Generating and Harnessing Learned Embeddings for Protein Design

Sanaa Mansoor

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

David Baker, Chair

Frank DiMaio

Phil Bradley

Program Authorized to Offer Degree:

Molecular Engineering

University of Washington

Abstract

Generating and Harnessing Learned Embeddings for Protein Design

Sanaa Mansoor

Chair of the Supervisory Committee:
Dr. David Baker
Biochemistry

The structure and function of proteins are encoded by their amino acid sequences. The field of protein design aims to uncover the fundamental connection between protein sequence, structure, and function to design novel proteins with important applications in fields such as medicine, biotechnology, and materials science. The complex relationship between protein sequence, structure, and function makes protein design a challenging task. In recent years, learned embeddings have emerged as a powerful tool to help deconvolute this relationship. Learned embeddings can convert high-dimensional protein data, such as protein sequences and structures, into small vectors of biologically relevant information. By capturing all the essential features of a protein in a compact form, embeddings enable the use of machine learning techniques for protein design. My PhD research has focused on generating meaningful learned embeddings of proteins and then harnessing them for various downstream predictions. For studying protein ensembles

and protein structure refinement, I developed embeddings through training generative models on two-dimensional structural data, followed by three-dimensional structural modeling. By incorporating sequence information, a joint representation of protein sequence and structure was developed for predicting the effects of single mutations on protein thermal stability. Finally, following the development and success of an accurate structure prediction model, RoseTTAFold, the embeddings learned from this model were used for "zero-shot" or unsupervised prediction of the effect of point mutations on protein stability and function. These successes demonstrate the importance of using learned protein embeddings for protein design and highlight the need for further research in this area to facilitate the creation of novel proteins with desired properties.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to express my sincerest gratitude to everyone who has supported me during my PhD. First, my supervisor, David Baker, for his invaluable guidance, insightful feedback, and tremendous support throughout this journey. I would like to thank Minkyung Baek for being a constant support, and an exceptional mentor to me for my entire PhD. She has set an amazing example for me, and I thank her for her guidance in every step of this journey. This work would not have been possible without her.

One of the highlights of my PhD was doing an internship with Dr. Eric Horvitz at Microsoft Research, which proved to be such an enriching experience for me. Eric was an incredible mentor and gave me constructive feedback at every step of the project. I would also like to give a massive thanks to my other mentors in the lab, specifically, Hahnbeom Park and Doug Tischer. Thank you for fielding all my questions over the years and for teaching me invaluable skills in machine learning and biochemistry. I would also like to thank the administrative staff at the IPD, especially Luki Goldschmidt for keeping the digs up and running for the entire lab to (mis)use.

I would also like to thank my colleagues in the lab for contributing to valuable discussions around my projects in lab: Justas Dauparas, Ivan Anishchenko, David Juergens, Jue Wang, Gyu-Rie Lee and Sergey Ovchinikov among others. You have all been inspirational scientists and role models for me to follow and I thank you for all your advice over the years.

I would like to thank my family for their continued support over the years and for urging me to take long walks every day. Thank you for also reminding me of how far I have come when

I lose sight of my progress. I would like to thank my brother, Hamid Mansoor, for his endless support and for encouraging me to take the CS101 course back in college which sparked my interest and eventually led me to pursue a career in this field.

Finally, I am so grateful to have met such warm and incredible people at the IPD who have become life-long friends. Adam Chazin-Gray for his infinite support, comfort, and humor. Areeb Shaukat and Meerit Said for their everlasting guidance and for being incredible friends-turned-family. I would also like to thank Sidney Lisanza, Ian Humphreys, Sam Pellock, Chad Miller and Basile Wicky for being tremendous friends along the way. Being around you all has always felt like sunshine in rainy Seattle.

