 doi:10.20944/preprints202110.0184.v1

Review

Self-attention based models for the extraction of molecular interactions from biological texts

Prashant Srivastava [1,‡], Saptarshi Bej [1,2,‡], Kristina Yordanova [1], and Olaf Wolkenhauer [1,2,*]

[1] Institute of Computer Science, University of Rostock, Germany; prashant.srivastava@uni-rostock.de; saptarshi.bej@uni-rostock.de; kristina.yordanova@uni-rostock.de; olaf.wolkenhauer@uni-rostock.de
[2] Leibniz-Institute for Food Systems Biology, Technical University of Munich, Freising, Germany;
* Correspondence: olaf.wolkenhauer@uni-rostock.de;
‡ These authors contributed equally to this work.

Abstract: For any molecule, network, or process of interest, to keep up with new publications on these, is becoming increasingly difficult. For many cellular processes, molecules and their interactions that need to be considered can be very large. Automated mining of publications can support large scale molecular interaction maps and database curation. Text mining and Natural Language Processing (NLP)-based techniques are finding their applications in mining the biological literature, handling problems such as Named Entity Recognition (NER) and Relationship Extraction (RE). Both rule-based and machine learning (ML)-based NLP approaches have been popular in this context, with multiple research and review articles examining the scope of such models in Biological Literature Mining (BLM). In this review article, we explore self-attention based models, a special type of neural network (NN)-based architectures that have recently revitalized the field of NLP, applied to biological texts. We cover self-attention models operating either at a sentence level or an abstract level, in the context of molecular interaction extraction, published from 2019 onwards. We conduct a comparative study of the models in terms of their architecture. Moreover, we also discuss some limitations in the field of BLM that identifies opportunities for the extraction of molecular interactions from biological text.

Keywords: Text-Mining; Self-Attention Models; Biological Literature Mining; Relationship Extraction; Natural Language Processing

1. Why text-mining?

Text mining techniques used for extracting information from text have been popularly used since 1992. Famous applications of text mining include IBM's Watson program, which performed spectacularly when competing against humans on the nightly game show Jeopardy [1]. Such techniques have played a significant role over the years in extracting and organising information from biological texts. For example, the popular STRING database uses automated text mining of the scientific literature to integrate all known and predicted associations between proteins, including both physical interactions and functional associations [2].

Biological systems are complex in nature. Years of research have produced a large volume of publications on the key molecular players involved in numerous cellular processes, disease phenotypes, and diseases. For example, a PubMed search for the molecule "p53", produces more than a hundred thousand hits, a PubMed search for the disease "colorectal cancer", produces more than two hundred thousand hits. For cell-level and tissue-level processes such as "apoptosis" and "metastasis", there are more than four hundred thousand hits.

In the 5-year span of 2016-2020, the average number of PubMed article hits per year for "p53", "colorectal cancer", "apoptosis" and "metastasis" are 4974, 13548, 29812, and 22305 respectively. One obvious application for text mining is the search for information

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2021 doi:10.20944/preprints202110.0184.v1

from the literature, as part of research projects. Since there are various databases and disease map projects that map out molecular interactions relevant to chosen diseases, the maintenance of such repositories requires substantial effort. A motivation for text mining is then also to assist the updating of data in repositories with new information from publications.

Modelling biological systems can have diverse motivations: investigating molecular interactions and their nature to understand regulatory mechanisms, investing associations between molecules and diseases or broader disease phenotypes, investing the consequences of genetic mutations and perturbations to cellular processes. Clearly, molecules such as genes, proteins, and drugs play a crucial role in such investigations. Rather than attempting to describe complex biological processes as a function of a handful of molecules, systems biologists increasingly appreciate the complexity of these systems, trying to visualise these processes as functions achieved through interactions among numerous molecular entities. These molecular entities (genes, proteins, or drugs) interact in harmony inside biological systems for each phenotype to be realised, be it a cellular process (e.g. cell signalling, metabolism, apoptosis), a disease phenotype (e.g. acute inflammation, metastasis), or even a disease (e.g. cancer, gaucher). However, a comprehensive systemic understanding requires extracting and integrating knowledge acquired from existing and new publications. In many cases, this results in large-scale models that require a lot of manual effort from the modellers, who laboriously hand-pick knowledge from hundreds of publications [37].

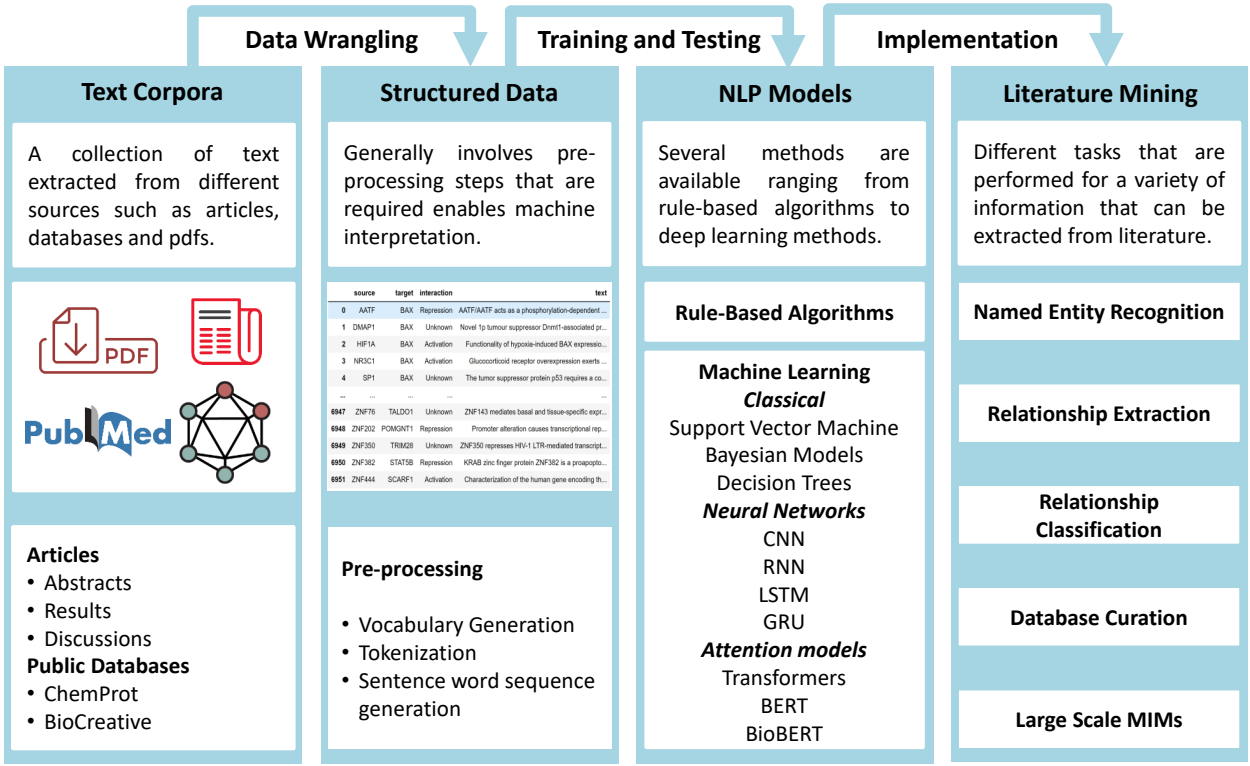

Figure 1. Workflow for biological literature mining (BLM). Starting with collection of text from different sources to processing them into structured data for modelling. Choosing from a plethora of NLP models such as BioBERT to perform BLM tasks such as NER, and RE to extract information from text.

Text-mining and natural language processing (NLP) based techniques are finding their applications for reducing the efforts of biologists to mine the biological literature for tasks like creation of large-scale models and keeping databases updated. This recent field of research focused on automatic knowledge extraction and mining from biomedical literature is known as biomedical literature mining (BLM). In this review article, we

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2021 doi:10.20944/preprints202110.0184.v1

focus on recent developments in BLM for the extraction of molecular interactions from biological texts.

BLM consists of several types of tasks, combinations of which can be realised as complex workflows to achieve the goal of knowledge extraction. An elementary task, for instance, is Named Entity Recognition (NER), which aims to identify biomedical concepts from given text-corpora. State-of-the-art models can perform this task with high accuracy. This upstream task is usually followed by Relationship Extraction (RE). Approaches for the RE task can be broadly categorised as rule-based approaches and machine learning (ML) approaches. Rule-based approaches depend on predefined rules based on inherent textual patterns in biomedical texts. The success of such approaches depends on the quality of the designed rules. In ML approaches, RE is usually posed as a classification problem. However, the design of the classification problem can vary with the motivation of the modeller. For example, there can be a binary classification problem that aims to model merely whether there exists an interacting pair of proteins in a document. More complicated multi-class classification problems investigate the nature of interactions among entities [35]. A general workflow for BLM is provided in Figure 1.

