Building closed-loop frameworks for AI-guided protein design

Thesis by
Alexandre Luiz Lourenço

In Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2025
Defended May 30, 2025

Alexandre Luiz Lourenço
ORCID: 0009-0005-0758-2968

ACKNOWLEDGEMENTS

There are a lot of people to thank for getting me to this place where I can be here, writing my thesis.

My dad, for being the ever present example of a scientist/engineer and my mom, for nurturing my science, technology, engineering, and math (STEM) affinities, balancing providing structure, and giving me the independence to creatively pursue my own path, from the science fair in 6th grade to today.

Vóvó, for her love and support for her only grandson.

Richard Murray and the rest of the Caltech 2014 iGEM team, for accepting me into the Caltech community 11 years ago as a high school student with much to learn. Thank you for believing in my potential.

Matt Thomson, for being the best scientific improv partner I can ask for and teaching me to be a better scientist and leader.

Thank you to my committee members, Kai Zinn, Pamela Bjorkman, and Steve Mayo, for forcing me to level up my rigor, not only helping me become a better presenter, but also a better thinker.

The Thomson lab members for lending an ear, a hand, or a tastebud whenever needed, whether it be for reasons scientific or otherwise. Special shoutout to Shichen Liu, Arjuna Subramanian, and Zach Martinez for being the best fellow intrepid explorers to navigate the protein design maze with.

Monica Barsever, for originally reaching out and reconnecting and leading the charge to make the 2024 IEA iGEM team happen and steering the ship around the constant obstacles and docking the team at our final destination of Paris, where they were able to present their work. Thank you to the Institute of Educational Advancement, especially Nicole Endacott and Deborah Monroe, for believing in us and providing the logistical support to make it happen.

Thank you to all of the backers who made the work possible, especially Linda J. McManus and William B. Bridges, for their generous financial support and Thomson lab and the other Thomson lab members who provided space and support.

To Mamadou Diallo, for generously donating his time to determine a scientific direction for the project both manageable in the short summer timeframe of the project, yet ambitious enough to make a dent in the universe.

Thank you to Ariel Furst, for lending her subject-matter expertise and lending the capacity to test our protein designs.

And of course, to the team members themselves: Lucas Garcia, Maya Kapur, Troy Reyes, Willow Nauber, Ray Chung, Jael Santos, and Camille Dahlgren, for their faith in me as a mentor and committing fully to the process. To their parents, thank you for the patience, understanding, and support to make it all happen.

Thank you to Tristan Jensen for being the best partner in getting the biotechnology out into the world in a form that is fun and makes the world a better place.

Last, but certainly not least, thank you to Maria Horbatenko for her love, patience, and support through the long days and nights of my scientific journey. Onward!

ABSTRACT

The design of proteins with tailored properties remains a central challenge in protein engineering, with profound implications for therapeutics, sustainable manufacturing, and environmental remediation. Recent advances in artificial intelligence have dramatically improved our ability to design novel proteins, yet the precision required for many applications remains elusive. This thesis details the development and implementation of closed-loop frameworks that integrate AI-guided protein design with quantitative experimental data to iteratively improve design outcomes.

First, I present Protein CREATE (Computational Redesign via an Experiment-Augmented Training Engine), a high-throughput platform that combines phage display with molecular counting techniques to generate quantitative binding data at scale. This platform enables rapid evaluation of thousands (and is in the process of being scaled to millions) of designed protein variants against multiple targets simultaneously.

In subsequent chapters, I explore two separate strands of protein design as they reach for each other to close the loop. One thread focuses on collecting data on binders I engineered to the interleukin 7 receptor alpha (IL7RA) and Insulin receptor while the other investigates the value data, even when limited, adds to improve the design process of enzymes to solve a pressing environmental remediation problem: cleaning up per and polyfluoroalkyl substances (PFAS).

While all of the targets discussed so far have benefited from developments in artificial intelligence, I explore one target where the benefits are limited, the human sweet taste receptor. Here, I leverage alternative computational methods coupled to experimental testing to chart a course for design.

Finally, I discuss the technologies we are integrating within the Protein CREATE framework to enable rapid *in vitro* and *in vivo* testing.

Throughout my PhD, I have been bringing the two threads of computational design and experimental characterization closer together for not only theoretically interesting, but also practically relevant, engineering cases. The methodologies developed here represent a significant advancement in our ability to design proteins with precisely tailored properties for diverse applications.

PUBLISHED CONTENT AND CONTRIBUTIONS

Lourenço, Alec L et al. (Jan. 2025). "Protein CREATE enables closed-loop design of de novo synthetic protein binders". In: *bioRxiv*, p. 2024.12.20.629847. DOI: 10.1101/2024.12.20.629847.
A.L.L. participated in conceiving the project, designing and performing the experiments, writing the code, and preparing the manuscript.

Garcia, Lucas et al. (Oct. 2024). "Design of Novel Dehalogenases using Protein Large Language Models". In: *bioRxiv*, p. 2024.10.28.620469. DOI: 10.1101/2024.10.28.620469.
A.L.L. participated in conceiving the project, designing and performing the experiments, and preparing the manuscript.

Subramanian, Arjuna M et al. (2023). "Unexplored regions of the protein sequence-structure map revealed at scale by a library of foldtuned language models". In: *bioRxiv*, p. 2023.12.22.573145. DOI: 10.1101/2023.12.22.573145.
A.L.L. designed and performed wetlab experiments.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

Chapter 1

DEVELOPMENT OF PROTEIN CREATE - A FRAMEWORK FOR CLOSED-LOOP PROTEIN BINDER DESIGN

1.1 Introduction: Motivating the need for improved data collection to further empower AI-driven protein design

The ability to produce proteins with defined functional properties is the central goal of protein engineering, holding enormous potential for a vast array of human endeavors such as therapeutic development, bioremediation, and sustainable manufacturing (Rocklin et al., 2017). Despite decades of progress, the astronomical size of protein sequence space continues to make systematic mapping of protein sequence to function a challenge. Recent breakthroughs in artificial intelligence have made the design of proteins with useful features increasingly feasible (Jumper et al., 2021; Lin, Akin, Rao, Hie, Zhu, Lu, Smetanin, et al., 2023). These approaches have made use of vast troves of natural sequences to learn internal patterns within the data, leading to the ability not only to predict key protein properties, such as three-dimensional structure and thermostability, but also to begin to control them (Ferruz, Schmidt, and Höcker, 2022).

Despite the tremendous progress in *de novo* design of proteins, current methods lack the precision necessary to optimize protein properties for many applications. For instance, taste and olfaction modulators require both high on and off rates so that the sensory receptors can quickly respond to the stimulus and reset. Other applications require optimization of multiple properties. For instance, therapeutics should have low immunogenicity and off-target binding in addition to their primary function of binding a given target. Collection and integration of experimental data of protein variant function should be able to refine these crude predictions, but both remain a challenge. Biological data collection suffers a tradeoff between throughput and data quality. High throughput display technologies such as phage and ribosome display may be able to screen $> 10^9$ variants, but the qualitative rather than quantitative nature of the collected data limits its use. While other methods, such as surface plasmon resonance (SPR), provide detailed kinetic data, they rely on purified proteins and repeated measurements under a variety of conditions, making them difficult to scale. Data integration into recently-developed protein design

models is hindered by the mismatch between the way these models are trained, often on auxiliary tasks such as next token prediction using readily available sequencing data, as opposed to the collected experimental data, which is directly indicative of protein function.

In my PhD, I set out to demonstrate how solving the collection and integration problems for a subset of protein functional data, binding data, could improve protein engineering. This led to the development of Protein CREATE (Computational Redesign via an Experiment-Augmented Training Engine), an integrated computational and experimental pipeline that bridges the gap between AI-driven protein design and quantitative experimental validation.

Recent advances in protein large language models (LLMs) have enabled the generation of diverse protein sequences predicted to fold into varied structures, bind to specific targets, and catalyze novel reactions. However, these computational models are often limited by the quality and quantity of training data, particularly when it comes to predicting and optimizing binding interactions. The "physics-free" approaches using artificial intelligence have shown remarkable success, with design success rates sometimes exceeding 10% without experimental optimization (Nijkamp et al., 2022). However, these algorithms often struggle to extrapolate their predictions to unnatural designed sequences, highlighting the need for experimental data to refine and improve the models.

While computational metrics such as the Alphafold interface predicted template modeling (iPTM) score have proven useful as computational predictors of binding (Bennett et al., 2023), they are far from sufficient to guarantee experimental binding success. Leading practitioners have emphasized achieving "one design, one binder" through algorithmic improvements while limiting experimental search. However, history suggests that leveraging search—specifically, the development of fast and cheap experimental data collection and integration—will be a more effective design approach in the long term.

This chapter describes the development of Protein CREATE, a framework designed to collect quantitative binding data at scale and integrate it into the protein design process, enabling a closed-loop approach to protein binder design. The system combines high-throughput experimental screening with computational modeling to iteratively improve protein binding predictions and designs.

1.2 Overview of Protein CREATE binding assay methodology

Protein CREATE represents a significant advancement in protein binding characterization by combining high-throughput experimental capabilities with quantitative readouts. The core of the platform is a novel "binding by sequencing" assay that enables rapid, parallel evaluation of thousands to millions of protein variants against multiple targets.

"Binding by sequencing"

The fundamental concept behind Protein CREATE is what we term "binding by sequencing" — a methodology that quantifies protein binding through DNA sequencing readouts. This approach provides several advantages over traditional binding assays:

1. **High throughput:** The assay can evaluate thousands to millions of variants simultaneously.
2. **Quantitative output:** Unlike traditional display methods that provide primarily qualitative binding information, Protein CREATE generates quantitative binding data.
3. **Multiplexed target testing:** Multiple targets can be screened in parallel, allowing assessment of both on-target binding and off-target interactions.

The binding by sequencing workflow begins with DNA libraries encoding the protein variants of interest. These libraries are first cloned into display vectors, expressed on the surface of a display platform (in our case, bacteriophage), and then assayed for binding against immobilized target proteins. The key innovation lies in how we process and analyze the bound phage to extract quantitative binding information, using next-generation sequencing with molecular counting techniques.

Why Phage? Genetically encoded, high throughput variant display

We selected bacteriophage, specifically T7 bacteriophage, as our display platform for several compelling reasons (Pande, Szewczyk, and Grover, 2010):

1. **Genetic linkage:** Phage display creates a physical link between the displayed protein (phenotype) and its encoding DNA (genotype), enabling direct identification of binding variants through DNA sequencing (Pande, Szewczyk, and Grover, 2010).

2. **High throughput capacity:** T7 phage libraries can reach titers of 10^{10} pfu/mL, enabling the screening of vast numbers of variants in a single experiment (Rosenberg et al., 1996; Krumpe and Mori, 2004). Even at 1000 copies per variant, this system could theoretically screen up to 10^7 unique variants.

3. **Robust display:** The T7 phage capsid can display a diverse range of protein sizes and structures, making it suitable for a wide variety of binder designs (Krumpe and Mori, 2004; Rosenberg et al., 1996).

4. **Speed and simplicity:** The T7 phage life cycle is rapid, allowing for fast library generation and amplification. The entire phage preparation process can be completed within 1-2 days.

5. **Compatibility with bacterial expression:** The use of E. coli as the host organism simplifies library preparation and reduces costs compared to eukaryotic display systems.

The process begins with cloning DNA libraries into T7 bacteriophage backbones, followed by packaging and infection of helper E. coli to display the protein library on the phage capsid surfaces (Rosenberg et al., 1996; Krumpe and Mori, 2004). After purification, the phage library is ready for binding assays against target proteins.

Beyond enrichment: counting individual binders using unique molecular identifiers (UMIs)

A significant limitation of traditional phage display is its qualitative nature, typically reporting binding as simple enrichment ratios after multiple rounds of selection. Protein CREATE overcomes this limitation through the incorporation of unique molecular identifiers (UMIs) into the sequencing preparation (Islam et al., 2014).

UMIs are short, random nucleotide sequences that are added to each DNA molecule prior to amplification (T. Smith, Heger, and Sudbery, 2022; Fu et al., 2011). Each phage particle receives a unique UMI tag, allowing us to:

1. **Count individual binding events:** Rather than measuring bulk enrichment, we can count the exact number of individual phage particles that bind to a target (Fu et al., 2011; T. Smith, Heger, and Sudbery, 2022).

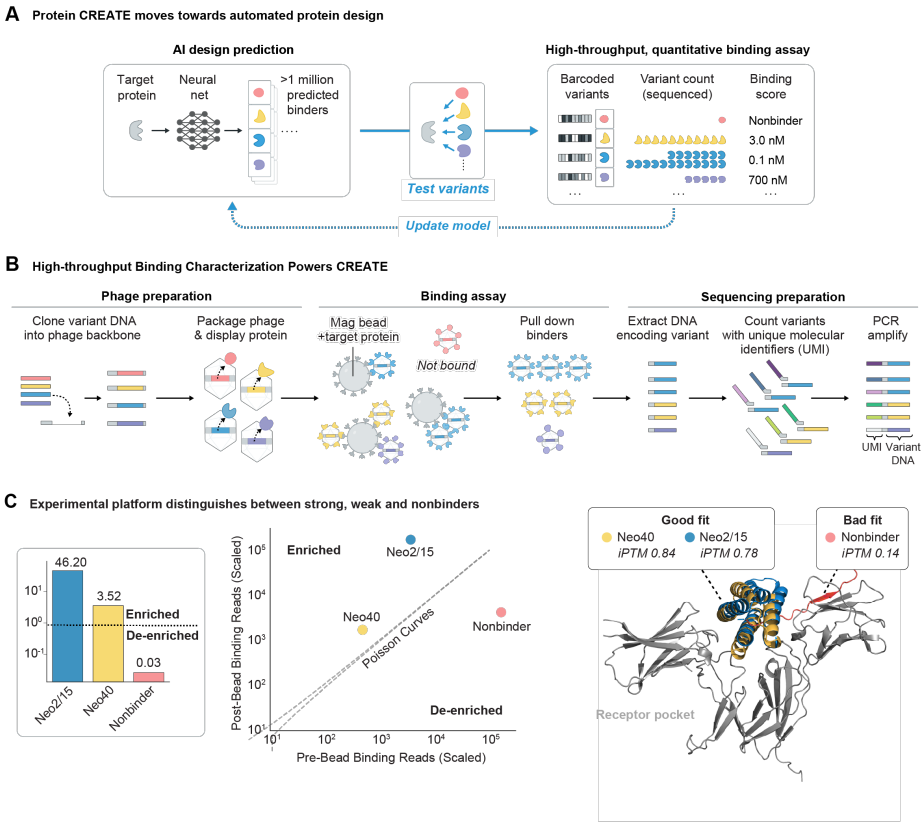


Figure 1.1: Overview of the Protein CREATE platform combining phage display with molecular counting techniques for quantitative assessment of protein binding. The workflow includes library generation, phage display, binding assays with UMI tagging, and computational analysis to extract quantitative binding information.

2. **Reduce amplification bias:** By grouping reads with the same UMI, we can correct for PCR bias that might otherwise skew quantification (Islam et al., 2014).

3. **Enable molecular counting:** The use of UMIs facilitates accurate molecular counting, which in turn makes pseudo-Kd estimation possible.

The workflow involves extracting DNA from bound phage, labeling it with UMIs, and then processing it for next-generation sequencing. By comparing the UMI-corrected counts of each variant before and after binding, we can calculate enrichment scores that more accurately reflect binding affinity.

Our validation experiments with proteins of known binding affinities demonstrated that this UMI-based counting approach provides enrichment values that correlate

well with actual Kd measurements (Silva et al., 2019). For instance, when testing IL-2 mimetics with known binding affinities, the stronger binder Neo2/15 (K_D = 18.8 nM) showed approximately 13-fold higher enrichment than the weaker binder Neo40 (K_D = 260 nM), closely matching their relative binding strengths (Silva et al., 2019).

Eye toward automation: production + screening on beads can be automated using a liquid handling robot (Opentrons)

A key consideration in the development of Protein CREATE was the potential for automation to further increase throughput and reproducibility. The assay was designed with compatibility with liquid handling robots, particularly the Opentrons platform, in mind.

The use of magnetic bead-based separations is particularly well-suited for automation, as it eliminates the need for centrifugation or filtration steps that can be challenging to automate. Our workflow involves:

1. **Magnetic bead preparation:** Streptavidin-coated magnetic beads are loaded with biotinylated target proteins.
2. **Automated binding assays:** Phage libraries are incubated with target-coated beads, followed by automated washing steps.
3. **Automated DNA extraction:** Bound phage DNA is extracted directly from the beads for sequencing preparation.

The entire process from target immobilization to DNA extraction can be programmed on liquid handling robots, allowing for increased throughput, reduced human error, and improved consistency between experiments. This automation potential is particularly valuable for implementing the closed-loop design-build-test cycles that Protein CREATE aims to enable, where rapid iteration is essential.

Data analysis pipeline (with simulations) to extract meaningful quantitative binding data

To translate the raw sequencing data into actionable binding information, we developed a comprehensive data analysis pipeline. This pipeline includes:

1. **Sequence processing:** Identification and filtering of reads with valid priming regions, followed by UMI detection and collapsing.

2. **Variant matching:** Translation of processed sequences to amino acid strings and matching against the design database.

3. **Enrichment calculation:** Normalization of read counts and calculation of enrichment values for each variant.

By analyzing data from multiple dilutions of the same sample, we identified the optimal range for UMI-based quantification, where coverage is sufficient but UMI collision events are minimized (Fu et al., 2011; T. Smith, Heger, and Sudbery, 2022).

The final output of the pipeline provides not just a ranked list of binders but quantitative enrichment scores that correlate with binding affinity, enabling more informed selection of candidates for further characterization and refinement.

1.3 Validation with known binders and quantitation of binding affinities

To validate the Protein CREATE platform's ability to accurately rank protein binding strengths, we tested a set of engineered IL-2 receptor binders (IL-2R$\beta\gamma$) with well-characterized binding affinities previously determined by surface plasmon resonance (SPR) (Silva et al., 2019; Cao et al., 2022).

Testing with protein variants of known affinity

We selected Neo2/15 and Neo40, two IL-2 mimetics designed by the Baker lab with known binding affinities of 18.8 nM and 260 nM, respectively (Silva et al., 2019), along with a non-binding control protein. These proteins were displayed on T7 bacteriophage and assayed for binding to IL-2R$\beta\gamma$-coated beads following our established protocol.

The results demonstrated clear separation between binders and non-binders (Cao et al., 2022; Silva et al., 2019):

1. **Strong binder (Neo2/15):** Showed high enrichment values consistent with its strong binding affinity (Silva et al., 2019).

2. **Weak binder (Neo40):** Showed moderate enrichment, approximately 13-fold lower than Neo2/15 (Silva et al., 2019).

3. **Non-binder control:** Showed de-enrichment (values < 1), indicating specific depletion during the binding assay.

Critically, the approximately 13-fold difference in enrichment between Neo2/15 and Neo40 closely matched their 14-fold difference in Kd values (18.8 nM vs. 260 nM) (Silva et al., 2019), demonstrating that Protein CREATE can not only distinguish between binders and non-binders but also accurately rank binding strengths.

1.4 Conclusion

The development of Protein CREATE represents a significant advancement in our ability to collect quantitative binding data at scale for protein engineering applications. By combining the high-throughput nature of phage display with the quantitative capabilities of UMI-based molecular counting (T. Smith, Heger, and Sudbery, 2022; Fu et al., 2011; Islam et al., 2014), we have created a platform that bridges the gap between computational design and experimental validation.

Key innovations of the Protein CREATE platform include:

1. The "binding by sequencing" methodology that enables quantitative assessment of binding strengths.
2. The incorporation of UMIs for accurate molecular counting and pseudo-Kd estimation.
3. A design compatible with automation for increased throughput and reproducibility.
4. A comprehensive data analysis pipeline that translates raw sequencing data into actionable binding information.

Validation with known binders demonstrates that Protein CREATE can accurately rank binding affinities (Silva et al., 2019), providing a reliable means to evaluate thousands to millions of designed variants in parallel. This capability is particularly valuable in the context of machine learning-based protein design, where large, high-quality datasets are essential for model training and refinement.

As demonstrated in subsequent chapters, Protein CREATE enables screening and discovery of novel binders to therapeutic targets with improved properties such as reduced off-target binding and maintained function despite low sequence homology to parent designs. We will now turn our attention to protein design algorithms and how they can be improved with data.

Chapter 2

LEARNING FROM DATA TO IMPROVE AI-BASED PROTEIN DESIGN

2.1 Introduction to language modeling for protein design

The emergence of language models for protein design represents one of the most significant recent advances in computational protein engineering (Wu et al., 2021). These models, inspired by techniques developed for natural language processing, treat protein sequences as "sentences" composed of amino acid "words" (Nijkamp et al., 2022; Ferruz, Schmidt, and Höcker, 2022). By training on vast databases of natural protein sequences, these models learn the underlying patterns and dependencies that govern protein structure and function, enabling them to generate novel sequences that adhere to these learned constraints.

Protein Language Models and Their Capabilities

Two prominent examples of protein language models that have transformed the field are ESM-2 and ProtGPT2, each with distinct architectures and capabilities. Although much of my work involves the former, I will highlight both here to draw attention to the relevant architecture differences.

ESM-2 (Evolutionary Scale Modeling) developed by Meta AI Research (formerly Facebook AI Research) represents a breakthrough in protein language modeling. Built on a transformer architecture with up to 15 billion parameters in its largest variant, ESM-2 was trained on the vast UniRef50 database containing millions of diverse protein sequences. This self-supervised training approach, predicting masked amino acids within sequences, enables ESM-2 to capture complex evolutionary patterns without requiring labeled data or structural information (Lin, Akin, Rao, Hie, Zhu, Lu, Smetanin, et al., 2023).

ESM-2's power extends beyond sequence generation to enable remarkable downstream applications. The most notable is ESMFold, which leverages the learned sequence representations to predict protein structure with accuracy rivaling AlphaFold2 but with significantly greater speed (Lin, Akin, Rao, Hie, Zhu, Lu, Santos Costa, et al., 2023). ESMFold can produce high-quality structural predictions in seconds to minutes rather than hours without relying on a multiple sequence align-

ment like AlphaFold (Jumper et al., 2021). Additionally, ESM-2's representations have proven valuable for predicting protein properties such as stability, function, and evolutionary fitness, making it a versatile tool for protein engineering.

ProtGPT2 takes a different approach, building on the GPT-2 architecture that has proven successful in text generation. Unlike ESM-2's masked language modeling, ProtGPT2 is trained using an autoregressive approach, generating proteins one amino acid at a time from left to right. This enables ProtGPT2 to generate completely novel protein sequences by sampling from the learned distribution. The model was trained on a curated dataset of natural protein sequences, learning to capture the sequential dependencies and patterns that characterize functional proteins (Ferruz, Schmidt, and Höcker, 2022).

ProtGPT2 has demonstrated remarkable capability in generating diverse, novel protein sequences that fold into stable structures despite having low sequence identity to any natural protein. This property makes it particularly valuable for exploring uncharted regions of protein sequence space that might harbor novel functions or improved properties.

Both ESM-2/ESMFold and ProtGPT2 represent significant advances in protein design, but they also have limitations. Most notably, while these models excel at capturing patterns from natural proteins, both struggle to design proteins with specific features, such as binding interfaces or catalytic functions, which often require precise spatial arrangements of amino acids. The proteins these models are trained on are drawn from sequence space in general, not a particular subpopulation, such as an enzyme subfamily. To draw an analogy to human text, these models may be used to understand if a protein has proper "grammar," but not what specifically is being "said" (the function of the protein). An additional limitation is that these models may struggle to generalize to sequences far from the natural distribution they were trained on (Ferruz, Schmidt, and Höcker, 2022). We have observed that many designed proteins often fail to express, possibly because these models fail to account for complex folding pathways or the amino acid frequencies that deviate significantly from those of the heterologous host.

While either model can serve as a useful foundation, additional data or feedback is critical to guide the design process. The following sections describe our efforts to augment these models. I start with a simple example — using three-dimensional data to devise a binder design strategy through context-based inverse folding. As the chapter progresses, I will explore more involved approaches involving critic models

and a closed-loop framework utilizing this critic to integrate experimental data.

2.2 Introducing spatial information into design using context-based inverse folding

A significant limitation of sequence-only language models is their inability to explicitly account for three-dimensional structure, particularly in designing proteins for specific interaction interfaces. To address this limitation, I developed and applied a simple approach we term "context-based inverse folding," which leverages structural information to explicitly design for a specific protein-protein binding interaction.

