Peptide Sequencing Via Protein Language Models

Thuong Le Hoai Pham[*]
Jillur Rahman Saurav[†*]
Aisosa A. Omere[*†]
Calvin J. Heyl[*†]
Mohammad Sadegh Nasr[*]
Cody Tyler Reynolds[*]
Jai Prakash Yadav Veerla[*]
Helen H Shang[‡]
Justyn Jaworski[*]
Alison Ravenscraft[*]
Joseph Anthony Buonomo[*§]
joseph.buonomo@uta.edu
Jacob M. Luber[*§]
jacob.luber@uta.edu

ABSTRACT

We introduce a protein language model for determining the complete sequence of a peptide based on measurement of a limited set of amino acids. To date, protein sequencing relies on mass spectrometry, with some novel Edman degradation based platforms able to sequence non-native peptides. Current protein sequencing techniques face limitations in accurately identifying all amino acids, hindering comprehensive proteome analysis. Our method simulates partial sequencing data by selectively masking amino acids that are experimentally difficult to identify in protein sequences from the UniRef database. This targeted masking mimics real-world sequencing limitations. We then modify and finetune a ProtBERT derived transformer-based model, for a new downstream task predicting these masked residues, providing an approximation of the complete sequence. Evaluating on three bacterial *Escherichia* species, we achieve per-amino-acid accuracy up to 90.5% when only four amino acids ([KCYM]) are known. Structural assessment using AlphaFold2 and TM-score validates the biological relevance of our predictions. The model also demonstrates potential for evolutionary analysis through cross-species performance. This integration of simulated experimental constraints with computational predictions offers a promising avenue for enhancing protein sequence analysis, potentially accelerating advancements in proteomics and structural biology by providing a probabilistic reconstruction of the complete protein sequence from limited experimental data.

CCS CONCEPTS

• **Applied computing** → **Sequencing and genotyping technologies**; *Bioinformatics*; Computational proteomics.

[*]University of Texas at Arlington, Texas USA.
[†]These authors contributed equally to this research.
[‡]UCLA Health, Los Angeles, California, USA.
[§]To whom correspondence should be addressed.

ACM-BCB, Nov. 22–25, 2024, Shenzhen, Guangdong Province, PR China

ACM ISBN 979-8-4007-1302-6/24/11
https://doi.org/10.1145/3698587.3701385

KEYWORDS

Computational Biology, Protein Sequencing, High Performance Computing, Machine Learning, Language Modeling, Deep Learning

ACM Reference Format:
Thuong Le Hoai Pham, Jillur Rahman Saurav, Aisosa A. Omere, Calvin J. Heyl, Mohammad Sadegh Nasr, Cody Tyler Reynolds, Jai Prakash Yadav Veerla, Helen H Shang, Justyn Jaworski, Alison Ravenscraft, Joseph Anthony Buonomo, and Jacob M. Luber. 2024. Peptide Sequencing Via Protein Language Models. In *Proceedings of The 15th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3698587.3701385

1 INTRODUCTION

Protein sequences are fundamental to understanding biological processes, disease mechanisms, and therapeutic developments [1, 2]. Despite significant advancements in genomics and proteomics, aided by machine learning (ML) techniques [3, 4], accurate and comprehensive protein sequencing remains a challenge in the field [5].

Protein sequencing methods primarily rely on techniques such as Edman degradation [6] and mass spectrometry (MS) [7], including liquid chromatography tandem mass spectrometry (LC-MS/MS) [8]. While these methods have advanced our understanding of proteins, they face significant limitations in accurately identifying all amino acids in a sequence, particularly for complex or low-abundance proteins [5]. Studies have shown hat only about 16% of peptides in complex samples are typically identified by data-dependent mass spectrometry [9], and we show in Figure 1 that only being able to identify amino acid sets KCYM or KCYMRHWST without supplemental computational methods provides only 10% and 32% of the sequence information. These limitations often result in partially known sequences, hindering comprehensive proteome analysis.

Despite these advancements, protein sequencing still faces significant challenges, including high error rates, complex data interpretation, and technological limitations [10, 11, 12]. Overcoming these hurdles requires further advancements in sequencing technologies, sophisticated data processing algorithms, and improved experimental protocols to enhance accuracy, reproducibility, and scalability [13, 12].

Recent advancements in click chemistry and bioorthogonal chemistry [14, 15, 16] have attempted to address this issue by enabling the identification of specific amino acids and their positions. For instance, Zheng et al. demonstrated the sequencing of short antibody peptides using targeted amino acid labeling [17]. However,

these techniques are still limited by the number of amino acids that can be correctly identified, resulting in partially masked sequences (e.g., xCxxCxxx, where C is the experimentally identifiable amino acid) [18]. Additionally, the click chemistry platform demonstrated in Zheng et al. only works with non-native peptides that have undergone *a priori* chemical modifications[17]; limitations in this step means that parts of the proteomic retinue are not measurable with this approach. Our language model can work with input from this non-native peptide platform, as well as hypothetical future developments in bioorthogonal chemistry that will allow for Edman degradation of native peptides.

To address this specific limitation, we propose a novel approach leveraging pretrained language models. Large language models (LLMs) have shown remarkable adaptability in interpreting protein sequences, excelling in predicting structures, functions, and evolutionary relationships [19, 20, 21]. We hypothesize that these models can be used to predict the identity of amino acids that are conditionally difficult to determine experimentally.

In this paper, we present a method that simulates partial sequencing data by selectively masking amino acids that are experimentally challenging to identify in protein sequences from the UniRef database. This targeted masking mimics real-world sequencing limitations. We then utilize ProtBERT [22], a transformer-based model, to predict these masked residues, providing a probabilistic reconstruction of the complete protein sequence.

