





A Transferability-Based Method for Evaluating the Protein Representation Learning

Fan Hu , Weihong Zhang , Huazhen Huang, Wang Li, Yang Li , and Peng Yin 

Abstract—Self-supervised pre-trained language models have recently risen as a powerful approach in learning protein representations, showing exceptional effectiveness in various biological tasks, such as drug discovery. Amidst the evolving trend in protein language model development, there is an observable shift towards employing large-scale multimodal and multitask models. However, the predominant reliance on empirical assessments using specific benchmark datasets for evaluating these models raises concerns about the comprehensiveness and efficiency of current evaluation methods. Addressing this gap, our study introduces a novel quantitative approach for estimating the performance of transferring multi-task pre-trained protein representations to downstream tasks. This transferability-based method is designed to quantify the similarities in latent space distributions between pre-trained features and those fine-tuned for downstream tasks. It encompasses a broad spectrum, covering multiple domains and a variety of heterogeneous tasks. To validate this method, we constructed a diverse set of protein-specific pre-training tasks. The resulting protein representations were then evaluated across several downstream biological tasks. Our experimental results demonstrate a robust correlation between the transferability scores obtained using our method and the actual transfer performance observed. This significant correlation highlights the potential of our method as a more comprehensive and efficient tool for evaluating protein representation learning.

Index Terms—Transferability, protein representation learning, optimal transport.

I. INTRODUCTION

PROTEINS are central to fundamental biological processes, making the development of effective protein representations crucial in computational biology. These representations serve to condense complex raw data into a more manageable, often low-dimensional space, capturing essential features for tasks like predictive modeling or interpretive exploration. Such a process is instrumental in enhancing model performance [1] and is pivotal for understanding protein functions in complex diseases and developing corresponding therapeutic medications.

Recently, the use of self-supervised pre-trained language models has become a prominent approach in protein representation learning. Demonstrating remarkable efficacy, these models have delivered exceptional performance across various biological tasks [2], [3], [4], [5], [6]. This method leverages transfer learning, beginning with extensive pre-training on large datasets to learn linguistic patterns and knowledge, subsequently applying this learned knowledge to specific downstream tasks [7], [8], [9].

Despite these advancements, there remains a degree of uncertainty regarding the extent to which pre-training enhances protein representation. This uncertainty stems from findings that some existing methods are suboptimal [10]. Traditional studies focus primarily on predictive performance in specific downstream benchmark tasks. While these benchmarks are valuable, they demand considerable computational resources and fail to comprehensively measure the models' transferability to a wide range of downstream tasks. This gap limits our ability to design more informed and effective strategies for protein representation. Consequently, there is an urgent need for a method that quantifies the transferability of pre-trained protein representations, effectively bridging the gap between theoretical expectations and practical outcomes. This need opens up two interrelated areas of inquiry:

A. Protein Representation Learning

All representation learning should prioritize the training objectives, which guide the direction of model optimization and determine the relevant information to be extracted. The pre-training objectives of existing protein language models are derived from similar tasks in natural language processing (NLP) [11], [12],

Manuscript received 8 November 2023; revised 24 January 2024; accepted 23 February 2024. Date of publication 28 February 2024; date of current version 7 May 2024. This work was supported in part by the National Key R&D Program of China under Grant 2022YFA1008300, in part by the Strategic Priority Research Program of Chinese Academy of Sciences under Grant XDB 38050100, in part by the Shenzhen Science and Technology Innovation Committee under Grant JCYJ20220818101216035, in part by Shenzhen Medical Research Funds under Grant A2303032, in part by the Shenzhen Science and Technology Program under Grant SGDX20201103095603009 and Grant JSGG20200225153023511, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515110647, and in part by the National Natural Science Foundation of China under Grant U22A2041. (Fan Hu and Weihong Zhang contributed equally to this work.) (Corresponding authors: Fan Hu; Peng Yin.)

Fan Hu, Weihong Zhang, Huazhen Huang, Wang Li, and Peng Yin are with the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: fan.hu@siat.ac.cn; wh.zhang2@siat.ac.cn; hz.huang@siat.ac.cn; w.li1@siat.ac.cn; peng.yin@siat.ac.cn).

Yang Li is with the Tsinghua Shenzhen International Graduation School, Tsinghua University, Shenzhen 518071, China (e-mail: yangli@sz.tsinghua.edu.cn).

Code and data are freely available at <https://github.com/SIAT-code/OTMTD>.

Digital Object Identifier 10.1109/JBHI.2024.3370680

such as masked language modeling (MLM) and next token prediction. Typically, a protein language model is based on the NLP similar distributional hypothesis, which assumes amino acids that frequently appear in the same contexts may have equivalent semantic information.

However, due to the complex characteristics of proteins, it is challenging to infer complete semantics from sequence data alone. There have been numerous attempts to incorporate structural information (such as a protein contact map) or functional knowledge (such as Gene Ontology, GO) into protein language models [6], [13], [14], [15], [16]. In addition to integrating modalities beyond protein sequences, these methods also incorporated new pre-training tasks, such as contact map prediction and protein function prediction.

The current mainstream methods are based on the “Pre-train and Fine-tune” paradigm. Specifically, protein representations are pre-trained on large-scale data and then fine-tuned for downstream tasks. However, a gap exists between the pre-training and fine-tuning processes. It remains uncertain whether the unique features required by a particular downstream task are effectively captured during the pre-training phase. The pre-train-fine-tune paradigm could potentially hinder performance if the most crucial information is omitted during the pre-training phase. Therefore, a method capable of quantifying the transferability from a pre-trained protein representation to downstream tasks is essential.