Table 1. Key abbreviations for text mining. Rows with text-mining problem names are marked in blue, model names are marked in red and biological interaction databases names are marked in green.

Abbreviation	Full name	Description
BLM	Biological Literature Mining	Mining information from biological literature/publications
NLP	Natural Language Processing	Ability of a computer program to understand human language
RE	Relationship Extraction	Extracting related entities and the relationship type from biological texts
NER	Named Entity Recognition	NLP-based approaches to identify context specific entity names from text
CNN	Convolutional Neural Network	A type of neural network popularly used in computer vision
RNN	Recurrent Neural Network	One of the neural network models designed to handle sequential data
LSTM	Long Short Term Memory	A successor of RNN useful for handling sequential data
GRU	Gated Recurrent Units	A successor of RNN useful for handling sequential data
BERT	Bidirectional Encoder Representations from Transformer	A pre-trained neural network popularly used for NLP tasks
KAN	Knowledge-aware Attention Network	A self-attention based network for RE problems
PPI	Protein-Protein Interaction	Interactions among proteins, a popular problem in RE
DDI	Drug-Drug Interaction	Interactions among drugs, a popular problem in RE
ChemProt	Chemical- Protein Interaction	Interactions among chemicals and proteins, a popular problem in RE

ML based approaches that are used for RE have several broad categorisations such as feature-based approaches, kernel-based approaches, and neural network-based approaches. Feature-based approaches involve the extraction of expert annotated lexical and syntactic features, and use the same for modelling. Kernel based approaches aim to map syntactic trees to higher-dimensional feature spaces by proper choice of kernels. Neural network (NN)-based approaches can learn latent feature representations from labelled data. Neural network architectures such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM), Gated

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2021 doi:10.20944/preprints202110.0184.v1

Recurrent Units (GRU) have been widely explored in this domain. Recent advances in the field of NLP are due to the introduction of a new class, known as self-attention-based models. These models account for long-range dependencies in text data and can learn contextual associations in text data better than previous neural network-based models [14]. Using transformer architectures as basis, several context-specific pre-trained models such as BERT and BioBERT have been built aimed at facilitating learning from biomedical texts [34–36].

Rule-based text mining approaches have been reviewed comprehensively by Zhao *et al.* [35]. Several deep learning based approaches have been covered by Zhang *et al.* [34], covering publications until 2019. Self-attention based models entered the stage around 2017; pre-trained networks for the extraction of molecular interactions at an abstract or sentence level, from biomedical texts, like BERT and BioBERT were published after 2018. In this review, we therefore focus on self-attention based models, both novel architectures and pre-trained networks, published from 2019 onwards. We first give a brief description of the philosophy behind self-attention. Next, we discuss the architectural aspects and compare the performances of some recent models proposed in the context of BLM. Finally, we discuss the pros and cons of using such models in the context of BLM before concluding our article.

Note that, given that there are a lot of abbreviations for terminologies relevant to this article, we provide some important abbreviations in Table 1, for convenience of the readers.

2. Evolution of Deep Learning models for NLP

Sequence-to-sequence models typically receive a sequence as input and generate a sequence as output. Input and output sequences can be numerical, time-dependent data or string data. Recurrent Neural Network (RNN) is a deep learning based model designed for learning from sequence data. At every learning step, RNNs take elements of a sequence as input and generate an output for that time step and updates a hidden state that can be associated with the "memory" of the network. For text-based data, RNNs once used to be the state-of-the-art models. However, RNNs proved to be less effective to learn from longer sequences, that is to create associations among elements of long sequences. This means that if there is a long sequence of text (a long sentence) and there is an association between two words, one located at the beginning of the sentence and the other towards the end, RNNs are unlikely to capture that information. LSTMs and GRUs were designed to mitigate this "memory" problem. The extremely popular LSTM model, for example, is designed to retain or forget information that is stored in the hidden state sequentially. Transformers, in contrast to the previous models, receive the whole sequence as input rather than taking elements of a sequence sequentially as inputs. To allow the model to recognise the sequential nature of the data, it employs the concept of positional encoding. The attention mechanism is then used to learn associations between elements of the sequence, which in turn is used to make decisions. Taking the entire sequence as input helps this model to learn relatively long-range associations between elements of a long sequence, which makes it apt for text data and thus applicable to NLP. Since the introduction of the attention model by Bahdanau *et al.* for Machine Translation in 2015, it has found applications in a wide range of NN-based architectures [43], while it received more recognition after the introduction of transformer models in 2017 [14]. However, apart from NLP, attention mechanism has been applied in computer vision, time-series analysis, and reinforcement learning [3–5]. In the NLP domain, attention models have helped improve machine translation, question-answering problems, text classification, representation learning, and sentiment analysis [6–10]. In what follows, we discuss some interesting aspects of the self-attention-based models. We briefly visualise the evolution of sequence-to-sequence models in Figure 2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2021 doi:10.20944/preprints202110.0184.v1

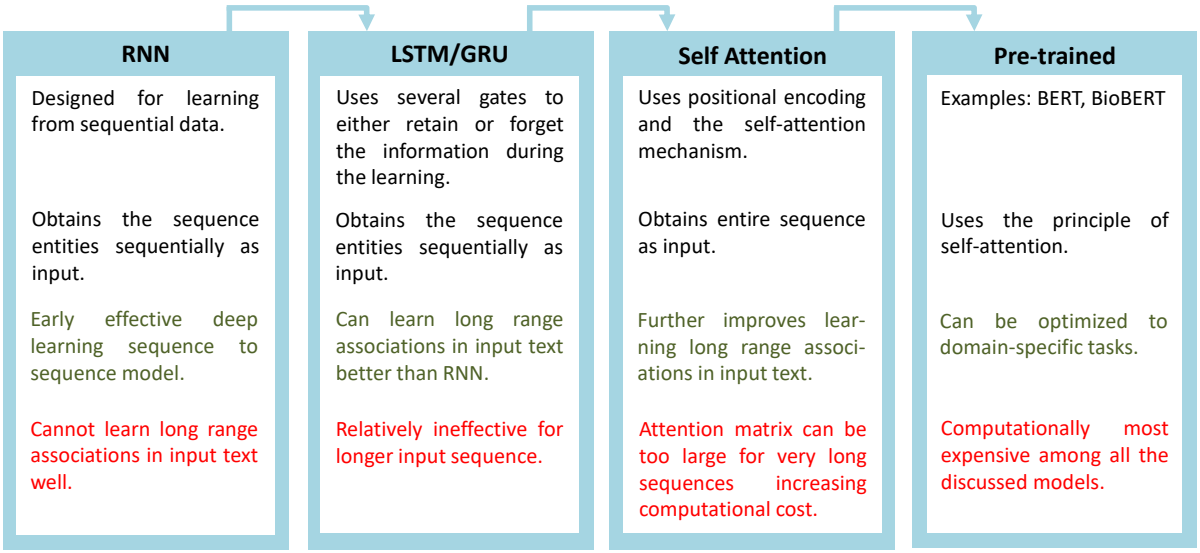

Figure 2. The evolution of sequence-to-sequence models for relationship extraction. Some pros and cons of the models are marked in green and red respectively.

2.1. Self-attention and its advantages

A typical sequence to sequence model consists of an encoder-decoder architecture [11]. Traditional encoder-decoder framework used in RNN, LSTM, or GRU have two main limitations as mentioned in Chaudhari *et al.* [13]:

- The encoder compresses all input information into a vector of fixed length which is passed to the decoder, causing significant information loss [13].
- Such models are unable to model the alignment between input and output vectors. "Intuitively, in sequence-to-sequence tasks, each output token is expected to be more influenced by some specific parts of the input sequence. However, decoder lacks any mechanism to selectively focus on relevant input tokens while generating each output token" [13].

The attention model tackles this issue by enabling the decoder to access the whole encoded sequence. The attention mechanism assigns attention weights over the input sequence, which captures the importance of each token in a sequence and prioritises them for generating output tokens at each step.

