FoldAlchemy: Protein Structures to Evolutionary Insights

Biswajit Banerjee[1], D. Eric Smith [1,2,3], Loren Dean Williams [1,2], Anton S. Petrov [1,2]
[1] School of Biology, Georgia Institute of Technology, Atlanta, GA 30332
[2] School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta, GA 30332
[3] Earth-Life Science Institute, Tokyo Institute of Technology, Meguro-ku, Tokyo, Japan, 〒152-8550

Introduction

At the advent of life, the establishment of the Central Dogma and the emergence of non-equilibrium metabolism were inextricably linked. This linkage is best symbolized by adenosine triphosphate (ATP), and ribonucleic acid (RNA) building block, an activator of amino acids in protein synthesis, and the primary metabolic energy currency. The Central Dogma proviso that RNA makes protein in the ribosome (a ribozyme), and that protein makes RNA in polymerases (protein enzymes), demonstrates deep integration and supports co-evolution of all biopolymer types and metabolism. Those linkages will be leveraged here to understand the origins of RNA, proteins and metabolism.

Understanding the relationship between protein folds and their catalytic functions is one of the fundamental questions in understanding of the events that took place at the dawn of life. Protein folds are the three-dimensional structures that proteins adopt, which are fundamental to their function. Despite the vast diversity of proteins, there are only a limited number of unique folds, each of which can be associated with specific catalytic activities. Mapping these folds to their corresponding functions can provide significant insights into the principles of protein design and the evolution of metabolic pathways. We hope that the universal and most conserved regions of metabolic networks and their respective enzymes contain a direct record of a very distant past.

Correlation between emergence of new folds and expansion of enzymatic networks (fold gating and fold injection) are critical concepts that intersect with the origin of life research, providing insights into how primitive metabolic systems might have evolved. Fold injection, the process of incorporating new folds into existing metabolic frameworks, offers a perspective on the evolution of metabolic complexity. Fold gating, by injecting new structural folds into primitive metabolic networks, early life forms could diversify their catalytic repertoire, facilitating the development of more sophisticated metabolic pathways. These processes underscore the importance of protein structure in the functionality and evolution of early metabolic systems, highlighting how structural innovations may have driven the transition from simple molecular interactions to complex, life-sustaining biochemical networks. Understanding these mechanisms can shed light on the origin of life, revealing how the first metabolic systems achieved the specificity and adaptability required for life's emergence.

While these are the most predominantly researched topic in fold association to metabolic reaction, in this research, I propose to study the enzymatic space and create a latent space representation of all the folds which associates to their functionality.

Motivation

Evolutionary Insights: Protein folds represent an ancient and conserved aspect of molecular evolution[2]. By mapping folds to their catalytic functions, we can trace the evolutionary history of metabolic pathways, understand how new functions emerge, and identify ancient enzymes that have been preserved across different species.

Functional Annotation: Many proteins with known structures but unknown functions can be better understood through fold to functionality mapping. This approach can help in annotating the functions of newly discovered proteins[3], leading to a more comprehensive understanding of protein folds.

Synthetic Biology: Fold to functionality mapping can assist in the rational design of enzymes with novel functions. By understanding the structural basis of enzyme activity, synthetic biologists can engineer proteins to perform desired reactions, paving the way for innovations in biotechnology and industrial processes.