Principles of Context-Based Inverse Folding

Inverse folding refers to the problem of designing a protein sequence that will fold into a specified three-dimensional structure—essentially the reverse of the protein folding problem. Recent advances in this area include models like ESM-IF, which can predict sequences likely to adopt a given structure (Hsu et al., 2022). Context-based inverse folding approach extends this concept by incorporating additional contextual information, such as binding partners or interaction interfaces.

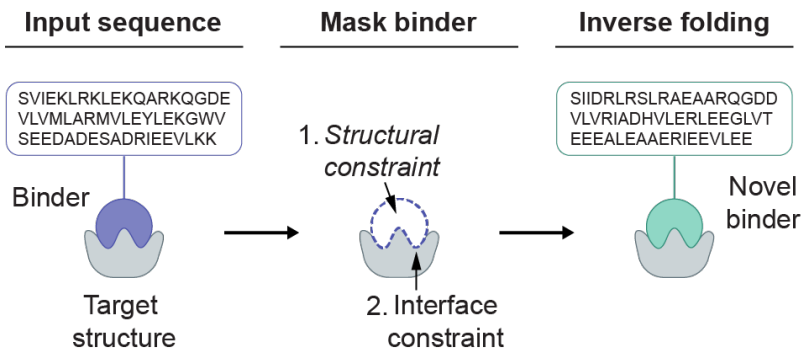


Figure 2.1: Context-dependent inverse folding approach. Starting with a known protein-protein complex structure, the binder sequence is masked while maintaining structural information and interface constraints, a modified inverse folding model then predicts novel binder sequences compatible with both the desired fold and target interaction.

The methodology involves:

1. **Starting with a known protein-protein complex structure:** We begin with the structure of a known binder-target pair, such as the IL7RA-antagonist

complex (PDB: 7OPB), or a predicted interaction pair (e.g. derived from AlphaFold Multimer (Jumper et al., 2021).

2. **Masking the binder sequence:** We mask only the sequence of the binder protein while retaining all structural information and interface constraints.

3. **Sequence prediction with context:** Using ESM-IF, we predict novel sequences for the binder that are compatible with both the desired fold and the interaction with the target protein.

This approach preserves critical binding interactions by explicitly incorporating the target protein during the design process, while allowing exploration of diverse sequences for the binder. The resulting designs maintain compatible surface residues for interaction while potentially having very different overall sequences from the parent molecule.

Implementation and Results

To evaluate the method, we applied it to the redesign of IL7RA binders, using the crystal structure of a previously designed mini-protein antagonist bound to IL7RA (PDB: 7OPB) as a template (Belarif et al., 2018). The context-based inverse folding model generated a diverse set of candidate binder sequences, many with less than 50% sequence identity to the parent binder while maintaining predicted binding capability, as evidenced by their interface predicted template modeling (iPTM) scores.

The iPTM has emerged as a valuable metric for assessing protein design success (Jumper et al., 2021). Values higher than 0.8 represent confident high-quality predictions, while scores below 0.6 suggest likely failed predictions. This creates a binary classification framework where iPTM can effectively discriminate between successful and unsuccessful protein designs.

When applied to protein design, iPTM scoring offers a computational pre-screening method to identify promising candidates before experimental validation. The iPTM score specifically evaluates the accuracy of predicted interfaces between protein chains, with higher values indicating robust, well-defined interactions between designed components, but, importantly, not the strength of such interactions. Still, this approach has transformed the protein design workflow, allowing researchers to rapidly filter designs and focus experimental resources on candidates with higher probability of success.

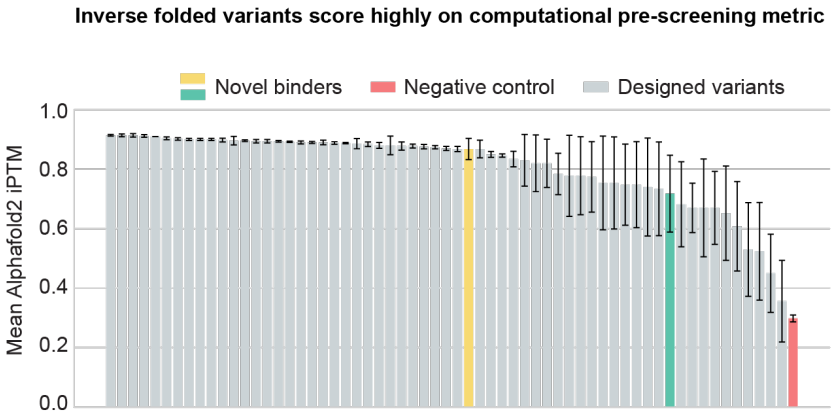


Figure 2.2: Interface predicted template modeling (iPTM) scores for variants using context-dependent inverse folding designed to bind to IL7RA. As will be discussed in a future chapter, two designs shown in the figure were characterized as binders using biochemical assays.

However, this binary classification approach has important limitations. iPTM scores can be artificially inflated by the presence of disordered regions or accessory domains that aren't part of the binding interface. Research has shown that current iPTM metrics don't always correlate with structural alignment RMSD scores, meaning some designs with high iPTM scores may still have poor structural quality when experimentally validated (Dunbrack Jr, 2025; Ferruz, Schmidt, and Höcker, 2023). Additionally, iPTM scores between 0.6 and 0.8 represent a "grey zone" where predictions could be either correct or incorrect, challenging the binary classification paradigm. The metric also struggles with intrinsically flexible regions or domains, potentially missing functionally important dynamic interactions.

Critic models for learning binding preferences

One approach to integrating experimental data involves training "critic" models that can predict the probability of binding success based on sequence features. Sequences can be tokenized and input into a multi-layer transformer network, where the final layer can give a single-dimensional output used to assign that sequence a "score" for some function (e.g. binding). Batch evaluation for many sequences using these models can be extremely fast, on the order of seconds compared to the minutes to hours per sequence co-folding and iPTM evaluation can take. These can also speed up the predictions for other lengthy computational methods, such as molecular dynamics (MD) simulations to which we turn to next.

Molecular Dynamics for Predicting Binding Affinities

Molecular dynamics (MD) simulations represent a computational method for studying the physical movements and interactions of atoms and molecules over time. The technique works by numerically solving Newton's equations of motion for a system of interacting particles, where forces between the particles and potential energy are calculated using molecular mechanics force fields(Abraham et al., 2015). This physics-based approach serves as a natural complement for the decidedly non-physical artificial intelligence algorithms used for sequence generation and evaluation previously discussed, providing time-resolved insights into conformational changes and non-covalent interaction strengths.

I collaborated with Jiapei Miao in Matt Thomson's lab to perform molecular dynamics simulations on the two context-dependent inverse fold-designed binders that were described in Figure 2.2. While both were binders, one (iPTM 0.7) was a significantly stronger binder than the other (iPTM 0.9).

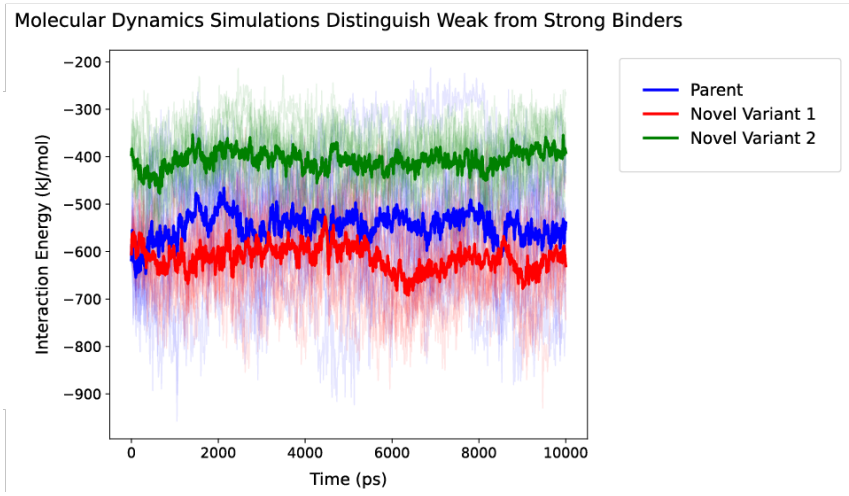


Figure 2.3: Molecular dynamics is able to explain binding strength of IL7RA binding proteins. Lower interaction energy corresponds to tighter binding, in line with the experimentally observed data.

Using GROMACS with the AMBER99SB force field, we simulated the behavior of both the parent mini binder (a strong binder) and our novel variants when bound to IL7Rα. The simulations corroborated our experimental Bioplex immunodetection assay results, revealing that the parent and novel variant 1 exhibited lower interaction energies compared to novel variant 2, which aligned with their respective binding strengths observed experimentally. These computational predictions were particularly effective at distinguishing between strong and weak binders, with the energy

profiles clearly separating variant 1 ($K_D \approx 88$ nM) from variant 2, which showed reduced binding in experimental assays. This orthogonal, non-binary prediction showcases MD as a useful tool in the protein design toolkit.

Actor-critic reinforcement learning for protein design

The critic models described above can be further integrated into a reinforcement learning (RL) framework to create a closed-loop system for protein design (Sutton and Barto, 2018). Reinforcement learning, which has demonstrated remarkable success in complex domains like game playing and robotic control, provides a natural framework for protein design by treating the design process as a sequential decision-making problem (Wu et al., 2021).

Background on actor-critic frameworks

Actor-critic methods represent a class of reinforcement learning algorithms that combine two components: an actor that determines which actions to take, and a critic that evaluates those actions (Sutton and Barto, 2018). In protein design, these components translate naturally:

- The **actor** is a generative model that produces protein sequences.
- The **critic** evaluates the quality of these sequences, typically by predicting their binding affinity or other desired properties.

This architecture addresses several key challenges in protein design:

1. **Exploration-exploitation balance**: The framework can systematically explore the vast protein sequence space while focusing computational resources on promising regions.
2. **Reality gap management**: By periodically validating with more rigorous computational methods or experimental data, the system can calibrate its predictions to better align with real-world outcomes.
3. **Sequence diversity**: Specific mechanisms can be incorporated to encourage diversity, preventing convergence to a single solution.
4. **Iterative improvement**: The feedback loop between generation and evaluation enables progressive refinement of designs.

Protein language model-based actor implementation

For our actor component, we implemented a novel approach based on ESM-2 that generates sequences through random position unmasking (Lin, Akin, Rao, Hie, Zhu, Lu, Smetanin, et al., 2023). Unlike traditional left-to-right generation, this approach:

1. Begins with a partially or completely masked protein sequence of predetermined length.
2. Iteratively unmasks random positions, with the language model predicting amino acids based on the surrounding context.
3. Continues until all positions are filled, creating a complete protein sequence.

This random unmasking strategy allows the language model to leverage bidirectional context when making each prediction, potentially capturing long-range dependencies more effectively than unidirectional generation (Lin, Akin, Rao, Hie, Zhu, Lu, Smetanin, et al., 2023). Early in training, reference sequences with high iPTM scores are partially masked and then reconstructed, helping the generator learn the characteristics of successful binders. As training progresses, the system gradually transitions to generating completely novel sequences from fully masked templates.

iPTM regression critic

The critic component consists of a regression model that predicts the interface predicted template modeling (iPTM) score for a given protein-receptor pair. This model:

1. Encodes both the candidate protein and receptor sequences using ESM-2 embeddings.
2. Combines these embeddings to capture potential interaction features.
3. Applies a regression model to predict the iPTM score, which serves as a computational proxy for binding affinity.

The iPTM metric, derived from AlphaFold predictions for protein-protein interfaces, provides a reasonably effective proxy for binding potential, though it has known limitations as discussed earlier (Dunbrack Jr, 2025). To overcome these limitations, the

critic undergoes periodic recalibration based on more rigorous structural predictions from state-of-the-art models like Boltz-1 (based on AlphaFold3), which provide a higher-fidelity assessment of binding potential.

Promoting diversity through algorithmic incentives

A central challenge in protein design is maintaining sufficient diversity to explore multiple solutions rather than converging to a single design. Our framework incorporates several mechanisms to promote sequence diversity:

1. **Diversity bonuses in the reward function**: Sequences that differ substantially from previously generated designs receive higher rewards.
2. **Similarity filtering**: Generated sequences are filtered to ensure sufficient difference from existing designs in the elite set.
3. **Adaptive temperature**: Early in training, higher sampling temperatures encourage exploration, gradually decreasing to favor exploitation later.
4. **Reference sequence decay**: Reliance on reference sequences decreases over time, encouraging the discovery of novel solutions.

These diversity-promoting mechanisms are crucial for exploring the protein binding landscape effectively and discovering multiple viable solutions rather than converging to a single local optimum.

Closed-loop training with reality calibration

The complete training algorithm integrates these components into a reinforcement learning loop:

1. The actor generates candidate protein sequences.
2. The critic evaluates these sequences and assigns rewards.
3. Periodically, selected sequences undergo more rigorous evaluation using structural prediction.
4. The critic is recalibrated based on these higher-fidelity assessments to reduce the reality gap.

5. The actor is updated to increase the probability of generating high-scoring sequences.

To validate our framework, we applied it to the design of protein binders for Human Serum Albumin (HSA), an abundant blood plasma protein. Binding to HSA can increase serum half life of many therapeutics, lowering the frequency of dosages.

Repeated runs of this algorithm yielded multiple sequences that achieved real iPTM scores above 0.75, with the best so far reaching 0.76. Furthermore, the sequences generated were diverse from one another and the training set at the sequence level and comprised various lengths ranging from 50–80 amino acids.

Additionally, as expected, the periodic retraining of the critic improved performance. Initially, the predicted iPTM scores significantly overestimated the real scores (MAE $\sim$0.2). As the training progressed, this gap narrowed to $\sim$0.17.

While our implementation used computationally-derived iPTM scores as a proxy for experimental data, the framework serves as a proof-of-concept for how experimental binding data could be integrated into computational protein design pipelines. The same architecture could incorporate wet-lab binding assays such as those developed in the Protein CREATE platform, providing a pathway to bridge the gap between computational prediction and experimental validation. In later chapters, we will see more examples of experimental data that can be integrated into this framework.

Chapter 3

PRACTICAL APPLICATION OF PROTEIN LANGUAGE MODELING: DESIGN OF NOVEL DEHALOGENASES USING PROTEIN LARGE LANGUAGE MODELS

3.1 Limited Data Makes Defluorinase Design Difficult

The PFAS Challenge

Per- and polyfluoroalkyl substances (PFAS) represent one of the most significant environmental challenges of the 21st century. These man-made chemicals have been manufactured since the 1940s and are characterized by their carbon-fluorine bonds, among the strongest covalent bonds in organic chemistry (Evich et al., 2022). This extraordinary bond strength provides PFAS with remarkable resistance to heat, water, and oil, making them valuable for numerous industrial and consumer applications ranging from non-stick cookware to firefighting foams, food packaging, and water-repellent textiles (Kwiatkowski et al., 2020).

However, the same chemical stability that makes PFAS useful also renders them persistent in the environment, earning them the unfortunate moniker "forever chemicals". They resist natural degradation processes, with estimated environmental half-lives ranging from decades to millennia (J. W. Washington et al., 2020). This persistence becomes particularly concerning when considering their bioaccumulative properties and documented toxicity (Agency for Toxic Substances and Disease Registry, 2024). Exposure to PFAS has been linked to serious health effects including increased cholesterol, hypertension, immune suppression, and various cancers, notably testicular and kidney cancers (Birnbaum et al., 2023). Recent research suggests that PFAS accumulate within cell membranes, altering critical properties such as membrane fluidity and plasticity, which may explain their diverse negative health outcomes.

The environmental ubiquity and health risks of PFAS have prompted regulatory action. In 2024, the U.S. Environmental Protection Agency (EPA) established stringent limits for two legacy PFAS compounds—perfluorooctanoic acid (PFOA) and perfluorooctane sulfonate (PFOS)—mandating their concentration in drinking water to be less than 4 parts per trillion (ppt), the minimum detectable limit. This regulation acknowledges the significant health risks posed by even trace amounts of

these compounds (U.S. Environmental Protection Agency, 2024).

Current Approaches to PFAS Remediation and Their Limitations

Traditional approaches to PFAS remediation have significant limitations (Meegoda et al., 2022). Physical methods like activated carbon adsorption and ion exchange can transfer PFAS from one medium to another but do not destroy the compounds. Chemical treatments such as advanced oxidation processes, electrochemical oxidation, and thermal decomposition can break down PFAS but are typically energy-intensive, costly, and may generate toxic byproducts, including shorter-chain PFAS that remain harmful.

Biological remediation offers a potentially more sustainable solution, but natural PFAS-degrading enzymes are extremely rare (Wackett, 2021). The scarcity of natural PFAS-degrading enzymes has a clear biophysical rationale: fluoride, a byproduct of PFAS degradation, is toxic to bacteria at relatively low concentrations, with inhibitory effects on essential enzymes observed at concentrations below 50 mM. This toxicity creates a paradoxical situation—a microorganism capable of degrading PFAS would need to evolve not only a dehalogenase enzyme but also a fluoride ion exporter to mitigate toxicity from the degradation byproduct. This dual requirement imposes selective pressure against the evolution of fast or efficient PFAS-degrading enzymes.

A6RdhA: A Rare Natural PFAS-Degrading Enzyme

In recent years, a significant breakthrough occurred with the discovery of a corrinoid iron-sulfur reductive dehalogenase, A6RdhA, from Acidimicrobium sp. Strain A6. This soil-dwelling microbe demonstrated the ability to partially defluorinate PFOA and PFOS when incubated with these substrates (S. Huang and Jaffé, 2019). While in vitro enzymatic activity has not been conclusively shown with isolated A6RdhA, knockout studies have demonstrated that when the gene coding for A6RdhA is deleted, Acidimicrobium sp. Strain A6 loses its ability to degrade PFOA. This finding strongly suggests that A6RdhA is the key catalyst in PFAS degradation (Jaffé et al., 2024; S. Huang, Fernández, and Summers, 2021).

As shown in Figure 3.1, the A6RdhA enzyme from Acidimicrobium sp. A6 plays a critical role in PFAS degradation. Panel A shows the predicted structure of the partially sequenced A6RdhA, highlighting its complex fold with alpha-helical regions (teal) and beta-sheets (magenta). Panel B demonstrates the ability of Acidimicrobium sp. A6 to degrade PFOA over time, converting it into shorter-chain PFAS

compounds. This conversion is evident from the decreasing concentration of PFOA and the increasing presence of shorter-chain derivatives like H-PFOA (partially defluorinated PFOA) and PFHpA (perfluoroheptanoic acid). Panel C provides compelling evidence for A6RdhA's essential role in this process, showing that wild-type strains release fluoride (a product of defluorination) while A6RdhA knockout strains show no fluoride release.

The discovery of A6RdhA represents a crucial starting point for enzymatic PFAS degradation. However, challenges remain. Only a partial sequence of A6RdhA is available, the enzyme has not been extensively characterized biochemically, and its activity, even in its native context, results in only partial defluorination of PFAS compounds. These limitations highlight the need for improved dehalogenase enzymes specifically designed for PFAS degradation.

The Challenge of Limited Training Data for Machine Learning

The scarcity of natural PFAS-degrading enzymes presents a significant challenge for computational protein design, particularly for approaches that rely on machine learning. Most successful applications of machine learning in protein design have benefited from large datasets of natural proteins with the desired function. In contrast, for PFAS degradation, we face what can be described as a "low N problem"—having extremely limited positive examples of enzymes that perform the desired catalysis.

This data limitation is particularly problematic for protein language models, which typically require extensive training data to learn the complex patterns and relationships that define protein structure and function. With only a single partial sequence of a confirmed PFAS-degrading enzyme, traditional machine learning approaches would seem inadequate for designing novel dehalogenases.

The challenge, therefore, is to develop strategies that can overcome this data limitation, enabling the design of novel dehalogenases despite the paucity of natural examples. As we will describe in the following sections, we addressed this challenge through an integrated approach that leverages structural similarity searches to expand our dataset, followed by focused fine-tuning of protein language models to learn the essential features of dehalogenases capable of PFAS degradation. I mentored the 2024 IEA iGEM team and we worked on this problem together, exploring computational approaches to overcome the limited training data available for PFAS-degrading enzymes while developing practical solutions for environmental

remediation.

3.2 Collection of Dehalogenases from Nature Using Efficient Database Search

Structure-Based Search Strategy Using FoldSeek

Given the limitations of having only a single partial sequence of a confirmed PFAS-degrading enzyme (A6RdhA), we developed a computational approach to identify additional candidate dehalogenases based on structural similarity rather than sequence homology. Our hypothesis was that enzymes with structures similar to A6RdhA would likely share functional properties, potentially including PFAS degradation capabilities, even if their sequences diverged significantly.

To implement this approach, we utilized FoldSeek, a powerful tool developed by van Kempen et al. (2024) that enables fast and accurate protein structure searching. Unlike traditional sequence-based search tools like BLAST, FoldSeek identifies proteins with similar three-dimensional structures, which is particularly valuable for finding functionally related proteins that may have low sequence identity (Kempen et al., 2024).

The workflow began with the predicted structure of the partial A6RdhA sequence. Despite being incomplete, this structure contained the core catalytic machinery and provided sufficient information for structural comparisons. We used AlphaFold3 (Abramson et al., 2024) to predict this structure, which was then used as the query for FoldSeek searches against comprehensive protein structure databases.

Identification and Analysis of Structurally Similar Candidates

The FoldSeek search yielded 68 diverse sequences with high structural similarity to A6RdhA. These candidates represented a significant expansion of our initial dataset from a single partial sequence to a collection of structurally related proteins. Importantly, while these proteins shared structural features with A6RdhA, they exhibited considerable sequence diversity, with many sharing less than 60% sequence identity with each other.

As shown in Figure 3.2A, the FoldSeek search identified several promising candidates with high structural similarity to A6RdhA despite low sequence identity. The figure highlights three representative candidates (ADA7X4IQ74, O68252, and AOA2E5N5L8) with TM-scores ranging from 0.965 to 1.06, indicating strong structural alignment. The heatmap on the right illustrates the sequence diversity among all selected candidates, with many sharing less than 40% sequence identity (dark

purple regions). This diversity is valuable for training protein language models, as it provides examples of how different sequences can adopt similar structures to perform related functions.

Most of the identified proteins were annotated as reductive dehalogenases in the UniProt database, though many had relatively low annotation scores, indicating limited characterization. This classification aligns with the function of A6RdhA and supports the hypothesis that these proteins might share catalytic capabilities, potentially including PFAS degradation.

The identification of these 68 structural homologs provided several advantages:

1. **Expanded training data:** The expanded dataset enabled more effective training of machine learning models for dehalogenase design.
2. **Structural insights:** Analysis of these structures revealed conserved features potentially important for dehalogenation activity, including cofactor binding sites and catalytic residues.
3. **Sequence diversity:** The sequence diversity among structural homologs highlighted regions of the protein that tolerate variation, informing our design strategy.

Figure 3.2B-C shows the successful expression of selected dehalogenase candidates. Panel B displays the SDS-PAGE gel for T7RdhA, a candidate dehalogenase, with a band at approximately 50 kDa matching its predicted molecular weight of 49.794 kDa. Panel C shows the expression of multiple dehalogenase candidates, demonstrating the successful implementation of our protein production pipeline.

Sequence and Structural Analysis of Dehalogenase Collection

To better understand the characteristics of our expanded dehalogenase dataset, we performed comprehensive sequence and structural analyses. Multiple sequence alignment revealed several highly conserved regions, particularly around the predicted active site and cofactor binding regions. These conserved elements likely represent essential components of the catalytic machinery.

Structural superposition of the AlphaFold-predicted models showed remarkable conservation of the overall fold despite sequence variations. The core architecture was preserved across all candidates. However, notable variations were observed

in loop regions and surface features, suggesting these areas might be amenable to optimization without disrupting catalytic function.

Of particular importance was the conservation of the corrinoid cofactor binding site, characterized by a specific arrangement of coordinating residues. The corrinoid cofactor is believed to be essential for the reductive dehalogenation mechanism, participating in electron transfer and potentially facilitating C-F bond cleavage.

Additionally, we identified a conserved iron-sulfur cluster binding motif, characterized by a specific pattern of cysteine residues. Iron-sulfur clusters often serve as electron transfer centers in redox enzymes and might play a crucial role in the reductive dehalogenation mechanism of A6RdhA and its homologs.

This detailed analysis provided valuable insights for the subsequent design of novel dehalogenases, highlighting regions that must be preserved to maintain catalytic function and identifying areas where variation might be introduced to enhance PFAS specificity or activity.

3.3 Rational Design of Chimeric Dehalogenases

Before employing protein language models for the design of completely novel sequences, we explored a rational design approach by creating chimeric enzymes that combine features from different known dehalogenases. This approach offers a bridge between the limited information available from natural PFAS-degrading enzymes and the generation of completely novel sequences.

