Text mining-based word representations for biomedical data analysis and machine learning tasks

Halima Alachram^{1,*}, Hryhorii Chereda¹, Tim Beißbarth¹, Edgar Wingender², Philip Stegmaier²

¹ Department of Medical Bioinformatics, University Medical Center, Göttingen, Lower Saxony, Germany

² geneXplain GmbH, Lower Saxony, Wolfenbüttel, Germany

* halima.alachram@bioinf.med.uni-goettingen.de

Abstract

Biomedical and life science literature is an essential way to publish experimental results. With the rapid growth of the number of new publications, the amount of scientific knowledge represented in free text is increasing remarkably. There has been much interest in developing techniques that can extract this knowledge and make it accessible to aid scientists in discovering new relationships between biological entities and answering biological questions. Making use of the *word2vec* approach, we generated word vector representations based on a corpus consisting of over 16 million PubMed abstracts. We developed a text mining pipeline to produce word2vec embeddings with different properties and performed validation experiments to assess their utility for biomedical analysis. An important pre-processing step consisted in the substitution of synonymous terms by their preferred terms in biomedical databases. Furthermore, we extracted

gene-gene networks from two embedding versions and used them as prior knowledge to train Graph-Convolutional Neural Networks (CNNs) on breast cancer gene expression data to predict the occurrence of metastatic events. Performances of resulting models were compared to Graph-CNNs trained with protein-protein interaction (PPI) networks or with networks derived using other word embedding algorithms. We also assessed the effect of corpus size on the variability of word representations. Finally, we created a web service with a graphical and a RESTful interface to extract and explore relations between biomedical terms using annotated embeddings. Comparisons to biological databases showed that relations between entities such as known PPIs, signaling pathways and cellular functions, or narrower disease ontology groups correlated with higher cosine similarity. Graph-CNNs trained with word2vec-embedding-derived networks performed best for the metastatic event prediction task compared to other networks. Word representations as produced by text mining algorithms like word2vec, therefore capture biologically meaningful relations between entities.

1 Introduction

The field of Natural Language Processing (NLP) is concerned with the development of methods and algorithms to computationally analyze and process human natural language. Solutions in this domain often have practical significance for everyday applications such as conversion between written and spoken language to enhance media accessibility, translation between languages, optical character recognition (OCR) for street sign detection in traffic assistance systems, or document/media content classification for recommendation systems. In biomedical research, NLP is of importance to extract reported findings, e.g., protein-protein interactions, from scientific texts. Several studies employed supervised machine learning algorithms to identify and extract

knowledge from scientific literature [1-4], which requires extensive manually labeled datasets for training.

A novel approach was recently introduced that applied neural networks (NNs) to learn high-dimensional vector representations of words in a text corpus that preserve their syntactic and semantic relationships [5]. The word2vec method, proposed by Mikolov et al. [5], embedded words in a vector space by predicting their co-occurrence so that words with similar meaning had a similar numerical representation. Word2vec can effectively cluster similar words together and predict semantic relationships. As produced by word2vec, word embedding allows computing relations between words obtained from a large unlabeled corpus, e.g., using their vector cosine similarity.

Several works in the biomedical research field have since adopted word vector presentations for various tasks like named entity recognition (NER) [6-7], medical synonym extraction [8], as well as extraction of chemical-disease relations [9], drug-drug interactions [10] or protein-protein interactions [11]. Different validation strategies have been proposed to assess the quality of word embeddings. Wang et al. evaluated word embeddings' performance generated from four different corpora, namely clinical notes, biomedical literature, Wikipedia, and news articles [12]. The evaluation was performed qualitatively and quantitatively. Their experimental results showed that embeddings trained on clinical notes are closer to human judgments of word similarity. They also demonstrated that word embeddings trained on general domain corpora are not substantially inferior in performance than those trained on biomedical or clinical domain documents. Zhang et al. assessed both the validity and utility of biomedical word embeddings generated using a sub-word information technique that combines unlabeled biomedical literature with domain knowledge in Medical Subject Headings (MeSH) [13]. They evaluated the effectiveness of their word

embeddings in BioNLP tasks, namely a sentence pair similarity task performed on clinical texts and biomedical relation extraction tasks. Their word embeddings have led to a better performance than the state-of-the-art word embeddings in all BioNLP tasks.

Several recent studies, particularly in molecular biology, have also considered word embeddings to represent biomedical entities and their functional relationships. Du et al. trained a gene embedding from human genes using gene co-expression patterns in data sets from the GEO databases and achieved an area under the curve score (AUC) of 0.72 in a gene-gene interaction prediction task [14]. Chen et al. employed NER tools to recognize and normalize biomedical concepts in a corpus consisting of PubMed abstracts. They trained four concept embeddings on the normalized corpus using different machine learning models. They assessed the concept embeddings' performance in both intrinsic evaluations on drug-gene interactions and gene-gene interactions and extrinsic evaluations on protein-protein interaction prediction and drug-drug interaction extraction. Their concept embeddings achieved better performance than existing methods in all tasks [15].

In other studies, word embeddings were used as input features to improve machine learning algorithms' performance. Kilimci et al. used different document representations, including term frequency-inverse document frequency (TF-IDF) weighted document-term matrix, mean of word embeddings, and TF-IDF weighted document matrix enhanced with the addition of mean vectors of word embeddings as features. They analyzed the classification accuracy of the different document representations by employing an ensemble of classifiers on eight different datasets. They demonstrated that the use of word embeddings improved the classification performance of texts [16].

