Accurate prediction of virus-host protein-protein interactions via a Siamese neural network using deep protein sequence embeddings

Graphical abstract

Highlights

- Deep learning approach (STEP) predicts virus protein to human host protein interactions
- It is based on recent deep protein sequence embeddings and Siamese neural network
- Prediction of PPIs of the JCV VP1 protein and of the SARS-CoV-2 spike protein
- Identify parts of sequences that most likely contribute to the PPI using explainable AI

Authors

Sumit Madan, Victoria Demina, Marcus Stapf, Oliver Ernst, Holger Fröhlich

Correspondence

sumit.madan@scai.fraunhofer.de (S.M.), holger.froehlich@scai.fraunhofer.de (H.F.)

In brief

Protein-protein interaction (PPI) databases that include already-known PPIs represent an important resource in bioinformatics. A major challenge is to extend our knowledge of PPIs, which are highly relevant for the development of novel virus-like particles that can deliver therapeutics to targeted cells and tissues. Here, we use these PPI databases and the protein sequence information to train deep Siamese neural network architecture while using transfer learning and apply them to predict new virus-host PPIs with high accuracy.
Accurate prediction of virus-host protein-protein interactions via a Siamese neural network using deep protein sequence embeddings

Sumit Madan, Victoria Demina, Marcus Stapf, Oliver Ernst, and Holger Fröhlich

1Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, 53757 Sankt Augustin, Germany
2Institute of Computer Science, University of Bonn, 53115 Bonn, Germany
3NEUWAY Pharma GmbH, In den Dauen 6A, 53117 Bonn, Germany
4Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, 53113 Bonn, Germany
5Lead contact
*Correspondence: sumit.madan@scai.fraunhofer.de (S.M.), holger.froehlich@scai.fraunhofer.de (H.F.)

https://doi.org/10.1016/j.patter.2022.100551

SUMMARY

Prediction and understanding of virus-host protein-protein interactions (PPIs) have relevance for the development of novel therapeutic interventions. In addition, virus-like particles open novel opportunities to deliver therapeutic compounds to targeted cell types and tissues. Given our incomplete knowledge of PPIs on the one hand and the cost and time associated with experimental procedures on the other, we here propose a deep learning approach to predict virus-host PPIs. Our method (Siamese Tailored deep sequence Embedding of Proteins [STEP]) is based on recent deep protein sequence embedding techniques, which we integrate into a Siamese neural network architecture. We use this approach to train models by using protein sequence information and known PPIs. We apply the models to two use cases to predict virus protein to human host interactions. Altogether our work highlights the potential of deep sequence embedding techniques as well as explainable artificial intelligence methods for the analysis of biological sequence data.

INTRODUCTION

Viral infections can cause severe tissue-specific damage to human health. In case of the infection of brain cells, severe neurological disorders can be the consequence. Accordingly, prediction and understanding of tissue-specific virus-host interactions is important for designing targeted therapeutic intervention strategies. At the same time virus-like particles (VLPs), such as John Cunningham VLPs, open novel opportunities to deliver therapeutic compounds to targeted brain cells and tissues, because...
these proteins have the ability to cross the blood-brain barrier.\(^2\) Hence, it is also relevant from a therapeutic perspective to know the binding of VLPs to potential drug receptors in the brain.

The knowledge about virus-host interactions covered in databases like VirHostNet\(^3\) is limited. While various experimental approaches exist to measure PPIs, including yeast two-hybrid screens, biochemical assays, and chromatography,\(^4\) these methods are often time consuming, laborious, costly, and difficult to scale to large numbers of possible PPIs. Thus, computational methods have been proposed that use various types of protein information to predict PPIs. Older approaches focused on predicting PPIs either using structure and/or genomic context of proteins.\(^5\) Other approaches\(^6,\) suggested classical machine learning algorithms (such as support vector machines) in combination with manually engineered features derived from protein sequences to predict PPIs.

In recent years, deep learning-based approaches\(^6,\) have become popular and have increasingly superseded traditional machine learning approaches for the prediction of PPIs. Often these approaches use known PPIs from established PPI databases (e.g., BioGrid, IntAct, STRING, human protein references database, VirHostNet)\(^3,\) to generate datasets to train deep neural network architectures. Some of these methods use recent network representation learning techniques to predict a known virus-host PPI graph.\(^6\) Other authors focused on protein sequences to predict PPIs. For example, Sun et al.\(^6\) and Wang et al.\(^9\) proposed using a stacked autoencoder. Chen et al.\(^9\) developed a deep learning framework using a Siamese neural architecture to predict binary and multi-class PPIs. Tsukiyama et al.\(^9\) recently proposed a long short-term memory (LSTM)-based model on top of a classical word2vec embedding of sequences to predict human-virus PPIs by using protein sequences. Using the same embedding technique, Liu-Wei et al.\(^9\) developed an approach that predicts host-virus PPIs for multiple viruses considering their taxonomic relationships.

In the last few years, transfer learning-based approaches from the natural language processing (NLP) area have massively impacted the field of protein bioinformatics.\(^9,\) These methods are trained on a huge amount of protein sequences to learn informative features of protein sequences. For instance, Elnaggar et al.\(^9\) used 2.1 billion protein sequences for the pre-training of ProtTrans, a collection of transformer models originally stemming from the NLP field. Such methods allow the transformation of a protein sequence into a vector representation, which can subsequently be used efficiently for various downstream tasks, e.g., protein family classification.\(^9\) There are several advantages of using the available pre-trained transformer models, such as avoiding the error-prone design of hand-crafted features to encode protein sequences and, correspondingly, a much more efficient development of new AI models with a potentially higher prediction performance.