B. Transferability

In transfer learning, the goal is to leverage knowledge or models acquired in one situation (source task) to improve performance in a different but related situation (target task). A transferability method quantitatively assesses how much of the knowledge transferred from the source task to the target task. It essentially acts as a roadmap for implementing transfer learning in practical applications, such as aiding in the selection of highly transferable tasks for joint training.

Research on transferability primarily falls into two categories: empirical and analytical studies. Empirical studies involve retraining the source model on new tasks and assessing task relationships using metrics like validation accuracy [17], [18]. This approach, while direct, can be computationally intensive and dependent on the specific models used. For example, Task2vec [18] is an empirical method where target data is processed through a probe network to compute a task embedding, predicting task similarities. However, such methods often require significant computational resources.

On the other hand, analytical methods provide a more computationally efficient way to estimate transferability. [19], [20], [21], [22]. These methods do not rely on retraining models but instead use mathematical and statistical techniques to predict how well knowledge from one task can transfer to another. For example, OTCE [23] [24], a recent analytical method, employs Wasserstein distance to estimate domain differences and conditional entropy to characterize task differences. Similarly, Liu et al. developed a method that calculates the similarity between

two tasks by embedding them into a vector space and using the Euclidean distance as a surrogate for the 2-Wasserstein distance [22].

However, these methods focus primarily on image and text classification so far. In these fields, the domain difference (e.g., different image styles) and task difference (e.g., classification from 3-category to 5-category) are much smaller than in the biological area (Fig. 2). More importantly, existing transferability methods are designed for simple classification tasks and are unsuitable for more complex and heterogeneous tasks (e.g., regression, long-tail multi-class) in the biological area.

To address these challenges, we propose a novel strategy to analytically predict the transfer performance of a multi-task pre-trained protein representation to downstream biological tasks. Specifically, this method quantitatively measures the similarities across multiple domains and heterogeneous tasks. To validate the effectiveness of the method, we devised a series of combinations of protein-specific pre-training tasks, yielding various pre-trained protein representations. These representations were then evaluated on several downstream biological benchmarks. As expected, a strong correlation (Spearman’s $R = 0.709$) was observed between predicted transferability and actual transfer performance.

The following are the main contributions of this paper:

- 1) We propose a strategy for quantifying the transferability from a pre-trained protein representation to downstream tasks, particularly in situations involving multiple heterogeneous tasks, which are extremely common in the biological area.
- 2) This method is capable of predicting transfer performance without the need for labor-intensive fine-tuning on downstream tasks. The effectiveness of this method has been confirmed by the high performance on the extensive experiments. This method can be utilized to guide the design of protein representation pre-training, particularly in the selection of pre-training objectives.

II. MATERIALS AND METHODS

A. Datasets

1) *Pretraining Data*: We assembled a multimodal protein dataset, comprising approximately 1 million entries, for pre-training. This dataset included three types of data: sequence, structure, and functional annotation. Specifically, the sequence and GO annotation were procured from UniProtKB Swiss-Prot (<https://www.uniprot.org/>), while the structure data was sourced from the AlphaFold Protein Structure Database (AlphaFold DB: <https://alphafold.ebi.ac.uk/download>) and the RCSB PDB (<https://www.rcsb.org/>). We collected fine-grained domain knowledge of proteins, including regions, motifs, and domains, from UniProtKB. Further details can be found in our previous study [6].

2) *Downstream Benchmark*: The selection of downstream tasks was guided by the objective to encompass a broad spectrum of protein-related tasks, each representing a distinct category of protein analysis. Specifically, we categorized these tasks into

three major groups based on their nature: 1) Protein properties: This category includes tasks such as predicting stability, fluorescence, and signal peptide [3]. 2) Protein structure-related tasks: This group encompasses tasks related to predicting secondary structures, remote homology, and fold classes [3]. 3) Protein interactions with other molecules: This category encompasses tasks like PDBbind [25] and Kinase [26], which involve interactions between proteins and other molecules. In selecting specific downstream tasks, we aimed to include widely-used benchmark datasets within each of these categories. The preprocessing code for these datasets, along with comprehensive annotations, is available for public access on our GitHub repository at <https://github.com/SIAT-code/OTMTD>.

B. Model Architecture and Pretraining Objectives

The model architecture is identical to that used in our previous study [6]. Briefly, the input protein sequences, structures, and GO annotations were processed by their respective encoders to obtain initial embeddings. Protein sequence and structure embeddings were aligned at the token level. This embedding was then globally aligned with the GO annotation embedding to produce the protein multimodal embedding. Subsequently, this protein multimodal embedding was pre-trained on a series of combinations of pre-training objectives to obtain various protein representations, which were then evaluated on a number of downstream biological tasks.

More specifically, five protein-specific pre-training objectives were used, including prediction of masked amino acids (MLM), prediction of masked GO terms (GO), and predictions of amino acids and locations within protein regions (R), motifs (M), and domains (D). For the MLM and GO tasks, a specific percentage of amino acids and Gene Ontology terms were masked, with the pre-training objectives being their accurate prediction. For the Regions, Motifs, and Domains tasks, we employed a strategy akin to the named entity recognition approach used in natural language processing. Specifically, we considered each category (e.g., motif 1, domain 2) as a named entity, subdividing each entity into combinations of individual amino acids, which were then classified. Then we pre-trained four models by taking different combinations of objectives: (1) MLM + RMD, (2) MLM + GO + D, (3) GO + RMD, (4) RMD.

