THE PROTEIN ENGINEERING TOURNAMENT: AN OPEN SCIENCE BENCHMARK FOR PROTEIN MODELING AND DESIGN

Chase Armer[1*] **Hassan Kane**[2*] **Dana Cortade**[3*] **Dave Estell**[4] **Adil Yusuf**[2]

Henning Redestig[4] **TJ Brunette**[3] **Pete Kelly**[3] **Erika DeBenedictis**[3,6]

∗**Designates equal contribution**
[1]Columbia University, New York City, United States • [2]Medium Biosciences
[3]Align to Innovate, Cambridge, United States • [4]International Flavors and Fragrances
[5]Boston University, Boston, United States • [6]Francis Crick Institute, London, United Kingdom
E-mail: tournament@alignbio.org

ABSTRACT

The grand challenge of protein engineering is the development of computational models that can characterize and generate protein sequences for any arbitrary function. However, progress today is limited by lack of 1) benchmarks with which to compare computational techniques, 2) large datasets of protein function, and 3) democratized access to experimental protein characterization. Here, we introduce the Protein Engineering Tournament, a fully-remote, biennial competition for the development and benchmarking of computational methods in protein engineering. The tournament consists of two rounds: a first in silico round, where participants use computational models to predict biophysical properties for a set of protein sequences, and a second in vitro round, where participants are challenged to design new protein sequences which will be experimentally characterized with open-source, automated methods to determine a winner. At the Tournament's conclusion, the experimental protocols and all collected data will be open-sourced for continued benchmarking and advancement of computational models. We hope the Protein Engineering Tournament will provide a transparent platform with which to evaluate progress in this field and mobilize the scientific community to conquer the grand challenge of computational protein engineering.

1 INTRODUCTION

The field of computational protein engineering aims to improve our understanding of protein function and design by developing predictive models, which infer biophysical properties from a protein's sequence (Hie & Yang, 2022), and generative models, which compose protein sequences possessing a desired set of properties (Strokach & Kim, 2022). However, several obstacles are currently limiting model development in both paradigms. Predictive modeling has been notably hampered by the lack of large, complex, and diverse datasets. Most available datasets, such as those documented in ProtaBank (Wang et al., 2019), are limited in scope, predominantly mapping simple sequence-function relationships through single point mutations. This simplicity restricts the ability of predictive models to accurately characterize a wide range of protein functions under varied conditions. On the other hand, generative modeling faces its own set of challenges, primarily due to computational scientists' limited capacity to experimentally validate and characterize their protein designs. This gap significantly hinders the development and benchmarking of new generative design methods, as there is no standardized protocol or infrastructure to reliably test and compare the efficacy of these novel protein sequences. Addressing these challenges is crucial for the advancement of computational methods in protein engineering, as it would enable more accurate predictions and innovative designs in protein function.

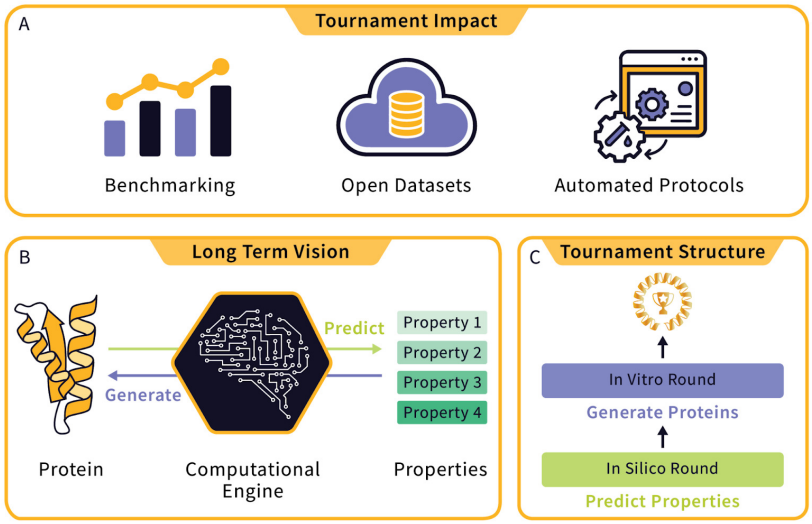

Figure 1: **Overview of the Tournament.** **(A)** The Tournament impacts the space by providing a transparent benchmark for computational methods, open datasets for the research community, and automated protocols for continued independent benchmarking. **(B)** The Tournament is designed to accelerate creation of computational techniques that can predict the biophysical properties for any given protein and generate a protein with any desired properties. **(C)** The Tournament consists of two sequential rounds: an in silico round for predicting properties of proteins and an in vitro round for generating proteins with specific properties.