We evaluate our approach on three *Escherichia* bacterial species: *E. coli*, *E. albertii*, and *E. fergusonii*. Our results demonstrate high prediction accuracy even with extremely limited known amino acids. We also validate the biological relevance of our predictions through structural assessment using AlphaFold2 [23], and standard structure evaluation metrics such as template modeling score (TM-score)[24, 25] and the local distance difference test (lDDT)[26].

This innovative integration of simulated experimental constraints with computational predictions offers a promising avenue for enhancing protein sequence analysis. By improving our ability to interpret partially sequenced data, we aim to accelerate advancements in proteomics and structural biology, potentially unlocking new insights into protein structure and function.

The remainder of this paper is structured as follows: In the Methods section, we detail our data preparation, model fine-tuning process, and evaluation metrics. The Results section presents our comprehensive analysis of our model's performance across various scenarios. Finally, we discuss the implications of our findings and potential future directions in the Discussion section.

2 PROBLEM STATEMENT

In an assumption that the partial sequencing can be acquired from Edman degradation enhanced by click chemistry, which provides the positions and identities of a limited set of amino acids within a protein, we aim to predict the complete protein sequence. It has been established in the literature that on non-native peptides, "AFG" can be distinguished from "AWG" [17]; we have also demonstrated how this approach could theoretically be extended to non-native peptides in Supplementary Figure **??**. [27, 28] to fill in the gaps from unknown amino acids, given the context provided from the known ones in combination with the protein's domain constraint, determined at the species level. Our goal is to develop a computational approach that can accurately predict the full protein sequence from this partial information, potentially revolutionizing protein sequencing methodologies.

3 METHODS

3.1 Protein Dataset

For model training, which we conducted on 8 NVIDIA DGX A100 80GB cards, we utilized the UniProt Reference Clusters (UniRef) database[29], specifically UniRef100, focusing on three bacterial species: *Escherichia coli* (NCBI taxID 562), *Escherichia albertii* (NCBI taxID 208962), and *Escherichia fergusonii* (NCBI taxID 564). In addition, we included a fourth species: *Listeria monocytogenes* (NCBI taxID 1639), which is phylogenetically distant, to assess the model's generalizability and its applicability to evolutionary studies. These UniRef100 datasets combined identical sequences and subfragments with 11 or more residues into one entry, reducing potential data leakage between training, evaluation, and testing datasets. Additionally, we removed sequences from hypothetical protein group to ensure the biological relevance of the dataset. The detailed report of dataset sizes is shown in Supplement Table **??**.

Following the pretrained model's data processing, we mapped non–canonical or unresolved amino acids ([BOUZ]) to *unknown* (X)[22]. The frequency distribution of 20 canonical amino acids and *unknown* (X) extracted from the three Escherichia children is presented in Supplement Figure **??**, and from *L. monocytogenes* is presented in Supplement Figure **??**.

We propose working on two cases of targeted sets of amino acids. The first set (**KCYM**) contains amino acids with two or more publications supporting successful identification: Lysine (K)[30, 31, 32], Cysteine (C)[33, 34, 35], Tyrosine (Y)[36, 37, 38, 39], and Methionine (M)[40, 41]. The second set (**KCYMRHWST**) includes the amino acids from the first set, with additional amino acids that have at least one publication supporting successful identification: Arginine (R)[42], Histidine (H)[43], Tryptophan (W)[44], Serine (S)[33], and Threonine (T)[45].

3.2 Training Model

We chose ProtBERT[28] as our pretrained model due to its well performance in general tasks, lightweight nature (420M parameters), and bidirectional property. However, we modified the architecture of the model to use a masked language modeling head for our training task, compatible with our problem formulation. We trained one model per domain (species) and per set of amino acids, **KCYM** and **KCYMRHWST**, resulting in a total of eight finetuned and architecturally modified models. For *E. coli*, *E. albertii* and *L. monocytogenes*, we performed training and evaluation on 50k and 25k sequences, respectively. Due to data limitation, *E. fergusonii* was trained and evaluated on 40k and 5k sequences, respectively. The pretrained and architecturally modified model was then finetuned using HuggingFace transformers[46].

3.3 Evaluation Strategies

The performance of the model predictions were evaluated based on two major aspects: prediction accuracy and secondary structure similarity. For prediction accuracy, we computed three measures to

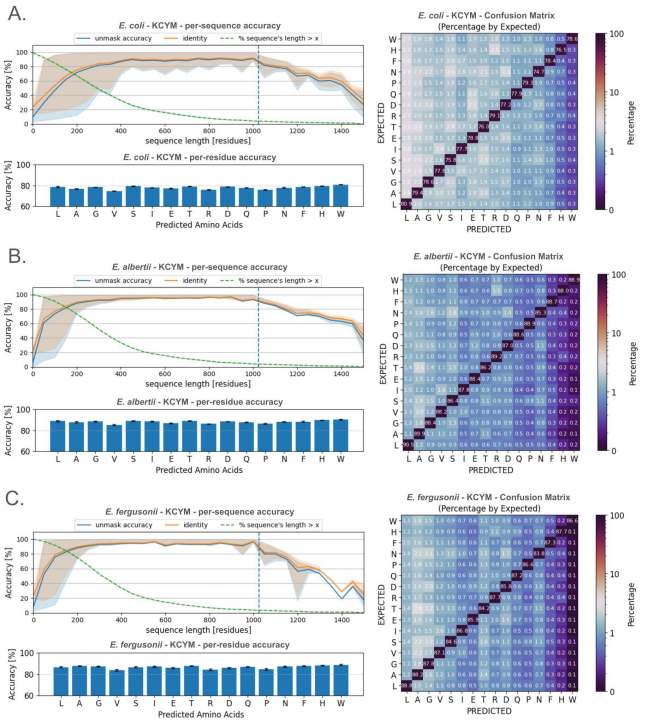

Figure 1: The accuracy per amino acid (right), and the per–sequence average accuracy of unmasking (top left) and identity (bottom left) of (A) *E. coli* (B) *E. albertii* and (C) *E. fergusonii* with known amino acids: KCYM