C. Optimal Transport-Based Multi-Task Distance

In this section, we introduce the Optimal Transport-based Multi-Task Distance (OTMTD), a novel approach for evaluating the transferability of multi-task pre-trained protein representations to various biological tasks. Unlike previous methods, OTMTD incorporates both feature and heterogeneous label information (e.g., regression, long-tail multi-class), offering an efficient and effective way to measure transferability across diverse and complex biological tasks. Here's a step-by-step breakdown of how OTMTD works:

As depicted in Fig. 1, our approach begins with the protein feature representations and multi-task labels from pre-training, coupled with those of a downstream task. Initially, we use multi-dimensional scaling (MDS) to transform the label-to-label

data into a simplified, embedded format. This streamlined label embedding is then merged with its corresponding feature representation to create a joint distribution. The next step involves implementing the Wasserstein Task Embedding framework [22], which projects this joint distribution into an embedding space. In this space, the Euclidean distances between task embeddings are indicative of transferability, reflecting the potential for effective knowledge transfer from pre-training to the downstream task. More specifically:

Preliminary The Kantorovich Optimal Transport problem can be seen as a method to optimally reshuffle the mass of one distribution (e.g., data points representing protein structures) into another, ensuring the least amount of 'work' is done. 'Work' here is quantified as the product of the mass moved and the distance it is moved. Specifically, let \mathcal{X} be a metric space, along with continuous or discrete probability measures $\alpha \in \mathcal{P}(\mathcal{X})$ and $\beta \in \mathcal{P}(\mathcal{X})$ [23]. The Kantorovich OT problem is defined as:

$$OT(\alpha, \beta) \triangleq \min_{\pi \in \Pi(\alpha, \beta)} \int_{\mathcal{X} \times \mathcal{X}} c(x, y) d\pi(x, y), \quad (1)$$

Here, $c(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}^+$ is a cost function, and $\Pi(\alpha, \beta)$ is a set of couplings consisting of joint distributions over the product space $\mathcal{X} \times \mathcal{X}$ with marginal distributions α, β that satisfy:

$$\Pi(\alpha, \beta) \triangleq \{\pi \in \mathcal{P}(\mathcal{X} \times \mathcal{X}) \mid P_{1\#}\pi = \alpha, P_{2\#}\pi = \beta\}. \quad (2)$$

When using the ground cost $c(x, y) = d_{\mathcal{X}}(x, y)^p$ for some $p \geq 1$, we can define the p -Wasserstein distance as:

$$W_p(\alpha, \beta) \triangleq OT(\alpha, \beta)^{1/p}. \quad (3)$$

In practice, it is almost infeasible to obtain the true marginal distributions. Discrete empirical measures $\hat{\alpha} = \sum_{i=1}^m \mathbf{a}_i \delta_{x^i}$ and $\hat{\beta} = \sum_{j=1}^n \mathbf{b}_j \delta_{y^j}$ are usually used instead, where \mathbf{a} and \mathbf{b} are vectors in the probability simplex.

MDS [27] is a dimension reduction method that projects points in Euclidean space into a subspace that best preserves their pairwise squared distances. Recent studies have demonstrated that MDS is also effective for Wasserstein distance [28]. In the case of metric MDS, given a set of high dimensional samples $\mathcal{X} = \{x_n\}_{n=1}^N$ and their distance matrix $D = \{d(x_i, x_j) \mid i, j \in [1, n]\} \in \mathbb{R}^{N \times N}$, the goal is to find an isometrical map $\psi : \mathcal{X} \rightarrow \mathbb{R}^l$ such that:

$$\min_{\psi} \sqrt{\frac{\sum_{i,j} (d(x_i, x_j) - \|\psi(x_i) - \psi(x_j)\|)^2}{\sum_{i,j} d(x_i, x_j)^2}}. \quad (4)$$

Label Embedding We define label-to-label distance as the p -Wasserstein Distance between their corresponding feature representations. Formally, let $N_{\mathcal{D}} := \{x \in \mathcal{X} \mid (x, y) \in \mathcal{D}\}$ be the set of feature representations with label y , and let n_y be its cardinality. Then the distance between label y, y' can be denoted as:

$$\mathcal{L}(y, y') = W_p^p(\alpha_y, \alpha_{y'}). \quad (5)$$

In our experiment, we use $p = 2$. After obtaining the label-to-label distances of all combinations of label pairs, we perform

