Aggregating Residue-Level Protein Language Model Embeddings with Optimal Transport

Navid NaderiAlizadeh[1,*] and Rohit Singh[1,2]

[1]Department of Biostatistics and Bioinformatics, Duke University, Durham, 27705, NC, USA
[2]Department of Cell Biology, Duke University, Durham, 27705, NC, USA

*Corresponding author. Email: navid.naderi@duke.edu

Abstract

Motivation: Protein language models (PLMs) have emerged as powerful approaches for mapping protein sequences into informative embeddings suitable for a range of applications. PLMs, as well as many other protein representation schemes, generate per-token (i.e., per-residue) representations, leading to variable-sized outputs based on protein length. This variability presents a challenge for protein-level prediction tasks, which require uniform-sized embeddings for consistent analysis across different proteins. Prior work has typically resorted to average pooling to summarize token-level PLM outputs. It is, however, unclear if such an aggregation operation effectively prioritizes the relevant information across token-level representations.
Results: Addressing this, we introduce a novel method utilizing sliced-Wasserstein embeddings to convert variable-length PLM outputs into fixed-length protein-level representations. Inspired by the success of optimal transport techniques in representation learning, we first conceptualize per-token PLM outputs as samples from a probabilistic distribution. We then employ sliced-Wasserstein distances to map these samples against a learnable reference set, creating a Euclidean embedding in the output space. The resulting embedding is agnostic to the length of the input and represents the entire protein. Across a range of state-of-the-art pre-trained ESM-2 PLMs, with varying model sizes, we show the superiority of our method over average pooling for protein-drug and protein-protein interaction. Our aggregation scheme is especially effective when model size is constrained, enabling smaller-scale PLMs to match or exceed the performance of average-pooled larger-scale PLMs. Since using smaller models reduces computational resource requirements, our approach not only promises more accurate inference but can also help democratize access to foundation models.
Availability and implementation: The implementation code can be found at `https://github.com/navid-naderi/PLM_SWE`.

Key words: Protein Language Models, Representation Learning, Optimal Transport, Sliced-Wasserstein Embedding, Set Learning, Protein-Drug Interaction, Protein-Protein Interaction

1. Introduction

Understanding the sequence–structure–function relationship for proteins is one of the grand challenges of biology. Among the problems defined around these relationships, a prominent class of problems comprises tasks where some structural or functional property of a protein is to be predicted from its sequence. The possible set of prediction tasks in this class is very diverse, including both classification (e.g., "is the protein a kinase?") and regression (e.g., the melting temperature of the protein), as well as tasks involving auxiliary inputs (e.g., small molecule SMILES representations, for predicting drug-target interactions). Any such problem can be formulated as a prediction task over a set of proteins, where the input consists of a protein sequence and the output (a label or number) is at the level of the *entire protein*, rather than its constituent amino acids. Any model that addresses such a problem formulation will necessarily have one or more steps where information across the constituent amino acids of the protein is summarized into a protein-level estimate.

The problem of appropriately aggregating amino acid-level information has become particularly pressing with the advent of protein language models (PLMs). PLMs, which are trained on massive corpora of protein sequence data using self-supervised learning, are able to build internal representations that capture evolutionary constraints on protein sequences. Since these representations comprehensively capture constraints on protein function and structure, PLMs have proven powerful in a wide range of tasks from structure prediction to interaction prediction and protein design. For many protein sequence-based property-prediction tasks, PLM-based approaches are now the state of the art (Littmann et al., 2021; Kaminski et al., 2023; Singh et al., 2023). Given a sequence of length n, a pre-trained PLM produces an embedding of dimensionality $\mathbb{R}^{n \times d}$, where d is the per amino acid (i.e., per-token) embedding dimensionality. The token-specific embedding captures not only biochemical information about the token (i.e., the amino acid) but also the local and global properties of the protein.

To aggregate these per-token embeddings into a protein-level representation, the most common approach has been to simply "average pool:" take the mean of each feature dimension along the length of the protein to produce an embedding in $\mathbb{R}^d$. While other pooling approaches, such as "max pooling" (taking the maximum of the set) or "softmax pooling" (i.e., average pooling after exponentiation, then logarithmized) are also sometimes used, average pooling is typically preferred for convenience, speed, and simplicity. However, it weighs each amino acid's representation equally. This is unrealistic—often, there are specific residues in the protein that are particularly important (e.g., the residues at an active site). Even when the per-token PLM representation did contain information distinguishing such residues from others, these distinctions would be lost during average pooling. We note that such considerations are not limited to PLM-based embeddings: in many task-specific neural network architectures, e.g., PIPR for PPI prediction (Chen et al., 2019), average pooling is used to summarize variable-length intermediate representations.

In this work, we present a novel approach to aggregate variable-length protein representations. Like average pooling, max pooling, and related approaches, our method is permutation invariant: it considers the per-token embeddings as a *set*, rather than a *sequence*, under the assumption that the PLM backbone fully embeds the sequential properties of the protein data in its output residue-level representations. Let $\mathbf{p}_j$ be a protein of length n_j, so that its PLM embedding can be represented as $\mathbf{x}_{\cdot j} = \{\mathbf{x}_{ij}\}_{i=1}^{n_j}$, with each $\mathbf{x}_{ij} \in \mathbb{R}^d$. Our work seeks to learn a set of m reference embeddings $\mathbf{x}_{\cdot 0} = \{\mathbf{x}_{i0}\}_{i=1}^{m}$, with $\mathbf{x}_{i0} \in \mathbb{R}^d$, that can characterize any variable-length representation. Conceptually, this is analogous to learning a task-specific basis representation in $\mathbb{R}^d$. The intuition underpinning our work is to think of $\mathbf{x}_{\cdot j}$ and $\mathbf{x}_{\cdot 0}$ as empirical probability distributions in $\mathbb{R}^d$ and to formulate $\mathbf{x}_{\cdot j}$'s distance from the reference set $\mathbf{x}_{\cdot 0}$ as an optimal transport calculation.

Our work builds upon optimal transport (OT) based approaches in computer vision to characterize sets of observations. We borrow from previous advances to deploy the OT intuition effectively and in a scalable fashion. In particular, we use "slices" in the embedding space, learnable directions in $\mathbb{R}^d$ onto which the input and reference token-sets $\mathbf{x}_{\cdot j}$ and $\mathbf{x}_{\cdot 0}$ are projected. These projections correspond to 1-D probability distributions. On such distributions, OT distances can be computed efficiently, and an ensemble of L slices serves to efficiently characterize the separation between input and reference sets.

The key conceptual advance of our work is unlocking task-specific learnability as a key component of PLM-based machine learning models. Broadly, these models can be thought of as pipelines of three segments: an initial *sequence* segment, a *summarization* segment, and a final *prediction* segment. For example, the sequence segment may consist of transformer layers, while the prediction segment may consist of a feed-forward network. However, if average pooling is used for summarization, no task-specific learning can happen there and will need to happen only in the sequence or the prediction segment. With our innovation, the summarization segment also becomes learnable, offering greater flexibility in matching the architecture of the neural network to the biological intuitions underlying the task.

Our work also has the potential to introduce interpretability into systems that would otherwise be opaque. The set of m reference embeddings learned by the system can serve as useful archetypes for the task at hand. For instance, in the computer vision context, it was shown that simply being able to associate the variable-length representation with one of the references can be informative about the typicality of the underlying object.

We apply our sliced-Wasserstein embedding (SWE) approach to three protein property prediction tasks: binary drug-target interaction prediction, out-of-domain drug-target affinity prediction, and protein-protein interaction prediction. SWE broadly outperforms average pooling, especially on small and moderate-sized ESM-2 models. Notably, the onerous GPU memory requirements of the largest ESM-2 models suggest that the performance boost of SWE could be critical to democratizing access to PLMs for researchers with limited GPU resources.