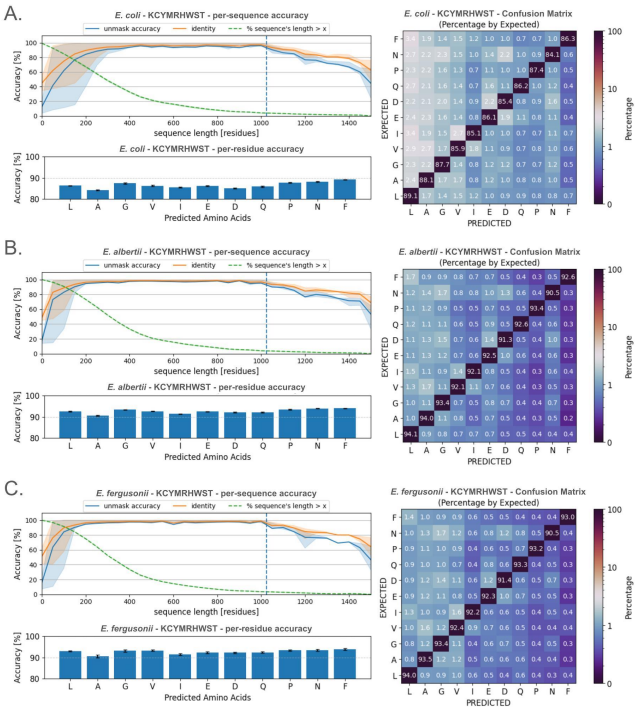

Figure 2: The accuracy per amino acid (right), and the per–sequence average accuracy of unmasking (top left) and identity (bottom left) of (A) *E. coli* (B) *E. albertii* and (C) *E. fergusonii* with known amino acids: KCYMTHWST

compare the primary sequence of the predicted and the true proteins: per–token accuracy, average per–sequence unmasking accuracy (i.e, excluded known amino acids), and average per–sequence total identity. Beside using an in–domain inference dataset to study the performance of models (see Figure 1 and 2), we also examined cross–domain accuracy among the three species. This aims to observe how taxonomic metrics (a prior knowledge about evolutionary distance in a phylogenic tree) correlates with the performance of our model predictions (see Table 1). For the two sets of amino acids (**KCYM** and **KCYMRHWST**), we performed testing inference on 5,000 sequences per species (randomly sampled from the inference dataset for 3 folds as shown in Supplement Table **??**).

To present useful amino acid suggestions/prioritization for experimental development in click chemistry based amino acid identification in the wet lab, we also performed training with amino acids from the small set and one additional amino acid from ([RHWST]), creating five additional study cases: **KCYMR**, **KCYMH**, **KCYMW**, **KCYMS**, and **KCYMT**. This configuration was only applied to the *E. coli* domain (with the same training protocols), and the inference was done using *E. coli* with the same inference setups as **KCYM** and **KCYMRHST** (see Table 2).

For the second aspect of measuring the quality of our predictions, we analyzed an important property of proteins: structure. AlphaFold2[23] is renowned for its high–accuracy prediction of protein three–dimensional structures from amino acid sequences using multiple sequence alignment in combination with a deep learning architecture. Recently, these AlphaFold2 predicted structures have been widely adapted inside large annotated databases, such as UniProt KnowledgeBase (UniProtKB)[47]. In this study, we used AlphaFold2 platform to examine how predicted sequences with less than 90% unmasking accuracy impact their structural integrity. Our study centered on sequences from the *E. coli* inference (fold–1), derived from the **KCYM** case, with unmasking accuracy bounded to the range [50–90]%. We used the AlphaFold Monomer v2.3.2 and its reduced database (with template date 2022-01-01) to generate structure predictions for our unmasked sequences. For true sequences (only those annotated in the UniProtKB), we retrieve the available AlphaFold2 structures from their database, which is last updated in AlphaFold DB version 2022-11-01, and created with the AlphaFold Monomer v2.0 pipeline. Filtering under these criteria yielded a total of 124 sequences for our structure analysis (see section 4.4). Supplement Figure **??** visualizes the protein structure derived from the predicted sequence and the actual UniProtKB sequence, these two structures overlaid, as well as the alignment between the predicted and actual amino acid sequence for one of these 124 proteins (UniProtKB ID A0A7H9QJ10). For additional validation on structures generated by AlphaFold2, we employed ESMFold v1.0, another protein structure prediction model with high accuracy [23, 48], on the sampled protein A0A7H9QJ10 for comparison in generated structures (see Supplement Figure **??**).

We computed the TM–score[24, 25] to compare the global similarity between the topologies of two structures. For local similarity, we computed the local difference distance test of the backbone atoms (lDDT–Cα)[26, 49] between the two structures, similar to the AlphaFold2 paper (see Figure 3).

4 RESULTS

4.1 Inference Accuracy

The accuracy of sequence predictions generated using the known amino acids set **KCYM** is presented in Figure 1, and the set **KCYMRHWST** is presented in Figure 2 for *Escherichia* species. As shown in the confusion heatmap for **KCYM**, even with a masking rate of 88.5%, the per–amino–acid accuracy reaches 74.7–80.9% in *E. coli*, 85.3–90.5% in *E. albertii*, and 83.8–88.8% in *E. fergusonii*. Accuracy analysis of *L. monocytogenes* is shown in Supplement Figure **??**.

