Training self-supervised peptide sequence models on artificially chopped proteins

Gil Sadeh[*]
Amazon Devices
gilsadeh@amazon.com

Zichen Wang[*]
Amazon Machine Learning Solutions Lab
zichewan@amazon.com

Jasleen Grewal[*]
AWS Professional Services
gjaslee@amazon.com

Huzefa Rangwala
Amazon Machine Learning Solutions Lab
rhuzefa@amazon.com

Layne Price[†]
Amazon Devices
prilayne@amazon.com

Abstract

Representation learning for proteins has primarily focused on the global understanding of protein sequences regardless of their length. However, shorter proteins (known as peptides) take on distinct structures and functions compared to their longer counterparts. Unfortunately, there are not as many naturally occurring peptides available to be sequenced and therefore less peptide-specific data to train with. In this paper, we propose a new peptide data augmentation scheme, where we train peptide language models on artificially constructed peptides that are small contiguous subsets of longer, wild-type proteins; we refer to the training peptides as "chopped proteins". We evaluate the representation potential of models trained with chopped proteins versus natural peptides and find that training language models with chopped proteins results in more generalized embeddings for short protein sequences. These peptide-specific models also retain information about the original protein they were derived from better than language models trained on full-length proteins. We compare masked language model training objectives to three novel peptide-specific training objectives: next-peptide prediction, contrastive peptide selection and evolution-weighted MLM. We demonstrate improved zero-shot learning performance for a deep mutational scan peptides benchmark.

1 Introduction

Proteins are integral to all cellular functions in living organisms. Proteins of length less than 50 amino acids, called peptides, can have different functional properties from larger proteins partly due to physical characteristics (shorter length and simpler three-dimensional folded structures) and partly due to their distinct biological or therapeutic roles[13, 20]. Naturally occurring peptides typically act as chemical messengers within and outside cells by interacting with other molecules like cellular receptors or antigens [2]. They also play key roles as interaction partners and messengers in the immune system of multicellular organisms [19]. Given their functional properties and ease of

[*]These authors contributed equally
[†]Corresponding author

Machine Learning for Structural Biology Workshop, NeurIPS 2022.

modification using laboratory techniques, peptide drug development has emerged as a promising area of therapeutic research and development [31].

While there has been several prior works in self-supervised protein models [24, 11, 12, 15, 1, 5, 14, 17, 9, 22], unfortunately the ability of these models to represent peptide sequences is understudied, as there are relatively few known naturally occurring peptides to experiment with. Few works [7, 25] have focused on peptide sequences, training transformer and convolutional models to predict peptide detectability by mass spectrometry. However, these methods involve supervised training over mass-spec data, and do not explore leveraging large-scale proteomic data with self-supervised learning.

In this work we train self-supervised learning models for peptides inspired from natural language modeling with the input representation being the primary sequence. We propose a new peptide-specific data augmentation framework, which we call "chopped proteins." Our data augmentation strategy involves simulating the primary sequence of novel peptides by randomly sampling contiguous subsets of longer protein primary sequences available in common databases, such as UniRef[27]. This sub-sampling scheme is inspired by intracellular proteasomal cleavage mechanisms[28], where proteins are degraded by being chopped into smaller peptides by protease enzymes. We train and evaluate the peptide models on chopped proteins using different training objectives. We use standard masked language modeling (MLM) [10] objective, and explore additional alternatives, exploiting pairwise peptide relationships and prior evolutionary information.

2 Methods

Baseline model, ESM1b.— ESM1b [24] is a protein transformer model, composed of 33 layers with embedding dimension $d = 1280$. It was trained over 30M cluster representative amino-acid sequences (i.e. primary structures), from the UniRef50 [27] dataset, using BERT-like MLM self-supervised training. ESM1b has previously achieved state-of-the-art performance, in comparison to competitive sequence based models, in mutational effect, secondary structure, and long-range contact prediction tasks. In addition, the authors have released the pretrained model weights [3]. Hence, we have chosen to use this model as our baseline, while utilizing the same architecture and initialization from the provided pretrained weights for our peptide-based training.

