bioRxiv preprint doi: https://doi.org/10.1101/2020.12.14.422772; this version posted December 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Transfer Learning with MotifTransformers for Predicting Protein-Protein Interactions Between a Novel Virus and Humans

Jack Lanchantin, Arshdeep Sekhon, Clint Miller, Yanjun Qi
Department of Computer Science
University of Virginia
{jjl5sw,clintm,yq2h}@virginia.edu

Abstract

The novel coronavirus SARS-CoV-2, which causes Coronavirus disease 2019 (COVID-19), is a significant threat to worldwide public health. Viruses such as SARS-CoV-2 infect the human body by forming interactions between virus proteins and human proteins that compromise normal human protein-protein interactions (PPI). Current in vivo methods to identify PPIs between a novel virus and humans are slow, costly, and difficult to cover the vast interaction space. We propose a novel deep learning architecture designed for in silico PPI prediction and a transfer learning approach to predict interactions between novel virus proteins and human proteins. We show that our approach outperforms the state-of-the-art methods significantly in predicting Virus–Human protein interactions for SARS-CoV-2, H1N1, and Ebola.

1 Introduction

Proteins are essential biomolecules that control diverse patho–physiological functions in human cells. A protein-protein interaction (PPI) denotes a critical process where two proteins come in contact with each other to carry out specific biological functions. PPIs between two human proteins are subject to direct or indirect attack by viral proteins, such as those from the 2019 novel coronavirus, also known as SARS-CoV-2. The interactions between virus proteins and human proteins (V–H interactions) are important for viruses to infect the human body, and ultimately overtake physiological functions (e.g., alveolar gas exchange). Accordingly, protein-protein interactions are often the subject of intense research by virologists and pharmaceutical scientists. Knowing and understanding which human proteins a virus may interact with is crucial for preventing viral infection.

Modern sequencing methods have made it cheap and fast to determine protein sequence information, yet it remains difficult to accurately uncover the full set of protein-protein interactions. Traditionally, PPIs have been studied individually

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.14.422772; this version posted December 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

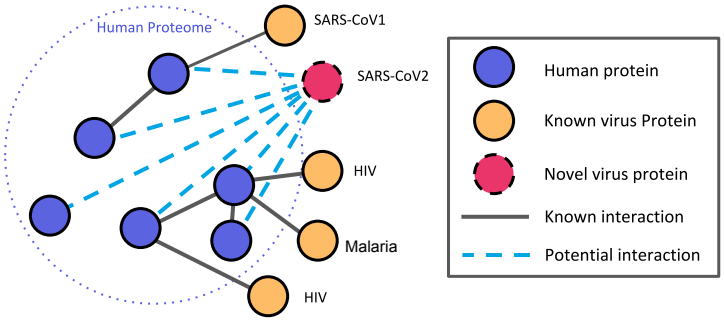

Figure 1: Protein-Protein Interactions (PPI). Overview of our task, where there is a set of previously known protein-protein interactions. Our goal is to predict all possible Virus–Human interactions for a novel virus protein, such as SARS-CoV-2.

through the use of genetic, biochemical, and biophysical techniques such as measuring natural affinity of binding partners in-vivo or in-vitro [44]. While accurate, these small-scale experiments have low sensitivity and are not suitable for full proteome analyses [69]. If there are $|P_h|\approx$20,000 total number of human proteins, and $|P_v|\approx$26 total virus proteins (in the case of SARS-CoV-2), then the potential resulting search space of V–H interactions is $|P_v|\times|P_h|=$ 0.5M.

High-throughput technologies, such as yeast two-hybrid screens (Y2H) [17] and Affinity-purification–mass spectrometry (AP-MS) [25, 19], are chiefly responsible for the relatively large amount of PPI evidence. Notably, the first experimental study for SARS-CoV-2 interactions used AP-MS [19]. However, these datasets are often incomplete, noisy, and hard to reproduce [36]. The resulting low sensitivity of high-throughput experiments is unfavorable when trying to fully understand how the virus interacts with humans. Additionally, these methods are expensive and somewhat time consuming. Fast and accurate computational methods are urgently needed to identify PPIs, especially in time-sensitive scenarios such as novel virus understanding where each day is crucial.

In this paper, we propose a scalable computational model to predict-protein interactions between a novel virus and humans using only sequence information. Motivated by the evidence that co-occurring short polypeptide sequences between interacting protein partners appear to be conserved across different organisms [45], we introduce a novel architecture to learn such short sequences, or "protein motifs". Additionally, we use a transfer learning approach to help generalize to unseen sequences. Our framework tests all possible interactions between a novel virus protein and human proteins. This setup is beneficial for three main reasons. First, our model can predict an initial set of interactions if experiments have not yet been done. Second, our model can expand the initial set of experimental interactions, resulting in a more complete interactome. Finally, such computational models enable us to easily test hypotheses such as the effect of mutations.

Most previous computational methods to predict PPIs have focused on within-

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.14.422772; this version posted December 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

species interactions [60, 6, 66, 18, 45, 35, 20, 68, 21]. These methods do not easily generalize to cross-species interactions (e.g., V–H) [67]. Few methods have attempted to predict cross-species protein interactions between humans and a novel virus [70, 67]. We argue that this is a much more realistic task when an unknown virus is discovered.

When using computational methods to predict V–H PPIs, two challenges stand out: (1) Existing sequence analysis tools focus on global alignment patterns while PPIs mostly depend on local binding motif patterns. (2) It is especially difficult for a novel virus since there is limited or no interaction data for those sequences. This requires the model to transfer knowledge from one domain (previously known sequences) to a new domain (novel virus sequences).

Recent work shows the transferability of large scale models on protein sequences for many downstream tasks [50, 55]. In this work, we propose a deep learning based pipeline to combine neural representation learning and transfer learning for solving the listed obstacles. The main contributions of our paper are:

1. We introduce the MotifTransformer, a novel deep neural architecture for protein sequence representation learning with a focus on sequence motifs. Our model is scalable, accurate, and flexible.
2. We present a transfer learning framework for PPI prediction in the context of a novel virus where no interactions are known.
3. We evaluate our predictions with validated interactions from three Virus–Human PPI datasets: our curated SARS-CoV-2 dataset based on interactions from BioGRID, as well as H1N1 and Ebola datasets.