As shown in Figure 3.3, we designed a chimeric enzyme (A6T7) by combining the core of A6RdhA with the C-terminal fragment of T7RdhA, another reductive dehalogenase identified in our structure-based search. Panel A illustrates the design concept, showing how the majority of the protein structure comes from A6RdhA (blue) with a specific C-terminal region from T7RdhA (magenta). The chimeric approach allowed us to test hypotheses about the role of different protein regions in substrate specificity and catalytic activity.

The A6T7 chimera was successfully expressed and purified, as confirmed by SDS-PAGE analysis (Figure 3.3B), with a band at the expected molecular weight of 46.865 kDa. This result demonstrated the feasibility of creating stable chimeric dehalogenases by combining features from different structural homologs.

3.4 Protein Language Model-Based Design of Novel Dehalogenases

Fine-Tuning ESM-2 on Dehalogenase Sequences

With our expanded dataset of structurally similar dehalogenases, we proceeded to develop a specialized protein language model for dehalogenase design. Our approach involved fine-tuning the ESM-2 model, which had been pre-trained on millions of diverse protein sequences, on our collected dehalogenase sequences.

The fine-tuning process aimed to shift the model's learned distribution towards sequences with dehalogenase-like features, enabling more targeted generation of novel dehalogenase candidates. We employed masked language modeling to fine tune the model by randomly masking a fraction of amino acids of each sequence in the data set according to the betalinear30 function (Lin, Akin, Rao, Hie, Zhu, Lu, Smetanin, et al., 2023). Each sequence was individually masked multiple times, increasing the total training data the fine-tuned model could see.

The fine-tuned model demonstrated the ability to generate sequences that shared key characteristics with the training dehalogenases while exhibiting novel variations. Analysis of the generated sequences revealed preservation of critical motifs, including cofactor binding sites and putative catalytic residues, while showing diversity in other regions.

Figure 3.4A illustrates our iterative unmasking approach for generating novel dehalogenase variants. Starting with a fully or partially masked sequence, we progressively unmasked positions in a random order, with each amino acid prediction informed by the surrounding sequence context. This approach leverages the bidirectional nature of the ESM-2 model, allowing it to use information from both sides of a masked position when making predictions.

Design and Selection of Candidate Dehalogenases

Using our fine-tuned protein language model, we generated a diverse set of candidate dehalogenase sequences. The generation process was guided by incorporating structural constraints derived from A6RdhA and its homologs, ensuring that the generated sequences would be compatible with the required three-dimensional fold and catalytic machinery.

We implemented a multi-step filtering pipeline to select the most promising candidates for experimental testing:

1. **Structure prediction:** AlphaFold2 was used to predict the structures of can-

didate sequences, and those with high confidence predictions that maintained the core dehalogenase fold were retained.

2. **Sequence evaluation:** Candidate structures were aligned to a known dehalogenase (PDB ID: 4UQU) and a putative defluorinase (T7RdhA) to check for the presence of essential motifs, including cofactor binding sites and catalytic residues.

3. **Active site analysis:** The predicted structures were analyzed for proper formation of the active site, with particular attention to the arrangement of catalytic residues and cofactor binding pockets.

The team identified three candidate dehalogenases for experimental characterization. These candidates showed considerable sequence diversity, with pairwise identities ranging from 40% to 70% compared to A6RdhA, while maintaining key structural and functional features.

3.5 Experimental Validation of Designed Dehalogenases

Expression and Purification Strategy

The experimental validation of our designed dehalogenases presented significant challenges, particularly given the complex cofactor requirements and potential oxygen sensitivity of these enzymes. Based on knowledge of related reductive dehalogenases, we anticipated that our designed enzymes would require a corrinoid cofactor (vitamin B12 derivative) and iron-sulfur clusters for activity.

We developed a comprehensive expression and purification strategy to address these challenges:

1. **Cell free expression:** Candidate genes were synthesized with codon optimization for expression in E. coli cell free lysate. A subsequent chapter will analyze expression in transcription translation (TX-TL) mix in more detail.

2. **Denaturation:** Post-expression, variants were formed as inclusion bodies. Prior to refolding with the relevant cofactors, proteins were denatured with 8 M urea.

3. **Anaerobic conditions:** Refolding steps were performed under anaerobic conditions to prevent oxidation of oxygen-sensitive cofactors, particularly iron-sulfur clusters.

4. **Cofactor supplementation:** Iron chloride, sodium sulfide, and cyanocobalamin were supplemented to ensure cofactor availability during refolding.

Structural and Biochemical Characterization of Novel Dehalogenases

Our AI-designed enzymes were characterized using a combination of structural prediction and biochemical analyses to confirm proper folding and cofactor incorporation. As shown in Figure 3.4B, we generated several novel sequences (designated AI-1, AI-4, and AI-10) with different levels of sequence similarity to known dehalogenases. These sequences were analyzed for their identity to closest BLAST matches, revealing that AI-1 had 41.71% identity to an unclassified Dehalobacter, AI-4 had 42.86% identity to Dehalobacter sp. 4CP, and AI-10 had 59.18% identity to an Acidimicrobiaceae bacterium.

AlphaFold3 structure predictions (Figure 3.4C) showed high confidence models for both T7RdhA (iPTM = 0.93) and AI-10 (iPTM = 0.77), suggesting proper folding of the designed proteins. Detailed analysis of the predicted structures revealed the expected cofactor binding sites for both norpseudo-cobalamin (Figure 3.4D) and iron-sulfur clusters (Figure 3.4E). The model also predicted a potential PFOA binding site (Figure 3.4F) with key residues (Y93, Y229, and W243 in T7RdhA) that may interact with the substrate.

UV-Vis Spectroscopy Confirms Successful Cofactor Incorporation

One of the critical aspects of validating our designed dehalogenases was confirming the successful incorporation of the essential cofactors. We used UV-Vis spectroscopy to assess the presence and integrity of both the iron-sulfur clusters and the corrinoid cofactor in our reconstituted enzymes.

As shown in Figure 3.4H, the UV-Vis spectrum of the AI-1 designed dehalogenase (orange line) exhibits distinct features that are consistent with successful cofactor incorporation. The spectrum shows a characteristic peak at approximately 390 nm, which is indicative of iron-sulfur clusters. Additionally, a broad absorption band between 450-550 nm is consistent with the presence of a corrinoid cofactor (vitamin B12 derivative). In contrast, the buffer control (blue line) shows no significant absorbance in these regions.

This spectroscopic evidence, combined with the successful expression of AI-1 confirmed by SDS-PAGE (Figure 3.4G), provides strong support for the proper folding and cofactor incorporation in our designed dehalogenase. The distinct spectral

features match those reported for other functional reductive dehalogenases in the literature, suggesting that our designed enzyme has successfully incorporated the cofactors required for catalytic activity.

The successful expression and cofactor incorporation of designed dehalogenases represents a significant achievement in our effort to create novel enzymes for PFAS degradation. These results demonstrate that protein language models can be effectively used to design complex enzymes with specific cofactor requirements, even when starting with limited examples of the target function. Future work will focus on assessing the catalytic activity of these enzymes against various PFAS compounds and further optimizing their performance through additional rounds of design and testing.

3.6 Substrate Co-folding

After the iGEM competition, the Boltz-1 model was published and made freely available to use (Wohlwend et al., 2024), which made it possible for us to computationally assess possible binding of our variants to PFOA. Under my guidance, one student, Maya, investigated whether or not putative defluorinase candidates could be screened for PFOA binding and possible catalytic activity using this new tool.

Maya first picked a positive and negative control within the dehalogenase family to co-fold with PFOA. The Tetrachloroethene Dehalogenase from *Geobacter* (Nakamura et al., 2018) served as a negative control while T7RdhA (Guo et al., 2023) was used as a positive control. Surprisingly, there was a noticeable difference in the ligand iPTM between the positive and negative controls. PFOA was predicted to bind in the active site of the positive control, T7RdhA, as expected.

Encouraged by these promising initial results, Maya co-folded each of the 3 language model designed dehalogenases with PFOA. While one had a lower iPTM value than the negative control and another had an intermediate value between the negative and positive controls, the third had an iPTM score higher than the positive control with PFOA docked in the enzyme active site. Future work will focus on expressing and refolding this enzyme with iron-sulfur clusters and cyanocobalamin cofactor to investigate catalytic activity.

3.7 Conclusions

In this chapter, we have demonstrated a novel approach to designing enzymes for a challenging environmental problem: the degradation of persistent PFAS pollutants.

By combining structure-based database searches with protein language model fine-tuning, we were able to overcome the "low N problem" that typically hinders machine learning approaches to protein design when few examples of the target function are available.

Our key achievements include:

1. Expanding a dataset from a single partial dehalogenase sequence to 68 structurally similar candidates using the FoldSeek structure search tool.
2. Successfully creating and expressing a chimeric enzyme (A6T7) that combines features from different dehalogenases.
3. Fine-tuning ESM-2 on dehalogenase sequences to create a specialized language model capable of generating novel dehalogenase candidates.
4. Designing multiple novel dehalogenase sequences with predicted structural features suitable for PFAS degradation.
5. Expressing and purifying designed dehalogenases with successful incorporation of both iron-sulfur clusters and corrinoid cofactors, as confirmed by UV-Vis spectroscopy.

This work illustrates how protein language models can be adapted to design complex enzymes even with limited training data. The approach demonstrates the value of incorporating structural information to guide the design process, particularly when sequence information alone is insufficient.

The successful design and expression of novel dehalogenases capable of incorporating complex cofactors represents an important step toward addressing the persistent challenge of PFAS contamination. By combining computational design with experimental validation, we have demonstrated an approach that could be applied to other challenging environmental problems where natural enzymatic solutions are limited or non-existent.

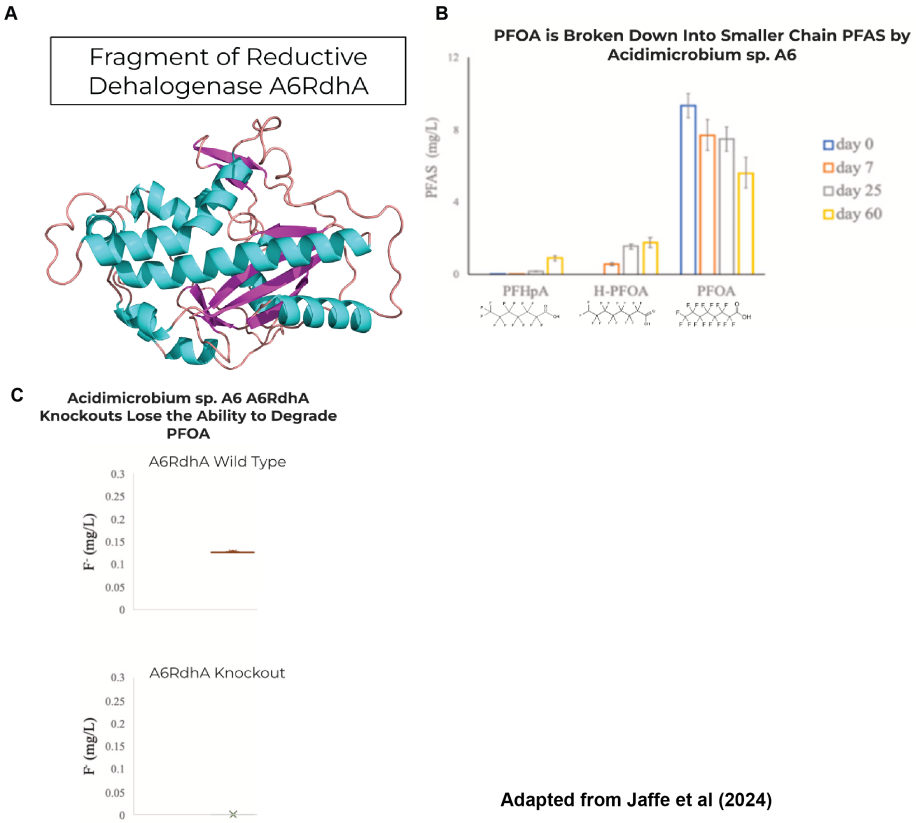


Figure 3.1: A) Predicted structure of the partially sequenced reductive dehalogenase A6RdhA from Acidimicrobium sp. A6. B) Degradation of PFOA by Acidimicrobium sp. A6 over time, showing the conversion of PFOA into shorter-chain PFAS compounds. C) Comparison of fluoride release between wild-type Acidimicrobium sp. A6 and an A6RdhA knockout strain, confirming the essential role of A6RdhA in PFAS degradation. Adapted from Jaffe et al. (2024).

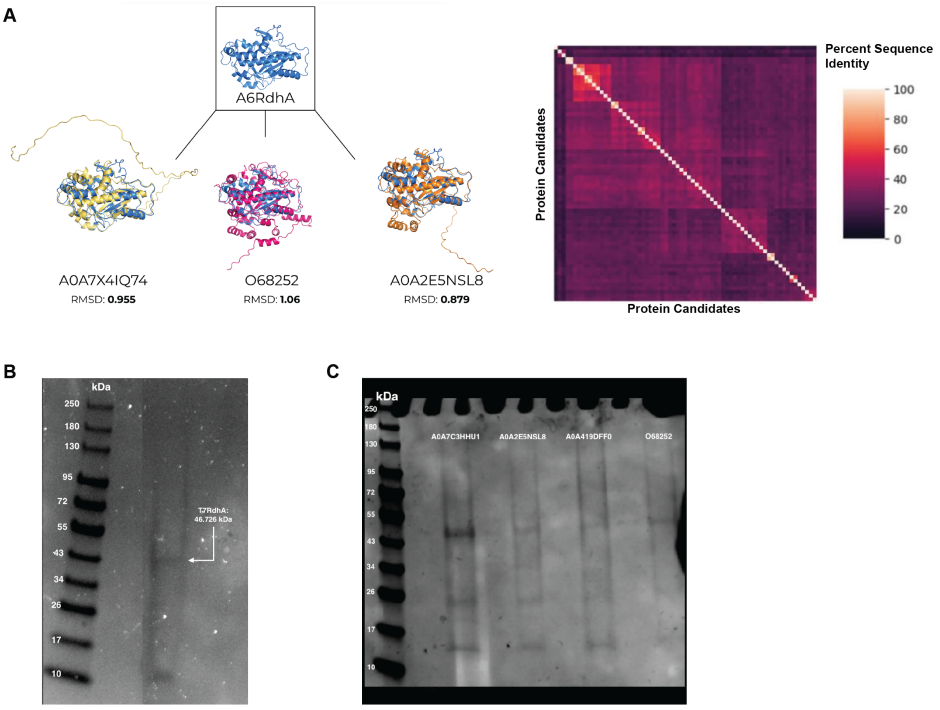


Figure 3.2: A) Identification of structurally similar dehalogenase candidates using FoldSeek. Three representative candidates are shown with their TM-scores, alongside a heatmap showing sequence identity among all selected candidates. High TM-scores (>0.5) indicate structural similarity despite low sequence identity. B) SDS-PAGE gel showing expression of T7RdhA (49.794 kDa). C) SDS-PAGE gel showing expression of multiple dehalogenase candidates.

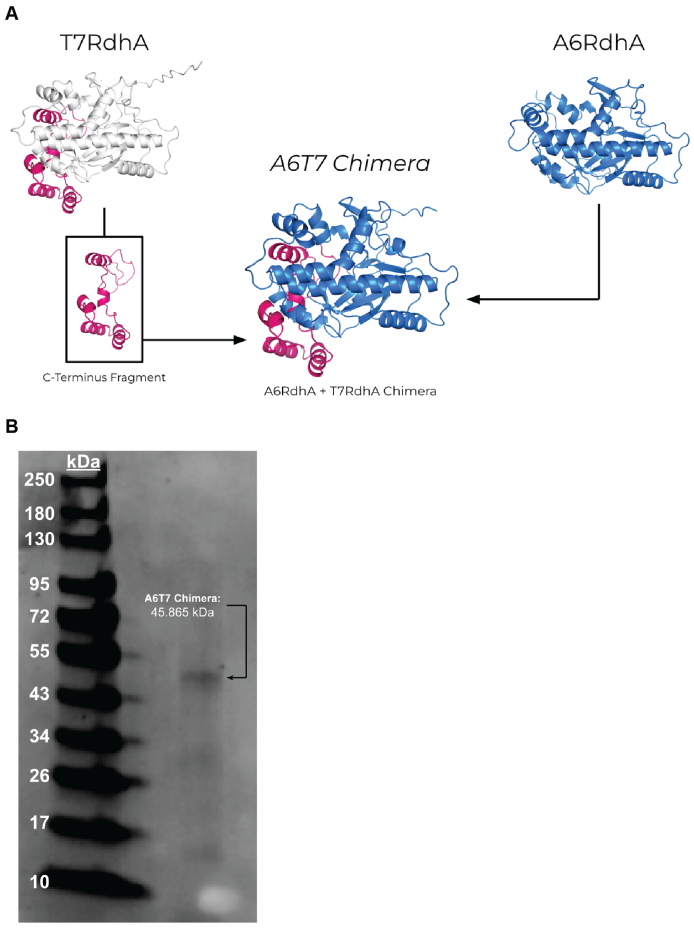


Figure 3.3: A) Design of the A6T7 chimeric enzyme by combining the core structure of A6RdhA (blue) with the C-terminal fragment of T7RdhA (magenta). B) SDS-PAGE gel confirming successful expression of the A6T7 chimera at the expected molecular weight of 46.865 kDa.

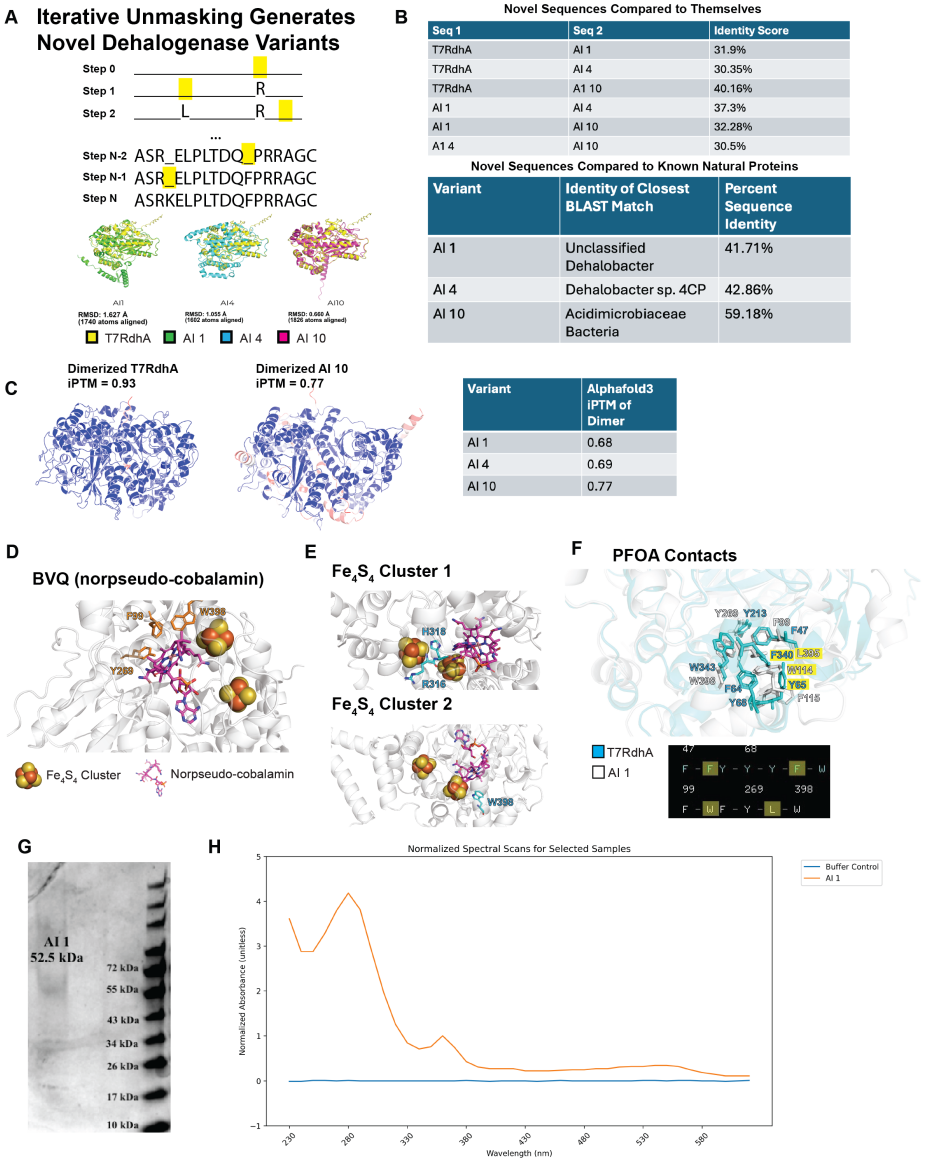


Figure 3.4: A) Iterative unmasking approach for generating novel dehalogenase variants. The process begins with a masked sequence and progressively reveals amino acids, guided by the fine-tuned protein language model. B) Comparison of novel dehalogenase sequences both among themselves and with natural proteins. C) AlphaFold3 structure predictions of designed dimerized dehalogenases T7RdhA (iPTM = 0.93) and AI-10 (iPTM = 0.77). D) Active site representation showing norpseudo-cobalamin and iron-sulfur clusters essential for dehalogenase activity. E) Detailed view of the iron-sulfur cluster coordination. F) Visualization of the PFOA binding site in designed dehalogenases. G) SDS-PAGE confirmation of AI-1 expression. H) UV-Vis spectral analysis showing characteristic absorption patterns of a successfully reconstituted iron-sulfur cluster and cobalamin cofactor in AI-1 compared to buffer control.

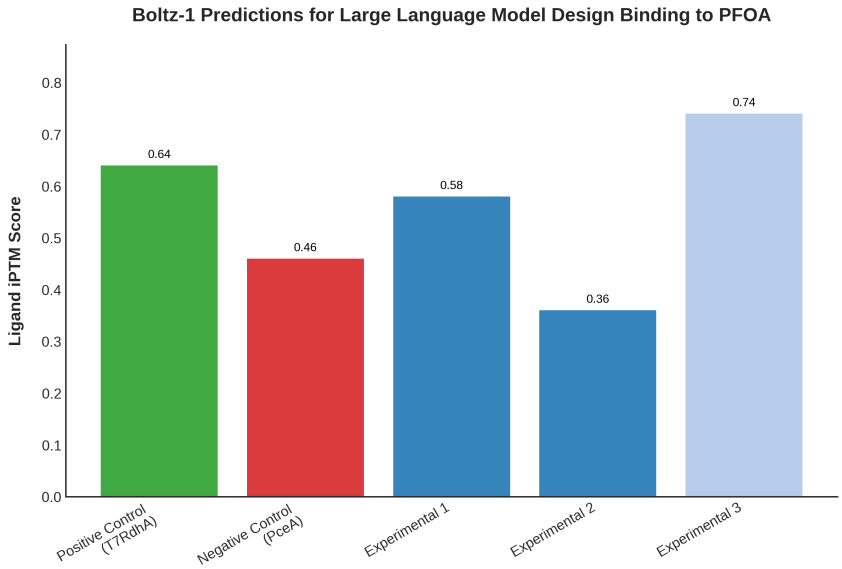


Figure 3.5: Boltz-1 predicted ligand iPTM scores between candidate enzymes and perfluorooctanoic acid (PFOA). A likely full-length defluorinase, T7RdhA with both high sequence and structural similarity to A6RdhA, while well-characterized Tetrachloroethene Dehalogenase from *Geobacter* was used as a negative control. The three language model derived sequences served as experimental conditions

Chapter 4

DEVELOPMENT OF NOVEL IL7RA BINDERS USING PROTEIN CREATE

4.1 Introduction to IL7RA as a Therapeutic Target

Interleukin-7 receptor alpha (IL7RA) represents a compelling therapeutic target with significant implications for both immunomodulation and treatment of autoimmune disorders. IL7RA is a cell surface receptor primarily expressed on lymphocytes that plays a critical role in T-cell development, homeostasis, and function (Calore and Petrara, 2023). The receptor's natural ligand, IL-7, initiates signaling by first binding to IL7RA, which then recruits the common gamma chain (γc) to form a heterodimeric complex. This interaction triggers downstream signaling through the JAK/STAT and PI3K/AKT pathways, ultimately promoting lymphocyte proliferation, survival, and differentiation (Figure 4.1).