Evaluating the validity of word embeddings by examining semantic relations can improve the transparency of word embeddings and facilitate the interpretation of the downstream applications using them. In this study, we applied word2vec to generate biomedical word embeddings and thoroughly validated their ability to capture biologically meaningful relations. The generated embeddings encompass biomedical terms for which synonyms have been replaced with their preferred terms before training. Our validation methods are similar to those used by Chen et al. [15]. However, their evaluations concentrated on genes by considering drug-gene and gene-gene interactions. Our assessment covers vector cosine similarity of relations in protein-protein interactions (PPIs), common pathways and cellular functions, or narrower disease ontology groups using existing knowledge in biomedical databases.

Most word embeddings have been trained using either the word2vec [5] or the GloVe [17] model, which uses information about each word's co-occurrence with its nearby words to represent it in a distinct vector. Word2vec employs negative sampling and sub-sampling techniques to reduce the computational complexity. Word embeddings, learned by word2vec or other methods such as Skip-Gram or Continuous Bag-of-Words that predict a word's context from raw text using a target word, are called static embeddings. Such static embeddings are useful for solving lexical semantic tasks, particularly word similarity and word analogy, and representing inputs in downstream tasks [18]. More recent techniques in language modeling were unsupervised pre-trained language models such as BERT (Bidirectional Encoder Representations from Transformers) [19] and ELMO (Embeddings from Language Models) [20] that create contextualized word representations. Such models support fine-tuning on specific tasks and have shown effective performance improvements in diverse NLP tasks such as question answering and text classification. BioBERT [21] (Bidirectional Encoder Representations from Transformers for

Biomedical Text Mining) was initialized with the BERT model to pre-train domain-specific language representations on large-scale biomedical articles. BERT outperformed the NLP state-of-the-art on a variety of biomedical and clinical tasks. However, training BERT is computationally expensive due to its high model complexity and a large amount of training data needed to achieve acceptable accuracy.

Moreover, convolutional neural networks (CNNs) are among the often applied deep-learning network architecture that have delivered a good performance in image recognition and classification. CNN models were also effective for NLP tasks such as text classification [22-23]. They have been used in bioinformatics [24], namely in drug discovery and genomics [25], and motivated further progress on graph structured prior information with promising results on the prediction of metastatic events [26-27].

Our word embeddings were assessed for their biological utility on a metastatic event prediction task. We trained Graph-CNNs on breast cancer gene expression data with gene-gene networks derived from word2vec embedding as prior knowledge to predict the occurrence of metastatic events. Graph-CNNs achieved the best performance with word2vec-embedding-derived networks compared to protein-protein interaction networks or networks derived using other word embedding algorithms.

2 Materials and methods

Word embeddings generated for this study were based on a corpus of 16,558,093 article abstracts from the public PubMed FTP repository.

2.1. Corpus pre-processing and training of word embeddings

Natural language text data usually require an amount of preparation before they can be fed into the model training. For the purpose of pre-processing of the text corpus we implemented a pipeline that conducted the steps depicted in Fig 1. The first phase applied classical processing steps such as lowercasing, lemmatization, and removal of punctuation and numerical forms. We assumed replacing synonyms of biomedical terms with their main terms can affect the similarity between words in a way to better capture functional relationships between biomedical entities. Thus, we introduced an optional step before starting training, in which synonymous terms were substituted by externally defined main terms. This corpus was then used to train the word2vec model. For training we used the word2vec implementation in Gensim [28] with context window of size 5, minimum count is equal to 5 and 300 neurons which is also the number of generated vector dimensions.

For this study, we generated the following two word2vec embeddings: ‘Embedding_v1’ in which synonymous terms of genes, diseases, and drugs were substituted by their preferred terms from HUGO [29], MeSH [30], and DrugBank [31], respectively; and ‘Embedding_v2’ for which the same preprocessing strategies and the same training process were applied but without replacing synonyms. Moreover, we assigned type labels using the same biomedical databases mentioned above in order to filter similarities. This enabled us to compare similarities between entities in obtained embeddings to existing knowledge in biomedical databases.

Fig 1. Applied workflow to generate word2vec embeddings. The process of generating word embeddings consists of two phases: a preprocessing and a training phase. Before training, an additional step conducts replacing synonymous terms with their main terms from biomedical databases.

2.2. Validation of word embeddings

To provide an expedient tool for biological research, relative locations of terms within the vector space of the embedding should exhibit agreement with existing biological knowledge. Our validation addressed protein-protein interactions, signaling pathways and biological processes, drug targets and human diseases, which have been and continue to be of interest in many biomedical research projects. The conducted validation experiments therefore examined whether vectors of members within groups defined by respective biological databases featured increased cosine similarities compared to randomly sampled entities.