Our goal is to expedite our understanding of virus mechanisms and to aid in the development of vaccines, diagnostics, therapeutics, and antibodies against a novel virus such as SARS-CoV-2. Predicting Virus–Human (V–H) interactions may enable a better biological understanding of the disease, and lead to the design of a drug target that blocks such a process from occurring. Thus, elucidating V–H protein interactions on a large-scale is a crucial first step for antiviral drug development.

2 Protein Sequences and the Protein-Protein Interaction Prediction Task.

Proteins are biomolecules in three different forms. The sequence form of proteins is a linear chain of amino acids (AAs). These consist of 20 possible standard AAs, two non-standard AAs: selenocysteine and pyrrolysine, two ambiguous AAs, and one unknown. In other words, proteins are strings built from a dictionary V of size ($|V|$=25). We represent a protein $\mathbf{x}$ as a sequence of characters x_1,x_2,...,x_L. Each character x_i is one possible amino acid from V.

Proteins rarely act in isolation but instead interact with other proteins to perform many biological processes (e.g., protein-ligand binding to its receptor to induce specific signaling pathways). This is referred to as a protein-protein interaction (PPI). Viruses infect the human body by forming Virus–Human

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.14.422772; this version posted December 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

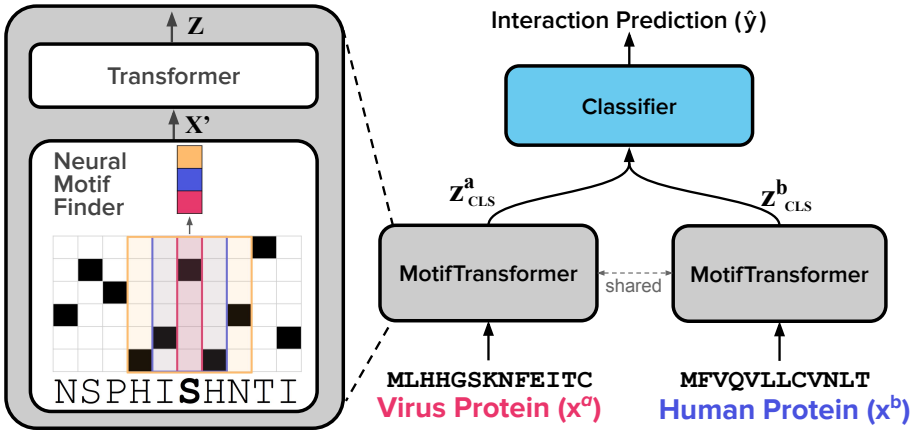

Figure 2: (left) Example of the motif finder in our MotifTransformer model. This example shows 3 convolutional filters of sizes 1x1 (red), 3x1 (blue), and 5x1 (orange) and demonstrates filter outputs for bolded amino acid **S**. This is repeated across all amino acids, which are used as the input to the Transformer. (right) Our MotifTransformer applied to protein-protein interaction prediction.

PPIs. Predicting which human proteins a virus protein will bind to is a key step in understanding viral pathogenesis[1] and designing viral therapies. This interaction prediction results in a binary classification problem: "given virus protein sequence $\mathbf{x}^q$ and human protein sequence $\mathbf{x}^k$, does the pair interact or not?".

Fig. 1 shows a visual representation of all types of PPIs that occur within the human body. There are three types of proteins in this diagram: Human proteins, previously known virus proteins, and a novel virus protein. As shown with solid lines, there are a set of known interactions between a Human and Human protein as well as between a known virus and Human proteins. Our target task is to predict all possible unknown sets of interactions between the novel virus and human proteins, as shown with a dashed line.

3 Proposed Model Architecture for PPI Prediction: MotifTransformer

Transformers [63] have obtained state-of-the-art results in many domains such as natural language [15], images [49], and protein sequences [55]. At a high level, Transformer encoder models "transform" the representations of input tokens (e.g., amino acids) by modeling dependencies between them in the form of attention. The importance, or weight, of token $\boldsymbol{x}_j$ with respect to $\boldsymbol{x}_i$ is learned through attention. The attention weight, α^t_{ij} between token i and j is computed as follows. First, we compute a normalized scalar attention coefficient α_{ij} between tokens i and j. Once we compute the attention weights between each i and j, we

[1] Mechanisms by which virus infection leads to disease in the target host

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.14.422772; this version posted December 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

update each $\boldsymbol{x}_i$ to $\hat{\boldsymbol{x}}_i$ using a weighted sum of all tokens followed by a nonlinear ReLU layer:

$$\alpha_{ij}=\text{softmax}\big((\mathbf{W}^a\boldsymbol{x}_i)^\top(\mathbf{W}^b\boldsymbol{x}_j)/\sqrt{d}\big) \tag{1}$$

$$\bar{\boldsymbol{x}}_i=\sum_{j=1}^{M}\alpha_{ij}\mathbf{W}^v\boldsymbol{x}_j \tag{2}$$

$$\hat{\boldsymbol{x}}_i=\text{ReLU}(\bar{\boldsymbol{x}}_i\mathbf{W}^r+\boldsymbol{b}_1)\mathbf{W}^o. \tag{3}$$

This update procedure can be repeated for ℓ layers where the updated vectors $\hat{\boldsymbol{x}}_i$ are fed as input to the successive layer. The learned weight matrices $\{\mathbf{W}^b,\mathbf{W}^a,\mathbf{W}^v,\mathbf{W}^r,\mathbf{W}^o\}\in\mathbb{R}^{d\times d}$ are not shared between layers. We represent the final output of the transformer encoder after ℓ layers as $\mathbf{Z}$. We refer the reader to [63, 56] for a detailed summary of the Transformer architecture, as we use the standard Transformer encoder modules.

