Extracting Gene Associations From Biomedical Literature Related To Wound Healing

Jayati Halder Jui
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260
jaj146@pitt.edu

Yoshita Buthalapalli
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15260
yrb3@pitt.edu

Abstract

Wound healing is a complex process which requires the collaborative efforts of different tissues and cell lineages. Genome wide transcription analysis helps to find potential targets to alleviate healing process via gene therapy. An accurate and comprehensive knowledge-base of gene and protein interactions can thus play an important role in understanding and identifying important interactions during different healing phases. In recent years, text mining methods are gaining much popularity to extract domain specific knowledge of biological process. The objective of our work is to identify such knowledge between genes and proteins interactions that play important role in wound healing.

1 Introduction

The wound healing process is well-understood on the cellular and tissue level; however, its complex molecular mechanisms are not yet uncovered in their entirety. Understanding the process of wound healing as a pattern of molecular networks provides the tools for analyzing and optimizing the healing process. It helps to answer specific questions that lead to better understanding of the complexity of the process. What are the gene involved in wound healing? How do these genes interact with each other during the different stages of wound healing?

A massive number of biological entities, such as genes and proteins, are mentioned in the biomedical literature. It is known that biomedical concepts have a high degree of ambiguity as the same words can be used to describe different types of concepts in free text. Predicting gene associations from such large corpus can prove to be helpful to many research tasks. To predict such associations, we followed the works of Chen et al. (2020) on BioConceptVec which covers over 400,000 biomedical concepts mentioned in the literature and is of the largest among the publicly available biomedical concept embeddings to date. Other works like Onto2Vec by Chen et al. (2018) introduced a method that can be used to produce feature vectors for biological entities based on their annotations to biomedical ontologies. They demonstrated how Onto2Vec representations could be used to improve predictive models for protein-protein interactions, while our paper focuses on predicting gene associations using a model trained on our data.

This workflow is divided into four stages. The first stage is data/abstracts collection and pre-processing which is discussed in section 2. Then we introduce our approach of using Word2Vec method to obtain vectors for each word in the data which is also covered in section 2. In section 3 we present an exploratory analysis of the word vectors. In section 4, we present intrinsic and extrinsic evaluation results to better understand whether the predicted associations are meaningful.

Preprint. Under review.

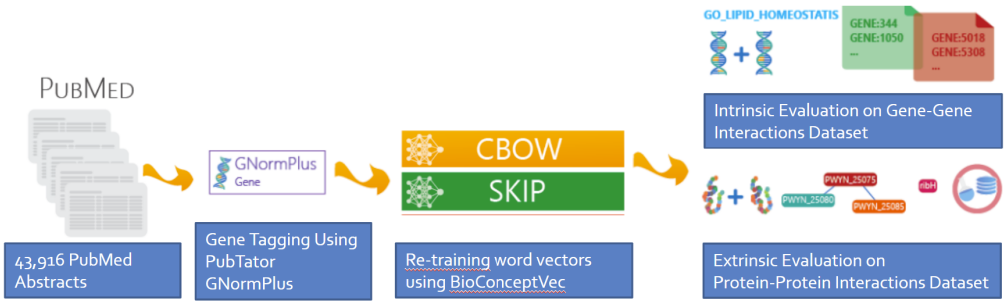


Figure 1: Workflow.

2 Training a word embedding model - Word2Vec

This section details the training procedure of the word embedding models.

2.1 Data Extraction

For training the word vectors, we used PubMed abstracts related to wound healing. We retrieved abstracts from PubTator API of NCBI for this purpose [Wei et al. (2013)]. Gene and protein full names are often a collection a words like *histone deacetylase 3* which are also represented by symbols (HDAC3 or HDAC-3). PubTator uses GNormPlus for tagging gene and protein names and can successfully identify full names as well as short symbolic forms [Wei et al. (2015)]. An example annotation is shown in figure 2.