2. Related Work

2.1. Protein Language Models

Large language models (LLMs), such as GPT-4 (Achiam et al., 2023) and Llama 2 (Touvron et al., 2023), have become the predominant tools for modeling sequential natural language data. The success of LLMs, which mostly rely on attention-based transformer architectures (Vaswani et al., 2017), has inspired researchers working with biological data to use similar ideas for analyzing protein sequences. In particular, the availability of massive protein sequence datasets has given rise to large-scale protein language models (PLMs), such as ESM (Rives et al., 2019; Lin et al., 2022), ProtBert (Elnaggar et al., 2021), ProtGPT2 (Ferruz et al., 2022), MSA Transformer (Rao et al., 2021), SaProt (Su et al., 2024), and xTrimoPGLM (Chen et al., 2024), to name a few. These models are mostly trained using unlabeled protein sequence data in a self-supervised way, where the goal is to train the model to predict a token that has been replaced with a special *mask* token using its surrounding context, i.e., other amino acids in the sequence. Such masking-based unsupervised training leads to token-level representations that have been shown to provide state-of-the-art performance in a wide array of downstream tasks, such as protein folding (Villegas-Morcillo et al., 2022), variant effect prediction (Brandes et al., 2023), peptide generation (Chen et al., 2023), antibody design (Wu and Li, 2023), and prokaryotic gene prediction (Tu et al., 2023).

The transformer architectures in LLMs, in general, and PLMs, in particular, produce residue-level representations that need to be summarized and aggregated for protein-level downstream tasks since different amino acid sequences have varying lengths. The question of aggregating a set of elements into a fixed-length representation is the key behind the research on *set representation learning*, which we discuss next.

2.2. Set Representation Learning

The goal of set representation learning is to map an unordered collection of elements into an embedding that is invariant to the permutation of the set elements and whose size is independent of the input set size (Ravanbakhsh et al., 2016; Wagstaff et al., 2019). Deep Sets (Zaheer et al., 2017) and Janossy Pooling (Murphy et al., 2019) are two seminal studies in this area, where the set embedding is modeled as a function of the sum or average of permutation-sensitive functions applied to all elements or all permutations of the input set. Follow-up work has leveraged ideas based on transformers (Lee et al., 2019), optimal transport (Kolouri et al., 2021; Naderializadeh et al., 2021b; Mialon et al., 2021; Naderializadeh et al., 2021a; Lu et al., 2024), and featurewise sorting (Zhang et al., 2020) for deep learning on sets, and demonstrated their efficacy in a variety of

learning settings, including point cloud classification (Qi et al., 2017a,b), graph representation (Maron et al., 2019), and multi-agent reinforcement learning (Sunehag et al., 2017; Naderializadeh et al., 2020; Kortvelesy and Prorok, 2022).

In the context of PLMs, prior work has, for the most part, used *average pooling* to summarize the token-level embeddings into universal protein-level embeddings (Bepler and Berger, 2021; Unsal et al., 2022; Singh et al., 2023; Sledzieski et al., 2023), and past research on other pooling methods is scarce (Sledzieski et al., 2021; Stärk et al., 2021; Iliadis et al., 2023). Averaging is a simple, unparameterized, permutation-invariant, and size-invariant function, which warrants its selection as the most natural and intuitive choice for aggregating PLM outputs. Nevertheless, it is unknown whether other, more sophisticated aggregation mechanisms could unlock additional performance gains compared to average pooling when used in conjunction with state-of-the-art PLMs. In this paper, we give an affirmative answer to this question by proposing a parameterized aggregation operation based on ideas from optimal transport to summarize residue-level embeddings generated by pre-trained PLMs into fixed-length protein-level embeddings.

3. Methods and Materials

3.1. Problem Formulation

Consider a protein's primary amino acid sequence of length $n \in \mathbb{N}$, denoted by $\mathbf{p} = (p_1, \dots, p_n) \in \mathcal{P}^n$, where $\mathbb{N}$ denotes the set of natural numbers, i.e., positive integers, $\mathcal{P}$ represents the residue alphabet. We gather the set of all possible protein sequences of arbitrary lengths into a set

$$\mathcal{X} = \bigcup_{n \in \mathbb{N}} \mathcal{P}^n. \tag{1}$$

The goal of protein-level representation learning is to find a function $\psi(\cdot; \theta_\psi) : \mathcal{X} \to \mathbb{R}^d$, parameterized by a finite-dimensional set of parameters $\theta_\psi \in \Theta_\psi$ (see Figure 1-a). Observe that the dimensionality of the embedding space, i.e., d, and the size of the model parameter space, i.e., $|\Theta|$, are *independent* of the length of the input protein sequence. In other words, the function $\psi(\cdot; \theta_\psi)$ should be able to map any given protein sequence of arbitrary length to a fixed-size representation in $\mathbb{R}^d$.

Assuming the availability of a (labeled) protein sequence dataset $\{(\mathbf{p}_j, y_j)\}_{j=1}^N$, where $\mathbf{p}_j \in \mathcal{X}, y_j \in \mathcal{Y}, \forall j \in \{1, \dots, N\}$, with $\mathcal{Y}$ denoting the set of all possible labels, the parameters of the representation function ψ are typically derived via an empirical risk minimization (ERM) problem,

$$\min_{\theta_\psi \in \Theta_\psi} \frac{1}{N} \sum_{j=1}^{N} \ell(\psi(\mathbf{p}_j; \theta_\psi), y_j), \tag{2}$$

where $\ell : \mathbb{R}^d \times \mathcal{Y} \to \mathbb{R}$ denotes a loss function. While resembling a supervised learning setting, note that the formulation in (2) also includes an unsupervised learning scenario, where the labels y_j are trivial/non-existent and the loss function ℓ does not depend on y.

A commonly used initial building block for the parameterization of the representation function ψ is a protein language model (PLM). We denote a PLM by a parameterized function $\phi(\cdot, \theta_\phi) : \mathcal{P}^n \to (\mathbb{R}^d)^n, \forall n \in \mathbb{N}$. This implies that a given PLM maps each amino acid in an input protein sequence into an individual embedding in $\mathbb{R}^d$. The parameters of the PLM, i.e., θ_ϕ, are usually trained by a masking-based objective, where some amino acid identities in the input sequences are masked by random tokens, and the model is trained to predict the correct amino acids using the corresponding output representations.

In this paper, our goal is on the aggregation, or pooling, function that bridges the gap between PLM-generated token-level outputs and the final protein-level representation. More formally, we are interested in an informative aggregation function $\pi(\cdot; \theta_\pi) : (\mathbb{R}^d)^n \to \mathbb{R}, \forall n \in \mathbb{N}$, parameterized by $\theta_\pi \in \Theta_\pi$. Taken together, the composition of the PLM ϕ and the aggregation function π constitutes the end-to-end protein-level representation learning pipeline (see Figure 1-b); for any given protein sequence $\mathbf{p} \in \mathcal{X}$, its representation can be derived as

$$\psi(\mathbf{p}; \theta_\psi) = \psi(\mathbf{p}; \theta_\phi, \theta_\pi) = \pi\Big(\phi(\mathbf{p}; \theta_\phi); \theta_\pi\Big) \in \mathbb{R}^d. \tag{3}$$

We assume that the PLM is pre-trained and its parameters, θ_ϕ, are frozen. Therefore, we are primarily interested in training the aggregation function $\pi(\cdot; \theta_\pi)$ for downstream prediction tasks.

