Adapting Protein Language Models for Explainable Fine-Grained Evolutionary Pattern Discovery

1st Ashley Babjac
Department of Computer Science
University of Tennessee
Knoxville, USA
ababjac@vols.utk.edu

1st Owen Queen
Department of Biomedical Informatics
Harvard University
Boston, USA
owen_queen@hms.harvard.edu

3rd Shawn-Patrick Barhorst
Department of Computer Science
University of Tennessee
Knoxville, USA
shaebahr@vols.utk.edu

4th Kambiz Kalhor
Department of Microbiology
University of Tennessee
Knoxville, USA
kkalhor@vols.utk.edu

5th Andrew D. Steen
Department of Microbiology
University of Tennessee
Knoxville, USA
asteen1@utk.edu

6th Scott J. Emrich
Department of Computer Science
University of Tennessee
Knoxville, USA
semrich@utk.edu

2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 979-8-3503-3748-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/BIBM58861.2023.10385976

***Abstract*—Organisms that live in different environments face different evolutionary pressures. As such, organisms that have more successful phenotypes reproduce more frequently, but differing selective pressures acting at the organismal level can influence genes, and thus proteins. Understanding how proteins adapt across environments may therefore be useful in engineering proteins for specific environments as well as to improve our understanding of basic biology. In this work, we explicitly compare homologous (read: paired) proteins from different environments. While previous studies have explored the relevant evolutionary pressures in one of these environments [11, 17] and genomic responses to those pressures [1, 28], no prior computational study of their proteins has been performed. We apply ESM-2 [20] and although there is no signal in our negative control (two divergent yeast strains) as expected, we obtain near perfect prediction accuracy for our selected environmental gradient–the well-established subsurface vs. surface biome. We further show that ESM-2 is able to capture relevant fine-grained biological patterns in its embedding space, even in its smallest model. Significantly, we demonstrate that these embeddings can be interpreted using a novel visualization pipeline built using explainable AI techniques.**

***Index Terms*—ESM-2, Classification, Explainability, Protein Sequence, Evolution**

I. Introduction

Proteins are an essential part of all organisms because of their impact on regulatory processes, gene expression, and other biological processes within cells. Proteins themselves are made up of amino acid chains and peptide bonds. There are 20 naturally occurring amino acids, and proteins can be represented as strings, where each character in the string represents an individual amino acid. It is well established that there is a correlation between protein structure and function [24, 30] and that amino acid order (and their underlying codons) can affect protein folding [15, 39]. As such, it is important to understand and interpret characteristics of proteins, especially ones linked to their functional evolution. Since proteins can be represented textually, modeling protein sequences naturally maps to natural language processing (NLP) tasks.

Many recent advances have been made in the field of NLP, and transformer networks have come to prominence because of their intrinsic self-attention and high-performing transfer learning capabilities [34]. For example, Bidirectional Encoder Representations from Transformers (BERT) architectures have shown good performance on a variety of down-stream tasks including classification [6]. More recently, the Evolutionary Scale Modeling (ESM) architecture was designed by Meta AI specifically for protein sequences [27]. The most recent variant, ESM-2, improves upon the earlier ESM model by including larger models, small architectural changes, and additional downstream tasks [20].

We seek to apply NLP methods for a specific niche downstream task: modeling evolutionary differences between structurally similar proteins produced in contrasting environments. Because detectable differences are more likely to occur in contrasting environmental conditions, and resulting environmental gradients are common, we focus on the surface/subsurface of the ocean since it has well-documented large-scale—but not well studied—functional differences between the two local environments [11]. We also consider two different strains of yeast (*Saccharomyces cerevisiae*) as a potential negative control, since the genetic adaptations in a bioreactor are likely a result of substantial gene expression differences [4] and therefore not large-scale differences in the proteins produced.

A. Related Work

From a computational perspective, many previous studies have shown adaptations of transformer models to modeling common protein language tasks. For example, Nambiar et al. used Bidirectional Encoder Representations from Transformers (BERT) to predict disordered regions based on protein sequence data [22]. BERT has also been used successfully

to look for signal peptides, which mark proteins for secretion across membranes in a cell [33]. Similarly, ESM has been modified for many downstream tasks including disordered protein modeling [26], functional prediction [13], and creation of functional annotations [35]. Other common protein language tasks such as fluorescence, stability, and protein folding have similarly been modeled, e.g., AlphaFold for protein folding [5]. Various protein language models have also been benchmarked for traditional downstream tasks in multiple studies [20, 36].

There have additionally been several non-computational studies looking at coarse-grained evolutionary patterns among paired species. With respect to the ocean, Jørgenson et al. describes several reasons for differences in surface/subsurface proteins including sedimentation, tectonic plate movement, chemical differences, pressure, and most importantly temperature [16]. Organisms in these low energy environments have also evolved to catabolize and subsist off much less energy than their surface counterparts [17]. Similarly, there have been in-depth metagenomics analysis of bioreactors. For example, Chirania et al. report that stress-response proteins are unregulated at higher levels of solids in a bioreactor, but there is an increased abundance in certain carbohydrate-active enzymes as well as other notable enzyme enrichment that correlates with solids levels [4]. While both of these problems have been analyzed from an environment perspective, no such study exists to examine the potential effects and adaptations with respect to differences in the proteins produced.