In this article, we introduce a novel deep learning architecture combining the recently published ProtBERT\(^9\) deep sequence embedding approach with a Siamese neural network to predict PPIs by using the primary sequences of protein pairs. While recent publications generally follow a similar strategy, they have used more traditional sequence embedding methods.\(^9\) To our knowledge, our work thus constitutes the first attempt to evaluate the use of the most recent, pre-trained transformer models to obtain a deep learning-based biological sequence embedding for PPI prediction. After evaluating the promising prediction performance of our method (Siamese Tailored deep Sequence Embedding of Proteins [STEP]), we use it for two cases: (i) predicting interactions of the John Cunningham polyomavirus (JCV) major capsid protein VP1 (UniProt:P03089) with human receptors in the brain, and (ii) predicting interactions of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) spike glycoprotein (UniProt:P0DTC2) with human receptors. Predicted interactions in both cases demonstrate a clear interpretation in the light of existing literature knowledge, hence supporting the biological relevance of predictions made by our method.

In this study, we make four contributions to the state-of-the-art. First, we construct a novel deep learning architecture STEP for virus-host PPI prediction that requires only the protein sequences as the input and discards the need of handcrafted or other types of features. Second, we demonstrate that using transformer-based models for PPI prediction achieves at least state-of-the-art performance for PPI prediction. In computer vision and NLP, such transformer-based models have shown that they are well suited for learning contextual relationships hidden in sequential data. However, these have not yet been applied to the field of PPI prediction. Hence, we use and build on the huge effort of Elnaggar et al.,\(^9\) who published a pre-trained ProtBERT model that was trained on more than 2 billion amino acid sequences. In addition, we demonstrate that using transfer learning in STEP achieves state-of-the-art performance, for which we evaluated STEP on multiple publicly available virus-host and host-host PPI datasets. Third, we predict interactions for two viruses that are known to cause serious diseases and provide an interpretation on those predictions demonstrating the support through existing literature knowledge. Last, we show how experimental explainable AI (XAI) techniques could be used to identify regions in protein amino acid sequences that attribute to the prediction of PPI.

RESULTS

Comparative evaluation of STEP with state-of-the-art work

We performed a head-to-head comparison of our STEP architecture (Figure 2) on three different datasets published by Tsukiyama et al.,\(^9\) Guo et al.,\(^9\) and Sun et al.\(^9\) Tsukiyama et al.\(^9\) recently published the LSTM-PHV Siamese model, which uses a more traditional word2vec sequence embedding. The dataset published by the authors consists of host-virus PPIs that were retrieved through the Host-Pathogen Interaction Database\(^2\) 3.0. In total, the dataset consists of 22,383 PPIs with 5,882 human and 996 virus proteins. Additionally, it includes artificially sampled negative instances with the positive to negative ratio of 1:10. The authors themselves compared LSTM-PHV on their dataset against a random Forest approach by Yang et al.,\(^2\) Guo et al.,\(^9\) published a yeast PPI dataset and used support vector machines to build a PPI detection model. Sun et al.\(^9\) created a dataset using human protein references database, which contains human-human PPIs. Tsukiyama et al.,\(^9\) and Guo et al.,\(^9\) performed a five-fold cross-validation (CV) experiment, whereas Sun et al.\(^9\) used a 10-fold CV setting. We evaluated our STEP...
architecture using the exact same datasets with the exact same data splits as the authors of the compared methods. STEP was initialized with the hyperparameters shown in Table S1. Table 1 shows the results of all experiments, demonstrating at least state-of-the-art performance of our method. Additionally, we can conclude that our approach compared on exactly the same data published by Tsukiyama et al.10 performs similar to their LSTM-PHV method and better than the approach by Yang et al.25

Finally, we also evaluated our STEP architecture on two additional tasks, namely, PPI type prediction and a PPI binding affinity estimation using the data and the CV setup provided by Chen et al.17 For both tasks, we reached at least state-of-the-art performances with our approach (see Note S1.1. and Table S4).

**Prediction of JCV major capsid protein VP1 interactions**

We split the brain tissue-specific interactome dataset including all positive and pseudo-negative interactions into training (60%), validation (20%), and test (20%) datasets. The validation set was used for tuning hyperparameters of the model only (see Table S5). After tuning on the validation set, we used our best model to make predictions on the hold-out test set. Figure 1 illustrates the area under receiver operator characteristic curve (AUC) and precision-recall curve (AUPR). The model achieved an AUC and AUPR of 88.78% and 88.32% on the unseen test set, respectively. Also, on an extended test set with a ratio 1:10 of positive to pseudo-negative samples the results are quite stable (see Table S6).

We used this STEP-brain model to predict interactions of the JCV major capsid protein VP1 with all human receptors. Table 2 shows the top 10 predicted interactions that are ranked by the score retrieved by the logistic output function of the model. File S3 contains all the predicted interactions. According to the method of integrated gradients, large parts of the VP1 sequence contribute to our model’s prediction of the PPI with the top ranked receptor KIAA1549 (Figure S4). More specifically, signal peptide N-regions in KIAA1549 negatively contribute to the predicted class, whereas the beginning of the non-cytoplasmic domain region is contributing positively.