3.2. Proposed Method

A simple and prominent example of an aggregation function is the *averaging* operation, i.e., $\pi_{\mathsf{avg}}(\mathbf{x}_1, \dots, \mathbf{x}_n) = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i$, which is indeed unparameterized (i.e., $\Theta_\psi = \emptyset$; see Figure 1-c). While used extensively in prior work in the protein representation learning literature (see, e.g., (Sledzieski et al., 2023; Singh et al., 2023)), average pooling may not capture the entirety of the information that is present in the per-token embeddings. This calls for methods that are able to capture such information in the aggregated representations.

In this paper, we propose to use sliced-Wasserstein distances from optimal transport (Deshpande et al., 2019; Kolouri et al., 2019), and in particular, *sliced-Wasserstein embedding* (SWE) (Naderializadeh et al., 2021a; Lu et al., 2024) to aggregate the token-level representations into a universal protein-level presentation. The main idea behind SWE is to treat the token-level embeddings as samples drawn from an underlying probability distribution, and then find the optimal transportation plan that maps that distribution to a trainable reference distribution (see Figure 1-d).

More formally, let $\{\mathbf{x}_{ij}\}_{i=1}^{n_j}$ denote the token-level embeddings of the j^{th} protein sequence in the dataset, $\mathbf{p}_j, j \in \{1, \dots, N\}$, of length n_j, which are produced by the pre-trained PLM; i.e., $(\mathbf{x}_{1j}, \dots, \mathbf{x}_{n_j j}) = \phi(\mathbf{p}_j; \theta_\phi)$. We assume that the embeddings $\{\mathbf{x}_{ij}\}_{i=1}^{n_j}$ are samples drawn from an underlying distribution $\mathfrak{D}_j$ supported on $\mathbb{R}^d$. We also consider a trainable reference set $\{\mathbf{x}_{i0}\}_{i=1}^m$ of m points in $\mathbb{R}^d$ that are drawn from a reference distribution $\mathfrak{D}_0$. Since there is no closed-form solution for calculating the optimal Monge coupling (Villani et al., 2009) between high-dimensional distributions in $\mathbb{R}^d$, we resort to *slicing* operations in order to map these distributions to several one-dimensional distributions, for which the Monge coupling has a closed-form solution. In particular, we consider trainable linear maps $\{\omega_l\}_{l=1}^L$, with $\omega_l \in \mathbb{R}^d, \forall l \in \{1, \dots, L\}$, through which each set of token-level embeddings $\{\mathbf{x}_{ij}\}_{i=1}^{n_j}, j \in \{0, \dots, N\}$ (including the reference embeddings with $n_0 = m$) is mapped to L *slices*, i.e., sets of one-dimensional points $\{u_{ij}^l\}_{i=1}^{n_j}, l \in \{1, \dots, L\}$, where

$$u_{ij}^l = \omega_l^T \mathbf{x}_{ij}, \ \forall i \in \{1, \dots, n_j\}, \forall j \in \{0, \dots, N\}, \forall l \in \{1, \dots, L\}.$$

Consider the l^{th} slice of a given set of token-level embeddings, i.e., $\{u_{ij}^l\}_{i=1}^{n_j}$ and the corresponding slice of the reference set, i.e., $\{u_{i0}^l\}_{i=1}^m$. These sets correspond to empirical 1-D distributions $\hat{\mathfrak{D}}_j^l(u) = \frac{1}{n_j} \sum_{i=1}^{n_j} \delta(u - u_{ij}^l)$ and $\hat{\mathfrak{D}}_0^l(u) = \frac{1}{m} \sum_{i=1}^m \delta(u - u_{i0}^l)$, respectively. The Monge coupling between $\hat{\mathfrak{D}}_j^l(u)$ and $\hat{\mathfrak{D}}_0^l(u)$ has

