

UC San Diego

UC San Diego Previously Published Works

Title

TUnA: an uncertainty-aware transformer model for sequence-based protein-protein interaction prediction.

Permalink

<https://escholarship.org/uc/item/9q42r5vh>

Journal

Briefings in Bioinformatics, 25(5)

Authors

Ko, Young

Parkinson, Jonathan

Liu, Cong

et al.

Publication Date

2024-07-25

DOI



10.1093/bib/bbae359

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial License, available at <https://creativecommons.org/licenses/by-nc/4.0/>

Peer reviewed

TUnA: an uncertainty-aware transformer model for sequence-based protein–protein interaction prediction

Young Su Ko ¹, Jonathan Parkinson¹, Cong Liu ¹, Wei Wang^{1,2,*}

¹Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA 92093-0359, United States

²Department of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, CA 92093-0359, United States

*Corresponding author. Department of Chemistry and Biochemistry, Department of Cellular and Molecular Medicine, UCSD, 9500 Gilman Drive, La Jolla, CA 92093-0359, United States. E-mail: wei-wang@ucsd.edu

Abstract

Protein–protein interactions (PPIs) are important for many biological processes, but predicting them from sequence data remains challenging. Existing deep learning models often cannot generalize to proteins not present in the training set and do not provide uncertainty estimates for their predictions. To address these limitations, we present TUnA, a Transformer-based uncertainty-aware model for PPI prediction. TUnA uses ESM-2 embeddings with Transformer encoders and incorporates a Spectral-normalized Neural Gaussian Process. TUnA achieves state-of-the-art performance and, importantly, evaluates uncertainty for unseen sequences. We demonstrate that TUnA's uncertainty estimates can effectively identify the most reliable predictions, significantly reducing false positives. This capability is crucial in bridging the gap between computational predictions and experimental validation.

Keywords: protein–protein interaction prediction; deep learning; uncertainty awareness

Introduction

Characterizing protein–protein interactions (PPIs) is fundamental to understanding many biological processes such as signal transduction, cellular metabolism, and the maintenance of cellular systems [1]. High-throughput techniques, such as yeast-two-hybrid [2] and tandem affinity purification [3], have greatly accelerated identification of PPIs, but these experiments are often time consuming and labor intensive. Recently, deep learning (DL) methods have emerged as a promising alternative [4]. While protein structure is critical for protein binding, DL models primarily relying on protein sequence, given its relative abundance over structural data, have achieved impressive performance [5, 6]. For example, protein–protein interaction prediction based on siamese residual recurrent convolutional neural network (PIPR) [5] utilizes a Siamese recurrent convolutional neural network to capture local and sequential features such as co-occurrence similarity of amino acids and electrostaticity- and hydrophobicity-based features. PIPR is outperformed in cross-species generalizability by D-SCRIPT [7], which combines linear and convolutional layers to learn a predicted contact map for a given PPI. Recently, Topsy-Turvy [8] combined D-SCRIPT with GLIDE [9], a network-based approach that considers the local shared-neighbor relationships together with the global network information and improved cross-species generalizability over D-SCRIPT.

Despite these advancements, a major challenge of DL-based models is its incapability to detect out-of-distribution (OOD) data points and avoid overfitting to training data. Overfitting is especially concerning for PPI prediction, where the vastness of the protein sequence space, and consequently the PPI space, cannot be fully captured in the training datasets. Prior state-of-the-art

(SoTA) sequence-based models have shown to be effective in predicting PPIs for species they were trained on, but have shown to be poor predictors when tested on untrained species [5–7]. This limitation in generalizability was further highlighted in a recent study that created a human dataset (referred to as the Bernett dataset) with strategically partitioned training, validation, and test datasets, all with the goal of minimizing sequence similarity and node-degree information [10]. When evaluated on the Bernett dataset, the PPI prediction models DeepFE, PIPR, D-SCRIPT, and Topsy-Turvy only achieved a balanced accuracy of 0.52, 0.52, 0.50, and 0.56 respectively, underscoring the urgent need for new methods with improved generalizability [10].

To avoid overfitting to the training data, a powerful strategy is to estimate the uncertainty of the predictions [11]. Uncertainty awareness in PPI prediction is particularly important because of the huge number of possible protein–protein pairs. Uncertainty-aware models provide a measure of confidence alongside its predictions, reporting lower confidence for predictions involving unfamiliar protein pairs or OOD samples, reflecting a self-awareness of its knowledge boundaries [12]. Uncertainty awareness is particularly important when PPI predictions are used for virtual screening, narrowing down specific protein pairs for experimental validation. Uncertainty estimates can serve as a filter, allowing model users to remove highly uncertain predictions to minimize false positives. As it has yet to be utilized for PPI prediction, the integration of uncertainty awareness into PPI prediction is a novel and necessary advancement, one that could enhance the reliability and applicability of DL in this field.