B. Our Contributions

Here we explore how protein language models can be utilized on a real-world microbiology task: discriminating potential fine-grained evolutionary patterns from paired proteins. We demonstrate the utility of ESM-2, a large protein language model, to correctly identify niche environmental patterns among our selected data. As expected, no patterns were identified in our negative control (highly divergent yeast strains), but, significantly, we are able to achieve nearly perfect performance in the subsurface/surface environmental gradient. To further validate these results we benchmark our methods using two additional protein language "control" datasets and found that ESM-2 outperforms all tested baselines across all tasks. Significantly, we also show that ESM-2 learns biologically meaningful features in its embedding space, a promising finding that could contribute to the use of protein language models for further studying of niche ecological tasks. Further, in order to analyze the fine-grained evolutionary patterns, we develop a novel visualization approach: we combine ESM-2, protein structural prediction (AlphaFold2), and explainable AI (XAI) to visually highlight biological differences between two extreme environments. We showcase a case study of five selected homologous protein pairs and give preliminary interpretations of the biological findings determined from ESM-2 embeddings of subsurface proteins.

II. Materials and Methods

A. Data

Organisms can be strongly impacted by environmental gradients. We model this on the protein level by selecting homologous paired sequences (see Figure 1). These paired sequences are ideal for studying evolutionary patterns because they are a direct product of natural environmental impacts on biological function and adaptation.

Our primary dataset (referred to as "Deep Surface") is a peer-reviewed, custom dataset of 230,456 homologous protein sequence pairs, where one sequence is present in the surface environment (labeled "surface" or 0) and the other sequence is from the deep subsurface water environment (labeled "subsurface" or 1) [41]. Based on metagenomics methods, the surface environment was defined by being less than 25 centimeters below the surface of the water, and the subsurface environment was defined by being greater than 15 meters below the surface of the water [41]. Prior to training, the data were cleaned to remove any sequences with ambiguous amino acids. When stacked, these pairs provide 460,912 total protein sequences for the dataset with perfectly balanced labels due to the pairs.

As a likely negative control, we also chose proteins from two divergent yeast strains where most adaptations to highly different environments are likely genetic (i.e., gene expression differences). This second dataset (referred to as "Yeast Biofuel") was generated by downloading the annotation for the reference lab strain (s288c) from the Saccromyces Genome Database (SGD) and matching genes from a biofuel strain (Jay291) using their SGD-provided gene names. This resulted in a dataset of 5,559 homologous protein sequence pairs, where one sequence is from the reference strain (labeled 0) and the other is selected/engineered for biofuel production (labeled 1). These data were cleaned in the same pipeline as the prior data, and when stacked provide 11,118 total sequences.

To further strengthen our model benchmarking, we additionally include two related protein datsets for performance comparison. We chose two standard protein benchmarking datasets available through torchdrug: 1) "Beta Lactamase"[1], a small dataset comprised of first-order mutants of the TEM-1 beta-lactamase protein with regression labels predicting the activity values; and 2) "Binary Localization"[2], a simplified subcellular localization prediction where the binary labels denote either membrane-bound/soluble.

B. Baseline Models

We selected four baseline models based on what is commonly used in similar bioinformatics studies [36]: 1) Random Forests; 2) Convolutional Neural Networks (CNNs); 3) Long Short Term Memory models (LSTMs); and 4) Transformers (BERT) . Our baseline models were built using the torchdrug API[3], with the exception of our Random Forest baseline that

[1]https://torchdrug.ai/docs/api/datasets.html#betalactamase
[2]https://torchdrug.ai/docs/api/datasets.html#binarylocalization
[3]https://torchdrug.ai/docs/api/models.html#protein-sequence-encoders

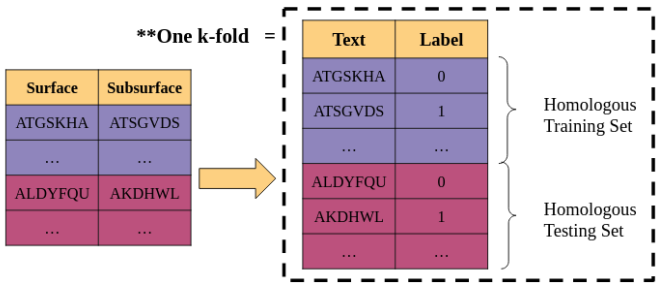

Fig. 1. An example of the homology that is preserved throughout our training and testing datasets for the Deep Surface dataset. Each row in the table on the left represents a homologous pair of proteins—one found at the surface while the other is found at the subsurface. When constructing each k-fold split, we ensure that there is no data leakage between training and testing sets by splitting such that the pairs remain together. The Yeast Biofuel data follows the same formulation, just using reference/biofuel pairs instead of surface/subsurface.

is implemented using Scikit-Learn. All code used for training is available on GitHub: https://github.com/owencqueen/DeepSurface.