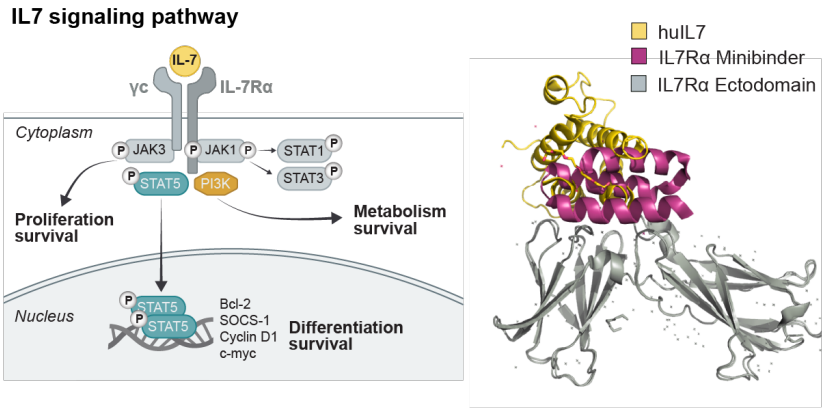


Figure 4.1: *Left:* IL7RA signaling pathway. Upon binding to its natural ligand IL-7, IL7RA recruits the common gamma chain (γc) to form a heterodimeric complex, activating downstream signaling through JAK/STAT and PI3K/AKT pathways. This signaling promotes lymphocyte proliferation, survival, and differentiation. *Right:* IL7RA bound to either human IL-7 or a synthetic designed antagonist

The IL-7/IL7RA signaling axis has dual therapeutic relevance. Enhancement of this pathway has potential applications in immunotherapy and treatment of immunodeficiency disorders, as IL-7 signaling stimulates T-cell proliferation and can enhance immune responses (Barata, Durum, and Seddon, 2019). Conversely, inhibition of IL7RA has emerged as a promising strategy for treating autoimmune conditions,

where pathological T-cell responses contribute to tissue damage. Studies have demonstrated that blocking IL7RA can effectively blunt antigen-specific memory T cell responses and reduce chronic inflammation in primates, suggesting potential efficacy in autoimmune diseases such as multiple sclerosis, rheumatoid arthritis, and inflammatory bowel disease (Belarif et al., 2018).

The therapeutic modulation of IL7RA presents several challenges. Natural IL-7 has limitations as a therapeutic agent, including a short half-life and complex manufacturing requirements. Antibody-based approaches to IL7RA blockade have demonstrated efficacy but face challenges related to production costs, immunogenicity, and delivery. Small protein binders, particularly engineered mini-proteins, offer an attractive alternative with the potential for improved tissue penetration, reduced immunogenicity, and more cost-effective production.

Previous efforts have successfully developed synthetic IL7RA antagonists that competitively inhibit IL-7 binding (Cao et al., 2022). These engineered mini-proteins bind to IL7RA in an orientation distinct from the natural ligand, effectively blocking IL-7 engagement and downstream signaling. However, we have observed these binders exhibit off-target binding to unrelated receptors, which can lead to undesired side effects and reduced therapeutic efficacy.

In this chapter, we describe the development of novel IL7RA binders with significantly reduced sequence homology to previously reported designs using the Protein CREATE platform. Our approach not only generates binders with unique sequences but also addresses the critical challenge of off-target binding, resulting in antagonists with substantially improved specificity profiles. We employ a "context-dependent inverse folding" approach described in Chapter 2 to generate candidate binders, followed by comprehensive screening, biochemical characterization, and functional validation. This work not only identifies specific novel binders with enhanced selectivity but also provides insights into the permissiveness of the IL7RA binding interface and implications for *in vivo* functionality, with implications for future therapeutic design.

4.2 Generation and Screening of IL7RA Binding Pool

Context-Dependent Inverse Folding and Off-target Binding

Previous designs, including the Baker lab IL7RA binder used as our template, have demonstrated strong on-target binding but also exhibit considerable off-target binding to unrelated proteins. This limitation is common in many designed protein

binders, where optimization focuses primarily on affinity for the target rather than specificity (Bennett et al., 2023). By exploring a broader sequence space through context-dependent inverse folding, we aimed to identify variants that not only maintained target engagement but also showed reduced off-target binding.

Using context-dependent inverse folding, we generated 42 candidate binders with high sequence diversity. These candidates were designed to maintain the structural features necessary for IL7RA binding while exploring substantially different sequences, with many sharing less than 60% sequence identity with the parent molecule. I hypothesized that off-target binding for these variants would be less, as context-dependent inverse folding should only preserve the desired binding interface, and these variants are not undergoing repeated rounds of screening, which can select for off-target behavior.

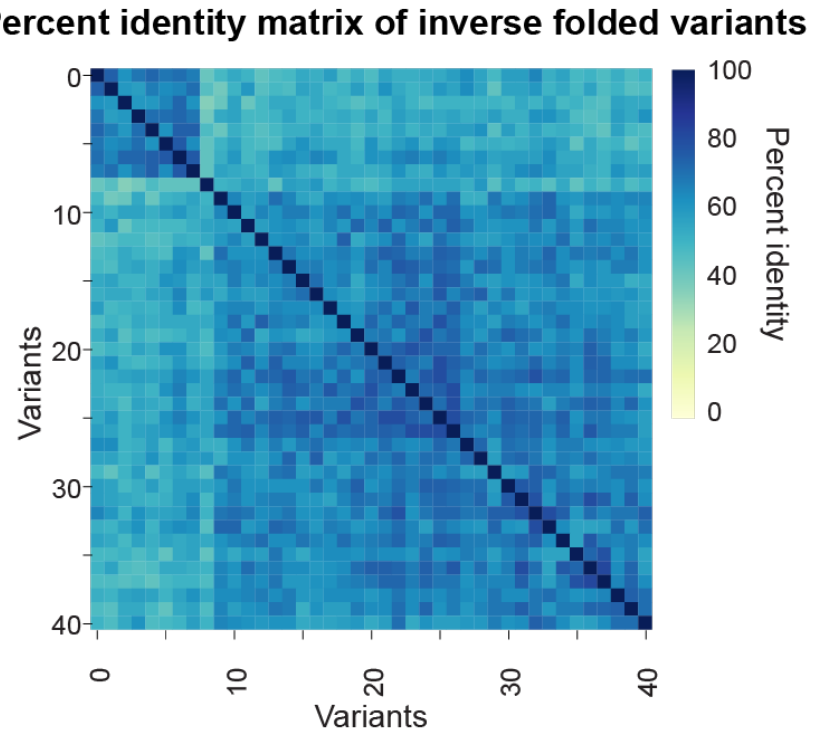


Figure 4.2: Percent Identity Matrix for Context-Dependent Inverse Folded Designs

Computational Pre-screening

Prior to experimental testing, we performed computational pre-screening of the designed variants using AlphaFold2-based metrics (Jumper et al., 2021; Bryant, Pozzati, and Elofsson, 2022). Specifically, we calculated the interface predicted template modeling (iPTM) score, which estimates the quality of the predicted inter-

face between the designed binder and IL7RA.

All designed variants scored highly on this metric (iPTM > 0.8), suggesting potential binding capability. However, as observed in previous studies and confirmed by our subsequent experimental results, iPTM scores alone are imperfect predictors of actual binding, highlighting the importance of experimental validation using platforms like Protein CREATE.

Experimental Screening Using Protein CREATE

To experimentally evaluate the binding of designed variants to IL7RA, we employed the Protein CREATE platform described in Chapter 1. Briefly, the platform uses a phage-based "binding by sequencing" assay to quantify the binding affinity of protein libraries at scale. DNA libraries encoding the designed variants were cloned into T7 bacteriophage backbones, expressed on the phage capsid, and evaluated for binding to immobilized IL7RA.

We screened the 42 designed variants along with a pool of off-target controls and previously characterized IL7RA binders (Cao et al., 2022). Enrichment scores, defined as the ratio of normalized sequence counts post-binding to pre-binding, were calculated for each variant (Figure 4.3). This screening identified multiple variants with significant enrichment, indicating successful binding to IL7RA.

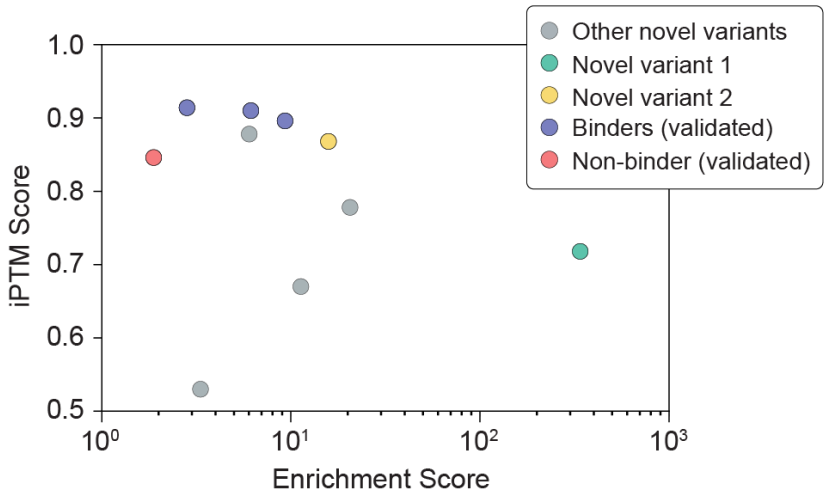


Figure 4.3: Results of Protein CREATE screening of designed IL7RA binders. Enrichment scores are shown for selected designed variants, previously characterized binders, and off-target controls. Several novel designs (highlighted) exhibited significant enrichment, indicating successful binding to IL7RA. No clear correlation was observed between computational iPTM scores and experimental enrichment.

Interestingly, we observed no strong correlation between the computational pre-screening metric (iPTM score) and experimental enrichment. Some variants with high iPTM scores showed modest enrichment, while others with similar iPTM scores demonstrated stronger binding. This observation reinforces the value of high-throughput experimental screening to complement computational predictions.

Based on the screening results, we selected two novel binders with sequence identity less than 60% to the parent sequence for further characterization. These variants, designated Novel Variant 1 (NV1) and Novel Variant 2 (NV2), represented promising candidates with distinct binding profiles.

4.3 Binding Analysis

Permissiveness of the Interface

To better understand the interactions between our novel binders and IL7RA, we performed detailed structural analysis using AlphaFold Multimer predictions (Jumper et al., 2021). As expected, NV1, NV2, and the parent all were predicted to fold into the same structure and bind the same location on the parent receptor. Interestingly, only around half of predicted receptor contacts for both variants were preserved from the parent design.

I investigated overall preservation patterns of the residues at positions predicted to constitute the receptor-binder interface as predicted by RING (Del Conte et al., 2024). Though the fraction of preserved contacts is greater among enriched contacts than among designs overall, the fraction of preserved contacts is overall quite low.

IL7RA is known to have a moderately hydrophobic interface, which led us to hypothesize that more nonpolar contacts will be preserved. We compared amino acid identities at positions identified by RING in variants enriched in the assay to those in the base design pool and found that, indeed, most (3/4) are nonpolar amino acids.

Additionally, CREATE allows discrimination of residues positions where the original amino acid need not be preserved (S45 and K55) from those that maintain the interface, even when the design methodology rarely preserves the amino acid identity in the base position (L21 and R48).

Surface Plasmon Resonance Validation

To quantitatively characterize the binding of selected variants to IL7RA, we employed surface plasmon resonance (SPR). The novel variants and parent molecule

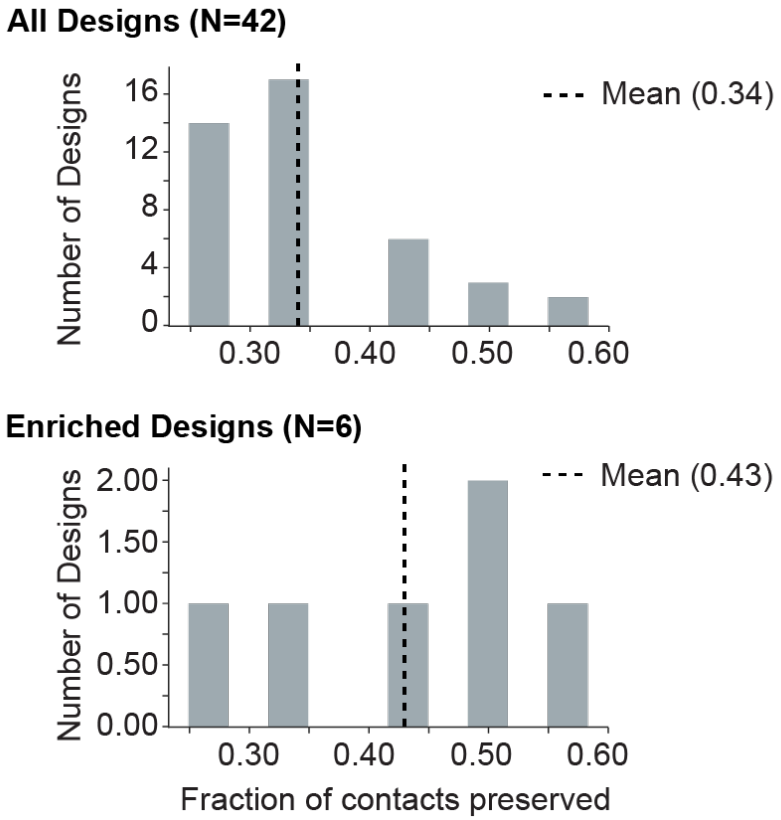


Figure 4.4: Fraction of Preserved Designs Among All Designs (Top) and Enriched Designs (Bottom)

were expressed in *E. coli* cell free lysate, purified to using Ni-NTA affinity chromatography, and tested for binding to immobilized IL7RA.

Due to its weak binding, we failed to extract a K_d for NV2, but obtained values of 3 nM and 88 nM for parent and NV1, respectively.

Luminex Immunodetection Assay

The Luminex immunodetection assay employed a multiplex bead-based approach for analyzing protein-protein interactions. Because of its multiplex nature, multiple targets, including off-targets, can be added easily measured in a single assay.

Bio-Plex magnetic COOH beads were functionalized with streptavidin or avidin, then incubated with biotinylated cytokine receptors (IL7Rα or IL2R$\beta\gamma$) in blocking buffer. Following biotin blocking and washing, the receptor-coated beads were exposed to His-tagged protein binders (1 μM) for 30 minutes. Detection was achieved through sequential 30-minute incubations with mouse anti-His monoclonal

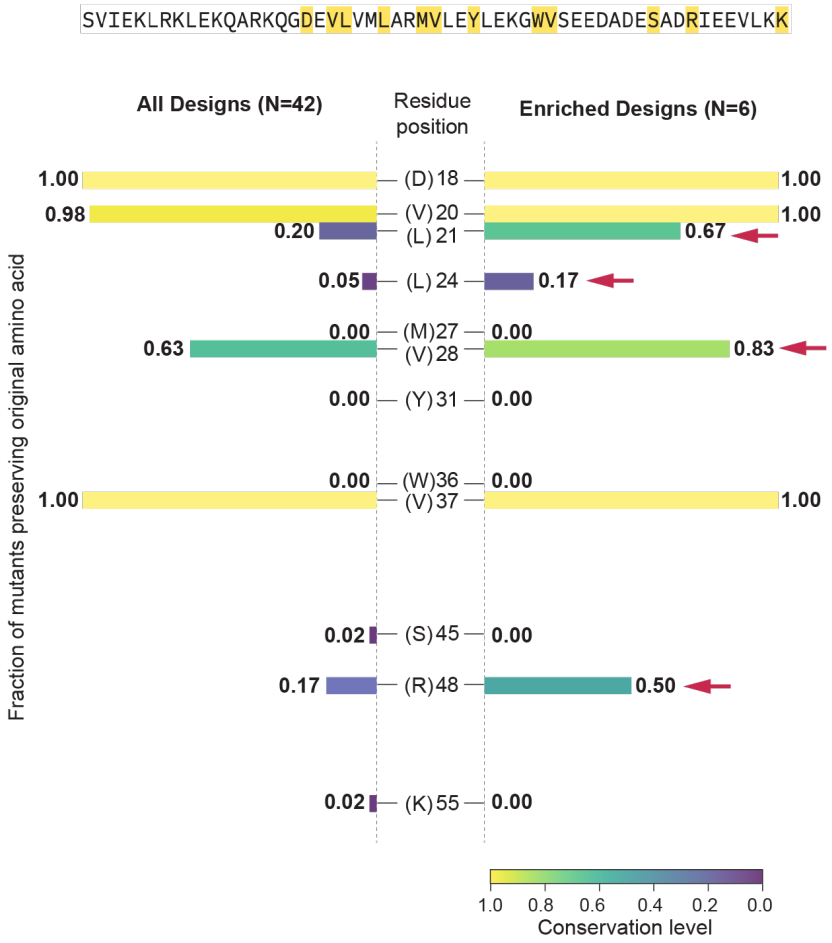
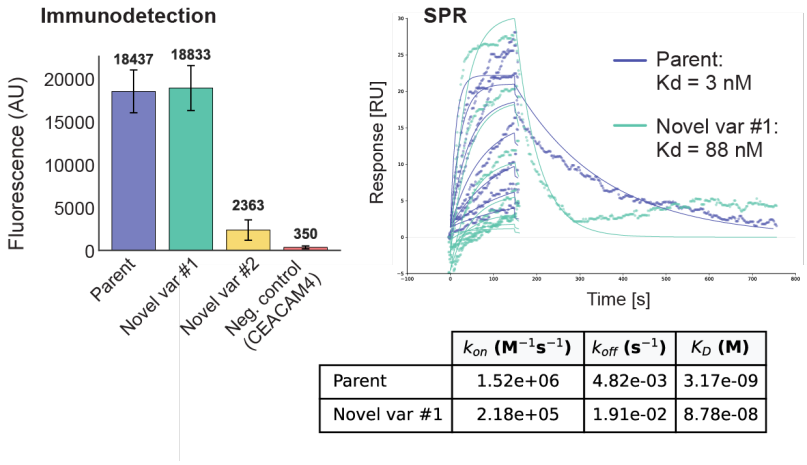


Figure 4.5: Identities of Enriched Variants Suggest Key Contacts for Successful Designs

antibody and phycoerythrin-conjugated anti-mouse IgG secondary antibody. After a final PBS wash, protein-protein interactions were quantified using the Bio-Plex 200 system, enabling efficient screening of multiple receptor-binder combinations.

While both the parent and NV1 showed substantial binding to IL7RA, NV2 (Figure 4.6, left), with its higher iPTM score than NV1, only showed weak binding, highlighting the binary nature of using iPTM as a heuristic.

When compared as a ratio of on target to off target binding, NV1 does almost an order of magnitude better than the parent sequence (Figure 4.7) as hypothesized due to the nature of context-dependent inverse folding design strategy. As we will see in the next section, small differences in biochemical properties can be significant for phenotype in functional assays



	k_{on} ($M^{-1}s^{-1}$)	k_{off} (s^{-1})	K_D (M)
Parent	1.52e+06	4.82e-03	3.17e-09
Novel var #1	2.18e+05	1.91e-02	8.78e-08

Figure 4.6: A dual antibody detection strategy was used to indicate binding of the enriched variant in an orthogonal assay when purified protein is allowed to bind to IL7RA immobilized on Bioplex beads. Additionally, binding for the parent binder and the stronger novel binder was characterized using surface plasmon resonance (SPR).

4.4 Functional Validation

Inhibition of IL-7-Induced Signaling

To more directly evaluate the functional activity of our novel binders, we assessed their ability to inhibit IL-7-induced signaling in a cell-based assay. The HEK-Blue^{TM} IL-7 reporter assay was utilized to assess IL-7 signaling inhibition in a cellular context when stimulated with IL-7 in the presence of increasing concentrations of each binder. These cells are engineered to express secreted embryonic alkaline phosphatase (SEAP) upon IL-7 pathway activation, which is then detected via conversion of a chromogenic substrate.

Both NV1 and the parent showed clear antagonistic activity in this engineered cell line, lining up with expectations from *in vitro* data in the literature (Cao et al., 2022).

in vitro PBMC Activation Assay

However, the story becomes more complex if we expose a population of human peripheral blood mononuclear cells (PBMCs) to the each of the mini-binder proteins. As human immune cells are highly responsive to endotoxin, I produced each variant in a cell-free lysate made of ClearColi®, a genetically modified *E. coli* strain engineered to lack immunogenic endotoxins, making it ideal for recombinant protein production with reduced endotoxin contamination and improved safety for therapeutic applications.

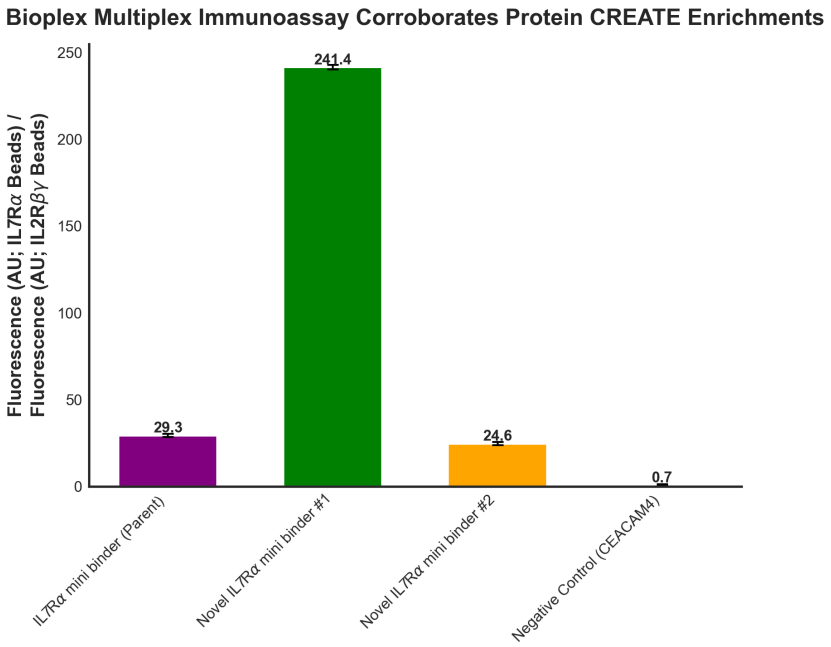


Figure 4.7: Binding to receptor when expressed as a ratio of on to off-target binding as measured in Luminex Immunodetection Assay

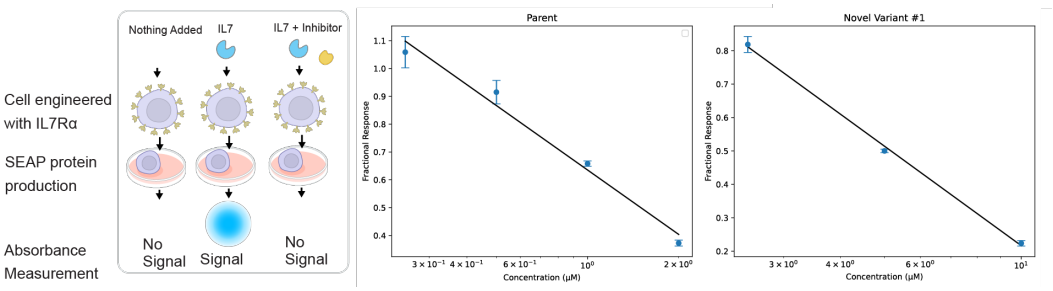


Figure 4.8: Increasing concentrations of each inhibitor are added to an engineered HEK cell line expressing secreted embryonic alkaline phosphatase (SEAP) in response to IL7RA activation and 17 pM IL7 added to the cell media. Cells express secreted embryonic alkaline phosphatase (SEAP) proportional to the degree of receptor activation, which is measured via conversion of a chromogenic substrate. Results were normalized to set the IL-7-stimulated control as 1 and the unstimulated condition as 0, enabling direct comparison of inhibitory potency across test compounds.

Either the parent or NV1 was added to a population of PBMCs for 24 hours with the following other conditions:

1. **Condition 1:** No other ligands added
2. **Condition 2:** IL-7 (200 ng/mL) added

3. **Condition 3:** IL-7 (200 ng/mL) and CD3/CD28 T-cell activator cocktail added

I performed flow cytometry on the fixed cells in collaboration with Yu-Jen Chen, staining for CD25 and CD80 to measure activation of T-cells and macrophages in the population.

