., Chikenji, G., & Koga, N. (2023). Exploration of novel αβ-protein folds through de novo design. Nature structural & molecular biology, 30(8), 1132–1140. https://doi.org/10.1038/s41594-023-01029-0
27. Sakuma, K., Kobayashi, N., Sugiki, T., Nagashima, T., Fujiwara, T., Suzuki, K., Kobayashi, N., Murata, T., Kosugi, T., Tatsumi-Koga, R., & Koga, N. (2024). Design of complicated all-α protein structures. Nature structural & molecular biology, 31(2), 275–282. https://doi.org/10.1038/s41594-023-01147-9
28. Koepnick, B., Flatten, J., Husain, T. et al. De novo protein design by citizen scientists. Nature 570, 390–394 (2019). https://doi.org/10.1038/s41586-019-1274-4
29. Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011.
30. Dieckhaus, H., Brocidiacono, M., Randolph, N. Z., & Kuhlman, B. (2024). Transfer learning to leverage larger datasets for improved prediction of protein stability changes. Proceedings of the National Academy of Sciences of the United States of America, 121(6), e2314853121. https://doi.org/10.1073/pnas.2314853121
31. Monteiro da Silva, G., Cui, J.Y., Dalgarno, D.C. et al. High-throughput prediction of protein conformational distributions with subsampled AlphaFold2. Nat Commun 15, 2464 (2024). https://doi.org/10.1038/s41467-024-46715-9
32. Li, B., & Ming, D. (2024). GATSol, an enhanced predictor of protein solubility through the synergy of 3D structure graph and large language modeling. BMC bioinformatics, 25(1), 204. https://doi.org/10.1186/s12859-024-05820-8
33. Ferrari, Á. J. R., Dixit, S. M., Thibeault, J., Garcia, M., Houliston, S., Ludwig, R. W., Notin, P., Phoumyvong, C. M., Martell, C. M., Jung, M. D., Tsuboyama, K., Carter, L., Arrowsmith, C. H., Guttman, M., & Rocklin, G. J. (2025). Large-scale discovery, analysis, and design of protein energy landscapes. bioRxiv : the preprint server for biology, 2025.03.20.644235. https://doi.org/10.1101/2025.03.20.644235