Chopped proteins from UniRef50.— In order to apply large scale self-supervised training for short length peptide sequences, we utilize available large protein data, and randomly "chop" small contiguous subsets from the given long protein sequences, as illustrated in Figure 3a. When "chopping", peptide sequence length is uniformly sampled from 8 to 50. As in ESM1b, we also utilize the UniRef50 [27] dataset. UniRef database provides clustered sets of sequences (from UniProt and UniParc) to obtain complete coverage of sequence space at several resolutions. UniRef50 consists of clustered sequences at 50% sequence identity level while hiding redundant sequences. Like ESM1b, we only utilize cluster representative sequences, which contain the most biological information. This dataset, also known as UniRef50-sparse (UR50-S), contains 30M cluster representative protein sequences. We use the test set provided by ESM1b and randomly split the remaining sequences between training (90%) and validation (10%) sets.

Masked Language Modeling (MLM).— This objective, originally proposed in BERT [10], is commonly used for learning language and protein representations [24, 11, 15, 1]. For each sequence x, we sample a set of indices M to mask, with probability p=0.15, replacing the true token at each index i of the input sequence with either (1) the <MASK> token (p=0.8); (2) a random amino acid token (p=0.1); or (3) the unchanged i^{th} token (p=0.1). For each index $i \in M$, we independently minimize the negative log likelihood of the true amino acid x_i given the masked input sequence $x_{/M}$ as context: $\mathcal{L}_{MLM} = -\mathbb{E}_M[\sum_{i \in M}(\log p(\hat{x}_i = x_i | x_{/M})]$, where $\log p(\hat{x}_i = x_i | x_{/M})$ is the logit corresponding to the true i^{th} token. Figure 3a illustrates applying MLM on chopped proteins.

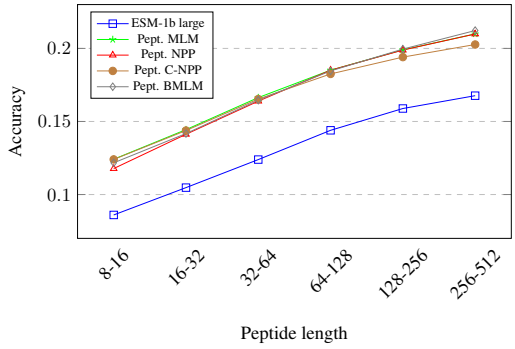
Alternative training objectives.— We additionally explore several alternative training objectives. We consider complementing the MLM objective with an additional, auxiliary, sequence-level, objective for learning pairwise peptide relationship. We propose the 'Next-Peptide-Prediction' (NPP) objective, which is equivalent to BERT's 'Next-Sentence-Prediction' objective, and an additional

[3] `https://github.com/facebookresearch/esm`

Table 1: **Exponentiated Cross Entropy (ECE) results for sequence models using an MLM or derivative objective.** Lower value is better, lowest value for each trained model is shown in bold.

Evaluation set	ESM1b	Pept. MLM	Pept. NPP	Pept. C-NPP	Pept. BMLM
peptides	6.6144±0.0073	6.3136±0.0063	**6.3131**±0.0074	6.3386±0.0048	6.5331±0.0071
proteins	4.8184±0.0087	4.4207±0.0095	**4.3937**±0.0061	4.5204±0.0020	4.5523±0.0049

Figure 1: **Peptide context prediction accuracy over different peptide length bins.** Average accuracy across 4 masked out context residues, beyond the peptide's edges, is shown.



contrastive extension of it (C-NPP). In addition, we propose the BLOSUM MLM (BMLM) objective, which incorporates evolutional information utilizing the Blosum62 substitution matrix[16] into the MLM training. These objectives are described in detail in appendix section .1.

3 Results

Unless stated otherwise, all our peptide models use the ESM1b backbone architecture and initialization, and were trained similarly, for 25 epochs over protein sequences, saving the best performing checkpoint over the validation set. We used Adam optimizer, with 0.0001 base learning rate, inverse square-root decay schedule, 10K warmup steps and an effective batch size of 8192 sequences.

Peptide models have improved generalization to hold-out peptides and proteins— The most straightforward evaluation for trained sequence models is to assess their performance on hold-out sequences using the exponentiated cross entropy (ECE) metric. ECE is the exponential of the model's MLM loss averaged per token ($2^{\mathcal{L}_{MLM}}$). The ECE metric is analogous to perplexity commonly used for evaluating language models. We report ECEs for all models over two subsets of UR50-S test set composed of peptides (short protein sequences) and proteins (of all lengths), respectively (Table 1). The peptide models achieved markedly better ECEs than ESM1b on both evaluation sets.