2.2.1. Signaling pathways, biological processes and human diseases

Reactome 72 [32] and TRANSPATH® 2020.2 [33] pathway-gene assignments as well as Gene Ontology (GO, release 2020-03-25) [34] biological process-gene assignments were extracted from

the geneXplain platform [35] version 6.0. Disease terms covered by the embedding were mapped to 139 groups of the Human Disease Ontology version 2019-05-13 [36] with more than 5 and less than 1000 member diseases. We calculated medians, lower and upper quartiles of cosine similarities for gene pairs within pathways and biological processes with at least 10 and not more than 3000 genes as well as for disease pairs within the 139 disease groups. In addition, we calculated medians, lower and upper quartiles for 2000 randomly sampled gene pairs and for 700 randomly sampled disease pairs that were not contained in selected groups. (Files S1-S4)

2.2.2. Protein-protein interactions

Known protein-protein interactions were extracted from Reactome 63 for 4254 genes (16727 interactions) with vector presentations in the embedding. For the purpose of comparison, we sampled 10,000 random gene pairs and the same number of gene pairs with known interactions (S5 File).

2.2.3. Drug-gene associations

The DrugBank [31] database combines detailed drug information with comprehensive drug target information. We extracted genes associated with each drug reported in DrugBank with type target. By considering 5234 drugs and their target genes, we created drug pairs based on the common genes that two drugs share in each pair. Drug-gene associations were obtained from DrugBank release 4.5.0 and cosine similarities of 50000 drug pairs with at least one shared target gene were compared to 50000 drug pairs without common target genes. Moreover, to examine the variability of the similarity distribution of drug pairs based on the number of genes they share, we sampled three drug pair groups (group1: no genes, group2: ≤ 5 genes, group3: ≤ 9 genes) (S6 File).

2.3. Breast cancer data

Graph-CNNs were trained on a breast cancer data set compiled by Bayerlová et al. [37]. The data consisted of 10 microarray data sets measured on Affymetrix Human Genome HG-U133 Plus 2.0 and HG-U133A arrays which are publicly available from the Gene Expression Omnibus (GEO) [38] under accession numbers GSE25066, GSE20685, GSE19615, GSE17907, GSE16446, GSE17705, GSE2603, GSE11121, GSE7390, and GSE6532. The RMA (robust multi-array average) probe-summary algorithm [39] was applied to normalize each data set separately after which they were combined and further normalized using quantile normalization applied over all datasets. When more than one probe was associated with a gene, the probe with the highest average expression value was chosen, leading to 12179 genes. Training set classes consisted of 393 patients with metastasis within the first 5 years and 576 patients without metastasis between 5 and 10 years after biopsy.

2.4. PPI networks

A broad range of machine learning models have been developed to analyze high-throughput datasets in the aim of predicting gene interaction and identifying prognostic biological processes. Recently, biomedical research has shown the ability of deep learning models in learning arbitrarily complex relationships from heterogeneous data sets with existing integrated biological knowledge. This biological knowledge is often represented by interaction networks. The high data dimensionality and the complexity of biological interaction networks are significant analytical challenges for modeling of the underlying systems biology. In this section, we present the PPI networks derived from different sources and used as prior knowledge to structure gene expression data.

2.4.1. Human Protein Reference Database

In a recent study, *Chereda et al.* [26] employed the Human Protein Reference Database (HPRD) protein-protein interaction (PPI) [40] network to structure gene expression data of breast cancer patients. Genes from gene-expression data were mapped to the vertices of the PPI network yielding an undirected graph with 7168 matched vertices consisting of 207 connected components. The main connected component had 6888 vertices, whereas the other 206 components each contained 1 to 4 vertices. Since the approach of utilizing prior network information in Graph CNNs required a connected graph [41] training was carried out on the gene set of the main connected component. In this study, we used the same PPI network with the same approach of *Chereda et al.* [26].

2.4.2. Word2vec embedding-based networks

We created two gene-gene networks (Embedding_net_v1 and Embedding_net_v2) from the embedding version that excluded synonyms (Embedding_net_v1) and from the other version where word synonyms were taken into account (Embedding_net_v2). Both networks consisted of gene pairs with edges weighted by their cosine similarity values. The cosine similarity threshold was set to 0.65 yielding the Embedding_net_v1 network with 10730 genes in 4397 connected components with a main component of 6092 vertices and the Embedding_net_v2 network with 10729 genes in 4399 components with a main component covering 6106 vertices. The main connected components of Embedding_net_v1 and Embedding_net_v2 networks shared 5750 genes, therefore overlapping in the majority of vertices.

2.4.3. STRING-derived network

The STRING database [42] is a collection of protein-protein associations which can be derived from one or more sources such as gene neighborhoods, gene co-occurrence, co-expression, experiments, databases, text-mining, and whose confidence is expressed by an aggregated score computed from scores of the individual interaction sources. We considered the text-mining score as well as the combined score of all the interaction sources to build weighted protein-protein interaction networks. This way, the classification performance of Graph CNNs trained on the STRING text-mining network could be compared to Graph CNNs with prior knowledge from word2vec embedding-based networks. Like with the HPRD PPI, we mapped the genes to the two constructed STRING networks and supplied their main components to the training process. Score thresholds were chosen to obtain comparable number of vertices as in the HPRD PPI.

2.4.4. BERT-embedding-derived Network

BERT (Bidirectional Encoder Representations from Transformers) [19] is a recently contextualized word representation model. The main technical innovation of BERT is the use of bidirectional transformers. BERT was pre-trained in English Wikipedia and Books Corpus as a general language representation model. BioBERT [21] is a language representation model based on BERT and designed for biomedical text mining tasks. It was initialized with the BERT model provided by Devlin *et al.* in 2019 [19] and pre-trained on PubMed abstracts and PubMed Central full-text articles (PMC). We used the pre-trained BioBERT weights of ‘BioBERT-Base v1.0’ that was trained using the same vocabulary of BERT base (12-layer, 768-hidden, 12-heads, 110M parameters) on English text and 200k PubMed abstracts in addition. We converted the pre-trained TensorFlow checkpoint model to Pytorch [43], extracted the numerical vectors of 768 dimensions

each and calculated the cosine similarities between entities to eventually extract a gene-gene network. The number of proteins in the main connected component was also kept according to the comparable number of vertices in the HPRD PPI.