The Protein Engineering Tournament aims to tackle the aforementioned obstacles by curating a series of tournaments which provide novel datasets for predictive modeling and experimental validation for generative design. By providing a transparent platform for benchmarking protein modeling and design methods, the Tournament hopes to reduce barriers to model development and validation (Figure 1A), and accelerate humanity's transition to reliable nano-scale engineering.

2 Related Approaches

Open datasets have long provided valuable opportunities for developing and benchmarking new methods in machine learning research. Computer vision datasets, such as MNIST (Deng, 2012) and ImageNet (Deng et al., 2009), not only provided individual research labs with a substrate for experimenting with new approaches but also created a yardstick with which to measure collective progress. Researchers in the protein engineering community have also utilized open datasets, such as FLIP (Dallago et al., 2021), TAPE (Rao et al., 2019), and ProteinNet (AlQuraishi, 2019), to encourage similar developments.

Science competitions take these efforts a step further by allowing researchers to test their computational methods on never-before-seen datasets. Perhaps the most notable example is the Critical Assessment of Structure Prediction (CASP) (Moult et al., 1995) a biennial event for computational protein structure prediction. Since its inception, CASP has become a crucial benchmark for the protein structure prediction community. By creating visibility around a single event, the competition has inspired an ambitious spirit among researchers to develop the best performing method, thereby encouraging a strong pace of method development. CASP has inspired the creation of similar competitions, like the Critical Assessment of Computational Hit-finding Experiments (CACHE) (Ackloo et al., 2022), which was created in the computational chemistry field to benchmark novel approaches for finding new small-molecule binders.

We believe there is a burgeoning opportunity to create a new scientific competition that addresses the unique challenges of predicting and engineering protein function. Computational research groups which lack the ability to experimentally characterize engineered proteins are currently unable to

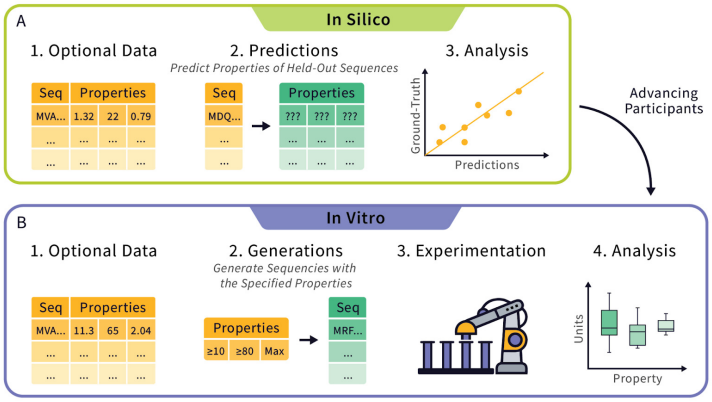

Figure 2: **Overview of Tournament Rounds.** The tournament consists of two rounds. **(A)** In the in silico round participants will predict properties for the provided protein sequences, with optional training data provided for specific events. Performance is evaluated by comparing the participant's predictions with the ground-truth values. Top performing participants will be selected to advance to the in vitro round. **(B)** In the in vitro round participants will generate protein sequences that possess a desired set of properties, which will then be expressed and characterized using cloud labs. Performance of the designed sequences will be evaluated as a weighted combination of the protein's properties; the exact evaluation metric will be event-dependent.

meaningfully evaluate the performance of their protein engineering methods. By introducing never-before-seen datasets on protein function and offering open-source experimental characterization of novel proteins, the Protein Engineering Tournament hopes to overcome this barrier and enable research groups from all backgrounds to participate in cutting-edge protein engineering research. In doing so, we expect the Tournament will become a unifying benchmark for the field.

3 THE PROTEIN ENGINEERING TOURNAMENT

3.1 TOURNAMENT STRUCTURE

The tournament will consist of two sequential rounds: the in silico round and the in vitro round (Figure 2).