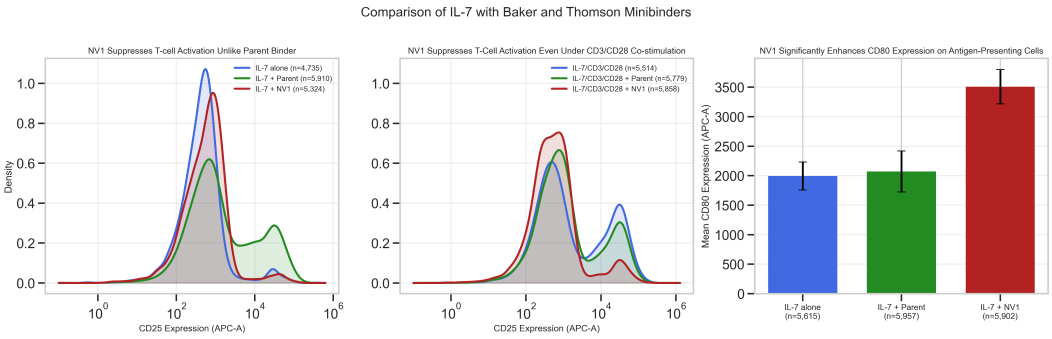


Figure 4.9: Flow cytometry analysis of immune cell activation markers in human PBMCs treated with IL7RA binders. **(A)** CD25 expression on T cells shows that the parent binder induces a high-expressing population that is absent with NV1 treatment. **(B)** Under CD3/CD28 co-stimulation conditions, NV1 selectively suppresses the high CD25-expressing population while preserving the main activation peak, demonstrating targeted immunomodulation. **(C)** CD80 expression on antigen-presenting cells reveals that NV1 significantly enhances co-stimulatory molecule expression compared to both IL-7 alone and IL-7 with parent binder, suggesting a unique dual immunomodulatory profile.

CD25 and CD80 expression profiles revealed distinct immunomodulatory effects of the parent binder and NV1. Flow cytometry analysis demonstrated that while neither IL7 alone nor IL7 with NV1 activated PBMCs as expected, the parent binder exhibited some immunostimulatory effect on a sub-population of PBMCs, possibly due to receptor clustering (Figure 4.9A).

This differential pattern persisted under stronger stimulation conditions. When PBMCs were treated with IL-7 plus CD3/CD28 co-stimulation, the parent binder had minimal impact on the resulting bimodal activation profile (Figure 4.9B). Remarkably, NV1 maintained its selective suppression of the high CD25-expressing population while preserving the main activation peak, demonstrating its ability to modulate specific aspects of T-cell activation even in the presence of strong co-stimulatory signals.

Most unexpectedly, NV1 significantly enhanced CD80 expression on antigen-presenting cells compared to both IL-7 alone and IL-7 with parent binder (Figure 4.9C). This

approximately 75% increase in CD80, a critical co-stimulatory molecule, suggests that NV1 possesses dual immunomodulatory properties: suppressing pathogenic T-cell responses while potentially enhancing antigen presentation and regulatory T-cell induction.

This distinctive immunomodulatory signature—selective T-cell suppression coupled with enhanced antigen-presenting cell activation—distinguishes NV1 from the parent design and may be directly attributable to its improved specificity for IL7RA over off-target receptors (as demonstrated in the Luminex assay). These findings illustrate how subtle differences in binding specificity can translate to profound alterations in functional activity within complex cellular environments, highlighting the potential of our context-dependent inverse folding approach to generate therapeutic candidates with unique and potentially advantageous biological properties.

4.5 Conclusion

Taken together, these results highlight the non-intuitive nature of protein-protein interactions and the value of combining computational design with comprehensive experimental validation. My use of context-dependent inverse folding specifically allowed me to make comparisons between structurally similar, sequentially dissimilar sequences, suggesting that the latter can play an important role in design, even when controlling for the former.

Given the unexpected results, the importance of high-throughput experimental test at multiple scales is increased even greater. If one protein design does not have the expected or desired behavior, another out of a set of structural mimics may. In short, this work not only delivers a promising new IL7RA antagonist with enhanced specificity and unique immunomodulatory properties but also opens the door and motivates alternative approaches for expanding the functional diversity of designed protein therapeutics.

Chapter 5

DESIGN AND TESTING OF INSULIN MIMICS WITH LOW SEQUENCE HOMOLOGY TO WILD-TYPE INSULIN

5.1 Introduction to Insulin Receptor Binding

Insulin and insulin-like peptides represent one of the most evolutionarily conserved protein families across eukaryotes, with evidence of ancestral insulin-like peptides dating back to the earliest unicellular eukaryotes (Weiss, 2009). Human insulin's structure consists of two peptide chains (A and B) connected by three disulfide bonds—two inter-chain and one intra-chain (Figure 5.1). This distinctive structural arrangement has remained remarkably preserved across species, suggesting fundamental importance to the molecule's function (Viola et al., 2023).

The binding interface between insulin and its receptor (IR) has been extensively characterized, with specific residues implicated in receptor recognition highly conserved across vertebrate species. This conservation initially suggested that these precise amino acid contacts might be essential for receptor binding (Menting et al., 2013). However, studies of insulin-like peptides from diverse organisms, such as *Drosophila melanogaster* insulin-like peptide 5 (DILP5), have challenged this assumption. Despite DILP5 retaining only two out of eight receptor contacts found in human insulin, it binds to human insulin receptor with impressive affinity (KD of 60 nM) (Viola et al., 2023). This remarkable finding suggests substantial plasticity in the insulin-receptor binding interface and opens exciting possibilities for designing novel insulin mimetics with low sequence homology to wild-type insulin.

The design of insulin mimetics with reduced sequence similarity to natural insulins has significant therapeutic implications. Current insulin therapies, while effective, face challenges such as physical stability. Novel insulin mimetics could potentially overcome these limitations while maintaining or even enhancing therapeutic efficacy. Additionally, such mimetics could offer improved pharmacokinetic profiles or novel modes of action.

In this chapter, we explore the design and testing of single-chain insulin mimetics with low sequence homology to wild-type insulin using two distinct computational approaches. We leverage the Protein CREATE platform described in earlier chapters to screen these designs against multiple targets, characterize key properties of

successful binders, and validate selected insulin mimics.

5.2 Overview and Comparison of Design Strategies

Given the diversity observed in natural insulin receptor binding solutions, we hypothesized that different computational design strategies would explore distinct regions of the viable binding space. To test this hypothesis, we employed two complementary design strategies: foldtuning and generator-scorer.

Foldtuning Design Strategy

Foldtuning leverages iterative refinement of a protein large language model to generate sequences that preserve structural features while reducing sequence identity to natural proteins. The approach works by fine-tuning a protein language model to produce sequences that maintain structural similarity to natural insulins and insulin-like peptides while decreasing sequence identity to natural sequences. This method emphasizes structural conservation while allowing substantial sequence divergence, mimicking natural evolutionary processes that have produced diverse insulin-like peptides across species.

Generator-Scorer Design Strategy

Our generator-scorer approach utilizes a two-step process:

1. A protein large language model proposes candidate sequences.
2. A scoring algorithm evaluates these candidates based on predicted binding affinity.

Specifically, candidate sequences are evaluated using AlphaFold2's interchain predicted template modeling (iPTM) score, which estimates the likelihood of interaction between the designed protein and the insulin receptor (Yin et al., 2022). Sequences that achieve high iPTM scores are selected for experimental testing, and the results feed back into the generator model to improve future designs.

Unlike foldtuning, which emphasizes structural conservation with sequence diversity, the generator-scorer approach directly optimizes for predicted binding affinity, potentially allowing greater structural variation.

Comparison of Key Properties

As hypothesized, each design strategy produced sequences with distinct biochemical characteristics (Figure 5.1). Foldtuned variants generally preserved most receptor contacts found in wild-type insulin, which aligns with the strategy's emphasis on structural conservation. However, interestingly, foldtuned variants contained on average 5 cysteine residues, rather than the 6 cysteines found in wild-type insulin that form its characteristic 3 disulfide bonds (Menting et al., 2013).

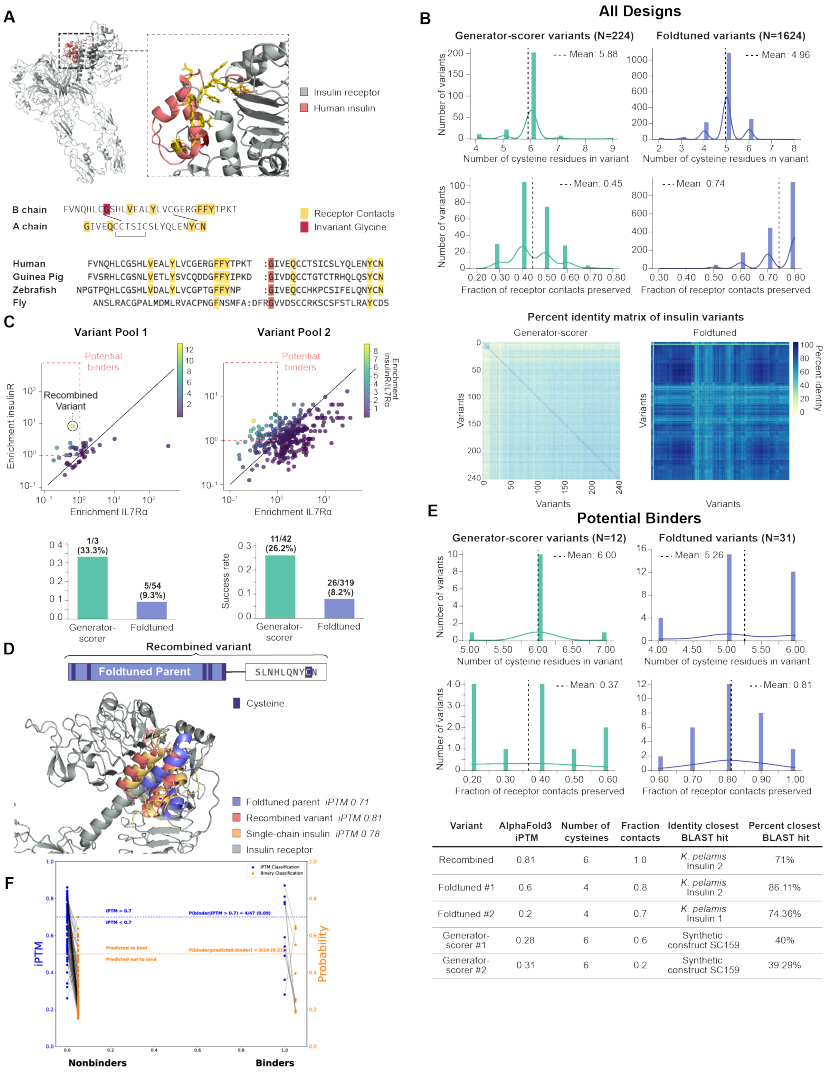
In contrast, generator-scorer variants typically contained 6 cysteines, matching wild-type insulin, but preserved fewer receptor contacts. This difference likely reflects the optimization targets of each strategy: foldtuning prioritizes structural similarity while allowing sequence divergence, while the generator-scorer approach directly optimizes for predicted binding, potentially discovering alternative binding modes.

The two design strategies also produced distinct sequence diversity patterns. Foldtuned variants clustered together in sequence space but remained distant from natural insulins, whereas generator-scorer variants showed greater dispersion but occasionally included sequences more similar to natural insulins. This pattern suggests the two approaches effectively sample different regions of viable sequence space for insulin receptor binders.

5.3 Screening Against Multiple Targets to Confirm Specificity

A crucial challenge in designing protein binders is ensuring specificity for the intended target. Given its 3 disulfide bonds, insulin-like peptides are particularly challenging to screen given the requirement for all disulfides to be formed correctly in order to have a functional protein. Assuming random chance of the proper disulfide bonds forming, that is a probability of 1/15 = 0.067. I hypothesized that misfolded variants may be non-specific, leading to high rates of binding on multiple targets. To specifically identify targets that bound to insulin receptor, I leveraged the Protein CREATE platform to screen our designed insulin mimetics against multiple targets in parallel, which provided a powerful approach to identify truly specific binders. Candidate insulin mimetics were displayed on T7 bacteriophage and screened against both the insulin receptor and IL7RA, an unrelated receptor that served as a specificity control (Figure 5.1). This parallel screening approach allowed us to distinguish between variants showing true specific binding to the insulin receptor versus those exhibiting non-specific binding behavior.

As expected, many variants were enriched on both IL7RA and Insulin receptor,



Variant	AlphaFold3 iPTM	Number of cysteines	Fraction contacts	Identity closest BLAST hit	Percent closest BLAST hit
Recombined	0.81	6	1.0	*K. pelamis* Insulin 2	71%
Foldtuned #1	0.6	4	0.8	*K. pelamis* Insulin 2	86.11%
Foldtuned #2	0.2	4	0.7	*K. pelamis* Insulin 1	74.36%
Generator-scorer #1	0.28	6	0.6	Synthetic construct SC159	40%
Generator-scorer #2	0.31	6	0.2	Synthetic construct SC159	39.29%

Figure 5.1: a) The human insulin sequences for the B-chain and A-chain are shown, linked via one intrachain and two interchain disulfide bridges. Residues implicated in binding to the human receptor are highlighted, along with an invariant glycine essential for proper folding. Sequence alignments of insulin homologs from other species are shown with preserved receptor contacts highlighted. The structure of human insulin with highlighted receptor contacts is also shown (PDB: 6SOF). b) Two different design strategies were used to generate predicted insulin receptor binders. Foldtuning preserves more receptor contacts but produces many variants with an unpaired cysteine relative to other design strategies.

indicating nonspecific binding. Potential binders were identified by selecting variants with enrichment > 1 for Insulin receptor, but < 1 for IL7RA. We identified 44 putative binders using this filter (12 generator scorer and 31 foldtuned variants along with 1 variant that was the product of a recombination event).

The most enriched variant in our first screen was not created by either design strategy, but rather was the product of a recombination event between two foldtuned variants (Figure 5.1d). This recombined variant is predicted to have superior structural and biochemical properties compared to either foldtuned parent, as it contains no unpaired cysteines, preserves all insulin receptor contacts, and has a structure similar to that of single-chain human insulin when co-folded with the insulin receptor.

Of the other 43 variants selected from our filter, we see a strong selection in favor of variants without unpaired cysteines for both design strategies, while each design strategy converges on different ways to make contact with the insulin receptor. Foldtuned variants that preserve receptor contacts are enriched, while generator-scorer variants undergo no such selective pressure (Figure 5.1e). These trends are also observed when picking out individual variants most enriched in the assay. While all variants contain no unpaired cysteines, the variants differ markedly in the fraction of receptor contacts preserved. Interestingly, a majority of the enriched variants have an iPTM score less than 0.6, indicating that a subset of viable designs may be excluded from design campaigns that use Alphafold-based filtering as a pre-screen. Additionally, as hypothesized, key sequence features are conserved within design strategies but differ between them as evidenced by the identity of the closest BLAST result for each of the top enriched variants

To demonstrate how Protein CREATE data can improve future rounds of design, we trained a binary classifier to predict insulin receptor binders from protein sequence using the data collected from our assay. We compared model performance on a held-out test set to predictions made using iPTM alone (Figure 5.1f). The model is able to reduce the high false positive rate of 91% (43/47) from iPTM scoring to 79% (11/14). This performance should be able to be further improved by collecting more data on binders due to the relative class imbalance currently seen in the data.

5.4 Biochemical Characterization of Novel Insulin Mimetics

Production and Folding of Variants

We selected eight variants of interest that were enriched in the Protein CREATE assay to synthesize with Amide Technologies. All eight variants were successfully

produced, but as of this writing, the success in refolding the variants into a functional form remains mixed. The below table emphasizes status of each variants:

Protein ID	**Status**
Recombined	Synthesized, but not refolded
ActorCritic77	Synthesized and refolded
ActorCritic81	Synthesized and refolded, with some higher MW adducts as impurities
ActorCritic173	Synthesized and refolded, with possibly desulfurization (-32da) as an impurity
Foldtuned2012	Synthesized, but not refolded
ActorCritic158	Synthesized and refolded
FoldtunedNeg145	Synthesized, but not refolded
ActorCritic209	Synthesized and refolded

Table 5.1: Protein Refolding Status Summary

The results thus far already show clear distinctions between the two methods in their capacity to generate properly folded, biologically active proteins in the conditions used to refold insulin industrially.

5.5 Functional Characterization of Insulin Mimetic

Synthetic insulins may form multiple isomers during the refolding process, resulting in different retention times for each. These can be isolated using Reverse Phased-High Pressure Liquid Chromatography (RP-HPLC). As refolded insulin mimics are tested, they will be evaluated for insulin receptor agonist activity using a cell-based phosphorylation assay. Cultured cells expressing the insulin receptor are exposed to increasing concentrations of the insulin analog (10^{-4} to 10^{4} nM) to generate a dose-response curve. Activation of the insulin signaling pathway is quantified by measuring downstream phosphorylation events using an enzyme-linked immunosorbent assay, with results reported as absorbance at 450 nm. Dose-response data is fitted to a four-parameter logistic model to determine potency. Characterization of other fractions are ongoing, as of this writing. Any detectable binding of new-to-nature sequences will be noteworthy. For context, human insulin-like growth factor 1 (IGF-1), which has approximately 50% identity to insulin, shows an EC_{50} of 30-100 nM to the insulin receptor (Daughaday et al., 1987).

5.6 Conclusion

This chapter has described the application of Protein CREATE toward the design of a particularly challenging peptide mimetic with historic scientific and therapeutic value. Our exploration of insulin mimetics with low sequence homology to wild-type

insulin has provided several key insights into both the biology of insulin-receptor interactions and the strengths of different computational design approaches.

The parallel application of foldtuning and generator-scorer design strategies yielded distinct populations of candidate insulin mimetics, each exploring different regions of viable sequence space. While foldtuned variants preserved more receptor contacts but often contained unpaired cysteines, generator-scorer variants maintained proper cysteine pairing potential but with fewer conserved receptor contacts. This divergence in design properties underscores the value of employing multiple complementary approaches when exploring challenging design spaces.

Our multi-target screening approach using Protein CREATE successfully identified specific insulin receptor binders while filtering out non-specific interactions, a critical capability when working with cysteine-rich proteins where misfolding can lead to promiscuous binding behavior. The serendipitous discovery of a recombined variant with superior predicted properties highlights the advantages of high-throughput experimental screening in identifying solutions that computational approaches alone might miss.

Post-screening analysis revealed important selection patterns, with specific enrichment for variants containing proper numbers of cysteines across both design strategies, suggesting the importance of disulfide bond formation for functional insulin mimetics. The observation that many enriched variants had relatively low iPTM scores challenges the conventional use of structure prediction metrics as primary pre-screening filters in computational protein design.

The differential success in refolding between generator-scorer and foldtuned variants suggests that sequence patterns optimized for predicted binding do not necessarily translate to favorable folding characteristics—an important consideration for therapeutic protein development.

Overall, this work demonstrates the power of combining diverse computational design approaches with high-throughput experimental screening to access unexplored regions of protein sequence space while maintaining specific functional properties. The work represents an important step toward the development of novel insulin therapeutics with reduced sequence identity to wild-type insulin, potentially addressing challenges in physical stability, pharmacokinetics, and immunogenicity that affect current insulin therapies and demonstrates the value of collecting data from diverse approaches to increase success rates on difficult targets.

Figure 5.1: c) The designs were assayed for insulin receptor binding using Protein CREATE. Due to potential misfolded variants that could show nonspecific binding, binders were determined by taking variants enriched on insulin receptor binding but de-enriched on an off-target receptor (IL7RA). d) A recombination event between two foldtuned variants was the most enriched variant in our first screen. Analysis indicated the addition of the C-terminal end shown improves predicted structural characteristics while restoring a possible disulfide bond.

Figure 5.1: e) Binders from two of the tested design strategies (inverse folded variants were not prevalent in the pool and thus not shown) were analyzed to determine the relative importance of key insulin properties. The prevalence of designs with an odd number of cysteines decreased for both designs; however generator-scorer designs showed a slight decrease in receptor contacts on average, while foldtuned variants showed a slight increase. The top five hits, based on having the highest on-target insulin receptor enrichment over off-target IL7RA ratio, were chosen for individual analysis. With the exception of the recombined variant, all variants show relatively low iPTM scores, suggesting using iPTM as a prescreen for binders may lead to false negatives. f) A held-out test set of variants from both design strategies and variant pools are classified into binders or nonbinders based on either iPTM scoring (blue) or a model trained on experimental data (orange). Variants above the blue and orange lines are predicted as binders by each respective method and nonbinders otherwise. True labels for all variants are given on the x-axis.

Chapter 6

DESIGNING BINDERS FOR THE HUMAN SWEET TASTE RECEPTOR - CHALLENGES AND METHODOLOGIES

6.1 Introduction: When Computational Models Fall Short

The advent of deep learning models such as AlphaFold, RoseTTAFold, and ESMFold has undeniably transformed structural biology and, by extension, the field of protein design (Jumper et al., 2021; Lin, Akin, Rao, Hie, Zhu, Lu, Smetanin, et al., 2023; Baek et al., 2021). These tools can predict protein structures with remarkable accuracy, often approaching experimental resolution. We have already seen how extracted metrics from these models, such as the interface predicted template modeling score (iPTM), can be translated into heuristics to guide and screen designs.

However, mapping sequence to function is still problematic, and reliance solely on predictive heuristics insufficient, particularly when tackling challenging biological targets (Wu et al., 2021). These generalized metrics may not capture the nuances required for specific, complex interactions or targets that deviate significantly from the training data distributions (Azzaz et al., 2022; Wallner et al., 2023). This limitation became apparent during efforts to design protein binders for a commercially significant and structurally complex target: the human sweet taste receptor.

This chapter details the investigation into designing binders for this receptor. It begins by outlining why the heuristics derived from structure prediction models proved inadequate for this specific design challenge. We will delve into the unique characteristics and complexities of the human sweet taste receptor (TAS1R2/TAS1R3) as a protein binding target (Chandrashekar et al., 2006; Temussi, 2009). Subsequently, an overview of the computational and experimental tools available for assessing protein designs will be presented. Finally, this chapter will chronicle the efforts undertaken in the Thomson lab to implement and integrate these tools—both individually and as part of a high-throughput pipeline—to characterize designed binders effectively.

6.2 Background: The Human Sweet Taste Receptor (TAS1R2/TAS1R3)

The perception of sweetness in humans is primarily mediated by a single receptor, a heterodimer composed of two Class C G protein-coupled receptors (GPCRs):

Taste receptor type 1 member 2 (TAS1R2) and Taste receptor type 1 member 3 (TAS1R3) (Chandrashekar et al., 2006). Like other members of the Class C family (which includes metabotropic glutamate and GABA receptors), TAS1R2 and TAS1R3 feature large N-terminal extracellular domains, known as Venus Flytrap Domains (VFDs), linked via a cysteine-rich domain (CRD) to the canonical seven-transmembrane helix domain (TMD) (Temussi, 2009).

This receptor exhibits promiscuity in ligand recognition, binding not only to canonical sugars like sucrose and glucose but also to a wide array of chemically diverse molecules, including small molecule sweeteners both natural (e.g. mogrosides, and steviol glycosides) and artificial (e.g. aspartame, sucralose) (Hao et al., 2024). Notably, several naturally occurring proteins can also trigger high intensity sweet responses (Kant, 2005). Eliciting a sweet taste response in the absence of sugar can help reduce sugar intake and combat its associated metabolic diseases (Zhang et al., 2010). While a variety of sweeteners can trigger this response, all either have non-sugar like onset and off-set characteristics (e.g. slower onset and after-taste) or activate other taste receptors, such as bitter receptors, producing off-notes (Temussi, 2009). Furthermore, the expression of TAS1R2/TAS1R3 extends beyond the oral cavity to tissues like the intestine and pancreas, where it participates in glucose sensing, regulation of glucose transporter expression (SGLT1, GLUT2), and overall glucose homeostasis, potentially influencing insulin release and GLP-1 secretion (Kochem, Hanselman, and Breslin, 2024). Activation of this receptor by non-nutritive sweeteners could lead to undesirable responses as well as impacts on the microbiome (Suez, Korem, et al., 2014; Suez, Cohen, et al., 2022).

6.3 Sweet Proteins: Nature's High-Potency Sweeteners

A fascinating subset of TAS1R2/TAS1R3 ligands are proteins that elicit a sweet sensation, often with potencies orders of magnitude greater than sucrose (Kant, 2005). These proteins represent attractive candidates for natural, low-calorie sugar substitutes. Key examples include:

- **Brazzein**: Originating from the West African plant *Pentadiplandra brazzeana*, brazzein is the smallest known sweet protein (54 amino acids, ~6.5 kDa). It is remarkably stable across wide pH and temperature ranges, largely due to its four intramolecular disulfide bonds. Its sweetness potency is estimated to be 500 to 2000 times that of sucrose on a weight basis. Its stability, small size, and desirable taste profile make it a prime target for commercial development

and protein engineering studies (Caldwell et al., 1998).