Peptide models predict amino acids flanking a peptide sequence— By padding peptide sequences with mask tokens, and attempting to recover the correct amino-acids found in corresponding positions in the origin protein sequences, we evaluate model's ability to predict peptide context (also referred to as peptide's flanking regions). Despite similarity to MLM, this task is more challenging due to the selection of consecutive regions on the edges as the masked tokens. In addition, this task has biological significance as a peptide's context plays an important role in proteasomal cleavage [28]. The evaluation was performed on randomly chopped sequences from a subset of long UR50-S test sequences on various sequence length bins, as shown in Figure 1. The results suggest that our peptide transformer models significantly outperform the baseline ESM1b protein model across short and long sequences. However, no statistically significant performance difference was found between the various training objectives. Further, it is important to note that context prediction accuracy consistently improves as sequence length increases, with all models, despite peptide transformer models being

Table 2: **Zero-shot mutational effect analysis (experimental measurement modality).** Average Spearman's ρ is shown, aggregated on the measurement modality of the DMS experiments. Number of studies for each modality shown in brackets. 'Winning' model for each modality is highlighted.

Measurement	ESM1b	Pept. BMLM	Pept. C-NPP	Pept. MLM	Pept. NPP
E1 reactivity (1)	0.1839	0.2124	0.2872	0.3332	0.3034
Enzyme function (3)	0.4290	0.4375	0.4258	0.4399	0.4307
Growth (20)	0.4410	0.4198	0.4465	0.4589	0.4575
Ligase activity (1)	0.4169	0.4153	0.4294	0.3955	0.4405
MIC (1)	0.6607	0.6374	0.6602	0.6703	0.6672
Peptide binding (2)	0.5477	0.4932	0.6146	0.5719	0.5857
Viral replication (5)	0.3920	0.4075	0.3624	0.3857	0.3951
Yeast growth (3)	0.3951	0.3930	0.3948	0.4049	0.4030

Table 3: **Clustering efficiency in recapitulating phyla memberships of proteins in three Pfam families.** For each family, performance is reported as average 1 nearest-neighbor generalization error for recovering phyla affiliations. Lower is better, lowest value for each Pfam family is shown in blue.

Pfam family	Untrained	ESM1b	Pept. MLM	Pept. NPP	Pept. C-NPP	Pept. BMLM
Beta-lactamase	0.250 40	0.093 55	0.093 46	0.097 50	0.123 81	0.093 33
SH3	0.336 47	0.168 54	0.169 41	0.206 33	0.236 35	0.173 71
WW	0.376 25	0.226 20	0.235 74	0.284 24	0.295 71	0.223 42

only fine-tuned on short peptides, thus suggesting these models may preserve information learned from full proteins when fine-tuned on short peptides.

Functional analysis with zero-shot mutation effect prediction.— Evaluating the effects mutant protein sequence over its Wild-Type (WT) counterpart is a fundamental problem for understanding and designing proteins. Previous study [18] found that pretrained protein language model can be used to score mutational effects without any training (zero-shot transfer). We took 36 previously published deep mutational scans (DMS) [23], each of which experimentally quantifies a set mutant sequences over the WT protein. The goal of this task is to regress the mutant effects over its WT with zero-shot transfer. The WT marginal scoring scheme previously used in the ESM1v publication [18] was used for this task. Results, calculated as Spearman's ρ for the effect measured in each study's set of mutated sequences versus the predicted likelihood ratios over the WT counterparts, are shown in Table 2 (Table A4 shows the same results when aggregated to the organism level). Distributions of sequence lengths are further described for each of these modality groups in Appendix Tables A2, A3.