In the in silico round, participants are presented with multiple events, each possessing a unique, never-before-seen dataset on protein function. Teams are tasked to develop models for inferring the biophysical properties of these protein sequences, under both supervised and zero-shot scenarios. Once the submissions are closed and the predictions evaluated, the final leaderboard will showcase the teams' performance in predicting the held-out experimental data.

In the in vitro round (Figure 2B), the teams will be asked to design protein sequences that maximize or satisfy certain biophysical properties. For example, an enzyme design challenge may ask for sequences which maximize enzymatic activity while staying above a specified threshold for protein stability and expression. Each team will submit a list of amino acid sequences that will be synthesized and experimentally characterized using automated laboratory protocols developed by the Tournament and its partners. Once characterized, a ranking algorithm will evaluate the submissions to produce a score for each participating team. The exact evaluation metric will depend on the protein target in question and will be tailored to the use-case for which it is being studied.

At the conclusion of the in vitro round, the Tournament will publish the final leaderboard, and the team with the highest performing proteins will be awarded the title of Protein Engineering Champion. The characterized protein sequences in the in vitro round and the datasets of the in silico round will be made publicly available. Furthermore, the automated protocols used to experimentally characterize the designed proteins will be made available to the public for continued use.

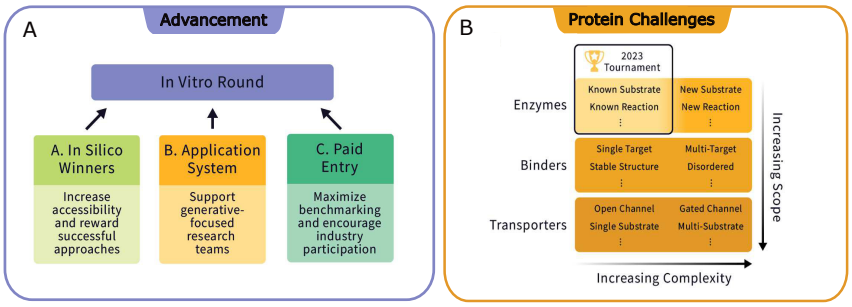

Figure 3: **Selecting Protein Design Challenges. (A)** There will be three avenues for participants to enter into the in vitro round, with each avenue catering to a unique audience. **(B)** The Tournament will continually expand to new domains of protein design; in each domain, we will continually select challenges that push the limits of current techniques. Our 2023 Pilot Tournament focused on Enzyme Design.

3.2 Participation

The first round of the Protein Engineering Tournament, the in silico round, will be open to any and all researchers interested in participating. Interested teams composed of one or more individuals will be able to download challenge data and upload final predictions. The in silico challenge will be open to the community for a specified amount of time, after which the submissions will be evaluated and the final leaderboard will be released.

To allow the greatest number of teams to participate in the in vitro round, while being conscious of the costs associated with DNA synthesis and experimental characterization, we propose three avenues for admission: 1) top performance in the in silico round, 2) submitting a written application, and 3) paid entry to cover the cost of experimentation (Figure 3A).

The first path focuses on rewarding innovative research teams who have demonstrated promising computational approaches in the in silico round. The second path is designed for groups with expertise in generative protein design but less experience in the in silico round's property prediction tasks. For the final path, a paid entry route will be available, primarily aimed at well-funded corporate labs, to broaden participation and allow a wider range of methods to be evaluated in the tournament.

3.3 Selecting our protein engineering challenges

The first Tournament will likely focus on a single function, with enzyme engineering and protein binder design as strong initial candidates. In future tournaments the design challenges will expand to encompass more domains of function (Figure 3B). The order in which we introduce new functions will be driven by practical application, technical feasibility, and amenability to high-throughput experimentation. As our computational methods improve, our challenges will expand into increasingly more difficult and complex domains, such that the frontier of scientific capabilities is always represented in the Tournament's challenges.

4 Tournament Creation and Outputs

4.1 Cloud laboratories and method development

To maximize our impact, we want the assays we develop for each tournament to be available long after the tournament has concluded. To accomplish this goal and capitalize on the benefits of automated experimentation, we will design and execute our experimental workflows in cloud laboratories.

Commercial cloud science laboratories, such as Emerald Cloud Labs, and academic cloud science laboratories, such as those found at Boston University and Carnegie Mellon, enable scientists to forgo the lab bench in favor of running experimental biology protocols in automation-enabled facili-

ties. In this paradigm researchers write their assays in a symbolic laboratory programming language that specifies the instructions for each step of their protocol. The scientists submit their protocols, wet-lab robots carry out each step of the protocol, and the resulting data is uploaded for analysis.