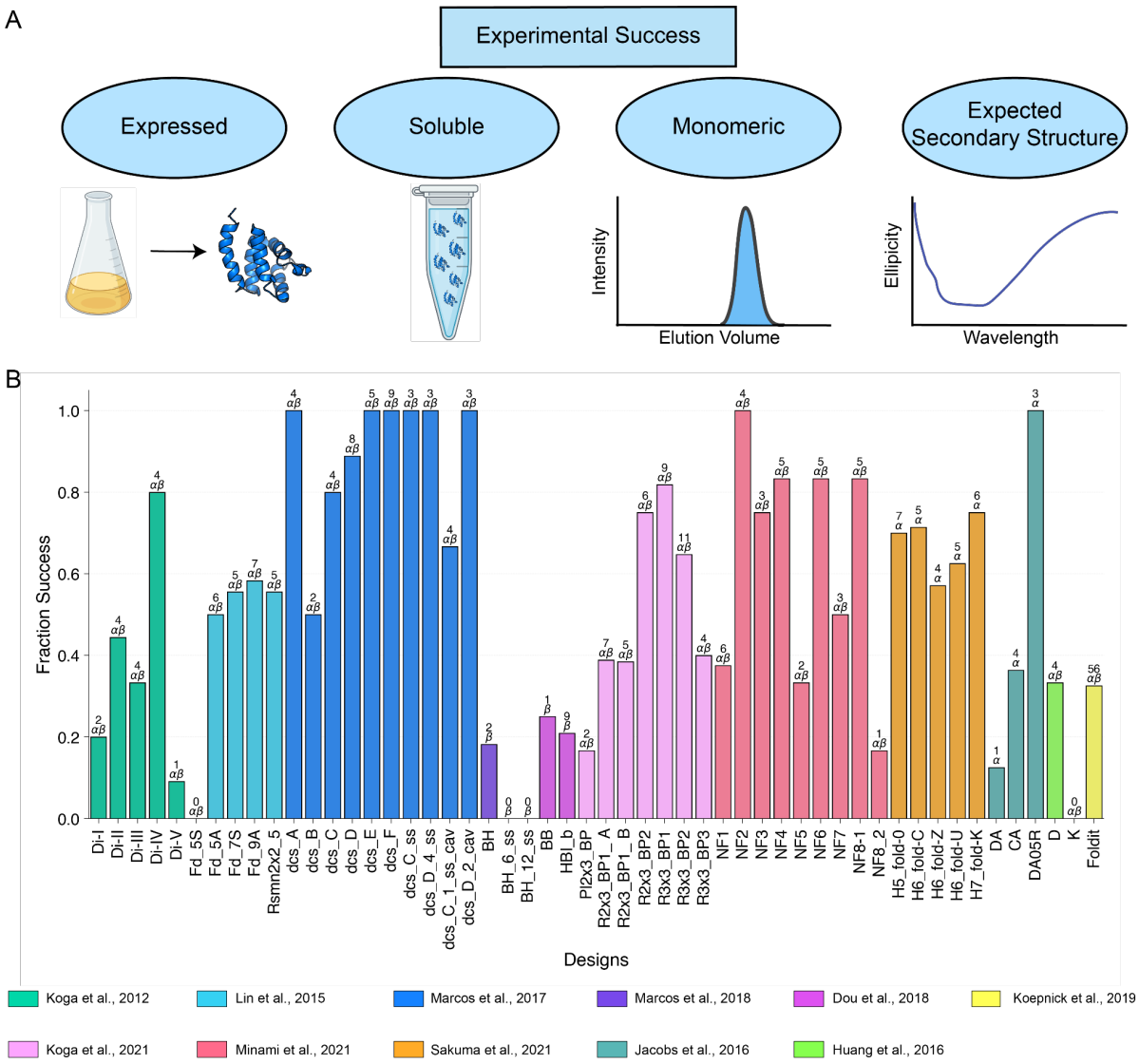

Figure 1: Experimental success rates vary depending on design objective and topology. A) Schematic depiction of criteria used to define experimentally successful de novo designs in this analysis. A design was considered successful if it was expressed, soluble, monomeric, and folded into the intended secondary structure **B)** Fraction of experimentally successful designs for each design topology. Design topologies consisted of all α-helical proteins (α), all β-sheet proteins (β), or mixed α/β proteins (α/β). The number of successful designs within each topology is reported above each bar. Bars for each design topology are colored according to the reported article.

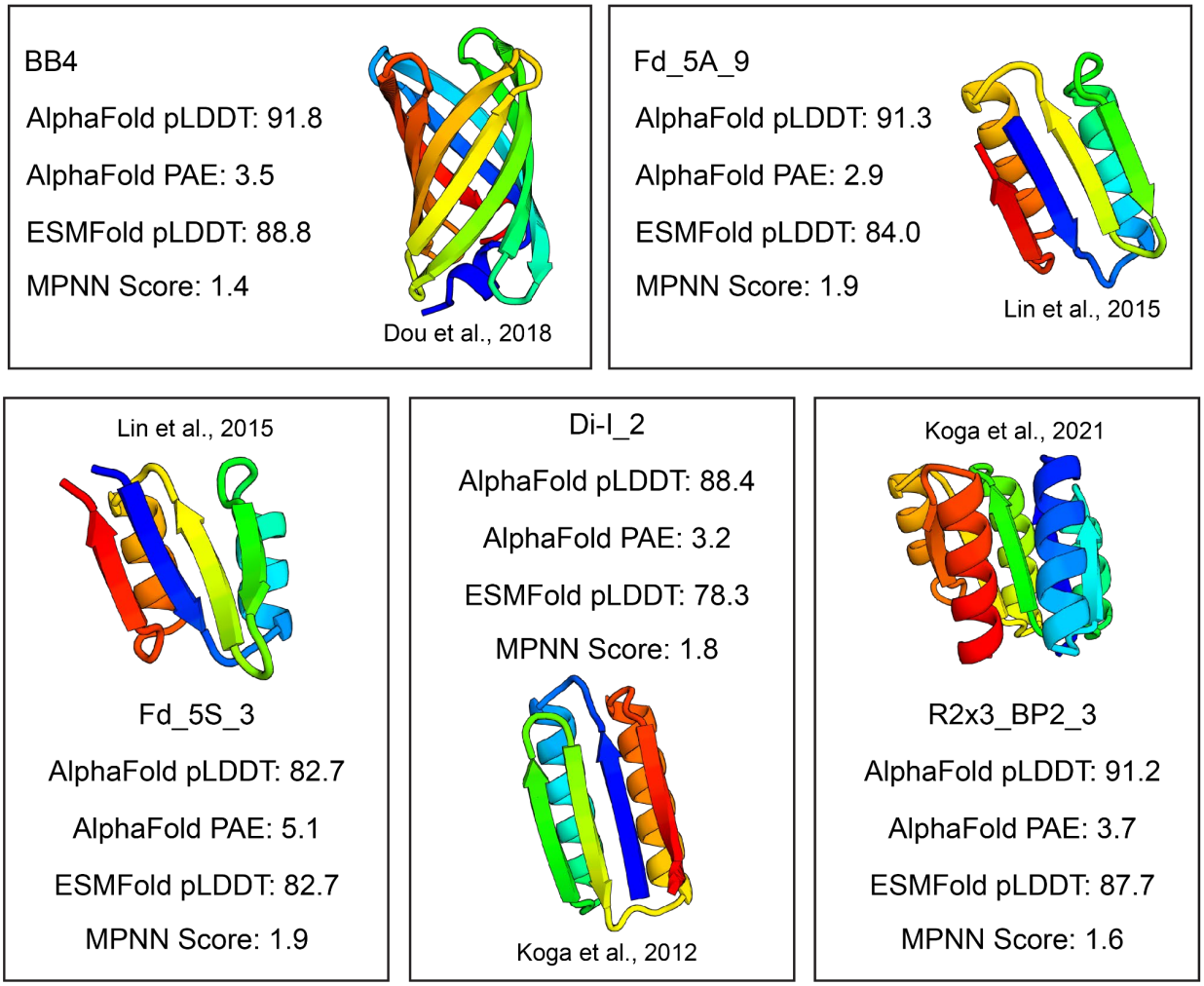

Figure 2: Soluble designs that failed to fold into intended secondary structures despite high AlphaFold confidence. High confidence (pLDDT > 82) AlphaFold predicted structure of designs that were experimentally soluble but failed to fold into intended secondary structures, assessed by circular dichroism.