Implementation details: We first retrieved all the PubMed IDs (PMID) of research papers related to wound healing. We divided the PMIDs in batches of 100 to send query to the PubTator API which returned a JSON object containing the offsets of gene names and their corresponding tags. Using the API we could annotate 43916 abstracts related to wound healing that has at least one gene/protein tags. After replacing the gene mentions with tags, there were 9151 unique genes tagged in the whole corpus. We shortened our corpus and vocabulary by discarding all the sentences in the abstracts that didn't have any tags. The final corpus had 244342 annotated sentences.

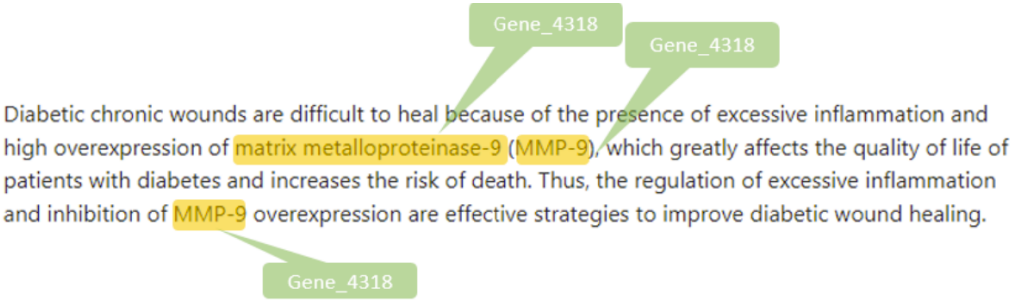


Figure 2: PubTator annotation of the first two lines abstract of Li et al. (2022)

We plotted a few graphs to understand the distribution of the corpus. Fig 3 show the word distributions, cumulative word distributions and the distribution of gene/protein tags in the corpus. From Fig 3 we can see that are about 130 words in each sample where a sample is defined as an abstract containing only the annotated sentences that have at least one gene/protein tag. Moreover, 50 percent of samples only contain a 100 words while each sample has at 10 gene/protein tags on average. However, most abstracts has only a few gene/protein tags.