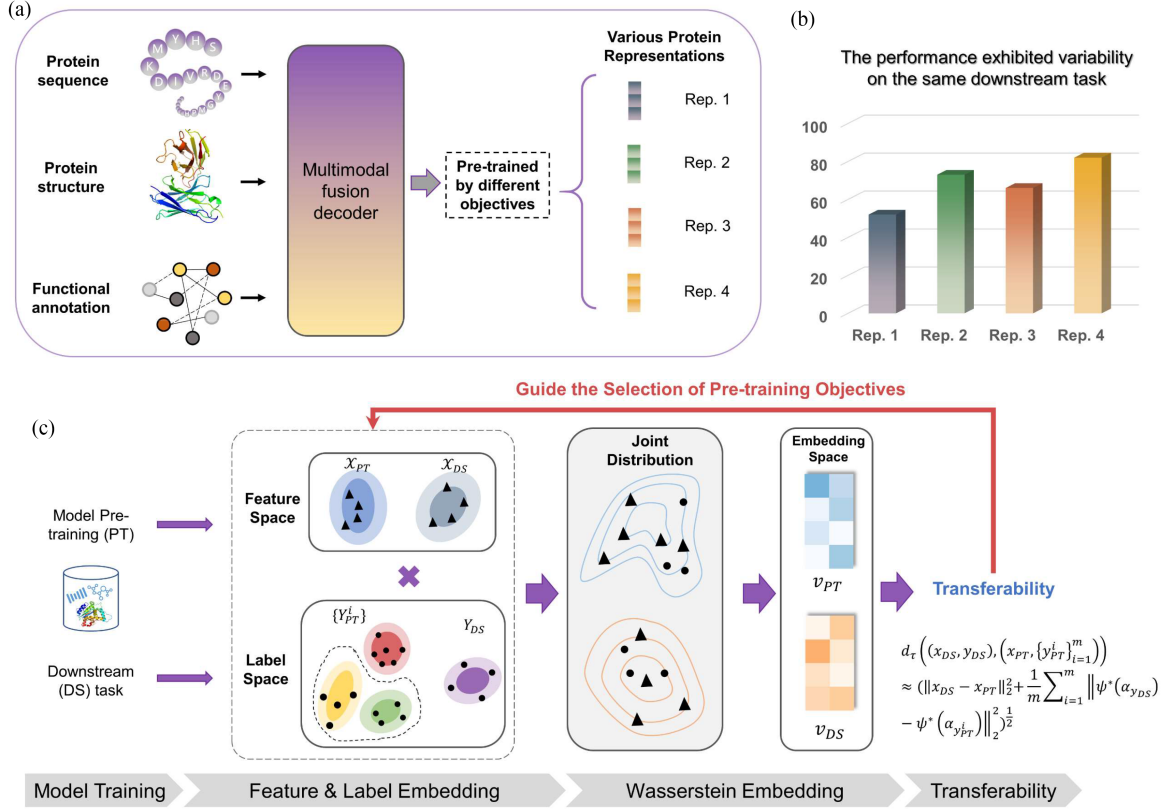


Fig. 1. Quantitative selection of the optimal pre-training model for downstream biological tasks. (a) Incorporation of multi-modal protein data with various pre-training objectives, yielding a variety of pre-trained protein representations. (b) Performance variation of these different pre-trained representations on the same downstream task. (c) Introduction of a novel method for quantifying the transferability from a pre-trained protein representation to downstream tasks, which can aid in the selection of pre-training objectives.

MDS to embed the label distribution into a low dimensional subspace \mathbb{R}^l . Suppose ψ^* is the best map that preserves label information, then we have:

$$\mathcal{L}(y, y') \approx \|\psi^*(\alpha_y) - \psi^*(\alpha_{y'})\|_2^2. \quad (6)$$

For the multi-task pre-training scenario, let $\{Y_{PT}^i\}_{i=1}^m$ be the set of heterogeneous labels of pre-training and let Y_{DS} be the label of a downstream task. Let \mathcal{X}_{PT} , \mathcal{X}_{DS} be the feature representations of pre-training and downstream task respectively. We define the label distance between the multiple pretraining tasks and the downstream task as the average of MDS embedded label-to-label distance:

$$\begin{aligned} \mathcal{L}(y_{DS}, \{y_{PT}^i\}_{i=1}^m) &= \frac{1}{m} \sum_{i=1}^m \mathcal{L}(y_{DS}, y_{PT}^i) \\ &= \frac{1}{m} \sum_{i=1}^m \|\psi^*(\alpha_{y_{DS}}) - \psi^*(\alpha_{y_{PT}^i})\|_2^2 \end{aligned} \quad (7)$$

Considering that:

$$\begin{aligned} d_\tau((x_{DS}, y_{DS}), (x_{PT}, \{y_{PT}^i\}_{i=1}^m)) \\ \approx \left(\|x_{DS} - x_{PT}\|_2^2 + \frac{1}{m} \sum_{i=1}^m \|\psi^*(\alpha_{y_{DS}}) - \psi^*(\alpha_{y_{PT}^i})\|_2^2 \right)^{\frac{1}{2}} \end{aligned}$$

$$= \left\| [x_{DS}, \psi^*(\alpha_{y_{DS}})] - \left[x_{PT}, \frac{1}{m} \sum_{i=1}^m \psi^*(\alpha_{y_{PT}^i}) \right] \right\|_2 \quad (8)$$

Here, $[\cdot, \cdot]$ denotes concatenation operation. We concatenate the feature representation and label embedding together to yield a global embedding over joint distribution $\tau \subseteq \mathbb{R}^{d+l}$, where d, l are the dimensions of feature and label embedding respectively. Note that for the pre-training with multiple tasks, the concatenation operation is performed on the feature embedding and the average of multiple label embeddings.

Barycentric Projection Once we have the global embedding, we utilize barycentric projection to project it to a Wasserstein embedding in a Hilbert subspace, where the Euclidean distance between these embeddings reveals the transferability. Let Z_0 be a fixed reference, then the optimal transport map that projects Z_0 to a global embedding Z_i , i.e., the Monge map, is approximated from the optimal transport plan via [29]:

$$T_i = N(\pi_i^* Z_i) \in \mathbb{R}^{N_0 \times (d+l)}, \quad (9)$$

Here, π_i^* is the optimal transport plan from Z_0 to Z_i , N_0 is the number of samples in the reference. Finally, the Wasserstein embedding for input Z_i can be calculated by:

$$\Phi(Z_i) = \frac{T_i - Z_0}{\sqrt{N}} \in \mathbb{R}^{N_0 \times (d+l)}. \quad (10)$$

