**Research Proposal**

# **Developing Computational Frameworks to Enhance Allostery Design through Directed Evolution**

### Burcin Acar (DeBenedictis Lab)

Allostery is a mechanism of communication between distant sites in a protein, allowing for changes in one site to impact the function of another. This remote control over function has great potential for biotechnology and medicine, and can be harnessed through a better understanding and design of allosteric mechanisms in proteins. Various techniques, including rational design, directed evolution, and *in silico* methods, are employed to design allostery in proteins exploiting recombination strategies ranging from point mutations; domain fusions, insertions, and tetherings; to protein splittings and segment swappings [(Köhler 2014; Singh et al. 2018)](https://paperpile.com/c/SwMIKX/YPb9+AJnd). However, despite their utility, none of these methods provide a definitive solution alone and a combination of computational and experimental approaches is often necessary for a successful design.

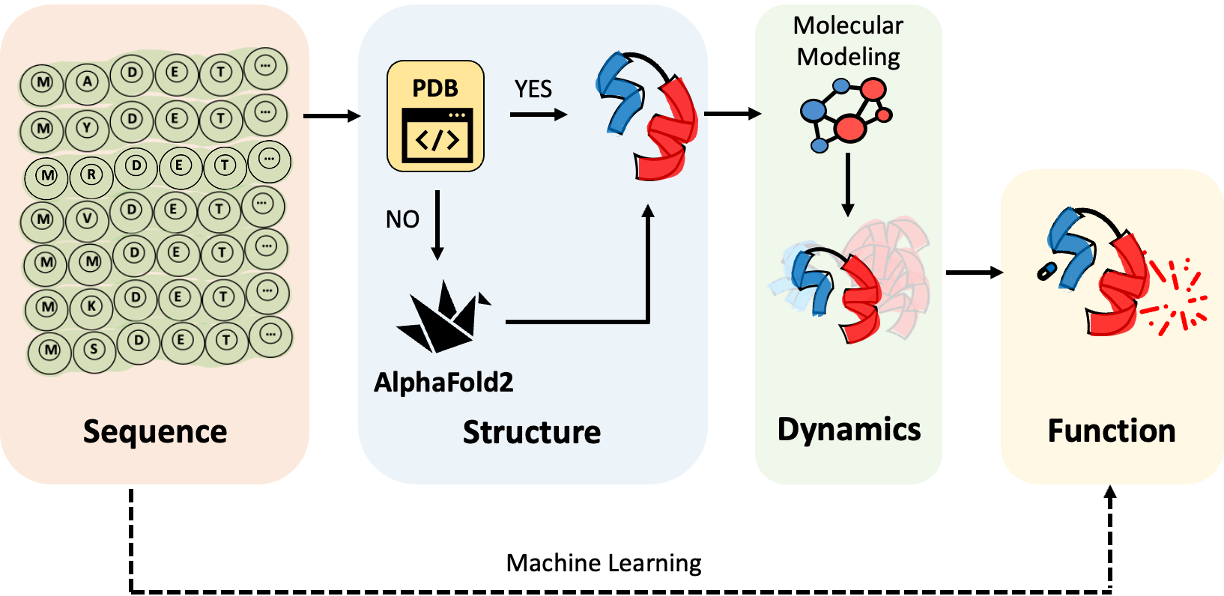
Computational tools enable the prediction of physicochemical characteristics of a protein to guide experiments and thus are useful to generate initial templates for protein design [(Kuhlman and Bradley 2019)](https://paperpile.com/c/SwMIKX/MYru). They are also useful in assessing the impact of mutations, or any other proposed change and engineering on these features. This aptness to scoring different protein sequences in regard to the final aim, expedites the design efforts [(Rosenfeld et al. 2016)](https://paperpile.com/c/SwMIKX/B0gn). Besides, with the ongoing enhancement of computational techniques, their importance in allostery design becomes increasingly prominent.