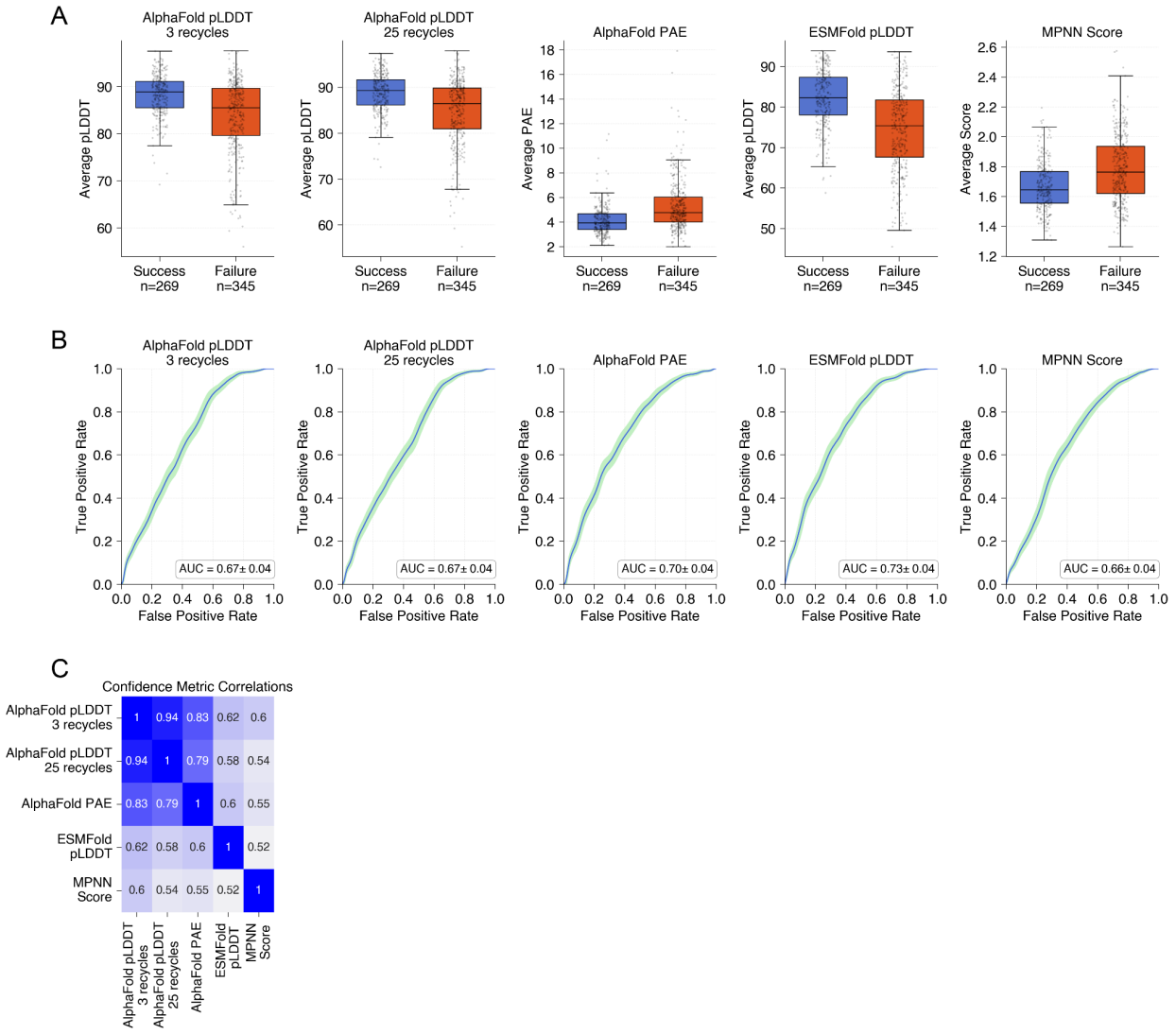

Figure 3: ESMFold pLDDT provides the largest discrimination at distinguishing between experimental successful and unsuccessful designs. A) Boxplots displaying distributions of confidence metrics between experimentally successful (blue) and unsuccessful (orange) designs with successful designs displaying better confidence metrics on average. **B)** ROC curves for each confidence metric displaying the average AUC values. The shaded region represents the 95% confidence interval computed using bootstrapping. **C)** Spearman correlation matrix showing absolute correlation between different confidence metrics.