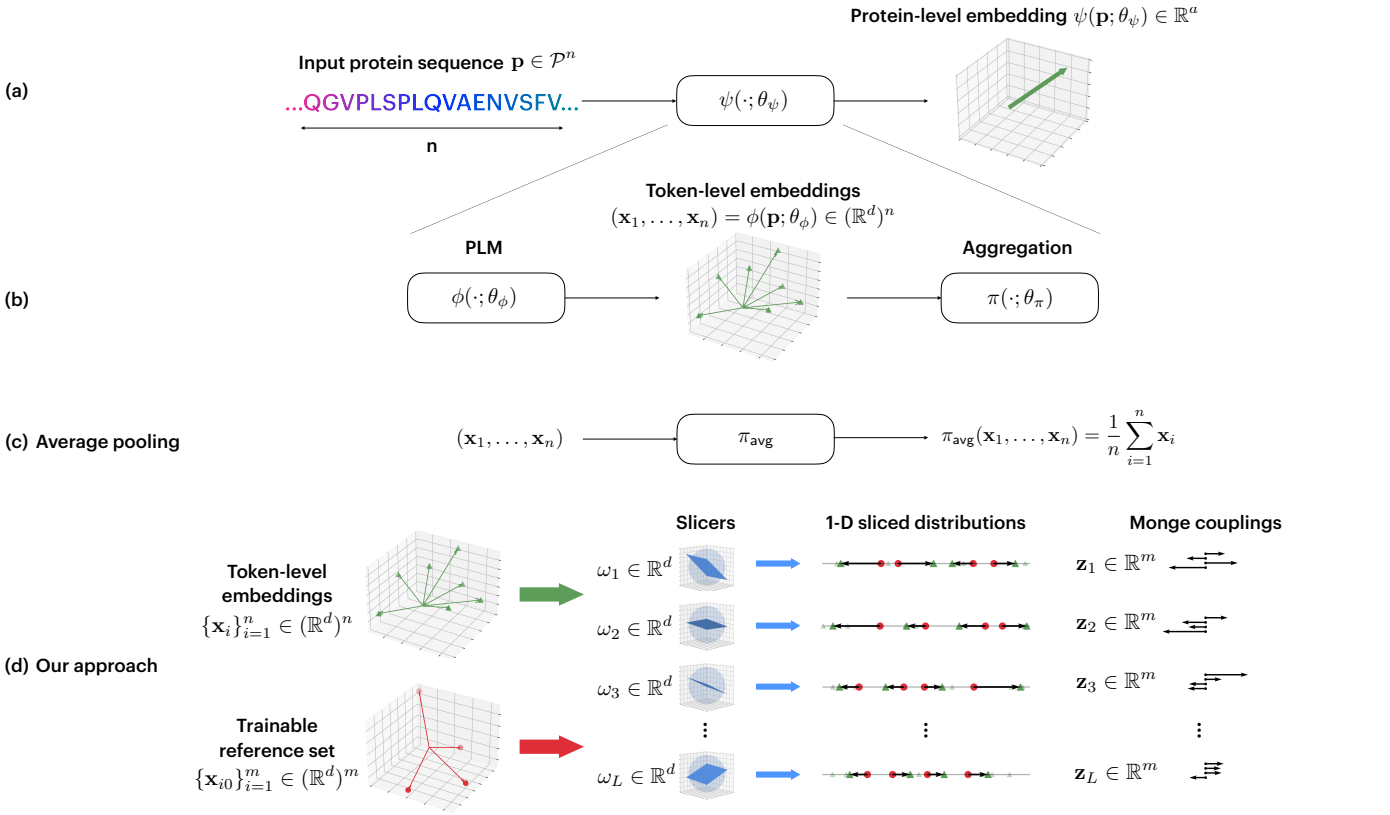

Fig. 1: Overview of the proposed method and comparison with average pooling. (a) We consider a parameterized protein representation learning function $\psi(\cdot;\theta_\psi)$, which takes as input a protein's amino acid sequence of arbitrary length and produces a fixed-length protein-level embedding at its output. (b) Breaking down the representation learning pipeline, we first pass the amino acid sequence through a pre-trained protein language model (PLM) $\phi(\cdot,\theta_\phi)$, thereby generating a set of token-level embeddings $(\mathbf{x}_1,\ldots,\mathbf{x}_n)$, each residing in $\mathbb{R}^d$. An aggregation function $\pi(\cdot,\theta_\pi)$ subsequently summarizes these embeddings into a protein-level embedding, whose size does not depend on the sequence length n. (c) Average pooling, which is most commonly used in the literature, simply takes the mean of the residue-level embeddings to derive the protein-level embedding. (d) Our proposed sliced-Wasserstein embedding aggregation module, which is based on comparing the probability distribution underlying the token-level embeddings, and a trainable reference probability distribution. Since such comparison is non-trivial in a high-dimensional space, we pass the token-level embeddings and the reference elements through a set of L trainable linear slicing operations $\{\omega_l\}_{l=1}^L$, which map the embeddings and reference elements into L pairs of 1-dimensional distributions. The Monge couplings between these 1-D distributions are then calculated based on sorting and interpolation operations, forming the basis for the final fixed-length protein-level embedding.

a closed-form solution, and in particular is an m-dimensional vector $\mathbf{z}_j^l = [z_{1j}^l,\ldots,z_{mj}^l]^T$ which relies solely on sorting and interpolation operations. In particular, depending on whether the length of the input protein sequence and the size of the reference set are identical, there are two cases:

- **Case 1** ($n_j = m$): In this scenario, the Monge map is simply the difference between the sorted sequences of points in the input slice and the reference slice. More precisely, let $\rho[\cdot]$ denote the permutation indices obtained by sorting $\{u_{ij}^l\}_{i=1}^{n_j}$. Moreover, let $\rho_0^{-1}[\cdot]$ denote the ordering that permutes the sorted set back to the original ordering based on sorting of elements in the reference set $\{u_{i0}^l\}_{i=1}^m$. Then, the Monge map elements are given by
$$z_{ij}^l = u_{i'j}^l - u_{i0}^l,\ \forall i \in \{1,\ldots,m\}, \tag{4}$$
where $i' = \rho[\rho_0^{-1}[i]]$.
- **Case 2** ($n_j \neq m$): The embedding procedure for this case follows similar steps as in Case 1, with the addition of an interpolation operation. Specifically, we derive the interpolated inverse cumulative distribution function (CDF) of the sliced token-level values, which we denote by F_{jl}^{-1}. This involves sorting $\{u_{ij}^l\}_{i=1}^{n_j}$, calculating the cumulative sum, and calculating the inverse through interpolation. Then, the elements of the Monge couplings can be derived as
$$z_{ij}^l = F_{jl}^{-1}\left(\frac{\rho_0^{-1}[i]}{m}\right) - u_{i0}^l,\ \forall i \in \{1,\ldots,m\}, \tag{5}$$
where ρ_0^{-1} is defined as in Case 1. Observe that (5) reduces to (4) if $n_j = m$.

Once the Monge couplings $\{\mathbf{z}_j^l\}_{l=1}^L$ have been derived for all slices, we concatenate them to form the embedding matrix $\mathbf{Z}_j = [\mathbf{z}_j^1,\ldots,\mathbf{z}_j^L]^T \in \mathbb{R}^{L\times m}$. To reduce the dimensionality of the embedding matrix, similarly to (Naderializadeh et al., 2021a), we perform a combination across the reference set elements using a learnable projection $\mathbf{w} \in \mathbb{R}^m$ to get $\mathbf{z}_j = \mathbf{Z}_j\mathbf{w} \in \mathbb{R}^L$. Finally, we use a linear mapping $\mathbf{V} \in \mathbb{R}^{d\times L}$, followed by an element-wise

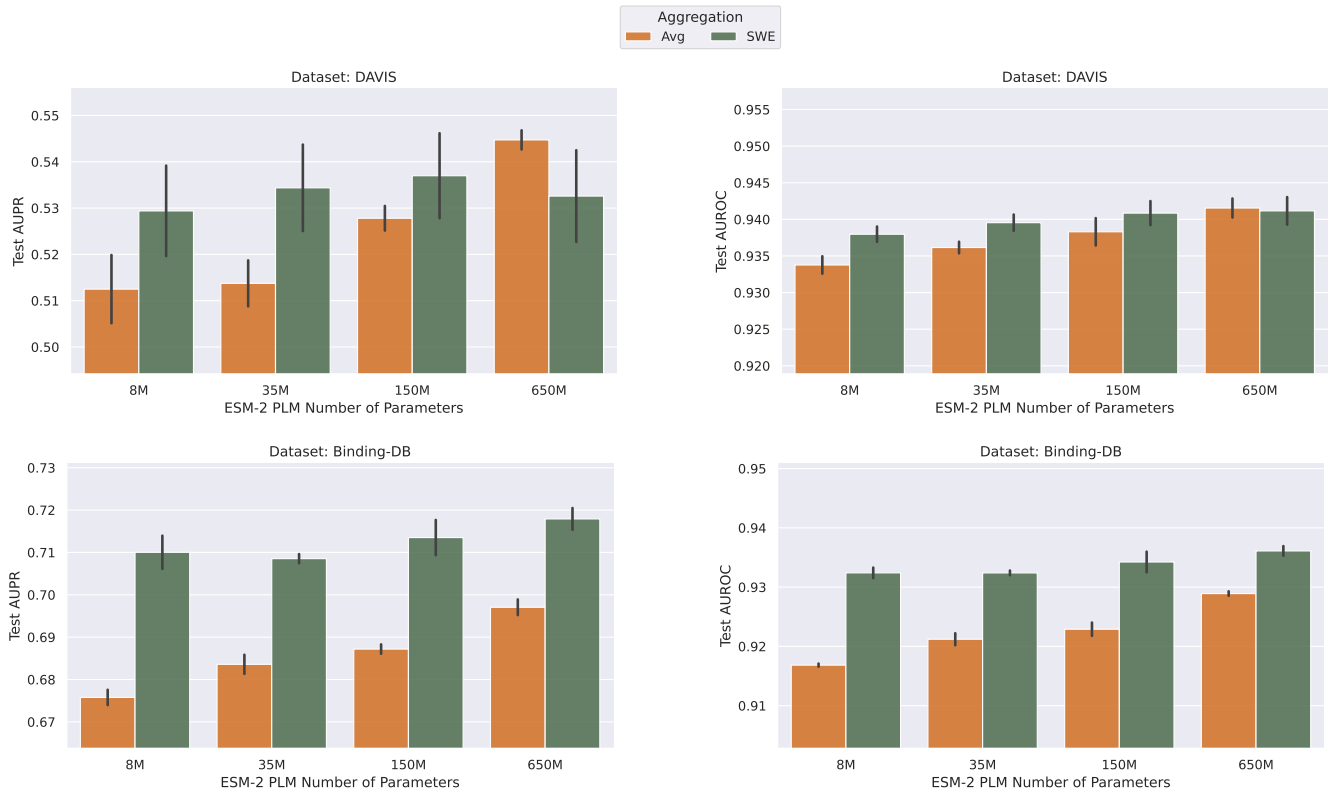

Fig. 2: Performance of the proposed SWE aggregation method as compared to average aggregation for the binary drug-target interaction task over the DAVIS (top) and BindingDB (bottom) datasets. Test performance is shown in terms of the area under the precision-recall curve (left) and the area under the ROC curve (right) across four different ESM-2 PLMs with increasing expressive power from left to right in each plot.

non-linearity $\sigma(\cdot)$, to derive the final protein-level embedding

$$\pi(\mathbf{x}_{1j}, \ldots, \mathbf{x}_{n_j j}; \theta_\pi) = \sigma(\mathbf{V}\mathbf{z}_j) \in \mathbb{R}^d, \tag{6}$$

with the parameters of the SWE aggregation function given by $\theta_\pi = \left(\{u_{i0}^l\}_{i=1,l=1}^{m,L}, \{\omega_l\}_{l=1}^{L}, \mathbf{w}, \mathbf{V}\right)$, where, to ease the optimization process, we learn the reference elements at the *output* of the slicers, training $\{u_{i0}^l\}_{i=1,l=1}^{m,L}$ directly instead of $\{\mathbf{x}_{i0}\}_{i=1}^{m}$.