Altogether, we observed a strong enrichment of VP1 interactions predicted with olfactory, serotonin, amine, taste, and acetylcholine receptors (Figure S2). Notably, neurotransmitter (and specifically serotonin) receptors have previously been suggested to be the entry of the virus into myelin-producing glial brain cells,26 causing progressive multifocal leukoencephalopathy as a fast progressing and life-threatening neurodegenerative disorder.27 Furthermore, we found an enrichment of tyrosine kinase activity (Figure S3), which is in line with the fact that tyrosine kinase inhibitors have been suggested as therapy against JCV.28,29

We further performed an enrichment analysis with InterPro30 protein domains for the predicted interactions between JCV major capsid protein VP1 and human receptors (Figure S5, Table S7). In line with the gene ontology (GO) enrichment analysis, the two top-ranked protein domains Inter-Pro:IPR006029 and Inter-Pro:IPR006029 are neurotransmitter-gated ion channel transmembrane domains that open transiently upon binding of specific ligands, which then allow transmission of signals at chemical synapses.31,32 Furthermore, the receptor-type tyrosine-protein phosphatase/carboxylic anhydrase domain is enriched, which is in line with the enrichment of tyrosine kinase activity found via GO analysis. The enriched domains Inter-Pro:IPR013106 (immunoglobulin V-set domain) and Inter-Pro:IPR006029 (immunoglobulin-like domain) are both immunoglobulin-like domains that are involved in cell-cell recognition, cell surface receptors, and immune system response,33 which play a role in the recognition of a virus protein.

### Table 1. Overview of the results of comparative evaluation of STEP on LSTM-PHV,16 yeast,23 and human PPI8 datasets

|                  | AUC          | AUPR         | F1            | MCC          |
|------------------|--------------|--------------|---------------|--------------|
| Comparative analysis on host-virus PPI dataset from Tsukiyama et al.10 via 5-fold CV |              |              |               |              |
| Tsukiyama et al.10 | 97.58% ±0.13% | 93.86% ±0.35% | 91.00% ±0.53% | 90.30% ±0.53% |
| STEP (ours)       | 98.72% ±0.16%* | 95.71% ±0.51%* | 91.53% ±0.65%* | 90.82% ±0.72%* |

|                  | AUC          | AUPR         | F1            | MCC          |
|------------------|--------------|--------------|---------------|--------------|
| Comparative analysis on single independent host-virus PPI test dataset from Tsukiyama et al.10 |              |              |               |              |
| Yang et al.25     | 96.30%       | 81.00%       | 72.40%        | 69.70%       |
| Tsukiyama et al.10 | 97.30%       | 93.80%       | 91.10%*       | 90.40%*      |
| STEP (ours)       | 98.50%*      | 94.50%*      | 89.69%        | 88.76%       |

|                  | AUC          | AUPR         | F1            | MCC          |
|------------------|--------------|--------------|---------------|--------------|
| Comparative analysis on Yeast PPI dataset from Guo et al.23 via 5-fold CV |              |              |               |              |
| Guo et al.23      | NA           | NA           | 87.34% ±1.33% | 75.09% ±2.51% |
| Chen et al.17     | NA           | NA           | 97.09% ±0.23% | 94.17% ±0.48% |
| STEP (ours)       | 99.61% ±0.10% | 99.58% ±0.17% | 97.37% ±0.27%* | 94.77% ±0.54%* |

|                  | AUC          | AUPR         | F1            | MCC          |
|------------------|--------------|--------------|---------------|--------------|
| Comparative analysis on Human PPI dataset from Sun et al.8 via 10-fold CV |              |              |               |              |
| Sun et al.8       | NA           | NA           | 97.15%        | NA           |
| STEP (ours)       | 99.74% ±0.03% | 99.66% ±0.04% | 98.84% ±0.09%* | 97.67% ±0.18% |

NA, not available in original publication.

For LSTM-PHV and Yeast PPI datasets, we applied a 5-fold CV similar to the authors of the given studies. For the Human PPI dataset of Sun et al.8, we applied a 10-fold CV for training the STEP models. The highest values are marked with asterisks. More details of each experiment can be found in Tables S1–S3.
Prediction of SARS-CoV-2 spike glycoprotein interactions

We performed a nested CV procedure on the given SARS-CoV-2 interactions dataset. We used five outer and five inner loops to validate the generalization performance and while performing the hyperparameter optimization in the inner loop. In each outer run, we created a stratified split of the interactome into train (4/5) and test (1/5) datasets. In the nested run, we further split the outer train dataset into train (1/5) and validation (1/5) datasets, which were used to optimize the hyperparameters of the model using the respective training data. The performance of the classifiers was evaluated with AUC and was averaged over all nested runs. The best identified hyperparameters (see Table S8) were used to train the models in the outer loop. We retrieved a final generalization performance of 83.42% (±3.91%) AUC and 84.02% (±4.58%) AUPR that was calculated by averaging the prediction results of the outer loop (see Table 3). On an extended test set with a ratio 1:10 of positive to pseudo-negative samples, the results are stable for the AUC; however, the AUPR decreases significantly (Tables S9 and S10).

We used the STEP-virus-host model obtained from the best outer fold to predict interactions of the SARS-CoV-2 spike protein (alpha, delta, and omicron variants) with all human receptors that were not already contained in VirHostNet (see Tables S11–S13). File S4 contains all the predicted interactions for the omicron variant. Interestingly, for all virus variants the sigma intracellular receptor 2 (GeneCards:TMEM97; UniProt:Q5BJF2) was the only one predicted with an outstanding high probability (of >70% in all cases) (Tables S11–S13). The sigma 1 and 2 receptors are thought to play a role in regulating cell survival, morphology, and differentiation.34,35 In addition, the sigma receptors have been proposed to be involved in the neuronal transmission of SARS-CoV-2.36 They have been suggested as targets for therapeutic intervention.37–39 Our results suggest that the antiviral effect observed in cell lines treated with sigma receptor binding ligands might be due to a modulated binding of the spike protein, thus inhibiting virus entry into cells. In this context, an analysis via the integrated gradients method shows that only parts of the sigma 2 receptor and the SARS-CoV-2 spike protein contribute to our model’s prediction of the PPI (Figure S6). More specifically, the non-cytoplasmic domain and EXPERA domains demonstrate positive integrated gradient scores, i.e., the existence of these domains influences our model to make the according prediction.