2.4.5. Random network

For further comparison, we created an unweighted, random network containing the same 6888 vertices that were mapped to HPRD PPI. Each vertex was connected to 8 randomly chosen vertices except itself. Repetitions of vertices were possible. As a result, the nodes' degrees form a unimodal distribution lying in the interval [8, 30] with a mean value of 15.991, median 16, and standard deviation 2.80.

2.5. Graph-Convolutional Neural Network (CNN)

One of the approaches for validation of the embedding networks is to analyze how the underlying molecular network influences performance of the machine learning method utilizing prior knowledge. The Graph-CNN [41] method was applied on the breast cancer dataset introduced in section 2.3 in recent studies [26-27]. We subtracted the minimal value (5.84847) of the data from each cell of the quantile normalized gene expression matrix to keep the gene expression values non-negative. The classification accuracy of Graph-CNNs was compared for different sources of network prior information: HPRD, Embedding_net_v1, Embedding_net_v2, STRING and BioBERT-based network. As for embedding networks, we utilized weighted and unweighted (taking into account only topology) versions. The vertices were mapped to the genes of gene expression data and weighted edges were filtered according to a threshold value. We considered thresholds higher than 0.5 for cosine similarity between vertices. The main connected component

of the underlying graph was used to structure the data. The performance was assessed by 10-fold cross-validation. For each of the data splits the model was trained on 9-folds and the classification was evaluated using 10-fold as validation set. For each underlying molecular network, the architecture and hyperparameters of Graph-CNN remained the same. For the majority of the cases, Graph-CNN was trained with 100 epochs, but for some versions of prior knowledge, a smaller number of epochs showed better results since convergence of gradient descent was happening faster. The most common evaluation metrics were used: area under curve (AUC), accuracy and F1-weighted score. The metrics were averaged over folds and the standard errors of their means were calculated. We compared performances based on weighted and unweighted networks.

2.6. Assessment of text corpus size effect

We tested how the text corpus size influences the variability of word representations and compared the similarities between given concepts obtained from four resulting embeddings. The embeddings were produced by applying word2vec to four text corpora of different sizes. The four text corpora were of size 4M, 8M, 12M, and $\sim >16$ M. We selected 10 terms of different entity types, the genes *brca1*, *psen1* and *egf*, the medical terms breast neoplasms, eczema, sleep and schizophrenia, and the molecular compounds ranitidine, lactose, and cocaine. For each entity term, we calculated its first 10 nearest neighbors and selected the ones that are commonly present in the four resulting embeddings (S7 File).

3 Results

3.1. Validation results

To demonstrate the utility of our word2vec embeddings in data analytical applications, we examined the agreement of cosine similarities between words according to their vector representations with information extracted from biomedical knowledgebases (see Materials and methods). As a result, pairs of genes with known interactions in the Reactome database showed on average higher cosine similarities than gene pairs without known interaction in the same database (Fig 2). Similarly, cosine similarities of drugs with overlapping target gene sets were, on average, higher than similarities between drugs without common target genes. Furthermore, cosine similarities within Reactome and TRANSPATH® pathways, as well as within GO biological processes, were increased compared to median cosine similarities of randomly sampled gene pairs (Fig 2). Regression curves estimated for the medians revealed a correlation between the number of pathways or GO category members and the median similarity, with higher values for smaller gene sets. We think that gene pairs in smaller pathway networks or biological processes were more likely to correspond to direct molecular interactors that share a close functional context than in pathway or functional categories with a higher number of members. The embedding, in many cases, indeed captured these relations. While disease-disease cosine similarities within Human Disease Ontology (HDO) groups also revealed such a trend for groups with less than 25 members, median similarities within groups were often smaller than for randomly chosen disease pairs (Fig 2). Therefore, disease-disease relations captured by broader HDO groups did not correspond well with vector presentations of the embedding. Better correspondence was observed for narrower disease groups but did not exceed similarities of random disease pairs. Full plots are provided as supplementary figures Figs S1-S6.

Fig 2. Validation of the word2vec embedding with existing knowledge from biomedical resources. Reactome pathways, TRANSPATH® pathways, GO biological processes and Human disease ontology present median cosine similarities as well as their lower and upper quartiles within groups of given number of members (genes or diseases, respectively) and for random samples (see Material and methods). Mean estimates were computed by fitting the decay function $f(x) = (x^a + b)^{-1}$ to medians, with the exception of the Human disease ontology comparison where a non-parametric local regression (Loess) was applied. Reactome protein-protein interactions and drug-gene associations show histograms of genes with or without known PPIs and of drugs with or without shared target genes, respectively.

