EVOSEQ-ML: ADVANCING DATA-CENTRIC MACHINE LEARNING WITH EVOLUTIONARY-INFORMED PROTEIN SEQUENCE REPRESENTATION AND GENERATION

Mehrsa Mardikoraem, Daniel Woldring
Chemical Engineering & Material Science, Michigan State University
Michigan, MI 48823, USA

ABSTRACT

In the rapidly evolving field of protein engineering, embracing advancements in machine learning (ML) has led to significant achievements, such as predicting protein structures (e.g., AlphaFold), representing protein sequences with language models (i.e., embeddings), and generating functional proteins from scratch. Despite these advances, the importance of data curation on ML model performance has not been thoroughly investigated. As we gather more sequence and structural data, evidence increasingly supports data-centric over model-centric approaches. Thus, our ML training strategy should prioritize high-quality, domain-specific data rather than focusing solely on model improvements. Examining the evolutionary trajectories and the myriad functional adaptations across millions of years reveals a vast, underleveraged dataset for protein engineering. Exploiting evolutionary insights, characterized by hyperstability and extensive functionality can overcome current limitations regarding data quality and quantity in ML frameworks. This paper presents a novel methodology that integrates evolutionary information beyond multiple sequence alignment (MSA) into ML models, setting a new standard for data-centric strategies in the field. Ancestral sequences, obtained by sampling from the probability distributions generated by ASR rather than selecting the single most probable sequence, enable the production of unprecedented scale data— up to billions. Our findings reveal that protein sequences generated by ASR-trained generative ML models can produce high stability and a wide variety of protein sequences. We further introduce family-specific protein representations by this evolutionary data to fine-tune the ESM protein language model and improve downstream classification tasks. The obtained sequence representations improved classification within multiple families tested. Therefore, we underscore the potential of evolutionary data in ML-driven protein engineering by providing datasets that are both extensive in quantity and unmatched in functional quality.

Keywords: Protein Engineering, Generative Models, Language Models, Fine-Tuning, Ancestral Sequence Reconstruction, Data-Centric Models

1 INTRODUCTION

The evolution of machine learning (ML) increasingly emphasizes the use of data-centric approaches. This shift recognizes that data quality and diversity are as vital as algorithm development for effective ML models (Singh, 2023; Adeoye et al., 2023). Unlike model-centric methods that are widely accessible through open-source platforms, data-centric methods offer tailored insights to specific applications which enhances model learning capabilities beyond scoring metrics (Miranda, 2023)(Zha & States, 2023). This trend is evident in the development of the GPT model, which evolved from focusing on architectural improvements to prioritizing data quality and data collection strategies. As a result, focusing on data-centric approaches promises a profound contribution toward ML-driven problem-solving in distinct domains. These approaches have already exhibited potential in fields ranging from finance to healthcare, improving model reliability, scalability, trustworthiness, and generalizability (Liu et al., 2023; Wang et al., 2023). This shift towards data-centric approaches is especially significant in areas demanding intricate analysis, like protein engineering, where such strategies could enhance the understanding and engineering of the complex protein fitness landscape (Romero & Arnold, 2009; Kauffman & Weinberger, 2018).

In protein engineering, the adoption of data-centric approaches is paramount due to the field's unique challenges, such as its complex and rugged fitness landscape, where minor alterations can significantly impact protein functionality and stability. Despite advances in applying ML that have facilitated the discovery of high-fitness proteins, these models struggle with imbalanced datasets and the scarcity of data across protein families. This underscores the necessity of reevaluating our strategies towards curating training data(Romero & Arnold, 2009; Kauffman & Weinberger, 2018; Mena & Daugherty, 2005; Gao et al., 2020; Deznabi et al., 2020; Meier et al., 2021). A focus on enhancing the diversity and quality of datasets is crucial, as it ensures that ML models are trained on data that accurately reflect the intricate biological properties of proteins. Thus, establishing a method that can provide a more effective training dataset for ML models will pave the way for strategically advancing protein engineering campaigns.