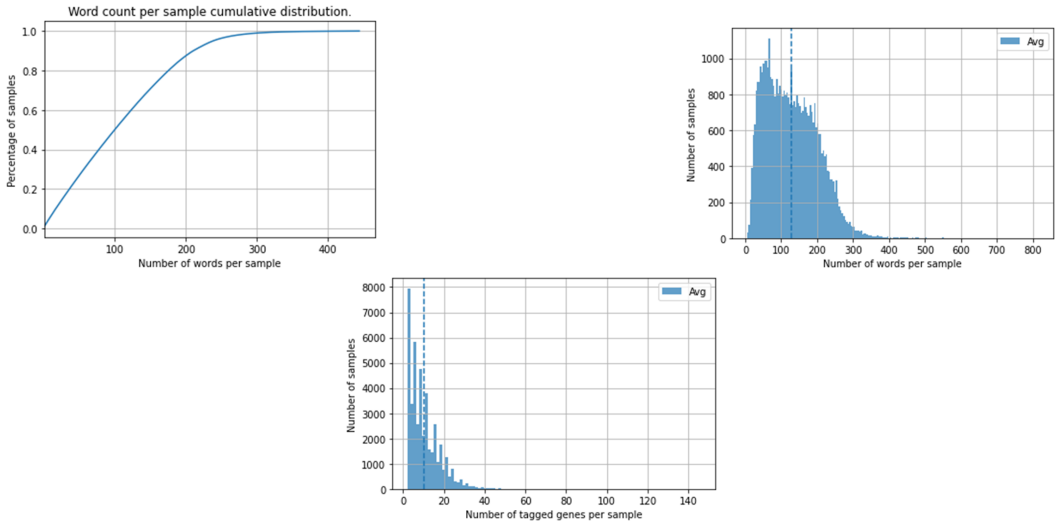


Figure 3: Distribution of words in the Corpus

2.2 Training Word Vectors

Word2Vec is a set of neural-network based tools that generate vector representations of words from large corpora. The vector representations are obtained in such a way that words with similar contexts tend to be close to each other in the vector space. We trained word vectors on the extracted corpus. To our knowledge, there is no agreement on which embedding model is the most effective in biomedical domains. Hence in this study, we used two Word2Vec embedding models, namely, **Continuous Bag-of-Words**, or CBOW model and **Continuous Skip-Gram** Model. The difference between CBOW and Skip-Gram is due to the difference in learning algorithm of Word2Vec. CBOW uses context-word to predict target-word while Skip-Gram uses target-word to predict context words.

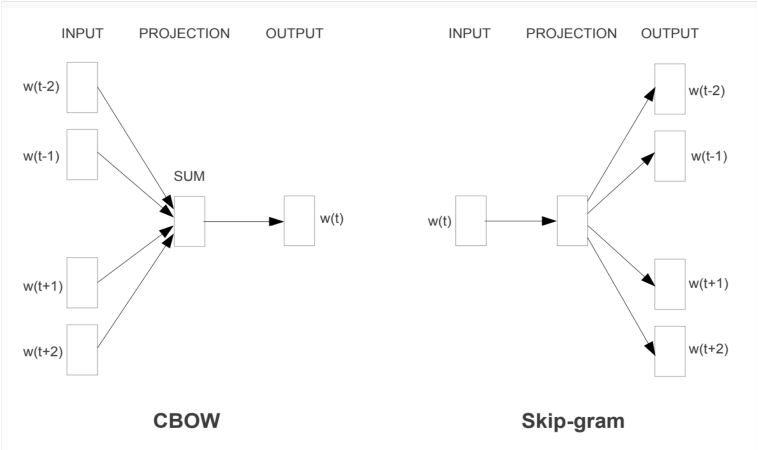


Figure 4: Word2Vec training models, Taken from "Efficient Estimation of Word Representations in Vector Space", 2013.

Word2Vec models utilize neural networks, each word is initially given a random vector which is tuned on each iteration. Since we have a limited corpus from which we need to learn embeddings of over 9k gene tags, we decided on incorporating prior knowledge about these genes. For this purpose, we used pre-trained wordvectors given by BioConceptVec and re-trained on our corpus without updating the Vocabulary with pre-trained words.

Following the word of Chen et. al., we trained our model using following hyper-parameters [Chen et al. (2020)]:

- min_count=5. All words with frequency below 5 are ignored.
- window=10. The maximum distance between current and predicted word.
- vector_size=100. The vector length for each word.
- alpha=0.025. The learning rate for weights
- epochs=30

3 Exploratory Analysis

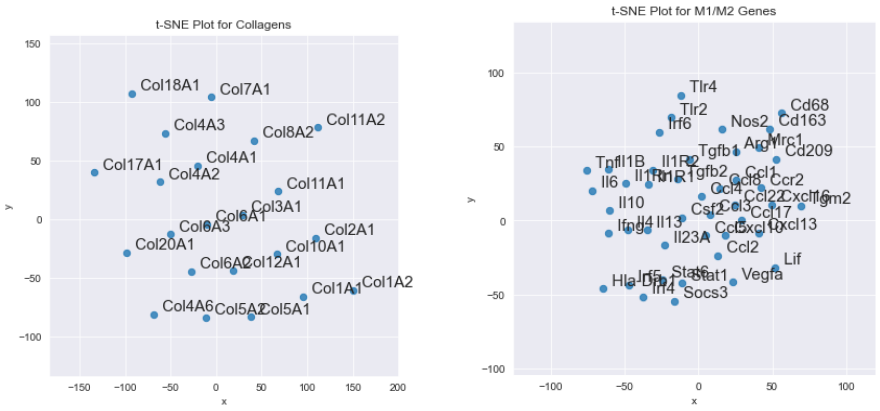
We posit that word vectors representing genes/proteins that are related through some biological processes should be similar. To explore the usefulness of the trained word vectors representing gene/proteins, we attempted to find the similarity of these word vectors in predefined sets related genes that are supposed to play important roles in wound healing. We manually created five genesets that are grouped based on their functions: collagens, fibrinogens, macrophage related genes, muscle regeneration genes and neural regeneration genes. Collagen is the most abundant protein in human body and the role of collagen in wound healing is to attract fibroblasts and encourage deposition of new collagen to the wound bed. Fibrinogens play an important role in blood clotting, fibrinolysis, cellular and matrix interactions and inflammation phase in wound healing. Macrophages play key roles in all phases of wound healing, which are inflammation, proliferation, and remodeling. As wounds heal, the local macrophage population transitions from predominantly pro-inflammatory (M1) to anti-inflammatory (M2). The final phase of wound healing a.k.a the closure of an wound can be realized by regeneration or repair. Both muscle and nerve regeneration genes play important roles in wound closure.

