Context-Aware Biomedical Word Embeddings Enhance ADR Prediction: A Shift from Word2Vec to BERT

Woohyuk Jeon, Minjae Park, Doyeon An, Wonshik Nam, Ju-Young Shin, Seunghee Lee, Suehyun Lee

Submitted to: JMIR Medical Informatics
on: October 14, 2024

Table of Contents

Original Manuscript 5
Supplementary Files 21
Figures 22
Figure 1 23
Figure 2 24
Figure 3 25
Figure 4 26
Figure 5 27
Figure 6 28
Figure 7 29
Figure 8 30
TOC/Feature image for homepages 31
TOC/Feature image for homepage 0 32

Context-Aware Biomedical Word Embeddings Enhance ADR Prediction: A Shift from Word2Vec to BERT

Woohyuk Jeon[1*]; Minjae Park[1*]; Doyeon An[1] MPH; Wonshik Nam[1]; Ju-Young Shin[2] PhD; Seunghee Lee[3] PhD; Suehyun Lee[1] PhD

[1]Department of Computer Engineering Gachon University Seongnam KR
[2]School of Pharmacy Sungkyunkwan University Suwon KR
[3]KonYang MEdical data ReseArch group Konyang University Hospital Daejeon KR
[*]these authors contributed equally

Corresponding Author:
Suehyun Lee PhD
Department of Computer Engineering
Gachon University
AI·Engineering Building, 317A
1342 Seongnam-daero, Sujeong-gu, Seongnam-si, Gyeonggi-do
Seongnam
KR

Abstract

Background: Adverse drug reactions (ADRs) pose serious risks to patient health, and effectively predicting and managing them is an important public health challenge. Given the complexity and specificity of biomedical text data, the traditional context-independent language model, Word2Vec, has limitations in fully reflecting the domain specificity of such data. Therefore, to predict drug-side effect relationships more accurately, we applied a Bidirectional Encoder Representations from Transformers (BERT) model specialized for biomedical applications.

Objective: This study aimed to propose a method for extracting drug-side effect relationships from embedding vectors generated by biomedical language models, specifically BERT-based models pre-trained on biomedical corpora. This approach aims to overcome the limitations of the traditional Word2Vec model in accurately capturing complex relationships in biomedical data.

Methods: Using data from 158,096 pairs of drug-side effect relationships from the Side Effect Resource (SIDER) database, we generated an adjacency matrix and calculated the cosine similarity between the word embedding vectors of drugs and side effects. Relation scores were calculated for a total of 8,235,435 drug-side effect pairs using this similarity. To evaluate the prediction accuracy of drug-side effect relationships, the area under the curve (AUC) value was measured using the calculated relation score and 158,096 known drug-side effect relationships provided by SIDER.

Results: The clagator/biobert_v1.1 model achieved an AUC of 0.915 at an optimal threshold of 0.289, largely outperforming the existing Word2Vec model, which had an AUC of 0.848. The BERT-based model pre-trained on the biomedical corpus outperformed the vanilla BERT model, with an AUC of 0.857. Furthermore, external validation with the FDA Adverse Event Reporting System (FAERS) data, using Fisher’s exact test based on 8,235,435 predicted drug-side effect pairs and 901,361 known relationships, confirmed high statistical significance (P<.001) with an odds ratio of 4.830. Additionally, a literature review was conducted for predicted drug-side effect relationships. This review reveals that these relationships have been reported in recent studies published after 2016.

Conclusions: This study introduces a method for extracting drug-side effect relationship data embedded in the pre-trained parameters of language models pre-trained on biomedical corpora and using this information to predict the probability of previously unknown drug-side effect relationships. We improved the accuracy of predicting drug-side effect relationships by using BERT-based models instead of the Word2Vec model. We found that BERT-based models pre-trained with biomedical corpora consider contextual information and achieve better performance in drug-side effect relationship prediction. External validation using the FAERS dataset combined with a literature review of certain cases confirmed high statistical significance, demonstrating the practical applicability of this approach. These results highlight the utility of natural language processing-based approaches for predicting and managing ADRs.

(JMIR Preprints 14/10/2024:67513)
DOI: https://doi.org/10.2196/preprints.67513

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Context-Aware Biomedical Word Embeddings Enhance ADR Prediction: A Shift from Word2Vec to BERT

Abstract

Background: Adverse drug reactions (ADRs) pose serious risks to patient health, and effectively predicting and managing them is an important public health challenge. Given the complexity and specificity of biomedical text data, the traditional context-independent language model, Word2Vec, has limitations in fully reflecting the domain specificity of such data. Therefore, to predict drug-side effect relationships more accurately, we applied a Bidirectional Encoder Representations from Transformers (BERT) model specialized for biomedical applications.