We have developed “TUnA” (Transformer-based Uncertainty Aware Model for PPI prediction), a sequence-based DL method

Received: March 30, 2024. Revised: May 31, 2024. Accepted: July 10, 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

that leverages implicit structural information for increased generalizability and uncertainty awareness for OOD detection. TUnA has three major components that make up the core framework: ESM-2 protein embeddings, use of Transformer encoder for learning intra- and interprotein relationships, and the incorporation of the Spectral-normalized Neural Gaussian Process (SNGP) method for uncertainty awareness [13]. SNGP enhances DL models by applying spectral normalization to the hidden layers and substituting the final fully connected layer with a Gaussian process (GP) layer. These modifications significantly improve models' uncertainty awareness while retaining their original predictive accuracy.

We assess TUnA on two widely used benchmark datasets: the cross-species dataset and the Bennett dataset. In the cross-species task, TUnA improves upon the previously best-performing Topsy-Turvy as well as the previously benchmarked PIPR [5] and D-SCRIPT [7]. Similarly on the Bennett dataset, TUnA is the most accurate and balanced model out of all evaluated methods. Additionally, we show that TUnA's uncertainty awareness improves calibration and we demonstrate a practical application of uncertainty awareness. Our findings suggest that TUnA not only advances the state of PPI prediction but also is the first model to emphasize uncertainty as a core component.

Methods

TUnA is an end-to-end framework designed to process two embedded protein sequences and output a binary prediction indicating noninteraction (0) or interaction (1) as well as a corresponding uncertainty estimate between 0 and 1, where 1 represents the highest possible uncertainty. Following Liu et al.'s methodology [13], we apply spectral normalization to the model's weights. Spectral normalization divides the hidden layer weights by their largest singular value, regularizing the amount of stretching or compression carried out by the hidden layers, ensuring approximately distance-preserving hidden mappings crucial for TUnA's uncertainty awareness. Additionally, we replace all instances of the ReLU activation function with the Swish activation function, as Ramachandran et al. have shown its effectiveness over ReLU [14].

Protein embedding

We utilize the ESM-2 pretrained protein language model to transform protein sequences into vector representations. While ESM-2 offers a range of pretrained models, varying in size from 8M to 15B parameters, we opt for the 150M parameter model due to computational limitations. Given an AA sequence with length N , the embedded representation is an $N \times 640$ matrix. Given that residue distances in sequence are not reflective of the residue distances in 3D structure, positional embeddings are not added.

Intraprotein feature extraction

To capture the intraprotein interactions, we individually process each protein in the input pair with a Transformer encoder. The protein sequences, represented as $N \times 640$ matrices, are first projected into $N \times d$ matrices, where d is the hidden dimension. A mask is applied to the padded regions such that padded regions are ignored during the self-attention block. The output of the intraprotein encoder is a set of encoded $N \times d$ representations for each protein, capturing essential intraprotein relationships and features.

Interprotein feature extraction

While the original Transformer decoder is tailored for text generation, our PPI prediction task requires feature extraction rather than sequential generation. Therefore, we propose the use of a secondary encoder in place of the decoder. This interprotein feature encoder takes as input the concatenated encoded representations from the intraprotein feature extraction step. For instance, if Protein A has length N and Protein B length M , their encoded outputs would be $N \times d$ and $M \times d$ matrices, respectively. Consequently, the input for the interprotein feature extraction module becomes a $(N + M) \times d$ matrix.

Gaussian process prediction module

As outlined by Liu et al. [13], the standard final fully connected layer is replaced by a Gaussian process layer that is conditioned on the symmetric interaction feature vectors during the final epoch of training. The kernel is approximated by the random Fourier features approximation [15]. At the last epoch of training, we calculate the covariance matrix, allowing us to generate both a mean logit and its variance for each example during evaluation. Using the mean and variance, the uncertainty-adjusted probability, P , is calculated using the mean-field approximation [13]. Following Liu et al., we define uncertainty as $(1 - P)/0.25$, meaning uncertainty is the highest when $P = .5$, indicating an OOD sample. To understand why, note that a GP with a radial basis function kernel begins with a prior mean of zero, updating this belief according to the training data [16]. For OOD samples, the GP reverts to its prior mean of 0, leading to an output logit of zero. Since $\text{sigmoid}(0)$ equals 0.5, distant examples are expected to yield a predicted probability P of .5 (Supplementary Fig. 1). We use the uncertaintyAwareDeepLearn 0.0.5 library (<https://github.com/Wang-lab-UCSD/uncertaintyAwareDeepLearn>) to implement the last layer GP.

Implementation and training details

We use the code for TransformerCPI [17], a Transformer-based protein-drug interaction prediction model, as a starting point and heavily adapt the architecture and workflow [17]. TUnA is implemented in PyTorch 1.13.1 with CUDA 11.6 and trained on a NVIDIA A6000 with 48 GB of memory. TUnA minimizes the binary cross-entropy loss with the Adam + Lookahead optimizer [18, 19]. While Adam does not require a learning rate scheduler, we observed adding a StepLR scheduler improved performance. To determine the number of epochs, we trained the TUnA for 20 epochs and then re-trained TUnA until the epoch when it achieved the lowest validation loss to minimize overfitting. Given the large number of hyperparameters and consequently the computational cost of traditional grid search, we identify the most important hyperparameters based on early validation performance. As the size of the hidden dimensions appeared to have a large impact on performance, we identified the optimal hidden dimension through grid search (Supplementary Tables 1 and 2). The selected hyperparameters for TUnA are described in Supplementary Table 3.