We analyzed the similarity of the word vectors in each of the aforementioned genesets. For computing the similarity of a set, we computed pair-wise similarities of all word vectors representing the genes/proteins in a set using cosine similarity. Since the range of cosine similarities can vary widely, we followed Chen et al. (2020) to normalize all pair-wise similarities before computing average. We applied Z-score standardization followed by Min-Max normalization to bring the pair-wise similarities in the range of 0 and 1 and then computed the average of the pair-wise similarities which we report as average set similarity. The similarities are summarized in Fig 5. We found that the collagen and macrophage related (M1/M2) sets obtained the highest similarities using both CBOW ans Skip-Gram embeddings. This is very meaningful because collagens and macrophage related genes are found to take part in wound healing from the beginning till the end. Hence, these genes are often mentioned in wound healing literature.

Geneset	Size	Avg. Sim CBOW	Avg. Sim SkipGram
Collagen	43	**0.6013**	**0.5627**
M1/M2	50	**0.5088**	**0.5619**
Muscle Regeneration	75	0.4339	0.4967
Fibrinogen	22	0.4033	0.4502
Nerve Regeneration	48	0.3989	0.4607

Figure 5: Average similarities of wound healing related genesets

Fig 6 shows the word vectors of the genes in Collagen and M1/M2 genesets in 2 dimensional t-SNE plots. The strong clusters as we can see in the plot confirms the high similarities in these sets as depicted in Fig. 5. We can further try to find the positively and negatively related genes in these wound healing related genesets. Fig. 7 shows our attempt of finding the most similar and least similar macrophage related and muscle regeneration related genes. We can assess the significance and correctness of these genes through biological literature. For example, according to Fig. 7 IL6 and IL1B are canonical pro-inflammatory (M1) marker genes while IL10 being an anti-inflammatory marker shows strong pro-inflammatory properties in wound healing and often grouped together with pro-inflammatory markers in several studies. While the least similar IL1R2 and VEGF-a are

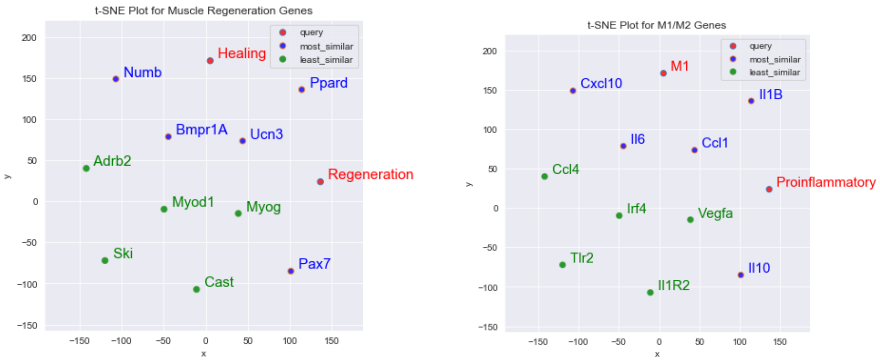


(a) t-SNE plot of Collagens

(b) t-SNE plot of M1/M2 genes

Figure 6: Visualizing word vectors in Collagen and M1/M2 genesets in lower dimensions using CBOW embeddings

canonical M2 markers which show anti-inflammatory properties as opposed to pro-inflammation [Ferrante and Leibovich (2012)].