The vanilla Transformer model for protein sequences uses amino acid characters as input tokens. In this formulation, amino acids $x_1,x_2,...,x_L$ are represented as token vectors $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_L)$ (one-hot or embedding vectors) and fed to the attention modules of the Transformer [55]. However, protein sequences have short, local features known as sequence motifs, which are important for many protein functions [52]. If we view amino acids as the protein analog of natural language characters, motifs are analogous to words. In particular, motifs are crucial for virus proteins to mimic human proteins and interact with other human proteins [14]. In order to better represent protein sequences, we introduce the MotifTransformer architecture. The key contribution of our model is a differentiable module for automatically learning sequence motifs, which we call the Neural Motif Finder. This module utilizes different length convolutional filters to find motifs directly from sequence end-to-end.

Specifically, given the protein character sequence as a one-hot matrix $\boldsymbol{X}\in\mathbb{R}^{\in L\times|V|}$, we apply six temporal convolutional filters of sizes $F=\{(1\times128), (3\times256), (5\times384), (7\times512), (9\times512), (11\times512)\}$, where the first number of each filter is the width and the second number is the token dimension. Each filter is zero-padded to preserve the original sequence length. We concatenate the output of the convolutional filters at each position to create a vector of size 2304 for each position. Then, it is fed to an MLP. This builds a $L\times d$ matrix, $\mathbf{X}'$. Finally, to encode positional information we add sinusoidal tokens [63] to the $\mathbf{X}'$ matrix, to be used as input to a Transformer encoder.

Using several convolutional filters of varying size allows the model to learn a diverse set of motifs. Specifically, in our implementation, the set of filters F allows the model to learn 2304 unique motifs of varying lengths. The Neural Motif Finder is illustrated in Fig. 2 (left). We then use a Transformer encoder that takes in matrix $\mathbf{X}'\in\mathbb{R}^{L\times d}$ and outputs $\mathbf{Z}\in\mathbb{R}^{L\times d}$ after ℓ Transformer layers.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.14.422772; this version posted December 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Dataset	Category	Output Shape	|Total|	|Train|	|Valid|	|Test|
Swiss-Prot	UPT	$L \times \|V\|$	562,280	562,280	N/A	N/A
Secondary Structure	SPT	$L \times 3$	11,361	8,678	2,170	513
Contact	SPT	$L \times L$	25,563	25,299	224	40
Homology	SPT	1195×1	13,766	12,312	736	718
SARS-CoV-2	PPI	1×1	815,279	199,346	49,836	610,950
H1N1 [70]	PPI	1×1	22,291	21,910	N/A	381
Ebola [70]	PPI	1×1	22,982	22,682	N/A	300

Table 1: Datasets: For each category of training: Language Model (UPT), Intermediate (SPT) and PPI, we provide the dataset output type and training/validation/test set sizes. L represents the sequence length, and $|V|$ represents the vocabulary size.

4 Proposed Training Formulation: Transfer Learning for Virus–Human PPI Prediction

Our target goal is to predict the interaction likelihood of a novel virus protein with a human protein. For a novel virus such as SARS-CoV-2, few or no interactions are known, making it difficult to train on labeled data. This requires a model that can generalize from other domains and tasks with larger amounts of labeled data. We introduce a two-step procedure for PPI prediction with a novel virus. The first step is to pretrain a base model to learn generic representations of protein sequences, and the second step is to finetune the model on V–H PPI data for previously known viruses. Pretraining the model allows us to learn representations that transfer well to the PPI task of a novel (i.e., unseen) virus.

4.1 Pretraining the MotifTransformer on Unsupervised and Supervised Tasks

Recent literature on learning self-supervised distributed representations of natural language have shown that pretraining using self-supervised and supervised methods encourage the model to learn semantics about the input domain that can help prediction accuracy on new tasks [7, 37, 11, 42, 15]. Notably, pretraining can help in transfer learning, where the model is given labeled training samples in one domain, and is required to predict samples in a new domain [58]. We propose several pretraining tasks for learning protein representations that generalize to our target Human-Virus PPI task.

To learn general $\mathbf{Z}$ representations of proteins that can be used for the V-H interaction task, we pretrain the MotifTransformer using two types of tasks: unsupervised pretraining (UPT) and supervised pretraining (SPT). First, we train the MotifTransformer on a large dataset of unlabeled protein sequences in an unsupervised pretraining (UPT) manner. We adopt the RoBERTa masked language model method [34]. Second, to improve the unsupervised protein representations, we further pretrain the model on a set of supervised structure and function prediction problems, which we denote SPT. We consider three tasks: (1) Secondary Structure (SS) prediction: a multi-class amino acid tagging task, (2) Contact prediction: a pairwise amino acid binary classification task, (3) Homology prediction: a multi-class protein classification task. In summary,

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.14.422772; this version posted December 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Method	SS	Contact	Homology
One-hot [50]	0.69	0.29	0.09
Alignment [50]	**0.80**	0.64	0.09
ResNet [50]	0.70	0.20	0.10
LSTM [50]	0.71	0.19	0.12
Transformer [50]	0.70	0.32	0.09
MotifTransformer	0.70	0.51	0.12
MotifTransformer + UPT	0.71	0.58	0.22
MotifTransformer (multi-task)	0.64	0.53	0.13
MotifTransformer + UPT (multi-task)	0.71	**0.70**	**0.38**

Table 2: Supervised pretraining task (SPT) results. For SS and Homology, accuracy is reported. For Contact, precision at $L/5$ for for medium and long-range contacts is reported.

the pretraining tasks allow the model learn good $\mathbf{Z}$ representations that we can finetune on the known V-H interaction data, as explained in the following subsection. We hypothesize that the representations learned will transfer across different domains when we test on novel virus sequences.

4.2 Finetuning the MotifTransformer on the Protein-Protein Interaction Task

After pretraining, the final task is to finetune the model on the known (i.e. experimentally validated) PPI data. Specifically, given protein a and protein b, we seek to estimate the probability that a and b interact. To obtain a single vector representation of each protein which will be used for the PPI task, we use a designated classification token, CLS from the MotifTransformer. We use the final CLS token as the vector representation for each protein. In other words, we use $\mathbf{z}^a_{\text{CLS}}$ to represent query protein $\mathbf{x}^a$, and $\mathbf{z}^b_{\text{CLS}}$ to represent key protein $\mathbf{x}^b$.