DISCUSSION

Huge advancements have been made recently by applying deep learning algorithms from NLP to protein bioinformatics. Protein language models such as ProtTrans and ProtBERT,19 which are trained on billions of protein sequences, learn informative features through the transformation of sequences to vector representations. These models previously showed their predictive power in various tasks such as prediction of secondary structure or classification of membrane proteins.19

In our work, we used ProtBERT within a specifically designed Siamese neural network architecture to predict PPIs by only using the primary sequences of protein pairs. We trained our models following a positive unlabeled (PU) learning scheme and performed an extensive evaluation and hyperparameter optimization of our models, demonstrating high prediction performances for virus protein to human receptor interactions of JCV and SARS-CoV-2. An additional head-to-head comparison with the recently published method by Tsukiyama et al.10 using a more traditional word2vec sequence embedding combined with an LSTM unit revealed state-of-the-art prediction performance of our STEP approach.

Interactions predicted by our proposed model between JCV major capsid protein VP1 and receptors in brain cells showed a strong enrichment of different neurotransmitters, including serotonin receptors, which is in line with the current literature. For the SARS-CoV-2 spike protein, our model interestingly predicted for all virus variants an interaction with the sigma intracellular receptor 2, which might explain the cytopathic effects of sigma receptor binding ligands reported in the literature.38–40 In both cases, recent techniques coming from the field of XAI allowed us to interpret model predictions and identify those parts of protein sequences that, according to our model, mostly influence the prediction of respective PPIs. Of course, a validation of these predictions would require experimental procedures that are beyond the scope of this article.

Altogether, our work demonstrates the potential of modern deep learning-based biological sequence embeddings and modern XAI techniques for bioinformatics. While in this article...
we focused on JCV and SARS-CoV-2, our proposed model could in future work be easily trained to predict interactions of other viruses as well and, thus, contribute to the emerging set of computational methods that might help to respond to future epidemic and pandemic situations more effectively. In addition, there is the potential to use our method in the context of modern drug development approaches, which use virus-like particles to deliver compounds to specific tissues and receptors.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for code and data should be directed to and will be fulfilled by the lead contact, Holger Fröhlich (holger.froehlich@scai.fraunhofer.de).

Materials availability

This study did not generate any physical materials.

Data and code availability

The data and source code are available at https://github.com/SCAI-BIO/STEP.

Construction of datasets

Primary data sources

The following primary resources were used to create training and test datasets in this work:

1. UniProt protein sequence dataset\(^41\) containing human protein sequences.
2. UniProt mapping dataset\(^41\) containing mappings to other databases.
3. VirHostNet dataset\(^1\) including virus-host interactions of SARS-CoV-2 spike glycoprotein.
4. PPT-Ohmnet dataset\(^2\) (https://snap.stanford.edu/biodata/datasets/10013/10013-PPT-Ohmnet.html, accessed November 18, 2021) containing brain tissue-specific protein-protein-interactions.
5. The GO\(^2\) receptor protein dataset containing annotation of proteins as receptors and parts of protein complexes.
6. Sequences of JCV major capsid protein VP1 (https://www.uniprot.org/uniprot/P03089, accessed on 18 November 2021) and SARS-CoV-2 spike glycoprotein (https://www.uniprot.org/uniprot/P0DTC2, accessed November 18, 2021).
7. Pathogen-host PPI training and test set provided by Tsukiyama et al.\(^10\) (http://kurata35.bio.kyutech.ac.jp/LSTM-PHV/download_page, accessed November 18, 2021) (used for comparative analysis).
8. Yeast PPI dataset from Guo et al.\(^28\) (used for comparative analysis).
9. Human PPI dataset from Sun et al.\(^8\) (used for comparative analysis).
10. PPI type prediction dataset SHS27k from Chen et al.\(^17\) (used for comparative analysis).
11. PPI binding affinity estimation dataset from Chen et al.\(^17\) (used for comparative analysis).

Construction of brain-specific protein-protein interactome dataset

We chose the PPT-Ohmnet database\(^2\) that includes tissue-specific human PPIs collected from various sources. PPT-Ohmnet only takes physical PPIs into account that are supported by experimental evidence (https://snap.stanford.edu/biodata/datasets/10013/10013-PPT-Ohmnet.html). More specifically, interactions contained in PPT-Ohmnet were collected from various curated databases such as TRANSFAC, IntAct, and MINT.\(^44\) The tissue information for an interaction was inferred through the low-throughput tissue-specific gene expression data.\(^45\) The protein-protein interactome can be considered as a graph, in which the proteins represent nodes and the interactions between them are considered as edges. Furthermore, every edge contains