The Random Forest approach builds upon a decision tree by performing bagging (building many iterations of trees through random sampling) and uses the majority vote of these trees as the final outcome [8]. In our Random Forest architecture, we iterate through a grid search of the following parameters to determine the best configuration for each model: estimators = [10, 25, 50, 100], max_depth = [10, 25, 50, 100,], learning_rate = [0.1, 0.01, 0.001, 0.0001]. The final forest had 100 estimators and a maximum depth of 25.

CNNs are neural networks which model the overlaps of inputs through mathematical convolutions. Each convolutional layer applies a ”filter” over the inputs, which allows the model to easily identify varying key information at each layer. We implement a CNN based on the parameters chosen in PEER [36]. We use two hidden layers of size 1024, each with a kernel size of five and padding of size two. The model is trained using a batch size of 64, adam optimizer, and learning rate of 1e-4 for a minimum of 100 epochs.

LSTMs are models which retain sequential (time) information throughout the network. LSTMs do this by retaining two threads of information: the “long-term” thread, which comes from a combination of information from the previous states, and the “short-term” thread that comes from information at the current time-step. [37]. Consistent with the benchmarking of PEER [36], we use a five layer LSTM, with three LSTM blocks (hidden_dim=640) and two MLP layers. The model is trained with a batch size of 32, adam optimizer and learning rate of 5e-5 for a minimum of 30 epochs.

Transformers are a family of models that are defined by an ability to be highly customized due to being built in blocks. We utilize the Bidirectional Encoder Representation Transformer (BERT) [6], which is a pre-trained encoder-only transformer architecture. We follow the default BERT configs (as tested by PEER [36]), which uses 12 attention heads and 12 hidden layers in the encoding layers, with two MLP layers for the output. We train with a batch size of 16, adam optimizer, and learning rate of 5e-5 for a minimum of 10 epochs.

C. ESM

Evolutionary Scale Models (ESM) are a family of large, general-purpose protein language models. We utilize the latest version (ESM-2), which has demonstrated superior performance over previous iterations of ESM on a variety of downstream tasks [20]. The ESM-2 transformer architecture is pre-trained on protein sequences from UniRef50 [32] using a masked language modeling (MLM) objective, where the model is asked to predict missing residues in the input protein sequence that are randomly masked-out during the pre-training process. We downloaded pre-trained weights provided by developers of ESM-2 through Huggingface[4].

We experiment with the 8 million (8m) parameter and 35 million (35m) parameter version of ESM-2, `esm2_t6_8M_UR50D` and `esm2_t12_35M_UR50D`, respectively. These are the two smallest ESM-2 variants in terms of number of parameters; we chose small models for efficiency and generalization purposes. If a small model is effective, this makes similar training more feasible for a broader segment of the research community given the computational resources required to run and train very large ESM-family models.

For our prediction task, we use the provided prediction network from Huggingface, a two-layer fully-connected network with hyperbolic tangent activation functions and dropout layers. This prediction network is randomly initialized, and it performs classification based on embeddings from the ESM-2 encoder. Following standard transfer learning methodologies, we experiment with two strategies for training this classifier on top of ESM-2: 1) fine-tuning only the prediction network and 2) fine-tuning the entire architecture. From here on, the first strategy is referred to as ”fixed ESM” and the second is referred to as ”fine-tuned ESM”. In general, the latter strategy is expected to yield better performance because a larger portion of model parameters are tuned for the prediction task.

For hyperparameters, we use a batch size of 8 and learning rate of 0.001 for both 8m fixed and 35m fixed models. For 8m and 35m models fine-tuned, we use batch size 4 and learning rate 0.0001. A smaller batch size is used for fine-tuning because of RAM limitations when tuning an entire model while a smaller learning rate is used for fine-tuning in order to avoid large deviance from original parameters when tuning the pre-trained layers.

D. Evaluation Criteria

To evaluate the classification performance of each model, we use a 5-fold cross validation strategy, i.e., we create five individual data splits and train five separate models then average the results to get the final performance metrics. To avoid potential data leakage across homologous proteins in the training and testing datasets, we use a homology-aware split, ensuring that no pair homologous proteins are present in both training and testing data (see Figure 1).