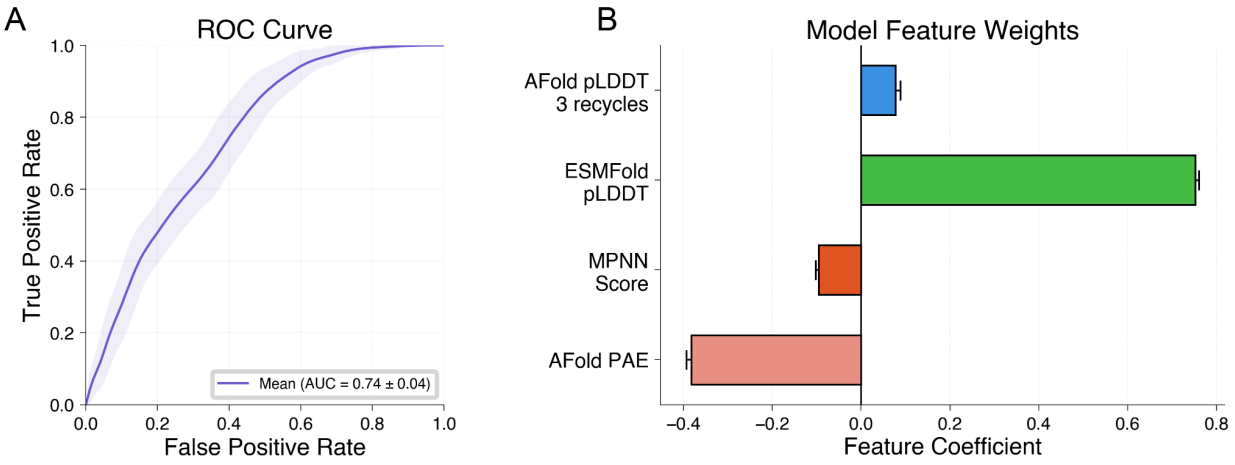

Figure 4: Combining confidence metrics into logistic regression model leads to minor improvement in average AUC. A) ROC Curve for logistic regression model trained on all confidence metrics to predict experimental success. The ROC Curve displays the mean (dark blue) and 95% confidence interval (shaded) generated through bootstrapping. **B)** Feature weights from the logistic regression model indicating the relative contribution of each metric to the model. ESMFold pLDDT showed the highest feature weight.

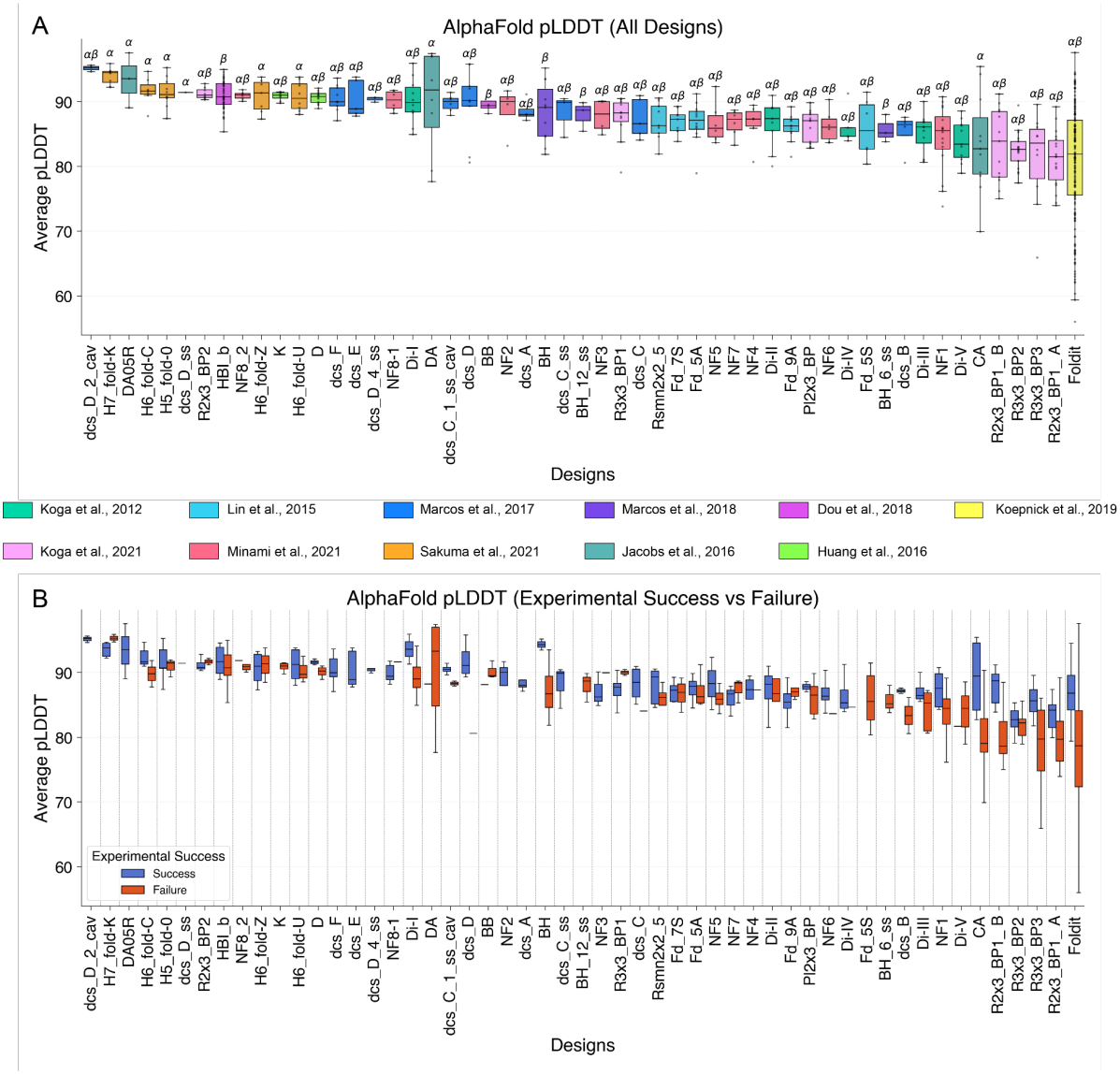

Figure 5: AlphaFold average pLDDT distributions vary across different design topologies. A) Distribution of AlphaFold average pLDDTs for designs included in this study. **B)** Distribution of AlphaFold average pLDDTs comparing experimentally successful designs (blue) and unsuccessful designs (orange) for each design topology.