(a) t-SNE plot of top Muscle Regeneration Related Genes

(b) t-SNE plot of top M1/M2 genes

Figure 7: Visualizing top genes in Muscle Regenration and M1/M2 genesets based on queries using CBOW embeddings

4 Evaluation

The evaluation of word embeddings can be broadly categorized into two types (i.e., intrinsic and extrinsic). Intrinsic evaluations are commonly accomplished via an unsupervised setting or using weakly supervised labels, whereas extrinsic evaluations are often performed via a supervised setting in downstream applications. In our study, we assess the effectiveness of gene embeddings through intrinsic evaluation. Through extrinsic evaluation we strengthen our argument about the applicability of these vectors in model learning.

4.1 Datasets

We use three datasets in particular assess the effectiveness of the learned gene embedding vectors. We had to reduce the all three datasets to match the genes found in this study.

4.1.1 Datasets for Intrinsin Evaluation

For intrinsic evaluation, we used two gene-gene relation datsets from MSigDB [Liberzon et al. (2011)]. MSigDB C2 is generated based on related genes that are found in biological pathways and reported in different biological literature. MSigDB C4 contain related genesets found through some cancer related micro-array data. Each entry in these datasets contain an identifier (pathway or concept) and a set of related gene to that identifier. Chen et al. (2020) extracted the C2 and C4 datasets and added a new column to these datasets namely negatively related genes. For each entry in these datasets, they created a set of negatively related genes by sub-sampling from a set of unrelated genes (or of unknown relations). The sizes of the positively and negatively related genes are same in each entry. We extracted the publicly available C2 and C4 datasets from Chen et al. (2020) and matched each entry to fit the genes in found in the vocabulary of our embedding models. We have extracted the entries in the datasets that had at least 10 genes and at most 50 genes while keeping the number of positively and negatively related genes the same. Our modified C2 dataset has 609 entries and C4 dataset has 141 entries after all processing.

4.1.2 Dataset for Extrinsic Evaluation

We used protein-protein interactions from the STRING database for extrinsic evaluation [Szklarczyk et al. (2021)]. We used STRINGdb package of *R* to extract the protein-protein interactions. We could map 3450 proteins in STRING from our vocabulary of 9151 tags. Each interaction in STRING is defined by a pair of proteins with a combined score that denotes the strength of the interaction. We extracted all interactions that have a combined score greater than 700. Our final dataset thus has 57126 interactions of 3450 proteins. For training a supervised model, we also needed some negative interactions. Hence we randomly created 16 negative interactions for each of the 3450 proteins using the genes whose relation to this protein is unknown (no relation in STRING db). Thus our final dataset has 112326 protein pairs with 57126 positive (interaction found) and 55200 negative (interaction not found) interactions.

4.2 Results

This section summarizes the evaluation results.

4.2.1 Intrinsic Evaluation Results

We evaluated the MSigDB C2 and C4 databases based on the difference of average positive and negative similarities. We first calculated the average similarities of each positive and negative sets in the dataset. Then we computed the average difference. The results are summarized in Fig. 8. We can see that C4 dataset has higher average difference in similarities. However, this dataset is also much smaller compared to C2. The difference in similarities is much smaller compared to the results described by Chen et al. (2020). However, a direct comparison is not possible because we trained our vectors in a very limited corpus. Some fine tuning of the word-vectors might improve the similarities.