During training, we limit the maximum sequence length to 512 AAs due to computational limitations. If a sequence exceeds this length, we randomly select a continuous 512 AA-long subset for each training instance, ensuring varied exposure to different sequence regions. Because we train using mini-batches, we zero-pad sequences shorter than 512 AAs. We note that this is only applied during training and that during testing, the model considers the entire sequence.

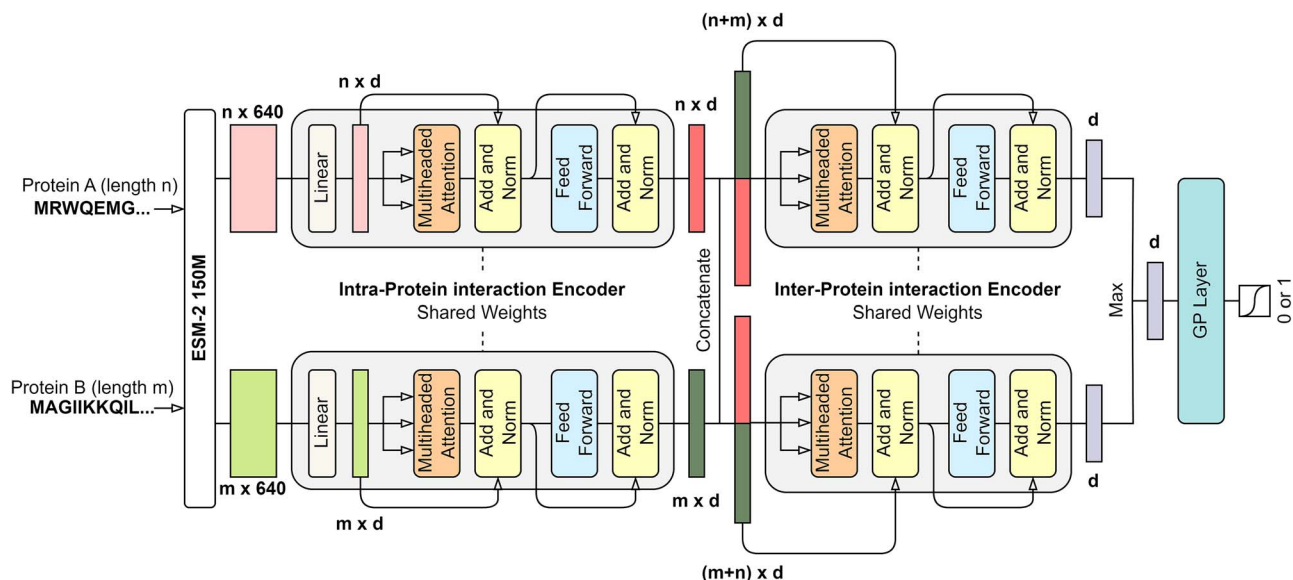


Figure 1. TUnA architecture overview. Protein sequences are embedded using ESM-2 then passed into the intraprotein interaction encoder composed of a linear transformation down to d -dimensions, followed by a standard Transformer encoder. The encoded protein representations are concatenated and passed through the interprotein interaction encoder. We use both concatenations (A||B and B||A) to ensure permutation invariance. The outputs of the interprotein interaction encoder are averaged along all nonpadded regions to output a d -dimension vector, referred to as the interaction feature vector. The interaction feature vectors are then max-pooled and used as the input for the GP layer. The GP layer returns an uncertainty-adjusted probability used to assign the label the protein pair as interacting or noninteracting.

Results

Model overview

The three core components of the TUnA architecture are shown in Fig. 1. First is the protein embedding method. Previous works have utilized such as one-hot encoding, hand-crafted physico-chemical features, or conjoint triad features [5, 20, 21]. Conversely, more recent models such as D-SCRIPT and Topsy-Turvy utilize pretrained protein language models to embed protein sequences. We utilize the SoTA ESM-2 protein language model, which has implicitly learned rich structural information via the masked language modeling objective. While only ever trained on sequence information, ESM-2 can learn structural information as predicting masked residues requires an understanding of evolutionary sequence patterns closely tied to biological structure [22]. ESM-2 provides a structural information-rich starting point for TUnA.

Second is the Transformer-based architecture. Transformers, widely used in natural language processing for their ability to capture rich long-range dependencies through the multiheaded self-attention mechanism, have shown to be impactful even in drug-protein interaction prediction and PPI prediction, as evidenced by TransformerCPI and TransformerGO, respectively [23–24]. Self-attention can be especially useful for protein sequences as the relationship between residues is not sequential in 3D. For example, a protein may have an important structural motif that is composed of residues distant in the amino acid (AA) sequence but close in the 3D structure. We first use the Transformer encoder twice, once per protein, with the goal of extracting an encoded description of the protein by considering the intraprotein interactions. By concatenating the encoded protein representations and passing it through the interprotein encoder, the goal is to create an informative representation of the entire protein-protein complex, combining information about the individual proteins and the interprotein interactions.