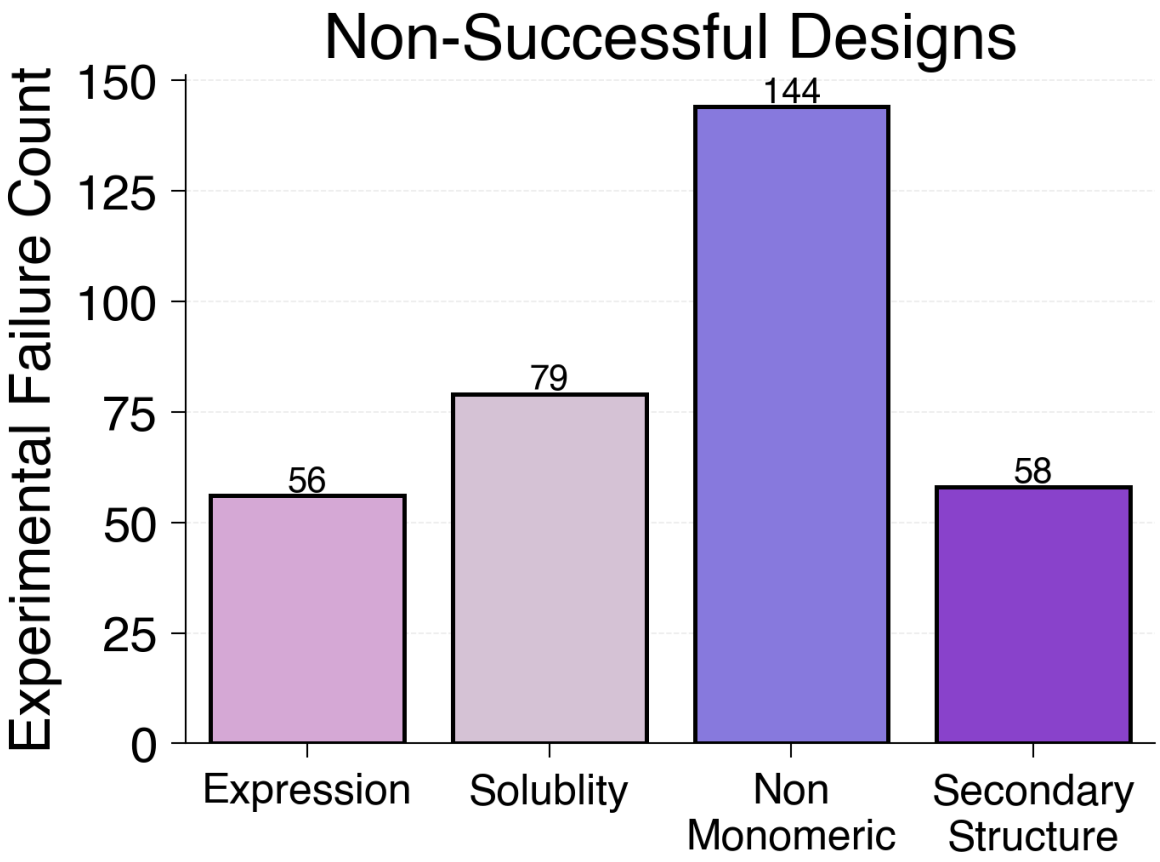

Figure S1. The majority of de novo designs failed experimentally due to not being monomeric. Counts of experimentally unsuccessful designs for each experimental criteria used to define design experimental success.