[4]https://huggingface.co/docs/transformers/model_doc/esm

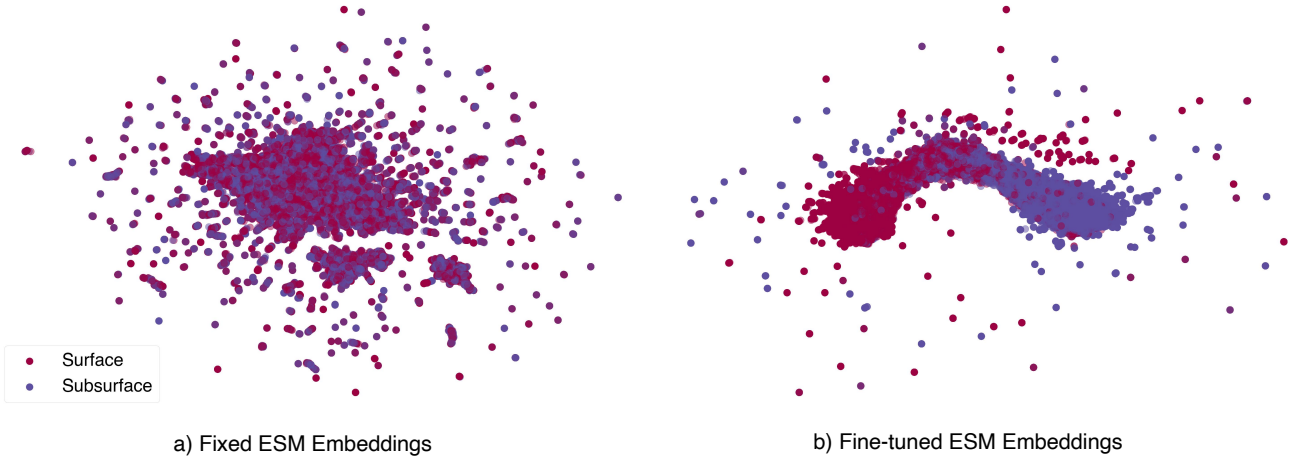

Fig. 2. Embedding spaces of ESM model for surface-subsurface prediction when fixed or when fine-tuned. Panel a) shows a UMAP [21] plot of the embeddings for ESM 35m without any fine-tuning on the surface vs. subsurface prediction task. Panel b) shows the UMAP visualization of the embedding space after fine-tuning ESM to predict surface vs. subsurface. As expected, ESM without fine-tuning cannot differentiate between subsurface and surface proteins because they are homologous and presumably have similar biological functions/structures. However, after fine-tuning (panel b), the embedding space separates into classes based on environmental gradient labels.

We compare the performance of all models using two metrics: AUROC (area under the receiver-operator curve) [23] and AUPRC (area under the precision-recall curve) [2]. AUROC is a number between 0 and 1 calculated as the area underneath the ROC curve (showing true vs false positive rate across all classification thresholds); 0.5 is considered to be random baseline performance. AUPRC represents the trade-off between precision (the ratio of correct positive predictions, i.e. low false positives) and recall (the ratio of correctly identifying actual positives, i.e. low false negatives). The value of this curve's area falls between 0 and 1, with higher numbers representing both high precision and high recall. The AUROC and AUPRC metrics have been shown to be more robust to misclassification and therefore better assessments of classification tasks in general [14].

E. Explainability

The field of explainable artificial intelligence (XAI) is concerned with elucidating the underlying behavior of machine learning models. Through this elucidation, users are able to understand model predictions, including patterns that are exploited in the data to make predictions. To interpret ESM embeddings, we use Integrated Gradients [31], a popular explainability method that produces attribution maps $E(\mathbf{x}_i) \in \mathbb{R}^L$ for a sequence $\mathbf{x}_i$ of length L. The explanations communicate relative importance scores: if $E(\mathbf{x}_i)[j_1] > E(\mathbf{x}_i)[j_2]$, then residue j_1 is said to be more important for making the prediction of $\mathbf{x}_i$ than residue j_2. We use a layer-wise variant of Integrated Gradients implemented by Captum [19]. Next, we map the associated gradients onto their predicted protein structure to visualize them. We use AlphaFold2 [18] with default settings to predict the proteins' 3D structures, and apply the average of each gradient as the coloring on top of each amino acid. The color for each amino acid is normalized to match with Matplotlib's sequential colormaps, and the appropriate color is applied on top of each atom using PyMol3D [29]. We choose five homologous surface-subsurface protein pairs to visualize. Each pair of proteins is chosen from a biologically relevant subset based on if our fine-tuned ESM-2 correctly predicts the surface/subsurface labels; this is to guard against faulty interpretations from a prediction on which the model was incorrect. The computed attributions come from the fine-tuned 8m-parameter ESM-2 model for our visualizations.

III. Results

A. Benchmarking PLM Methods

Table I shows the results of our benchmarking across all models. To confirm the integrity of our model training, we note the results of the Beta Lactamase and Binary Localization tasks match that of previously published benchmarking using the same data [36]. When looking at subsurface/surface predictions, fine-tuned ESM-2 performs the best of all models, with the 35m-parameter model achieving an AUROC of 0.925 and AUPRC of 0.927. Outside of the ESM models, CNNs perform the next best at 1.96% and 2.70% worse AUROC and AUPRC, respectively, than the best-performing ESM model. CNNs are similar in performance to that of the 8m-parameter ESM-2, with all other models (LSTMs, BERT, random forests) performing comparatively lower. Surprisingly, the worst performing model is BERT, which performs only slightly better than random chance (AUROC and AUPRC equal to 0.502). Similar close-to-random AUROC and AUPRCs are obtained for yeast using the baselines and ESM (data not shown), as expected, given strain differences are likely stemming largely from gene expression differences and not differences in produced proteins.