The concept of self-attention came into prominence after the introduction of the Transformer model. "Intra-attention, also known as self-attention is an attention mechanism relating different positions of a single sequence to compute a representation of the sequence" [14]. Vaswani *et al.* demonstrated that the transformer architecture has shorter training time and higher accuracy for machine translation without any recurrent component [14]. Transformers have become a state-of-the-art approach for NLP tasks, and they have been adopted for a variety of NLP problems such as Generative Pre-Training Transformer (GPT, GPT-2) for language modelling, Universal Transformer for Question Answering, and Bidirectional Encoder Representations from Transformer (BERT) for language representation [9,15,16]. The transformer model has two key aspects:

1. **Positional encoding:** Given an input sentence in a transformer model, the model first creates a vectorised representation of the sentence S, such that each word in the sentence is represented by a vector of a user-defined dimension. The vectorised version of the sentence S is then integrated with positional encoding. Recall that, unlike sequence-to-sequence models such as RNNs and LSTMs which would feed the sequence elements (words in a sentence) as input sequentially, self-attention based models feed the entire sequence (sentence) as input at a time. This requires a mechanism that can account for the sequential structure of the input sequence/sentence.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2021 doi:10.20944/preprints202110.0184.v1

This is achieved through positional encoding. The formal expression for positional encoding is given by a pair of Equations:

$$P_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \tag{1}$$

$$P_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \tag{2}$$

In Equations (1) and (2), the expression pos is used to denote the position of a word in a sentence and d denotes the dimension of user-defined dimensions for the word-embeddings. That is, each word is essentially perceived by the model as a d-dimensional vector. The index i runs over the dimensions of these word embeddings and take values in the range $[1, d]$. Note that, Equations (1) and (2) propose two different functions over the vector, depending on whether one is calculating an odd index or even index of the word embedding vector. The dependence of the positional encoding functions on $\frac{2i}{d}$, given that these functions are periodic functions by design, ensures that several frequencies are captured over several dimensions of the word-embedding vectors. "The wavelengths form a geometric progression from 2π to $10000 \cdot 2\pi$" [14]. Intuitively, proximal words in a sentence are likely to have a similar P value in a lower frequency, but can still be differentiated in the higher frequencies. For far-apart words in a sentence, the case is just the opposite. Equations (1) and (2) also ensure robustness and uniformity of the positional encoding function P, over all sentences, independent of their length [14].

2. **The self-attention mechanism:** Once the positional encoding is integrated with the word embedding of an input sentence S, the resultant vector W is fed into the mechanism of self-attention. There is a popular analogy used by many data scientists to explain the concepts of Query (Q), Key (K) and Value (V), that are central to the idea of self-attention. When we search for a particular video on YouTube, we submit a query to the search engine, which then maps our query to a set of keys (video title and descriptions), associated to existing videos in the database. The algorithm then presents to us the best possible values as the search result we see. For a self-attention mechanism [14],

$$K = Q = V = W \tag{3}$$

A dot product between Q and K in the form $Q \cdot K^T$, can measure the attention between pairwise words in a sentence, to generate attention weights. The attention weights are used to generate a weighted mean over the word vectors in V, to obtain relevant information from the input as per the given task. As these vectors are learnt through the training procedure of the model, the framework can help the model to retrieve relevant information from an input, for a given task. The Equation governing the process is given by [14]:

$$A(K, Q, V) = \text{softmax}\left(\frac{K \cdot Q^T}{\sqrt{d}}\right) V \tag{4}$$

In practice, however, a multi-headed attention mechanism is used. The idea of multi-heads is again often compared to the use of different filters in CNNs, where each filter learns latent features from the input. Similarly, in multi-headed attention, different heads learn different latent features from the input. The information from all heads is later integrated by a concatenation operation. To account for multiple heads, Equation (3) is violated of course, and the dimension of the positionally encoded word vector W is distributed over the multiple heads. Equation (4), is

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2021 doi:10.20944/preprints202110.0184.v1

also adjusted accordingly by replacing the denominator of $\frac{K \cdot Q^T}{\sqrt{d}}$ by d_k, where d_k is the dimension of the keys considering multiple heads. Several other concepts like layer normalisation and masking are also used in transformer models, which we will not discuss in detail here. A representation of the transformer architecture and attention map over a sentence are provided in Figure 3 [14].

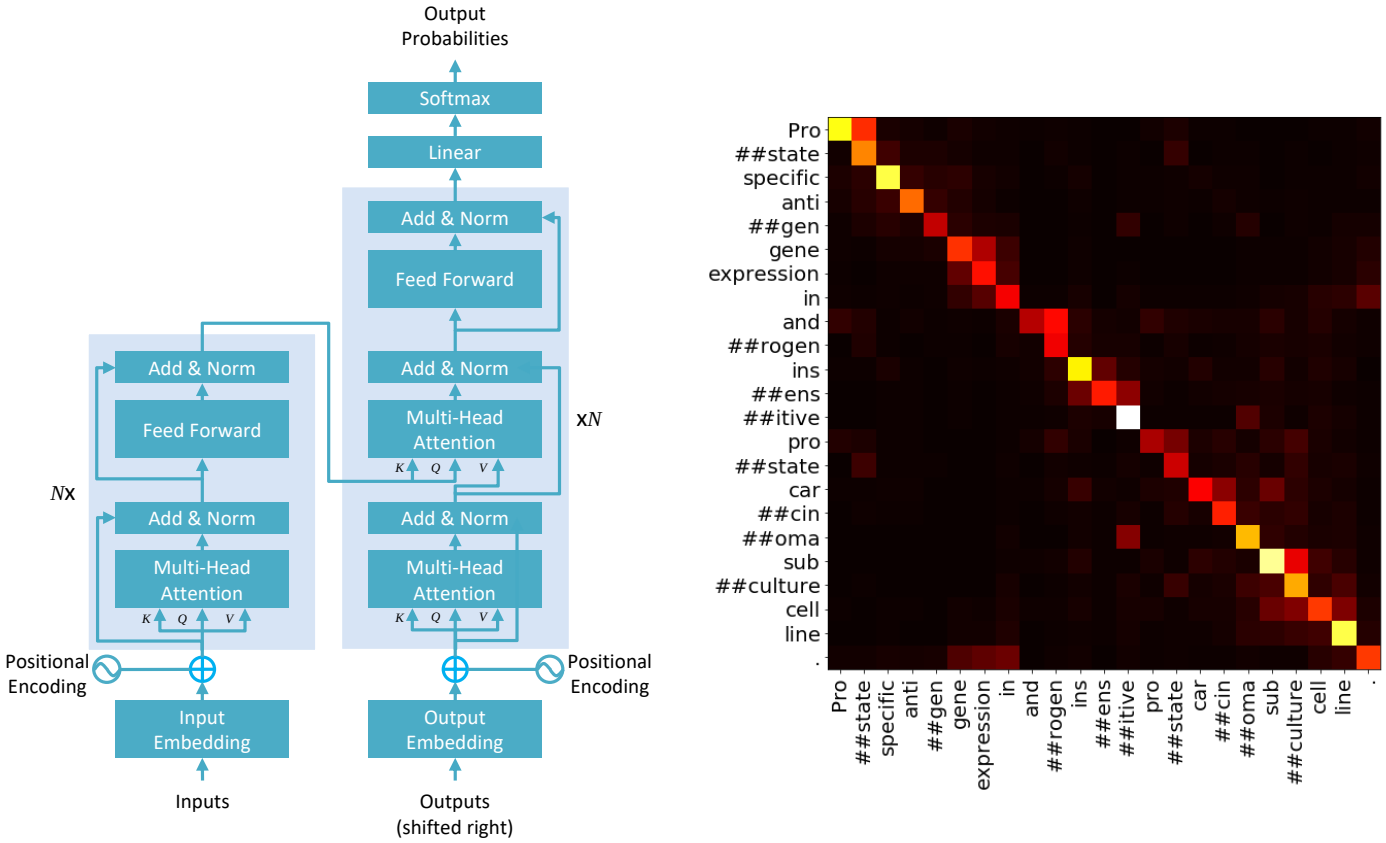

Figure 3. Left: A schematic of the transformer model architecture with attention-based encoder-decoder architecture. The encoder's output is passed into the decoder to be used as Key and Query for the second attention layer. The symbol $N\times$ beside the transformer blocks in the encoder and the decoder represent N layers of the transformer block. Right: An example heatmap of attention mechanism. The heatmap shows pairwise attention weights between pieces of strings in a sentence for a trained model. A hotter hue for a block in the heatmap, corresponds to higher attention between the string in the row and the string in the column, respective to the block.

2.2. *Pre-trained models*

Pre-training models have been in existence for a long time. The idea behind pre-trained language models is to create a black box that can understand language and can be used for specific tasks in that language. These language models are usually pre-trained on very large datasets to generate embeddings which are used in various NLP models. These learned word embeddings are generalised and do not represent any task-specific information. Hence, to utilise them properly, they are fine-tuned on task-specific datasets. Using these pre-trained language representations can help decrease model size and achieve state-of-the-art performance.