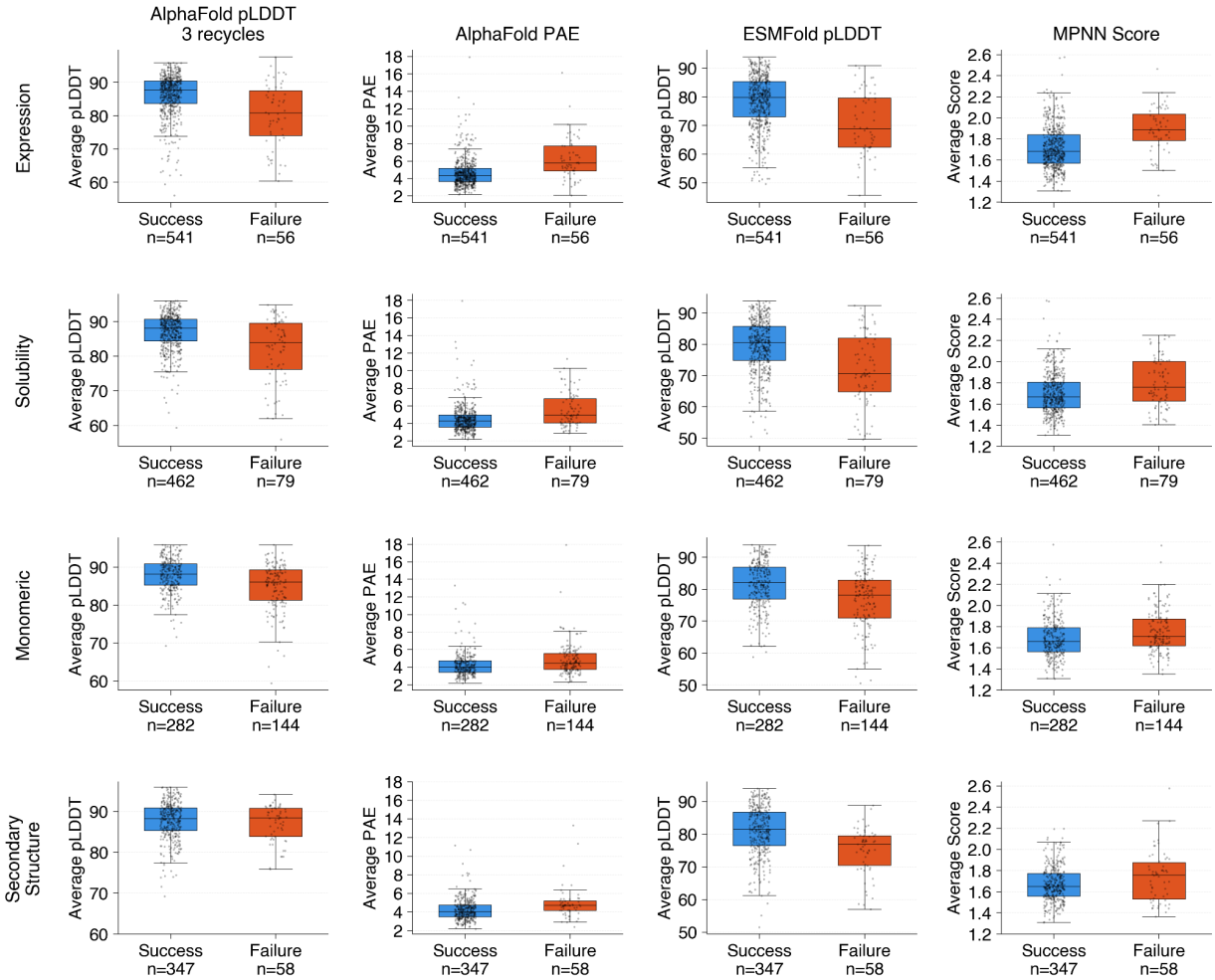

Figure S2. Distribution of confidence metrics for experimental outcomes of designs. Distribution of confidence metrics comparing successful (blue) and unsuccessful designs (orange) across different experimental outcomes: expression, solubility, monomeric state, and secondary structure.

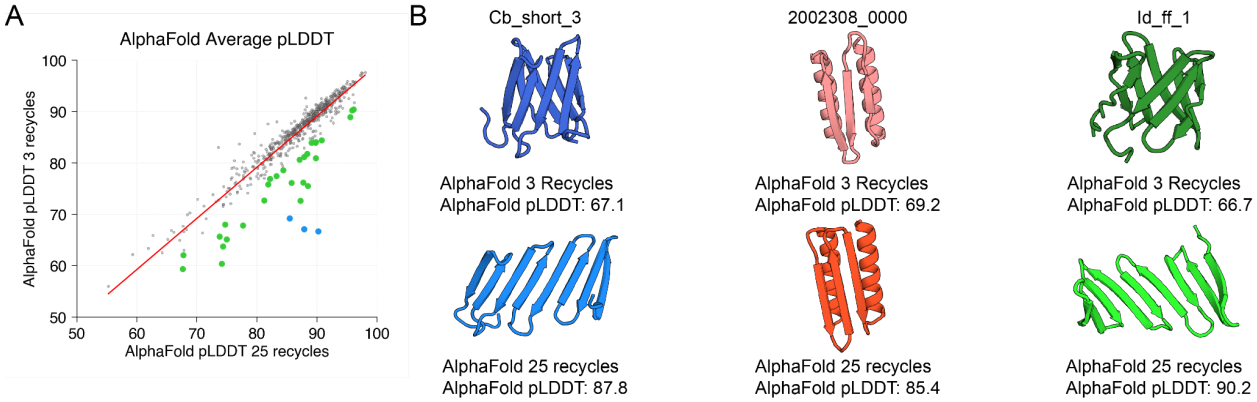

Figure S3. Average AlphaFold pLDDT values are highly correlated with different recycle parameters. A) Scatterplot capturing the relationship between AlphaFold average pLDDT for de novo designs computed using different recycle parameters. Green highlighted points indicate designs that increased their average pLDDT by >5 after increasing the recycles, and blue highlighted points are designs that increased their average pLDDT by >15 after increasing recycles. **B)** AlphaFold predicted structures of designs that increased their average pLDDT >15 with increased recycles.

.