Computational approaches to protein design can be roughly classified into two categories; those utilizing general molecular modeling and bioinformatics tools, and those employing machine learning techniques. Classical bioinformatic methods are useful due to their lower computational requirements, robustness and also detailed mechanistic output they provide in case of a specific system. These methods can be exemplified with multiple-sequence alignments providing evolutionary information, homology and *ab initio* modeling yielding structure predictions, molecular dynamics and docking simulations along with network models estimating dynamic mechanism and energy-based scoring of protein sequences on characteristics like stability and solubility. For example, the commonly used protein design server RosettaDesign [(Liu and Kuhlman 2006)](https://paperpile.com/c/SwMIKX/5Pap) combines all these aspects to identify optimum amino acid sequences for target protein structures and functions. On the other hand, machine learning based approaches gain popularity as available experimental data for input grows and the applied computational tools are refined. This is evident in the remarkable success of AlphaFold2 [(Jumper et al. 2021)](https://paperpile.com/c/SwMIKX/yt14) in predicting protein structures from sequence.

The general aim in computational allostery design is to achieve a protein sequence with projected features, be it a target structure, ligand-binding characteristics, stability or backbone flexibility. So the computational methods used in allostery design may vary depending on its goal, which can range from optimizing protein catalytic activity, enhancing ligand specificity, to combining naturally independent processes for biosensor design. While the first might benefit from quantum mechanical modeling of the active site, the followings might gain from the ligand-docking simulations. Additionally, the computational methods of preference might diversify based on the characteristics of the available experimental data to be integrated with. For example, in order to design a thermostable version of a protein, different computational tools can be selected. As one of them, FRESCO (Framework for Rapid Enzyme Stabilization by Computational libraries) combines molecular dynamics simulations, energy minimization, and statistical optimization with functional assays under high temperatures to enhance the rational design of thermostable enzymes [(Wijma, Fürst, and Janssen 2018)](https://paperpile.com/c/SwMIKX/IciQ). Alternatively, the KnowVolution (knowledge gaining directed evolution) method uses structural modeling to guide the directed evolution process by assessing the positions of mutations and minimizing the efforts for screening of the clone library [(Cheng, Zhu, and Schwaneberg 2015)](https://paperpile.com/c/SwMIKX/razf). In parallel with this example, this project aims to develop computational tools to enhance directed evolution experiments performed with phage- and robotics-assisted near-continuous evolution (PRANCE; [(DeBenedictis et al. 2022)](https://paperpile.com/c/SwMIKX/ml6j)) toward the ultimate goal of designing allostery in proteins.

PRANCE is a high-throughput application of the well established directed evolution method PACE (Phage-assisted Continuous Evolution; [(Popa et al. 2020)](https://paperpile.com/c/SwMIKX/SDQK)) through robotics, on protein design. Both PRANCE and PACE utilize the M13 bacteriophage as a vector to carry the target protein's DNA and to infect host *E. coli* cells. The M13 phage has been designed to lack the gene encoding its crucial coat protein, pIII, which is necessary for phage maturation and infectivity. However, the relevant gene is provided on an accessory plasmid that has transformed the host cells and its expression is contingent on the target protein to attain the desired function through cycles of random mutagenesis and selection. In this process, the random mutations occur at a higher rate than what occurs naturally and are put under selection pressure established by pIII expression achieved through the desired activity of target protein. Only phages with a fit target protein can get pIII expressed from the host cells, proceed to the next infection cycle, and continue the directed evolution process. The experiments are stopped when the target protein has acquired enough advantageous mutations to attain the desired property. While both methods are effective in optimizing protein properties, PRANCE provides several improvements on PACE, including the ability to evolve many independent populations in parallel, significantly reducing researcher intervention and consumables, making it efficient for the extensive exploration of the evolutionary landscape.