Further, while ESM-2 has the best results, these results vary across different numbers of parameters and varying training

TABLE I
MODEL RESULTS ACROSS MOST TASKS. THE DATA "BETA LACTAMASE" AND "BINARY LOCALIZATION" COME FROM A STANDARD SET OF BENCHMARKING TASKS. THE POOR PERFORMING YEAST STRAIN RESULTS ARE NOT SHOWN. THE REMAINING RESULTS AS REPORTED HERE ARE IN LINE WITH OTHER BENCHMARKS [36]. ACROSS ALL MODELS THE LARGER 35M-PARAMATER ESM-2 PERFORMS THE BEST FOR ALL METRICS.

Data	Beta Lactamase		Binary Localization		Deep Surface	
Metrics	RMSE	Spearman ρ	AUROC	AUPRC	AUROC	AUPRC
RF-baseline	0.790	-0.228	0.541	0.596	0.820	0.800
CNN-baseline	0.181	0.794	0.889	0.894	0.907	0.902
LSTM-baseline	0.316	0.099	0.923	0.926	0.891	0.885
BERT-baseline	0.323	0.210	0.513	0.585	0.502	0.502
ESM-2 (8m, fixed)	0.313	0.269	0.933	0.943	0.542	0.534
ESM-2 (8m, fine-tuned)	0.176	0.782	0.953	0.958	0.909	0.910
ESM-2 (35m, fixed)	0.322	0.187	0.948	0.953	0.545	0.538
ESM-2 (35m, fine-tuned)	**0.142**	**0.847**	**0.958**	**0.958**	**0.925**	**0.927**

strategies. When ESM-2 is fine-tuned (layers unlocked) we achieve around 0.4 better AUROC and AUPRC for both the 8m- and 35m-parameter models when compared to the fixed (locked layer) models. Across numbers of parameters, a slight gain is observed when scaling from 8 million to 35 million parameters; this matches intuition observed in the original ESM work where parameter scaling tends to improve performance [20]. However, the gain in performance between 8m and 35m models is only slight – a 338% parameter increase results in only 1.76% gain in AUROC and only 1.87% gain in AUPRC.

B. Explainability analysis

Having demonstrated that deep learning, particularly ESM-2, can accurately discriminate differences where they should (Beta Lactamase, Binary Localization) and likely should not occur (yeast strains), we next confirm the surface/subsurface learned contexts relate to the well-established environmental gradient. Figure 2 shows UMAP reductions of the 35m ESM-2 learned embeddings for both fixed (left) and fine-tuned (right) models. We can see that the fine-tuned model has distinct separation between surface and subsurface proteins, whereas the fixed model contains no separation. This means that the fine-tuned ESM embeddings contain ecological patterns not found in the original ESM pre-training that are determined by using fine-tuning. This is significant because no previous study has been able to pinpoint niche ecological patterns within sequence embeddings.

To demonstrate the learned ecological patterns visually, we use five case studies to highlight the regions of input that ESM-2 learned to be important for predicting proteins that exist in two very different abiotic environments (Figure 3). Despite only slight differences in overall structure, there is distinct differentiation among the colors mapped on top of each protein. For example, the first column shows prominent differentiation throughout the entire protein structure, with the subsurface protein having noticeably more color in its coils than the surface protein. This is particularly evident in the "tails" of the paired proteins. The subsurface protein (top) appears to have lower attribution (red) in its tail when compared to its surface counterpart with higher attribution (blue). Similarly, the third and fifth columns show visibly stark contrasts in color between the subsurface and surface proteins – with both subsurface proteins (top row) showing importance (blue) among the elongated regions of the protein. It is biologically plausible that the substructural information identified from these visualizations are evolutionarily important. We will discuss potential implications of these mappings in Section IV-B.

IV. DISCUSSION

A. ESM is state-of-the-art for niche downstream task prediction

As discussed in Section III-A, fine-tuned ESM-2 performs very well. We were not able, however, to detect differences in our negative control (divergent yeast strains). These results reinforce the intuition that there is a "Goldilocks" zone in understanding more closely related proteins, i.e. two optimized strains are likely not diverged enough on the protein level while the pre-existing ESM-2 model is too generalized (see Figure 3). It follows what we see in that fine-tuning pre-trained weights performs better, and we see this result across all four datasets considered.