| Rank | Receptor protein ID | Receptor protein name | Score (in %) | Associated GO molecular function |
|------|---------------------|-----------------------|--------------|----------------------------------|
| 1    | Q9HCM3              | UPF0606 protein KIAA1549 | 99.31        | –                                |
| 2    | O94991              | SLIT and NTRK-like protein 5 | 99.09        | protein binding                  |
| 3    | Q7Z443              | polycystic kidney disease protein 1-like 3 | 98.68        | calcium channel activity, sour taste receptor activity |
| 4    | O60840              | voltage-dependent L-type calcium channel subunit alpha-1F | 98.63        | high voltage-gated calcium channel activity, metal ion binding |
| 5    | P13611              | versican core protein   | 98.51        | calcium ion binding, hyaluronic acid binding, glycosaminoglycan binding, extracellular matrix structural constituent conferring compression resistance |
| 6    | P23471              | receptor-type tyrosine-protein phosphatase zeta | 98.33        | protein tyrosine phosphatase activity, integrin binding, protein binding, phosphatase activity, hydrolase activity, phosphoprotein phosphatase activity, transmembrane receptor protein tyrosine phosphatase activity |
| 7    | Q8N2Q7              | neurologin-1            | 98.33        | neurexin family protein binding, signaling receptor activity, identical protein binding, cell adhesion molecule binding, scaffold protein binding, PDZ domain binding, amyloid-beta binding |
| 8    | Q9BZV3              | interphotoreceptor matrix proteoglycan 2 | 98.23        | heparin binding, hyaluronic acid binding, extracellular matrix structural constituent |
| 9    | P41968              | melanocortin receptor 3 | 98.19        | peptide hormone binding, G protein-coupled receptor activity, melanocyte-stimulating hormone receptor activity, neuropeptide binding, melanocortin receptor activity |
| 10   | P23470              | receptor-type tyrosine-protein phosphatase gamma | 98.14        | protein tyrosine phosphatase activity, identical protein binding, phosphatase activity, transmembrane receptor protein tyrosine phosphatase activity, hydrolase activity, phosphoprotein phosphatase activity |
Table 3. Results of the outer loop folds retrieved during the nested CV of STEP-virus-host model by using the test set with a ratio of 1:1 positive to pseudo-negative instances

| Outer fold | AUC      | AUPR     |
|------------|----------|----------|
| 1          | 88.17%   | 89.93%   |
| 2          | 86.83%   | 88.62%   |
| 3          | 77.03%   | 77.73%   |
| 4          | 82.52%   | 81.67%   |
| 5          | 82.56%   | 82.15%   |
| Mean       | 83.42% (± 3.91%) | 84.02% (± 4.58%) |

The architecture and transfer learning of STEP

We used a deep Siamese neural network architecture while using transfer learning to learn relevant, latent features of PPI pairs based on protein sequences. **ProtBERT** 

**ProtBERT** is a pre-trained model trained on approximately 2 billion protein sequences using a masked language modeling objective. It is based on the BERT model that was developed for the natural language domain. Hereby, ProtBERT considers protein sequences as sentences and the so-called building blocks of proteins—amino acids—as vocabulary. The ProtBERT model, specifically the BFD variant used in this work, consists of 30 layers with 16 attention heads and 1,024 hidden layers. It was trained by using the Lamb optimizer for around 23.5 days on 128 compute nodes each containing 1,024 tesnor processing units. During training, the language model learns to predict the biophysical characteristics of proteins from billions of protein sequences.

**Siamese neural network architecture**

Given a pair of proteins, we first obtained their sequences. These sequences were then fed into a Siamese model architecture (Figure 2), in which the pre-trained ProtBERT model was used to obtain embeddings of both protein sequences. There are various ways to infer the relation between sequence embeddings. Some researchers focus on concatenation and others focus on element-wise multiplication (also known as Hadamard product) of both sequence embeddings. In this work, we implemented an integration layer that uses the Hadamard product to combine the sequence embeddings, as it is often found to be the most effective way to model symmetric characteristics of proteins.

**Classification head for PU learning**

On top of the integration layer, we added a classification head represented by multiple hidden layers (Figure 2). We designed the classification head as a bottleneck-shaped architecture with a combination of dropout and linear layers, which ended in an output layer using a logistic function and thus allowed to rank protein pairs as either more likely to interact (positive) or not (negative). Notably, a network with bottleneck structure introduces a gradual decrease of the number of neurons per layer that allows the network to focus on relevant information and discards redundant or irrelevant information.

**Evaluation criteria**

We evaluated our models using an independent test dataset. This consisted of a defined fraction of known PPIs taken at random and excluded from training plus a specified fraction of pseudo-negatives that were not part of the training set. The performance was measured using the AUC and the AUPR.

It should be re-emphasized that in our data negative samples are those protein pairs for which an interaction is unknown. Therefore, we evaluated the

The architecture and transfer learning of STEP

We used a deep Siamese neural network architecture while using transfer learning to learn relevant, latent features of PPI pairs based on protein sequences. **ProtBERT** is a pre-trained model trained on approximately 2 billion protein sequences using a masked language modeling objective. It is based on the BERT model that was developed for the natural language domain. Hereby, ProtBERT considers protein sequences as sentences and the so-called building blocks of proteins—amino acids—as vocabulary. The ProtBERT model, specifically the BFD variant used in this work, consists of 30 layers with 16 attention heads and 1,024 hidden layers. It was trained by using the Lamb optimizer for around 23.5 days on 128 compute nodes each containing 1,024 tesnor processing units. During training, the language model learns to predict the biophysical characteristics of proteins from billions of protein sequences.

**Siamese neural network architecture**

Given a pair of proteins, we first obtained their sequences. These sequences were then fed into a Siamese model architecture (Figure 2), in which the pre-trained ProtBERT model was used to obtain embeddings of both protein sequences. There are various ways to infer the relation between sequence embeddings. Some researchers focus on concatenation and others focus on element-wise multiplication (also known as Hadamard product) of both sequence embeddings. In this work, we implemented an integration layer that uses the Hadamard product to combine the sequence embeddings, as it is often found to be the most effective way to model symmetric characteristics of proteins.