Evolutionary information enrichment in sequence embeddings.— Previous studies [24, 11, 9] have illustrated that the representation space learned by protein models reflect evolutionary information. As such, without further supervision, these models are able to cluster evolutionarily related protein sequences closer than less related ones. We probe the impact of the continued training of protein models on chopped sequences on this ability. We sample common Pfam protein families [4], including beta-lactamase (PF00144), SH3 domain (PF00018), and WW domain (PF00397), and apply t-SNE[29], for each family (Fig 5), on the extracted protein representations. We follow van der Maaten et al.[30] and calculate the generalization error from 1-nearest neighbor classifiers that are trained on the low-dimensional data representation. Our results (Table 3) show that the BMLM peptide model outperforms ESM1b on beta-lactamase and WW families, while achieving competitive performance on SH3 family. Notably, the three representative protein families cover a range of sequence lengths: beta-lactamase (324 AAs), SH3 (47 AAs), and WW (30 AAs).

Table 4: **Sequence type generalization analysis.** Comparison of models trained on natural vs. artificial ("chopped") short sequences. Models were trained with MLM objective. (FT) stands for models finetuned from ESM1b. We compare ECE (Lower is better) across test sets of each data type.

Evaluation Set	ESM1b	UR50-S Chopped (FT)	Peptide Atlas (FT)	UR50-S Shorts (FT)	UR50-S Chopped	Peptide Atlas	UR50-S Shorts
UR50-S Chopped	7.0837	**5.9511**	6.3835	6.3526	6.2687	7.6378	6.5923
Peptide Atlas	7.4915	6.1710	**5.3595**	6.5857	6.3551	5.3868	6.7195
UR50-S Shorts	6.6286	6.3135	7.0923	**5.9930**	6.5024	8.3490	6.8436

Sequence type analysis - Chopped vs. Natural we compare our chopped-sequence models with models trained on natural peptides, derived either from short UR50-S sequences ($\sim 1M$ peptides) or from Peptide Atlas [8] ($\sim 3.5M$ sequences), which is the largest collection of mass-spectrometry identified peptides. In comparison to the chopped-sequence training, in which all $\sim 30M$ UR50 sequences are considered, and the random on-the-fly chopping adds sequence variability between different epochs, we get a much larger data scale of almost $\sim 30M * \#epochs$ (we use 25 epochs in our experimental setup, and assume here that most sequences are long enough to produce different unique chopped sequences at each epoch). We compare performance of each model across the various test sets and observe, at Table 4, that while each model performs best on it's dedicated data type, the models trained on chopped sequences generalize better than others to unseen sequence types.

Next, we seek to quantify the intrinsic differences of between chopped and natural peptide sequences and how such differences are perceived by different models (trained on different sequence types). We performed a balanced sampling of peptides from natural and chopped sequence test sets, and obtained their embeddings from models. To quantify the differences of those embeddings from the chopped and natural sequences, we examine the t-SNE projections of the embeddings and calculate the nearest-neighbor (NN) generalization error. Lower NN-error indicates the two populations are more distinctive whereas higher NN-error indicates the two populations are more intermixed in the embedding space. We use one-hot encoded sequence representations as the baseline trend for the intrinsic differences. ESM1b without finetuning was used as the baseline model. As shown in Fig 2, both ESM1b and its chopped-sequence finetuned variant learn to bring the embeddings from two populations closer than the baseline (one-hot encoding), whereas language models finetuned on naturals alone, learnt distinct embeddings for either set. All models showed a trend of generating more distinct embeddings for chopped and natural sequences as their length increases.

4 Conclusion

Proteins perform many essential functions in biological systems and can be successfully developed as bio-therapeutics. Here we presented an approach for self-supervised training on shorter chopped protein sequences. Our empirical results showed that these learned representations are beneficial for downstream peptide-related tasks, while slightly sacrificing the performances on some of the longer protein-focused tasks. Specifically, we observed peptide models outperformed ESM1b on context prediction accuracy in all length ranges, with improvement ranging from 24.5% on the longest to 41.7% for the shortest sequence lengths. We also found albeit having some differences with natural peptides, chopped protein sequences leads to models with improved generalization performance on out-of-distribution data. This indicates chopped proteins is a powerful data augmentation method for training protein language models.