The top left panel of Figure 1 shows that the average per–sequence accuracies (unmasking and identity) vs. sequence length, averaged per 50-residue bin and highlighted using the 75th percentile interval. The average per–sequence unmasking accuracy and identity are 73.53% and 76.75% for *E. coli*, 88.46% and 89.87% for *E. albertii*, 88.33% and 89.73% for *E. fergusonii*. The performance of the model decayed when the protein sequence length exceeds the model's maximum length of 1024 residues. This behavior is expected due to the property of the BERT model, which has linear positional embedding and the training maximum length is set to be 1024 residues. Note that only about 5% of sequences in the data had length exceeding this threshold.

After taking this into account, the line plots (left) indicate that the performance of the prediction is more stable and accurate for longer protein sequences. Specifically, with just the four known amino acids **KCYM**, the unmasking accuracy for sequences longer than 300 residues reached approximately 80% for *E. coli* and over 90% for *E. albertii* and *E. fergusonii*. However, it should be noted that only about half of the protein sequences in these species are longer than 300 residues.

In the case of knowing nine amino acids **KCYMRHWST** (see Figure 2), where the masking rate is 67.1%, the per–amino–acid accuracy reaches 84.1–89.1% in *E. coli*, 90.5–94.1% in *E. albertii*, and 90.5–94.0% in *E. fergusonii*. The average per–sequence unmasking accuracy and identity are 83.26% and 88.96% in *E. coli*, 93.38% and 95.62% in *E. albertii*, 93.49% and 95.69% in *E. fergusonii*. For proteins with length longer than 200 residues, representing 80% of protein sequences, the unmasking accuracy of *E. coli* exceeds 80%, while *E. albertii* and *E. fergusonii* exceed 90% accuracy.

4.2 Cross–species Performance

We evaluated the model's performance in capturing evolutionary information by cross–inferring each species' protein sequences using models trained on each other species. Our three species: *E. coli*, *E. albertii*, and *E. fergusonii*, are all members of the *Escherichia* genus, and thus are expected to share a significant amount of genetic information, indicating a decent homology in protein sequences.

As shown in Table 1, when the model only knows the four amino acids **KCYM**, the unmasking accuracy is high only when the training and inference are from the same domain (see section 4.1). The unmasking accuracy is significantly lower when the model tries predicting out–of–domain sequences, proportional to phylogenetic distance between domains. The results of the **KCYM** case reveal that, in the condition where the domain is specified, the model predictions only need a small set of known amino acids (in this case, **KCYM**) to capture the characteristics of the domain's protein sequences, achieving an average accuracy of at least 73%. However, with this size of amino acids set, our model fails to capture the nuance of sequences beyond the domain specified.

In the case of knowing nine amino acids **KCYMRHWST**, the in–domain unmasking accuracy increased by 5–10% compared to the previous **KCYM** case. Besides the high in–domain accuracy, the model predictions for out–of–domain sequences also performed much better, with the lowest accuracy at 64.02% when training on *E. fergusonii* and inferring on *E. coli*, and the highest accuracy at 82.35% when training on *E. coli* and inferring on *E. albertii*.

Overall, the model trained on *E. coli* performed best on out–of–domain inference, followed by *E. albertii*, and lastly *E. fergusonii*. This outcome is expected due to the high yield of protein sequences available from *E. coli* and *E. albertii* compared to *E. fergusonii*. In summary, the cross–inference results indicate that knowledge of the species to which a sequence belongs increases prediction accuracy. Furthermore, they demonstrate the model's potential for predicting protein sequences based on another related species when the sequences' species identity may not be known.

4.3 Generalizability

From previous results (see section 4.1 and 4.2), our work suggests that transformer models like BERT can predict protein sequences with high accuracy, given prior knowledge of limited sets of amino acids and the species domain. Additionally, the accuracy of the predictions increases significantly with a larger set of known amino acids. However, expanding the set of identifiable amino acids introduces exponential challenges in Edman degradation. This process requires peptides to undergo more chemical identification cycles, leading to an increased noise in sequencing and a higher risk of unstable peptide degradation. Therefore, we investigated our model performance on sequence prediction by using five different sets of known amino acids as a guide for prioritizing amino acids to develop future click chemistry based identification for; in essence we are comparing how unmasking new amino acids ameliorate model performance, and comparing these results to prioritize future wet lab efforts. We evaluated the inference of five additional models, which are trained on five known amino acids: four being **KCYM** and one from the set ([RHWST]) amino acids (see Table 2).

Among the five experiments, the one with **KCYMS** has the highest sequence coverage from known amino acids, at 17.69% (corresponding to 82.31% masking rate). However, the case of **KCYMR** demonstrates the best, with average per–sequence unmasking accuracy at 76.18% (2.65% more than **KCYM**), average per–sequence identity at 80.48% (3.73% more), and per–token accuracy at 80.11% (1.94% more). The other four cases show comparable results to **KCYMR** (within 2% differences).