**Classification head for PU learning**

On top of the integration layer, we added a classification head represented by multiple hidden layers (Figure 2). We designed the classification head as a bottleneck-shaped architecture with a combination of dropout and linear layers, which ended in an output layer using a logistic function and thus allowed to rank protein pairs as either more likely to interact (positive) or not (negative). Notably, a network with bottleneck structure introduces a gradual decrease of the number of neurons per layer that allows the network to focus on relevant information and discards redundant or irrelevant information.

**Evaluation criteria**

We evaluated our models using an independent test dataset. This consisted of a defined fraction of known PPIs taken at random and excluded from training plus a specified fraction of pseudo-negatives that were not part of the training set. The performance was measured using the AUC and the AUPR.

It should be re-emphasized that in our data negative samples are those protein pairs for which an interaction is unknown. Therefore, we evaluated the
Figure 2. Architecture of our STEP model that uses the Siamese neural network while using the ProtBERT embeddings.
ability of our models to enrich true positives at the beginning of a predicted ranking of potential PPIs. This ability is exactly reflected by AUC and AUPR measures, which are thus frequently used in the literature about PU learning.\textsuperscript{47} Notably, from a theoretical point of view the AUC estimated via PU learning and the one from a fully labeled dataset are provably linearly correlated.\textsuperscript{51}

**Hyperparameter optimization**

To tune our system, we performed an extensive Bayesian hyperparameter optimization\textsuperscript{52} using the training data. Owing to the huge amount of training time for a single trial, hyperparameter candidates were evaluated using a single validation set consisting of a specified fraction of known PPIs plus an equal amount of sub-sampled negatives. For each trial, intermediate and final performances were assessed using the AUC measure and captured in an SQL database for later analyses. The captured data were also used by the pruning process of Optuna to stop unpromising trials at an early stage.\textsuperscript{53} Each optimization trial was executed on a 2 $\times$ A100 NVIDIA GPUs with VMEM of 32 GB and five trials were executed parallelly by using 10x GPUs in total. The whole optimization process took 10 full days by executing 116 trials in total. The evaluated hyperparameter ranges and the best parameters are illustrated in Tables S5 and S6.

**Making STEP models explainable: An analysis of integrated gradients**

One of the main criticisms of modern deep learning approaches is their often-perceived black box character. To address this concern, we aimed to understand the influence of individual amino acids on model predictions. For that purpose, we used the integrated gradients method,\textsuperscript{34} which offers an intuitive and mathematically sound approach to explain predictions made by a deep neural network. Integrated gradients require no modifications to the trained model. Given an input sample ($x \in \mathbb{R}^d$), integrated gradients rely on a baseline/reference input sample ($x' \in \mathbb{R}^d$), which we constructed using the concatenation of one class, multiple padding, and one separator token. For a STEP model $F : \mathbb{R}^d \rightarrow [0, 1]$, integrated gradients are then obtained by accumulating the partial derivatives $\frac{\partial F}{\partial x_i}$ with respect to input feature $i$ while moving from the reference $x'$ to the observed input $x$:\textsuperscript{54}

$$\text{IntegratedGrad}_i(x) = \sum_{i=0}^{d} \int_{x'}^{x} \frac{\partial F(x' + a \times (x - x'))}{\partial x_i} ds$$

We used 1,000 steps to approximate the integrated gradients, as suggested by Sundararajan et al.\textsuperscript{34} for highly nonlinear networks.

**Gene set enrichment analysis**

To better understand the biology of all ranked predictions in the individual use cases, we performed a gene set enrichment analysis to investigate an enrichment of gene sets listed in the Molecular Signatures Database.\textsuperscript{55} (MsigDB). We downloaded molecular function gene sets of the GO included as the collection C5 from MsigDB (v7.4, MsigDB/c5.go.mf.v7.4.gmt) and MsigDB/c5.go.bp.v7.4.gmt. We considered a GO term to be statistically significant if, after applying the multiple hypothesis testing correction with the Benjamini-Hochberg method,\textsuperscript{56} its adjusted p value was less than 0.01.

**SUPPLEMENTAL INFORMATION**

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2022.100551.

**ACKNOWLEDGMENTS**

We thank André Gemund for his support regarding the computational infrastructure of Fraunhofer SCAI. We thank NEUWAY Pharma GmbH who provided the funding for the work presented in this study.

**AUTHOR CONTRIBUTIONS**

Conceptualization, O.E. and H.F.; Methodology, H.F. and S.M.; Data Curation, Formal Analysis, Visualization, Investigation, Validation, S.M.; Supervision, H.F.; Project Administration, V.D., M.S., O.E., and H.F.; Writing—Original Draft, S.M. and H.F.; Writing—Review and Editing, S.M., V.D., M.S., O.E., and H.F.

**DECLARATION OF INTERESTS**

V.D., M.S., and O.E. are employees of NEUWAY Pharma GmbH. The company funded the work presented in this article but had no influence on scientific results.

Received: March 11, 2022
Revised: March 28, 2022
Accepted: June 16, 2022
Published: July 28, 2022