Ancestral sequence reconstruction (ASR) offers one possible data-centric approach to protein engineering. ASR utilizes computational techniques to infer ancient protein sequences from modern descendants, thereby enriching our datasets with high-quality, diverse, and stable sequences (Joho et al., 2022; Gumulya & Gillam, 2017; Zaugg et al., 2014; Pauling et al., 1963). Built on the foundation of evolutionary biology and molecular phylogenetics, ASR involves constructing phylogenetic trees using substitution models, a process that maps out evolutionary relationships and predicts ancestral states. ASR's approach to assigning posterior probabilities to amino acids at various positions allows for the exploration of a vast array of sequence combinations. For example, with just 15 positions that each have four high-likelihood amino acids, ASR can generate up to a billion unique sequences (refer to Figure 1). Recent studies across various protein families have highlighted ASR's effectiveness in dealing with statistical uncertainties. These studies demonstrate that sequences generated from ASR predictions, which sample a broad distribution of amino acids, can be functional and in some cases, more functional or offer novel functionalities, compared to sequences based solely on the most likely amino acid predictions. This suggests that relying solely on the maximum likelihood in each position may not always yield the best outcomes (Eick et al., 2017; Bar-Rogovsky et al., 2015). Therefore, with the potential to generate many high-quality sequences, ASR represents a potentially impactful approach to incorporating data-centric strategies in protein engineering.

In this study, we explore how evolutionary data, particularly that gleaned from ASR and ancestral proteins, can be integrated into state-of-the-art ML models to address two primary objectives: (1) the generation of novel protein sequences through generative modeling and (2) the enhancement of protein classification via fine-tuned language models. Our findings demonstrate the untapped potential of evolutionary information in refining and advancing the capabilities of ML models in protein engineering, paving the way for more biologically informed computational strategies.

2 METHOD

2.1 GENERATIVE MODEL FOR NOVEL PROTEIN SEQUENCE GENERATION

We aimed to test whether incorporating evolutionary data into the training of generative models yields protein sequences that are both novel and structurally stable. To this end, we selected the PK2 protein family as our focus and employed a Variational Autoencoder (VAE) as our generative ML model. Subsequent computational analyses (i.e., AlphaFold, FoldX, UMAP visualization) were conducted on the sequences generated by this model to rigorously assess their quality.

2.1.1 DATA PROCESSING

PK2 is an ethylene-forming enzyme (EFE) in which we extracted its evolutionary information using the AP-LASR (VanAntwerp et al., 2023) software(`https://github.com/WoldringLabMSU/AP-LASR`), a tool designed to reconstruct evolutionary information by leveraging the phylogenetic tree of the query protein sequence. This tool has facilitated ASR by fully automating the reconstruction process from multiple sequence alignment by MAFFT (Katoh & Standley, 2013) to tree phylogeny predictions via IQ-Tree2 (Minh et al., 2020). AP-LASR outputs several key datasets: sequences of modern proteins, ancestral proteins (ASR-Max), and near-ancestral proteins (ASR-Dist). The ASR_Dist dataset is created by sampling from a posterior probability distribution generated via the ASR.state file in IQ-Tree. For our project, the threshold for picking amino acids from this distribution was set at 0.2 (Eick et al., 2017; Sennett & Theobald, 2022) to strike an optimal balance between maintaining ancestral properties and sequence diversity (i.e.,