- **Thaumatin**: Found in the katemfe fruit (*Thaumatococcus daniellii*), thaumatin is significantly larger than brazzein and possesses exceptional sweetness potency, reported to be up to 100,000 times sweeter than sucrose on a molar basis (though estimates vary). It features eight disulfide bonds and is approved as a sweetener and flavour enhancer (E957) for decades in numerous countries (EFSA Panel on Food Additives and Flavourings (FAF), 2021).
- **Other Sweet and Taste-Modifying Proteins**: The repertoire also includes Mabinlin, Monellin, Pentadin, and the taste-modifying proteins Curculin (which is also sweet) and Miraculin (which converts sour tastes to sweet) (Kant, 2005).

A common characteristic distinguishing these proteins from simple sugars is their temporal sweetness profile. Unlike the rapid onset and fast decay of sucrose, sweet proteins often exhibit a noticeable delay in sweetness perception followed by a lingering sweet aftertaste. For instance, thaumatin is known for its slow onset and persistent sweetness (Joseph et al., 2019).

6.4 Complex Ligand Interactions and Activation Mechanisms

The interaction between ligands and the TAS1R2/TAS1R3 receptor is complex and not fully elucidated. While small molecules like sucrose bind within the Venus Flytrap domain (VFD) clefts (sucrose potentially interacting with both TAS1R2 and TAS1R3 VFDs, while many artificial sweeteners bind primarily to the TAS1R2 VFD), the sheer size of sweet proteins precludes their binding entirely within these clefts. Early models proposed that proteins like brazzein might insert 'fingers' or specific loops into the VFD cleft to trigger activation (Laffitte et al., 2022).

Recent computational work, including studies from the Goddard group, suggests even greater complexity and alternative binding sites. Steviol glycosides, for example, are predicted to bind to multiple distinct sites on the receptor heterodimer — potentially up to four, encompassing regions within both the VFDs and the transmembrane domains (TMDs). This multi-site interaction model could help rationalize previously confusing experimental data, particularly from competition binding assays, and suggests that different ligands might stabilize distinct receptor conformations or utilize different binding locations (Hao et al., 2024).

Receptor activation involves significant conformational changes. Ligand binding, predominantly initiating at the TAS1R2 VFD (VFD2), is thought to trigger a cascade of structural rearrangements, starting at VFD2, and then to TAS1R3 VFD (VFD3), and the TAS1R3 cysteine-rich domain (CRD3), followed by the TAS1R3 TMD (TMD3), leading to G protein activation. Mutational and docking studies suggest that brazzein, one of the sweet proteins, binds directly to CRD3, perhaps triggering activation by bypassing the VFD2 and VFD3 conformation changes from small molecule sweeteners (Assadi-Porter et al., 2010).

Recently developed artificial intelligence tools, such as AlphaFold Multimer, have been useful to predict as well as design protein-protein interactions. Not only do these tools give relative confidence in a predicted interaction, but also the likely interacting spots of a protein when co-folded together. The interaction predicted template modeling score (iPTM) is predictive of the presence of an interaction, though not its strength. Scores closer to 1 indicate a higher likelihood of an interaction, while those closer to 0 indicate the lack of predicted interactions. Given the stochastic nature of AlphaFold, lower scores tend to have varied predicted protein-protein interaction sites (Bennett et al., 2023).

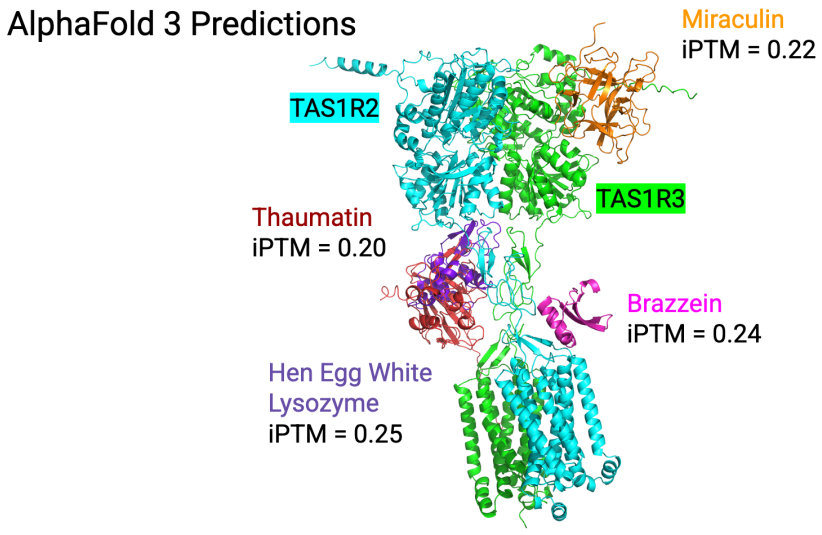


Figure 6.1: AlphaFold 3 iPTM scores for known sweet proteins with the human sweet taste receptor. Despite experimental evidence of interaction, all tested sweet proteins show low iPTM scores (<0.25), indicating high false negative rates when using computational prediction for sweet protein design.

I co-folded several known sweet proteins with the human sweet taste receptor using the latest model developed by DeepMind and Isomorphic labs, AlphaFold 3

(Abramson et al., 2024; Wallner et al., 2023). Despite the fact that the four proteins interact with the sweet taste receptor, none of them have an iPTM score greater than 0.25 (Figure 6.1). The known sweet or sweet-associated proteins (thaumatin, monellin, brazzein, mabinlin, miraculin, curculin, lysozyme, and pentadin) share minimal sequence homology and diverse tertiary structures despite their similar sweet-inducing properties (Kant, 2005), suggesting convergent evolution of their taste-modifying functions through different structural solutions (Caldwell et al., 1998). The failure of AlphaFold to predict their interactions with the human sweet taste receptor indicates a likely high false negative rate and overall lack of model power for the design of novel sweet proteins (Bryant, Pozzati, and Elofsson, 2022; Yin et al., 2022).

6.5 Sweet Nothings: The Challenge of Reagent Authenticity

A significant, pragmatic obstacle encountered during this research, and one that threatens the broader field of sweet taste receptor biology, is the questionable authenticity of commercially available sweet protein reagents. Of the intensely sweet proteins, thaumatin is the most widely available. However, we found many of these "thaumatin" samples contain no protein and only a mix of alternative sweeteners.

To systematically investigate this, I collaborated with the on-campus Proteome Exploration Laboratory (PEL) to analyze the composition of various sweet protein samples obtained from different commercial vendors. Surprisingly, samples from many vendors contained either no thaumatin, such as the one from Sigma Aldrich (which contained a mixture of sucrose and neotame, a highly potent artificial sweetener), or a mix of thaumatin with another high-intensity sweetener, such as the sample from Tokyo Chemical Industry (TCI), which not only contained thaumatin protein, but also neotame. Only the sample from Apura Ingredients contained thaumatin without any small molecule sweetener impurities.

As it is approved both as a sweetener and a flavoring agent, thaumatin is also present in consumer products. According to Mintel's Global New Products Database (GNPD), thaumatin (E957) was labeled on only a small number of products (n = 34) between January 2016 and May 2021. However, it may have more widespread use. Because such small amounts are required to either sweeten products or enhance their flavor, it often falls below labeling thresholds (EFSA Panel on Food Additives and Flavourings (FAF), 2021).

I hypothesized that, given the prevalence of fake thaumatin as a scientific reagent,

that adulteration may be occurring in the consumer sector as well. To investigate this claim, I purchased a can of SORTED, a thaumatin-sweetened beverage sold in Australia. Mass spectroscopy failed to detect any thaumatin-associated peptides, but did detect the presence of sucralose, another high-intensity artificial sweetener.

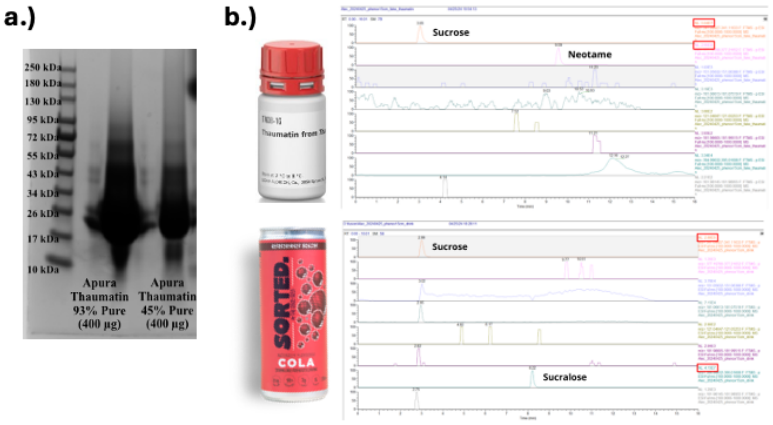


Figure 6.2: a.) SDS-PAGE gel showing protein of the expected size (23 kDa) from one supplier, indicating genuine thaumatin is present in the sample. b.) Retention times for various small molecule sweeteners were analyzed using HPLC chromatography to compare "thaumatin" samples from two different sources: Sigma Aldrich (shown in the top chromatogram) and the consumer soda company SORTED (shown in the bottom chromatogram). The chromatograms reveal distinct sweetener profiles, with the Sigma Aldrich sample containing sucrose and neotame, while the SORTED sample contains sucrose and sucralose.

The widespread fraud uncovered makes more sense given the relative obscurity of thaumatin within the food industry (EFSA Panel on Food Additives and Flavourings (FAF), 2021). While mass spectroscopy based analysis of samples is out of reach for many small formulators, food manufacturers, and laboratories, it was largely unnecessary to judge the authenticity of the source. Genuine thaumatin appears as a brown powder, derived from its natural plant origins and minimal processing, while counterfeit thaumatin is white. This suggests most buyers are unfamiliar with thaumatin's basic characteristics. As it is still extracted from the arils of the katemfe fruit grown in West Africa, it is an expensive sweetener relative to the artificial sweeteners found in counterfeit thaumatin samples, making fraud economically advantageous. Both of these characteristics — relative obscurity in the public's imagination and high cost relative to other sweeteners — incentivize fraud and are true for other sweet proteins as well, which are even more scarce and difficult to obtain. The presence of fake thaumatin samples from scientific reagent companies

such as Sigma Aldridge and Tokyo Chemical Industry may also hinder research meant to uncover sweet protein – sweet taste receptor binding interactions, as small molecule sweeteners likely have alternative pathways for activating the human sweet taste receptor than larger sweet protein activators like thaumatin.

6.6 Methodologies for Assessing Designed Binders

Evaluating the success of protein design efforts, especially for complex targets like TAS1R2/TAS1R3, requires a multi-faceted approach combining biochemical, cellular, and computational methods. Our efforts focused on establishing and integrating several key techniques:

Receptor Purification and Biochemical Characterization

Directly studying the interaction between designed binders and the receptor necessitates obtaining purified receptor components. Both TAS1R2/TAS1R3 are G-protein coupled receptors with seven-transmembrane domains (Chandrashekar et al., 2006). This makes them exceptionally difficult to produce recombinantly and purify. TAS1R3 in particular is known to have low expression and does not localize to the membrane in the absence of TAS1R2. Furthermore, TAS1R2 forms homodimers in addition to the heterodimer with TAS1R3 necessary for sweet taste perception (Belloir et al., 2021). The development and usage of lauryl maltose neopentyl glycol (LMNG) in 2010 has made it possible to solubilize and stabilize many membrane proteins, including TAS1R2 and TAS1R3 (Chae et al., 2010). Belloir et al were able to purify the TAS1R2 homodimer from a HEK293S cell line in 2021 (Belloir et al., 2021). Recently, the TAS1R2/TAS1R3 structure was determined using a similar purification strategy (Madsen et al., 2024).

Though structure resolution requires large amounts of protein of a single species, I reasoned that a small scale purification from Expi293 cells with minimal clean-up may yield enough receptor to be used in a Protein CREATE screen. In collaboration with Zhilin from the Voorhees lab at Caltech, I expressed and purified FLAG-tagged TAS1R2 and TAS1R3 solubilized with LMNG from Expi293 cells. Western blot analysis showed the presence of multiple products as expected. T7 bacteriophage displaying both wild-type brazzein and thaumatin sequences was mixed with phage displaying off-target sequences and allowed to bind the crude purification product. A slight enrichment of both brazzein displaying phage and thaumatin displaying phage, distinguishable from the off-target phage provided initial proof of concept results that the Protein CREATE platform can be used to screen for novel sweet

proteins.

Cell-Based Functional Assays

To assess whether designed binders can functionally activate the sweet taste receptor, cell-based assays are indispensable. This involved developing or utilizing cell lines (commonly HEK293 cells) engineered to stably or transiently express the functional TAS1R2/TAS1R3 heterodimer (Belloir et al., 2021). Upon ligand binding and receptor activation, downstream signaling cascades are initiated, typically involving gustducin activation, phospholipase C-β2 (PLC-β2) stimulation, and subsequent generation of second messengers like inositol 1,4,5-triphosphate (IP3) and diacylglycerol (DAG) (Chandrashekar et al., 2006), often leading to measurable changes in intracellular calcium levels.

Assays monitoring these downstream events (e.g., calcium response assays) allow for the quantitative assessment of ligand potency and efficacy.

Cell lines tying GPCR activation to downstream signaling and detecting the calcium mobilization response have been reported in the literature (Belloir et al., 2021). These HEK293 cells express TAS1R2 and TAS1R3 along with a mutant G-protein, gustducin G16-gust44. Intracellular calcium levels are measured using either a fourth construct expressing GCAMP6, a calmodulin – GFP fusion protein or an intracellular calcium responsive dye.

Those cell lines reported in the literature typically rely on cotransfection of TAS1R2 and TAS1R3 expressing plasmids into a cell line stably expressing G16-gust44, a chimeric G-protein containing the last 44 amino acids of gustducin (Ueda et al., 2003). To address concerns about co-transfecting multiple plasmids and increasing repeatability of performing functional sweet taste receptor activation in the future, I designed plasmids to integrate the TAS1R2 and TAS1R3 genes into the genome using lentiviral vectors. Efforts to test these constructs are still underway as of the time of this writing.

Computational Modeling: Molecular Dynamics and Docking

Complementing experimental approaches, computational modeling provides atomic-level insights into receptor-ligand interactions. Computational studies are easier and faster to perform than experimental ones, facilitating rapid design-build-test loops. As we have previously seen, recent AI-based modeling approaches fail when it comes to predicting sweet protein binding. As a result, I used traditional docking

and molecular dynamics (MD) simulations to investigate if sweet protein designs can be pre-screened for function (Honorato et al., 2024; Abraham et al., 2015). Scouring the literature, I found that many mutants of brazzein have been created and their sweetness levels are characterized relative to wild type (Jin et al., 2003; Walters et al., 2009; Singarapu et al., 2016). Both mutants with higher and lower levels of sweetness have been produced (Caldwell et al., 1998; Kant, 2005). I curated a list of 23 brazzein mutants with varying levels of sweetness compared to wild-type. After folding each variant using AlphaFold 3 (Abramson et al., 2024), I collaborated with Jiapei Miao from our lab to run molecular dynamics simulations on each variant and extracted the internal energy of each (Abraham et al., 2015). We found that the stability of the protein, as measured by having a lower potential energy, increased the likelihood of a variant having increased sweetness (Figure 6.3). This provides a promising strategy for creating and optimizing new sweet protein variants computationally before running any wet lab experiments.

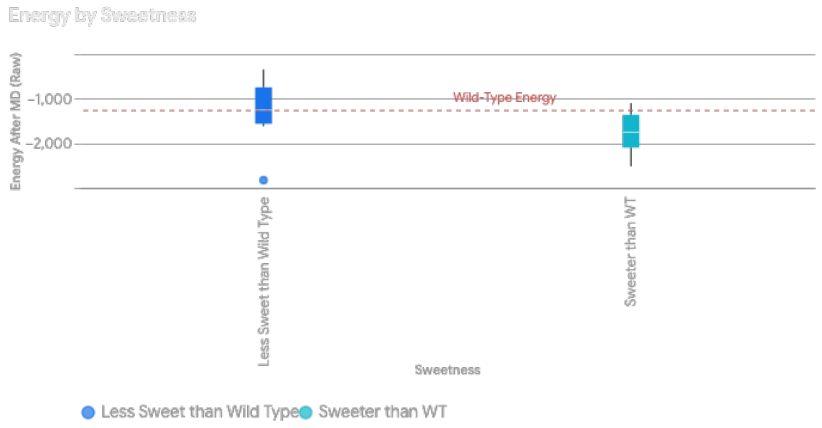


Figure 6.3: Relationship between protein stability (measured by potential energy from molecular dynamics simulations) and sweetness for brazzein variants. More stable variants (lower potential energy) tend to exhibit higher sweetness, providing a computational metric for pre-screening sweet protein designs.

6.7 Production of Sweet Protein Variants

A critical bottleneck in testing designed proteins is their production at sufficient scales for testing. Sweet proteins of commercial interest that are thermostable — brazzein and thaumatin — also have made disulfide bonds — four and six, respectively, which must be properly formed to ensure proper function. Despite efforts to produce it recombinantly, most (authentic) sources of thaumatin are extracted from the katemfe fruit. Partly due to its smaller size and fewer disulfide bonds, brazzein

is comparatively easier to biomanufacture. However, with eight cysteine residues, there are 105 possible arrangements, with only one configuration yielding the functionally active protein. This seems daunting, but is not actually a concern. Izawa et al (1996) were able to synthesize brazzein via peptide synthesis, reduce the protein using DTT, and later refold it via air oxidation in buffer without a glutathione redox system or specialized folding agents. This demonstrates that the brazzein fold is the most thermodynamically favored one out of all accessible folds (Izawa et al., 1996).

As part of my investigation to find ways to produce and assay sweet proteins, I conducted an experiment to validate the findings of Izawa et al. (1996) using chemically synthesized brazzein. I ordered 3 mg of lyophilized brazzein peptide obtained from Elim Biopharmaceuticals. To fully denature and reduce the protein:

1. The lyophilized brazzein was reconstituted in 100 µL of denaturing buffer containing 6 M guanidine hydrochloride and 0.2 M ammonium acetate.
2. Dithiothreitol (DTT) was added to a final concentration of 500 mM (7.7 mg) to ensure complete reduction of all disulfide bonds.
3. The mixture was incubated at room temperature for 30 minutes to allow complete reduction.
4. The reduced protein was concentrated using a 3 kDa molecular weight cutoff filter.
5. The sample was washed within the filter three times with 0.2 M ammonium acetate to remove the denaturant and reducing agent.
6. After the final concentration step, approximately 60 µL of protein solution remained, with an estimated concentration of 50 mg/mL based on the starting material.

To monitor the spontaneous refolding process, I evaluated the recovery of sweetness, which serves as a functional assay for correctly folded brazzein:

1. The first taste assessment was conducted approximately 15 hours after initiating the refolding process. No discernible sweetness was detected at this timepoint.

2. A second assessment at 69 hours post-reduction suggested a slight sweet taste, indicating the beginning of functional recovery.

3. The sample was then passed through a 7 kDa Zeba desalting column and resuspended in 0.2 M magnesium acetate to remove any remaining denaturants or reducing agents.

4. At 112 hours (approximately 4.7 days), a definite sweet taste was detected, confirming partial recovery of the functional protein. The protein concentration was measured at 1.03 mg/mL. This is several factors higher than the 0.152 mg/mL level detection threshold for sweetness.

Though the recovery of the sweet taste indicated successful production of functional brazzein, the cost of the synthesis ($540), limits the throughput of peptide synthesis for large-scale variant screening.

Production of brazzein in heterologous hosts presents multiple challenges. The first is that modifications to brazzein's structure, such as purification tags or even an N-terminal methionine, can significantly reduce the sweetness of the final product. A brazzein variant with methionine as the starting codon reduced the perceived sweetness by more than half. Second is the need to produce any variants in an oxidizing environment in order to successfully form the 4 disulfide bonds necessary for function. In an effort to investigate cheaper options for medium throughput production of variants, I implemented a previously protocol to produce brazzein within the E. coli periplasm. This approach sidesteps the concerns listed above by using a pelB leader sequence, which is cleaved upon successful export of brazzein to the periplasm, where the oxidizing environment facilitates proper folding of the wild-type, functional brazzein sequence. I designed and transformed expression plasmids for both wild type brazzein and a triple mutant (H31R/E36D/E41A) previously found to be ~18 times as sweet as the wild type into ClearColi E. coli, which do not produce viable endotoxin. These variants were grown in Studier media (ZYM-5052) to auto-induce brazzein expression during growth at 37ºC. The brazzein was extracted using a combination of osmotic shock to exclusively release proteins from the periplasm followed by a heat denaturation to precipitate out non-heat stable proteins.

Specifically, after overnight growth of 0.4 L of each culture, the supernatant of each variant (1.50 g of cells containing wild type brazzein, 1.44 g of the triple mutant brazzein) was resuspended in 30 mM Tris-HCl, pH 8.0 for a wash step. The pellet

was then treated with 30 mM Tris-HCl, pH 8.0 and 5 mM CaCl2 and incubated for 5 minutes at room temperature before being re-pelleted and the supernatant removed. The pellet was then treated with 20 mL of hypertonic sucrose solution (30% sucrose and 30 mM Tris-HCl, pH 8.0). Following resuspension, EDTA was added to the pellet directly to a final concentration of 5 mM and the solution incubated at room temperature for 15 minutes at room temperature. The cells were pelleted and the supernatant was collected for analysis while the pellet was resuspended in 20 mL of ice-cold deionized water. The resuspension was added to an ice bath for 10 minutes before the supernatant was collected and subjected to heat treatment at 80ºC for 1 hour. After centrifugation to remove precipitate, the resulting protein was concentrated and washed in deionized water. The presence of sweet taste confirmed successful expression and purification of brazzein while the relatively low cost ($122.50) for DNA synthesis of a cloned expression plasmid makes this approach a more viable option for medium-throughput sweet protein production than peptide synthesis.

Another approach investigated briefly due to rapid protein production and use of linear DNA obtained via PCR or synthesis was the use of E. coli cell free extracts. As we will see in the next chapter, this type of protein production is amenable to automation, making it potentially higher throughput than the other sweet protein production methods discussed. Several disulfide enhancer supplements exist for these systems, such as the myTXTL Antibody/DS Kit from Arbor Biosciences or PURExpress® Disulfide Bond Enhancer supplement from New England Biolabs. However, the absence of cellular compartments in cell free systems complicates the use of signal sequences like pelB that are cleaved during sequence processing. As a result, I chose to systematically investigate the former approaches rather than this later one for sweet protein screening. However, for other types of protein screening, cell free production was immensely useful for streamlining workflows and confirming function and production of protein variants for biochemical and *in vitro* assays within 1-2 days.

6.8 Conclusion

The design of protein binders for the human sweet taste receptor presents unique challenges that push the boundaries of current computational and experimental methods. The complex nature of the receptor, the diversity of its ligands, and the subtleties of its activation mechanisms all contribute to making this an exceptionally difficult target for protein design.

Our investigations have highlighted several key findings:

1. **Limitations of computational prediction**: Current AI-based structural prediction tools, including AlphaFold, show significant limitations in predicting interactions between sweet proteins and the sweet taste receptor, with high false negative rates observed.

2. **Reagent authenticity challenges**: The widespread fraud in commercially available sweet protein reagents represents a significant hurdle for research in this field, necessitating careful verification of source materials.

3. **Multi-faceted approach required**: Successful design efforts for complex targets like the sweet taste receptor require integration of multiple approaches, including biochemical characterization, cell-based functional assays, and traditional computational modeling techniques.

4. **Production bottlenecks**: Efficient production of properly folded sweet proteins with correct disulfide bonds remains a challenge, though several promising approaches have been identified.

Despite these challenges, our work has established several important foundations for future sweet protein design efforts. The development of receptor purification protocols, cell-based assays, and production methods for sweet protein variants provides essential tools for screening and validating designed binders. Furthermore, our findings regarding the correlation between protein stability and sweetness intensity offer a promising computational strategy for pre-screening designs.

Looking ahead, the integration of these approaches with the Protein CREATE platform described in earlier chapters holds significant promise for accelerating the discovery and optimization of novel sweet protein variants. By enabling high-throughput screening of designed libraries against purified receptor components, this platform could overcome many of the current bottlenecks in sweet protein discovery and design.