This approach offers the potential to greatly improve the accessibility and reproducibility of life science research by enabling scientists to share experimental protocols as easily as we share software. At the conclusion of each Tournament, the protocols we developed to execute characterization assays in the in vitro round will be made openly available to the scientific community. Therefore, researchers will be able to continually benchmark their computational methods on the same standardized assays even after the Tournament has concluded. Since each Tournament will introduce new protein engineering challenges, this approach will lead to an ever-expanding corpus of open-source protein engineering workflows to help benchmark new computational methods for years to come.

4.2 Tournament Datasets

The automated experimental protocols discussed above will be used to generate the datasets for the in silico and in vitro rounds. Furthermore, the Tournament will work with academic and corporate research entities to make unpublished datasets of protein function available as predictive challenges in the in silico round.

4.3 Analysis of the State of the Field

We will aggregate the learnings from our tournament's results into a State of the Field report. This report will analyze the performance of our participant's computational approaches, noting the relative standing of different techniques across different challenges, and discussing our collective progress throughout the various domains of protein engineering.

4.4 Governance

The Protein Engineering Tournament is operated by Align to Innovate, a non-profit dedicated to improving the reproducibility, scalability, and shareability of life science research through community-driven initiatives. The Tournament will be run by a combination of Align to Innovate employees and volunteers from the protein engineering community.

5 PILOT TOURNAMENT

A pilot tournament began May 1st 2023 and will be completed by the end of March 2024 with the theme of Enzyme Design based on six multi-objective datasets received from both industry and academic groups. Initial interest in the pilot tournament led to the registration of over 90 individuals assembled into 30 teams, representing a mix of academic (55%), industry (30%), and independent (15%) teams. For the pilot tournament, the in vitro round experimentation is performed in-house by a corporate partner. Development of the cloud laboratory assays for future tournaments is currently underway within Align to Innovate.

6 CONCLUSION

The Protein Engineering Tournament introduces an innovative, community-driven platform to accelerate the advancement of computational protein engineering. This open science initiative combines in silico and in vitro methods, employing cloud laboratories to ensure reproducibility and continued access to experimental workflows. By creating a unified benchmark, the Tournament will stimulate collaboration and competition among researchers and create opportunities for the community to evaluate their computational models on novel protein engineering challenges. Further, through its integration with Align to Innovate, the Tournament builds upon the strengths of a diverse scientific community, fostering transparency and sharing of knowledge. As we look ahead to the first official Tournament, we anticipate that this initiative will contribute significantly to the evolving landscape of protein engineering and enable the scientific community to conquer the grand challenge of protein design.

REFERENCES

S. Ackloo et al. Cache (critical assessment of computational hit-finding experiments): A public–private partnership benchmarking initiative to enable the development of computational methods for hit-finding. *Nature Reviews Chemistry*, 6:287–295, 2022.

M. AlQuraishi. Proteinnet: a standardized data set for machine learning of protein structure. *BMC Bioinformatics*, 20:311, 2019.

C. Dallago et al. Flip: Benchmark tasks in fitness landscape inference for proteins. *bioRxiv*, pp. 2021.11.09.467890, 2021. doi: https://doi.org/10.1101/2021.11.09.467890.

J. Deng, W. Dong, R. Socher, L-J Li, K. Li, and F-F Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–55. IEEE, 2009.

L. Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–2, 2012.

B. L. Hie and K. K. Yang. Adaptive machine learning for protein engineering. *Curr. Opin. Struct. Biol.*, 72:145–152, 2022.

J. Moult, J. T. Pedersen, R. Judson, and K. Fidelis. A large-scale experiment to assess protein structure prediction methods. *Proteins: Structure, Function, and Genetics*, 23(3):ii–v, 1995.

R. Rao et al. Evaluating protein transfer learning with tape. *Adv Neural Inf Process Syst*, 32: 9689–9701, 2019.

A. Strokach and P. M. Kim. Deep generative modeling for protein design. *Curr. Opin. Struct. Biol.*, 72:226–236, 2022.

C. Y. Wang et al. Protabank: A repository for protein design and engineering data. *Protein Sci.*, 28: 672, 2019.