Word2Vec	Dataset	Avg. Pos Similarity	Avg. Neg Similarity	Avg. Difference (%)
CBOW	C2	0.4178	0.3270	9.07%
	C4	0.4305	0.3293	**10.12**%
SkipGram	C2	**0.**4247	0.3303	9.44%
	C4	0.**4331**	0.**3324**	10.07%

Figure 8: Results of intrinsic evaluation

4.2.2 Extrinsic Evaluation Results

For extrinsic evaluation we generated five models for both the W2V algorithms. We split our obtained PPI(protein-protein interaction) dataset into train and test set by a 80%-20% split. Our training data has 89860 interactions while the test data has 22466 interactions. The results of extrinsic evaluation is depicted in Fig. 9. The baseline model simply takes the cosine similarity between two proteins, and reports a positive interaction if the similarity is higher than 0.5. The F1 score of the baseline models for both CBOW and Skip-Gram models are greater than 0.65. This shows that related proteins have high similarity in their vector representations. We generated four machine learning models for both CBOW and Skip-Gram.

The Logistic regression models achieve 0.71 and 0.7 F1 score for CBOW and SKip-gram respectively. We used saga solver with Ridge regulation which is better suited for large datasets. Logistic Regression (LR) fails when the decision boundary between two classes is non linear. As we have seen in previous graphs, the gene similarity tends to form a cluster/ spherical shapes. Hence, it is reasonable for the LR models to fail. Logistic Regression can model nonlinearities using feature/basis functions, the limitation is on how to define the right set of basis functions and many basis functions means many weights to learn. We also have a large feature vector size of 200 achieved by concatenating the word vectors of the protein pairs. Hence, we decided to move on to try Gaussian Naive Bayes method which is a generative approach to classification. But it perform poorly compared to LR. Naive Bayes makes an independence assumption between the features. Since the proteins are correlated, the independence assumption of Naive Bayes fails. This has resulted in poor performance of the Naive bayes model which is even worse than the base Skip-Gram model. To avoid any prior assumption of the feature space, we moved on to trying SVM.

SVM tries to find the **best margin**, while logistic regression does not, instead it can have different decision boundaries with different weights that are near the optimal point. Also, SVM has been known to perform well with textual data as it works well with unstructured and semi-structured data. The kernel choice of SVM makes a huge difference in the final model performance. The kernel selection is very crucial for higher dimensional data space. We achieved best performance for both CBOW and SKip-Gram embeddings with SVM classifiers and radial basis function (RBF) kernel. We wanted to try an ensemble technique in the hope of getting improved performance over SVM. We used XGBoost with considerable fine-tuning since XGBoost has many parameters. We tuned the number of estimators, the maximum depth of a tree and the learning rate. The best performance of XGBoost models were achieved with 500 estimators, 0.1 learning rate and a maximum tree depth of 10. The performance of the best XGBoost models were the same as the SVM models for both embedding models. We could not improve the performance of XGBoost over SVM. Some fine-tuning of the other parameters might improve XGBoost performance slightly.

W2V Algo	Model	Precision	Recall	F1_Score
CBOW	Cosine Sim > 0.5	0.76	0.69	0.66
	Logistic Regression	0.71	0.71	0.71
	Naïve Bayes	0.66	0.66	0.66
	SVM	**0.90**	**0.90**	**0.90**
	XGBoost	**0.90**	**0.90**	**0.90**
SkipGram	Cosine Sim > 0.5	0.75	0.68	0.68
	Logistic Regression	0.70	0.70	0.70
	Naïve Bayes	0.66	0.66	0.66
	SVM	**0.90**	**0.90**	**0.90**
	XGBoost	**0.90**	**0.90**	**0.90**