BERT was introduced in 2019 by Devlin *et al.* , which is a bi-directional pre-trained transformer network, trained on unlabelled texts. BERT aims to generate a language representation by utilising the encoder network of the Transformer model. BERT can be used in a variety of NLP tasks such as question answering, text classification, language inference, sentiment analysis, etc. Pre-trained BERT model can be fine-tuned with one

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2021 doi:10.20944/preprints202110.0184.v1

additional output layer to create NLP models without requiring task specific architecture engineering [9].

The BERT's authors presented two BERT models, $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$. $\text{BERT}_{\text{BASE}}$ consists of 12 Transformer blocks, 12 self-attention heads, hidden units of size 768, and a total of 110M trainable parameters. Whereas, $\text{BERT}_{\text{LARGE}}$ has 24 Transformer blocks, 24 self-attention heads, with a hidden unit size of 1024 and a total of 340M parameters. BERT can take as input both a single sentence and a pair of sentences as one token sequence, allowing it to handle a variety of NLP tasks. The first token of every sequence is a classification token ($[CLS]$). To separate sentence pairs, a token ($[SEP]$) is used. Moreover, a learned embedding is added to every token, indicating its belonging to a sentence. The input representation is obtained by adding token embeddings, sentence embeddings, and positional embeddings.

Devlin *et al.* used two pre-training strategies for BERT, the first is the masked language model (MLM) and the second is Next Sentence Prediction (NSP). The Masked Language Model randomly chooses 15% of input tokens and masks them by replacing the chosen tokens with $[MASK]$ token. These masked tokens are then predicted by BERT based on the context of other non-masked tokens. The MLM task enables bidirectional Transformer pre-training, which allows the model to learn the context of a word based on both of its left and right surrounding words [9]. In the next sentence prediction task, the model receives pairs of sentences as an input and predicts whether the first sentence is followed by the second sentence. When choosing sentences A and B for NSP pre-training task, 50% of the pre-training examples are chosen such that A is followed by B and labelled as $IsNext$. The other 50% of pre-training examples are chosen such that A is not followed by B and labelled as $NotNext$. For the pre-training corpus, the authors used BooksCorpus having 0.8B words and text passages of English Wikipedia having 2.5B words [17]. WordPiece embedding was used to create a vocabulary of 30,000 words [46]. BERT obtained state-of-the-art performance on eleven NLP tasks including General Language Understanding Evaluation (GLUE) benchmark, Stanford Question Answering Dataset (SQUAD), and Situations With Adversarial Generations (SWAG) dataset [18–20].

Since BERT's release, several BERT based models have been released for domain specific tasks, for example, ALBERT, BERTweet, CamenBERT, RoBERTa, SciBERT, and BioBERT. BioBERT presented by Lee and Yoon *et al.*, is a pre-trained language model for biomedical text mining [22–27]. During pre-training, BioBERT was initialised with weights from BERT and then trained on biomedical domain corpora. The biomedical corpora consists of PubMed abstracts having 4.5B words and PMC full articles having 13.5B words. To ensure BERT's compatibility with BioBERT, the original vocabulary of BERT was used. WordPiece tokenization was applied for words that were not present in BERT's vocabulary (for example, Immunoglobulin is tokenized as I ##mm ##uno ##g ##lo ##bul ##in) [21]. BioBERT outperforms BERT and other state-of-the-art models in three biomedical NLP tasks: NER, RE, and QA. BioBERT achieves state-of-the-art performance requiring only minimal architectural modification. Since its introduction, BioBERT has been used in various NLP tasks [26].

3. Applications of self-attention based models in BLM

3.1. Commonly used datasets

Interactions among genes, proteins, chemicals, and drugs is a well-explored field. These types of studies have been one of the cornerstones of Systems Biology, as they help in visualising complex biological processes at a higher level of complexity. As a result, there are quite a few well-maintained and organised databases in these directions. As we have observed in our review, the most popular ones used for self-attention-based models are BioGRID, IntAct, DrugBank and ChemProt datasets. In addition, many other PPI based databases such as STRING, MINT, BIND, TRRUST and AIR are publicly available [40,41,44,45]. These databases contain annotations for numerous proteins and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2021 doi:10.20944/preprints202110.0184.v1

interactions. Interestingly, however, these datasets are all annotated differently. For example, for the BioGRID database, there are fifteen types of annotated interactions: direct interaction, synthetic lethality, physical association, association, co-localisation, dosage lethality, dosage rescue, phenotypic enhancement, phenotypic suppression, synthetic growth defect, synthetic rescue, dosage growth defect, negative genetic interaction, synthetic haploinsufficiency, and positive genetic interaction [40]. In contrast, for datasets like TRRUST or AIR, there are only three types of mentioned interactions: activation (positve), repression (negative) and unknown (undefined). Some works also prefer to curate customised datasets for their studies [37,42]. Elangovan *et al.* considered IntAct database as the basis of their training data creation. Their annotation is based on chemical characterisations of the interactions [28,41]. They design their study as a classification problem on eight classes: acetylation, methylation, demethylation, phosphorylation, dephosphorylation, ubiquitination deubiquitination, and negative. Su *et al.* and Giles *et al.* on the other hand, use two and five types of annotations respectively for PPI interactions. Moreover, as Giles *et al.* explored in their study, even for human-annotated data, ambiguities do persist [29].

3.2. Architectural comparison of some recent attention-based models

A summary of all discussed models have been provided in Table 2. We will now discuss the architectural aspects of the models in detail.

***Elangovan et al.* (2020)** [28] The motivation of the work by Elangovan *et al.* lies in the fact that popular PPI databases such as IntAct, despite containing a large amount of information on PPIs, only 4% of these interactions are functionally annotated. The functional annotations of two interacting proteins can however be found in relevant publications. Given relevant text data (e.g. abstracts of publications) Elangovan *et al.* focuses on extracting functional annotations of interacting proteins [28].

For this particular work, the authors select PPIs from the IntAct dataset having seven types of functional annotations, namely: phosphorylation, dephosphorylation, methylation, demethylation, ubiquitination, deubiquitination, and acetylation. The task addressed in the article is, therefore, to determine the type of PPI interaction, rather than solely to determine whether two proteins interact. PPIs for which the type of interaction is explicitly mentioned in the abstract of a relevant article are termed as *typed interactions* [28].

Assuming that the type of the interaction of a PPI can appear anywhere in the abstract, possibly across multiple sentences, the authors have used an abstract level annotation of the PPIs. Due to this coarse-grained annotation method, where the data is labelled as per the co-existence of the PPI interaction and the interaction type word in an abstract and not by precise causation between the two entities, the model has been described by the authors as a "weakly supervised" one [28].

The authors are also careful to state their assumption that the annotated PPI interaction is described in the abstract of the article, although in practice this information may prevail in any part of the text. It is further assumed that if, for an annotated PPI interaction in the IntAct database, the type of interaction does not appear in the abstract, then it is annotated somewhere in the full text. Such data instances motivate the authors to define negative samples in the training data. Given a protein pair (p_1, p_2), if there is no associated interaction word in the abstract against the IntAct annotation(s) of the pair, then the protein pair and the IntAct annotation form a negative sample. Note that this implies that a negative sample does not necessarily mean that the protein pair does not interact with each other, but merely that the abstract of the relevant article does not mention this interaction. This rather strong assumption also makes the data noisy, as mentioned by the authors. This, on the other hand, implies that *untyped interactions*, or interactions whose type is not known, would also be a subset of the negative samples [28].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2021 doi:10.20944/preprints202110.0184.v1

The model used by the authors for this paper is a fine-tuned version of the BioBERT model. The fine-tuning process enabled BioBERT to adapt to the typed PPI classification task. The authors refer to this model as PPI-BioBERT in the article. To further improve the probability estimate of each prediction, the authors use an ensemble of 10 PPI-BioBERT models for decision-making [28].

Table 2. Table summarising several aspects of the compared studies. Several publications investigate different variants of the proposed models. We present the performance of only the best model among them.