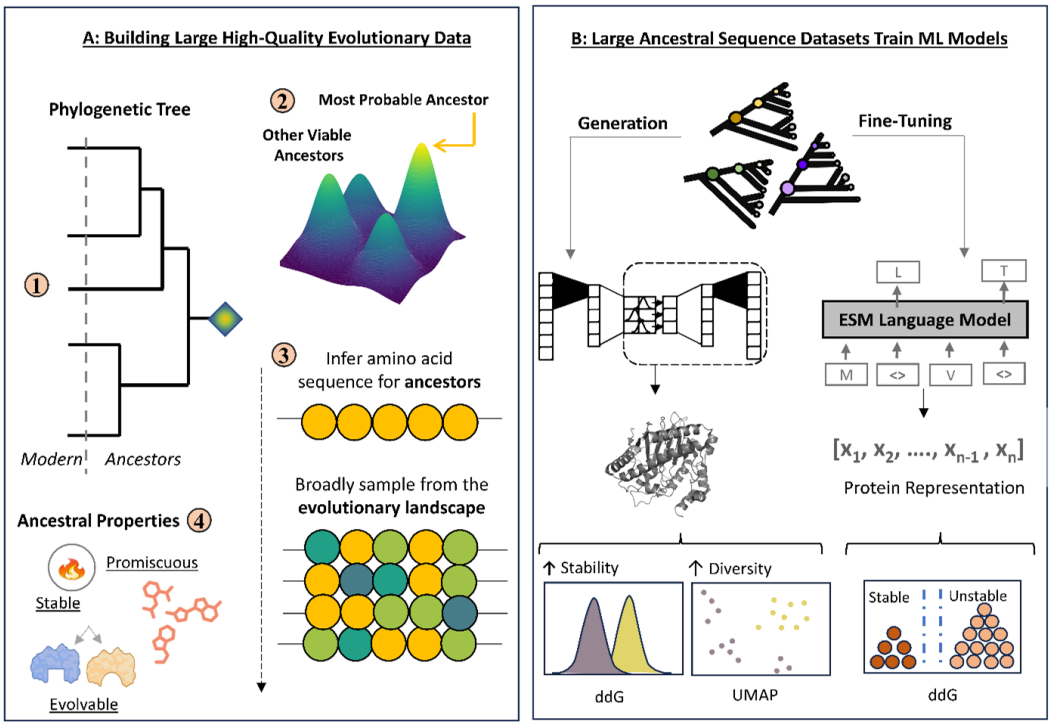

Figure 1: **Large Ancestral Datasets are Compiled from the Evolutionary Landscape to Train ML Models**. A. Represents the generation of high-quality evolutionary data. The phylogenetic tree for the family of interest is generated via ASR to access the ancestral sequences. Ancestral sequences are often known to be stable, promiscuous, and evolvable. It is important to know each node in the tree is not just one ancestor but a predicted distribution of sequences when the most probable amino acid having the highest likelihood is shown in yellow. B. Obtained evolutionary information was used as training data for sequence generation and family-specific protein sequence representation.

amino acids with a posterior probability greater than 0.2 for a given position were included for sampling in the dataset). From the data generated by AP-LASR, we sampled the sequences from four nodes that represented high-stability ancestors obtained from various evolutionary timescales Node10, Node13, Node 253, and Node 384. Then two distinct datasets for training our ML model were crafted: the "Homogeneous Dataset," which comprises sequences equally sampled from ancestral nodes, and the "Diverse Dataset," created by passing the initial dataset through CD-Hit (Li & Godzik, 2006) with a 0.9 similarity threshold. This process was aimed at enhancing the diversity of the sequences in our training set, a critical step for ensuring the robustness and generalizability of our model.

2.1.2 Model Training

A Variational Autoencoder (VAE) was selected for its proficiency in generating new data points (in this case, protein sequences) that are coherent with the training data. We employed one-hot encoding to transform the sequences into a format suitable for computational processing. Feature extraction from these one-hot encoded sequences was performed using a 1D Convolutional Neural Network (CNN) layer, allowing us to capture the local sequence patterns effectively. The architecture of our VAE was designed with a latent space dimensionality of 100, ensuring sufficient complexity to capture the nuances of protein sequence variability. Additionally, we incorporated batch normalization within the network to facilitate smoother and more stable learning dynamics. This combination of 1D CNN for feature learning and batch normalization for optimization contributed to refining the model's ability to generate meaningful protein sequences.

2.1.3 EVALUATION

The ultimate test of our generative model's efficacy lies in its ability to produce viable and thermostable sequences of the PK2 enzyme. To assess this, we utilized AlphaFold2 to predict the 3D structures of the generated PK2 sequences. Results from AlphFold2 were visualized and superimposed in PyMol. This allowed verification that a similar folding pattern to the wild-type sequence was maintained. Following structural prediction, stability calculations were carried out using FoldX. This evaluation phase was crucial, as it allowed us to ascertain not just the novelty of the generated sequences, but to also assess their practical applicability in terms of structural integrity and thermal stability. We then compared the distribution of generated structure stability measurements followed by the Dunn statistical test for measuring the obtained results' significance. The generated sequences using sequences generated on a model trained with the ASR-Dist sequence were compared to modern sequences via random sampling and UMAP visualizations. This plot is a proficient dimensional reduction technique that represents the data manifold in lower dimensions.