Using these representations, we predict $\hat{y}$, the likelihood of the two proteins interacting using an order-independent classifier inspired by [61]:

$$\mathbf{v}=[|\mathbf{z}_q-\mathbf{z}_k|;\mathbf{z}_q\odot\mathbf{z}_k] \tag{4}$$

$$\hat{y}=\mathbf{w}\big(\text{GELU}(\mathbf{W}\mathbf{v}+b)\big), \tag{5}$$

where $\mathbf{W}\in\mathbb{R}^{2d\times 2d}$, and $\mathbf{w}\in\mathbb{R}^{1\times 2d}$ are projection matrices, $\mathbf{v}\in\mathbb{R}^{2d\times 1}$. $\hat{y}$ is then fed through a sigmoid function to obtain the interaction probability.

Once this model is trained, we can test it on pairs of Virus–Human interaction sequences where the virus was not used during training. The pretraining method on generic tasks learn structures of proteins which generalize well to unseen proteins.

5 Related Work

Protein-Protein Interaction Prediction. Many previous PPI works focus on developing intra-species interactions [60, 64, 22, 54, 28, 9]. In other words, they would have one model for only Human–Human interactions and another

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.14.422772; this version posted December 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

model for only Yeast–Yeast interactions. Cross-species interaction prediction instead relates to where each protein in the interaction comes from a different species. Many works predict cross-species PPIs where the testing set contains proteins that are in the training set [62, 47, 13, 39, 5]. These methods do not reflect the real-world setting for a novel virus since we don't have training proteins available for the virus. Additionally, PPI prediction methods generally perform much better for test pairs that share components with a training set than for those that do not [41].

Few works have focused on the more difficult task of cross-species interaction prediction where one of the protein species is completely unseen during training, which is what our work tackles. DeNovo [16] used an SVM for cross species interaction prediction. Yang et al. [67] introduce a deep learning embedding method combined with a random forest. Zhou et al. [70] improved DeNovo's SVM for novel Virus–Human interaction.

Protein Sequence Classification. Machine learning methods have achieved considerable results predicting properties of proteins that have yet to be experimentally validated by experimental studies. [46, 33] introduce multitask deep learning models for sequence labeling tasks such as SS prediction. [55, 50, 38] focus on methods of language model pretraining for generalizable representations of sequences. In particular, [50, 55] and [8] showed that self-supervised pretraining can produce protein representations that generalize across protein domains.

Transformers. Transformers [63] obtained state-of-the-art results on several NLP tasks [15]. One problem with the vanilla Transformer model on token level inputs is that locality is not preserved. [4] used varying convolutional filters on characters at the word level and took the mean of the output to get a single vector representation for each word. Since proteins have no inherent "words" we use the convolutional output for each character as its local word. Instead of using character level inputs, word or byte-pair encodings can be used in order to preserve the local structure of words in text [59].

Transfer Learning Our work relates to several others in natural language processing where pretraining is used to transfer knowledge from both unlabeled and related labeled datasets [3, 32, 43]. Transfer learning is closely tied with few-shot learning [51, 31], which typically aims to use representations from prior tasks to generalize. Transformers are particularly well-fitted for transfer learning as their parallelizable architecture allows for fast pretraining on large datasets [15, 48]. It has been shown that this large-scale pretraining generalizes well enough for accurate few-shot learning [10].

6 Experimental Setup and Results

6.1 Model Details and Evaluation Metrics

MotifTransformer Variations and Details We evaluate three variants of our model: (1) MotifTransformer: this is the base model which uses no pretraining, only training on the target PPI task. (2) MotifTransformer+UPT: this

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.14.422772; this version posted December 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Method	AUROC	AUPR	F1(%)	P@100
Embedding+RF [67]	0.748	0.071	**0.116**	0.126
MotifTransformer	0.740	0.065	0.104	0.129
MotifTransformer+UPT	0.751	0.070	0.110	0.147
MotifTransformer+UPT+SPT	**0.753**	**0.076**	0.114	**0.151**

Table 3: Human and SARS-CoV-2 Interaction Predictions. Each metrics is reported as the mean across all virus proteins. Best results are reported in bold.

variant uses the language model pretraining and finetuning on the PPI task. (3) MotifTransformer+UPT+SPT: this uses both language model pretraining and supervised structure/family prediction pretraining before finetuning on the PPI task. We test both single task and multi-task models on the 3 SPT tasks. We use the multi-task trained model for all PPI tasks. We train all models using a 12-layer transformer with 8 attention heads and GELU activations [24].

For the language model and intermediate tasks, we clip all sequences to a maximum length of 1024. For the protein interaction task, we clip sequences to length 1600. For language model pretraining, we use a batch size of 1024, a linear warmup, and max learning rate of 1e-3. For all other tasks, we use a batch size of 16 with max learning rate of 1e-5. The language model is trained for 60 epochs, and all others are trained for 100. All models are trained with an Adam optimizer [29] and 10% dropout. Our models are implemented in PyTorch and we run each model on 4 NVIDIA Titan X GPUs. Language model pretraining (UPT) takes approximately 3 days, SPT pretraining takes 1 day, and the PPI task takes 3 days. Testing on $\sim$0.5M PPI pairs takes about 1 hour.

Metrics. For the supervised pretraining tasks, we use the metrics reported by previous work [50]. For the PPI task, we are largely focused on ranking interaction predictions based on probability, so we report two non-thresholding metrics: area under the ROC curve (AUROC), and area under the precision-recall curve (AUPR). We additionally report F1 scores where we consider thresholds [0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9]. As done in previous works, the results are selected on the best performing test epoch for each metric. For the SARS-CoV-2 dataset, we evaluate each metric for each virus protein individually since we are interested in the accuracy of predicting human interactions for specific virus proteins. The reported results are the mean value across all 25 virus proteins. For this dataset, we also report precision at 100 (P@100).