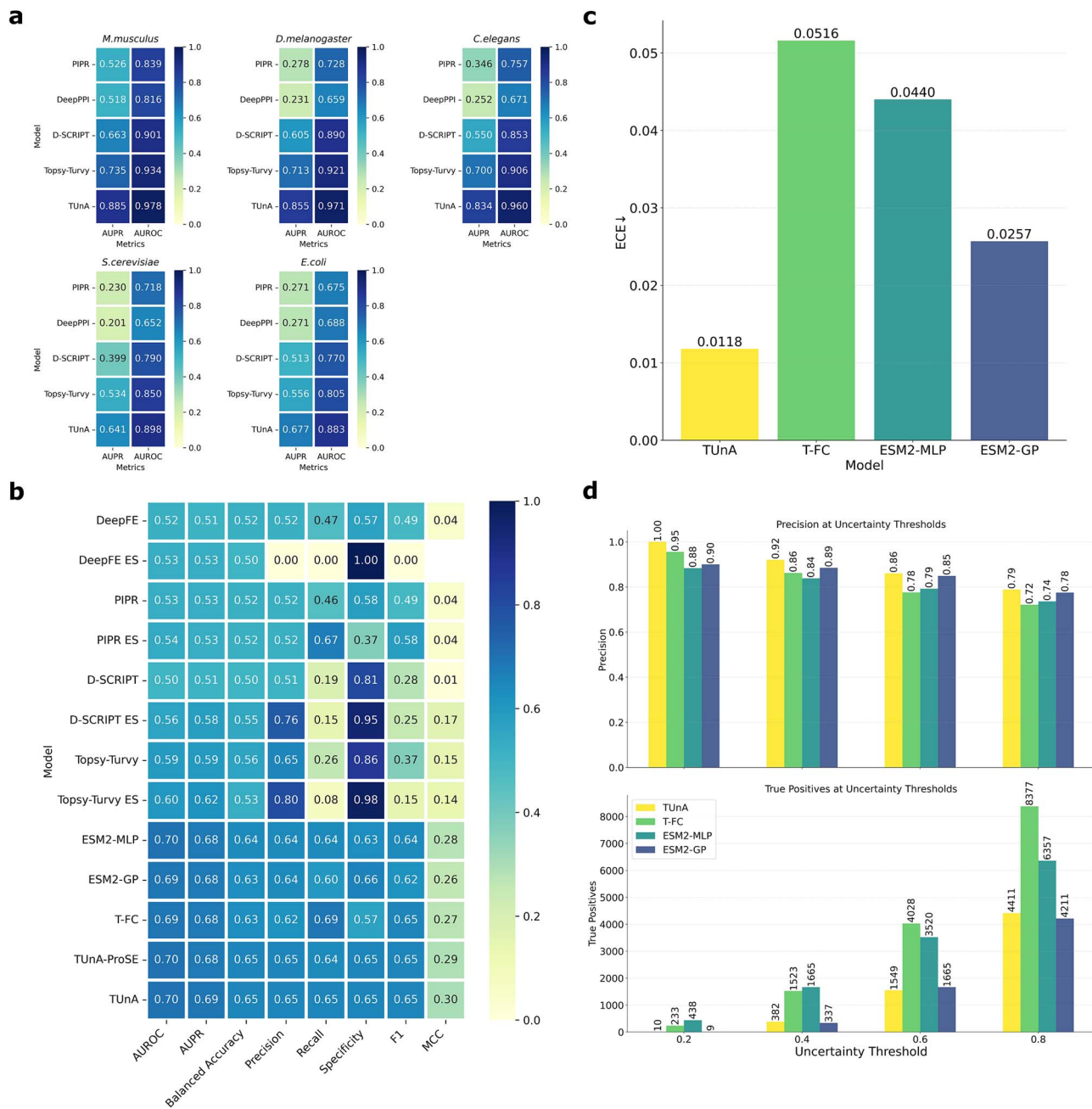
Lastly, we implement the SNGP method outlined by Liu *et al.* to introduce distance and uncertainty awareness [13]. SNGP involves applying spectral normalization to the hidden layers for

approximately distance-preserving hidden mappings and replacing the final fully connected layer with a Gaussian process approximated using random Fourier features. Compared to other uncertainty estimation methods such as Deep Ensembles [25] and Monte Carlo Dropout [26], SNGP only requires only a single network, thus offering a low-cost uncertainty estimate, and also combines the flexibility of neural network models and better uncertainty calibration of GP. During inference, using the learned covariance matrix, TUnA outputs an uncertainty-adjusted probability, P . The uncertainty is a function of P , where uncertainty is highest when $P = .5$.