2.2 FINE-TUNING LANGUAGE MODELS FOR FITNESS PREDICTION

In this section, we detail our approach to creating family-specific protein representations which is a promising route for improved prediction scores in downstream tasks. We crafted four datasets that incorporate modern and evolutionary sequences in fine-tuning the ESM2 protein language model. The fine-tuned representations obtained from these datasets were compared against the ESM base representation and against each other in stability prediction task proficiency.

2.2.1 DATA PROCESSING

To generate precise, family-specific protein representations for downstream tasks, we concentrated on two proteins, Endolysin and Lysozyme C, in which we obtained their labeled datasets for stability prediction in FireProtDB (`https://loschmidt.chemi.muni.cz/fireprotdb/`). Our data processing involved assembling four distinct unlabeled datasets to fine-tune the ESM model for each protein family of interest: (i) Modern sequences sourced with NCBI's BLAST, (ii) a collection of Interpro-derived sequences that encompass an expanded set of modern proteins based on family affiliations, (iii) ancestral sequences inferred through maximum likelihood estimations in ASR via AP-LASR (ASR-Max), and (iv) near-ancestral sequences (ASR-Dist). The latter were meticulously derived not only from the most probable sequences but also from those showing promising likelihood in ancestral inference, thereby ensuring a comprehensive dataset that integrates a wide spectrum of evolutionary insights and is substantially higher in data quantity, compared to ASR-Max. We sampled 1000 sequences from each high-quality ancestral node (SH-aLRT $> 80\%$ and ultrafast bootstrapping $> 95\%$) reconstructed in ASR and removed repeat sequences. The prediction task was designed to be stability classification—determining if a given sequence is stable ($\Delta\Delta G < -0.5$ kcal/mol) or unstable ($\Delta\Delta G > 0.5$ kcal/mol).

2.2.2 MODEL TRAINING

For model fine-tuning, we employed the ESM2 model (esm2_t12_35M_UR50D) trained with 35M parameters which generates 480 embedding dimensions and contains 12-layer representations. We unfroze its last two layers to adapt its learning to our specific datasets. A batch size of 32 sequences was utilized to optimize the training process, alongside the implementation of early stopping to mitigate the risk of over-fitting. This fine-tuning phase was critical, allowing us to tailor the model to our evolutionary-informed datasets. Post-tuning, we extracted the embedding from each of the four datasets to further refine our approach to protein family classification, employing KNN, Random Forest, and XGBoost algorithms to assess the model's predictive performance within each representation derived from distinct fine-tuning methods.

2.2.3 EVALUATION

The evaluation of our fine-tuned language model focused on its ability to classify protein families accurately, employing a suite of classification metrics to gauge performance comprehensively. Precision, recall, balanced accuracy, Area Under the Curve (AUC), and the F1 score were calculated for each protein family (Endolysin and Lysozyme C) across the representations. For more robust training, we performed 5-fold cross-validation for the datasets. Then the trained models were tested on a held-out test set which was 30% of the initial data. Note that, for robustness, we repeated this

on 20 distinct random states and reported the mean and standard deviation for the obtained results among all the classification scores.

3 RESULTS

The critical importance of data in the realm of protein engineering, particularly in structure prediction and the enhancement of generalization metrics, is acknowledged. This paper sets out to illuminate the potential of integrating underutilized yet rich evolutionary information to elevate the performance of the generative model and language models in the field. It is important to note that the efficacy of fine-tuning, recognized as a state-of-the-art method for ML predictive tasks, is intrinsically linked to the caliber of the dataset upon which it is refined. For generative models, the selection of training data that aptly captures the prior distribution holds substantial weight in determining the quality of the sequences generated. Accordingly, our investigation seeks to discern how this evolutionary information can be strategically employed in ML platform to facilitate a more informed exploration and subsequently navigation of the protein fitness landscape.

3.1 GENERATED SEQUENCES VIA EVOLUTIONARY INFORMATION WERE NOVEL & STABLE

Novel sequences were generated after loss minimization in the validation set following the sampling from the learned latent representations. Three different sets of training data (called modern, homogeneous, and diverse described in detail in the method section) were used to generate distinct sets of sequences. We sampled sets of sequences both in training and generation populations within these datasets. Our study's findings are quite promising, revealing that: (i) our datasets not only augment the volume of training data through the innovative integration of uncertainty in ML but also (ii) provide a richer set of sequences with inherently higher stability for training purposes. Moreover, (iii) the stability distribution of the sequences generated using near ancestors aligns closely with those of the training set, attesting to the potential of our method to replicate high-quality protein stability profiles in novel sequence creation.