PRANCE has proven to be a potent method for protein design [(DeBenedictis et al. 2022)](https://paperpile.com/c/SwMIKX/ml6j), nonetheless it can be further enhanced through the incorporation of computational tools. Specifically, computational models that can predict potential functionality and allostery of proteins from their sequence would be instrumental to identifying optimal initial sequences for PRANCE to be evolved toward the desired functionality. Although machine learning enables us to predict functionality out of sequence information only [(Hou et al. 2022)](https://paperpile.com/c/SwMIKX/geoW), the resulting model linking sequence to function is nearly a black box [(Grau, Nowé, and Vranken 2021)](https://paperpile.com/c/SwMIKX/fdQ1) and might yield inconsistent predictions [(Petti et al. 2023)](https://paperpile.com/c/SwMIKX/ANg1). An alternative approach is to develop mechanistic models, using structure and dynamic information as a guiding intermediate step from sequence to function (Figure 1). The purpose of this project is to develop such models that can predict functionality and allostery along with evolvability from a given initial sequence to give insight on prospective PRANCE outputs. These models might economize already-prudent PRANCE experiments by sampling functionally relevant parts of the evolutionary landscape instead of a brutal search on the overall landscape.

Figure 1- Two different approaches to protein function prediction from sequence. Direct prediction from sequence (dashed line) and prediction using structural and dynamic information (regular lines).

Following steps are proposed to develop allostery prediction models:

1. A training dataset from exemplary designed proteins with a diverse set of sequence libraries and corresponding functional annotations such as functional enrichment scores will be curated both from the literature [(Nadler et al. 2016)](https://paperpile.com/c/SwMIKX/vQOL) and PRANCE outputs that will be obtained at our lab.
2. The three dimensional structures of the compiled protein sequences will be extracted from Protein Data Bank [(Berman, Henrick, and Nakamura 2003)](https://paperpile.com/c/SwMIKX/0602) if available, otherwise it will be predicted by AlphaFold 2 [(Jumper et al. 2021)](https://paperpile.com/c/SwMIKX/yt14).
3. Dynamics of the resulting structures will be modeled by Elastic Network Models as it provides a fast estimation of the equilibrium dynamics of proteins [(Chennubhotla et al. 2005)](https://paperpile.com/c/SwMIKX/ZRP8) required for the large number of sequences. Finer approaches yet with increased requirements of computational power and time such as molecular dynamics [(Karplus and McCammon 2002)](https://paperpile.com/c/SwMIKX/OGBP) will be exploited on structures of specific interest, by which more detail such as side chain information or solvent effects can also be incorporated into the model.
4. Based on the derived structural and dynamical properties, an allostery score will be developed to predict functionality and allostery of a given sequence, later to be compared with experiments. This score will benefit from assessing a potential dynamic coupling between input and output domains of the designed protein along with penalizing large disruption of the structural and dynamical characteristics of the constituent protein segments.
5. The developed allostery score will be tested on various available designs and used to assess the functionalities of novel protein designs.

Integration of this sequence to allostery prediction model with PRANCE will provide a comprehensive hybrid platform for allostery design in proteins.

References

[Berman, Helen, Kim Henrick, and Haruki Nakamura. 2003. “Announcing the Worldwide Protein Data Bank.” *Nature Structural Biology* 10 (12): 980.](http://paperpile.com/b/SwMIKX/0602)

[Cheng, Feng, Leilei Zhu, and Ulrich Schwaneberg. 2015. “Directed Evolution 2.0: Improving and Deciphering Enzyme Properties.” *Chemical Communications*  51 (48): 9760–72.](http://paperpile.com/b/SwMIKX/razf)

[Chennubhotla, Chakra, A. J. Rader, Lee-Wei Yang, and Ivet Bahar. 2005. “Elastic Network Models for Understanding Biomolecular Machinery: From Enzymes to Supramolecular Assemblies.” *Physical Biology* 2 (4): S173–80.](http://paperpile.com/b/SwMIKX/ZRP8)

[DeBenedictis, Erika A., Emma J. Chory, Dana W. Gretton, Brian Wang, Stefan Golas, and Kevin M. Esvelt. 2022. “Systematic Molecular Evolution Enables Robust Biomolecule Discovery.” *Nature Methods* 19 (1): 55–64.](http://paperpile.com/b/SwMIKX/ml6j)

