ProtSTonKGs: A Sophisticated Transformer Trained on Protein Sequences, Text, and Knowledge Graphs

Helena Balabin[1,2][0000−0002−6392−9306], Charles Tapley Hoyt[3][0000−0003−4423−4370], Benjamin M Gyori[3][0000−0001−9439−5346], John Bachman[3][0000−0001−6095−2466], Alpha Tom Kodamullil[1][0000−0001−9896−3531], Martin Hofmann-Apitius[1][0000−0001−9012−6720], and Daniel Domingo-Fernández[1,4,5][0000−0002−2046−6145]

[1] Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, Sankt Augustin 53757, Germany
[2] Bonn-Rhein-Sieg University of Applied Sciences, 53757, Sankt Augustin, Germany
[3] Laboratory of Systems Pharmacology, Harvard Medical School, 02115, Boston, MA, USA
[4] Fraunhofer Center for Machine Learning, Germany
[5] Enveda Biosciences, Boulder, CO, 80301, USA

Abstract. While most approaches individually exploit unstructured data from the biomedical literature or structured data from biomedical knowledge graphs, their union can better exploit the advantages of such approaches, ultimately improving representations of biology. Using multimodal transformers for such purposes can improve performance on context dependent classification tasks, as demonstrated by our previous model, the Sophisticated Transformer Trained on Biomedical Text and Knowledge Graphs (STonKGs). In this work, we introduce ProtSTonKGs, a transformer aimed at learning all-encompassing representations of protein-protein interactions. ProtSTonKGs presents an extension to our previous work by adding textual protein descriptions and amino acid sequences (i.e., structural information) to the text- and knowledge graph-based input sequence used in STonKGs. We benchmark ProtSTonKGs against STonKGs, resulting in improved F_1 scores by up to 0.066 (i.e., from 0.204 to 0.270) in several tasks such as predicting protein interactions in several contexts. Our work demonstrates how multimodal transformers can be used to integrate heterogeneous sources of information, paving the foundation for future approaches that use multiple modalities for biomedical applications.

Keywords: Natural Language Processing · Knowledge Graphs · Transformers · Bioinformatics · Machine Learning.

1 Introduction

While machine learning approaches have recently been applied in biomedical applications such as drug discovery and protein structure prediction, they tend

to be tailored toward specific applications, and the resulting models often do not generalize well. Thus, transfer learning approaches have enormous potential, since information from one generic setting can be exploited to improve generalization in another specific application. Language models used in natural language processing leverage the transfer learning paradigm to learn general representations of unstructured text data. For instance, the Bidirectional Encoder Representations from Transformers for biomedical text mining (BioBERT) [9] model has been pre-trained on millions of articles from PubMed to represent biomedical knowledge. A complementary approach that represents knowledge in a structured way is the use of knowledge graphs (KGs), which aggregate facts (in the form of (source, relation, target) triples) from heterogeneous data sources. For example, a KG can be constructed to represent all protein-protein interactions (i.e., an interactome). The goal of such KGs is to model biology at the protein-level in order to better understand the underlying processes regulating the cell. By combining the advantages of both approaches (text and KG), we can better represent biology by modelling the interdependencies between the information comprised in unstructured text (e.g., amino acid sequences or protein descriptions) and structured KG data (e.g., known protein interactions).

2 Related Work

Building on the success of the Bidirectional Encoder from Transformers (BERT) model introduced by Devlin et al. [3], several natural language processing approaches have extended the original transformer model architecture through auxiliary information from KGs. However, most approaches are either restricted to the general domain [12], or they require explicit alignments between text and KG entities [7]. Multimodal transformers pose as a generalization to the incorporation of multiple modalities (e.g., text, image and video data) based on the transformer model architecture. Inspired by the the cross encoder presented in the Modulated Detection Transformer (MDETR) model [8], we previously introduced STonKGs, a Sophisticated Transformer trained on biomedical text and Knowledge Graphs [2]. In more detail, STonKGs uses concatenated embedding sequences derived from unstructured text data from biomedical text corpora as well as from structured information from KGs (referred to as text-triple pairs) as input to a joint transformer. However, the concept of multimodal transformers can be extended to further modalities to incorporate additional biological data sources.