Runtime and Memory Considerations. Note that the total number of parameters in the proposed SWE aggregation function is $mL + 2dL + m = \mathcal{O}((m+d)L)$. For the settings we consider in our numerical evaluations, as we describe next, this is a negligible overhead as compared to the size of the backbone PLMs.

3.3. Evaluation Settings

We use the state-of-the-art transformer-based ESM-2 family (Lin et al., 2022) as the PLMs $\phi(\cdot, \theta_\phi)$ that generate token-level embeddings. ESM-2 models have been pre-trained (via unsupervised mask-based objectives) using tens of millions of unique protein sequences and are shown to encode evolutionary patterns and significantly outperform baseline PLMs in structure prediction tasks. In particular, we evaluate our SWE aggregation method when applied to the outputs of four distinct pre-trained ESM-2 models, with 8M, 35M, 150M, and 650M parameters, in increasing order of complexity and expressive power.

For the SWE aggregation method, we consider four options for the number of slices, $L \in \{128, 256, 512, 1024\}$, as well as four options for the number of reference points, $m \in \{128, 256, 512, 1024\}$, leading to 16 different SWE configurations. In all experiments, we set the dimensionality of the final embedding space as $d = 1024$.

We evaluate the efficacy of our proposed SWE embedding method across three different tasks and four different datasets as described below:

- **Drug-Target Interaction (DTI) Prediction**: In this binary classification task, the goal is to predict whether or not a given drug interacts with a target protein. As in (Singh et al., 2023), for a given drug, we first find its Morgan fingerprint, denoted by $\mathbf{f} \in \mathbb{R}^c$, which we then map to the same embedding space as the target protein (i.e., $\mathbb{R}^d$), using a learnable projector $\mathbf{S} \in \mathbb{R}^{d\times c}$ and a non-linearity $\sigma(\cdot)$. We then use the cosine similarity between the drug and target embeddings to calculate their interaction probability, which we then optimize using the cross-entropy loss. In particular, for a given training dataset of (drug, target, label) triplets $\{(\mathbf{d}_j, \mathbf{p}_j, y_j)\}_{j=1}^N$, the ERM problem in (2) is reformulated as

$$\min_{\theta_\pi \in \Theta_\pi, \mathbf{S}\in\mathbb{R}^{d\times c}} \frac{1}{N}\sum_{j=1}^{N} \ell_{\mathsf{CosBCE}}\Big(\psi(\mathbf{p}_j; \theta_\phi, \theta_\pi), \sigma\left(\mathbf{S}\mathbf{f}_j\right), y_j\Big), \tag{7}$$

where the optimization occurs over the parameters of the drug projector, $\mathbf{S}$, and the SWE aggregation function, θ_π, while the parameters of the PLM, θ_ϕ, are kept frozen. The loss function $\ell_{\mathsf{CosBCE}} : \mathbb{R}_+^d \times \mathbb{R}_+^d \times \{0,1\} \to \mathbb{R}$ in (7) is defined as

$$\ell_{\mathsf{CosBCE}}(\mathbf{x}_1, \mathbf{x}_2, y) = -\Bigg[y \log\left(\frac{\mathbf{x}_1^T\mathbf{x}_2}{\|\mathbf{x}_1\|\|\mathbf{x}_2\|}\right) + (1-y)\log\left(1 - \frac{\mathbf{x}_1^T\mathbf{x}_2}{\|\mathbf{x}_1\|\|\mathbf{x}_2\|}\right)\Bigg]. \tag{8}$$

We use $\mathsf{ReLU}(\cdot)$ as the non-linearity $\sigma(\cdot)$ to produce both the drug and target embeddings in order to ensure they are constrained to the non-negative orthant of the embedding space, i.e., $\mathbb{R}_+^d$. We use two of the datasets used by Singh et al. (2023), namely DAVIS (Davis et al., 2011) and Binding-DB (Liu et al., 2007), to evaluate our proposed SWE-based embedding mechanism in this task.

- **Out-of-Domain Drug-Target Affinity Prediction**: In this regression task, we use the Therapeutics Data Commons (TDC) Drug-Target Interaction Domain Generalization (DTI-DG) Benchmark (Huang et al., 2021), where the goal is to predict the affinity of the interaction between drugs and protein targets patented between 2019 and 2021 by training on DTI interaction affinities patented in the preceding 5-year window (i.e., 2013–2018). We use the inner product between the drug and target embeddings to predict the interaction affinity and optimize the prediction model parameters using mean squared error (MSE). Especially, we replace the loss function in (7) with $\ell_{\mathsf{MSE}} : \mathbb{R}_+^d \times \mathbb{R}_+^d \times \mathbb{R}_+ \to \mathbb{R}_+$, defined as

$$\ell_{\mathsf{MSE}}(\mathbf{x}_1, \mathbf{x}_2, y) = \left[y - \left(\mathbf{x}_1^T \mathbf{x}_2\right)\right]^2 . \tag{9}$$

- **Protein-Protein Interaction (PPI) Prediction**: In the third task, we focus on predicting whether or not two given proteins will interact with each other. Specifically, we embed each protein's amino acid sequence using the same PLM and aggregation pipeline in parallel, and then we leverage the cosine similarity between the protein-level representations to estimate their interaction probability. For a given PPI training dataset consisting of N (protein, protein, label) triplets $\{(\overline{\mathbf{p}_j}, \underline{\mathbf{p}_j}, y_j)\}_{j=1}^N$, this task seeks to solve the following ERM reformulation of (2),

$$\min_{\theta_\pi \in \Theta_\pi, \mathbf{S} \in \mathbb{R}^{d \times c}} \frac{1}{N} \sum_{j=1}^{N} \ell_{\mathsf{CosBCE}} \left(\psi(\overline{\mathbf{p}_j}; \theta_\phi, \theta_\pi), \psi(\underline{\mathbf{p}_j}; \theta_\phi, \theta_\pi), y_j \right).$$

We use the "gold standard" dataset provided by Bernett et al. (2023) to evaluate our proposed SWE aggregation mechanism in this task, which comprises balanced PPI data without any leakage between training, validation, and testing samples.

We train the SWE aggregation parameters using the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 10^{-4} and cosine annealing schedule with a restart duration of 10 epochs. We run all our numerical experiments with five different random seeds for 50 epochs, using a batch size of 32 (except for PPI experiments with $L = 1024$ slices, where we use a reduced batch size of 24 due to computational limitations). We report the mean and standard deviation of the test/validation performance for the SWE configuration (i.e., (L, m) pair) with the highest target validation performance across the 50 training epochs. The target validation performance metric for the binary DTI and PPI prediction tasks is set to the validation AUPR (area under the precision-recall curve), while it is set to the validation PCC (Pearson correlation coefficient) for the DTI affinity prediction task.