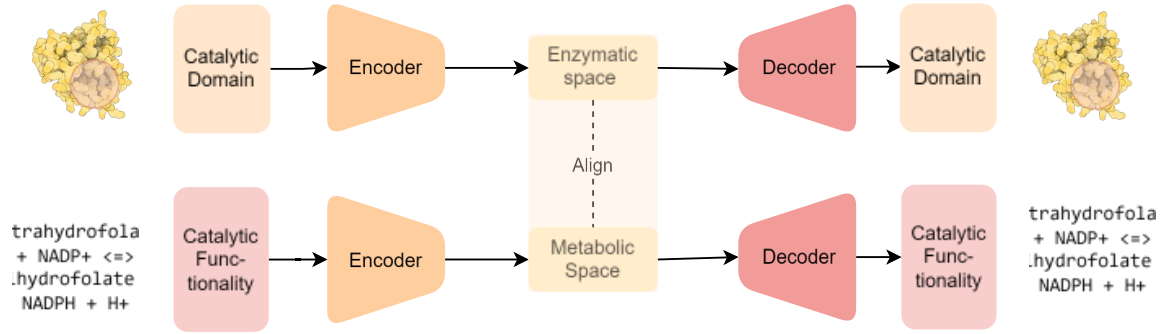

Figure 1: Demonstration of aligning with the help of contrastive training[1] in deep learning for enzyme dihydrofolate reductase (DHFR) and it's catalytic domain highlighted in red, the green highlight represents the latent vector space that represents both the

Aims of proposed work

The primary goal of this project is to create a latent space leveraging the power of encoder and decoder architecture which aligns both the spaces. This latent vector space will not only make both spaces directly searchable but also help generate intermediate folds that were necessary to reach the current generation but went extinct or not currently available in the dataset. Below is the proposed outcome in more elaborated manner:

1. Enzymatic space representation:

I propose representing the enzymatic space as a latent continuous n-dimensional space where intermediate folds can be identified and mapped to their possible functionalities. By developing a model that captures the nature of enzyme folds and their associated activities, we can predict intermediate structures and their potential catalytic roles. This approach will facilitate the discovery of new enzymes and the engineering of existing ones for novel functions. The representation of enzymatic space as a continuous landscape will be a powerful tool for both fundamental research in protein evolution and practical applications in biotechnology and medicine, enabling the rational design of enzymes with tailored properties.

2. Searchable Enzymatic space:

With the help of my previous research, we can easily find enzymes and the catalytic domains associated with a reaction. However, we still lack the capability of searching all the reactions based on the fold that catalyzes them. This proposal aims to develop a comprehensive database that links enzyme folds to their respective catalytic activities, enabling a fold-based search capability. This enhancement will allow researchers to explore the enzymatic space more effectively, identifying potential enzymes based on their structural folds and uncovering new catalytic activities. Such a tool would be invaluable in understanding the structural basis of enzyme function and could accelerate the discovery of novel enzymes for biotechnological applications.

3. Optimize Pathway Search:

Identifying the constraints and preferences that nature follows when establishing biochemical routes is crucial for optimizing pathway searches. My research will focus on elucidating these natural constraints and preferences, as well as identifying alternative biochemical pathways. By understanding the possible evolutionary trajectories and the minimal sets of reactions and conditions necessary for the emergence of biochemical functionality, we can better predict and design metabolic pathways. This knowledge can shed light on the origins of metabolism and guide the engineering of synthetic pathways for industrial and medical applications, ensuring they are both efficient and robust.

4. Conservation of patterns:

Understanding which folds are conserved and finding patterns within these conserved folds is essential for unraveling the evolutionary history of proteins. My research will investigate the conservation of folds and identify patterns that signify evolutionary transitions between highly conserved folds and those that are less conserved. By mapping these transitions, we can gain insights into the structural and functional evolution of proteins. This information will be crucial for evolutionary biology studies and can inform the design of proteins with desired properties by mimicking natural evolutionary processes.