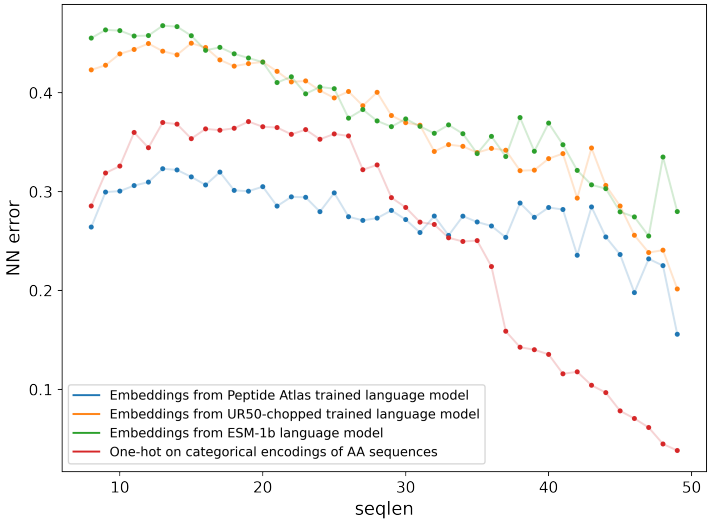


Figure 2: **Protein sequence representations encode evolutionary information.** The differences in peptide representations quantified by 1-NN generalization error on test sequences is plotted at peptide lengths ranging from 8 to 50. Each line represents a distinct model. Lower error indicates the model embeddings from the two sequence populations (chopped and naturals) are more distinct.

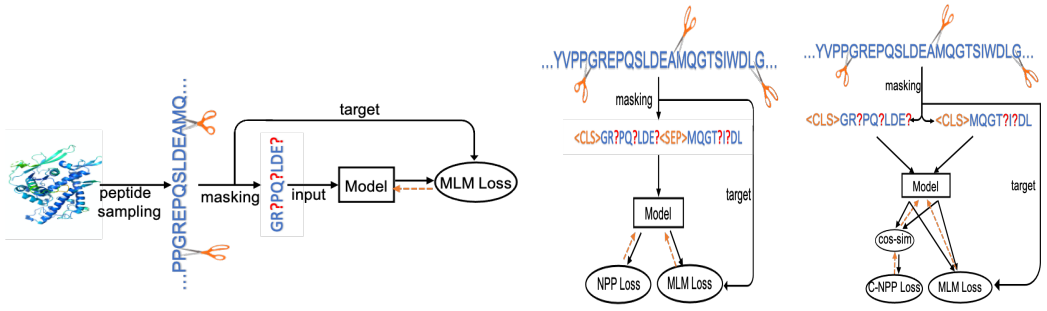
Appendix

.1 Alternative training objectives

Next Peptide Prediction (NPP).— For the NPP objective, firstly we reinterpret the 'next sentence prediction' task from NLP, as proposed in the BERT paper [10], as 'next peptide prediction'. As in Ref. [10], we concatenate two peptide sequences from the chopped protein dataset, with a special <SEP> token in between, as the input sequence to the model. Given the first peptide before the <SEP> token, the second peptide is either the correct next peptide in the parent protein sequence or a randomly sampled one. The objective is a binary cross-entropy loss applied over a linear projection from the <CLS> token output representation. Despite the 'next-sentence-prediction' objective being criticized in recent NLP publications [32, 3], no analogous research has been done for amino-acid sequences. We believe that due to the nature of the task, vocabulary and underlying chemical and structural considerations, NPP might be more challenging and hence lead to improved learned representations.

Contrastive NPP (C-NPP).— We introduce a contrastive extension of the NPP objective, where the model needs to learn to detect the correct next peptide sequence from the parent protein out of a large pool of candidates. Specifically, a batch of N peptide samples will be composed of $N/2$ 'first peptides' and their corresponding $N/2$ 'next peptides'. The <CLS> token output representation is projected to a lower dimension, using two separate sets of learned projection weights. Cosine similarity is computed to provide an estimated compatibility score between each possible pair of 'first' vs 'next' peptide sequences. The objective is composed of cross-entropy loss, following the contrastive objective defined in SimCLR [6], defined as follows for each positive pair of first and next peptides: $\mathcal{L}_{\mathrm{C-NPP}}(i) = -\log \frac{\exp(S(f_i, n_i)/\tau)}{\sum_{k=0}^{N/2} \exp(S(f_i, n_k)/\tau)}$, where f_i is the projected representation of the ith first peptide, n_i is the projected representation of the corresponding ith next peptide, S stands for cosine similarity between the two elements, and $\tau \in \mathbb{R}^+$ corresponds to a real temperature scaling factor. For efficient evaluation, the potential candidate pool is restricted to the true next peptides from different samples within the same batch. In addition, we extended this objective with an equivalent loss term for predicting the previous (first) peptide, from the pool of first peptide candidates, for each next peptide. MLM loss is applied in addition to these two loss terms. A similar contrastive objective has been shown to improve performance of protein language-style models trained on full protein sequences [26].