4. Results

4.1. Binary Drug-Target Interaction Prediction

Figure 2 shows the performance of the proposed SWE-based aggregation method for the binary DTI task, evaluated over the DAVIS and Binding-DB datasets. In most scenarios, our proposed SWE aggregation function performs on par with or better than simply averaging the token-level embeddings. Our method especially shines when i) the PLM is smaller and has lower expressive power (e.g., 8M vs. 650M ESM-2 PLMs), and ii) there is more data to train the SWE aggregation parameters (Binding-DB vs. DAVIS). While the latter point is intuitive since SWE aggregation introduces additional trainable parameters and is, therefore, prone to overfitting, the former point is very significant from an efficiency point of view. In particular, given limited computational resources, where one is constrained to using smaller PLMs, our proposed aggregation operation can, in most cases, push the performance of the smaller PLMs to similar, or even higher, levels than larger PLMs whose outputs are aggregated using averaging. Given the presence of enough training data, especially in the case of Binding-DB, our results demonstrate that there are also gains to be achieved for larger PLM backbones with hundreds of millions of parameters when using SWE-based aggregation as compared to average pooling.

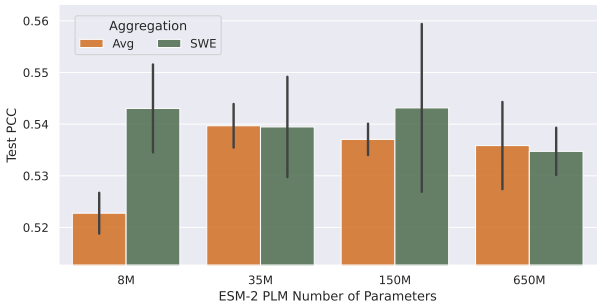

Fig. 3: Performance comparison of the proposed SWE aggregation method and average pooling in the drug-target affinity prediction task across four different ESM-2 models. Test performance is shown in terms of the Pearson correlation coefficient.

4.2. Out-of-Domain Drug-Target Affinity Prediction

Figure 3 shows the test PCC levels achieved by our proposed SWE aggregation method as compared to average pooling. We observe that in this task, the SWE pooling mechanism performs similarly to, or better than, average pooling for all the ESM-2 PLMs. With a test PCC of 0.543 ± 0.015, our proposed method ranks second in the TDC DTI-DG leaderboard [1] as of this writing.

4.3. Protein-Protein Interaction Prediction

Table 1 compares the validation and test performance of our proposed SWE aggregation method with average pooling across the considered ESM-2 PLMs. As the table shows, SWE generally outperforms average pooling in terms of F1-score and recall, while performing on par with average pooling in terms of accuracy, and underperforming for the other metrics. These results are consistent with the ones reported by Sledzieski et al. (2023), where more expressive models and heavier fine-tuning lead to superior F1 and recall levels, while lowering the other metrics. Further research is required on how to strike the right balance among the different metrics through which PPI prediction quality is measured.

4.4. Impact of the Number of Slices and the Reference Set Size

Figure 4 shows the validation AUPR gains that our proposed SWE-based aggregation method achieves over average pooling for the considered ESM-2 PLMs. Three remarks are in order: i) As

[1] https://tdcommons.ai/benchmark/dti_dg_group/bindingdb_patent/

	ESM-2 PLM	Aggregation	Accuracy	F1	MCC	AUPR	Precision	Recall	Specificity
Validation	8M	SWE	0.607 ± 0.003	0.626 ± 0.008	0.216 ± 0.006	0.652 ± 0.004	0.619 ± 0.006	0.671 ± 0.023	0.701 ± 0.013
		Avg	0.606 ± 0.002	0.614 ± 0.004	0.216 ± 0.004	0.653 ± 0.002	0.637 ± 0.004	0.643 ± 0.012	0.731 ± 0.008
	35M	SWE	0.617 ± 0.002	0.632 ± 0.009	0.235 ± 0.004	0.669 ± 0.003	0.627 ± 0.002	0.693 ± 0.024	0.690 ± 0.009
		Avg	0.617 ± 0.001	0.628 ± 0.008	0.237 ± 0.002	0.667 ± 0.001	0.648 ± 0.001	0.666 ± 0.024	0.735 ± 0.005
	150M	SWE	0.623 ± 0.002	0.633 ± 0.006	0.248 ± 0.003	0.677 ± 0.003	0.633 ± 0.005	0.663 ± 0.021	0.681 ± 0.011
		Avg	0.624 ± 0.002	0.635 ± 0.006	0.251 ± 0.004	0.679 ± 0.003	**0.661 ± 0.004**	0.676 ± 0.023	**0.754 ± 0.008**
	650M	SWE	**0.629 ± 0.002**	**0.668 ± 0.001**	**0.259 ± 0.004**	**0.685 ± 0.002**	0.630 ± 0.008	**0.810 ± 0.004**	0.643 ± 0.023
		Avg	0.626 ± 0.001	0.650 ± 0.004	0.251 ± 0.002	0.684 ± 0.002	0.651 ± 0.003	0.732 ± 0.030	0.732 ± 0.008
Test	8M	SWE	0.637 ± 0.004	0.638 ± 0.013	0.274 ± 0.008	0.688 ± 0.005	0.636 ± 0.007	0.642 ± 0.030	0.632 ± 0.027
		Avg	0.639 ± 0.000	0.637 ± 0.005	0.279 ± 0.001	0.690 ± 0.000	0.641 ± 0.004	0.633 ± 0.013	**0.645 ± 0.013**
	35M	SWE	0.646 ± 0.003	0.661 ± 0.010	0.294 ± 0.004	0.696 ± 0.005	0.635 ± 0.011	0.690 ± 0.035	0.602 ± 0.040
		Avg	0.648 ± 0.000	0.651 ± 0.010	0.297 ± 0.001	0.699 ± 0.000	0.646 ± 0.008	0.658 ± 0.029	0.639 ± 0.028
	150M	SWE	0.649 ± 0.002	0.664 ± 0.007	0.299 ± 0.005	0.703 ± 0.002	0.636 ± 0.005	0.696 ± 0.020	0.601 ± 0.020
		Avg	0.651 ± 0.001	0.654 ± 0.009	0.303 ± 0.003	0.708 ± 0.002	**0.649 ± 0.006**	0.659 ± 0.025	0.643 ± 0.023
	650M	SWE	0.650 ± 0.005	**0.683 ± 0.007**	0.307 ± 0.007	0.705 ± 0.005	0.625 ± 0.012	**0.754 ± 0.033**	0.545 ± 0.042
		Avg	**0.657 ± 0.002**	0.670 ± 0.006	**0.315 ± 0.003**	**0.717 ± 0.001**	0.646 ± 0.007	0.696 ± 0.022	0.618 ± 0.024

Table 1. PPI validation and test results on the "gold standard" dataset by Bernett et al. (2023) for the SWE and average pooling methods across four different ESM-2 PLM backbones. Following (Sledzieski et al., 2023), for a comprehensive evaluation, we report the performance using seven different metrics, including accuracy, F1-score, Matthews correlation coefficient (MCC), AUPR, precision, recall, and specificity, with mean and standard deviation across five different random seeds. Underlined numbers indicate the best aggregation performer in each metric for each phase (validation/test) and each PLM. Bold numbers indicate the best performer in each metric for each phase (validation/test).