Performance on cross-species dataset

The cross-species dataset (Supplementary Table 4) was constructed by Sledzieski *et al.*, with PPI data originating from the STRING database (v11) filtered to only include experimentally determined physical binding interactions [7, 27]. Furthermore, Sledzieski *et al.* used CD-HIT [28] to cluster nonhuman sequences with human sequences at 40% similarity. Proteins with high similarity to training set proteins were removed to prevent the model from abusing sequence similarity to make predictions [7]. The datasets are purposely imbalanced, a 10:1 negative to positive ratio, based on the assumption that positive interactions are very rare. Lastly, Sledzieski *et al.* only include PPIs involving proteins between 50 and 800 AAs.

Following D-SCRIPT and Topsy-Turvy, we report the average and standard deviation of performance metrics across three random initializations. Figure 2a shows the area under the precision recall curve (AUPR) and area under the receiver operating curve (AUROC) for each model. Given the large class imbalance, AUPR is an appropriate metric for model evaluation. TUnA demonstrates a clear improvement over Topsy-Turvy, achieving the highest AUPR and AUROC scores across all five evaluated species. TUnA achieves higher performance at a significantly reduced computational cost compared to Topsy-Turvy. While Topsy-Turvy required ~ 79 h for 10 epochs of training, TUnA took ~ 15 h for



18 epochs. In addition, while Topsy-Turvy utilizes the ProSE [29] protein language model’s $N \times 6165$ embedding for each length N protein sequence, TUnA uses ESM-2’s $N \times 640$ embedding, requiring ~ 10 times less memory. In total, the embeddings for the unique mouse sequences for Topsy-Turvy requires ~ 367 Gb, while TUnA only requires ~ 35 Gb. This reduction in computational cost is important for providing a more accessible and practical tool for PPI prediction.

Performance on Bennett dataset

While the cross-species task can provide a measure of a model’s generalizability, particularly regarding the applicability

of knowledge learned from human PPIs to other species, the Bennett dataset [30] serves as another rigorous benchmark. As a human-only dataset, the Bennett dataset lacks evolutionarily conserved information across species that may inflate model performance. The Bennett dataset aims to minimize data leakage. Starting with positive PPIs from the HIPPIEv2.3 [31] human PPI database, negative PPIs were randomly sampled. Then, the PPIs were partitioned into three parts with the graph partitioning framework KaHIP “such that there are no overlapping sequences” [32]. For each partition, an equal amount of negative PPIs was generated by random sampling. Additionally, CD-HIT was used to reduce redundancy “both” within “and between” partitions,

removing any proteins with >40% pairwise sequence similarity. Furthermore, to minimize the node degree bias, in which models may predict a PPI based solely on the fact that one of the proteins is frequently involved in positive interactions, the node degree is balanced between the positive and negative examples.

For our experiments, we adhered to the predetermined partitions for training, validation, and testing in the *Bernett et al.* study [10]. Due to computational limitations in the protein embedding step, we only used protein sequences with a length of 1500 amino acids or less for the training and validation sets. However, the full test set without this length limitation was used for a fair benchmark comparison. Details of the dataset are shown in [Supplementary Table 5](#).

Recently, *Sledzieski et al.* [33] showed that ESM-2 embeddings combined with a multilayer perceptron (MLP) classifier can achieve SoTA performance on the *Bernett* dataset. Because an official implementation of this model is not publicly available, we train a model closely following *Sledzieski et al.*'s architecture [33], referred to as ESM2-MLP, for comparison. Additionally, we aimed to highlight potential differences between using ESM-2 embeddings and ProSE embeddings by training TUnA using ProSE embeddings (TUnA-ProSE). To better understand the individual contributions of Transformer encoder and SNGP, we train two additional models. First, we remove the SNGP components from TUnA, removing spectral normalization and replacing the GP layer with a fully connected layer. We refer to this model as the Transformer with fully-connected (T-FC). Second, we train a model in which we add SNGP to ESM2-MLP, which we refer to as ESM2-GP. For ESM2-GP, we add spectral normalization to ESM2-MLP and replace the last fully connected layer with a GP layer. For T-FC and ESM2-GP, we use the same hyperparameters of their respective original models, except the number of epochs trained for. We determine the number of epochs using the same methodology previously described. As the final part of this benchmark, we compare the runtimes of TUnA with other models with top performance on this dataset to provide a quantitative assessment of the cost associated with each model. The hyperparameters for these models are shown in [Supplementary Table 6](#).

We show that TUnA achieves SoTA performance on the *Bernett* dataset and shows the best balance in the quality and quantity of predictions, evidenced by the highest MCC ([Fig. 2b](#)). While *Topsy-Turvy ES* excels in precision and specificity, it has the lowest recall out of all models, suggesting it is heavily biased toward predicting negative interactions. Based on the performances of TUnA-ProSE, ESM2-MLP, ESM2-GP, and T-FC we get a deeper insight on the influence of the ESM-2 embeddings, the Transformer encoder, and SNGP.