[Grau, Isel, Ann Nowé, and Wim Vranken. 2021. “Interpreting a Black Box Predictor to Gain Insights into Early Folding Mechanisms.” *Computational and Structural Biotechnology Journal* 19 (August): 4919–30.](http://paperpile.com/b/SwMIKX/fdQ1)

[Hou, Qingzhen, Katharina Waury, Dea Gogishvili, and K. Anton Feenstra. 2022. “Ten Quick Tips for Sequence-Based Prediction of Protein Properties Using Machine Learning.” *PLoS Computational Biology* 18 (12): e1010669.](http://paperpile.com/b/SwMIKX/geoW)

[Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature* 596 (7873): 583–89.](http://paperpile.com/b/SwMIKX/yt14)

[Karplus, Martin, and J. Andrew McCammon. 2002. “Molecular Dynamics Simulations of Biomolecules.” *Nature Structural Biology* 9 (9): 646–52.](http://paperpile.com/b/SwMIKX/OGBP)

[Köhler, Valentin. n.d. *Protein Design*. Springer New York. Accessed March 8, 2023.](http://paperpile.com/b/SwMIKX/YPb9)

[Kuhlman, Brian, and Philip Bradley. 2019. “Advances in Protein Structure Prediction and Design.” *Nature Reviews. Molecular Cell Biology* 20 (11): 681–97.](http://paperpile.com/b/SwMIKX/MYru)

[Liu, Yi, and Brian Kuhlman. 2006. “RosettaDesign Server for Protein Design.” *Nucleic Acids Research* 34 (Web Server issue): W235–38.](http://paperpile.com/b/SwMIKX/5Pap)

[Nadler, Dana C., Stacy-Anne Morgan, Avi Flamholz, Kaitlyn E. Kortright, and David F. Savage. 2016. “Rapid Construction of Metabolite Biosensors Using Domain-Insertion Profiling.” *Nature Communications* 7 (July): 12266.](http://paperpile.com/b/SwMIKX/vQOL)

[Petti, Samantha, Nicholas Bhattacharya, Roshan Rao, Justas Dauparas, Neil Thomas, Juannan Zhou, Alexander M. Rush, Peter Koo, and Sergey Ovchinnikov. 2023. “End-to-End Learning of Multiple Sequence Alignments with Differentiable Smith-Waterman.” *Bioinformatics*  39 (1). https://doi.org/](http://paperpile.com/b/SwMIKX/ANg1)[10.1093/bioinformatics/btac724](http://dx.doi.org/10.1093/bioinformatics/btac724)[.](http://paperpile.com/b/SwMIKX/ANg1)

[Popa, Serban C., Ichiro Inamoto, Benjamin W. Thuronyi, and Jumi A. Shin. 2020. “Phage-Assisted Continuous Evolution (PACE): A Guide Focused on Evolving Protein–DNA Interactions.” *ACS Omega* 5 (42): 26957–66.](http://paperpile.com/b/SwMIKX/SDQK)

[Rosenfeld, Lior, Michael Heyne, Julia M. Shifman, and Niv Papo. 2016. “Protein Engineering by Combined Computational and In Vitro Evolution Approaches.” *Trends in Biochemical Sciences* 41 (5): 421–33.](http://paperpile.com/b/SwMIKX/B0gn)

[Singh, Raushan K., Jung-Kul Lee, Chandrabose Selvaraj, Ranjitha Singh, Jinglin Li, Sang-Yong Kim, and Vipin C. Kalia. 2018. “Protein Engineering Approaches in the Post-Genomic Era.” *Current Protein & Peptide Science* 19 (1): 5–15.](http://paperpile.com/b/SwMIKX/AJnd)

[Wijma, Hein J., Maximilian J. L. J. Fürst, and Dick B. Janssen. 2018. “A Computational Library Design Protocol for Rapid Improvement of Protein Stability: FRESCO.” *Methods in Molecular Biology*  1685: 69–85.](http://paperpile.com/b/SwMIKX/IciQ)