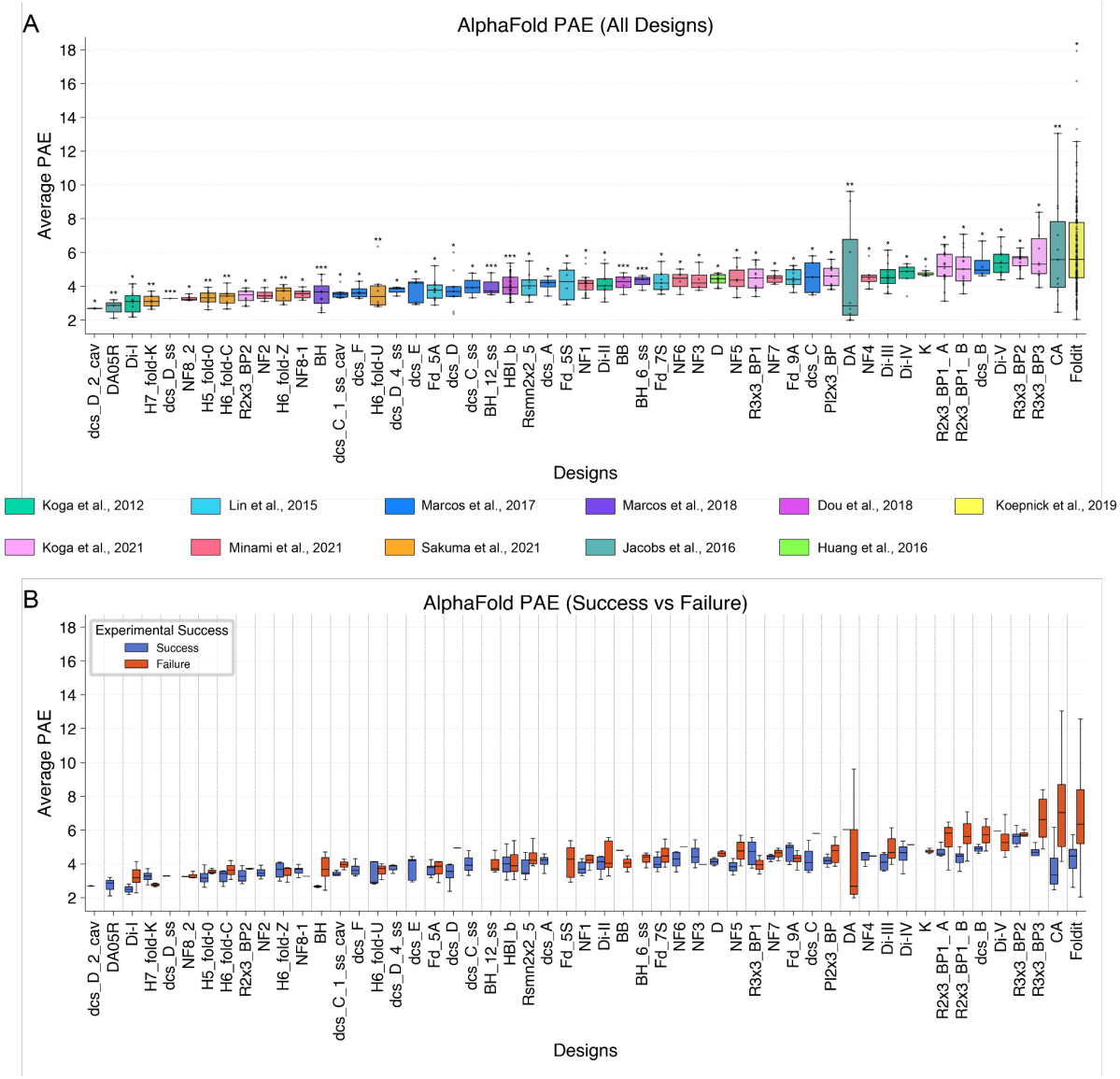

Figure S4. AlphaFold average PAE distributions for designed topologies. A) Distributions of AlphaFold average PAE for designs included in this study. **B)** Distribution of AlphaFold average PAE for experimentally successful designs (blue) and unsuccessful designs (orange) for each design topology.