Surprisingly, of all of the models tested, the BERT-baseline model—which ESM is most similar to—consistently did not perform well. This is somewhat expected given the torchdrug API relies on a version of BERT that is randomly initialized. We attempt to combat this by considering DR-BERT as an alternative, given it is a BERT model that was trained on a corpus of amino acid sequences – albeit for the purpose of disorder prediction by residue [22]. This model appears to partially account for some protein language patterns (as it was pre-trained on protein data), but overall performance on predicting subsurface/surface proteins is similar to results obtained using our fixed-ESM architectures, where the AUROC/AUPRC falls around the range ~0.53 - 0.54. We conclude that patterns contained in DR-BERT are more beneficial than initializing from random weights, but could be still insufficient for more specialized contexts, especially since other deep learning approaches (e.g., CNNs and LSTMs) perform much better on the same data.

Similarly, ESM without fine-tuning is also unable to discriminate the subsurface gradient (see Table I); however, after fine-tuning, ESM achieves state-of-the-art performance for this

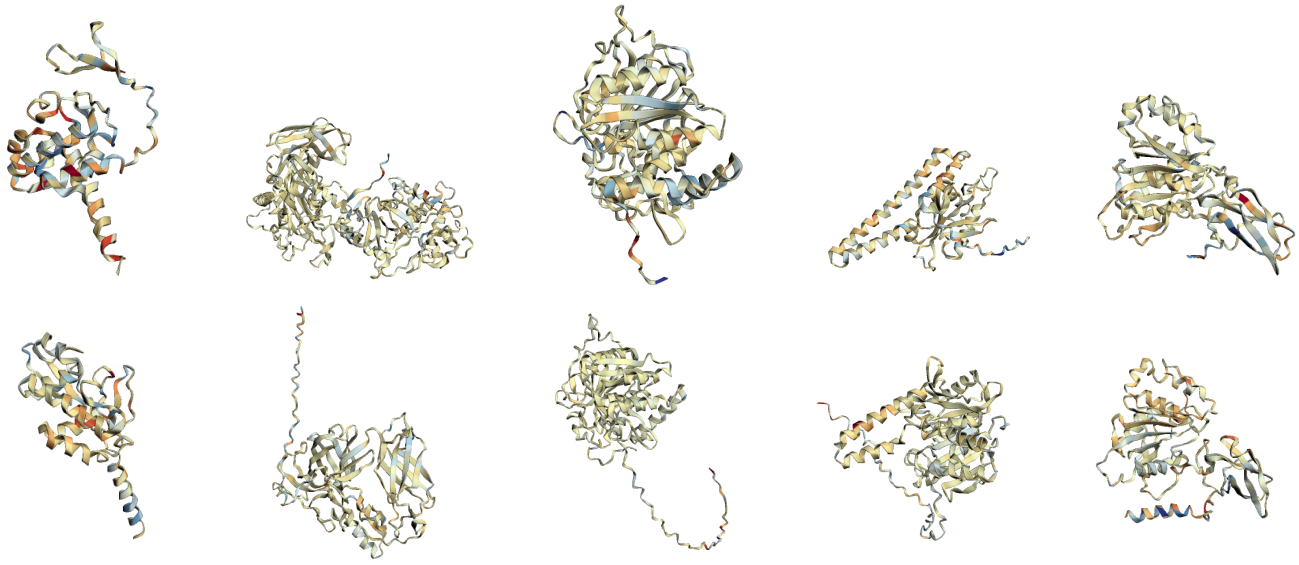

Fig. 3. A demonstration of explainable AI methods for qualitative assessment of paired proteins chosen from Zinke et al. [41]. The top row represents subsurface proteins , and the bottom row represents their corresponding homolog surface proteins, where all proteins were predicted correctly by our model. We map the gradients returned from Integrated Gradients onto the predicted structure from AlphaFold2 by amino acid. Each atom inside an amino acid is colored using the same color. We use Pymol3D and Matplotlib for visualization of the proteins. The colormap is Matplotlib's "RdYlBu", where blue represent higher attributions and red represent lower attributions.

task. This further reinforces the idea of a "Goldilocks" zone. The ESM-2 patterns, while helpful in understanding general protein patterns, are too coarse-grained to fully understand this difficult ecological task alone. This is supported by the fact that the fixed embedding space (Figure 2) holds no discriminative geometry about the prediction task—as expected given these proteins differ only in the environment they are produced in, not their inferred function(s)—while the fine-tuned embedding space separates clearly, leading to better prediction performance. Our findings suggests there are consistent amino acid differences that may have allowed proteins to adapt to more extreme environments such as those in the subsurface of the ocean, but more work is needed to verify this claim.

B. Potential implications of mapped ESM-2 attributions

We use a custom pipeline built on Integrated Gradients [31] to compute a singular attribution array per sequence. Hence the interpretation of a single attribution value would be the average importance each amino acid contributes to the model prediction in the context of that particular sequence. When combined with mapped protein structure predictions, we can effectively visualize never before seen fine-grained patterns that are important to the predicted task – in this case, evolutionary patterns. We discuss the novel patterns identified in our case study visualizations below which demonstrate the efficacy of our technique.