4.4 Structure Analysis

The comparison of lDDT–Cα vs. TM–score between the predicted and true sequences' AlphaFold2 structures is shown in Figure 3, in

Table 1: Inference unmasking accuracy and standard deviation across species and sets of amino acids

finetuning model on known tokens: KCYM (3 folds)					
	Pretrained model	***E. coli* model**	***E. albertii* model**	***E. fergusonii* model**	***L. monocytogenes* model**
E. coli	7.28% (0.01)	**73.53% (0.49)**	51.38% (0.17)	45.22% (0.09)	7.58% (0.03)
E. albertii	7.35% (0.02)	65.15% (0.49)	**88.46% (0.18)**	50.16% (0.64)	7.62% (0.03)
E. fergusonii	7.35% (0.03)	60.06% (0.35)	50.61% (0.34)	**88.33% (0.15)**	7.65% (0.01)
L. monocytogenes	6.58% (0.03)	7.60% (0.03)	7.70% (0.002)	7.67% (0.01)	**88.22% (0.22)**
finetuning model on known tokens: KCYMRHWST (3 folds)					
	Pretrained model	***E. coli* model**	***E. albertii* model**	***E. fergusonii* model**	***L. monocytogenes* model**
E. coli	9.94% (0.04)	**83.26% (0.45)**	69.08% (0.33)	64.02% (0.17)	18.18% (0.08)
E. albertii	9.89% (0.02)	82.35% (0.45)	**93.38% (0.05)**	71.04% (0.67)	19.07% (0.16)
E. fergusonii	9.79% (0.05)	79.58% (0.46)	72.49% (0.18)	**93.49% (0.13)**	19.29% (0.01)
L. monocytogenes	9.57% (0.01)	21.66% (0.10)	20.29% (0.17)	19.77% (0.20)	**91.92% (0.21)**

Table 2: Inference accuracy report and masking ratio of different sets of known amino acids

	KCYM	**KCYMR**	**KCYMH**	**KCYMW**	**KCYMS**	**KCYMT**	**KCYMRHWST**
Per–token acc. [%]	78.17% (0.37)	**80.11% (0.63)**	78.67% (0.70)	79.35% (0.71)	78.58% (0.71)	80.07% (0.69)	86.82% (0.27)
Per–seq acc. [%]	73.53% (0.60)	**76.18% (0.81)**	74.81% (0.94)	75.61% (0.91)	75.10% (0.89)	75.52% (0.90)	83.26% (0.55)
Per–seq identity [%]	76.75% (0.53)	**80.48% (0.65)**	78.41% (0.79)	78.95% (0.78)	79.76% (0.73)	79.90% (0.73)	88.96% (0.37)
Masking ratio [%]	88.47% (0.01)	82.74% (0.09)	86.27% (0.07)	86.91% (0.05)	**82.31% (0.07)**	82.91% (0.03)	67.09% (0.03)

which the left panel is colored by per-sequence unmasking accuracy, and the right panel is colored by sequence length.

The TM–score evaluates the global similarity between two structures, while the lDDT–Cα assesses the local distances of the backbones. According to the plot, the high lDDT–Cα can happens even with low TM–score, but not the opposite where the TM–score is high but lDDT–Cα is low. This is often caused by the structures having the large difference in the bending angles at the coil regions, yield a divergence in the structure's global shapes, and hence resulted in low TM–score. But the local structure conformations, such as alpha–helices and beta–sheets, are conserved, leading to high value of lDDT–Cα. An AlphaFold2 and ESMFold example result (UniProtKB ID: A0A7H9QJ10) is presented in Figure **??**, showing the molecular view of two structures (using py3Dmol[50]) and their pairwise alignment, colored by unmasking matches (green) and mismatches (red). An additional illustrative example of a different protein is presented in the appendix.

To know how our model prediction's quality (measured by unmasking accuracy) impacts the predicting protein structure, we need to understand how the value of TM–score approximately corresponds to whether the protein pairs sharing the same topology. Xu et al.'s paper, studied on the CATH and SCOP databases, reported that the high posterior probability of two structures having the same topology corresponds to a TM–score roughly between 0.4 and 0.6, with the specific threshold varies by datasets[24]. In our structure results, reported from 124 samples of *E. coli* with known set as **KCYM**, we also observed the decrease in general lDDT–Cα values when the TM–score lower than 0.6, and hence 0.6 is our evaluation threshold. This means for TM–score > 0.6, we have a high statistical confidence that the two structures are the same topology. And for TM–score < 0.6, we need to evaluate auxiliary metrics such as lDDT–Cα, unmasking accuracy, sequence length, etc. to conclude the similarity in topology.

The Figure 3's left panel suggests that we have a high confidence in structure similarity between our predicted sequences and true sequences (TM–score > 0.6) when the unmasking accuracy is above 75%. For outliers where the unmasking accuracy > 75% but TM–score < 0.6, we notices that their sequence length are often long (see Figure 3's right panel). Because of the sequence length, these protein are thus expected to have higher chance having local divergence, leading to a sensitive TM–score, but the high lDDT–Cα.

While intuition may suggest that the low accuracy predictions have low structure similarity, it is not the case. For TM–core > 0.6, many of the sequences has unmasking accuracy < 75%, and some are even less than 65%. These sequences are observed to often have lower lDDT–Cα compared to ones with high accuracy.

5 FUTURE DIRECTIONS

Peptide sequencing enabled by our language model will have many important implications for the development of liquid biopsies, which could yield more information for treatment decisions in the oncology clinic. Liquid biopsy is a minimally invasive tool to identify cancer biomarkers within fluids such as blood plasma and urine. These liquid samples have been readily explored as sources of nucleotide biomarkers such as non-coding RNAs and tumor-specific DNA, but creating diagnostics based on proteins has been limited by

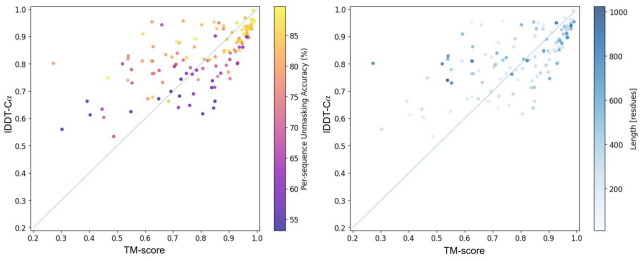

Figure 3: Cα LDDT vs TM–score generated from comparing *E. coli* – KCYM's predicted vs real sequence's AlphaFold2 structures, colored in unmasking accuracy (left) and sequence length (right)

signal to noise ratios for the detection of low-abundance hits, and difficulty discerning the source of proteins to cancer- specific cells without first isolating the circulating cancer cells [51]. However, liquid biopsies are advantageous due to their safety, high repeatability, ability to monitor disease progression and prognosis, all without the need for an inpatient procedure [52].