TABLE I
DATASETS AND TRANSFORMER-BASED EMBEDDERS USED IN THREE FIELDS

Field	Datasets	Embedder
CV	MNIST, FashionMNIST, EMNIST, KMNIST, USPS	Vision Transformer
NLP	AG News, DBPedia, Yelp Reviews, Amazon Reviews, YAHOO Answers	BERT
Bio	Stability, Fluorescence, Secondary structure, PDBbind, and Kinase set	Our multimodal model

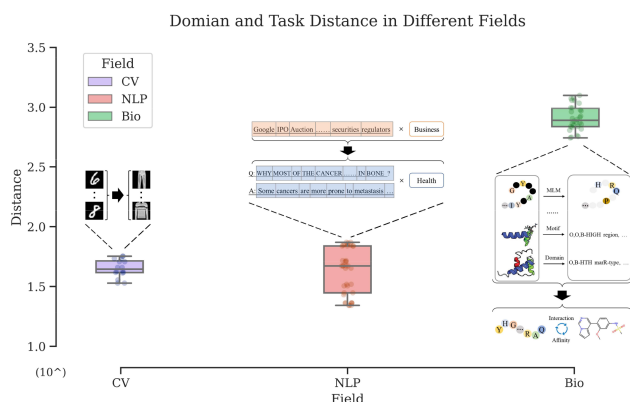


Fig. 2. Comparison of the differences in cross-domain cross-task transfer in the CV, NLP, and computational biology. Three transfer examples from different fields are displayed. In CV, the task involves the classification of handwritten character digits to daily wear items. In NLP, the task is the classification of news articles to question-answering tasks. In computational biology, the task is a multi-task protein pre-training task leading to a protein-drug interaction prediction task.

The transferability between multi-task pre-training and the downstream task can be quantitatively measured by computing the Euclidean distance between their vectors derived from the flattened Wasserstein embeddings.

III. RESULTS

A. Greater Domain and Task Differences Within the Biological Area

To our knowledge, studies on transferability have primarily focused on computer vision (CV) and natural language processing (NLP), which are intuitively simpler to comprehend and visualize. We hypothesized that quantifying transferability in computational biology is more challenging than in other fields, and that current methods may not be optimal. To intuitively visualize this gap, we quantitatively calculated and compared the cross-domain and cross-task distances of these fields. Specifically, we collected multiple cross-domain and cross-task datasets from the CV, NLP, and biological fields. These datasets were uniformly embedded with a field-specific Transformer model to generate field-specific embeddings (Table I). We then calculated the transfer distance of embedding combinations separately for each field. As shown in Fig. 2, the inter-dataset distance in the Bio field is approximately 1.5 orders of magnitude greater than in the

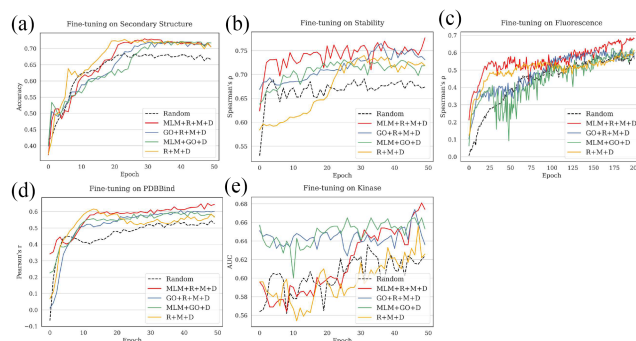


Fig. 3. Transferability of various pre-trained protein representations to downstream tasks. The pre-trained protein representation groups include MLM+RMD, GO+RMD, MLM+GO+D, and RMD. A control group, denoted as ‘random’, employed randomly initialized model weights. These groups were fine-tuned on several biological benchmarks, including (a) secondary structure, (b) stability, (c) fluorescence, (d) PDBbind, and (e) Kinase.

CV and NLP fields. These results validated our hypothesis that quantifying transferability in the biological field is more difficult than in other fields. Clearly, the large amount of redundant information in natural images facilitates the transferability of CV tasks. The semantic and syntactic differences between text sequences in natural language are moderate, whereas protein sequences not only possess heterogeneous semantic information but also carry an abundance of biological evolutionary knowledge, thereby exhibiting the greatest variability.

B. Pretraining and Finetuning of Protein Representations

In this study, we utilized a multi-modal model architecture that integrated information from protein sequence, protein structure, and Gene Ontology. We implemented five pre-training tasks: Masked Language Modeling (MLM), Region (R), Domain (D), Motif (M), and Gene Ontology (G). Leveraging these tasks, we devised a series of combinations and pre-trained a variety of models (i.e., the model structure remained consistent, but the pre-training tasks varied). Each model was subjected to pre-training on an RTX 3090 GPU for 150 epochs, spanning nearly 28 days. This procedure resulted in four unique pre-trained protein representations: MLM+RMD, G+RMD, MLM+G+D, and RMD.