BLOSUM MLM (BMLM).— For the BLOSUM MLM objective, as in MLM, each input sequence is modified by replacing a fraction of the amino acid tokens with a special mask token. The network is trained to predict the missing tokens from the modified sequence. The idea is to "smooth" the



(a) (b) (c)

Figure 3: **Self-supervised training procedures on chopped proteins.** (a) Masked language modeling on chopped proteins. (b) NPP - a pair of chopped sequences is concatenated and processed jointly. A binary cross-entropy loss, over projected <CLS> token representation, optimizes classification of true vs fake next peptide. (c) C-NPP - first and next chopped peptides are processed separately. Compatibility between all first and next candidate peptides in the batch is computed via cosine similarity over projected <CLS> token representations. A contrastive cross-entropy loss is optimized to ensure correct pairs will receive higher compatibility scores than non-matching pairs.

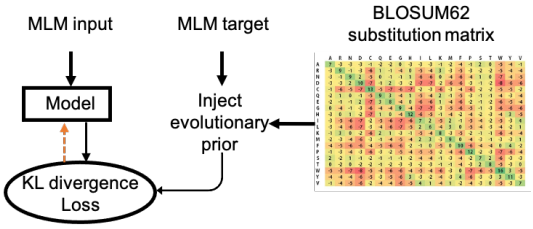


Figure 4: **Training procedure incorporating evolutionary information utilizing the BLOSUM62 substitiution matrix**. In BLOSUM MLM (BMLM), we transform the one-hot MLM target into a probability distribution over tokens, based on the BLOSUM62 substitution scores of the target amino-acid, and minimize the KL-divergence loss.

one-hot removed token targets, based on the BLOSUM62 substitution matrix, and assign pseudo-probabilities to each amino-acid token based on the probability that it could substitute the removed token. The final objective is a KL-divergence loss between target and prediction distribution as follows: $\mathcal{L}_{BMLM} = \mathbb{E}_M[\sum_{i \in M} D_{KL}(p(\hat{x}_i|x_{/M})||blosum(x_i))]$, where M is the set of masked elements, D_{KL} denotes KL-divergence and $blosum(x_i)$ is the "pseudo-probability" function for the amino acid x_i. It returns a 1-d vector with 20 elements corresponding to a row in the BLOSUM matrix. The BLOSUM rows are transformed into pseudo-probabilities with Softmax.

.2 Additional results

Supervised evaluation on TAPE Stability task.— We evaluated the performance of protein and peptide language models on the TAPE stability prediction task [21]. This task was selected because it contains exclusively peptides of length 45 amino acids. For this task, models were evaluated over 3 random seeds. Training was done with early stopping until validation loss plateaued over 10 epochs. A warm-up schedule was applied for the learning rate with 5000 warm-up steps.

Comparison of fine-tuning and linear probing on the TAPE stability task indicated improved generalization with linear probing. Results reported in this paper are from linear probing. Two peptide transformers, namely those trained with MLM and BMLM objectives, significantly outperformed the ESM1b baseline protein language model on the stability task. ESM1b is the current state-of-the art on this task, achieving a Spearman's ρ of 0.7365. When averaging over 3 random seeds, the Peptide BMLM and MLM models yield $\rho = 0.7650 \pm 0.0067$ and $\rho = 0.7673 \pm 0.0154$, respectively. The other peptide objectives did not exceed the performance of ESM1b, giving $\rho = 0.6707 \pm 0.0185$, and $\rho = 0.6460 \pm 0.0141$, using C-NPP, and NPP, respectively. To our knowledge, BMLM and MLM peptide transformers achieve a new state-of-the-art performance on this task.