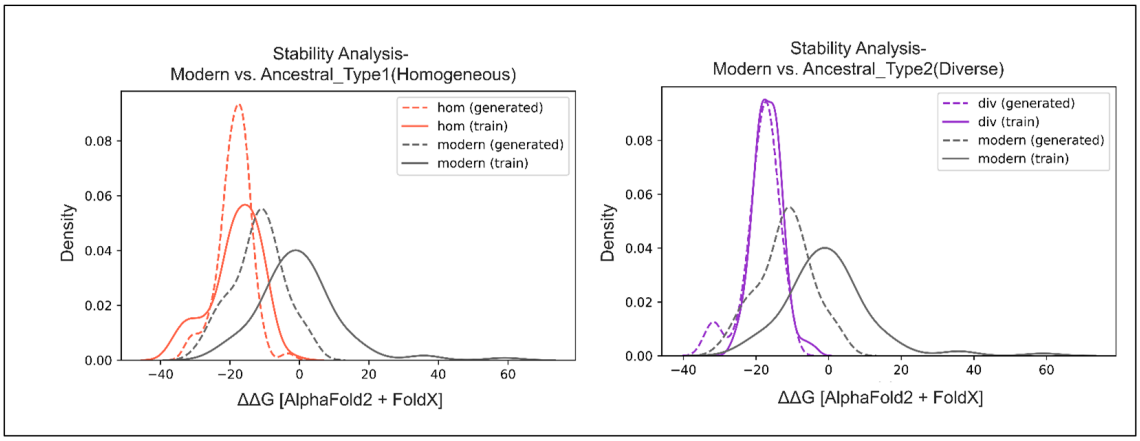

Figure 2: **ASR-derived sequences exhibited higher thermal stability compared to modern sequences. Moreover, generated sequences maintained stability profiles with the training data**. Stability comparison of PK2 protein sequences derived from various training datasets. The top-left panel illustrates the shift in thermodynamic stability ($\Delta\Delta G$ values) when incorporating ancestral information into the training data, with ancestral sequences demonstrating increased thermal stability. The remaining panels compare the $\Delta\Delta G$ distributions of novel sequences, generated post-training, to those of the training data across modern, homogeneous ancestral (AncType1), and diverse ancestral (AncType2) datasets, underscoring our method's effectiveness in producing novel, yet thermally stable protein variants.

3.2 EVOLUTIONARY-DRIVEN PROTEIN REPRESENTATIONS IMPROVED CLASSIFICATIONS

As detailed in the methods section, we fine-tuned the ESM2 model on four distinct datasets to obtain protein representations termed modern, Inter-Pro, ancestral, and near-ancestors. Intriguingly, in both predictive stability tasks—determining if a given sequence is stable ($\Delta\Delta G < -0.5\ kcal/mol$)

or unstable ($\Delta\Delta G > 0.5\ kcal/mol$) for Endolysine and Lysozyme C proteins—the representations derived from fine-tuning ESM2 with ancestral data exhibited enhanced performance relative to those derived from modern data. Furthermore, they showed comparable or superior performance to those obtained from Inter-Pro-derived sequences. The outcomes for both protein families are presented in Table 1 and Table 2. Our Method derived from ASR and uncertainty estimations (ASR-Dits) has shown improved performance across other representations in KNN and comparable or improved performance over InterPro-derived representations in ensemble-based classifiers (i.e. Random Forest, XGBoost).

Table 1: Comparison of Classifier Performance Across Fine-Tuned Representations– Endolysin