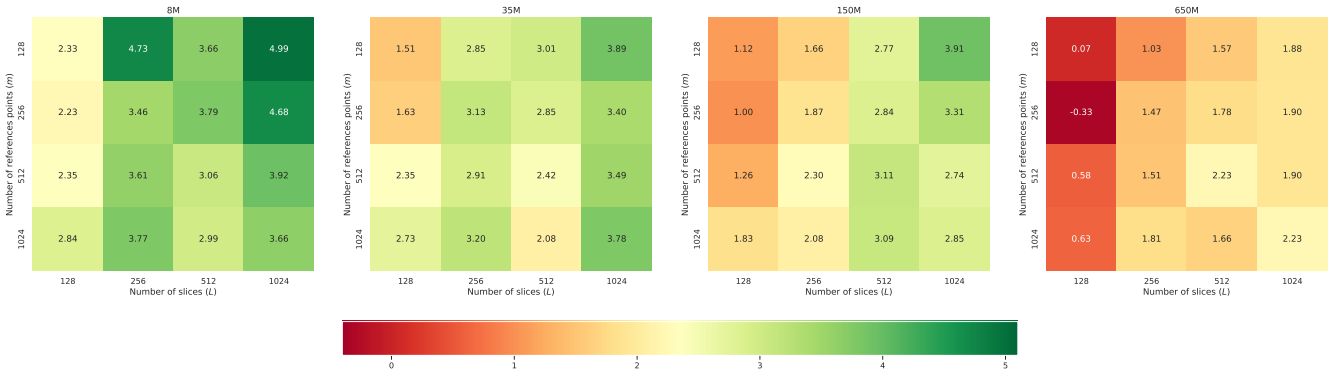

Fig. 4: Average validation AUPR gains (%) of the proposed SWE aggregation method over average pooling for different numbers of slices and reference points across different ESM-2 PLMs in the Binding-DB DTI prediction task.

we mentioned earlier, the performance gains are more pronounced when the backbone PLM is less powerful and vice versa; ii) On average, increasing the number of slices (i.e., L) boosts the downstream performance of the proposed SWE aggregation, which is expected since such an increase leads to a more accurate Monte-Carlo approximation of the sliced-Wasserstein distances between the distributions induced by the per-token embeddings (Kolouri et al., 2019); and, iii) The correlation between the downstream performance and the size of the reference set (i.e., m) is less clear and depends on the context. As suggested in prior work (Kolouri et al., 2021; Naderializadeh et al., 2021a), a reasonable choice for the number of reference points is the average length of the amino acid sequences in the training set, but in general, this is an important hyperparameter than needs to be optimized.

4.5. Interpretability of the Learned SWE Representations

One of the desirable properties of the proposed embeddings is that the sliced-Wasserstein distance of the distributions underlying the token-level embeddings of two given protein sequences can be approximated by the average distance of their Monge couplings to the reference across different slices (Naderializadeh et al., 2021a). In particular, for two proteins $\overline{\mathbf{p}}$ and $\underline{\mathbf{p}}$ with Monge coupling matrices (as defined in Section 3.2) $\overline{\mathbf{Z}} = [\overline{\mathbf{z}}^1, \dots, \overline{\mathbf{z}}^L]^T \in \mathbb{R}^{L\times m}$ and $\underline{\mathbf{Z}} = [\underline{\mathbf{z}}^1, \dots, \underline{\mathbf{z}}^L]^T \in \mathbb{R}^{L\times m}$, respectively, we can approximate their pairwise sliced-Wasserstein distance as

$$\mathcal{SW}_2(\overline{\mathbf{p}}, \underline{\mathbf{p}}) \approx \left(\frac{1}{L} \sum_{l=1}^{L} \left\| \overline{\mathbf{z}}^l - \underline{\mathbf{z}}^l \right\|_2^2 \right)^{\frac{1}{2}}, \tag{10}$$

where, with a slight abuse of notation, we use $\mathcal{SW}_2(\overline{\mathbf{p}}, \underline{\mathbf{p}})$ to denote the sliced-Wasserstein distance between token-level representations of $\overline{\mathbf{p}}$ and $\underline{\mathbf{p}}$ at the output of the PLM backbone.

The approximation in (10) allows us to visualize the pairwise distance of interacting and non-interacting proteins in the embedding space. Figure 5 shows the distributions of (approximate) SW distances between proteins in the training dataset separated by whether or not they interact, where the embeddings are generated by the 650M-parameter ESM-2 backbone and a trained SWE aggregation module with $m = 1024$ reference points and $L = 1024$ slices. As the figure demonstrates, there is a separation between the two histograms, with interacting proteins landing closer to each other in the embedding space from a sliced-Wasserstein distance point of view as compared to non-interacting pairs of proteins.

5. Discussion

We introduce a novel approach for protein representation based on optimal transport principles. We anticipate that our method could have applications beyond PLMs to other foundation models in biology (e.g., for DNA sequence models). Our interpretable

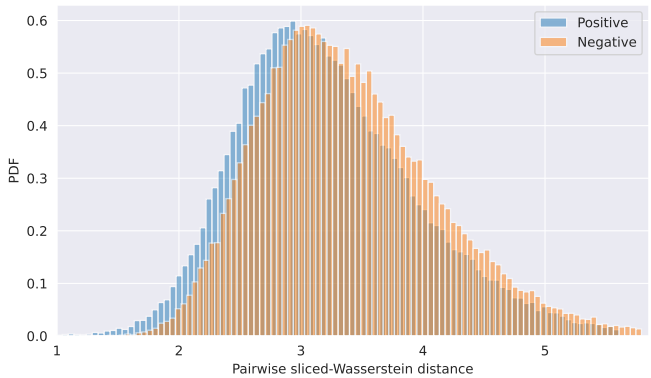

Fig. 5: Histograms of sliced-Wasserstein distance between embeddings of interacting and non-interacting proteins at the output space of a pre-trained ESM-2 PLM with 650M parameters. The SW distance is calculated using a trained SWE pooling module with $m = 1024$ reference points and $L = 1024$ slices.

approach summarizes per-token embeddings in terms of their distance from a set of reference embeddings that are learned from the data. Compared to the conventional approach of average pooling, this learnability provides our approach with greater flexibility in capturing task-specific knowledge. We observed that pooling work acceptably well for very large PLMs, where the greater complexity of the sequence representation compensates for the lack of learnability in the summarization layer. However, these large models have memory requirements that are beyond the capabilities of many GPUs. Sophisticated summarization approaches will be crucial in maximizing the power of the smaller models that can fit on common GPUs and could further unlock fine-tuning opportunities for such pre-trained models.

Future work could focus on further exploring the interpretability of our method. One direction of research could be to refine the selection of reference embeddings, potentially incorporating an auxiliary loss function to encode greater biological intuition into these models. Such an approach would enhance the biological relevance of the representations produced by our method. Our work thus opens up new pathways for enhancing the interpretability and efficiency of PLM-based approaches in biology.

References

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

T. Bepler and B. Berger. Learning the protein language: Evolution, structure, and function. *Cell Systems*, 12(6):654–669.e3, 2021. ISSN 2405-4712. doi: https://doi.org/10.1016/j.cels.2021.05.017.

J. Bernett, D. B. Blumenthal, and M. List. Cracking the black box of deep sequence-based protein-protein interaction prediction. *bioRxiv*, 2023. doi: 10.1101/2023.01.18.524543.

N. Brandes, G. Goldman, C. H. Wang, C. J. Ye, and V. Ntranos. Genome-wide prediction of disease variant effects with a deep protein language model. *Nature Genetics*, 55(9):1512–1522, 2023. doi: 10.1038/s41588-023-01465-0.