3.2. Graph-CNNs results

Graph-CNN models were trained on the breast cancer data from section 2.3 to predict an occurrence of a metastatic event, utilizing different prior knowledge. The models with gene-gene networks derived from the embeddings showed best performance compared to HPRD PPI,

STRING-derived networks, BioBERT-derived network, and random network, in classifying patients into two groups, metastatic and non-metastatic. The networks were compared based on the similarity threshold and the number of vertices included. The architecture of Graph-CNN consists of 2 convolutional layers. Two convolutional layers were used with 32 convolutional filters. Maximum pooling of size 2 applies to both of the convolutional layers. Two fully connected layers have 512 and 128 nodes, consequently. The training was performed on 90% of data (872 patients) and 10% was saved as a test set (97 patients). The predicted patient data were classified into two groups: metastatic and non-metastatic. Table 1 presents the performance of Graph-CNNs trained with the word2vec-embedding networks (Embedding_net_v1 and Embedding_net_v2) and STRING derived network incorporating the edge weights. For STRING, the edge weights are the scores computed based on text-mining techniques (see Materials and Methods). We didn't consider a weighted BioBERT-derived network since the minimal weight was already 0.938 for around 6000 vertices, which is not that much different from 1.0. We can see that Embedding_net_v1 demonstrated a better performance than Embedding_net_v2 for almost the same number of vertices and better than the text-mining-based STRING network.

Table 1. The results of how weighted underlying networks influence the performance of Graph CNNs trained on gene expression data.

Network	Vertices	Similarity Threshold	AUC [%]	Accuracy [%]	F1-weighted [%]	Epochs
Embedding_net_v1 (weighted)	6092	0.65	83.09±0.97	76.67±1.14	76.45±1.14	100
Embedding_net_v2 (weighted)	6086	0.68932	82.05±1.08	75.03±0.70	74.74±0.76	100
Embedding_net_v1 (weighted)	6775	0.63	82.53±1.46	75.62±1.72	75.29±1.73	100
Embedding_net_v2 (weighted)	6774	0.6774	81.87±1.39	75.33±1.22	75.16±1.28	40
STRING (text mining weighted)	6840	0.744	81.76±1.95	75.97±1.99	75.63±2.02	100

The networks were compared based on the similarity threshold and number of vertices included. ‘Vertices’ are the vertices in the main connected component. ‘Similarity Threshold’ is the minimum weight value to keep connections between genes. AUC, Accuracy, and F1 weighted are the evaluation metrics used. ‘Epochs’ is the number of passing the entire dataset through the neural network.

In Table 2, we compared how unweighted network topologies influence the classifier’s performance depending on the similarity threshold and the number of vertices included. The baseline performance corresponds to HPRD PPI prior knowledge. STRING (combined) and BioBERT-based networks were considered only as unweighted since the weight thresholds to reach a comparable number of vertices were close to 1, 0.938 and 0.952, respectively.

The embedding networks have a threshold value allowing to change the strength of similarity between vertices. Change of threshold for Embedding_net_v1 from 0.63 to 0.65 increased the classification result in weighted and unweighted cases. We can also observe that for embedding networks, the incorporation of weight’s edges increased slightly, although not substantially, the classification performance. Meanwhile, STRING and BioBERT-based networks do not bring any

improvements compared to HPRD PPI or random network. Thus, Graph-CNNs showed the best results on our dataset, incorporating weighted Embedding_net_v1 with a threshold of 0.65.

Table 2. Influence of unweighted underlying networks on the performance of Graph CNNs trained gene expression data.

Network	Vertices	Similarity Threshold	AUC [%]	Accuracy [%]	F1-weighted [%]	Epochs
HPRD	6888	-	82.57±1.25	76.07±1.30	75.82±1.33	100
Embedding_net_v1	6092	0.65	83.02±1.09	76.38±1.44	76.14±1.47	40
Embedding_net_v1	6775	0.63	82.41±1.20	76.03±1.53	75.69±1.53	100
Embedding_net_v2	6874	0.675	82.37±1.36	75.04±1.25	74.84±1.17	40
STRING (text mining)	6840	0.744	81.67±2.01	76.07±1.50	75.61±1.57	25
STRING (combined)	6862	0.938	81.77±1.17	74.62±1.56	74.33±1.64	100
BioBERT-based network	6865	0.95245	82.26±1.27	74.81±1.53	74.61±1.50	40
Random network	6888	-	81.89±0.10 9	75.65±0.99	75.43±0.99	40

The networks were compared based on the similarity threshold and number of vertices included. ‘Vertices’ are the vertices in the main connected component. ‘Similarity Threshold’ is the minimum weight value to keep connections between genes. AUC, Accuracy, and F1 weighted are the evaluation metrics used. ‘Epochs’ is the number of passes the entire dataset through the neural network.

3.3. Effect of text corpus size

We generated four embeddings trained with corpora of different sizes to examine the variability of cosine similarity values depending on the amount of training data. Fig 3 illustrates the results for selected terms of different types and their nearest neighbors. The cosine similarities are varying between the terms. For example, the similarity value between the “breast neoplasms” and “ovarian neoplasms” had increased slightly as follows: 0.851 (4M), 0.855 (8M), 0.861 (12M), 0.862

(~16M). Breast neoplasm is one of the most frequent diagnosed neoplasms reported in biomedical literature. Many studies have also reported similarities between breast and ovarian cancer since they share similar mutations (tumor suppressors). On the other hand, the similarity between "brca1" and "brca2" is almost the same in the four embeddings (0.898, 0.893, 0.891, 0.898 for 4M, 8M, 12M, and ~16M, respectively) with a very high similarity compared to the other nearest neighbors of "brca1". BRCA1 and BRCA2 genes are the most common genes defined in literature with certain mutations and lead to an increased risk of breast and ovarian neoplasms. Similarly, for "schizophrenia" and "bipolar_disorder", their similarity has changed slightly (0.822, 0.825, 0.828, 0.829 for 4M, 8M, 12M, and ~16M, respectively). An overlap between schizophrenia and bipolar disorder has been commonly reported in the literature.