**REFERENCES**

1. Swanson, P.A., and McGavern, D.B. (2015). Viral diseases of the central nervous system. Curr. Opin.Virol. 11, 44–54. https://doi.org/10.1016/j.coviro.2014.12.009.
2. Ye, D., Zimmermann, T., Demina, V., Sotnikov, S., Ried, C.L., Rahn, H., Stapf, M., Untucht, C., Rohe, M., Terstappen, G.C., et al. (2021). Trafficking of JC virus-like particles across the blood-brain barrier. Nanoscale Adv. 3, 2488–2500. https://doi.org/10.1039/d0na0879f.
3. Guirind, T., Delmote, S., and Navratti, V. (2019). VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. Nucleic Acids Res. 43, D583–D587. https://doi.org/10.1093/nar/gku1121.
4. Lalonde, S., Ehrhardt, D.W., Loqué, D., Chen, J., Rhee, S.Y., and Frommer, W.B. (2008). Molecular and cellular approaches for the detection of protein–protein interactions: latest techniques and current limitations. Plant J. 53, 610–635. https://doi.org/10.1111/j.1365-313x.2007.03332.x.
5. Skrabanek, L., Saini, H.K., Bader, G.D., and Enright, A.J. (2008). Computational prediction of protein–protein interactions. Mol. Biotechnol. 38, 1–17. https://doi.org/10.1007/s12033-007-0069-2.
6. Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y., and Jiang, H. (2007). Predicting protein–protein interactions based only on sequences information. Proc. Natl. Acad. Sci. USA 104, 4337–4341. https://doi.org/10.1073/pnas.0607879104.
7. Zhou, X., Park, B., Choi, D., and Han, K. (2018). A generalized approach to predicting protein-protein interactions between virus and host. BMC Genom. 56, 588. https://doi.org/10.1186/s12864-018-4924-2.
8. Sun, T., Zhou, B., Lai, L., and Pei, J. (2017). Sequence-based prediction of protein interaction using a deep-learning algorithm. BMC Bioinf. 18, 277. https://doi.org/10.1186/s12859-017-1700-2.
9. Wang, Y.-B., You, Z.-H., Li, X., Jiang, T.-H., Chen, X., Zhou, X., and Wang, L. (2017). Predicting protein–protein interactions from protein sequences by a stacked sparse autoencoder deep neural network. Mol. Biosyst. 13, 1336–1344. https://doi.org/10.1039/c7mb00188f.
10. Tsukiyama, S., Hasan, M.M., Fuji, S., and Kurata, H. (2021). LSTM-PHV: prediction of human-virus protein–protein interactions by LSTM with word2vec. Briefings Bioinf. 22, bbab228. https://doi.org/10.1093/bib/bbab228.
11. Xu, W., Gao, Y., Wang, Y., and Guan, J. (2021). Protein–protein interaction prediction based on ordinal regression and recurrent convolutional neural networks. BMC Bioinf. 22, 485. https://doi.org/10.1186/s12859-021-04369-0.
12. Oughred, R., Rust, J., Chang, C., Breitkreutz, B.-J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolias, N., Zhang, F., et al. (2021). The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci. 30, 187–200. https://doi.org/10.1002/pro.3978.
13. Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., Del-Toro, N., et al. (2014). The MintAct project—IntAct as a common curation platform for 11 molecular interaction databases. Nucleic Acids Res. 42, D358–D363. https://doi.org/10.1093/nar/gkt1115.
14. Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019).
STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 47, D607–D613. https://doi.org/10.1093/nar/gky1131.

15. Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopalan, A., et al. (2009). Human protein reference database—2009 update. Nucleic Acids Res. 37, D767–D772. https://doi.org/10.1093/nar/gkn982.

16. Du, H., Chen, F., Liu, H., and Hong, P. (2021). Network-based virus-host interaction prediction with application to SARS-CoV-2. Patterns 2, 100242.

17. Chen, M., Ju, C.-J.-T., Zhou, G., Chen, X., Zhang, T., Chang, K.-W., Zaniole, C., and Wang, W. (2019). Multifaceted protein–protein interaction prediction based on Siamese residual RCNN. Bioinformatics 35, i305–i314. https://doi.org/10.1093/bioinformatics/btz328.

18. Liu-Wei, W., Kafkas, S., et al. (2009). Human protein reference database—2009 update. Nucleic Acids Res. 37, D767–D772. https://doi.org/10.1093/nar/gkn982.

19. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Liu, S., Maslov, S., Hopkins, M., and Ritz, A. (2020). Generation of two forms of the gamma-aminobutyric acidA receptor gamma 2-subunit in mice by alternative splicing. J. Neurochem. 56, 713–715. https://doi.org/10.1111/jnc.14159.1991.tb08209.x.

20. Wagner, K., Edson, K., Heginbotham, L., Post, M., Huganir, R.L., and Czernik, A.J. (1991). Determination of the tyrosine phosphorylation sites of the nicotinic acetylcholine receptor. J. Biol. Chem. 266, 23784–23789. https://doi.org/10.1016/s0021-9258(16)34351-9.

21. Teichmann, S.A., and Chothia, C. (2000). Immunoglobulin superfamily proteins in Caenorhabditis elegans 1 Edited by G. von Heijne. J. Mol. Biol. 296, 1387–1383. https://doi.org/10.1006/jmbj.1999.3497.

22. Huang, Y.-S., Lu, H.-L., Zhang, L.-J., and Wu, Z. (2014). Sigma-2 receptor ligands and their perspectives in cancer diagnosis and therapy: sigma-2 receptor ligands. Med. Res. Rev. 34, 532–566. https://doi.org/10.1002/med.21297.

23. Guo, L., and Zhen, X. (2015). Sigma-2 receptor ligands: neurobiological effects. Comput. Mater. Continua 22, 989–1003. https://doi.org/10.2174/0929867322666150114163607.

24. Das, A.B., 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. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug-repurposing. Nature 583, 459. https://doi.org/10.1038/s41586-020-2175-4.

25. Ostrov, D.A., Bluhm, A.P., Li, D., Khan, J.Q., Rohamare, M., Rajamanickam, K., Bhanumathy, K., Lew, J., Falzarano, D., Vizeacoumar, F.J., et al. (2021). Highly specific sigma receptor ligands exhibit anti-viral properties in SARS-CoV-2 infected cells. Pathogens 10, 1514. https://doi.org/10.3390/pathogens10111514.