As previously shown by *Sledzieski et al.* [33], ESM2-MLP's strong performance demonstrates ESM-2 embeddings play a critical role in generalizing to unseen sequences. This aligns with our prior belief that the embeddings may contain implicit learned structural and evolutionary information that are important for predicting PPIs. As unseen sequences are different in sequence but potentially similar in structure, ESM-2 provides additional information leverageable by the model. The similar performance between TUnA and TUnA-ProSE indicates that both ESM-2 and ProSE embeddings are informative for PPI prediction. Given that both protein language models are trained on a similarly sized Uniref-based dataset [22, 29], this outcome is not surprising. However, as ProSE embeddings are much larger than ESM-2 embeddings and do not offer a corresponding performance increase, we prefer

Table 1. Comparison of wall-clock runtimes in minutes.

Model	Training runtime	Average time per epoch	Inference runtime
TUnA	219.52	15.58	7.30
T-FC	73.07	14.23	6.08
ESM2-MLP	19.27	1.65	0.82
ESM2-GP	24.57	2.05	2.07

ESM-2 embeddings for their superior performance and lower computational overhead on this benchmark.

Next, a comparison between the no-GP models (ESM2-MLP and T-FC) and GP models (ESM2-GP and TUnA) illustrates the impact of a last GP layer on performance. For the ESM models, a GP layer performs worse than MLP. However, for the transformer models, TUnA has better overall performance than T-FC. The difference in response to the GP underscores the GP's sensitivity to the input features used to update the covariance matrix. Given the Transformer encoder's capacity to generate more informative interaction feature vectors compared to the MLP, incorporating a GP with T-FC (leading to TUnA) is a more lucrative strategy than incorporating GP to ESM2-MLP. We note that the inputs to the GP are the same dimension in both ESM2-GP and TUnA, suggesting the difference is not due to a difference in dimensionality but rather the quality of the input features. Thus, we believe the higher computational cost of the Transformer is justified when used in conjunction with SNGP.

Based on the wall-clock runtimes for training and inference of TUnA, T-FC, ESM2-MLP, and ESM2-GP, we see how computational overhead Transformer-based models require ([Table 1](#)). Unsurprisingly, one training epoch for TUnA and T-FC is longer on average than one training epoch for ESM2-MLP and ESM2-GP as the Transformer architecture is deeper than the MLP architectures. For both training and inference, TUnA requires the most time while ESM2-MLP is the least time consuming.

Effect of uncertainty awareness

We highlight our second novel contribution to the field of PPI prediction—uncertainty awareness. In general, DL models can be overconfident for unseen and OOD data [34]. Overconfidence results in misleading and unreliable predictions. Thus, SNGP is a core part of TUnA, enabling it to make predictions reflecting its knowledge and confidence. Confidence calibration, measured by the expected calibration error (ECE) [35] assesses the quality of a model's uncertainty awareness, where better calibration results in a lower ECE. The *Bernett* test set, given it has no sequences seen during training, is OOD with respect to sequence and thus ideal to evaluate the models' response to OOD data. We calculate the ECE for best-performing models TUnA, T-FC, ESM2-MLP, and ESM2-GP.

While [Fig. 2b](#) suggests T-FC, ESM2-MLP, and ESM2-GP have comparable performance metrics, their respective ECes reveal a drastic difference in calibration. As shown in [Fig. 2c](#), models without SNGP (T-FC and ESM2-MLP) have significantly higher ECes compared to their counterparts with SNGP (TUnA and ESM2-GP). These results suggest SNGP can be an effective method for adding uncertainty awareness to PPI prediction models. Additionally, we observe a similar pattern seen in [Fig. 2b](#) where the improvement in ECE between T-FC and TUnA is greater than the improvement between ESM2-MLP and ESM2-GP, further justifying and highlighting the advantage of the Transformer-GP combination. Overall, TUnA is the best calibrated and most uncertainty-aware model, evidenced by the lowest ECE.

In a study involving protein engineering, Parkinson *et al.* use uncertainty to select and narrow down the most certain predictions for experimental evaluation to save cost and time [36]. While we cannot experimentally validate predictions in this study, we describe a practical application of uncertainty awareness using the Bennett test set. For selecting experimental candidates, precision is a key metric considering the number of false positives must be minimized. In other words, the quality of the predictions can be a more important factor than the number of predictions. For TUnA, T-FC, ESM2-MLP, and ESM2-GP, we calculate the precision after removing predictions above the uncertainty thresholds 0.2, 0.4, 0.6, and 0.8, where 0.2 represents the most stringent threshold. For all models, we use the predictive uncertainty, defined as $(1 - P)/0.25$, where P is the probability of interaction. In addition, we count the number of true positives within each threshold.

In all models, the precision increases as we remove uncertain predictions (Fig. 2d). Furthermore, the models incorporating SNGP see a higher precision across different thresholds compared to the models without, TUnA notably having perfect precision at the 0.2 threshold. In reality, downstream validations are often time consuming where the precision of model predictions is the top priority and TUnA is particularly useful for filtering out predictions based on uncertainty.