In contrast, the similarity between "eczema" and "atopy" had changed differently. It had decreased from 0.713 in the embedding with the corpus of size 4M to 0.669 in the one with 8M, to continue increasing again to 0.691 (12M) and 0.701 (~16M) while staying lower than their similarity in the embedding pre-trained with the smallest corpus.

Overall, we observed that the nearest neighbors of selected terms were assigned similarity values as well as a similar ranking that were varying slightly in the four embeddings for the majority of the selected terms. However, for common terms such as breast neoplasms, BRCA1, and schizophrenia and their nearest neighbors with which they tend to appear more frequently in biomedical literature, the similarity was more robust.

Fig 3. Assessment of similarities between selected terms and their nearest neighbors present in 4 embeddings. The selected terms are the genes brca1, psen1 and egf, the medical terms breast neoplasms, eczema, sleep and schizophrenia, and the molecular compounds ranitidine, lactose, and cocaine; and their nearest neighbors present in 4 embeddings trained on a text corpus of different sizes. Each numbered subplot represents the cosine similarities between a selected term and its nearest neighbors from the four embeddings. Each plotted line represents one embedding and each dot on a line is the similarity between a selected term and one nearest neighbor. A point on the line is the similarity value between a selected term and one nearest neighbor. Nearest neighbors are arranged according to decreasing cosine similarity for the 4M corpus.

4 Discussion

In this paper, we focused on demonstrating the utility of word embedding-derived knowledge in uncovering valuable biological relationships and its application in machine learning tasks. While the text corpus consisted entirely of PubMed abstracts, our approach can be applied to complete scientific or other types of natural language texts. We performed a computational analysis to validate similarities between biomedical entities, namely, genes, diseases, and drugs, using existing knowledge in biomedical databases. Comparisons showed that relations between known PPIs, common pathways and cellular functions, or narrower disease ontology groups correlated with higher vector cosine similarity. Gene pairs with known PPIs in Reactome have generally shown higher cosine similarities. Gene embeddings seem to be rich with semantic information about gene function. On the other hand, gene pairs with high cosine similarities shown without known interactions in Reactome would lead to new investigations to uncover hidden functional relationships. Besides, gene pairs sharing common pathways in Reactome and TRANSPATH®, as well as common biological processes in GO, showed increased cosine similarities compared to the median of randomly sampled gene pairs. Similarities were also increased with smaller group sizes, which more likely represent direct molecular interactions. Disease pairs also showed increased cosine similarities within smaller human disease ontology (HDO) terms/groups, e.g., ≤ 20 diseases, which probably represent more specific disease classes. However, disease embeddings did not correspond well on the basis of median random similarity. This is potentially interesting to investigate why semantic relations between diseases differ from the HDO, although it would suggest that the embedding could harbor new insights for disease ontologists.

Corpus size effect assessment showed that similarities between selected terms were substantially affected by the corpus size. In general, we noticed that the first nearest neighbor for most terms

was not strongly influenced by the corpus size even though it was not changing proportionally with the corpus size. The highest and strongest similarities were observed between "breast_neoplasms" and "ovarian_neoplasms" as well as between "brca1" and "brca2". This might be justified by the fact that these terms are very common in the present literature and words in each pair tend to occur more frequently in the same context. This could validate the ability to extract meaningful functional relationships between biomedical terms.

Additionally, to demonstrate the utility of the embedding in machine learning tasks, we assumed that similarities between biological entities might help create networks of a specific type. The results of Graph-CNN showed that the embedding-based networks are topologically meaningful. Weighted and unweighted Embedding_net_v1 allowed to increase the classification performance to predict breast cancer's metastatic event achieving the highest F1 score, AUC, and accuracy. This has demonstrated the biological utility of the embedding as prior knowledge. This utility of the biomedical embedding can be an advantage to create other networks of different types of entities such as disease or drug networks. Moreover, the change of similarity threshold of edge weights from 0.63 to 0.65 has led to an increase in performance; this could be due to the fact that "weak genes" contained in the network were filtered out. Besides, Random network-based demonstrated lower performance, although it is still to be investigated how simulated networks with different degree distributions would influence the classification error rate. It was also shown that the model trained with the Embedding_net_v1 network had performed better than with the Embedding_net_v2. The former was produced from the embedding that considered synonymous terms replaced with their main terms. Such a procedure has surely influenced the embedding information and, in particular, the semantic relations between terms. For example, considering the gene WNT4 and its nearest neighbor WNT7a, the cosine similarity between them has increased

from 0.798 in Embedding_v2 to 0.811 in Embedding_v1, respectively. Although the increase in similarity was small, this has led to changing WNT7a from being the third neighbor of WNT4 to becoming its first neighbor. Although our examination was based only on gene-gene relations, it can be extended to cover other types of relations, e.g., disease-disease, gene-disease, etc. Our validation analysis was performed based on Embedding_v1, while the similarities between biomedical entities can also be evaluated by checking the influence of Embedding_v2.