Work	Datasets	Model	Tasks performed	Performance
Elangovan et al. (2020) [28]	Processed version of IntAct dataset with seven types of interaction	Ensemble of fine-tuned BioBERT models; No External knowledge used	Typed and Untyped RE with relationship types such as, phosphorylation, acetylation etc.	Typed PPI: 0.540; Untyped PPI: 0.717; Metric: F1 Score
Giles et al. (2020) [29]	Manually curated from MEDLINE database	Fine-tuned BioBERT model; Used STRING database knowledge during dataset curation	Classification problem with classes coincidental mention, positive, negative, incorrect entity recognition and unclear	Curated data & BioBERT: 0.889; Metric: F1 Score
Su et al. (2020) [32]	Processed versions of BioGRID, DrugBank and IntAct dataset	Fine-tuned BERT model integrated with LSTM and additive attention; No External knowledge used	Classification tasks on PPI (binary), DDI (multi-class) and ChemProt (multi-class)	PPI: 0.828; DDI: 0.807; ChemProt: .768; Metric: F1-Score
Su et al. (2021) [33]	Processed versions of BioGRID, DrugBank and IntAct dataset	Contrastive learning model; No External knowledge used in dataset curation or as a part of the model	Classification tasks on PPI (binary), DDI (multi-class) and ChemProt(multi-class)	PPI: 0.827; DDI: 0.829; ChemProt: .787; Metric: F1-Score
Wang et al. (2020) [30]	Processed versions of BioCreative VI PPI dataset	A multitasking architecture based on BERT, BioBERT, BiLSTM and Text CNN; No External knowledge used	Document triage classification, NER (auxiliary tasks) and PPI RE (main task).	NER task: .936, PPI RE (exact match evaluation): 0.431; Metric:F1-Score
Zhou et al. (2019) [31]	Processed versions of BioCreative VI PPI dataset	KAN; TransE, is used to integrate prior-knowledge from BioGRID and IntAct datasets on triplets to the model	PPI-RE classification task from BioCreative VI	PPI RE (exact match evaluation): 0.382 PPI RE: (HomoloGene evaluation): 0.404; Metric:F1-Score

***Giles et al.* (2020)** [29] While conventional string matching is used to search for co-occurrences of entities (gene or protein names) in a sentence, it results in the inclusion of large amounts of noise in the results. For instance, as the authors of this particular research work point out, in case of the PPI detection problem, about 75 % of the sentences containing co-occurring names of possibly interacting proteins do not describe any causal relationship between them. With this motivation, the authors investigate the possibility of using fine-tuned BioBERT to analyse these co-occurrences and thereby to accurately determine the functional association between the co-occurring proteins in a given sentence [29].

An interesting experiment conducted by the authors during the data preparation is the investigation of inter-annotator agreement. Three independent expert curators curated PPIs from 925 sentences identified by NER tagging within papers drawn from MEDLINE. Surprisingly, concordance between all three curators was observed in only 48.8 % of the cases, which demonstrates the complexity of the problem [29].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2021 doi:10.20944/preprints202110.0184.v1

Moreover, the authors experimented with the need of a semi-supervised preprocessing step for training data curation. This experiment was necessary due to an inherent class imbalance between positive protein interactions and coincidental mention of proteins. The authors repeated the data curation step after filtering the sentences such that, only those which contained two genes identified to have a strong likelihood of interacting, signified by a StringDB high combined score, were retained. Even with high reliability scores from StringDB, no improvement in the rate of identification of positive interactions was found. However, for some other cases, such as the drug-drug interaction problem, this step proved to be more effective. The authors concluded that this type of preprocessing approach can assist in cases of balanced training data curation in specific problems.

As far as predictive models are concerned, the authors compared some rule-based approaches with a fine-tuned version of BioBERT [29].

Su et al. (2020 & 2021) [32,33] We now discuss two research papers that are related to each other and share two common authors. The first paper investigates the scope of the BERT and BioBERT model in general BLM problems. The second paper improves on the result of the first one by improving the performance of pre-trained BERT model by using a pre-training step involving contrastive learning. Both papers use very similar study designs. The effectiveness of the models is demonstrated by applying them to three types of RE tasks from the biomedical domain: chemical-protein (ChemProt, using BioGRID database), drug-drug interactions (DDI, using DrugBank database) and protein-protein interactions (PPI, using IntAct database). The PPI classification task is considered a binary classification, indicating that the authors refrain from a more function-oriented classification as explored by Elangovan *et al.*, whereas the ChemProt and BioGRID classification tasks are multi-label classification tasks with five and four annotated interaction types in the respective databases [32,33].

In the first paper, Su *et al.* (2020) propose some new fine-tuning mechanisms for the BERT model. They point out that the RE problems are posed as classification problems, and pre-trained models like BERT rely on a specific [CLS] token from the last layer to make decisions. "The [CLS] token is used to predict the next sentence (NSP task) during the pre-training, which usually involves two or more sentences, but the inputs of our relation extraction tasks only contain one sentence. This indicates that the [CLS] output might ignore important information about the entities and their interaction because it is not trained to capture this kind of information [32]". As a solution to this, the authors propose to add a new module that can summarise all outputs from the last layer and concatenate that information with the [CLS] output as an extra fine-tuning step. The authors have experimented with the choice of the new module used to summarise information using LSTM and additive attention [32].

In the second paper, Su *et al.* (2021) propose a contrastive learning based approach to improve performance of pre-trained models. The term contrastive learning is used for a family of methods to construct a discriminative model comparing pairs of inputs. The training process for such models is designed such that similar input instances have "positive" labels whereas, dissimilar input instances are labelled as "negative" instances. The goal is to learn a text representation by maximising the agreement between inputs from positive pairs via a contrastive loss in the latent space, and the learned representation can then be used for relation extraction. The authors point out the lack of exploitation of the potential of such contrastive models for text data in general and RE problems from biomedical natural language processing specifically. The reason behind this, as explained by the authors, is that it is more challenging to design a general and efficient data augmentation method to construct positive and negative pairs necessary to train such models [33].

Moreover, in Su *et al.* (2021), the authors propose a new metric, "prediction shift", to measure the sensitivity degree to which the small changes of the inputs will make the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2021 doi:10.20944/preprints202110.0184.v1

model change its prediction, thereby arguing that the proposed model is more robust compared to simply using BERT for classification of interaction words [33].

To generate a positive pair of samples compatible to the training design of the contrastive model, the authors resort to simplistic data augmentation techniques. The goal is to slightly alter the original sentence using methods such as synonym replacement, random swap of words, or random deletion of words. Given a sentence s, two entity mentions (chemical or gene names) e_1 and e_2 in s and a relation type r also mentioned in s, the authors hypothesise that the shortest dependency path (SDP) between the two entity mentions (e_1 and e_2) in the sentence s, captures the required information to assert the relationship of the two entities. Keeping the SDP fixed, the authors therefore alter the rest of the word tokens in the text to generate augmented data, to ultimately generate positive samples. The hypothesis related to SDP is not novel in itself and has been explained in related research articles "If entities e_1 and e_2 are arguments of the same predicate, then the shortest path between them will pass through the predicate, which may be connected directly to the two entities, or indirectly through prepositions." Given a training batch of N sentences, the authors create an alternative "view" of each sentence (making a pool of $2N$ sentences) and then for every sentence s, they consider $< s, s' >$ as a positive pair. The other $2N - 1$ sentences are considered to be a negative sample, each compared to the sentence s [33].

The general architecture of the model is fairly similar to the general structure of Siamese neural networks. Training samples (sentences) are fed into the neural network in pairs (labelled positive or negative), each input sentence in the pair, goes through two independent channels of identical architecture. The final output is then generated by combining the outputs from these two independent channels, which is used to calculate the loss, which is optimised to be less for similar sentences (positive pairs). Each independent channel has a neural network encoder used to create encoding for the input sentences corresponding to the channel, and a projection head (a multi-layered perceptron) to transform the encoding to a desired dimension, which is known to improve the representation quality during training [33].

Wang et al. **(2020)** [30] RE among proteins is affected by mutations, implying that interactions among proteins may vary from one study to another depending on these mutations as well as the context of the study. To this end, the Biocreative VI challenge consists of two subtasks.

- Identifying documents describing mutations affecting PPI.
- Extracting relevant PPI through RE.

The first task, also referred to as Document Triage by the authors as Document Triage, clearly improves the practicality of using NLU based models for RE in the context of PPI. The second task can extract interacting protein pairs from documents containing a triage. The term "Triage", refers to a tuple of a source protein, a target protein, and their relevant interactions. Although RE is the main task addressed in this research article, the authors argue that the introduction of auxiliary tasks, such as Document Triage classification (whether a document describes genetic mutations affecting protein-protein interactions) and gene recognition task (NER), significantly improves the RE task [30].

The experiments for triage and RE tasks are performed on the BioCreative VI Track 4 corpus, containing 4082 articles in the training set, of which 1729 are relevant to PPI interactions involving mutations. Standard preprocessing approaches such as replacing mentions of gene names by predefined strings are employed [30].