Discussion

We introduced TUnA, a novel uncertainty-aware sequence-based model for PPI prediction. TUnA utilizes ESM-2 embeddings as well as Transformer encoders for extracting intra- and inter-protein interactions. In addition, we incorporated SNGP to add uncertainty awareness. To the best of our knowledge, TUnA is the first method to incorporate uncertainty awareness to the PPI prediction task.

First, we showed TUnA improves upon the existing methods for predicting cross-species PPIs as well as for predicting human PPIs without sequence similarity-based data leakage, node degree bias, or any evolutionarily conserved information available in the cross-species task. Second, we explored TUnA's uncertainty awareness as well as the contributions of the ESM-2 embeddings, the Transformer encoder, and SNGP in model performance. We compared TUnA against three different models all utilizing ESM-2 embeddings, T-FC, ESM2-MLP, and ESM2-GP. We found ESM-2 embeddings are a rich starting point for any model given its implicitly learned structural and evolutionary information, even at the 150M parameter level. However, the differences in the models became more apparent when looking at each model's level of uncertainty awareness, measured by the ECE. Incorporation of SNGP improved uncertainty awareness in all cases but more significantly for the Transformer-based model than the MLP-based model, highlighting the advantage of the Transformer-GP combination. We demonstrated that uncertainty awareness is not only a theoretical advantage but also has practical implications for improving precision and reducing the risk of false positives, which is crucial for selecting targets for the follow-up experimental validations.

In conclusion, TUnA represents an advancement of the PPI prediction field as a state-of-the-art, uncertainty-aware method. Through uncertainty awareness, we hoped to bridge the gap between computational predictions and practical application—TUnA's uncertainty estimates provide a simple and effective way for anyone to select the most promising PPI candidates. Future work can continue to improve upon robustness against unseen sequences, continuing to push the boundary of what is possible with sequence alone.

Key Points

- Current deep learning methods for protein–protein interaction (PPI) prediction cannot generalize to proteins different from those in the training set and do not provide reliable uncertainty estimates for their predictions.
- We aimed to bridge the gap between computational predictions and practical application by developing a model that measures the uncertainty of its predictions. This uncertainty enables the identification and prioritization of the most promising PPI candidates for experimental validation.
- We introduce TUnA, the first uncertainty-aware model for PPI prediction. TUnA not only achieves state-of-the-art performance on existing benchmarks but also demonstrates superior uncertainty awareness.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Funding

This work was partially supported by the National Institutes of Health (R21AI158114, R01AI150282).

Data availability

All data and source code used to generate the results in this paper is deposited and publicly available at <https://github.com/Wanglab-UCSD/TUnA>.

Author contributions

W.W. contributed to project conceptualization and project administration. J.P. contributed to supervision and software. C.L. contributed to software. Y.K. contributed to methodology, software, and investigation. All authors contributed to writing, editing, and reviewing.

References

1. Braun P, Gingras A. History of protein–protein interactions: from egg-white to complex networks. *Proteomics* 2012;**12**:1478–98. <https://doi.org/10.1002/pmic.201100563>.
2. Fields S, Song O, kyu. A novel genetic system to detect protein–protein interactions. *Nature* 1989;**340**:245–6. <https://doi.org/10.1038/340245a0>.
3. Gavin AC, Bösch M, Krause R. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 2002;**415**:141–7. <https://doi.org/10.1038/415141a>.
4. Tang T, Zhang X, Liu Y. *et al.* Machine learning on protein–protein interaction prediction: models, challenges and trends. *Brief Bioinform* 2023;**24**:bbad076. <https://doi.org/10.1093/bib/bbad076>.
5. Chen M, Ju CJT, Zhou G. *et al.* Multifaceted protein–protein interaction prediction based on Siamese residual RCNN. *Bioinformatics* 2019;**35**:i305–14. <https://doi.org/10.1093/bioinformatics/btz328>.
6. Hashemifar S, Neyshabur B, Khan AA. *et al.* Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics* 2018;**34**:i802–10. <https://doi.org/10.1093/bioinformatics/bty573>.