Furthermore, the influence of embedding-based networks can also be examined by considering text-mining-based networks other than STRING and BioBERT. One could also derive networks from BioBERT using different hidden layers. For our BioBERT-derived network, the vectors of words were extracted from the last hidden layer. The BERT authors extracted vector combinations from different layers. They tested word-embedding strategies by feeding those vector combinations to a BiLSTM (bidirectional long short-term memory used on a named entity recognition task and observing the resulting F1 scores [19]. The concatenation of the last four layers produced the best results on this specific task. It is generally advisable to test different combinations in a particular application since results may vary.

We have also developed a web service based on this work to explore biomedical concepts present in our generated embeddings. It can be accessed under the link: <https://ebiomecon.genexplain.com/>. The service facilitates accessing the embedding information, and it provides functions to explore similarities of biomedical concepts, including the possibility to extract network vertices. Our embedding networks were extracted in the form of vertices. The implementation and the functionality of the web service will be described in more depth in a separate publication (manuscript in preparation).

Graph-CNNs, trained with word2vec-embedding-based networks, have shown a good enough performance to classify metastatic events vs. non-metastatic events. Predictions of Graph-CNN applied to the same gene expression data used in this study with the HPRD PPI were explained in a recent study and provided patient-specific subnetworks [27]. An interesting research question brought up by this study is whether patient-specific subnetwork genes predicted using an embedding-based gene-gene network would give different insights into the tumor biology of a patient than those predicted using PPI networks. This might also provide biological insights into the molecular interactions in the network and help validate the biological utility of the embedding-based network information.

5 Conclusion

In this study, we leveraged a state-of-the-art text-mining tool to learn the semantics of biomedical concepts presented in biomedical literature and to demonstrate the utility of learned embeddings for biomedical research and analysis. Our learned embeddings have been validated in computational analyses addressing protein-protein interactions, signaling pathways and biological processes, drug targets, and human diseases and showed agreement with existing biological knowledge. The results demonstrated that vector representations of biomedical concepts as produced by word2vec can capture meaningful biological relationships between biomedical entities. We also showed that semantic relations extracted from vast literature could be applied as prior knowledge in machine learning tasks.

Acknowledgments

This work was supported by the German Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF) project iDDSEM *MyPathSem* (ID 031L0024A+B).

Supporting information

S1 Fig. Gene-gene cosine similarities within Reactome pathways of given number of genes.
(TIF)

S2 Fig. Gene-gene cosine similarities within GO biological processes of given number of genes.
(TIF)

S3 Fig. Gene-gene cosine similarities within TRANSPATH® pathways of given number of genes.
(TIF)

S4 Fig. Disease-disease cosine similarities within human disease ontology groups of given number of diseases.
(TIF)

S5 Fig. Histograms of genes with or without known Reactome protein-protein interactions.
(TIF)

S6 Fig. Histograms of drug-gene associations with or without shared target genes.
(TIF)

S7 Fig. Drug-drug cosine similarity distributions with shared genes of given number in DrugBank.
(TIF)

S1 File. Summary of gene-gene cosine similarities within biological processes in Gene Ontology (GO).
(CSV)

S2 File. Summary of gene-gene cosine similarities within pathways in Reactome.

547 (CSV)

548 **S3 File. Summary of gene-gene cosine similarities within pathways in TRANSPATH®.**

549 (CSV)

550 **S4 File. Summary of disease-disease cosine similarities within disease groups in Human**

551 **Disease Ontology (HDO).**

552 (CSV)

553 **S5 File. Gene pairs with or without known protein-protein interactions in Reactome.**

554 (CSV)

555 **S6 File. Drug pairs with or without shared target genes.**

556 (CSV)

557 **S7 File. Cosine similarities of selected terms and their nearest neighbors in four corpora of**

558 **different sizes.**

559 (CSV)

560

561 **References**

562

- 563 1. Müller H-M, Kenny EE, Sternberg PW. Textpresso: an ontology-based information
564 retrieval and extraction system for biological literature. PLoS Biol. 2004;2(11).
- 565 2. Swain MC, Cole JM. ChemDataExtractor: a toolkit for automated extraction of chemical
566 information from the scientific literature. J Chem Inf Model. 2016;56(10):1894–904.
- 567 3. Spangler S, Wilkins AD, Bachman BJ, Nagarajan M, Dayaram T, Haas P, et al.
568 Automated hypothesis generation based on mining scientific literature. In: Proceedings of
569 the 20th ACM SIGKDD international conference on Knowledge discovery and data
570 mining. 2014. p. 1877–86.
- 571 4. Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A. GENIES: a natural-language
572 processing system for the extraction of molecular pathways from journal articles. In:
573 ISMB (supplement of bioinformatics). 2001. p. 74–82.
- 574 5. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of
575 words and phrases and their compositionality. In: Advances in neural information
576 processing systems. 2013. p. 3111–9.
- 577 6. Tang B, Cao H, Wang X, Chen Q, Xu H. Evaluating word representation features in
578 biomedical named entity recognition tasks. Biomed Res Int. 2014;2014.