Surface proteins are generally more all around yellow in color (of median importance to the model). Interestingly, subsurface proteins—which are produced in the more extreme environment—tend to have far more visual coloring when compared to their surface counterparts. This is in terms of both blue coloring (high importance) and red coloring (low importance). For example, the slight structural differences among protein pairs are highlighted in blue in Figure 3 (important), e.g., surface proteins (bottom row) tend to have elongated tail regions that the subsurface proteins lack. As a good control of the accuracy of our XAI methodology, regions of near identical structure among pairs appear red in the visualizations, indicating low importance. In all five sets of proteins visualized in Figure 3, the coiled regions (among others) tend to be marked unimportant if they are shared among proteins. While these findings are already known to biologists, they are also a good qualitative "sanity check" that our model is understanding the correct biological patterns.

From an interpretation standpoint, this means that there are regions that are almost completely predictive of being a subsurface protein as well as regions that have almost no information, which is expected given these pairs of proteins presumably have very similar biological functions. Hence, we have either discovered artifacts of evolution or more likely biological differences that are induced by the more extreme environment of the subsurface of the ocean. In either scenario, this is exciting due to the fact that these coloring's are not necessarily defined by structural or functional differences but environmental context. This will lend itself to future experimental studies to better design proteins to be more stable under different and especially extreme conditions.

C. Future Work

To fully understand the biological implications of more specialized ESM models more in-depth studies are needed. One such option is to use molecular dynamic simulations [12]

to test whether differentiated regions could impact structural stability or any other important protein characteristics. Another technique that has the ability to help better understand these patterns is sequence to sequence translation [38]. While there has been much work performing sequence to sequence translation with spoken languages (e.g. [3, 40]), there has been little work in applying these techniques to protein language modeling. Some work has been done in other bioinformatics applications to try and model multiple sequence alignment [7], expedite drug development [10], and generate protein sequences [9]. We would like to apply this technique to translate between subsurface and surface sequences (or vice versa) because organisms in these low energy environments have also evolved to catabolize and subsist off much less energy [17] and therefore their proteins are likely much more stable. We leave to future work an in depth comparison of protein structures and stability (molecular dynamics) between two given environments.

V. Conclusion

We established ESM-2 as being the best architecture for multiple sequence related protein language tasks and especially our new target task of studying proteins in a well-established environmental gradient: surface vs. subsurface of the ocean. This is in line with a large body of work demonstrating that protein pre-training significantly improved performance in many downstream tasks (e.g. [25, 36]). Significantly, we discovered that the smaller 8m parameter model performs comparably to the 35m parameter model when fine-tuned. This is a positive result for researchers lacking computational resources, as accessible, smaller models are still able to produce good results under certain circumstances. Further, we have shown that our custom visualization pipeline using mapped attribution scores highlights interesting ecological patterns. In our case study of five protein pairs, we have found that the model gradients identify patterns likely to contain significant biological information – especially in reference to subsurface proteins produced in the more extreme environment. This is a significant finding for biologists looking to interpret evolutionary information across varying environments. We foresee this pipeline being useful to future protein studies due to ESM-2's robust application area and the ability to interpret fine-grained contextual patterns via XAI elucidation.

Acknowledgment

ANB was supported by the NSF GRFP. We thank the UTK High-Performance Computing team for their assistance in acquiring hardware to run our experiments.

References

[1] Jordan T Bird et al. "Uncultured microbial phyla suggest mechanisms for multi-thousand-year subsistence in Baltic Sea sediments". In: *MBio* 10.2 (2019), e02376–18.

[2] Kendrick Boyd, Kevin H Eng, and C David Page. "Area under the precision-recall curve: point estimates and confidence intervals". In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*. Springer. 2013, pp. 451–466.

[3] Necati Cihan Camgoz et al. "Sign language transformers: Joint end-to-end sign language recognition and translation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10023–10033.

[4] Payal Chirania et al. "Metaproteomics reveals enzymatic strategies deployed by anaerobic microbiomes to maintain lignocellulose deconstruction at high solids". In: *Nature Communications* 13.1 (2022), p. 3870.

[5] Patrick Cramer. "AlphaFold2 and the future of structural biology". In: *Nature structural & molecular biology* 28.9 (2021), pp. 704–705.

[6] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[7] Edo Dotan et al. "Multiple sequence alignment as a sequence-to-sequence learning problem". In: *The Eleventh International Conference on Learning Representations*. 2023. URL: https://openreview.net/forum?id=8efJYMBrNb.

[8] Khaled Fawagreh, Mohamed Medhat Gaber, and Eyad Elyan. "Random forests: from early developments to recent advancements". In: *Systems Science & Control Engineering: An Open Access Journal* 2.1 (2014), pp. 602–609.

[9] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. "ProtGPT2 is a deep unsupervised language model for protein design". In: *Nature communications* 13.1 (2022), p. 4348.

[10] Daria Grechishnikova. "Transformer neural network for protein-specific de novo drug generation as a machine translation problem". In: *Scientific reports* 11.1 (2021), pp. 1–13.