3 Approach

In this work, we present ProtSTonKGs, a protein-specific extension of the STonKGs model architecture with an additional modality representing protein sequences as well as further textual information. Given the focus of the model on proteins, we generated a subset of the statements from the Integrated Network and Dynamical Reasoning Assembler (INDRA) [6] used for pre-training STonKGs

by filtering for text-triple pairs in which both the source and the target nodes represent proteins. For these text-triple pairs, we augmented the text evidence with textual node descriptions for source and target nodes obtained from Entrez Gene [10] and the respective amino acid sequences from UniProt [1], resulting in the overall input sequence format shown in Figure 1. In total, we employed 666, 334 protein-specific multimodal inputs, based on statements for which complete information could be obtained for all modalities.

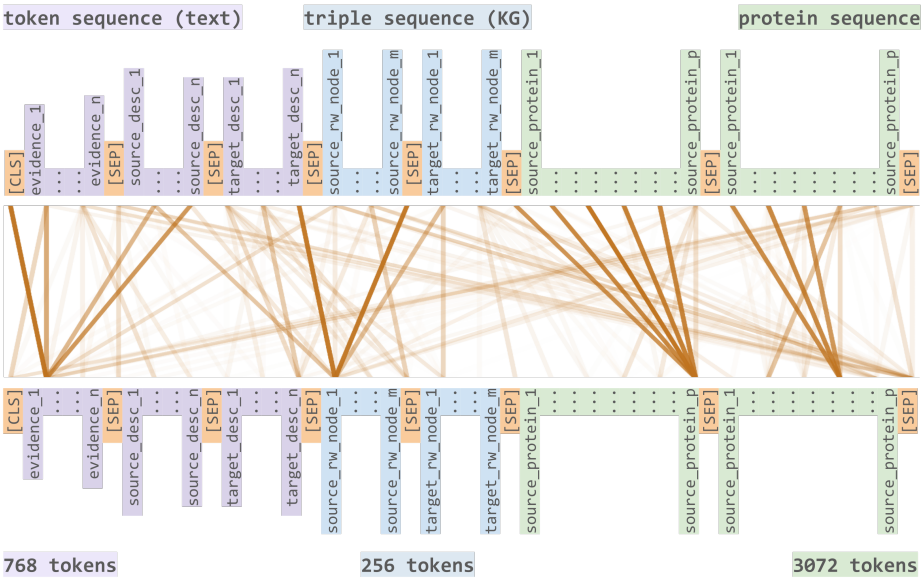

Fig. 1. Cross-modal attention between text, KG, and protein data. Each input is a concatenation of a token, triple, and protein sequence. In pre-training, three different heads are used to convert text, KG, and protein-based inputs to probabilities over the respective vocabularies. These three vocabularies represent the total set of possible text tokens, nodes in the KG, and amino acids.

Since the inclusion of complete amino acid sequences results in input sequences that exceed the maximum input length of BERT [3], we used BigBird [11] as a basis for the cross encoder in ProtSTonKGs instead, as this model is particularly well-suited for handling longer sequence lengths. In parallel to the original STonKGs model, the initial embeddings for text and KG nodes were derived from BioBERT [9] and node2vec [5] (i.e., a re-trained model on the protein-specific subgraph of the INDRA KG), respectively. Moreover, the initial embeddings for the protein sequences were generated using ProtBERT [4]. ProtSTonKGs was pre-trained for $n = 15,000$ training steps on the protein-specific multimodal inputs with a batch size of $b = 256$, and the remaining hyperparameters are equivalent to those used in STonKGs. We evaluated ProtSTonKGs and compared it against STonKGs using the same pre-training and fine-tuning procedures introduced in [2] using weighted F_1 scores. However, given the in-

Model	Relation type		Context annotation				Annotation error	
	1) Polarity	2) Interaction type	3) Cell line	4) Disease	5) Location	6) Species	7) Binary	8) Multiclass
STonKGs_{300k}	**0.838**	**0.988**	0.228	0.204	0.377	**0.863**	**0.957**	**0.968**
ProtSTonKGs	0.802	0.987	**0.251**	**0.270**	**0.381**	0.843	**0.957**	**0.968**
Absolute performance gain	-0.036	-0.001	0.023	0.066	0.004	-0.020	0.000	0.000
Relative performance gain	-4.29%	-0.10%	10.09%	32.25%	1.06%	-2.32%	0.00%	0.00%