- 579 7. Liu S, Tang B, Chen Q, Wang X. Effects of semantic features on machine learning-based
580 drug name recognition systems: word embeddings vs. manually constructed dictionaries.
581 Information. 2015;6(4):848–65.
- 582 8. Jagannatha A, Chen J, Yu H. Mining and ranking biomedical synonym candidates from
583 Wikipedia. In: Proceedings of the sixth international workshop on health text mining and
584 information analysis. 2015. p. 142–51.
- 585 9. Jiang Z, Jin L, Li L, Qin M, Qu C, Zheng J, et al. A CRD-WEL system for chemical-
586 disease relations extraction. In: The fifth BioCreative challenge evaluation workshop.
587 2015. p. 317–26.
- 588 10. Liu S, Tang B, Chen Q, Wang X. Drug-drug interaction extraction via convolutional
589 neural networks. Comput Math Methods Med. 2016;2016.
- 590 11. Jiang Z, Li L, Huang D. A general protein-protein interaction extraction architecture based
591 on word representation and feature selection. Int J Data Min Bioinform. 2016;14(3):276–
592 91.
- 593 12. Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, et al. A comparison of
594 word embeddings for the biomedical natural language processing. J Biomed Inform.
595 2018;87:12–20.
- 596 13. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word
597 embeddings with subword information and MeSH. Sci data. 2019;6(1):1–9.
- 598 14. Du J, Jia P, Dai Y, Tao C, Zhao Z, Zhi D. Gene2vec: distributed representation of genes
599 based on co-expression. BMC Genomics. 2019;20(1):82.
- 600 15. Chen Q, Lee K, Yan S, Kim S, Wei C-H, Lu Z. BioConceptVec: Creating and evaluating
601 literature-based biomedical concept embeddings on a large scale. PLoS Comput Biol.
602 2020;16(4):e1007617.
- 603 16. Kilimci ZH, Akyokus S. Deep learning-and word embedding-based heterogeneous
604 classifier ensembles for text classification. Complexity. 2018;2018.
- 605 17. Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In:
606 Proceedings of the 2014 conference on empirical methods in natural language processing
607 (EMNLP). 2014. p. 1532–43.
- 608 18. Wang Y, Cui L, Zhang Y. Using Dynamic Embeddings to Improve Static Embeddings.
609 arXiv Prepr arXiv191102929. 2019;
- 610 19. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional
611 transformers for language understanding. arXiv Prepr arXiv181004805. 2018;
- 612 20. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. Deep contextualized
613 word representations. arXiv Prepr arXiv180205365. 2018;

21. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234–40.
22. Kim Y. Convolutional neural networks for sentence classification. *arXiv Prepr arXiv14085882*. 2014;
23. Hu B, Lu Z, Li H, Chen Q. Convolutional neural network architectures for matching natural language sentences. In: *Advances in neural information processing systems*. 2014. p. 2042–50.
24. Min S, Lee B, Yoon S. Deep learning in bioinformatics. *Brief Bioinform*. 2017;18(5):851–69.
25. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–44.
26. Chereda H, Bleckmann A, Kramer F, Leha A, Beissbarth T. Utilizing Molecular Network Information via Graph Convolutional Neural Networks to Predict Metastatic Event in Breast Cancer. *Stud Health Technol Inform*. 2019;267:181–6.
27. Chereda H, Bleckmann A, Menck K, Perera-Bel J, Stegmaier P, Auer F, et al. Explaining decisions of Graph Convolutional Neural Networks: patient-specific molecular subnetworks responsible for metastasis prediction in breast cancer. *bioRxiv*. 2020;
28. gensim: Introduction [Internet]. Available from: <https://radimrehurek.com/gensim/intro>
29. Povey S, Lovering R, Bruford E, Wright M, Lush M, Wain H. The HUGO gene nomenclature committee (HGNC). *Hum Genet*. 2001;109(6):678–80.
30. Dhammi IK, Kumar S. Medical subject headings (MeSH) terms. Springer; 2014.
31. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res*. 2008;36(suppl_1):D901–6.
32. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. *Nucleic Acids Res*. 2020;48(D1):D498–503.
33. Choi C, Krull M, Kel A, Kel-Margoulis O, Pistor S, Potapov A, et al. TRANSPATH®—a high quality database focused on signal transduction. *Comp Funct Genomics*. 2004;5(2):163–8.
34. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. *Nat Genet*. 2000;25(1):25–9.
35. Home - geneXplain geneXplain [Internet]. Available from: <http://genexplain.com/>
36. Schriml LM, Mitraka E, Munro J, Tauber B, Schor M, Nickle L, et al. Human Disease

- 647 Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids*
648 *Res.* 2019;47(D1):D955–62.
- 649 37. Bayerlová M, Menck K, Klemm F, Wolff A, Pukrop T, Binder C, et al. Ror2 signaling
650 and its relevance in breast cancer progression. *Front Oncol.* 2017;7:135.
- 651 38. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI
652 GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.*
653 2012;41(D1):D991–5.
- 654 39. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al.
655 Exploration, normalization, and summaries of high density oligonucleotide array probe
656 level data. *Biostatistics.* 2003;4(2):249–64.
- 657 40. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et
658 al. Human protein reference database—2009 update. *Nucleic Acids Res.*
659 2009;37(suppl_1):D767–72.
- 660 41. Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs
661 with fast localized spectral filtering. In: *Advances in neural information processing*
662 systems. 2016. p. 3844–52.
- 663 42. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11:
664 protein–protein association networks with increased coverage, supporting functional
665 discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 2019;47(D1):D607–
666 13.
- 667 43. PyTorch-Transformers | PyTorch [Internet]. Available from:
668 <https://pypi.org/project/pytorch-transformers/>