Chapter 7

PRODUCTION OF PROTEIN LIBRARIES IN E. COLI CELL FREE EXTRACTS

7.1 Introduction to Cell-Free Protein Production

Cell-free transcription-translation (TXTL) systems represent a powerful platform for protein synthesis that circumvents many limitations of traditional in vivo expression systems (Silverman, Karim, and Jewett, 2020; Thornton et al., 2024). By extracting the cellular machinery necessary for transcription and translation from E. coli cells while excluding cellular barriers such as membranes and cell walls, TXTL systems provide an open and highly controllable environment for rapid protein production (Carlson et al., 2012). These systems contain ribosomes, tRNAs, aminoacyl-tRNA synthetases, translation factors, RNA polymerases, and metabolic enzymes necessary for energy regeneration (Silverman, Karim, and Jewett, 2020).

The advantages of cell-free systems include rapid protein expression (typically within hours rather than days), the ability to produce toxic proteins that would otherwise kill host cells, and simplified purification processes. Additionally, these systems allow for precise control over reaction conditions and direct access to the reaction environment, enabling real-time monitoring and manipulation of protein synthesis. This chapter explores how we leveraged these advantages to develop novel approaches for protein library production and optimization.

7.2 Cell Free Production Pipeline

Cell-free expression systems often utilize regulatory elements that function differently than their *in vivo* counterparts, as they lack cellular barriers and feedback mechanisms present in intact *E. coli* cells (Silverman, Karim, and Jewett, 2020; Carlson et al., 2012). Promoter strengths, ribosome binding site efficiencies, and termination signals can exhibit significantly altered behaviors in the cell-free environment, requiring specific optimization and characterization distinct from traditional *in vivo* expression systems (Karig et al., 2011). I identified regulatory elements for high expression individually reported in the literature and combined them to form a standard expression cassette for producing high amounts of protein. This cassette consists of four main parts:

1. A dual promoter made up of both a high expression native sigma 70 promoter from E. coli and a T7 RNA polymerase annealing site for transcription if T7 RNA polymerase is present.

2. A ribosome binding site identified as strong in both E. coli cells and within cell free extract. Natively, this ribosome binding site drives capsid expression of T7 bacteriophage.

3. An N-terminal polypeptide tag to allow for efficient folding and purification of protein product using downstream nickel affinity chromatography. This sequence, MSHHHHHHHHSENLYFQSGGG, was used to produce < 100 amino acid computationally designed proteins. I also found that this sequence is required for proper folding of at least one previously published protein mini binder sequence. The sequence of this previously published binder (SVIEKL-RKLEKQARKQGDEVLVMLARMVLEYLEKGWVSEEDADESADRIEEVLKK) was chemically synthesized and allowed to bind to its binding partner, IL7RA (Cao et al., 2022). No binding was detected when SPR was performed with this synthesized ligand. However, when the ligand was chemically synthesized with the aforementioned leader sequence, MSHHHHHHHHSEN-LYFQSGGGSVIEKLRKLEKQARKQGDEVLVMLARMVLEYLEKGWVSEEDADESADRIEEV (used to produce the sequence in a bacterial expression system), the mini protein bound with an affinity equal to that measured when it is expressed in bacteria.

4. A terminator sequence derived from T7 bacteriophage to abort transcription.

This cassette is flanked by Twist adaptor sequences that can be amplified using a standard set of primers. Post amplification, the product can be cleaned up using either a commercial PCR cleanup kit or SPRI magnetic beads. This DNA can be added to a cell free reaction master mix, such as Arbor myTXTL, and incubated at 29ºC for >6 hours. For small scale protein purification reactions, 50 to 60 µL of total reaction volume is typically set up. Following cell-free expression, proteins are purified using Ni-NTA affinity chromatography, typically using NEBExpress® Ni Spin Columns. A simplified workflow is as follows:

1. The binding column is washed with binding buffer (20 mM NaH_2PO_4, 300 mM NaCl, pH 7.4) prior to addition of the sample.

2. Cell-free reactions are diluted with the binding buffer.
3. The diluted sample is incubated with Ni-NTA resin on the binding column for 2 minutes with gentle agitation.
4. The column is washed three times with wash buffer (20 mM NaH_2PO_4, 300 mM NaCl, 5 mM imidazole, pH 7.4).
5. Proteins are eluted with elution buffer (20 mM NaH_2PO_4, 300 mM NaCl, 500 mM imidazole, pH 7.4).

This streamlined purification process typically yields 10-50 µg of purified protein per 50 µL cell-free reaction, sufficient for many biochemical assays, such as surface plasmon resonance (SPR), performed elsewhere in the thesis. The entire pipeline, from PCR to purified protein, can be completed within a single day, representing a significant advantage over traditional expression systems.

7.3 Cell Free Production of Endotoxin-free Proteins

A major limitation of conventional E. coli-based expression systems is the presence of endotoxins, primarily lipopolysaccharides (LPS) from the outer membrane. These endotoxins can trigger strong immune responses when introduced into mammalian systems, making standard E. coli extracts unsuitable for producing proteins intended for therapeutic applications or immunological studies. This is an important consideration for functional assays we have performed with cytokine mimics or cytokine receptor inhibitors, as these experiments involve exposing the purified protein to populations of peripheral blood mononuclear cells (PBMCs), including immune cells. As even small amounts of endotoxin can cause a response, producing the recombinant protein of interest in a non-endotoxin containing host strain (such as a mammalian expression cell line such as HEK293 or endotoxin-free bacteria such as Bacillus subtilis). Optimizing production and cloning plasmid for alternative host strain production is laborious, so finding an alternative would be highly desirable.

Through a collaborative effort with Han Zhang co-advised by Kaihang Wang and Richard Murray, we developed endotoxin-free cell-free extracts using ClearColi™, an engineered E. coli strain with modified LPS structures that do not elicit immune responses (Mamat et al., 2015).

ClearColi Cell Extract Preparation

ClearColi BL21(DE3) were cultured in 2YT media until reaching mid-to-late-log phase (OD~1.4), then harvested by centrifugation. The cell pellet was washed with S30 buffer (14 mM Magnesium Glutamate, 60 mM Potassium Glutamate, 2 mM DTT, pH 8.2). Cell disruption was performed via sonication at 50% amplitude using a 5 s on/5 s off cycle until reaching a total energy input of ~1000 J per 1 mL of suspension. Following lysis, cellular debris was removed by centrifugation to clarify the extract. A run-off reaction was conducted at 37°C for 1 hour to reduce background expression, followed by 2 hours of dialysis in S30 buffer. A final centrifugation step was performed to remove any remaining precipitates, yielding cell-free extract suitable for subsequent experiments.

Cell-free protein synthesis reactions were formulated using optimized component concentrations following established protocols(Sun et al., 2013). Each reaction mixture contained ClearColi cell lysate at a final protein concentration of 10 mg/mL, supplemented with 2% PEG8000 as a macromolecular crowding agent and 10 mM maltose as an energy source. The reaction buffer included 50 mM HEPES (pH 8) and was enriched with a complete amino acid mixture (1.5 mM each), nucleotides (4.8 nM NTP mix containing 1.5 mM each of ATP and GTP, 0.9 mM each of CTP and UTP, pH adjusted to 7.5 with KOH), and 0.2 mg/mL tRNA. Additional cofactors and energy regeneration components were incorporated: 0.26 mM coenzyme A, 0.33 mM NAD^+, 0.75 mM cyclic AMP (cAMP), 0.068 mM folinic acid, 1 mM spermidine, and 30 mM 3-phosphoglyceric acid (3PGA) as the primary energy source. Magnesium and potassium ion concentrations were individually calculated and optimized for each reaction batch. Cell free samples were careful to only undergo a single freeze-thaw cycle when they were used to maximize yields.

Validation of Protein Production

To express endotoxin free cytokine mimics for experiments in Chapter 4, I mixed 40 µL of ClearColi extract with 80 µL of energy premix solution. This solution was divided into 40 µL aliquots for 3 separate reactions, where I added 15 µL of linear cytokine mimic DNA. All reactions were conducted at 29°C under controlled conditions to ensure reproducible protein synthesis.

After overnight incubation, the samples were purified using Ni-NTA affinity chromatography as described previously in the chapter. To confirm expression, I ran an SDS-PAGE gel on the purified proteins:

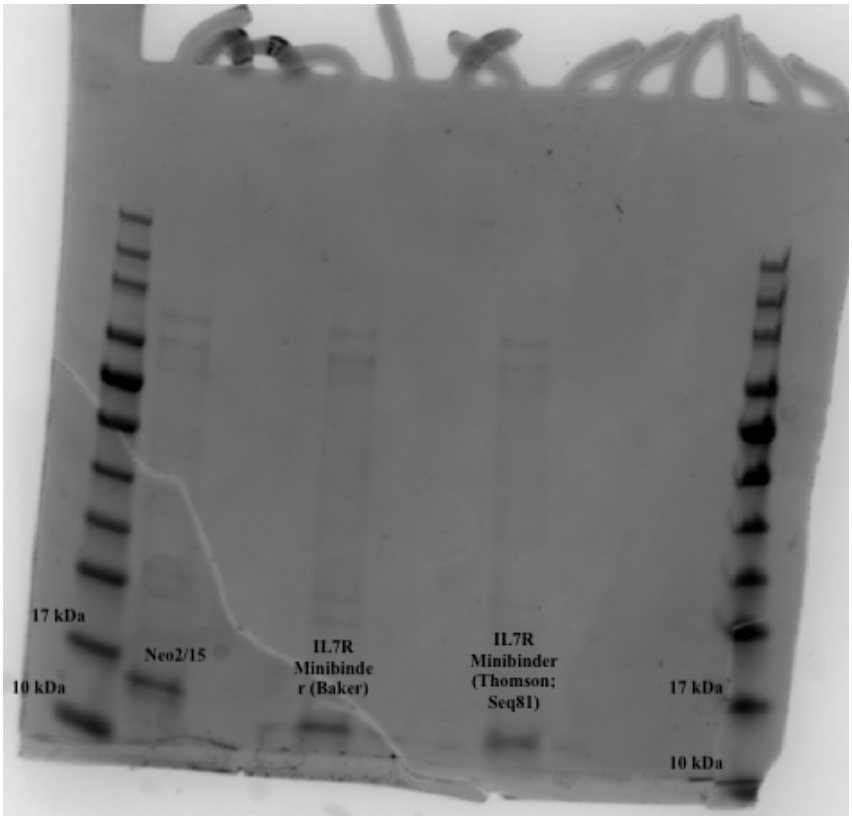


Figure 7.1: SDS PAGE gel of proteins produced from ClearColi extracts. From left to right - Neo2/15, an IL-2 mimic (Silva et al., 2019), parent IL7RA minibinder (Cao et al., 2022), and NV1

Given that bands are visible close to the expected sizes of 11 kDa and 8 kDa for Neo2/15 and the IL7RA minibinders, respectively, I concluded that the computationally designed small proteins can be successfully expressed and purified with high yield and purity. Importantly, the streamlined production protocol—from initial expression to final purification—was completed within a 24 hour timeframe and can be scaled to tens or hundreds of variants, dramatically reducing production time and increasing throughput compared to conventional methods.

7.4 Cell Free Production of Infectious T7 variants

Bacteriophage T7 is a well-characterized lytic phage that infects E. coli. The T7 virion consists of an icosahedral capsid containing the 40 kb double-stranded DNA genome, a short tail, and tail fibers. The capsid is primarily composed of the major capsid protein gp10, with approximately 415 copies per phage particle (Dunn, Studier, and Gottesman, 2013).

Interestingly, the capsid contains two forms of the major capsid protein: gp10A and

gp10B. The gp10B form results from a -1 programmed translational frameshift that occurs at approximately 10% frequency during translation of gene 10 (Condron, Atkins, and Gesteland, 1991). This frameshift produces a C-terminal region in gp10B that differs from the full-length gp10A protein. The natural variation at the C-terminus of gp10 makes this region particularly tolerant to modifications without disrupting capsid assembly.

This tolerance for variation, combined with the high copy number of gp10 proteins in each phage particle (415 copies), makes the T7 capsid an attractive platform for protein display applications. Researchers can fuse foreign peptides or even entire proteins to the C-terminus of gp10 to create phage display libraries, enabling applications ranging from antibody discovery to vaccine development (Rosenberg et al., 1996).

However, a common problem with fusing protein to the C-terminus of gp10 is the toxicity of the resulting fusion protein. Disruption of the capsid can significantly reduce assembly and, thus, infectivity of the resulting virion. The T7Select system from Novagen addresses this issue by using specialized vectors that reduce the copy number of fusion proteins on the phage capsid. For example, the T7Select 10-3 vector displays only 5-15 copies of the fusion protein per phage particle instead of 415 due to a mutated promoter region upstream of the T7 gene 10. The phage is viable and displays limited amounts of fusion protein by growing the phage in a complementing host strain (BLT5403) that provides wild-type gp10A protein from a plasmid. The resulting phage capsids contain mostly unmodified capsid proteins, with only a small number of fusion proteins, maintaining phage viability while still allowing for effective display of the target protein (Krumpe and Mori, 2004).

A key advantage of T7 phage over filamentous phages like M13 is its assembly mechanism. T7 phage particles are assembled entirely within the bacterial cytoplasm and released through cell lysis, rather than through secretion. In contrast, M13 phage, traditionally used for phage display, assembles in the bacterial periplasm, requiring all displayed proteins to be secreted through the inner membrane, a significant limitation for displaying many cytoplasmic proteins. The cytoplasmic assembly of T7 phage makes it particularly suitable for displaying proteins that cannot be efficiently secreted or that might misfold during membrane translocation. This feature allows T7 phage display systems to accommodate a wider range of protein types, especially larger proteins and those with complex folding requirements. Additionally, T7 phage replicates more rapidly than M13, with plaques forming within 3 hours at

37°C, which significantly decreases the time needed for multiple rounds of selection (Pande, Szewczyk, and Grover, 2010).

Another remarkable advantage of T7 phage is its ability to be synthesized in cell-free systems, a capability not shared by M13 phage. Cell-free bacteriophage synthesis allows complete T7 phage assembly *in vitro* using bacterial lysates containing transcription-translation machinery combined with the T7 genomic DNA as a template. This approach can produce up to 10^{11} infectious T7 phage particles per milliliter within hours (Levrier et al., 2024). The lytic, cytoplasmic assembly nature of T7 makes this possible, whereas M13 phage requires intact cellular membrane structures for assembly due to its secretion-dependent lifecycle. This cell-free capability enables rapid engineering, production, and selection of T7 phage variants without requiring live bacterial cultures, offering significant advantages for protein engineering.

I replicated existing protocols for producing wild-type T7 bacteriophage in E. coli cell free lysates. This involves adding in at least 50 ng of wild-type T7 bacteriophage genome to a cell free reaction in a total volume of 10 µL. This was incubated at 29ºC for at least 6 hours to allow for the phage to form, followed by a plaque assay to check for production of infectious phage.

This was typically done by adding 0.4 mL of E. coli grown to mid-log phase and adding them to 4 mL of top agar (LB broth with 0.8% agar) kept at 48ºC before adding the combined mixture to a petri dish. Once the agar solidifies, 10-20 µL of sample is spotted on the plate and allowed to dry. The plate is then incubated at 37ºC for several hours until the agar turns cloudy, indicating E. coli growth. Regions without growth represent areas with infectious phage.

I observed that the transition from no phage production to phage production in cell free is highly non-linear. Less than 50 ng of wild-type genome leads to no plaques observed, while 50 ng or even slightly greater leads to titers upward of 10^9 plaque forming units (PFU) / mL. Note that these lower reported titers are likely due to withholding addition of dNTPs, which if added can support additional T7 genome replication. This extreme nonlinearity likely reflects cooperativity in assembly and is supported by the fact that addition of crowding agents to the cell free mixture, such as polyethylene glycol (molecular weight 8000 g/mol), enhances the yield of infectious T7 bacteriophage in lysate (Levrier et al., 2024).

Given the previously listed constraints around displaying variants as part of capsid

fusions in T7 bacteriophage, I first devised a method to produce viable T7Select 10-3, specifically T7Select 10-3b, phage within a cell free system. Addition of even high amounts (500 ng) of the T7Select 10-3b genome does not lead to plaque formation. I designed and ordered primers to amplify the fragment of the wild-type genome encoding gene 10 as well as regions directly upstream and downstream to fully capture the native promoter, ribosome binding site, and terminator regions. Addition of 100 ng of this fragment along with at least 160 ng of T7Select 10-3b led to plaques that produce infectious T7Select 10-3b when spotted on plates containing complementing host strain (BLT5403) that provides wild-type gp10A protein from a plasmid. We build on this result to construct libraries of capsid-displayed variants in the next subsection.

7.5 Assembling Libraries of Cell-free Produced Phages

Cell free rebooting of phage can presumably result in more well-balanced libraries due to the lack of multiple rounds of infection and replication inherent in in vivo propagation (Levrier et al., 2024). This combined with a simplified, faster phage production workflow amenable to automation combined with high phage titers that result makes cell free production of phage libraries immensely attractive for downstream assay with the Protein CREATE platform (Kristensen et al., 2024; Higashi et al., 2024).

A concern with any high-throughput library preparation method for display is that the functional variant screened (the "phenotype") matches the genetically encoded sequence measured (the "genotype") (Botta, Chemiakine, and Gennarino, 2022). Preserving this genotype to phenotype linkage is essential if the assay is to give any meaningful data readout (T. Smith, Heger, and Sudbery, 2022; Guyer, Laustsen, and Ledsgaard, 2024). Because individual phages infect individual cells in vivo, this linkage is automatically preserved during normal phage propagation in E. coli (Rosenberg et al., 1996; Krumpe and Mori, 2004). In contrast, if the cell free lysate were a well-mixed, dilute solution, fusion capsid produced by one variant would be as likely to be assembled with any phage genome in solution as it were to with the genome it was transcribed and translated from (Levrier et al., 2024). If true, this fact alone poses a significant challenge to the construction of viable phage libraries in a cell free lysate mixture.

A key finding is that phage gene expression and assembly does not behave like one would expect in a well-mixed solution (Levrier et al., 2024). Levrier et al (2024) re-

port that "the fast kinetics of T7 phage coat and tail proteins' cooperative assembly to encompass the phage genome, following their coupled transcription and translation, may limit their diffusion and cross-binding to non-self-phage genome in the viscous transcription-translation (TXTL) mix, leading to genotype to phenotype linkage." The scientists tested this linkage by co-expressing two different T7 phage variants (one with a tail fiber mutation allowing infection of E. coli strains containing the smallest form of rough lipopolysaccharide (ReLPS)) in the same TXTL reaction. Through a series of experiments, they demonstrated that the resulting phage populations showed a much higher proportion of "pure" genotype-phenotype coupling than would be expected from random assembly. The fraction of phages with matching genomes and proteins was orders of magnitude greater than predicted by a randomized hypothesis, indicating significant coupling in the cell-free system (Levrier et al., 2024). This genotype-phenotype linkage is particularly valuable for selecting phages with altered specificity, such as those with the ability to infect bacteria with modified lipopolysaccharides. The researchers successfully used PHEIGES to rapidly identify tail fiber mutations that enable T7 phages to infect ReLPS E. coli strains that are normally resistant to wild-type T7.

This finding that tail fiber mutants can be produced and screened in bulk cell free while preserving the genotype to phenotype linkage inspired me to attempt screening of libraries of phage capsid mutants. This requires solving both the technical challenges of constructing and assembling the phage libraries and then screening them using the Protein CREATE platform.

Design of Library Constructs

Our library constructs incorporate two key elements:

1. A variable region encoding diverse capsid fusion proteins.
2. A downstream barcoded region that replaces the unique molecular identifier (UMI) used in traditional approaches. Because the phages no longer replicate (no dNTPs are added to the reaction mixture), these barcodes should be unique to each phage. This greatly simplifies the downstream processing of the Protein CREATE assay collapsing the two PCR steps into a single one.

Given the large size of the T7 bacteriophage genome, 36,249 bp for T7Select 10-3b, library construction can be a challenge. Traditional assembly methods, such as

Gibson and Golden Gate Assembly, are not compatible with TXTL due to buffer chemistry conflicts. Levrier et al (2024) solve this problem by making use of an assembly method that joins homologous ends of fragments after digestion with exonuclease III during a 1 minute incubation at 75ºC which concurrently deactivates the enzyme (Levrier et al., 2024; Nozaki, 2022). The product of this reaction can go directly into the TXTL reaction without any purification needed and has been used to construct high titers of largely monocultures of phages with one or multiple edits. I wanted to see if this approach could extend toward assembling barcoded libraries of phage.

To do so, I broke up the T7Select 10-3b genome into a set of six fragments in contrast to the four used for wild-type phage assembly in Levrier et al (2024). The two extra fragments are required due to the fact that the library must be inserted within the third fragment (Fragment C) that contains the capsid gene. The primers used to generate fragments T7A, T7B, and T7D remain unchanged as described in Levrier et al (2024), while the additional three fragments are described below:

1. A shortened fragment T7C_1 spanning the original start of T7C described in Levrier et al (2024), but ending at the C-terminus of the phage capsid.
2. The library to be assembled into the phage using primers to amplify Protein CREATE libraries as described in previous chapters.
3. An additional fragment spanning the end of the T7C fragment described in Levrier et al (2024), but with a string of 15 random bases incorporated into the forward primer to serve as the unique barcode.

Each of the fragments is amplified using polymerase chain reaction (PCR) using the T7Select 10-3b or the ordered oligo pool as a template and purified using a PCR cleanup kit. To increase DNA concentration, the fragments may be spun in a vacuum centrifuge. The fragments are then mixed into an equimolar ratio (final concentration of ~9 nM) and treated with the exonuclease buffer for 1 minute at 75ºC followed by incubation at room temperature for 5 minutes before addition to the cell free lysate (myTXTL Pro Master Mix, Arbor Biosciences) alongside 100 ng of linear DNA expressing helper capsid as previously described in a further section to form viable, infectious phage for screens. Plaque assays and subsequent sequencing (described in more detail below) confirm assembly of viable phages containing the desired inserts.

Sequencing of Library Constructs

During the process of library construction, I also implemented a next generation sequencing pipeline using Oxford Nanopore to help debug challenges encountered during assembly protocol optimization (Logsdon, Vollger, and Eichler, 2020). While traditional Illumina sequencing offers high accuracy with error rates below 1%, it requires extensive library preparation including adapter ligation, size selection, and often PCR amplification while being limited to short DNA fragments (T. Smith, Heger, and Sudbery, 2022; Fu et al., 2011). In contrast, Oxford Nanopore technology provided critical advantages for troubleshooting the assembly process: it allowed direct sequencing of native DNA without adaptors, accommodated the full-length phage genome fragments without size limitations, and delivered real-time results with minimal sample preparation. This streamlined approach enabled rapid identification of assembly junction errors and verification of correct constructs, significantly accelerating the iterative optimization process that would have been severely constrained by Illumina's shorter read lengths and more complex preparation requirements. Additionally, the ability to reuse the Nanopore MinION flow cell multiple times for amplicon analysis made the debugging cheap.

The Nanopore library preparation protocol began with end-repair and A-tailing of purified PCR products. For each sample, 11.5 µL of DNA was combined with 1.75 µL Ultra II End-prep Reaction Buffer and 0.75 µL Ultra II End-prep Enzyme Mix, then incubated at 20°C for 5 minutes followed by 65°C for 5 minutes. After AMPure XP bead purification, samples were barcoded using the Native Barcoding kit (barcodes NB1-NB24) to enable multiplexing. For each sample, 7.5 µL of end-prepped DNA was ligated to 2.5 µL native barcode using 10 µL Blunt/TA Ligase Master Mix at room temperature for 20 minutes. This barcoding strategy allowed simultaneous analysis of multiple samples, significantly reducing sequencing time and cost.

Barcoded samples were pooled and adapter-ligated using 5 µL native adapter, 10 µL NEBNext Quick Ligation buffer, and 5 µL Quick T4 DNA Ligase for 20 minutes at room temperature. Following purification with AMPure beads and Short Fragment Buffer washes, the final library was eluted in 15 µL Elution Buffer. For sequencing, the MinION flow cell was primed with a mixture of Flow Cell Flush buffer, BSA, and Flow Cell Tether, followed by loading of the sequencing library prepared with 37.5 µL Sequencing Buffer, 25.5 µL Library Beads, and 12 µL DNA library. This optimized loading protocol maximized sequencing yield while minimizing pore

oversaturation.

The real-time nature of Nanopore sequencing enabled observation of results as they generated, allowing immediate assessment of sample quality and composition. This approach proved invaluable for rapidly detecting mixed populations in phage libraries and identifying synthesis biases.