The architecture of the model is compatible with the multitask (main and auxiliary tasks) learning strategy as proposed by the authors. For creating meaningful vector representation of the input text, the authors use BERT and BioBERT models. The BERT layer is shared as an embedding layer for all downstream layers. For the main RC task and auxiliary Document Triage task, a downstream Text CNN model is added to the model. Independent BiLSTM layers are used as a downstream layer for the gene

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2021 doi:10.20944/preprints202110.0184.v1

recognition auxiliary task. The authors argue that introduction of the auxiliary learning tasks improves the classification performance of the main RE task [30].

Zhou et al. **(2019)** [31] In this research work, the authors propose the Knowledge-aware Attention Network (KAN) for PPI extraction. The motivation of this work, published in 2019, is the fact that pre-existing methods needed extensive feature engineering and could not make full use of the prior knowledge available in the form of knowledge bases [31].

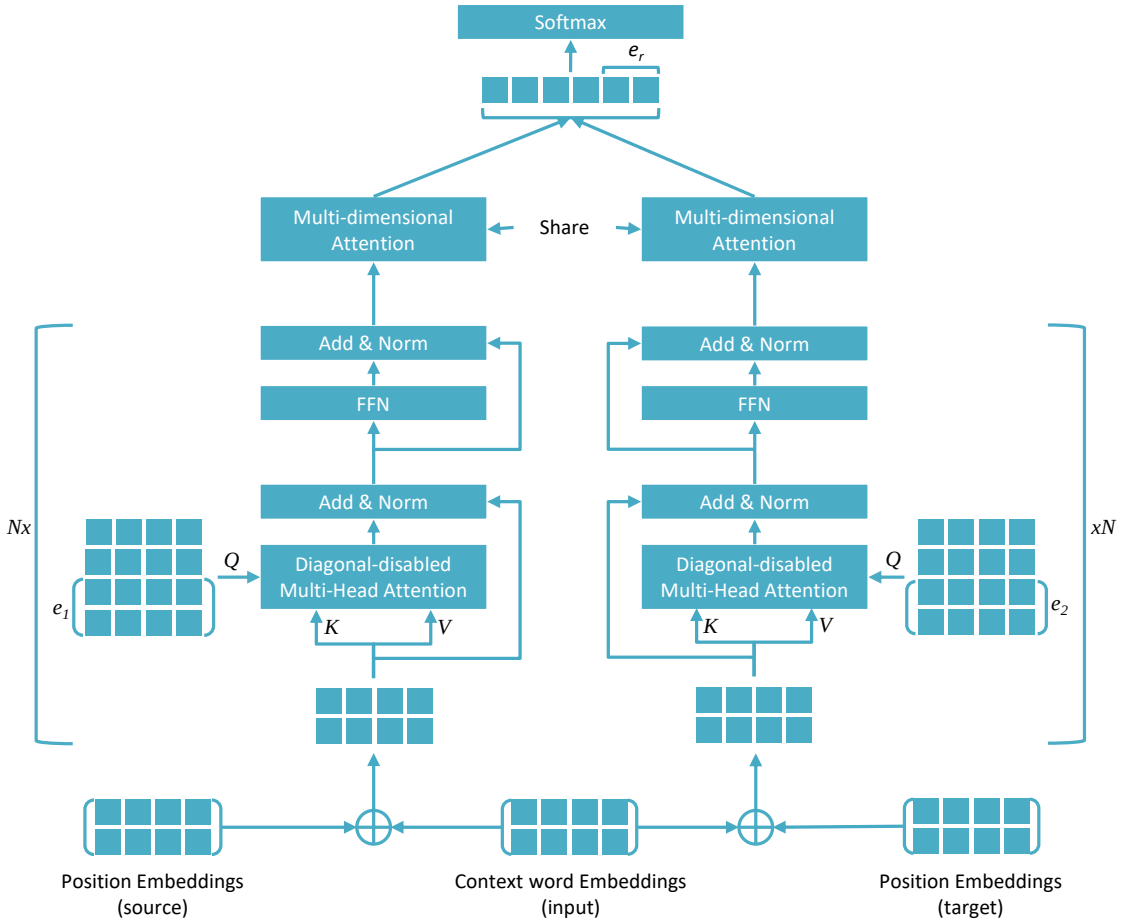

Figure 4. Architecture of Knowledge-aware Attention Networks (KAN). The symbol $N\times$ beside the marked blocks represents N copies of the respective blocks where, N can be defined by a modeller. The architecture has two parallel input channels, taking information on the source and target entity as input along with the relevant text. External knowledge is integrated into the model in the form of entity specific representations e_1 and e_2 and a representation of their known relationship e_r.

Experiments with the model are conducted on the BioCreative VI Track 4 PPI extraction task corpus. PPI relation triplets are extracted from two knowledge bases, IntAct and BioGRID, both of which contain 45 relation types. A total of 1,518,592 triples and 84,819 protein entities are obtained for knowledge representation training, i.e. they are fed as prior knowledge to the model during training. Like other approaches, the KAN model has elaborate preprocessing protocols. Some assumptions adopted during the preprocessing seem to be rather strict. For example, the authors write "To reduce the number of inappropriate instances, the sentence distance between a protein pair should be less than three." In addition to this, other general protocols such as replacing gene/protein names and context-specific words (interactions) by pre-defined strings are also employed [31].

As far as the model architecture is concerned, KAN is innovative. A schematic representation of the model is shown in Figure 4. KAN has two architectural components

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 12 October 2021 doi:10.20944/preprints202110.0184.v1

that are identical in structure, one for processing information relevant to a source protein and the other for processing information relevant to the corresponding target protein, given a source-target protein pair in a sentence. The information on the positions of the source and target proteins is encoded along with the sentence while the input is fed into the model. This is done by modifying the general idea of positional embedding that is employed in self-attention based model in general. In this case, position encoding is encoded with respect to the positions of the source and target proteins in a given sentence. Respective positional encodings for the source and target proteins are fed into the respective architectural components along with the encoded sentence. Next, these inputs are passed through a diagonal-disabled multi-headed attention layer in each architectural component. Generally, self attention based processes are represented as some mathematical operations among a query (Q), Key (K) and Value (V) vector. The same vector (vectorised form of the word sequence in a sentence added to the positional encoding) is considered for Q, K and V. The multiplication of K and V produces a square attention matrix, which is then multiplied by Q. In the KAN model, however, the authors use a different form of Q. As the model aims to exploit the entity-relation triplets recorded in triplets as prior knowledge, TransE (a typical knowledge representation approach, which represents the relation between two entities as a translation in a representation space) is used to create vector representations of these triplets. The vector representations of the source and target (e_1 and e_2) proteins are used as a part of Q in the respective architectural components. After passing the outputs of the attention layers through a feed-forward network and a multi-dimensional attention layer in each architectural component, the outputs from two architectural components are concatenated to obtain the final feature representation. At this stage, the vector representation of the relation between the source and the target proteins (e_r) is also concatenated with the vector representation of the proteins to take advantage of prior knowledge. The results from this layer are then passed through a softmax activation to obtain the final outputs for the classification task. The authors also experimented with several variations of the KAN model [31].

3.3. Performance comparison among the discussed models

The models we have discussed in the previous section are designed to perform diverse tasks varying from document triage finding to RE problems (PPI, DDI, ChemProt etc.) to detection of "typed" PPI. Moreover, they operate on different datasets and have different preprocessing approaches involved. In addition to these factors, they are also often evaluated on different performance measures. It is therefore difficult, if not impossible, to come up with a fair way of comparing their performances. However, one can still observe patterns in the results which can be of significance.

The results of Wang, *et al.* and Zhou *et al.* are comparable. Both of them address the same dataset, that is, BioCreative VI dataset. An evaluation criterion called "Exact Match evaluation" is also similar in both cases. It is defined as: "A predicted relation only counts when the GeneIDs are the same as human-annotated GeneIDs." In this regard, Wang *et al.* with a F1-Score of 43.14 clearly outperforms the KAN model by Zhou *et al.*, with a F1-Score of 38.23, which confirms that learning auxiliary tasks along with the principal task could play a role in improving model performances. It is also noteworthy that the preprocessing protocols of Wang *et al.* are comparatively simpler. Although these two papers deal with extraction of interacting protein pairs from documents, they do not emphasise on specifying the type of interaction. Knowing the type of interaction can be extremely useful while creating a large-scale disease map such as the Atlas of Inflammation Resolution [37] or the Parkinsons Disease Map [48].

Elangovan *et al.* on the other hand, address the type of interaction among protein pairs. Their "typed" classification problem deals with seven different kinds of interactions: phosphorylation, dephosphorylation, methylation, demethylation, ubiquitination, deubiquitination, and acetylation. Thus, their model is a multi-class classification model

 doi:10.20944/preprints202110.0184.v1

PPI-BioBERT, which produced an F1-Score of 35.4 %. Interestingly, an ensemble of 10 PPI-BioBERT models improve the F1-Score to 54 %, which shows that there is a scope of ensemble-based models in RE based problems [28].