These pre-trained protein representations were then fine-tuned on five downstream tasks, including three protein property benchmarks (Secondary Structure, Stability, and Fluorescence) and two protein-ligand interaction benchmarks (PDBbind and Kinase). Additionally, we established a random control group in which the fine-tuning phase was conducted with model weights initialized randomly. Fig. 3 illustrates the varying performance levels of these pre-trained protein representations on downstream tasks. The protein representation that was randomly initialized consistently showed the lowest performance on all downstream tasks during the fine-tuning processes, indicating that the other protein representations gained advantages from pre-training. The MLM+RMD group outperformed others in many downstream tasks. For example, the MLM+RMD group

TABLE II

EFFICACY COMPARISON BETWEEN FINE-TUNING AND OTMTD WHEN EVALUATING PRETRAINED PROTEIN REPRESENTATIONS ON DOWNSTREAM TASKS

Pre-training models	Fine-tuning time cost (min) (Empirical method)								OTMTD time cost (min) (Quantitative method)							
	SS	ST	FL	PDB	Kinase	RH	SP	FC	SS	ST	FL	PDB	Kinase	RH	SP	FC
MLM+	137	217	205	177	650	887	373	395	2.45	3.25	2.73	2.47	3.80	2.38	3.83	2.45
R+M+D	± 10	± 3	± 15	± 8	± 16	± 21	± 20	± 6	± 0.02	± 0.13	± 0	± 0	± 0	± 0	± 0.02	± 0.10
GO+	146	285	278	176	645	912	406	393	2.03	2.98	2.30	2.07	2.98	2.00	3.04	2.01
R+M+D	± 11	± 2	± 17	± 6	± 9	± 24	± 13	± 6	± 0	± 0.05	± 0	± 0.13	± 0.03	± 0.07	± 0.08	± 0
MLM+	140	215	200	173	646	897	399	389	2.00	2.80	2.23	2.00	2.87	1.96	2.90	2.05
GO+D	± 9	± 3	± 15	± 7	± 20	± 17	± 8	± 5	± 0	± 0.02	± 0.02	± 0.03	± 0	± 0.04	± 0.05	± 0.08
RMD	140	288	252	169	663	891	377	384	1.50	2.22	1.73	1.52	1.85	1.41	1.90	1.58
	± 1	± 3	± 4	± 3	± 9	± 17	± 14	± 3	± 0.02	± 0.13	± 0.02	± 0	± 0.02	± 0.03	± 0.07	± 0

achieved nearly Pearson's $R = 0.35$ before fine-tuning on PDBbind (at epoch 0), compared to the near-zero value of the random group. After fine-tuning, these values increased to nearly 0.65 and 0.51, respectively (at epoch 150). These findings suggest that the knowledge gained during the corresponding pre-training was effectively transferred to the downstream tasks.

C. Efficiency of OTMTD Compared to Fine-Tuning Evaluation

To quantitatively compare the efficiency of our proposed quantitative method with the empirical approach for assessing the transferability of pretraining to downstream tasks, we recorded the computation time for both methods. We conducted each method's experiments 5 times using 5 different random seeds and calculated the mean and standard deviation of the time cost. As detailed in Table II, the quantitative method demonstrates significantly higher efficiency compared to the empirical method. Specifically, the average time costs for the five downstream tasks under the four pretraining models are 2.00, 2.81, 2.25, 2.01, 2.88, 1.94, 2.92, and 2.02 minutes when using the quantitative method. In contrast, the corresponding time costs using the empirical method are substantially higher, ranging from 140.75 to 897 minutes. Notably, for data-intensive downstream tasks like Kinase, where fine-tuning often requires considerably more time, the quantitative OTMTD's time costs are remarkably close to those of the stability task, despite the former's data volume being over three times larger. Overall, these results underscore the superior efficiency of the quantitative method, highlighting its practical value in evaluating transferability and its potential to efficiently guide the selection of optimal pretraining tasks.

D. High Accuracy of OTMTD on Biological Benchmarks

The relative change in accuracy or correlation coefficient on the test set served as a proxy for empirical transferability from multi-task pre-training to the downstream task. We repeated the fine-tuning experiments five times using distinct random seeds. We analytically computed the OTMTD scores for each pair of pre-trained representations and downstream tasks, yielding a total of 20 points in Fig. 4. The performance of the proposed OTMTD was evaluated using the Spearman correlation coefficient. The horizontal axis in Fig. 4 represents the transfer distance (OTMTD) between the pre-training representation and the

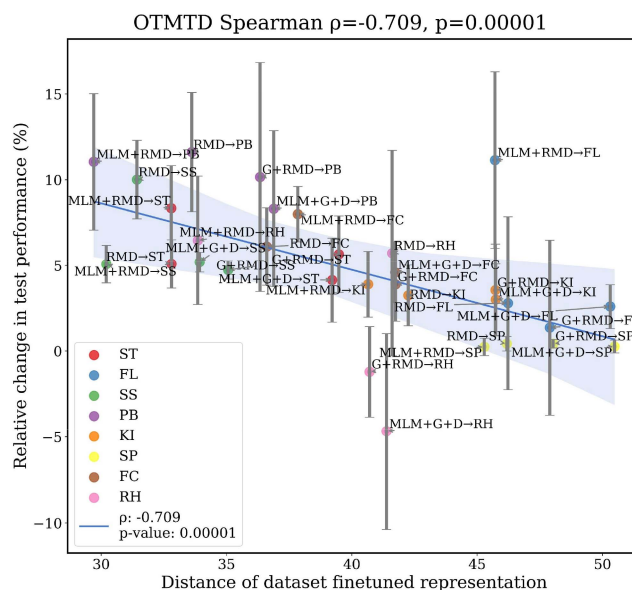


Fig. 4. Performance of the proposed OTMTD in relation to various biological benchmarks. The horizontal axis represents the transfer distance (OTMTD) between the pre-training representation and the downstream task (e.g., MLM+RMD to FL). The vertical axis represents the corresponding actual performance. A higher Spearman's ρ indicates a stronger correlation between predicted transferability and actual performance. The pre-training objectives include MLM (masked amino acid), G (masked Gene Ontology), and RMD (domain/motif/region placement capture). The biological benchmarks include ST (stability), FL (fluorescence), SS (secondary structure), PB (PDBbind), KI (Kinase), SP (signal peptide), FC (fold classes), and RH (remote homology).

downstream task (e.g., MLM+RMD to FL), while the vertical axis represents the corresponding actual performance (relative change of performance from randomly initialized to pre-trained representations).