Classifier	**Dataset**	**Score (Mean±STD)**				
		Balanced Accuracy	F1	Precision	Recall	ROC_AUC
KNN	Modern	0.75±0.04	0.63±0.08	0.80±0.12	0.53±0.08	0.91±0.04
	Interpro	0.77±0.05	0.66±0.09	0.84±0.10	0.55±0.10	0.90±0.03
	ASR-Max	0.80±0.05	0.69±0.08	0.78±0.09	0.62±0.09	0.91±0.04
	ASR-Dist	0.82±0.05	0.73±0.08	0.84±0.09	0.65±0.10	0.94±0.03
Random Forest	Modern	0.83±0.05	0.74±0.07	0.83±0.08	0.68±0.09	0.96±0.02
	Interpro	0.82±0.04	0.73±0.07	0.85±0.06	0.65±0.09	0.96±0.02
	ASR-Max	0.83±0.05	0.74±0.07	0.83±0.07	0.67±0.09	0.96±0.02
	ASR-Dist	0.84±0.04	0.77±0.06	0.85±0.07	0.70±0.08	0.97±0.02
XGBoost	Modern	0.84±0.04	0.75±0.07	0.83±0.07	0.70±0.08	0.95±0.03
	Interpro	0.84±0.05	0.76±0.07	0.85±0.07	0.70±0.09	0.95±0.04
	ASR-Max	0.84±0.05	0.75±0.07	0.83±0.07	0.69±0.10	0.95±0.03
	ASR-Dist	0.85±0.04	0.78±0.06	0.84±0.07	0.72±0.07	0.94±0.04

Table 2: Comparison of Classifier Performance Across Fine-Tuned Representations– Lysozyme C

Classifier	**Dataset**	**Score (Mean±STD)**				
		Balanced Accuracy	F1	Precision	Recall	ROC_AUC
KNN	Modern	0.71±0.05	0.57±0.10	0.85±0.13	0.44±0.10	0.83±0.06
	Interpro	0.70±0.05	0.54±0.10	0.74±0.10	0.44±0.12	0.81±0.04
	ASR-Max	0.71±0.05	0.57±0.10	0.90±0.12	0.43±0.10	0.81±0.04
	ASR-Dist	0.73±0.05	0.61±0.09	0.83±0.13	0.50±0.09	0.86±0.05
Random Forest	Modern	0.74±0.05	0.63±0.09	0.83±0.09	0.51±0.12	0.94±0.02
	Interpro	0.75±0.06	0.64±0.10	0.80±0.09	0.55±0.13	0.92±0.02
	ASR-Max	0.73±0.06	0.60±0.10	0.78±0.11	0.50±0.12	0.93±0.02
	ASR-Dist	0.75±0.05	0.64±0.09	0.82±0.11	0.54±0.11	0.94±0.02
XGBoost	Modern	0.76±0.05	0.65±0.08	0.80±0.12	0.57±0.11	0.92±0.03
	Interpro	0.78±0.05	0.67±0.08	0.80±0.10	0.60±0.13	0.94±0.02
	ASR-Max	0.77±0.05	0.67±0.09	0.78±0.12	0.60±0.09	0.94±0.03
	ASR-Dist	0.78±0.05	0.67±0.07	0.77±0.10	0.60±0.10	0.93±0.03

4 CONCLUSION

In this work, we explored the augmentation of ML models for protein engineering by incorporating evolutionary information. Our methodology involved generating novel protein sequences with a VAE model and fine-tuning language models for improved fitness prediction. The results demonstrated that the sequences generated from evolutionary-informed datasets were not only novel but also exhibited higher thermal stability, showcasing the potential of ASR to enhance the quality and diversity of training data for generative models. Additionally, fine-tuning the ESM2 model with these datasets led to improved classification accuracy in protein family prediction tasks. This pioneering approach underscores the significant impact of integrating evolutionary insights into machine learning frameworks, marking a substantial advancement in our ability to navigate the complex protein fitness landscape with greater precision and efficiency.

REFERENCES

J. Adeoye, L. Hui, and Y.X. Su. Data-centric artificial intelligence in oncology: A systematic review assessing data quality in machine learning models for head and neck cancer. *J. Big Data*, 2023. doi: 10.1186/s40537-023-00703-w.

Hila Bar-Rogovsky, Adi Stern, Osnat Penn, Itay Kobl, Tal Pupko, and Dan S Tawfik. Assessing the prediction fidelity of ancestral reconstruction by a library approach. *Protein Eng Des Sel*, 28: 507–518, 2015. doi: 10.1093/protein/gzv038.

Iman Deznabi, Bulent Arabaci, Mehmet Koyuturk, and Oznur Tastan. Deepkinzero: Zero-shot learning for predicting kinase-phosphosite associations involving understudied kinases. *Bioinformatics*, 36:3652–3661, 2020. doi: 10.1093/bioinformatics/btaa013.