Table 1. Benchmark performances of STonKGs_{300k} and ProtSTonKGs. The reported scores are weighted F_1 scores on the test partition of each protein-specific fine-tuning dataset, using a single (80/20) train-test split. Additionally, both absolute and relative ($\frac{\text{ProtSTonKGs} - \text{STonKGs}_{300k}}{\text{STonKGs}_{300k}} * 100$) performance gains are reported as well.

creased computational cost of the longer input sequence lengths used in ProtSTonKGs, we used a single (80/20) train-test split instead of cross-validation for evaluating the models. We created protein-specific subsets of the benchmark datasets used in STonKGs based on text-triple pairs consisting of proteins with textual descriptions from Entrez Gene [10] as well as amino acid sequences from UniProt [1]. Finally, the implementation and the pre-trained ProtSTonKGs model are available at https://github.com/stonkgs/stonkgs and https://huggingface.co/stonkgs/protstonkgs.

4 Experimental Results

As shown in Table 1, ProtSTonKGs outperformed STonKGs on three out of eight classification tasks, and achieved equal F_1 scores on two additional tasks. While ProtSTonKGs resulted in only a minor improvement on task 5 (i.e., a relative performance gain of 1.06%), it led to considerable improvements on task 3 and 4 (i.e., relative performance gains of 10.09% and 32.25%, respectively). The improvement of ProtSTonKGs on these three context classification tasks indicates the potential benefit of including protein-specific information for the disambiguation of various biological contexts of a given text-triple pair. On the two relation type tasks (task 1 and 2), as well as the species task (task 6), the original STonKGs_{300k} performed better than ProtSTonKGs. However, STonKGs outperformed ProtSTonKGs by a smaller margin (a relative difference in performance of less than 5%) on these tasks. Moreover, there is no difference between STonKGs and ProtSTonKGs on the two annotation error tasks (task 7 and 8), which is expected due to the lack of additional informative value (with regards to the prediction of (in)correctly extracted text-triple pairs) of the protein-specific information added in ProtSTonKGs.

5 Conclusion and Future Work

We have presented an extension of our previous STonKGs model, ProtSTonKGs, focused on proteins by incorporating another modality (i.e., protein sequences) as well as additional text data (i.e., textual node descriptions). While this is one of the first efforts towards generating multimodal single stream transformers with more than two modalities in the biomedical field, we envision several possibilities to expand the presented work. For instance, we plan to incorporate other biological entities in the future (e.g., chemicals with node descriptions and simplified molecular-input line-entry system (SMILES) sequences). Furthermore, the pre-trained or fine-tuned models can be used to predict the role of novel proteins in a specific context. Finally, the same multimodal cross encoder can be further pre-trained on other data sources.

References

[1] Rolf Apweiler et al. "UniProt: the Universal Protein knowledgebase". In: *Nucleic Acids Research* (2004).

[2] Helena Balabin et al. "STonKGs: A Sophisticated Transformer Trained on Biomedical Text and Knowledge Graphs". In: *bioRxiv* (2021).

[3] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2019.

[4] Ahmed Elnaggar et al. "ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Learning". In: *bioRxiv* (2021).

[5] Aditya Grover and Jure Leskovec. "node2vec: Scalable Feature Learning for Networks". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016.

[6] Benjamin Mate Gyori et al. "From word models to executable models of signaling networks using automated assembly". In: *Molecular Systems Biology* (2017).

[7] Bin He et al. "BERT-MK: Integrating Graph Contextualized Knowledge into Pre-trained Language Models". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2020.

[8] Aishwarya Kamath et al. "MDETR - Modulated Detection for End-to-End Multi-Modal Understanding". In: *arXiv* (2021).

[9] Jinhyuk Lee et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining". In: *Bioinformatics* (2020).

[10] Donna Maglott et al. "Entrez Gene: gene-centered information at NCBI". In: *Nucleic Acids Research* (2011).

[11] Manzil Zaheer et al. "Big Bird: Transformers for Longer Sequences". In: *Advances in Neural Information Processing Systems 33*. 2020.

[12] Zhengyan Zhang et al. "ERNIE: Enhanced Language Representation with Informative Entities". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019.