6.2 Pretraining Tasks

Datasets. For the language model (LM) task, we train MotifTransformer on all Swiss-Prot protein sequences [12]. Swiss-Prot is a collection of 562,253 manually reviewed, non-redundant protein sequences from 9,594 organisms. It includes most human and known virus proteins, allowing our model to learn the distribution of both types. We train the self-supervised model using cross-entropy between the predicted token and the true token. We then further pretrain the model with three intermediate pretraining (SPT) tasks from [50]: secondary structure prediction, contact prediction and remote homology prediction. We explain each task in detail in the Appendix.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.14.422772; this version posted December 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

	H1N1			Ebola		
Method	AUROC	AUPR	F1(%)	AUROC	AUPR	F1(%)
SVM [70]	0.886	-	76.2	0.867	-	76.0
MotifTransformer	0.903	0.898	84.4	0.912	0.953	86.0
MotifTransformer+UPT	0.908	0.903	85.5	0.953	0.961	**90.4**
MotifTransformer+UPT+SPT	**0.945**	**0.948**	**86.5**	**0.968**	**0.974**	89.6

Table 4: Virus–Human PPI Tasks from Zhou et al. [70]. Best results are in bold. "-" indicates the metric was not reported.

Baselines. We compare our model against three deep learning methods: a vanilla Transformer, an LSTM [26], and ResNet [23], all run by [50]. Our MotifTransformer uses the same Transformer model size (number of trainable parameters) as the vanilla transformer, and similar model size to the LSTM and ResNet. We do not compare to the pretrained models from [50] since we use a different pretraining dataset. We also compare our method against two baseline methods from [50]. "One-hot" uses one-hot feature inputs that are fed to simple classifiers such as an MLP or 2-layer ConvNet. "Alignment" uses sequence alignment features (BLAST or HHBlits), which are matrices that encode evolutionary information about the protein [57, 53].

Results. Here we investigate the benefits of the MotifTransformer on baseline datasets. Table 2 shows results on the three SPT tasks. Our proposed MotifTransformer performs as well or better than baseline methods, aside from alignment methods on SS prediction. Self supervised language modeling adds improvement over the base MotifTransformer. Multi-task training with language model pretraining outperforms all other non-alignment methods.

6.3 SARS-CoV-2–Human PPI Task

Dataset. While there may be no known Virus–Human interactions for a novel virus, there are many experimentally validated interactions for previous viruses. Our proposed approach is to train an interaction model on known V–H interactions (as indicated by solid lines in Fig 1), and then test on all possible V–H interactions (as indicated by dotted lines in Fig 1). We explain our full training and testing setup below. A summary of the datasets used is provided in Table 1.

Training Data: We use the V–H dataset from [65], which is based on data from the Host-Pathogen interaction Database (HPIDB; version 3.0) [2]. This dataset excludes interactions from large-scale MS experiments, non-physical interactions, redundant PPIs, and interactions between proteins with less than 30 amino acids, more than 5000 amino acids or non-standard amino acids. This resulted in 22,653 experimentally verified human-virus PPIs as a positive sample set. The authors chose negative pairs based on the 'Dissimilarity-Based Negative Sampling' [16]. The selected ratio of positive to negative samples is 1:10. Following Yang et al., we use the same training set (80%) and an independent validation set (20%) for model training and hyperparameter selection, respectively.

Testing Data: For our Virus–Human PPI task, we use the 13,947 known

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.14.422772; this version posted December 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

SARS-CoV-2–Human interactions from the BioGRID database (Coronavirus version 4.1.190) [40]. All BioGRID interactions are experimentally validated, most from [19]. Considering all 20,365 SwissProt human proteins, we label all other pairs from the total space of 20,365*26 to be non-interacting (a total of 529,490).

It is important to note that the labeled "negative" samples contain many pairs of proteins that interact but are not known to do so. This results in an overestimation of the false positive rate. However, this overestimation is currently unavoidable [18].

Baselines: We compare to two baseline models. For the BLAST baseline, we use BLASTp and BioGRID to find potential interactions. In this method, for a given test protein, we search all related proteins from SwissProt. Next, for each found protein we find all its human protein interactions from BioGrid. We consider those proteins to be positive interaction predictions, and all others to be negative. For the Embedding+RF baseline, we use the pretrained model from [67], which uses Doc2Vec to embed protein sequences and then a random forest to classify pairs. No other methods provide code or a browser service to run novel protein interactions.

Results Across most metrics, our method MotifTransformer outperforms the baselines and previous state-of-the-art methods. UPT and SPT pretraining help generalize to our target PPI task, where the non-pretrained models aren't as accurate. We note that our testing dataset is highly imbalanced. In other words, most interactions (99.7%) are negative, or non-interacting. Thus, the AUROC metric is not indicative of good results. We turn our attention to the AUPR metrics, where our method performs the best. This confirms our hypothesis that pretraining on large protein datasets learn evolutionary structures of proteins which generalize well to unseen (i.e. zero-shot) proteins. This is a promising result not only for SARS-CoV-2, but for potential future novel viruses.

6.4 Zhou et al. Virus–Human PPI Tasks

Datasets. In our SARS-CoV-2 dataset, we explain a testing scenario where we have no knowledge of the true V–H interactions, resulting in a large possible interaction space (all possible $|P_v|\times|P_h|$ interactions). Zhou et al. [70] created V–H datasets where they hand selected the negative interactions based on the known positives, making sure that they had an even positive/negative split. While this setup is unrealistic for a true novel virus (because we don't know which ones are positive), we compare to their results to show the strength of our method. Zhou et al. SPTroduce two H–V datasets, H1N1 and Ebola, as explained below.

In the H1N1 dataset, the training set contains 10,955 true PPIs between human and any virus except H1N1 virus, plus an equal amount (10,955) negative interaction samples. The testing set contains 381 true PPIs between human and H1N1 virus, and 381 negative interactions. Similarly, in the Ebola dataset, the training set contains 11,341 true PPIs between human and any virus except Ebola virus, plus an equal amount (11,341) negative interaction samples. The

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.14.422772; this version posted December 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

testing set contains 150 true PPIs between human and Ebola virus, and 150 negative interactions.

Baseline. For these datasets, we use the baseline from [70], which showed that their SVM was the state-of-the-art method.