Su *et al.* (2020 & 2021) have both explored the same three BLM based problems, PPI, DDI, and ChemProt. The first article written in 2020 compares several variations of the BioBERT model (such as using LSTM and attention layers and utilising the classification token [CLS] in their model). Using their models, they achieved F1-Scores of 82.8, 80.7, and 76.8 for PPI, DDI and ChemProt tasks, respectively. However, from the results, it is difficult to conclude which kind of architecture (using LSTM or Attention layers) in particular is beneficial. Perhaps that is why, in the follow-up article in 2021, the authors use none of the ideas of using attention or LSTM layers. Rather, it focuses on contrastive learning. The best results were produced by a model that used contrastive learning in addition to adding information from external knowledge bases. In this case, they achieved F1-Scores of 82.7, 82.4, and 76.9 for PPI, DDI and ChemProt tasks, respectively, which is not a significant improvement on the previous work [32,33].

4. Discussion

The classification approach popularly used for RE hinders the transfer of knowledge across databases and corresponding datasets. This is due to the different annotations used in different databases. Knowledge transfer and integration across databases, in accordance to the classification approach, therefore requires the amalgamation of corresponding datasets with different annotations. This can create a multi-label classification problem with multiple classes, making a model difficult to train. Intrinsic imbalance persistent in such datasets along with ambiguous manual annotations across datasets make it even harder to effectively train classification models. This reduces the practical usability of the classification models as a modeller cannot customise these models as per their modelling needs, and they have to rely on pre-annotated databases to train their models.

Moreover, modelling directed interactions requires knowledge on the source entity, the target-entity and the relationship among them. Usually, this used to be designed as a relationship triplet finding problem. Most classification models for RE, do not consider the sense of directionality that is associated with the related entities. Only a few models, like KAN, consider taking the source and target entities as a part of the input and tries to predict the corresponding interaction. But still, KAN is not trained in such a way such that it can differentiate between a source and a target entity in case of a directed interaction.

Even if such a classification model exists which can differentiate between a source and target entity of a directed interaction, while practically using such a model to extract relationships from new data, a modeller has to know from a relevant text, which entity is the source and which entity is the target. Without knowing this information, directed interactions cannot be modelled. While, NER based models can identify the entity names from new literature, the issue of annotating source and target entities among them is not addressed in the discussed approaches, in general. This again hinders the practical applicability of such models in knowledge extraction.

What makes the practical use of many of these models difficult are the diverse preprocessing protocols and strict assumptions adopted by the models. Almost every model that we discussed replaces protein or chemical names by specific strings (e.g. Su *et al.*, Wang *et al.*, Elangoven *et al.*, Zhou *et al.*) [28,30–33]. The model by Zhou *et al.*, for example, adopted elaborate protocols while curating training data such as [31]:

- Reducing the number of inappropriate instances, the sentence distance between a protein pair is assumed to be less than three;
- Selecting the words between a protein pair and three expansion words on both sides as the context word sequence with respect to the protein pair;
- Removing protein names from input string;

 doi:10.20944/preprints202110.0184.v1

- Replacing numeric entries by predefined strings;

For attention-based models, even though there have been some attempts of making models explainable by observing the attention matrices, such attempts are rare in the case of BLM. For example, Su *et al.* (2020) and Zhou *et al.* have made some limited efforts to explain the behaviour of their models [31,32].

In the field of computer vision, the concept of explainable models is quite popular. Being able to explain decisions made by a model can be important in the case of BLM as well. A biproduct of explainability could be, for example, a knowledge graph, which is a compact way of summarising a lot of information as well as discovering new information [47]. Biological information can be represented in its most general form as knowledge graphs. A model which can be used to curate and represent from new literature entities such as genes, proteins, phenotypes, etc., and their relationships in the form of knowledge graphs can address some issues of the classification approach discussed before. The nodes of the knowledge graph represent the entities, and edges are annotations of directed or undirected relations among the entities. Customised edge annotations, as per the interest of the modeller, can be fed into a model, making the modeller adaptable to the need of the modeller. Given the positional information on a word representing an edge annotation (e.g. activation, repression, phosphorylation) in a sentence, self-attention-based models can be used to predict the positions of the source and target node entities (e.g. source gene, target gene) for that particular edge annotation. In case a modeller is not interested to model inter-entity relationships in particular and is simply interested in modelling whether there is an association between two entities (gene-phenotype association), such a model can account for this by learning the position of the target entity, given the position of the source entity or vice-versa.

A knowledge-graph based model, as discussed above, could be used in a pipeline with other NLP tasks to develop an end-to-end approach for customised knowledge extraction and knowledge discovery. For example, NER and Document triage can be used as preceding tasks in a pipeline. Discovery of relationships among new entities can be achieved through models operating on knowledge graphs generated by the model. For example, Liu *et al.* proposed a model for the discovery of new relationships between compounds and diseases from knowledge graphs using a reinforcement learning approach on knowledge graphs [47].

5. Conclusion

Clearly, attention-based models, both novel architectures and pre-trained networks, are being explored widely in the domain of BLM. Complex algorithms have been constructed to handle a wide variety of tasks such as NER, RE, document classification, triage mining. Some publications have proposed coherent workflows attempting to make the algorithms more practically usable. However, challenges like diversely annotated datasets, transfer of knowledge for trained models across datasets, lack of explainability, complex preprocessing protocols, and the large amount of computational power required to tune pre-trained models reveal the scope of further research in this domain with a goal of a more generalistic and practically useful approach.

Author Contributions: P.S. and S.B. designed, drafted and revised the manuscript; K.Y. and O.W. critically discussed and revised the content of the manuscript. All the authors have read and agreed to the contents of the manuscript.

References

1. Kotu V., Deshpande B., Chapter 9 - Text Mining, Editor(s): Vijay Kotu, Bala Deshpande, Data Science (Second Edition), Morgan Kaufmann, 2019, Pages 281-305, ISBN 9780128147610, https://doi.org/10.1016/B978-0-12-814761-0.00009-5
2. Szklarczyk, D., Gable, A.L., Nastou, K.C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N.T., Legeay, M., Fang, T., Bork, P., Jensen, L., & Mering, C.V. (2021). The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, 49, D605 - D612, https://doi.org/10.1093/nar/gkaa1074

3. Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., & Houlsby N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR 2021*, https://arxiv.org/pdf/2010.11929.pdf
4. Tran D. T., Iosifidis A., Kanniainen J. & Gabbouj M., Temporal Attention-Augmented Bilinear Network for Financial Time-Series Data Analysis, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 5, pp. 1407-1418, May 2019, https://doi.org/10.1109/TNNLS.2018.2869225.
5. Choi J., Lee B., & Zhang B. (2017). Multi-focus Attention Network for Efficient Deep Reinforcement Learning. *ArXiv*, https://arxiv.org/ftp/arxiv/papers/1712/1712.04603.pdf.
6. Luong T., Pham H., Manning C.D., Effective Approaches to Attention-based Neural Machine Translation, 2015 *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1412–1421, https://aclanthology.org/D15-1166.pdf
7. Hermann K.M., Kočiský T., Grefenstette E., Espeholt L., Kay W., Suleyman M., and Blunsom P. 2015. Teaching machines to read and comprehend. *In Proceedings of the 28th International Conference on Neural Information Processing Systems*, Volume 1 (NIPS'15). MIT Press, Cambridge, MA, USA, 1693–1701, https://proceedings.neurips.cc/paper/2015/file/afdec7005cc9f14302cd0474fd0f3c96-Paper.pdf
8. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E. (2016). Hierarchical Attention Networks for Document Classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, 1480-1489, https://aclanthology.org/N16-1174.pdf
9. Devlin, J., Ming-Wei C., Kenton L. and Kristina T.(2019), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT Vol 1, 4171-4186, https://aclanthology.org/N19-1423.pdf
10. Ambartsoumian, A. and Popowich, F. (2018), Self-Attention: A Better Building Block for Sentiment Analysis Neural Network Classifiers, *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Association for Computational Linguistics, pp. 130-139, https://aclanthology.org/W18-6219.pdf
11. Cho K., van Merrienboer B., Gülçehre C., Bahdanau D., Bougares F., Schwenk H., Bengio Y. (2014), Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724-1734, https://aclanthology.org/D14-1179.pdf
12. Cho K., van Merrienboer B., Bahdanau D., Bengio Y. (2014), Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pp. 103-111, https://aclanthology.org/W14-4012.pdf
13. Chaudhari, S., Polatkan, G., Ramanath, R., & Mithal, V. (2019). An Attentive Survey of Attention Models. *ArXiv*, https://arxiv.org/pdf/1904.02874.pdf
14. Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. *31st Conference on Neural Information Processing Systems* pp. 6000–6010, https://papers.nips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
15. Radford, A., & Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training. https://openai.com/blog/language-unsupervised/
16. Dehghani M., Azarbonyad H., Kamps J., and de Rijke M. 2019. Learning to Transform, Combine, and Reason in Open-Domain Question Answering. *In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19)*. Association for Computing Machinery, New York, NY, USA, 681–689. https://doi.org/10.1145/3289600.3291012
17. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. 2015 *IEEE International Conference on Computer Vision (ICCV)*, 19-27. https://doi.org/10.1109/ICCV.2015.11
18. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S.R. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353–355, https://aclanthology.org/W18-5446.pdf
19. Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100, 000+ Questions for Machine Comprehension of Text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392, https://aclanthology.org/D16-1264.pdf.
20. Zellers, R., Bisk, Y., Schwartz, R., & Choi, Y. (2018). SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 93–104, https://aclanthology.org/D18-1009.pdf
21. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J.R., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., & Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv*, https://arxiv.org/pdf/1609.08144.pdf
22. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *International Conference on Learning Representations 2020*, https://arxiv.org/pdf/1909.11942.pdf
23. Nguyen, D.Q., Vu, T., & Nguyen, A. (2020). BERTweet: A pre-trained language model for English Tweets.*Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Association for Computational Linguistics,9-14, https://aclanthology.org/2020.emnlp-demos.2.pdf