[11] Tori M Hoehler and Bo Barker Jørgensen. "Microbial life under extreme energy limitation". In: *Nature Reviews Microbiology* 11.2 (2013), pp. 83–94.

[12] Scott A Hollingsworth and Ron O Dror. "Molecular dynamics simulation for all". In: *Neuron* 99.6 (2018), pp. 1129–1143.

[13] Mingyang Hu et al. "Exploring evolution-based &-free protein language models as protein function predictors". In: *arXiv preprint arXiv:2206.06583* (2022).

[14] Jin Huang and Charles X Ling. "Using AUC and accuracy in evaluating learning algorithms". In: *IEEE Transactions on knowledge and Data Engineering* 17.3 (2005), pp. 299–310.

[15] W.M. Jacobs and E.I. Shakhnovich. "Evidence of evolutionary selection for cotranslational folding". In: *Pro-*

ceedings of the National Academy of Sciences 114.43 (2017), pp. 11434–11439.

[16] Bo Barker Jørgensen and Antje Boetius. "Feast and famine—microbial life in the deep-sea bed". In: *Nature Reviews Microbiology* 5.10 (2007), pp. 770–781.

[17] Bo Barker Jørgensen and Steven D'Hondt. "A starving majority deep beneath the seafloor". In: *Science* 314.5801 (2006), pp. 932–934.

[18] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873 (2021), pp. 583–589.

[19] Narine Kokhlikyan et al. *Captum: A unified and generic model interpretability library for PyTorch*. 2020. arXiv: 2009.07896 [cs.LG].

[20] Zeming Lin et al. "Evolutionary-scale prediction of atomic-level protein structure with a language model". In: *Science* 379.6637 (2023), pp. 1123–1130.

[21] Leland McInnes, John Healy, and James Melville. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[22] Ananthan Nambiar et al. "DR-BERT: A Protein Language Model to Annotate Disordered Regions". In: *bioRxiv* (2023), pp. 2023–02.

[23] Sarang Narkhede. "Understanding auc-roc curve". In: *Towards Data Science* 26.1 (2018), pp. 220–227.

[24] Gregory A Petsko and Dagmar Ringe. *Protein structure and function*. New Science Press, 2004.

[25] Roshan Rao et al. "Evaluating protein transfer learning with TAPE". In: *Advances in neural information processing systems* 32 (2019).

[26] Istvan Redl et al. "ADOPT: intrinsic protein disorder prediction through deep bidirectional transformers". In: *bioRxiv* (2022), pp. 2022–05.

[27] Alexander Rives et al. "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences". In: *Proceedings of the National Academy of Sciences* 118.15 (2021), e2016239118.

[28] Jenna M Schmidt et al. "Potential activities and long lifetimes of organic carbon-degrading extracellular enzymes in deep subsurface sediments of the Baltic sea". In: *Frontiers in Microbiology* (2021), p. 2546.

[29] Schrödinger, LLC. "The PyMOL Molecular Graphics System, Version 1.8". Nov. 2015.

[30] Jeffrey Skolnick and Jacquelyn S Fetrow. "From genes to protein structure and function: novel applications of computational approaches in the genomic era". In: *Trends in biotechnology* 18.1 (2000), pp. 34–39.

[31] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks". In: *International conference on machine learning*. PMLR. 2017, pp. 3319–3328.

[32] Baris E Suzek et al. "UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches". In: *Bioinformatics* 31.6 (2015), pp. 926–932.

[33] Felix Teufel et al. "SignalP 6.0 predicts all five types of signal peptides using protein language models". In: *Nature biotechnology* 40.7 (2022), pp. 1023–1025.

[34] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[35] Shaojun Wang et al. "NetGO 3.0: Protein Language Model Improves Large-scale Functional Annotations". In: *Genomics, Proteomics & Bioinformatics* (2023).

[36] Minghao Xu et al. "Peer: a comprehensive and multi-task benchmark for protein sequence understanding". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 35156–35173.

[37] Yong Yu et al. "A review of recurrent neural networks: LSTM cells and network architectures". In: *Neural computation* 31.7 (2019), pp. 1235–1270.

[38] Chulhee Yun et al. "Are transformers universal approximators of sequence-to-sequence functions?" In: *arXiv preprint arXiv:1912.10077* (2019).

[39] G. Zhang, M. Hubalewska, and Z. Ignatova. "Transient ribosomal attenuation coordinates protein synthesis and co-translational folding". In: *Nature Structural & Molecular Biology* 16.3 (2009), pp. 274–280.

[40] Shiyu Zhou et al. "Syllable-based sequence-to-sequence speech recognition with the transformer in mandarin chinese". In: *arXiv preprint arXiv:1804.10752* (2018).

[41] Laura A Zinke et al. "Microbial organic matter degradation potential in Baltic Sea sediments is influenced by depositional conditions and in situ geochemistry". In: *Applied and Environmental Microbiology* 85.4 (2019), e02164–18.