7. Sledzieski S, Singh R, Cowen L. et al. D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions. *Cell Syst* 2021;**12**:969–982.e6. <https://doi.org/10.1016/j.cels.2021.08.010>.
8. Singh R, Devkota K, Sledzieski S. et al. Topsy-Turvy: integrating a global view into sequence-based PPI prediction. *Bioinformatics* 2022;**38**:i264–72. <https://doi.org/10.1093/bioinformatics/btac258>.
9. Devkota K, Murphy JM, Cowen LJ. GLIDE: combining local methods and diffusion state embeddings to predict missing interactions in biological networks. *Bioinformatics* 2020;**36**:i464–73. <https://doi.org/10.1093/bioinformatics/btaa459>.
10. Bernett J, Blumenthal DB, List M. Cracking the black box of deep sequence-based protein-protein interaction prediction. *Brief Bioinform* 2024;**25**:bbae076. <https://doi.org/10.1093/bib/bbae076>.
11. Gawlikowski J, CRN T, Ali M. et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:210703342*. 2022.
12. Parkinson J, Wang W. Linear-scaling kernels for protein sequences and small molecules outperform deep learning while providing uncertainty quantitation and improved interpretability. *J Chem Inf Model* 2023;**63**:4589–601. <https://doi.org/10.1021/acs.jcim.3c00601>.
13. Liu J, Lin Z, Padhy S. et al. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *Adv Neural Inf Process Syst* 2020;**33**:7498–512.
14. Ramachandran P, Zoph B, Le QV. Searching for activation functions. *arXiv preprint arXiv:171005941*. 2017. Available from: <http://arxiv.org/abs/1710.05941>.
15. Rahimi A, Recht B. Random features for large-scale kernel machines. In: Platt J, Koller D, Singer Y et al. (eds.), *Advances in Neural Information Processing Systems 20 [Neural Information Processing Systems, NIPS 2007, December 3–6, 2007. Vancouver and Whistler, British Columbia, Canada]*, pp. 1177–84. Red Hook, NY, USA: Curran Associates, 2007. https://papers.nips.cc/paper_files/paper/2007/hash/013a006f03dbc5392effeb8f18fda755-Abs-tract.html.
16. Rasmussen CE, Williams CKI. *Gaussian Processes for Machine Learning*. Cambridge, MA: The MIT Press, 2005. Available from: <https://direct.mit.edu/books/book/2320/gaussian-processes-for-machine-learning>.
17. Chen L, Tan X, Wang D. et al. TransformerCPI: improving compound–protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. Elofsson a, editor. *Bioinformatics* 2020;**36**:4406–14. <https://doi.org/10.1093/bioinformatics/btaa524>.
18. Kingma DP, Ba J. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. 2017. Available from: <http://arxiv.org/abs/1412.6980>.
19. Zhang MR, Lucas J, Hinton G. et al. Lookahead optimizer: k steps forward, 1 step back. *arXiv preprint arXiv:1907.08610*. 2019. Available from: <http://arxiv.org/abs/1907.08610>.
20. Zhang SW, Hao LY, Zhang TH. Prediction of protein–protein interaction with pairwise kernel support vector machine. *Int J Mol Sci* 2014;**15**:3220–33. <https://doi.org/10.3390/ijms15023220>.
21. Wang J, Zhang L, Jia L. et al. Protein-protein interactions prediction using a novel local conjoint triad descriptor of amino acid sequences. *Int J Mol Sci* 2017;**18**:2373. <https://doi.org/10.3390/ijms18112373>.
22. Lin Z, Akin H, Rao R. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**:1123–30. <https://doi.org/10.1126/science.ade2574>.
23. Vaswani A, Shazeer N, Parmar N. et al. Attention is all you need. *arXiv preprint arXiv:1706.03762*. 2017.
24. Ieremie I, Ewing RM, Niranjan M. TransformerGO: predicting protein–protein interactions by modelling the attention between sets of gene ontology terms. Martelli PL, editor. *Bioinformatics* 2022;**38**:2269–77. <https://doi.org/10.1093/bioinformatics/btac104>.
25. Lakshminarayanan B, Pritzel A, Blundell C. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *arXiv preprint arXiv:1612.01474*. 2017.
26. Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. *arXiv preprint arXiv:1506.02142*. 2016.
27. Szklarczyk D, Gable AL, Nastou KC. et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res* 2021;**49**:D605–12. <https://doi.org/10.1093/nar/gkaa1074>.
28. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;**22**:1658–9. <https://doi.org/10.1093/bioinformatics/btl158>.
29. Bepler T, Berger B. Learning the protein language: evolution, structure, and function. *Cell Syst* 2021;**12**:654–669.e3. <https://doi.org/10.1016/j.cels.2021.05.017>.
30. Bernett J. PPI prediction from sequence, gold standard dataset. *figshare* 2022. Available from: https://figshare.com/articles/dataset/PPI_prediction_from_sequence_gold_standard_dataset/21591618/3.
31. Alanis-Lobato G, Andrade-Navarro MA, Schaefer MH. HIPPIE v2.0: enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Res* 2017;**45**:D408–14. <https://doi.org/10.1093/nar/gkw985>.
32. Sanders P, Schulz C. KaHIP v3.00–Karlsruhe high quality partitioning–user guide. *arXiv preprint arXiv:13111714*. 2013.
33. Sledzieski S, Kshirsagar M, Baek M. et al. Democratizing protein language models with parameter-efficient fine-tuning. *Proc Natl Acad Sci U S A* 2024;**121**:e2405840121.
34. Nguyen A, Yosinski J, Clune J. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images. *arXiv preprint arXiv:1412.1897*. 2015.
35. Guo C, Pleiss G, Sun Y. et al. On calibration of modern neural networks. *arXiv preprint arXiv:1706.04599*. 2017. Available from: <http://arxiv.org/abs/1706.04599>.
36. Parkinson J, Hard R, Wang W. The RESP AI model accelerates the identification of tight-binding antibodies. *Nat Commun* 2023;**14**:454. <https://doi.org/10.1038/s41467-023-36028-8>.