B. Chen, X. Cheng, P. Li, Y.-a. Geng, J. Gong, S. Li, Z. Bei, X. Tan, B. Wang, X. Zeng, et al. xtrimopglm: unified 100b-scale pre-trained transformer for deciphering the language of protein. *arXiv preprint arXiv:2401.06199*, 2024.

M. Chen, C. J.-T. Ju, G. Zhou, X. Chen, T. Zhang, K.-W. Chang, C. Zaniolo, and W. Wang. Multifaceted protein–protein interaction prediction based on siamese residual rcnn. *Bioinformatics*, 35(14):i305–i314, 2019.

T. Chen, P. Vure, R. Pulugurta, and P. Chatterjee. AMP-diffusion: Integrating latent diffusion with protein language models for antimicrobial peptide generation. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*, 2023. URL `https://openreview.net/forum?id=145TM9VQhx`.

M. I. Davis, J. P. Hunt, S. Herrgard, P. Ciceri, L. M. Wodicka, G. Pallares, M. Hocker, D. K. Treiber, and P. P. Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature Biotechnology*, 29(11):1046–1051, 2011. doi: 10.1038/nbt.1990.

I. Deshpande, Y.-T. Hu, R. Sun, A. Pyrros, N. Siddiqui, S. Koyejo, Z. Zhao, D. Forsyth, and A. G. Schwing. Max-sliced Wasserstein distance and its use for GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10648–10656, 2019.

A. Elnaggar, M. Heinzinger, C. Dallago, G. Rehawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, M. Steinegger, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2021.

N. Ferruz, S. Schmidt, and B. Höcker. ProtGPT2 is a deep unsupervised language model for protein design. *Nature Communications*, 13(1):4348, 2022. doi: 10.1038/s41467-022-32007-7.

K. Huang, T. Fu, W. Gao, Y. Zhao, Y. Roohani, J. Leskovec, C. W. Coley, C. Xiao, J. Sun, and M. Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *Proceedings of Neural Information Processing Systems, NeurIPS Datasets and Benchmarks*, 2021.

D. Iliadis, B. D. Baets, T. Pahikkala, and W. Waegeman. A comparison of embedding aggregation strategies in drug-target interaction prediction. *bioRxiv*, 2023. doi: 10.1101/2023.09.25.559265.

K. Kaminski, J. Ludwiczak, K. Pawlicki, V. Alva, and S. Dunin-Horkawicz. plm-blast: distant homology detection based on direct comparison of sequence representations from protein language models. *Bioinformatics*, 39(10):btad579, 2023.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. Rohde. Generalized sliced Wasserstein distances. *Advances in neural information processing systems*, 32, 2019.

S. Kolouri, N. Naderializadeh, G. K. Rohde, and H. Hoffmann. Wasserstein embedding for graph learning. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=AAes_3W-2z`.

R. Kortvelesy and A. Prorok. QGNN: Value function factorisation with graph neural networks. *arXiv preprint arXiv:2205.13005*, 2022.

J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y. W. Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International Conference on Machine Learning*, pages 3744–3753. PMLR, 2019.

Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

M. Littmann, M. Heinzinger, C. Dallago, K. Weissenow, and B. Rost. Protein embeddings and deep learning predict binding residues for various ligand classes. *Scientific Reports*, 11(1): 23916, 2021.

T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson. BindingDB: A web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research*, 35(suppl_1):D198–D201, 2007.

Y. Lu, X. Liu, A. Soltoggio, and S. Kolouri. Slosh: Set locality sensitive hashing via sliced-wasserstein embeddings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2566–2576, 2024.

H. Maron, H. Ben-Hamu, N. Shamir, and Y. Lipman. Invariant and equivariant graph networks. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Syx72jC9tm`.

G. Mialon, D. Chen, A. d'Aspremont, and J. Mairal. A trainable optimal transport embedding for feature aggregation and its relationship to attention. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=ZK6vTvb84s`.

R. L. Murphy, B. Srinivasan, V. Rao, and B. Ribeiro. Janossy pooling: Learning deep permutation-invariant functions for variable-size inputs. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=BJluy2RcFm`.

N. Naderializadeh, F. H. Hung, S. Soleyman, and D. Khosla. Graph convolutional value decomposition in multi-agent reinforcement learning. *arXiv preprint arXiv:2010.04740*, 2020.

N. Naderializadeh, J. F. Comer, R. Andrews, H. Hoffmann, and S. Kolouri. Pooling by sliced-Wasserstein embedding. *Advances in Neural Information Processing Systems*, 34:3389–3400, 2021a.

N. Naderializadeh, S. Kolouri, J. F. Comer, R. W. Andrews, and H. Hoffmann. Set representation learning with generalized sliced-Wasserstein embeddings. *arXiv preprint arXiv:2103.03892*, 2021b.

C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017a.

C. R. Qi, L. Yi, H. Su, and L. J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017b.

R. M. Rao, J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, and A. Rives. Msa transformer. In *International Conference on Machine Learning*, pages 8844–8856. PMLR, 2021.

S. Ravanbakhsh, J. Schneider, and B. Poczos. Deep learning with sets and point clouds. *arXiv preprint arXiv:1611.04500*, 2016.

A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, and R. Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*, 2019. doi: 10.1101/622803.

R. Singh, S. Sledzieski, B. Bryson, L. Cowen, and B. Berger. Contrastive learning in protein language space predicts interactions between drugs and protein targets. *Proceedings of the National Academy of Sciences*, 120(24):e2220778120, 2023. doi: 10.1073/pnas.2220778120.

S. Sledzieski, R. Singh, L. Cowen, and B. Berger. D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions. *Cell Systems*, 12(10):969–982.e6, 2021. ISSN 2405-4712.

S. Sledzieski, M. Kshirsagar, M. Baek, B. Berger, R. Dodhia, and J. L. Ferres. Democratizing protein language models with parameter-efficient fine-tuning. *bioRxiv*, 2023. doi: 10.1101/2023.11.09.566187.

H. Stärk, C. Dallago, M. Heinzinger, and B. Rost. Light attention predicts protein location from the language of life. *Bioinformatics Advances*, 1(1):vbab035, 11 2021. ISSN 2635-0041. doi: 10.1093/bioadv/vbab035.

J. Su, C. Han, Y. Zhou, J. Shan, X. Zhou, and F. Yuan. Saprot: Protein language modeling with structure-aware vocabulary. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=6MRm3G4NiU`.

P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls, et al. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*, 2017.

H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

T. Tu, G. Krishna, and A. Aghazadeh. Protigeno: a prokaryotic short gene finder using protein language models. *arXiv preprint arXiv:2307.10343*, 2023.

S. Unsal, H. Atas, M. Albayrak, K. Turhan, A. C. Acar, and T. Doğan. Learning functional properties of proteins with language models. *Nature Machine Intelligence*, 4(3):227–245, 2022. doi: 10.1038/s42256-022-00457-9.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

C. Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

A. Villegas-Morcillo, A. M. Gomez, and V. Sanchez. An analysis of protein language model embeddings for fold prediction. *Briefings in Bioinformatics*, 23(3):bbac142, 04 2022. ISSN 1477-4054. doi: 10.1093/bib/bbac142.

E. Wagstaff, F. Fuchs, M. Engelcke, I. Posner, and M. A. Osborne. On the limitations of representing functions on sets. In *International Conference on Machine Learning*, pages 6487–6494. PMLR, 2019.

F. Wu and S. Z. Li. A hierarchical training paradigm for antibody structure-sequence co-design. *arXiv preprint arXiv:2311.16126*, 2023.

M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.

Y. Zhang, J. Hare, and A. Prügel-Bennett. FSPool: Learning set representations with featurewise sort pooling. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=HJgBA2VYwH`.