Chapter 1. INTRODUCTION

Proteins are the molecular machines that perform the most critical functions in living organisms. They consist of a sequence of amino acids that spontaneously fold into unique three-dimensional structures to carry out biochemical functions. Understanding this sequence-structure-function relationship is integral in designing new proteins that carry out important and prespecified functions. Advancements in computational protein design have led to remarkable developments in new drug therapies [1], [2], biosensors [3], and small molecule binder proteins [4]. Until recently, most computational methods for understanding and analyzing protein function and dynamics have used either a first principles-based approach involving structural simulations or sequence modeling approaches which identify existing co-evolutionary pressures on proteins.

Rosetta [5], a physics-based protein design software, is guided by an all-atom heuristic energy function that estimates the free energy of a given protein conformation. The energy function is made up of hydrogen bonding, ionic interactions, pairwise inter-atomic terms describing the Van der Waals interactions, as well as solvation and statistical terms. The Rosetta modeling suite still has considerable success in designing proteins since the energy function continues to be refined and reparametrized over the years. Rosetta draws on our foundational understanding of the physics behind protein design and folding. However, it and similar first-principles-based approaches are computationally expensive and require expert domain knowledge to be set up properly.

Another example of a physics-based approach to study protein structure, function, and dynamics is Molecular Dynamics (MD) simulation. MD software, such as GROMACS [6], have been used to simulate the natural motions of proteins and other biomolecules in an all-atom setting. These simulations can capture multiple conformational states that a protein can adopt. Capturing these different conformations is valuable in understanding the functional mechanism of a protein and performing fine-grained structure prediction. This approach, like Rosetta, also requires extensive computational time and expert domain knowledge to set up and interpret.

Statistical sequence modeling has long been used to study protein structure and function. Through this method, conserved regions and motifs of protein sequences are identified which implies conserved function. Sequence based algorithms have used *k*-mer counts, calculated amino acid composition, and predicted secondary structure [7]. Previous work has also been done in using the covariation between amino acids at pairs of positions in the sequence (co-evolution) to predict the three-dimensional protein structure [8]. However, these methods fail to make use of the growing databases of sequence, structure, and functional information deposited online.

Deep-learning methods that are used to predict three-dimensional protein structures from multiple sequence alignments (MSAs) have gained a lot of attention recently [9], [10] because of their near experimental-level accuracy. The use of these models can be extended and adapted for protein design and protein function prediction through activation maximization and unsupervised techniques [11], [12]. AlphaFold and RoseTTAFold are trained on MSAs that have many protein

sequences that are similar enough to be aligned and diverse enough to contain distant coevolutionary information. These models learn on the MSA and structural template information to form an embedding or representation of proteins that is used for the final three-dimensional structure prediction output.

Following the success of large-scale models in the field of Natural Language Processing (NLP) [13]–[15], active research is ongoing in developing deep-learning models for protein representation learning. Raw protein data is transformed into vector representations with the assumption that the functional information is encoded in the input features. The main objective of representation learning is to preserve the semantic similarity between the data points as a function of distance in the vector embedding space. High dimensional data such as protein distance maps or protein structural domains can be converted to low-dimensional representations using methods such as Principal Component Analysis (PCA) [16] or t-SNE [17] . This learned low-dimensional representation can then be exploited to navigate search and sampling of specific features desired in the output of the model. Efficiently limiting the search space of protein data is a hard problem due to the immense size of the conformational states of a single protein. Recent efforts have been made into employing deep learning-based methods for this problem.

To learn a meaningful and low-dimensional representation of the three-dimensional structure of a protein, it can be represented as two-dimensional pairwise distance map between all backbone atoms. This distance map was used to train a Generative Adversarial Network (GAN) to generate fixed-length full-atom protein backbones through a learned embedding [18]. Three-dimensional coordinates have been used as input to a Graph Convolutional Network

(GCN) to generate an embedding for use in protein function prediction [19]. However, the relatively small number of structurally validated proteins, as compared to the expansive set of sequences available, motivates the focus to date on protein sequences for generating useful embeddings and representations for downstream tasks such as structure and function prediction.