A strong and statistically significant correlation of -0.709 Spearman's ρ exists between OTMTD and empirical test results, indicating that OTMTD is highly predictive of transferability between heterogeneous biological multi-task pre-training representation and downstream task. By utilizing this method, we were able to efficiently compute the transfer distance and select the optimal pre-trained protein representation for a specific downstream task. This approach is more efficient than comparing all models after fine-tuning on all downstream tasks. For example, the transfer distance of MLM+RMD was predicted to

As we confront increasingly complex downstream tasks, the application of transferability methods becomes crucial. The model's pre-training is optimized towards the pre-training task, but there may be significant heterogeneity in the distribution between the pre-training task and the downstream application task. This heterogeneity could introduce bias and impact the effectiveness of the transfer from pre-training to downstream tasks. There is a substantial gap between the pre-training and fine-tuning processes. Furthermore, the distributions of various downstream tasks may significantly differ from the commonly used downstream benchmark datasets for evaluating pre-trained models, making it difficult to select the most suitable pre-training model based on known results. In such cases, running all models would undoubtedly be inefficient. Therefore, using transferability quantitative methods to quickly screen for the optimal model would be a more efficient choice.

In response to this challenge, we have introduced a novel transferability-based method in this study. This method can be used to compute the transfer distance and select the optimal pre-trained protein representation for a specific downstream task, which is more efficient than comparing all models after fine-tuning on all downstream tasks. It addresses a critical gap in the current understanding of how to effectively transfer pre-trained protein representations to downstream tasks. Our OTMTD is unique in its ability to predict the transfer performance from a pre-trained protein representation to downstream tasks by assessing similarities between pre-training and downstream features across domains and multiple heterogeneous tasks. This approach is a significant departure from traditional methods, which often struggle to accurately predict transfer performance due to the complexity and heterogeneity of protein data.

Our approach outperforms other transferability methods such as H-Score, Wasserstein Distance, OTDD, and OTCE. These methods primarily focus on image and text classification, where the domain and task differences are much smaller than in the biological area. Moreover, these methods are designed for simple classification tasks and are unsuitable for more complex and heterogeneous tasks in the biological area. Our method overcomes these limitations by being able to handle multiple heterogeneous tasks, which are extremely common in the biological area. In addition to its effectiveness in protein representation learning, our method exhibits a promising adaptability to other domains within computational biology and bioinformatics. The unique challenges in these fields, such as the analysis of genomic sequences, structural bioinformatics, or systems biology, often involve complex, multi-dimensional data that require sophisticated interpretation. By leveraging the inherent flexibility of our approach, which accommodates the complex nature of biological data, we foresee its applicability extending to these diverse areas. This adaptability not only underscores the versatility of our method but also opens avenues for its application in a wider range of bioinformatics tasks, thus providing a comprehensive tool for assessing transferability across various subfields of computational biology.

In conclusion, our study represents a significant advancement in the field of computational biology and bioinformatics. Our

transferability-based approach enables efficient prediction and assessment of protein pretrained representation performance in downstream tasks, saving substantial fine-tuning time across diverse tasks. This efficiency enhances the evaluation of protein representation learning models and provides a more effective quantitative method for their design. We are excited to see the future applications and developments of our approach. This work sets a new standard for the evaluation and application of pre-trained protein representations, and we anticipate that it will inspire further innovations in this rapidly evolving field.