26. Abate, C., Niso, M., Abatematteo, F.S., Contino, M., Colabufo, N.A., and Berardi, P. (2020). PB28, the sigma-1 and sigma-2 receptors modulator with potent anti–SARS-CoV-2 activity: a Review about its pharmacological properties and structure affinity relationships. Front. Pharmacol. 11, 598910. https://doi.org/10.3389/fphar.2020.598910.

27. Yeasilkaya, U.H., Balcioglu, Y.H., and Sahin, S. (2020). Reissuing the sigma receptors for SARS-CoV-2. J. Clin. Neurosci. 80, 72–73. https://doi.org/10.1016/j.jocn.2020.08.014.

28. Dant, C., Niso, M., and Berardi, P. (2020). PB28, the sigma-1 and sigma-2 receptors modulator with potent anti–SARS-CoV-2 activity: a Review about its pharmacological properties and structure affinity relationships. Front. Pharmacol. 11, 598910. https://doi.org/10.3389/fphar.2020.598910.

29. Abate, C., Niso, M., Abatematteo, F.S., Contino, M., Colabufo, N.A., and Berardi, P. (2020). PB28, the sigma-1 and sigma-2 receptors modulator with potent anti–SARS-CoV-2 activity: a Review about its pharmacological properties and structure affinity relationships. Front. Pharmacol. 11, 598910. https://doi.org/10.3389/fphar.2020.598910.

30. Blum, M., Chang, H.-Y., Chuguransky, S., Grego, T., Kandasamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., et al. (2021). The InterPro protein families and domains database: 20 years on. Nucleic Acids Res. 49, D344–D354. https://doi.org/10.1093/nar/gkaa977.

31. Kofuji, P., Wang, J.B., Moss, S.J., Huganir, R.L., and Burt, D.R. (1991). Molecular biology, epidemiology, and pathogenesis of progressive multifocal leukoencephalopathy, the JC virus-induced demyelinating disease of the human brain. Clin. Microbiol. Rev. 4, 471–506. https://doi.org/10.1128/cm.05031-11.

32. Boothpur, R., and Brennan, D.C. (2010). Human polymyositis and disease with emphasis on clinical BK and JC. J. Clin. Virol. 47, 306–312. https://doi.org/10.1016/j.jcv.2009.12.006.

33. Teichmann, S.A., Chothia, C. (2000). Immunoglobulin superfamily proteins in Caenorhabditis elegans 1 Edited by G. von Heijne. J. Mol. Biol. 296, 1387–1383. https://doi.org/10.1006/jmbj.1999.3497.

34. Owen, J.M., Rezelj, V.V., Guo, J.Z., Swaney, D.L., et al. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug-repurposing. Nature 583, 459. https://doi.org/10.1038/s41586-020-2175-4.

35. Ostrov, D.A., Bluhm, A.P., Li, D., Khan, J.Q., Rohamare, M., Rajamanickam, K., Bhanumathy, K., Lew, J., Falzarano, D., Vizeacoumar, F.J., et al. (2021). Highly specific sigma receptor ligands exhibit anti-viral properties in SARS-CoV-2 infected cells. Pathogens 10, 1514. https://doi.org/10.3390/pathogens10111514.

36. Abate, C., Niso, M., Abatematteo, F.S., Contino, M., Colabufo, N.A., and Berardi, P. (2020). PB28, the sigma-1 and sigma-2 receptors modulator with potent anti–SARS-CoV-2 activity: a Review about its pharmacological properties and structure affinity relationships. Front. Pharmacol. 11, 598910. https://doi.org/10.3389/fphar.2020.598910.

37. Yeasilkaya, U.H., Balcioglu, Y.H., and Sahin, S. (2020). Reissuing the sigma receptors for SARS-CoV-2. J. Clin. Neurosci. 80, 72–73. https://doi.org/10.1016/j.jocn.2020.08.014.

38. Dant, C., Niso, M., and Berardi, P. (2020). PB28, the sigma-1 and sigma-2 receptors modulator with potent anti–SARS-CoV-2 activity: a Review about its pharmacological properties and structure affinity relationships. Front. Pharmacol. 11, 598910. https://doi.org/10.3389/fphar.2020.598910.
47. Sansone, E., De Natale, F.G.B., and Zhou, Z.-H. (2019). Efficient training for positive unlabeled learning. IEEE Trans. Pattern Anal. Mach. Intell. 41, 2584–2598. https://doi.org/10.1109/tpami.2018.2860995.
48. Eid, F.-E., ElHefnawi, M., and Heath, L.S. (2016). DeNovo: virus-host sequence-based protein–protein interaction prediction. Bioinformatics 32, 1144–1150. https://doi.org/10.1093/bioinformatics/btv737.
49. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. Preprint at arXiv. https://doi.org/10.48550/arXiv.1810.04805.
50. You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., and Hsieh, C.-J. (2019). Large batch optimization for deep learning: training bert in 76 minutes. Preprint at arXiv. https://doi.org/10.48550/arXiv.1904.00962.
51. Menon, A., Rooyen, B.V., Ong, C.S., and Williamson, B. (2015). Learning from corrupted binary labels via class-probability estimation. In Proceedings of the 32nd International Conference on Machine Learning (PMLR), pp. 125–134.
52. Bergstra, J., Yamins, D., and Cox, D. (2013). Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures. In International Conference on Machine Learning (PMLR), pp. 115–123.
53. Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: a next-generation hyperparameter optimization framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2623–2631.
54. Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. Preprint at arXiv. https://doi.org/10.48550/arXiv.1703.01365.
55. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. USA 102, 15545–15550. https://doi.org/10.1073/pnas.0506580102.
56. Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc. B 57, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.