Sequence embeddings created via semi-supervised training have demonstrated strong performance over a broad range of biologically relevant downstream tasks [20], [21]. Through the semi-supervised training objective, these models capture long range dependencies between unrelated families of proteins. Early contextualized sequence embedding models included ELMO which uses representations from the hidden states of bi-directional LSTMs [22]. This model was then applied to supervised-training tasks such as prediction of subcellular localization or structure prediction [23]. More recently, semi-supervised models trained on the huge amounts of sequence data available have achieved state-of-the-art performance on a wide variety of benchmark datasets such as protein contact map prediction and function prediction [24], [25]. This early work on protein language models demonstrated the power and potential that these methods would have for protein design.

These language models also learn very informative embeddings that have been used to better capture coevolutionary information and can link sequence to function through transfer learning [13], [15], [24]–[26]. These models continue to increase in accuracy with more compute time and data. This motivates the need to add strong biological priors to aid in making these models more data-efficient, possibly through addition of structural features.

Overall, this introduction outlines and motivates the need for better and more interpretable learned protein embeddings for downstream prediction tasks. Here, I will present my efforts to generate, and harness learned protein embeddings using generative models, semi-supervised and unsupervised approaches. First, I will describe my work using a two-dimensional structure-based generative model to study and explore protein structure refinement and K-Ras ensemble generation. I also present a semi-supervised approach to building a joint embedding on both protein sequence and structure that I used for supervised protein thermal stability prediction. Finally, I used an already existing model, RoseTTAFold, for "zero-shot" or unsupervised mutation effect prediction on protein stability and function.

Chapter 2. HARNESSING JOINT REPRESENTATIONS OF SEQUENCE AND STRUCTURE FOR SUPERVISED THERMAL STABILITY PREDICTION

This section contains content previously published as: Mansoor, S., Baek, M., Madan, U., & Horvitz, E. (2021). Toward More General Embeddings for Protein Design: Harnessing Joint Representations of Sequence and Structure. *BioRxiv*, 2021.09.01.458592. DOI: https://doi.org/10.1101/2021.09.01.458592

2.1 ABSTRACT

Protein embeddings learned from aligned sequences have been leveraged in a wide array of tasks in protein understanding and engineering. The sequence embeddings are generated through semi supervised training on millions of sequences with deep neural models defined with hundreds of millions of parameters, and they continue to increase in performance on target tasks with increasing complexity. For this project, we chose to use a more data-efficient approach to encode protein information through joint training on protein sequence and structure in a semi-supervised manner. We show that the method can encode both types of information to form a rich embedding space which can be used for downstream prediction tasks. We show that the incorporation of rich structural information into the context under consideration boosts the performance of the model by predicting the effects of single mutations. We attribute increases in accuracy to the value of leveraging proximity within the enriched representation to identify sequentially and spatially close residues that would be affected by the mutation, using experimentally validated or predicted structures.

2.2 INTRODUCTION

Proteins consist of a linear chain of amino acids that fold to form a three-dimensional structure to carry out vital processes all living organisms. The sequence to structure to function relationship is integral in understanding how to design proteins for specific functions. Most methods that seek to understand the complex sequence to structure to function relationship take either a first principles approach with structural simulations [5] or leverage sequence embeddings trained through an adaptation of semi-supervised machine learning methods used to construct large-scale neural models developed for natural language processing (NLP) tasks [13], [15], [24], [27].

Sequence embeddings created via semi-supervised training have demonstrated strong performance over a broad range of biologically relevant downstream tasks, particularly in the realm of protein engineering [20], [21]. Through the semi-supervised training objective, these models capture long-range dependencies between unrelated families of proteins. Protein structure is more informative for predicting function than sequence [28]. However, because of the cost and time needed to experimentally validate protein structures, there is a relatively small database of them publicly available. In contrast, as the cost of sequencing continues to decrease, the amount of sequence data generated grows and thus motivating most machine learning research to be focused on using this growing data for protein function prediction tasks. Language models centered on semi-supervised training on sequences continue to grow and become increasingly accurate with larger architectures, more compute time, and more data. In this project, I explored an alternate approach, promising greater data-efficiency via the explicit