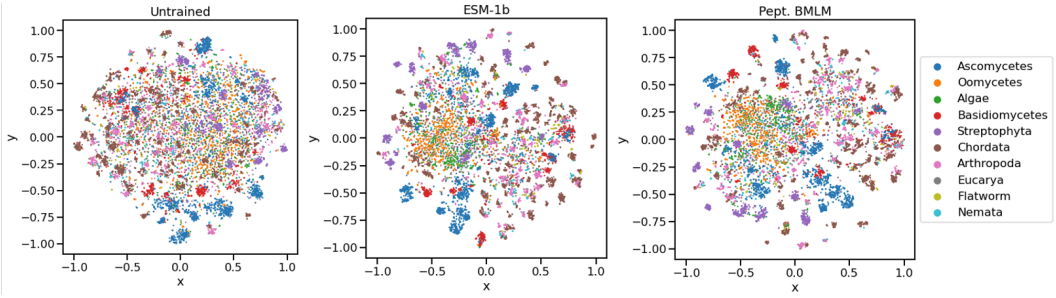


Figure 5: **Protein sequence representations encode evolutionary information.** Each point represents a protein/peptide sequence from WW domain (PF00397), and each sequence is colored by the phylum it belongs to (dimensionality is reduced by t-SNE). The panels correspond to sequence representations from untrained (left), ESM1b (middle), and peptide transformer with BMLM objective.

Table A1: **Sequence length statistics for our UR50-S dataset splits.** Number of protein sequences shown in brackets in the first column.

Split	Average length	Minimum length	Maximum length
Train (24.4M)	311.7	16	36 991
Validation (2.7M)	311.6	16	34 984
Test (3M)	311.6	16	34 674

Table A2: **Sequence length statistics for DMS datasets, based on experimental measurement modality.** Number of studies for each modality shown in brackets in the first column.

Measurement	Average length	Minimum length	Maximum length
E1 reactivity (1)	76	76	76
Enzyme function (3)	345.3	189	501
Growth (20)	220.5	72	439
Ligase activity (1)	104	104	104
MIC (1)	263	263	263
Peptide binding (2)	68.5	36	101
Viral replication (5)	456.4	114	686
Yeast growth (3)	164.3	101	243

Table A3: **Sequence length statistics for DMS datasets, based on model system for experiments.** Number of studies for each modality shown in brackets in the first column.

Model system	Average length	Minimum length	Maximum length
ECOLX (4)	215.3	72	263
env (2)	686	686	686
HUMAN (8)	178.1	36	360
Other (12)	331.9	114	565
RODENT* (2)	102.5	101	104
YEAST (8)	119.4	75	240

Table A4: **Zero-shot mutational effect analysis (experimental model system).** Average Spearman's ρ is shown, aggregated on the model system used in the DMS experiments. Number of studies for each modality shown in brackets in the first column. 'Winning' model for each modality is highlighted.

Model system	ESM1b	Pept. BMLM	Pept. C-NPP	Pept. MLM	Pept. NPP
HUMAN (8)	0.3948	0.3778	0.3954	0.3933	0.3907
ECOLX (4)	0.4358	0.4121	0.4245	0.4303	0.4287
env (2)	0.4655	0.4336	0.3681	0.4668	0.4441
RODENT* (2)	0.4174	0.3896	0.4789	0.4512	0.4715
Other (12)	0.4536	0.4551	0.4536	0.4520	0.4596
YEAST (8)	0.4374	0.4246	0.4765	0.5072	0.5027

References

[1] Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 16(12):1315–1322, 2019.

[2] Vasso Apostolopoulos, Joanna Bojarska, Tsun-Thai Chai, Sherif Elnagdy, Krzysztof Kaczmarek, John Matsoukas, Roger New, Keykavous Parang, Octavio Paredes Lopez, Hamideh Parhiz, et al. A global review on short peptides: frontiers and perspectives. *Molecules*, 26(2):430, 2021.

[3] Stéphane Aroca-Ouellette and Frank Rudzicz. On losses for modern language models. *arXiv preprint arXiv:2010.01694*, 2020.

[4] Alex Bateman, Lachlan Coin, Richard Durbin, Robert D Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, Mhairi Marshall, Simon Moxon, Erik LL Sonnhammer, et al. The pfam protein families database. *Nucleic acids research*, 32(suppl_1):D138–D141, 2004.