There is a significant technological gap between current diagnostic assays and the proteoform resolution necessary to characterize and quantifiably identify cancer-specific biomarkers and prognosis indicators [53]. Taken together, an improvement in proteoform identification and quantification with resolution to the single molecule would foster rapid development and implementation of utilizing well- studied proteoforms as both diagnostic and prognosis biomarkers. Optimally, such a technology would enable the detection of protein analytes in fecal, urine, or plasma samples.

Thus, in the future, we aim to expand peptide sequencing via the language model presented in this paper to develop a platform to directly sequence proteins within a complex milieu through highly-specific chemical ligation of amplifiable DNA barcodes for amino acid identity, sequence position, and peptide identity which provides a quantifiable readout with higher sensitivity than mass spectrometry alone [54][5]. This future platform will driven by closely entwined advancements in machine learning algorithm design and chemical reaction development and characterization.

6 DISCUSSION

We present a protein language model designed to determine the complete sequence of a peptide based on the measurement of a limited set of amino acids. Traditional protein sequencing primarily relies on mass spectrometry, with some novel Edman degradation-based platforms capable of sequencing non-native peptides. However, these techniques face significant limitations in accurately identifying all amino acids, thus hindering comprehensive proteome analysis. Our approach simulates partial sequencing data by selectively masking amino acids that are experimentally challenging to identify in protein sequences from the UniRef database, thereby mimicking real-world sequencing limitations. By modifying and fine-tuning a ProtBERT-derived transformer-based model, we predict these masked residues, providing an approximation of the complete sequence. Unlike traditional multiple sequence alignment (MSA) approaches, our model views sequence data as partial sequences, providing a new perspective and methodology for protein sequence analysis.

Our method, evaluated on three bacterial *Escherichia* species, achieves per-amino-acid accuracy of up to 90.5% when only four amino acids ([KCYM]) are known. Structural assessments using AlphaFold2 and TM-score validate the biological relevance of our predictions, and the model demonstrates potential for evolutionary analysis through cross-species performance. This integration of simulated experimental constraints with computational predictions offers a promising avenue for enhancing protein sequence analysis. By improving our ability to interpret partially sequenced data, our approach has the potential to accelerate advancements in proteomics and structural biology, enabling a probabilistic reconstruction of complete protein sequences from limited experimental data.

Oxford Nanopore's DNA/RNA sequencing platform, which makes inferences from incomplete signal (squiggles) and converts them algorithmically to sequence space, initially performed with lower accuracy when first introduced [55, 56] than our model in terms of accuracy of inferred sequence. This suggests that our computational model paired with a few additional wet lab improvements has the potential to yield the first clinically useful protein sequencing platform.

However, several challenges remain. Experimental verification is essential to validate our computational predictions and ensure their biological relevance. Additionally, successfully implementing the hypothetical Edman degradation pipeline requires effective peptide immobilization techniques without C-terminus modification. Overcoming these hurdles will be crucial for the practical application of our method.

This integration of simulated experimental constraints with computational predictions offers a promising avenue for enhancing protein sequence analysis. By improving our ability to interpret partially sequenced data, we aim to accelerate advancements in proteomics and structural biology, potentially unlocking new insights into protein structure and function.

The future directions for our research involve expanding peptide sequencing via our language model to develop a platform capable of directly sequencing proteins within complex milieus. This will involve highly specific chemical ligation of amplifiable DNA barcodes for amino acid identity, sequence position, and peptide identity, providing a quantifiable readout with higher sensitivity than mass spectrometry alone [54][5]. Such advancements will drive closely intertwined developments in machine learning algorithm design and chemical reaction characterization, ultimately fostering rapid implementation of proteoform-based diagnostics and prognostics.

In summary, our computational approach, validated through predicted structures generated using AlphaFold2 and evaluated using TM-score and lDDT–Cα, demonstrates significant potential for improving protein sequence analysis. By integrating these predictions with experimental techniques, we aim to bridge the gap between current technologies and the high-resolution identification required for advanced proteomics.

ACKNOWLEDGMENTS

This work was supported by the Cancer Prevention and Research Institute of Texas First Time Faculty Award (J.M.L.) and a University of Texas System Rising STARs Award (J.M.L.). We thank Yash Tobre for assistance with data curation.

7 CODE AVAILABILITY

Please use the following URL to access the code and model checkpoints: https://github.com/jacobluber/protein-sequencing-LLMs.

REFERENCES

[1] Elizabeth L Lieu, Tu Nguyen, Shawn Rhyne, and Jiyeon Kim. 2020. Amino acids in cancer. *Experimental & molecular medicine*, 52, 1, 15–30.

[2] Oliver DK Maddocks et al. 2017. Modulating the therapeutic response of tumours to dietary serine and glycine starvation. *Nature*, 544, 7650, 372–376.

[3] Maxwell W Libbrecht and William Stafford Noble. 2015. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16, 6, 321–332.