Figure 9: Results of extrinsic evaluation

5 Conclusion and Future Work

In this study, we attempted to assess the associations of genes mentioned in biological literature related to wound healing. We computed the vector representation of these genes through standard text mining approaches. The work of developing word vectors for clinical domains have gained much attention from the research community for many years. But a standard set of word vectors for biological domains is still to be developed. We analyzed the prior works in the domain for finding biological entity embeddings in detail and took an approach of finding gene relations from wound healing literature from the insights obtained through literature review. We skipped to write a thorough literature review because of the space limitation. The effectiveness of the obtained gene vectors through this study have been confirmed by the exploratory analysis and the intrinsic and extrinsic evaluation. Our extrinsic evaluation results outperformed the results reported by Chen et al. (2020) and Chen et al. (2018). The association of the found targets with wound healing is still subject to in depth studies. For example, gene regulatory networks can be obtained through the found association of the gene vectors and be confirmed with experimental studies.

6 Github link

https://github.com/juijayati/Data-Mining-Project-Spring-22.git

7 Member Contribution

Yoshita Buthalapalli:

- Initially Applied word2vec on non-annotated abstracts and derived interactions and plotted graphs. Understood that bigrams or multi word gene names weren't noticed by the model and this lead to the conclusion that we had to use gene tagging.
- Wrote code for annotating the abstracts using the pmid's. Parsed JSON response, replaced gene names with tags, dropped sentences which didn't have any gene tag.
- Wrote base code for using Word2Vec model and using spacy library to remove stop words and lemmatize data.
- Wrote code for plotting t-SNE plots where query is marked red and most-similar are marked blue and least-similar are marked green.
- Analyzing the performance of each algorithm as to why an algorithm might fail and why to try others in the extrinsic evaluation step.
- Wrote base code for fetching gene interactions from STRINGdb, preprocessed it and mapped the names to tags. This was later created into a dataset that helped in extrinsic evaluation; loaded embeddings/vectors and performed cosine evaluation, calculated precision, recall, fscore.

Jayati Jui:

- Plotted the word distribution for the abstract dataset. This was helpful in deciding the word vector size.
- Used the base code for Word2Vec to perform both CBOW and skipgram methods using pretrained BioConceptVec.
- Manually collected all the five genesets for performing the exploratory analysis and analyzed the average similarity for each of the gene sets. Plotted the t-SNE graphs for visualizing the word vectors.
- Performed intrinsic evaluation after extracting C2 and C4 datasets from MSigDB. Calculated the average positive and negative similarity and the average difference for the datasets using our trained model.
- Used R STRINGdb package to collected PPI dataset. Randomly generated negative interactions for the training and testing dataset. Applied different supervised algorithms on the dataset for extrinsic evaluation.

References

Chen, M., Tian, Y., Chen, X., Xue, Z., and Zaniolo, C. (2018). On2vec: Embedding-based relation prediction for ontology population. In *Proceedings of the 2018 SIAM International Conference on Data Mining*, pages 315–323. SIAM.

Chen, Q., Lee, K., Yan, S., Kim, S., Wei, C.-H., and Lu, Z. (2020). Bioconceptvec: Creating and evaluating literature-based biomedical concept embeddings on a large scale. *PLoS computational biology*, 16(4):e1007617.

Ferrante, C. J. and Leibovich, S. J. (2012). Regulation of macrophage polarization and wound healing. *Advances in wound care*, 1(1):10–16.

Li, Y., Zhang, X., He, D., Ma, Z., Xue, K., and Li, H. (2022). 45s5 bioglass® works synergistically with sirna to downregulate the expression of matrix metalloproteinase-9 in diabetic wounds. *Acta Biomaterialia*.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740.

Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N. T., Legeay, M., Fang, T., Bork, P., et al. (2021). The string database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, 49(D1):D605–D612.

Wei, C.-H., Kao, H.-Y., and Lu, Z. (2013). Pubtator: a web-based text mining tool for assisting biocuration. *Nucleic acids research*, 41(W1):W518–W522.

Wei, C.-H., Kao, H.-Y., and Lu, Z. (2015). Gnormplus: an integrative approach for tagging genes, gene families, and protein domains. *BioMed research international*, 2015.