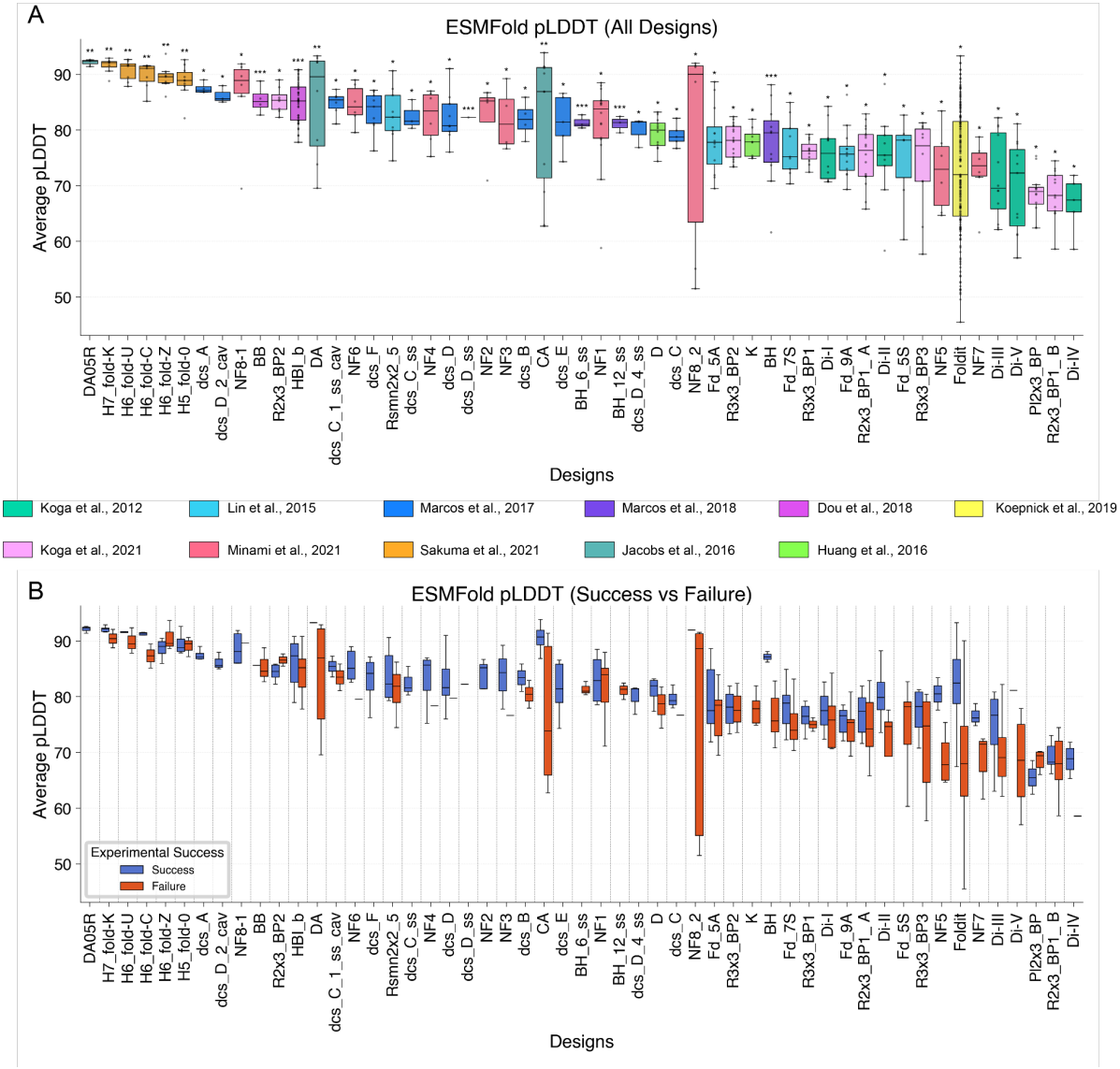

Figure S5. ESMFold average pLDDT distributions for designed topologies. A) Distribution of ESMFold average pLDDT for designs included in this study. **B)** Distribution of ESMFold average pLDDT for experimental successful designs (blue) and unsuccessful designs (orange) for each design topology

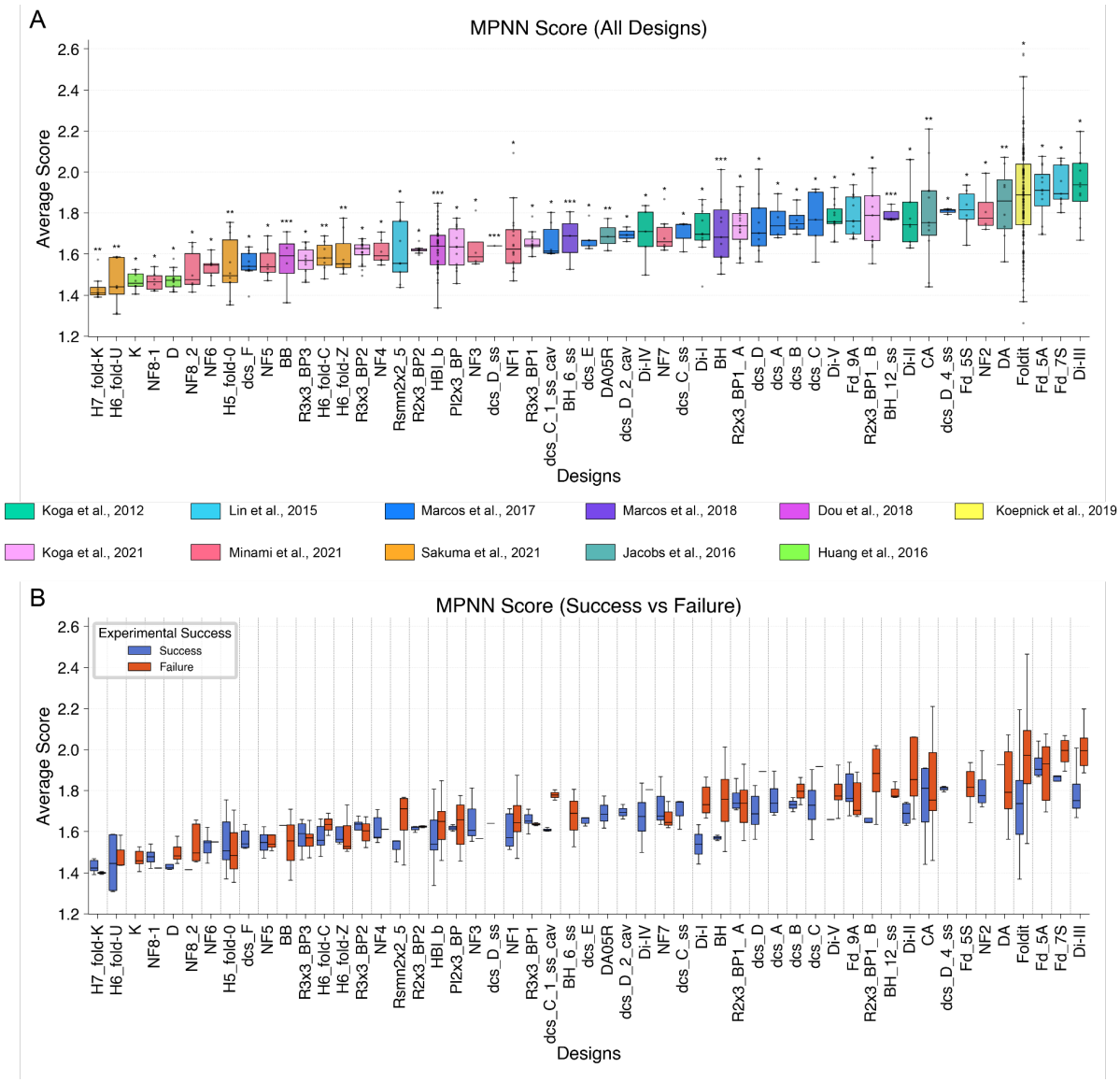

Figure S6. ProteinMPNN score distributions for designed topologies. A) Distribution of MPNN scores for designs included in this study. **B)** Distribution of MPNN score for experimental successful (blue) and unsuccessful designs (orange) for each design topology.