24. Martin, L., Muller, B., Ortiz S. J. P., Dupont, Y., Romary, L., De la Clergerie, E., Seddah, D. & Sagot, B. (2019). CamemBERT: a Tasty French Language Model. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7203–7219, https://aclanthology.org/2020.acl-main.645.pdf
25. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*, https://arxiv.org/pdf/1907.11692.pdf.
26. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36, 1234 - 1240. https://doi.org/10.1093/bioinformatics/btz682
27. Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A Pretrained Language Model for Scientific Text. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3615–3620, https://doi.org/10.18653/v1/D19-1371
28. Elangovan, A., Davis, M.J., & Verspoor, K. (2020). Assigning function to protein-protein interactions: a weakly supervised BioBERT based approach using PubMed abstracts. *ArXiv*, https://arxiv.org/ftp/arxiv/papers/2008/2008.08727.pdf
29. Giles, O., Karlsson, A., Masiala, S., White, S., Cesareni, G., Perfetto, L., Mullen, J., Hughes, M., Harland, L., & Malone, J. (2020). Optimising biomedical relationship extraction with BioBERT. *bioRxiv*, https://www.biorxiv.org/content/10.1101/2020.09.01.277277v1.full
30. Y. Wang, S. Zhang, Y. Zhang, J. Wang and H. Lin, "Extracting Protein-Protein Interactions Affected by Mutations via Auxiliary Task and Domain Pre-trained Model," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Seoul, Korea (South)*, 2020 pp. 495-498. https://doi.ieeecomputersociety.org/10.1109/BIBM49941.2020.9313120
31. Huiwei Zhou, Zhuang Liu, Shixian Ning, Chengkun Lang, Yingyu Lin, Lei Du, Knowledge-aware attention network for protein-protein interaction extraction, *Journal of Biomedical Informatics*, Volume 96, 2019, 103234, ISSN 1532-0464, https://doi.org/10.1016/j.jbi.2019.103234
32. Su, P., & Vijay-Shanker, K. (2020). Investigation of BERT Model on Biomedical Relation Extraction Based on Revised Fine-tuning Mechanism. 2020 *IEEE International Conference on Bioinformatics and Biomedicine*, 2522-2529. https://arxiv.org/pdf/2011.00398.pdf
33. Su, P., Peng Y., & Vijay-Shanker, K. (2021).Improving BERT Model Using Contrastive Learning for Biomedical Relation Extraction. 2021 *ArXiv*, https://arxiv.org/pdf/2104.13913.pdf.
34. Zhang, Y., Lin, H., Yang, Z., Wang, J., Sun, Y., Xu, B., & Zhao, Z. (2019). Neural network-based approaches for biomedical relation classification: A review. *Journal of Biomedical Informatics*, 103294 https://doi.org/10.1016/j.jbi.2019.103294
35. Zhao, S., Su, C., Lu, Z., & Wang, F. (2021). Recent advances in biomedical literature mining. *Briefings in Bioinformatics*. 22(3), pp. 1-19, https://doi.org/10.1093/bib/bbaa057
36. Papanikolaou, N., Pavlopoulos, G., Theodosiou, T., & Iliopoulos, I. (2015). Protein-protein interaction predictions using text mining methods. *Methods*, 74, 47-53 https://doi.org/10.1016/j.ymeth.2014.10.026
37. Serhan, C., Gupta, S., Perretti, M., Godson, C., Brennan, E., Li, Y., Soehnlein, O., Shimizu, T., Werz, O., Chiurchiù, V., Azzi, A., Dubourdeau, M., Gupta, S., Schopohl, P., Hoch, M., Gjorgevikj, D., Khan, F., Brauer, D., Tripathi, A., Cesnulevicius, K., Lescheid, D., Schultz, M., Särndahl, E., Repsilber, D., Kruse, R., Sala, A., Haeggström, J., Levy, B., Filep, J., & Wolkenhauer, O. (2020). The Atlas of Inflammation Resolution (AIR). *Molecular aspects of medicine*, vol 74, pp 47-53, https://doi.org/10.1016/j.mam.2020.100894
38. Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., & Cesareni, G. (2007). MINT: a Molecular INTeraction database. *FEBS Letters*, 513(1), pp 135-140 https://doi.org/10.1016/S0014-5793(01)03293-8
39. Bader, G.D., Donaldson, I., Wolting, C., Ouellette, B., Pawson, T., & Hogue, C. (2001). BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Research*, 31 1, 248-50, https://doi.org/10.1093/nar/gkg056
40. Oughtred, R., Rust, J.M., Chang, C.S., Breitkreutz, B., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F., Dolma, S., Coulombe-Huntington, J., Chatr-aryamontri, A., Dolinski, K., & Tyers, M. (2020). The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Science : A Publication of the Protein Society*, 30, 187 - 200 https://doi.org/10.1002/pro.3978
41. Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S.K., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., & Apweiler, R. (2004). IntAct: an open source molecular interaction database. *Nucleic Acids Research*, 32 Database issue, D452-5, https://doi.org/10.1093/nar/gkh052
42. Han, H., Shim, H., Shin, D., Shim, J., Ko, Y., Shin, J., Kim, H., Cho, A., Kim, E., Lee, T., Kim, H., Kim, K., Yang, S., Bae, D., Yun, A., Kim, S., Kim, C.Y., Cho, H.J., Kang, B., Shin, S., & Lee, I. (2015). TRRUST: a reference database of human transcriptional regulatory interactions. *Scientific Reports*, Article number: 11432, https://doi.org/10.1038/srep11432
43. Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. *Arxiv*, https://arxiv.org/pdf/1409.0473.pdf
44. Wishart, D., Knox, C., Guo, A., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z., & Woolsey, J. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Research*, 34, D668 - D672.
45. Taboureau, O., Nielsen, S.K., Audouze, K., Weinhold, N., Edsgärd, D., Roque, F.S., Kouskoumvekaki, I., Bora, A., Curpan, R., Jensen, T., Brunak, S., & Oprea, T. (2011). ChemProt: a disease chemical biology database. *Nucleic Acids Research*, 39, D367 - D372.
46. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J.R., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., & Dean, J. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *ArXiv*, https://arxiv.org/pdf/1609.08144.pdf

 doi:10.20944/preprints202110.0184.v1

47. Liu, Y., Hildebrandt, M., Joblin, M., Ringsquandl, M., Raissouni, R., & Tresp, V. (2021). Neural Multi-Hop Reasoning with Logical Rules on Biomedical Knowledge Graphs, The Semantic Web (page 375-391) 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Lecture Notes in Computer Science book series (LNCS, volume 4825), Springer.
48. Fujita, K.A., Ostaszewski, M., Matsuoka, Y., Ghosh, S., Glaab, E., Trefois, C., Crespo, I., Perumal, T.M., Jurkowski, W., Antony, P.M., Diederich, N., Buttini, M., Kodama, A., Satagopam, V.P., Eifes, S., del Sol, A., Schneider, R., Kitano, H., & Balling, R. (2013). Integrating Pathways of Parkinson's Disease in a Molecular Interaction Map. Molecular Neurobiology, 49, 88 - 102.