Results. For the V–H datasets from [70], we can see that our method outperforms the previous state-of-the-art SVM baseline. While this dataset is not indicative of a real novel virus setting since the test set negatives are hand-selected, we can use it to compare different methods. Since this dataset has an even positive/negative testing split, AUROC is a good metric to compare methods, and we can see that across both novel viruses, our method outperforms the SVM. We see notable performance increase using the pretrained MotifTransformer.

7 Conclusion

Computational methods predicting PPIs are critical for a novel virus that threatens widespread public health. Most previous methods are developed for intra-species interactions, and do not generalize to novel viruses. In this paper, we introduce the MotifTransformer for protein interaction prediction between a novel virus and humans. We propose a transfer learning approach for predicting protein interactions. We introduce a new testing setup using protein interactions from SARS-CoV-2. We show that our method can help accurately predict Virus–Human interactions early on in the virus discovery and experimentation pipeline. This can help biologists better understand how the virus attacks the human body, allowing researchers to potentially develop effective drugs more quickly. By providing a computational model for interaction prediction, we hope this will accelerate experimental efforts to define a reliable network of Virus–Human protein interactions. While this work is focused on SARS-CoV-2, H1N1, and Ebola, our framework is applicable for any virus. In the case of a future novel virus, our framework will be able to rapidly produce-protein interaction predictions.

Acknowledgements This work was partly supported by the National Science Foundation under NSF CAREER award No. 1453580 to Y.Q, as well as the Google Cloud COVID-19 Credits Program. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect those of the National Science Foundation.

References

[1] AlQuraishi, M.: End-to-end differentiable learning of protein structure. Cell systems **8**(4), 292–301 (2019)

[2] Ammari, M.G., Gresham, C.R., McCarthy, F.M., Nanduri, B.: Hpidb 2.0: a curated database for host–pathogen interactions. Database **2016** (2016)

[3] Ando, R.K., Zhang, T.: A framework for learning predictive structures from multiple tasks and unlabeled data. Journal of Machine Learning Research **6**(Nov), 1817–1853 (2005)

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.14.422772; this version posted December 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[4] Baevski, A., Edunov, S., Liu, Y., Zettlemoyer, L., Auli, M.: Cloze-driven pretraining of self-attention networks. arXiv preprint arXiv:1903.07785 (2019)

[5] Barman, R.K., Saha, S., Das, S.: Prediction of interactions between viral and host proteins using supervised machine learning methods. PloS one **9**(11), e112034 (2014)

[6] Ben-Hur, A., Noble, W.S.: Kernel methods for predicting protein–protein interactions. Bioinformatics **21**(suppl_1), i38–i46 (2005)

[7] Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. Journal of machine learning research **3**(Feb), 1137–1155 (2003)

[8] Bepler, T., Berger, B.: Learning protein sequence embeddings using information from structure. arXiv preprint arXiv:1902.08661 (2019)

[9] Bitbol, A.F.: Inferring interaction partners from protein sequences using mutual information. PLoS computational biology **14**(11), e1006401 (2018)

[10] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners (2020)

[11] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of machine learning research **12**(Aug), 2493–2537 (2011)

[12] Consortium, U.: Uniprot: a worldwide hub of protein knowledge. Nucleic acids research **47**(D1), D506–D515 (2019)

[13] Cui, G., Fang, C., Han, K.: Prediction of protein-protein interactions between viruses and human by an svm model. In: BMC bioinformatics. vol. 13, p. S5. Springer (2012)

[14] Davey, N.E., Travé, G., Gibson, T.J.: How viruses hijack cell regulation. Trends in biochemical sciences **36**(3), 159–169 (2011)

[15] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[16] Eid, F.E., ElHefnawi, M., Heath, L.S.: Denovo: virus-host sequence-based protein–protein interaction prediction. Bioinformatics **32**(8), 1144–1150 (2016)

[17] Fields, S., Song, O.k.: A novel genetic system to detect protein–protein interactions. Nature **340**(6230), 245–246 (1989)

[18] Gomez, S.M., Noble, W.S., Rzhetsky, A.: Learning to predict protein–protein interactions from protein sequences. Bioinformatics **19**(15), 1875–1881 (2003)

[19] Gordon, D.E., Jang, G.M., Bouhaddou, M., Xu, J., Obernier, K., White, K.M., O'Meara, M.J., Rezelj, V.V., Guo, J.Z., Swaney, D.L., et al.: A sars-cov-2 protein interaction map reveals targets for drug repurposing. Nature pp. 1–13 (2020)

[20] Guo, Y., Yu, L., Wen, Z., Li, M.: Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. Nucleic acids research **36**(9), 3025–3030 (2008)

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.14.422772; this version posted December 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[21] Hamp, T., Rost, B.: Evolutionary profiles improve protein–protein interaction prediction from sequence. Bioinformatics **31**(12), 1945–1950 (2015)

[22] Hashemifar, S., Neyshabur, B., Khan, A.A., Xu, J.: Predicting protein–protein interactions through sequence-based deep learning. Bioinformatics **34**(17), i802–i810 (2018)

[23] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

[24] Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)

[25] Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., et al.: Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. Nature **415**(6868), 180–183 (2002)

[26] Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)

[27] Hou, J., Adhikari, B., Cheng, J.: Deepsf: deep convolutional neural network for mapping protein sequences to folds. Bioinformatics **34**(8), 1295–1303 (2018)

[28] Karunakaran, K.B., Balakrishnan, N., Ganapathiraju, M.K.: Interactome of sars-cov-2/ncov19 modulated host proteins with computationally predicted ppis (2020)

[29] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

[30] Klausen, M.S., Jespersen, M.C., Nielsen, H., Jensen, K.K., Jurtz, V.I., Soenderby, C.K., Sommer, M.O.A., Winther, O., Nielsen, M., Petersen, B., et al.: Netsurfp-2.0: Improved prediction of protein structural features by integrated deep learning. Proteins: Structure, Function, and Bioinformatics **87**(6), 520–527 (2019)

[31] Li, Z., Zhou, F., Chen, F., Li, H.: Meta-sgd: Learning to learn quickly for few-shot learning. arXiv preprint arXiv:1707.09835 (2017)