Geeta N Eick, Jamie T Bridgham, David P Anderson, Michael J Harms, and Joseph W Thornton. Robustness of reconstructed ancestral protein functions to statistical uncertainty. *Mol Biol Evol*, 34:247–261, 2017. doi: 10.1093/molbev/msw223.

W. Gao, S.P. Mahajan, J. Sulam, and J.J. Gray. Deep learning in protein structural modeling and design. *Patterns*, 1:100142, 2020. doi: 10.1016/j.patter.2020.100142.

Yoseph Gumulya and Elizabeth MJ Gillam. Exploring the past and the future of protein evolution with ancestral sequence reconstruction: The "retro" approach to protein engineering. *Biochem J*, 474:1–19, 2017. doi: 10.1042/BCJ20160507.

Yuichi Joho, Varodom Vongsouthi, Michelle A Spence, Jessica Ton, Chelsea Gomez, Lian L Tan, Jake A Kaczmarski, Anthony T Caputo, Shalini Royan, Colin J Jackson, et al. Ancestral sequence reconstruction identifies structural changes underlying the evolution of ideonella sakaiensis petase and variants with improved stability and activity. *Biochemistry*, 2022. doi: 10.1021/acs.biochem.2c00323.

K. Katoh and D.M. Standley. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780, April 2013. doi: 10.1093/molbev/mst010. Epub 2013 Jan 16.

S.A. Kauffman and E.D. Weinberger. The nk model of rugged fitness landscapes and its application to maturation of the immune response. In *Mol. Evol. Rugged Landscapes Proteins, RNA Immune Syst.*, pp. 135–175, 2018. doi: 10.1201/9780429498879-9.

Weizhong Li and Adam Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006. doi: 10.1093/bioinformatics/btl158.

H. Liu, M. Chaudhary, and H. Wang. Towards trustworthy and aligned machine learning, 2023.

Joshua Meier, Roshan Rao, Reinier Verkuil, Jesse Liu, Tom Sercu, and Antoine Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv*, 2021 (07):2021.07.09.450648, 2021.

M.A. Mena and P.S. Daugherty. Automated design of degenerate codon libraries. *Protein Eng. Des. Sel.*, 18:559–561, 2005. doi: 10.1093/protein/gzi061.

Bui Quang Minh, Heiko A Schmidt, Olga Chernomor, Dominik Schrempf, Michael D Woodhams, Arndt von Haeseler, and Robert Lanfear. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, 37(5):1530–1534, 02 2020. ISSN 0737-4038. doi: 10.1093/molbev/msaa015. URL `https://doi.org/10.1093/molbev/msaa015`.

L. Miranda. Study notes on data-centric machine learning, 2023.

Linus Pauling, Emile Zuckerkandl, Thorkil Henriksen, and Rolf Lövstad. Chemical paleogenetics. *Acta chem scand*, 17:S9–S16, 1963.

P.A. Romero and F.H. Arnold. Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.*, 10:866–876, 2009. doi: 10.1038/nrm2805.

Michael A Sennett and Douglas L Theobald. Ancestral sequence reconstructions evaluated by extant sequence cross-validation. *bioRxiv*, pp. 1–23, 2022.

P. Singh. Systematic review of data-centric approaches in artificial intelligence and machine learning. *Data Sci. Manag.*, 6:144–157, 2023. doi: 10.1016/j.dsm.2023.06.001.

James VanAntwerp, Mehrsa Mardikoraem, Nathaniel Pascual, and Daniel Woldring. Ap-lasr: Automated protein libraries from ancestral sequence reconstruction. *bioRxiv*, 2023. doi: 10.1101/2023.10.09.561537. URL https://www.biorxiv.org/content/early/2023/10/12/2023.10.09.561537.

A.X. Wang, S.S. Chukova, C.R. Simpson, and B.P. Nguyen. Data-centric ai to improve early detection of mental illness. In *IEEE Work. Stat. Signal Process. Proc.*, volume 2023-July, pp. 369–373, 2023. doi: 10.1109/SSP53291.2023.10207938.

Jonas Zaugg, Yoseph Gumulya, Elizabeth MJ Gillam, and Mikael Bodén. Chapter 21 of prospective and retrospective strategies. 2014. ISBN 9781493910533.

D. Zha and U. States. Data-centric artificial intelligence: A survey, 2023.