[4] Bo Wen, Wen-Feng Zeng, Yuxing Liao, Zhiao Shi, Sara R Savage, Wen Jiang, and Bing Zhang. 2020. Deep learning in proteomics. *Proteomics*, 20, 21-22, 1900335.

[5] Javier Antonio Alfaro et al. 2021. The emerging landscape of single-molecule protein sequencing technologies. *Nature methods*, 18, 6, 604–617.

[6] Hugh D Niall. 1973. [36] automated edman degradation: the protein sequenator. In *Methods in enzymology*. Vol. 27. Elsevier, 942–1010.

[7] Donald F Hunt, JR Yates 3rd, Jeffrey Shabanowitz, Scott Winston, and Charles R Hauer. 1986. Protein sequencing by tandem mass spectrometry. *Proceedings of the National Academy of Sciences*, 83, 17, 6233–6237.

[8] M Vogeser and KG Parhofer. 2007. Liquid chromatography tandem-mass spectrometry (lc-ms/ms)-technique and applications in endocrinology. *Experimental and clinical endocrinology & diabetes*, 115, 09, 559–570.

[9] Annette Michalski, Juergen Cox, and Matthias Mann. 2011. More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent lc- ms/ms. *Journal of proteome research*, 10, 4, 1785–1793.

[10] Matthew Beauregard Smith, Kent VanderVelden, Thomas Blom, Heather D Stout, James H Mapes, Tucker M Folsom, Christopher Martin, Angela M Bardo, and Edward M Marcotte. 2024. Estimating error rates for single molecule protein sequencing experiments. *PLOS Computational Biology*, 20, 7, e1012258.

[11] Mike Filius et al. 2024. Full-length single-molecule protein fingerprinting. *Nature Nanotechnology*, 1–8.

[12] Brian C Searle. 2024. Nanopore protein sequencing achieves significant new milestones. *Clinical Chemistry*, hvae041.

[13] Morgan M Brady and Anne S Meyer. 2022. Cataloguing the proteome: current developments in single-molecule protein sequencing. *Biophysics Reviews*, 3, 1.

[14] Bernhard Stump. 2022. Click bioconjugation: modifying proteins using click-like chemistry. *ChemBioChem*, 23, 16, e202200016.

[15] Oleksandr Koniev and Alain Wagner. 2015. Developments and recent advancements in the field of endogenous amino acid selective bond forming reactions for bioconjugation. *Chemical Society Reviews*, 44, 15, 5495–5551.

[16] Samuel L Scinto et al. 2021. Bioorthogonal chemistry. *Nature Reviews Methods Primers*, 1, 1, 30.

[17] Liwei Zheng, Yujia Sun, Michael Eisenstein, and Hyongsok Tom Soh. 2024. Peptide sequencing via reverse translation of peptides into dna. *bioRxiv*. eprint: https://www.biorxiv.org/content/early/2024/06/03/2024.05.31.596913.full.pdf. DOI: 10.1101/2024.05.31.596913.

[18] Jagannath Swaminathan, Alexander A Boulgakov, Erik T Hernandez, Angela M Bardo, James L Bachman, Joseph Marotta, Amber M Johnson, Eric V Anslyn, and Edward M Marcotte. 2018. Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. *Nature biotechnology*, 36, 11, 1076–1082.

[19] Tristan Bepler and Bonnie Berger. 2021. Learning the protein language: evolution, structure, and function. *Cell systems*, 12, 6, 654–669.

[20] Jeffrey A Ruffolo and Ali Madani. 2024. Designing proteins with language models. *nature biotechnology*, 42, 2, 200–202.

[21] Liuzhenghao Lv, Zongying Lin, Hao Li, Yuyang Liu, Jiaxi Cui, Calvin Yu-Chian Chen, Li Yuan, and Yonghong Tian. 2024. Prollama: a protein large language model for multi-task protein language processing. *arXiv preprint arXiv:2402.16445*.

[22] Ahmed Elnaggar et al. 2021. Prottrans: towards cracking the language of lifes code through self-supervised deep learning and high performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1. DOI: 10.1109/TPAMI.2021.3095381.

[23] John Jumper et al. 2021. Highly accurate protein structure prediction with alphafold. *Nature*, 596, 7873, 583–589.

[24] Jinrui Xu and Yang Zhang. 2010. How significant is a protein structure similarity with tm-score= 0.5? *Bioinformatics*, 26, 7, 889–895.

[25] Yang Zhang and Jeffrey Skolnick. 2005. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research*, 33, 7, 2302–2309.

[26] Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. 2013. Lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29, 21, 2722–2728.

[27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[28] Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. 2022. ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, 38, 8, (Feb. 2022), 2102–2110. eprint: https://academic.oup.com/bioinformatics/article-pdf/38/8/2102/49009610/btac020.pdf. DOI: 10.1093/bioinformatics/btac020.

[29] Baris E. Suzek, Yuqi Wang, Hongzhan Huang, Peter B. McGarvey, Cathy H. Wu, and the UniProt Consortium. 2014. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31, 6, (Nov. 2014), 926–932. eprint: https://academic.oup.com/bioinformatics/article-pdf/31/6/926/49011550/bioinformatics_31_6_926.pdf. DOI: 10.1093/bioinformatics/btu739.

[30] Ajcharapan Tantipanjaporn and Man-Kin Wong. 2023. Development and recent advances in lysine and n-terminal bioconjugation for peptides and proteins. *Molecules*, 28, 3, 1083.

[31] George W Anderson, Joan E Zimmerman, and Francis M Callahan. 1964. The use of esters of n-hydroxysuccinimide in peptide synthesis. *Journal of the American Chemical Society*, 86, 9, 1839–1842.