[32] Lin, D., Wu, X.: Phrase clustering for discriminative learning. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. pp. 1030–1038. Association for Computational Linguistics (2009)

[33] Lin, Z., Lanchantin, J., Qi, Y.: Must-cnn: a multilayer shift-and-stitch deep convolutional architecture for sequence-based protein structure prediction. In: Thirtieth AAAI conference on artificial intelligence (2016)

[34] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

[35] Martin, S., Roe, D., Faulon, J.L.: Predicting protein–protein interactions using signature products. Bioinformatics **21**(2), 218–226 (2005)

[36] von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., Bork, P.: Comparative assessment of large-scale data sets of protein-protein interactions. Nature **417**(6887), 399–403 (2002)

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.14.422772; this version posted December 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[37] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)

[38] Min, S., Park, S., Kim, S., Choi, H.S., Yoon, S.: Pre-training of deep bidirectional protein sequence representations with structural information (2019)

[39] Nourani, E., Khunjush, F., Durmuş, S.: Computational approaches for prediction of pathogen-host protein-protein interactions. Frontiers in microbiology **6**, 94 (2015)

[40] Oughtred, R., Stark, C., Breitkreutz, B.J., Rust, J., Boucher, L., Chang, C., Kolas, N., O'Donnell, L., Leung, G., McAdam, R., et al.: The biogrid interaction database: 2019 update. Nucleic acids research **47**(D1), D529–D541 (2019)

[41] Park, Y., Marcotte, E.M.: Flaws in evaluation schemes for pair-input computational predictions. Nature methods **9**(12), 1134 (2012)

[42] Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)

[43] Peters, M.E., Ammar, W., Bhagavatula, C., Power, R.: Semi-supervised sequence tagging with bidirectional language models. arXiv preprint arXiv:1705.00108 (2017)

[44] Phizicky, E., Fields, S.: Protein-protein interactions: methods for detection and analysis. Microbiol Rev. **59**(1), 94–123 (1995)

[45] Pitre, S., Hooshyar, M., Schoenrock, A., Samanfar, B., Jessulat, M., Green, J.R., Dehne, F., Golshani, A.: Short co-occurring polypeptide regions can predict global protein interaction maps. Scientific reports **2**, 239 (2012)

[46] Qi, Y., Oja, M., Weston, J., Noble, W.S.: A unified multitask architecture for predicting local protein properties. PloS one **7**(3), e32235 (2012)

[47] Qi, Y., Tastan, O., Carbonell, J.G., Klein-Seetharaman, J., Weston, J.: Semi-supervised multi-task learning for predicting interactions between hiv-1 and human proteins. Bioinformatics **26**(18), i645–i652 (2010)

[48] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners

[49] Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., Shlens, J.: Stand-alone self-attention in vision models. arXiv preprint arXiv:1906.05909 (2019)

[50] Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, X., Canny, J., Abbeel, P., Song, Y.S.: Evaluating protein transfer learning with tape. arXiv preprint arXiv:1906.08230 (2019)

[51] Ravi, S., Larochelle, H.: Optimization as a model for few-shot learning (2016)

[52] Redhead, E., Bailey, T.L.: Discriminative motif discovery in dna and protein sequences using the deme algorithm. BMC bioinformatics **8**(1), 385 (2007)

[53] Remmert, M., Biegert, A., Hauser, A., Söding, J.: Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. Nature methods **9**(2), 173 (2012)

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.14.422772; this version posted December 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[54] Richoux, F., Servantie, C., Borès, C., Téletchéa, S.: Comparing two deep learning sequence-based models for protein-protein interaction prediction. arXiv preprint arXiv:1901.06268 (2019)

[55] Rives, A., Goyal, S., Meier, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., Fergus, R.: Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. bioRxiv p. 622803 (2019)

[56] Rush, A.M.: The annotated transformer. In: Proceedings of Workshop for NLP Open Source Software (NLP-OSS). pp. 52–60 (2018)

[57] Schäffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V., Altschul, S.F.: Improving the accuracy of psi-blast protein database searches with composition-based statistics and other refinements. Nucleic acids research **29**(14), 2994–3005 (2001)

[58] Schick, T., Schütze, H.: Exploiting cloze questions for few-shot text classification and natural language inference. arXiv preprint arXiv:2001.07676 (2020)

[59] Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units. arXiv preprint arXiv:1508.07909 (2015)

[60] Sun, T., Zhou, B., Lai, L., Pei, J.: Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC bioinformatics **18**(1), 1–8 (2017)

[61] Tai, K.S., Socher, R., Manning, C.D.: Improved semantic representations from tree-structured long short-term memory networks. arXiv preprint arXiv:1503.00075 (2015)

[62] Tastan, O., Qi, Y., Carbonell, J.G., Klein-Seetharaman, J.: Prediction of interactions between hiv-1 and human proteins by information integration. In: Biocomputing 2009, pp. 516–527. World Scientific (2009)

[63] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)

[64] Wang, L., Wang, H.F., Liu, S.R., Yan, X., Song, K.J.: Predicting protein-protein interactions from matrix-based protein sequence using convolution neural network and feature-selective rotation forest. Scientific reports **9**(1), 1–12 (2019)

[65] Yang, K.K., Wu, Z., Arnold, F.H.: Machine-learning-guided directed evolution for protein engineering. Nature methods **16**(8), 687–694 (2019)

[66] Yang, L., Xia, J.F., Gui, J.: Prediction of protein-protein interactions from protein sequence using local descriptors. Protein and Peptide Letters **17**(9), 1085–1090 (2010)

[67] Yang, X., Yang, S., Li, Q., Wuchty, S., Zhang, Z.: Prediction of human-virus protein-protein interactions through a sequence embedding-based machine learning method. Computational and structural biotechnology journal **18**, 153–161 (2020)

[68] You, Z.H., Zhu, L., Zheng, C.H., Yu, H.J., Deng, S.P., Ji, Z.: Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set. In: BMC bioinformatics. vol. 15, p. S9. Springer (2014)

[69] Zhang, S.W., Wei, Z.G.: Some remarks on prediction of protein-protein interaction with machine learning. Medicinal Chemistry **11**(3), 254–264 (2015)

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.14.422772; this version posted December 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[70] Zhou, X., Park, B., Choi, D., Han, K.: A generalized approach to predicting protein-protein interactions between virus and host. BMC genomics **19**(6), 568 (2018)

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.14.422772; this version posted December 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Method	AUROC	AUPR	F1
SVM [70]	0.858	-	79.2
Embedding + RF [67]	0.871	-	79.8
MotifTransformer+UPT+SPT	**0.886**	**0.806**	**80.6**

Table 5: Virus–Human PPI Tasks from Barman et al. [5]. Best results are in bold. "-" indicates the metric was not reported.