REFERENCES

- [1] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013, doi: [10.1109/TPAMI.2013.50](https://doi.org/10.1109/TPAMI.2013.50).
- [2] A. Rives et al., "Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences," *Proc. Nat. Acad. Sci.*, vol. 118, no. 15, Apr. 2021, Art. no. e2016239118, doi: [10.1073/pnas.2016239118](https://doi.org/10.1073/pnas.2016239118).
- [3] R. Rao et al., "Evaluating protein transfer learning with TAPE," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 9689–9701, doi: [10.1101/676825](https://doi.org/10.1101/676825).
- [4] R. M. Rao et al., "MSA transformer," in *Proc. Int. Conf. Mach. Learn.*, 2021, vol. 139, pp. 8844–8856, doi: [10.1101/2021.02.12.430858](https://doi.org/10.1101/2021.02.12.430858).
- [5] S. Unsal, H. Atas, M. Albayrak, K. Turhan, A. C. Acar, and T. Doğan, "Learning functional properties of proteins with language models," *Nature Mach. Intell.*, vol. 4, no. 3, pp. 227–245, Mar. 2022, doi: [10.1038/s42256-022-00457-9](https://doi.org/10.1038/s42256-022-00457-9).
- [6] F. Hu, Y. Hu, W. Zhang, H. Huang, Y. Pan, and P. Yin, "A multimodal protein representation framework for quantifying transferability across biochemical downstream tasks," *Adv. Sci.*, vol. 10, no. 22, pp. 1–14, Aug. 2023, doi: [10.1002/adv.202301223](https://doi.org/10.1002/adv.202301223).
- [7] Y. Zhang et al., "Computer-aided diagnosis of complications after liver transplantation based on transfer learning," *Interdiscipl. Sci.: Comput. Life Sci.*, vol. 16, pp. 123–140, 2024, doi: [10.1007/s12539-023-00588-6](https://doi.org/10.1007/s12539-023-00588-6).
- [8] D. Zhou, S. Peng, D.-Q. Wei, W. Zhong, Y. Dou, and X. Xie, "LUNAR: Drug screening for novel coronavirus based on representation learning graph convolutional network," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 18, no. 4, pp. 1290–1298, Jul./Aug. 2021, doi: [10.1109/TCBB.2021.3085972](https://doi.org/10.1109/TCBB.2021.3085972).
- [9] S. Lin, G. Zhang, D.-Q. Wei, and Y. Xiong, "DeepPSE: Prediction of polypharmacy side effects by fusing deep representation of drug pairs and attention mechanism," *Comput. Biol. Med.*, vol. 149, Oct. 2022, Art. no. 105984, doi: [10.1016/j.combiomed.2022.105984](https://doi.org/10.1016/j.combiomed.2022.105984).
- [10] N. S. Detlefsen, S. Hauberg, and W. Boomsma, "Learning meaningful representations of protein sequences," *Nature Commun.*, vol. 13, no. 1, Dec. 2022, Art. no. 1914, doi: [10.1038/s41467-022-29443-w](https://doi.org/10.1038/s41467-022-29443-w).
- [11] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, vol. 30, pp. 1–14.
- [12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, Oct. 2018, vol. 1, pp. 4171–4186, doi: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423).
- [13] V. Gligorijević et al., "Structure-based protein function prediction using graph convolutional networks," *Nature Commun.*, vol. 12, no. 1, Dec. 2021, Art. no. 3168, doi: [10.1038/s41467-021-23303-9](https://doi.org/10.1038/s41467-021-23303-9).
- [14] T. Beppler and B. Berger, "Learning protein sequence embeddings using information from structure," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–17.
- [15] N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, and M. Linial, "ProteinBERT: A universal deep-learning model of protein sequence and function," *Bioinformatics*, vol. 38, no. 8, pp. 2102–2110, Apr. 2022, doi: [10.1093/bioinformatics/btac020](https://doi.org/10.1093/bioinformatics/btac020).
- [16] N. Zhang et al., "OntoProtein: Protein pretraining with gene ontology embedding," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–18.
- [17] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3712–3722.
- [18] A. Achille et al., "Task2vec: Task embedding for meta-learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 6430–6439.

- [19] Y. Bao et al., "An information-theoretic approach to transferability in task transfer learning," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 2309–2313, doi: [10.1109/ICIP.2019.8803726](https://doi.org/10.1109/ICIP.2019.8803726).
- [20] A. T. Tran, C. V. Nguyen, and T. Hassner, "Transferability and hardness of supervised classification tasks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1395–1405, doi: [10.1109/ICCV.2019.00148](https://doi.org/10.1109/ICCV.2019.00148).
- [21] C. Nguyen, T. Hassner, M. Seeger, and C. Archambeau, "Leep: A new measure to evaluate transferability of learned representations," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 7294–7305.
- [22] X. Liu, Y. Bai, Y. Lu, A. Soltoggio, and S. Kolouri, "Wasserstein task embedding for measuring task similarities," 2022, *arXiv:2208.11726*.
- [23] Y. Tan, Y. Li, and S.-L. Huang, "OTCE: A transferability metric for cross-domain cross-task representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15774–15783, doi: [10.1109/CVPR46437.2021.01552](https://doi.org/10.1109/CVPR46437.2021.01552).
- [24] Y. Tan, Y. Li, S.-L. Huang, and X.-P. Zhang, "Transferability-guided cross-domain cross-task transfer learning," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Feb. 05, 2024, doi: [10.1109/TNNLS.2024.3358094](https://doi.org/10.1109/TNNLS.2024.3358094).
- [25] R. Wang, X. Fang, Y. Lu, and S. Wang, "The PDBbind database: Collection of binding affinities for protein–ligand complexes with known three-dimensional structures," *J. Med. Chem.*, vol. 47, no. 12, pp. 2977–2980, Jun. 2004, doi: [10.1021/jm030580l](https://doi.org/10.1021/jm030580l).
- [26] L. Chen et al., "TransformerCPI: Improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments," *Bioinformatics*, vol. 36, no. 16, pp. 4406–4414, Aug. 2020, doi: [10.1093/bioinformatics/btaa524](https://doi.org/10.1093/bioinformatics/btaa524).
- [27] M. A. A. Cox and T. F. Cox, "Multidimensional scaling," in *Handbook of Data Visualization*. Berlin, Germany: Springer, 2008, pp. 315–347.
- [28] K. Hamm, N. Henscheid, and S. Kang, "Wassmap: Wasserstein isometric mapping for image manifold learning," *SIAM J. Math. Data Sci.*, vol. 5, pp. 475–501, Apr. 2023.
- [29] S. Kolouri, N. Naderializadeh, G. K. Rohde, and H. Hoffmann, "Wasserstein embedding for graph learning," in *Proc. Int. Conf. Learn. Representations*, 2021, pp. 1–20.
- [30] C. Frogner, F. Mirzazadeh, and J. Solomon, "Learning embeddings into entropic wasserstein spaces," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–12.
- [31] D. Alvarez-Melis and N. Fusi, "Geometric dataset distances via optimal transport," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 21428–2143.
- [32] J. Jumper et al., "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021, doi: [10.1038/s41586-021-03819-2](https://doi.org/10.1038/s41586-021-03819-2).