[32] George W Anderson, Joan E Zimmerman, and Francis M Callahan. 1963. N-hydroxysuccinimide esters in peptide synthesis. *Journal of the American Chemical Society*, 85, 19, 3039–3039.

[33] Julien C Vantourout et al. 2020. Serine-selective bioconjugation. *Journal of the American Chemical Society*, 142, 41, 17236–17242.

[34] Kévin Renault, Jean Wilfried Fredy, Pierre-Yves Renard, and Cyrille Sabot. 2018. Covalent modification of biomolecules through maleimide-based labeling strategies. *Bioconjugate Chemistry*, 29, 8, 2497–2513.

[35] Gregory A Grant. 2017. Modification of cysteine. *Current protocols in protein science*, 87, 1, 15–1.

[36] Hitoshi Ban, Masanobu Nagano, Julia Gavrilyuk, Wataru Hakamata, Tsubasa Inokuma, and Carlos F Barbas III. 2013. Facile and stabile linkages through tyrosine: bioconjugation strategies with the tyrosine-click reaction. *Bioconjugate chemistry*, 24, 4, 520–532.

[37] Tanzeela Abdul Fattah, Aamer Saeed, and Fernando Albericio. 2018. Recent advances towards sulfur (vi) fluoride exchange (sufex) click chemistry. *Journal of Fluorine Chemistry*, 213, 87–112. DOI: https://doi.org/10.1016/j.jfluchem.2018.07.008.

[38] Peter A Szijj, Kristina A Kostadinova, Richard J Spears, and Vijay Chudasama. 2020. Tyrosine bioconjugation–an emergent alternative. *Organic & Biomolecular Chemistry*, 18, 44, 9018–9028.

[39] Feng Liu, Hua Wang, Suhua Li, Grant AL Bare, Xuemin Chen, Chu Wang, John E Moses, Peng Wu, and K Barry Sharpless. 2019. Biocompatible sufex click chemistry: thionyl tetrafluoride (sof4)-derived connective hubs for bioconjugation to dna and proteins. *Angewandte Chemie International Edition*, 58, 24, 8029–8033.

[40] Shixian Lin et al. 2017. Redox-based reagents for chemoselective methionine bioconjugation. *Science*, 355, 6325, 597–602.

[41] Jia Zang, Yulin Chen, Wenxuan Zhu, and Shixian Lin. 2019. Chemoselective methionine bioconjugation on a polypeptide, protein, and proteome. *Biochemistry*, 59, 2, 132–138.

[42] Maheshika SK Wanigasekara, Xiaojun Huang, Jayanta K Chakrabarty, Alejandro Bugarin, and Saiful M Chowdhury. 2018. Arginine-selective chemical labeling approach for identification and enrichment of reactive arginine residues in proteins. *ACS omega*, 3, 10, 14229–14235.

[43] Chuan Wan et al. 2022. Histidine-specific bioconjugation via visible-light-promoted thioacetal activation. *Chemical Science*, 13, 28, 8289–8296.

[44] Klaas W Decoene, Kamil Unal, An Staes, Olivier Zwaenepoel, Jan Gettemans, Kris Gevaert, Johan M Winne, and Annemieke Madder. 2022. Triazolinedione protein modification: from an overlooked off-target effect to a tryptophan-based bioconjugation strategy. *Chemical Science*, 13, 18, 5390–5397.

[45] Alexandra M Webster, Christopher R Coxon, Alan M Kenwright, Graham Sandford, and Steven L Cobb. 2014. A mild method for the synthesis of a novel dehydrobutyrine-containing amino acid. *Tetrahedron*, 70, 31, 4661–4667.

[46] Thomas Wolf et al. 2019. Huggingface's transformers: state-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

[47] 2023. Uniprot: the universal protein knowledgebase in 2023. *Nucleic acids research*, 51, D1, D523–D531.

[48] Zeming Lin et al. 2022. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*.

[49] Marco Biasini et al. 2013. Openstructure: an integrated software framework for computational structural biology. *Acta Crystallographica Section D: Biological Crystallography*, 69, 5, 701–709.
[50] David Koes et al. 2020. Py3dmol: a python interface for 3dmol.js. Accessed: 2024-07-09. (2020). https://pypi.org/project/py3Dmol/.
[51] Michele Marchioni et al. 2021. Biomarkers for renal cell carcinoma recurrence: state of the art. *Current Urology Reports*, 22, 6, 31.
[52] Paulo G Bergerot, Andrew W Hahn, Cristiane Decat Bergerot, Jeremy Jones, and Sumanta Kumar Pal. 2018. The role of circulating tumor dna in renal cell carcinoma. *Current treatment options in oncology*, 19, 1–11.
[53] Ashley Di Meo, Ihor Batruch, Marshall D Brown, Chuance Yang, Antonio Finelli, Michael A Jewett, Eleftherios P Diamandis, and George M Yousef. 2020. Searching for prognostic biomarkers for small renal masses in the urinary proteome. *International journal of cancer*, 146, 8, 2315–2325.
[54] Winston Timp and Gregory Timp. 2020. Beyond mass spectrometry, the next step in proteomics. *Science Advances*, 6, 2, eaax8978.
[55] Clive G Brown and James Clarke. 2016. Nanopore development at oxford nanopore. *Nature biotechnology*, 34, 8, 810–811.
[56] Thomas Laver, J Harrison, PA O'neill, Karen Moore, Audrey Farbos, Konrad Paszkiewicz, and David J Studholme. 2015. Assessing the performance of the oxford nanopore technologies minion. *Biomolecular detection and quantification*, 3, 1–8.

Received 15 July 2024