[5] Tristan Bepler and Bonnie Berger. Learning protein sequence embeddings using information from structure. *arXiv preprint arXiv:1902.08661*, 2019.

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[7] Hao Cheng, Bing Rao, Lei Liu, Lizhen Cui, Guobao Xiao, Ran Su, and Leyi Wei. Pepformer: End-to-end transformer-based siamese network to predict and enhance peptide detectability based on sequence only. *Analytical Chemistry*, 93(16):6481–6490, 2021.

[8] Frank Desiere, Eric W Deutsch, Nichole L King, Alexey I Nesvizhskii, Parag Mallick, Jimmy Eng, Sharon Chen, James Eddes, Sandra N Loevenich, and Ruedi Aebersold. The peptideatlas project. *Nucleic acids research*, 34(suppl_1):D655–D658, 2006.

[9] Nicki Skafte Detlefsen, Søren Hauberg, and Wouter Boomsma. Learning meaningful representations of protein sequences. *Nature communications*, 13(1):1–12, 2022.

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

[11] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv:2007.06225*, 2020.

[12] Noelia Ferruz, Steffen Schmidt, and Birte Höcker. A deep unsupervised language model for protein design. *bioRxiv*, 2022.

[13] Keld Fosgerau and Torsten Hoffmann. Peptide therapeutics: current status and future directions. *Drug Discovery Today*, 20(1):122–128, 2015.

[14] Liang He, Shizhuo Zhang, Lijun Wu, Huanhuan Xia, Fusong Ju, He Zhang, Siyuan Liu, Yingce Xia, Jianwei Zhu, Pan Deng, et al. Pre-training co-evolutionary protein representation via a pairwise masked language model. *arXiv preprint arXiv:2110.15527*, 2021.

[15] Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, Dmitrii Nechaev, Florian Matthes, and Burkhard Rost. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC bioinformatics*, 20(1):1–17, 2019.

[16] Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

[17] Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R Eguchi, Po-Ssu Huang, and Richard Socher. Progen: Language modeling for protein generation. *arXiv preprint arXiv:2004.03497*, 2020.

[18] Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Language models enable zero-shot prediction of the effects of mutations on protein function. *Advances in Neural Information Processing Systems*, 34, 2021.

[19] Bruce H Morimoto. Therapeutic peptides for cns indications: Progress and challenges. *Bioorganic & Medicinal Chemistry*, 26(10):2859–2862, 2018.

[20] Markus Muttenthaler, Glenn F King, David J Adams, and Paul F Alewood. Trends in peptide drug discovery. *Nature Reviews Drug Discovery*, 20(4):309–325, 2021.

[21] Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

[22] Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. In *International Conference on Learning Representations*, 2020.

[23] Adam J Riesselman, John B Ingraham, and Debora S Marks. Deep generative models of genetic variation capture the effects of mutations. *Nature methods*, 15(10):816–822, 2018.

[24] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15), 2021.

[25] Guillermo Serrano, Elizabeth Guruceaga, and Victor Segura. Deepmspeptide: peptide detectability prediction using deep learning. *Bioinformatics*, 36(4):1279–1280, 2020.

[26] Hongyu Shen, Layne C Price, Taha Bahadori, and Franziska Seeger. Improving generalizability of protein sequence models with data augmentations. *bioRxiv*, 2021.

[27] Baris E Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H Wu. Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23(10):1282–1288, 2007.

[28] Keiji Tanaka. The proteasome: overview of structure and functions. *Proceedings of the Japan Academy, Series B*, 85(1):12–36, 2009.

[29] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[30] Laurens Van Der Maaten, Eric Postma, Jaap Van den Herik, et al. Dimensionality reduction: a comparative. *J Mach Learn Res*, 10(66-71):13, 2009.

[31] Lei Wang, Nanxi Wang, Wenping Zhang, Xurui Cheng, Zhibin Yan, Gang Shao, Xi Wang, Rui Wang, and Caiyun Fu. Therapeutic peptides: current applications and future directions. *Signal Transduction and Targeted Therapy*, 7(1):1–27, 2022.

[32] Weidi Xu, Xingyi Cheng, Kunlong Chen, and Taifeng Wang. Symmetric regularization based bert for pair-wise semantic reasoning. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1901–1904, 2020.