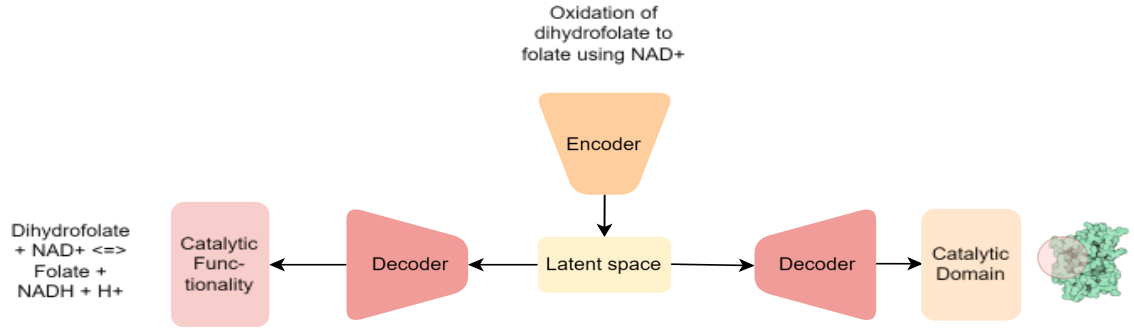

Figure 2: Encoder encoding user query to latent space which is used by decoders to decode as Fold and Functionality for the enzyme

Data Collection and Modeling:

Data: A manually curated, comprehensive collection of databases is available with represents all 102 generations of available compounds, encompassing 5,836 distinct reactions governed by 2,939 unique enzymes across more than 160 metabolic pathways. This extensive dataset provides a detailed snapshot of the intricate biochemical landscape, including catalytic domains and corresponding functions for each enzyme. Also consisting of network expansion information for these reactions and their generations.

Modeling: Leveraging the power of transformer models and Self-Supervised Learning (SSL)[4], I propose to train a hybrid encoder-decoder architecture with cross-attention[5] mechanisms. This contrastive training approach will yield a latent vector space that is both searchable and capable of providing specific information about catalytic domains, enzymes, and their functionalities. One significant advantage of this modality training is its ability to query and predict results for enzyme folds not currently existing in the database, thereby enhancing the scope and applicability of the research.

Pipeline: As illustrated in Figure 2, the proposed pipeline utilizes an advanced encoder-decoder architecture to transform user queries into a latent vector seamlessly aligned with the metabolic and enzymatic space. The encoder effectively encodes the query, creating a latent representation that guides the decoder in identifying the catalytic domain and functional purpose of the enzyme under investigation. The novelty of this approach lies in its ability to handle queries for enzymes and folds do not present in the existing database. By leveraging contrastive training and cross-attention mechanisms, the encoder can generalize beyond known data, enabling the decoder to predict the domains and functionalities of previously uncharacterized enzymes. This innovative method significantly enhances the predictive power and applicability of the search system within the biochemical research domain.

Conclusion

This proposed research aims to create a searchable enzymatic space based on folds, optimize pathway searches by understanding natural constraints and preferences, uncover conservation patterns within protein folds, and represent the enzymatic space as a continuous landscape. These advancements will provide deep insights into the structural and functional evolution of enzymes, enhance our ability to design novel metabolic pathways, and drive forward the fields of synthetic biology and biotechnology. Through this research we will identify basal protein segments, and trace protein evolution back in time, to understand both evolutionary pressures and the primitive forms of proteins. This research will provide insight into the more primitive protein structures that led to the diversity of life we see around us today. This research could be a tiny step forward towards finding the origin of life from the view of catalytic protein folds and how they contributed to the evolution and sustenance of life.

References

1. Liu, S., et al., *Multi-modal molecule structure–text model for text-based retrieval and editing*. Nature Machine Intelligence, 2023. **5**(12): p. 1447-1457.
2. Longo, L.M. and M. Blaber, *Protein design at the interface of the pre-biotic and biotic worlds*. Archives of biochemistry and biophysics, 2012. **526**(1): p. 16-21.
3. Van den Broeck, L., et al., *Functional annotation of proteins for signaling network inference in non-model species*. Nature Communications, 2023. **14**(1): p. 4654.
4. Balestriero, R., et al. *A Cookbook of Self-Supervised Learning*. in *International Conference on Learning Representations*. Transactions of Machine Learning Research.
5. Vaswani, A., et al., *Attention is all you need*. Advances in neural information processing systems, 2017. **30**.