The Protein CREATE assay was also modified to work with an input of cell-free produced phage and a Nanopore readout. The sample prep was as follows:

1. I first isolated packaged phage DNA from TXTL reactions by DNase digestion to eliminate any unpackaged DNA fragments. For each sample, 10 µL of TXTL product was diluted with 34 µL PBS, then treated with 1 µL DNase I and 5 µL 10× DNase buffer for 30 minutes at 37°C. This critical step ensured only packaged phage genomes remained for downstream analysis, providing an accurate representation of successfully assembled constructs.

2. Following DNase treatment, I employed PCR amplification with primers targeting specific regions of interest and sometimes indices used for demultiplexing. Reactions contained DNase-treated samples, 0.2 µM of each primer, 25 µL 2× GxL PrimeStar Master Mix, and nuclease-free water to 50 µL. This approach allowed for selective amplification of regions crucial for verifying correct library assembly.

3. The DNA samples were quantified and used for Nanopore-specific library preparation as described above.

After sequencing, custom python scripts were used to look for regions of interest within the inserts — the specific regions varied based on the identity of the fragment and the purpose of the sequencing. For instance, for verifying proper assembly of phages, I checked for and extracted the unique barcodes and the identities of the library-displayed variants to convert raw reads into variant counts. If a binding assay was performed, these values can be used in a similar fashion to calculate a pseudo-dissociation constant as is done in the original Protein CREATE assay.

While I am still in the process of further optimizing this version of the Protein CREATE assay, I have already detected more than 15,000 unique variants and have detected enrichment for a mock library consisting of a variant of phage displaying a binder for interleukin 7 receptor alpha (IL7RA) diluted in a library of phages

containing an empty vector that should not bind to the target. As the protocol is further refined, it will be packaged into an automated format that will enable the validation of design panels against a wide array of purified targets, generating the promised fuel for the training engine.

7.6 Conclusion

Cell-free protein production systems offer powerful advantages for protein engineering and library screening applications. This chapter has demonstrated several innovations in leveraging these systems for efficient protein production and high-throughput library construction:

1. **Optimized expression constructs**: By combining carefully selected regulatory elements, we developed a standard expression cassette that enables robust protein production in cell-free systems, facilitating rapid testing of individual protein variants.

2. **Endotoxin-free protein production**: Through collaboration, we developed ClearColi-derived cell-free extracts that maintain high protein yields while dramatically reducing endotoxin levels, enabling applications requiring endotoxin-free preparations.

3. **Cell-free phage assembly**: We established protocols for producing functional T7 phage in cell-free systems, including the more challenging T7Select display system, opening new possibilities for rapid phage engineering.

4. **Library assembly and barcode integration**: By adapting the PHEIGES approach, we developed methods to create barcoded libraries of phage display variants, streamlining downstream Protein CREATE assays.

5. **Nanopore sequencing integration**: Implementing Oxford Nanopore technology provided critical advantages for troubleshooting assembly processes and characterizing library diversity.

These advances collectively contribute to a more efficient protein engineering pipeline, enabling rapid iterations of the design-build-test cycle that drives protein engineering. The ability to produce and screen thousands of variants in parallel, with integrated barcode tracking and simplified workflow, represents a significant step toward the goal of closed-loop protein design .

The combination of cell-free production with the Protein CREATE platform creates a powerful system for generating the experimental data needed to train and refine computational protein design models. As these methods continue to be optimized and automated, they will increasingly enable the data-driven approach to protein design outlined in the earlier chapters of this thesis.

BIBLIOGRAPHY

Abraham, M J et al. (2015). "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers". In: *SoftwareX* 1–2, pp. 19–25. DOI: 10.1016/j.softx.2015.06.001.

Abramson, Josh et al. (May 2024). "Accurate structure prediction of biomolecular interactions with AlphaFold 3". In: *Nature* 630, pp. 493–500. DOI: 10.1038/s41586-024-07487-w.

Agency for Toxic Substances and Disease Registry (Jan. 2024). *Potential health effects of PFAS chemicals*. URL: https://www.atsdr.cdc.gov/pfas/health-effects/index.html.

Assadi-Porter, Fariba M et al. (2010). "Key amino acid residues involved in multi-point binding interactions between brazzein, a sweet protein, and the T1R2-T1R3 human sweet receptor". In: *Journal of Molecular Biology* 398.4, pp. 584–599. DOI: 10.1016/j.jmb.2010.03.017.

Azzaz, Fodil et al. (2022). "The Epigenetic Dimension of Protein Structure Is an Intrinsic Weakness of the AlphaFold Program". In: *Biomolecules* 12.10, p. 1527. DOI: 10.3390/biom12101527.

Baek, Minkyung et al. (2021). "Accurate prediction of protein structures and interactions using a three-track neural network". In: *Science* 373.6557, pp. 871–876. DOI: 10.1126/science.abj8754.

Barata, João T., Scott K. Durum, and Benedict Seddon (2019). "Flip the coin: IL-7 and IL-7R in health and disease". In: *Nature Reviews Immunology* 19.11, pp. 676–693. DOI: 10.1038/s41577-019-0168-5.

Belarif, Lyssia et al. (2018). "IL-7 receptor blockade blunts antigen-specific memory T cell responses and chronic inflammation in primates". In: *Nature Communications* 9.1, p. 4483. DOI: 10.1038/s41467-018-06804-y.

Belloir, Christine et al. (2021). "Biophysical and functional characterization of the human TAS1R2 sweet taste receptor overexpressed in a HEK293S inducible cell line". In: *Scientific Reports* 11.1, p. 22238. DOI: 10.1038/s41598-021-01731-3.

Bennett, Nathaniel R et al. (2023). "Improving de novo protein binder design with deep learning". In: *Nature Communications* 14.1, p. 3602. DOI: 10.1038/s41467-023-38328-5.

Birnbaum, Linda S et al. (2023). "Environmental per-and polyfluoroalkyl substances: An introduction to their classification and their impact on human health". In: *Andrology* 11.5, pp. 850–860. DOI: 10.1111/andr.13428.

Botta, Salvatore, Alexei Chemiakine, and Vincenzo A Gennarino (2022). "Dual antibody strategy for high-resolution imaging of murine Purkinje cells and their dendrites across multiple layers". In: *STAR Protocols* 3.2, p. 101545. DOI: 10.1016/j.xpro.2022.101545.

Bryant, Patrick, Gabriele Pozzati, and Arne Elofsson (2022). "Improved prediction of protein-protein interactions using AlphaFold2". In: *Nature communications* 13.1, p. 1265. DOI: 10.1038/s41467-022-28865-w.

Caldwell, JE et al. (1998). "Solution structure and sweetness determinants of brazzein, a small, heat-stable, sweet-tasting protein." In: *Nature Structural and Molecular Biology* 5.6, pp. 427–431. DOI: 10.1038/nsb0698-427.

Calore, Elisa and Maria Raffaella Petrara (2023). "IL-7/IL-7R axis: From bench to bedside". In: *Journal of Leukocyte Biology* 113.3, pp. 173–189. DOI: 10.1002/JLB.5MR0722-408R.

Cao, Longxing et al. (2022). "Design of protein-binding proteins from the target structure alone". In: *Nature* 604.7907, pp. 541–548. DOI: 10.1038/s41586-022-04654-9.

Carlson, Erik D et al. (2012). "Cell-free protein synthesis: applications come of age". In: *Biotechnology advances* 30.5, pp. 1185–1194. DOI: 10.1016/j.biotechadv.2011.09.016.

Chae, Pil Seok et al. (2010). "Maltose-neopentyl glycol (MNG) amphiphiles for solubilization, stabilization and crystallization of membrane proteins". In: *Nature Methods* 7.12, pp. 1003–1008. DOI: 10.1038/nmeth.1526.

Chandrashekar, Jayaram et al. (2006). "The receptors and cells for mammalian taste". In: *Nature* 444.7117, pp. 288–294. DOI: 10.1038/nature05401.

Condron, Brian G, John F Atkins, and Raymond F Gesteland (1991). "Frameshifting in gene 10 of bacteriophage T7". In: *Journal of Bacteriology* 173.21, pp. 6998–7003. DOI: 10.1128/jb.173.21.6998-7003.1991.

Daughaday, William H. et al. (Nov. 1987). "On the nomenclature of the somatomedins and insulin-like growth factors". In: *Endocrinology* 121.5, pp. 1911–1912. DOI: 10.1210/endo-121-5-1911.

Del Conte, Alessio et al. (July 2024). "RING 4.0: faster residue interaction networks with novel interaction types across over 35,000 different chemical structures". In: *Nucleic Acids Research* 52.W1, W306–W312. DOI: 10.1093/nar/gkae337.

Dunbrack Jr, Roland L (Feb. 2025). "Rēs ipSAE loquunt: What's wrong with AlphaFold's ipTM score and how to fix it". In: *bioRxiv*, p. 2025.02.10.637595. DOI: 10.1101/2025.02.10.637595.

Dunn, James, F William Studier, and Max Gottesman (2013). "Structural and biochemical analyses of bacteriophage T7 tail fiber protein gp17". In: *Journal of Molecular Biology* 425.11, pp. 1979–1988. DOI: 10.1074/jbc.M113.491209.

EFSA Panel on Food Additives and Flavourings (FAF) (2021). "Re-evaluation of thaumatin (E 957) as food additive". In: *EFSA Journal* 19.11, e06884. DOI: 10.2903/j.efsa.2021.6884.

Evich, Marina G. et al. (Feb. 2022). "Per- and polyfluoroalkyl substances in the environment". In: *Science* 375.6580, eabg9065. DOI: 10.1126/science.abg9065. URL: https://www.science.org/doi/10.1126/science.abg9065.

Ferruz, Noelia, Steffen Schmidt, and Birte Höcker (2022). "ProtGPT2 is a deep unsupervised language model for protein design". In: *Nature communications* 13.1, p. 4348. DOI: 10.1038/s41467-022-32007-7.

– (2023). "A Framework for Structure-Based Screening of Protein-Protein Interactions Using AlphaFold". In: *eLife* 12, e98179. DOI: 10.7554/eLife.98179.

Fu, Gregory K et al. (2011). "Counting individual DNA molecules by the stochastic attachment of diverse labels". In: *Proceedings of the National Academy of Sciences* 108.22, pp. 9026–9031. DOI: 10.1073/pnas.1017621108.

Guo, Hao-Bo et al. (Mar. 2023). "Accurate prediction by AlphaFold2 for ligand binding in a reductive dehalogenase and implications for PFAS (per- and polyfluoroalkyl substance) biodegradation". In: *Scientific Reports* 13.1, p. 4082. DOI: 10.1038/s41598-023-30310-x. URL: https://www.nature.com/articles/s41598-023-30310-x.

Guyer, Thomas, Andreas Hougaard Laustsen, and Line Ledsgaard (2024). "Phage display technology and its impact in the discovery of novel protein-based drugs". In: *Expert Opinion on Drug Discovery* 19.5, pp. 455–471. DOI: 10.1080/17460441.2024.2367023.

Hao, S et al. (2024). "Steviol rebaudiosides bind to four different sites of the human sweet taste receptor (T1R2/T1R3) complex explaining confusing experiments". In: *Communications Chemistry* 7.1, p. 236. DOI: 10.1038/s42004-024-01324-x.

Higashi, Katsuaki et al. (2024). "Construction of a T7 phage random peptide library by combining seamless cloning with in vitro translation". In: *The Journal of Biochemistry* 175.1, pp. 85–93. DOI: 10.1093/jb/mvad104.

Honorato, Rodrigo V et al. (2024). "The HADDOCK2.4 web server: A leap forward in integrative modelling of biomolecular complexes". In: *Nature Protocols*. DOI: 10.1038/s41596-024-01011-0.

Hsu, Chloe et al. (2022). "Learning inverse folding from millions of predicted structures". In: *bioRxiv*, pp. 2022–04. DOI: 10.1101/2022.04.10.487779.

Huang, Shan, Luis Antonio Fernández, and Zarath M Summers (2021). "Comparative Analysis of Bacterial Defluorination of Perfluorocarboxylic Acids (PFCAs) and Perfluoroalkyl Sulfonic Acids (PFSAs)". In: *Environmental Science & Technology* 55.22, pp. 15289–15297. DOI: 10.1021/acs.est.1c03874.

Huang, Shan and Peter R. Jaffé (Oct. 2019). "Defluorination of Perfluorooctanoic Acid (PFOA) and Perfluorooctane Sulfonate (PFOS) by *Acidimicrobium* sp. Strain A6". In: *Environmental Science & Technology* 53.19, pp. 11410–11419. DOI: 10.1021/acs.est.9b04047. URL: https://pubs.acs.org/doi/10.1021/acs.est.9b04047.

Islam, Saiful et al. (2014). "Quantitative single-cell RNA-seq with unique molecular identifiers". In: *Nature Methods* 11.2, pp. 163–166. DOI: 10.1038/nmeth.2772.

Izawa, H et al. (1996). "Synthesis of brazzein, a sweet protein from pentadiplandra brazzeana. 1. Preparation of a synthetic precursor, des-(54-Asp)-brazzein by a combination of solid-phase and solution methods". In: *Bioscience, biotechnology, and biochemistry* 60.6, pp. 1018–1020. DOI: 10.1002/(sici)1097-0282(199607)39:1<95::aid-bip10>3.0.co;2-b.

Jaffé, Peter R. et al. (Jan. 2024). "Defluorination of PFAS by *Acidimicrobium sp.* strain A6 and potential applications for remediation". In: *Methods in Enzymology*. Ed. by Randy B. Stockbridge. Vol. 696. Academic Press, pp. 287–320. DOI: 10.1016/bs.mie.2024.01.013.

Jin, Zheyuan et al. (2003). "Monkey Electrophysiological and Human Psychophysical Responses to Mutants of the Sweet Protein Brazzein: Delineating Brazzein Sweetness". In: *Chemical Senses* 28.6, pp. 491–498. DOI: 10.1093/chemse/28.6.491.

Joseph, Jewel Ann et al. (Apr. 2019). "Bioproduction of the Recombinant Sweet Protein Thaumatin: Current State of the Art and Perspectives". In: *Frontiers in Microbiology* 10, p. 695. DOI: 10.3389/fmicb.2019.00695.

Jumper, John et al. (2021). "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873, pp. 583–589. DOI: 10.1038/s41586-021-03819-2.

Kant, Ravi (2005). "Sweet proteins–potential replacement for artificial low calorie sweeteners". In: *Nutrition journal* 4.1, pp. 1–6. DOI: 10.1186/1475-2891-4-5.

Karig, David K et al. (2011). "Expression optimization and synthetic gene networks in cell-free systems". In: *Nucleic Acids Research* 40.8, pp. 3763–3774. DOI: 10.1093/nar/gkr1191.

Kempen, Michel van et al. (Feb. 2024). "Fast and accurate protein structure search with Foldseek". en. In: *Nature Biotechnology* 42.2, pp. 243–246. ISSN: 1546-1696. DOI: 10.1038/s41587-023-01773-0. URL: https://www.nature.com/articles/s41587-023-01773-0 (visited on 10/19/2024).

Kochem, Matthew C, Emily C Hanselman, and Paul A S Breslin (May 2024). "Activation and inhibition of the sweet taste receptor TAS1R2-TAS1R3 differentially affect glucose tolerance in humans". In: *PLoS ONE* 19.5, e0298239. DOI: 10.1371/journal.pone.0298239.

Kristensen, Camilla S et al. (2024). "Cell-free synthesis of infective phages from in vitro assembled phage genomes for efficient phage engineering and production of large phage libraries". In: *Synthetic Biology* 9.1, ysae012. DOI: 10.1093/synbio/ysae012.

Krumpe, Lauren RH and Takeshi Mori (2004). "T7 lytic phage-displayed peptide libraries: construction and diversity characterization". In: *Methods in molecular biology (Clifton, NJ)* 1248, pp. 51–66. DOI: 10.1007/978-1-4939-2020-4_4.

Kwiatkowski, Carol F et al. (2020). "Scientific basis for managing PFAS as a chemical class". In: *Environmental science & technology letters* 7.8, pp. 532–543. DOI: 10.1021/acs.estlett.0c00255.

Laffitte, Anni et al. (2022). "Functional Characterization of the Venus Flytrap Domain of the Human TAS1R2 Sweet Taste Receptor". In: *International Journal of Molecular Sciences* 23.16, p. 9216. DOI: 10.3390/ijms23169216.

Levrier, Arnaud et al. (2024). "PHEIGES: all-cell-free phage synthesis and selection from engineered genomes". In: *Nature Communications* 15.1, p. 2223. DOI: 10.1038/s41467-024-46585-1.

Lin, Zeming, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Allan dos Santos Costa, et al. (2023). "Language models of protein sequences at the scale of evolution enable accurate structure prediction". In: *bioRxiv*, pp. 2022–07. DOI: 10.1101/2022.07.20.500902.

Lin, Zeming, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, et al. (Mar. 2023). "Evolutionary-scale prediction of atomic-level protein structure with a language model". en. In: *Science* 379.6637, pp. 1123–1130. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.ade2574. URL: https://www.science.org/doi/10.1126/science.ade2574.

Logsdon, Glennis A, Mitchell R Vollger, and Evan E Eichler (2020). "Long-read human genome sequencing and its applications". In: *Nature Reviews Genetics* 21.10, pp. 597–614. DOI: 10.1038/s41576-020-0236-x. URL: https://www.nature.com/articles/s41576-020-0236-x.

Madsen, Joshua J et al. (2024). "The structure of human sweetness". In: *Cell* 187.11, pp. 2591–2603. DOI: 10.1016/j.cell.2024.04.019.

Mamat, Uwe et al. (2015). "Detoxifying Escherichia coli for endotoxin-free production of recombinant proteins". In: *Microbial Cell Factories* 14, p. 57. DOI: 10.1186/s12934-015-0241-5.

Meegoda, Jay N. et al. (Dec. 2022). "A Review of PFAS Destruction Technologies". In: *International Journal of Environmental Research and Public Health* 19.24, p. 16397. DOI: 10.3390/ijerph192416397. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC9778349/.

Menting, John G et al. (2013). "How insulin engages its primary binding site on the insulin receptor". In: *Nature* 493.7431, pp. 241–245. DOI: 10.1038/nature11781.

Nakamura, Ryuki et al. (Aug. 2018). "Functional Expression and Characterization of Tetrachloroethene Dehalogenase From Geobacter sp." In: *Frontiers in Microbiology* 9. DOI: 10.3389/fmicb.2018.01774. URL: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2018.01774/full.

Nijkamp, Erik et al. (2022). "ProGen: Language modeling for protein generation". In: *arXiv preprint arXiv:2004.03497*. DOI: 10.1101/2020.03.07.982272.

Nozaki, Shingo (2022). "Rapid and Accurate Assembly of Large DNA Assisted by In Vitro Packaging of Bacteriophage". In: *ACS Synthetic Biology* 11.12, pp. 4113–4122. DOI: 10.1021/acssynbio.2c00419.

Pande, Jyotsna, Magdalena M Szewczyk, and Amit K Grover (2010). "Phage display: concept, innovations, applications and future". In: *Biotechnology advances* 28.6, pp. 849–858. DOI: 10.1016/j.biotechadv.2010.07.004.

Rocklin, Gabriel J et al. (2017). "Global analysis of protein folding using massively parallel design, synthesis, and testing". In: *Science* 357.6347, pp. 168–175. DOI: 10.1126/science.aan0693.

Rosenberg, Alan et al. (1996). "T7 select phage display system: a powerful new protein display system based on bacteriophage T7". In: *inNovations* 6, pp. 1–6.

Silva, Daniel-Adriano et al. (2019). "De novo design of potent and selective mimics of IL-2 and IL-15". In: *Nature* 565.7738, pp. 186–191. DOI: 10.1038/s41586-018-0830-7.

Silverman, Adam D, Ashty S Karim, and Michael C Jewett (2020). "Cell-free gene expression: an expanded repertoire of applications". In: *Nature reviews genetics* 21.3, pp. 151–170. DOI: 10.1038/s41576-019-0186-3.

Singarapu, Kiran K et al. (2016). "Structure-function relationships of brazzein variants with altered interactions with the human sweet taste receptor". In: *Protein Science* 25.3, pp. 711–719. DOI: 10.1002/pro.2870.

Smith, Tom, Andreas Heger, and Ian Sudbery (2022). "Applications of unique molecular identifiers in next-generation sequencing". In: *Human Molecular Genetics* 31.R1, R11–R19. DOI: 10.1093/hmg/ddac097.

Suez, Jotham, Yotam Cohen, et al. (2022). "Personalized microbiome-driven effects of non-nutritive sweeteners on human glucose tolerance". In: *Cell* 185.18, 3307–3328.e19. DOI: 10.1016/j.cell.2022.07.016.

Suez, Jotham, Tal Korem, et al. (2014). "Artificial sweeteners induce glucose intolerance by altering the gut microbiota". In: *Nature* 514.7521, pp. 181–186. DOI: 10.1038/nature13793.

Sun, Zachary Z et al. (2013). "Protocols for Implementing an Escherichia coli Based TX-TL Cell-Free Expression System for Synthetic Biology". In: *Journal of Visualized Experiments* 79, e50762. DOI: 10.3791/50762.

Sutton, Richard S and Andrew G Barto (2018). *Reinforcement Learning: An Introduction*. 2nd ed. Cambridge, MA: MIT Press.

Temussi, Piero A (2009). "Sweet, bitter and umami receptors: a complex relationship". In: *Trends in biochemical sciences* 34.6, pp. 296–302. DOI: 10.1016/j.tibs.2009.02.005.

Thornton, Savannah R et al. (2024). "Applications of cell free protein synthesis in protein design". In: *Protein Science* 33.4, e5148. DOI: 10.1002/pro.5148.

U.S. Environmental Protection Agency (2024). *EPA Finalizes First Ever National Drinking Water Standard for PFAS*. URL: https://www.epa.gov/newsreleases/epa-finalizes-first-ever-national-drinking-water-standard-pfas (visited on 04/10/2024).

Ueda, Takashi et al. (Aug. 2003). "Functional Interaction between T2R Taste Receptors and G-Protein Alpha Subunits Expressed in Taste Receptor Cells". In: *The Journal of Neuroscience* 23.19, pp. 7376–7380. DOI: 10.1523/JNEUROSCI.23-19-07376.2003.

Viola, Cristina M. et al. (Oct. 2023). "Structural conservation of insulin/IGF signalling axis at the insulin receptors level in Drosophila and humans". en. In: *Nature Communications* 14.1, p. 6271. ISSN: 2041-1723. DOI: 10.1038/s41467-023-41862-x. URL: https://www.nature.com/articles/s41467-023-41862-x (visited on 11/27/2024).

Wackett, Lawrence P. (Oct. 2021). "Why Is the Biodegradation of Polyfluorinated Compounds So Rare?" In: *mSphere* 6.5. Ed. by Katherine McMahon, e00721–21. DOI: 10.1128/mSphere.00721-21. URL: https://journals.asm.org/doi/10.1128/mSphere.00721-21.

Wallner, Björn et al. (2023). "Improved multimer prediction using massive sampling with AlphaFold in CASP15". In: *Proteins: Structure, Function, and Bioinformatics* 91.12, pp. 1608–1619. DOI: 10.1002/prot.26562.

Walters, D Eric et al. (2009). "Design and Evaluation of New Analogs of the Sweet Protein Brazzein". In: *Chemical Senses* 34.8, pp. 679–690. DOI: 10.1093/chemse/bjp048.

Washington, John W et al. (2020). "Degradation of perfluoroalkyl acids by a combined photocatalytic and electrochemical approach". In: *Environmental Science & Technology Letters* 7.5, pp. 351–357. DOI: 10.1021/acs.estlett.0c00052.

Weiss, Michael A. (2009). "The Structure and Function of Insulin: Decoding the TR Transition". en. In: *Vitamins and Hormones* 80, p. 33. DOI: 10.1016/S0083-6729(08)00602-X. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC3297421/ (visited on 11/27/2024).

Wohlwend, Jeremy et al. (Nov. 2024). “Boltz-1: Democratizing Biomolecular Interaction Modeling”. In: *bioRxiv*. DOI: 10.1101/2024.11.19.624167.

Wu, Zachary et al. (2021). “Protein sequence design with deep generative models”. In: *Current Opinion in Chemical Biology* 65, pp. 18–27. DOI: 10.1016/j.cbpa.2021.04.004.

Yin, Rui et al. (2022). “Benchmarking AlphaFold for protein complex modeling reveals accuracy determinants”. In: *Protein Science* 31.8, e4379. DOI: 10.1002/pro.4379.

Zhang, Feng et al. (2010). “Molecular mechanism of the sweet taste enhancers”. In: *Proceedings of the National Academy of Sciences* 107.10, pp. 4752–4757. DOI: 10.1073/pnas.0911660107.