Method	AUROC	AUPR	F1
DeNovo [16]	-	-	81.9
SVM [70]	0.897	-	84.2
MotifTransformer+UPT+SPT	**0.989**	**0.991**	**95.9**

Table 6: SLiM PPI Tasks from Eid et al. [16]. Best results are in bold. "-" indicates the metric was not reported.

A Appendix

A.1 Additional PPI Experiments

We also report the results for our model on two extra baseline datasets. The first is the host–virus datasets from Barman et al. [5]. Although this dataset is not for a completely unseen test virus, we see our method outperforms the baselines. In Table 5, we see that our method outperforms previous methods including an SVM and random forest.

The second is the SLiMs dataset from Eid et al. [16]. This dataset was constructed specifically to evaluate how well the model learns Short Linear Motifs (SLiMs) that are transferable across train/test splits. In Table 6, we see that our model is significantly better than the baselines. We attribute this to the neural motif finder.

A.2 Details of the Supervised Pretraining (SPT) Tasks

The three tasks we use for supervised pretraining are all from well-defined protein structure datasets. For Secondary Structure (SS) Prediction, the original data is from [30] and report accuracy on the testing set. For contact prediction, data is from [1]. We train the contact model using binary cross-entropy between the predicted contact and the true contact. We report precision of the $L/5$ most likely contacts (where L is the sequence length) for medium and long-range contacts. For homology prediction, the data is from [27]. We train the homology model using cross-entropy between the predicted vector and the true class, and we report accuracy. In all 3 tasks, we use the train/validation/test splits from [50]. In the multi-task setting, we sample a new task uniformly each batch, and update the model parameters. We evaluate the non-pretrained and language model (LM) pretrained MotifTransformer for the supervised pretraining tasks.

We compare our model against three deep learning methods: a vanilla Transformer, an LSTM [26], and ResNet [23], all run by [50]. Our MotifTransformer uses the same Transformer model size as the vanilla transformer, and similar model size to the LSTM and ResNet. We do not compare to the pretrained models from [50] since we use a different pretraining dataset. We also compare our method against two baseline methods from [50]. The first uses one-hot feature inputs that are fed to simple classifiers

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.14.422772; this version posted December 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

such as an MLP or 2-layer ConvNet. The second uses sequence alignment features, which are matrices that encode evolutionary information about the protein [57, 53].

We also examine the effectiveness of the MotifTransformer on the supervised pretraining tasks with and without language model pretraining. Table **??** shows our results on the three SPT tasks. Our proposed MotifTransformer performs as well or better than baseline methods, aside from alignment methods on SS prediction. Language modeling adds improvement over the base MotifTransformer. Multi-task training with language model pretraining outperforms all other non-alignment methods.

Secondary Structure Prediction. While the protein sequence is the primary representation, proteins fold into three-dimensional structures which are crucial for their function. Secondary structure is the three-dimensional form of local protein segments. Each character in the sequence can be labeled by its secondary structure, which is one of $|C|$ classes where $C=$ {Helix, Strand, Other}. This results in a sequence tagging task where each input amino acid character x_i is mapped to a label $y_i \in C$. We predict the likelihood of each class for x_i using the following linear mapping:

$$\hat{\mathbf{y}}_i = \mathbf{W}\mathbf{z}_i + b, \tag{6}$$

with learned matrix $\mathbf{W} \in \mathbb{R}^{|C| \times d}$, bias b, and MotifTransformer output $\mathbf{z}_i \in \mathbb{R}^{d \times 1}$. $\hat{y}$ is then fed through a softmax function to obtain class probabilities.

Contact Prediction. Contact prediction is an auxiliary 3D structure task which aims to predict the contact of each set of amino acid pairs in the sequence. Pair (x_i , x_j) of input amino acids from sequence $\mathbf{x}$ is mapped to a label $y_{ij} \in \{0,1\}$ indicating whether or not the amino acids are physically close (< 8Å apart) to each other.

To produce the contact likelihood of pair (x_i,x_j), we use the following formula which preserves non-directionality of contacts:

$$\hat{y} = \big((\mathbf{z}_i \mathbf{W}_k \cdot \mathbf{z}_j \mathbf{W}_q) + (\mathbf{z}_i \mathbf{W}_q \cdot \mathbf{z}_j \mathbf{W}_k)\big)/2, \tag{7}$$

where $\{\mathbf{W}_k, \mathbf{W}_q\} \in \mathbb{R}^{d \times d}$. $\hat{y}$ is then fed through a sigmoid function to obtain the contact probability.

Remote Homology Detection. The goal of remote homology detection is predict the structural and functional class of a protein. Since proteins evolve, many proteins are structurally (and thus, functionally) similar, although their sequences are slightly different. Accurately predicting the homology of a protein would allow the model to group similar structural proteins together. This is a protein classification task where each input sequence $\mathbf{x}$ is mapped to a label $y \in C$, where $|C|$=1195 different possible protein folds.

We use a designated *CLS* token from the MotifTransformer to predict one of the $|C|$ labels for a given sequence. We use a single linear layer mapping $\mathbf{z}_i$ to a $|C|$-dimensional vector:

$$\hat{\mathbf{y}} = \mathbf{W}\mathbf{z}_{CLS} + b, \tag{8}$$

where $\mathbf{W} \in \mathbb{R}^{|C| \times d}$ is a projection matrix and $\mathbf{z}_i \in \mathbb{R}^{d \times 1}$ is the *CLS* token output vector from the Transformer. $\hat{y}$ is fed through a softmax function to obtain class probabilities.