Objective: This study aimed to propose a method for extracting drug-side effect relationships from embedding vectors generated by biomedical language models, specifically BERT-based models pre-trained on biomedical corpora. This approach aims to overcome the limitations of the traditional Word2Vec model in accurately capturing complex relationships in biomedical data.

Methods: Using data from 158,096 pairs of drug-side effect relationships from the Side Effect Resource (SIDER) database, we generated an adjacency matrix and calculated the cosine similarity between the word embedding vectors of drugs and side effects. Relation scores were calculated for a total of 8,235,435 drug-side effect pairs using this similarity. To evaluate the prediction accuracy of drug-side effect relationships, the area under the curve (AUC) value was measured using the calculated relation score and 158,096 known drug-side effect relationships provided by SIDER.

Results: The clagator/biobert_v1.1 model achieved an AUC of 0.915 at an optimal threshold of 0.289, largely outperforming the existing Word2Vec model, which had an AUC of 0.848. The BERT-based model pre-trained on the biomedical corpus outperformed the vanilla BERT model, with an AUC of 0.857. Furthermore, external validation with the FDA Adverse Event Reporting System (FAERS) data, using Fisher's exact test based on 8,235,435 predicted drug-side effect pairs and 901,361 known relationships, confirmed high statistical significance (*P*<.001) with an odds ratio of 4.830. Additionally, a literature review was conducted for predicted drug-side effect relationships. This review reveals that these relationships have been reported in recent studies published after 2016.

Conclusions:
This study introduces a method for extracting drug-side effect relationship data embedded in the pre-trained parameters of language models pre-trained on biomedical corpora and using this information to predict the probability of previously unknown drug-side effect relationships. We improved the accuracy of predicting drug-side effect relationships by using BERT-based models instead of the Word2Vec model. We found that BERT-based models pre-trained with biomedical corpora consider contextual information and achieve better performance in drug-side effect relationship prediction. External validation using the FAERS dataset combined with a literature review of certain cases confirmed high statistical significance, demonstrating the practical applicability of this approach. These results highlight the utility of natural language processing-based approaches for predicting and managing ADRs.

Keywords: adverse drug reaction; ADR prediction; NLP; BERT; word embedding;

Introduction

An adverse drug reaction (ADR) is a harmful, unintended reaction that occurs despite the proper use of medication [1]. In addition to causing serious health problems, ADRs are known to be one of the leading causes of prolonged patient hospitalization and increased healthcare spending [2]. Approximately 2 million cases of serious ADRs are reported annually in the United States, resulting in 100,000 deaths [3]. Therefore, early prediction and prevention of ADRs during drug development is a critical challenge for patient safety and public health.

Traditionally, ADR prediction has been based on approaches that analyze the chemical structure, mechanism of action, and pharmacokinetic properties of drugs [4]. However, with recent advances in natural language processing (NLP) techniques, attempts have been made to automatically extract and predict drug-side effect relationships from vast amounts of biomedical literature data [5-7]. The development of machine learning-based ADR prediction models is accelerating, especially with the advent of word embedding techniques such as Word2Vec [8], which can effectively vectorize semantic information embedded in textual data.

However, biomedical text data are characterized by a much more specialized and complex set of terms and concepts compared with the general literature, and the interactions between them are also highly diverse and dynamic [9]. Therefore, a general-purpose language model that does not adequately reflect these domain specificities is limited in its ability to accurately capture drug-side effect relationships. In fact, it has been pointed out that traditional word embedding models such as Word2Vec, which do not consider contextual information, do not sufficiently represent the relationships between complex biomedical concepts [10].

One solution to this problem is to use language models based on Bidirectional Encoder Representations from Transformers (BERT) [11] to perform word embedding. BERT is a language model based on the transformer [12] architecture, which has recently gained attention; unlike traditional one-way language models, it has richer language expressiveness by learning context in both directions. In addition, because we trained on large corpora, domain-specific pre-trained models using large biomedical corpora can fully reflect the domain specificity of the biomedical text data.

Recently, several BERT-based models have been proposed that utilize large biomedical corpora such as PubMed and PMC for domain-specific pre-training. Examples include BioBERT [13], BioMedBERT [14], and PharmBERT [15], which have demonstrated high performances in various bio-NLP tasks.

Therefore, in this study, based on the ADR prediction methodology proposed by Lim [16], we replace the word-embedding model with a biomedical domain-specific BERT model in Word2Vec based on the performance improvement of ADR prediction that occurs through comparative experiments.

Methods

System Overview

Figure 1 presents an overview of this study and illustrates the overall research flow from data collection to validation.

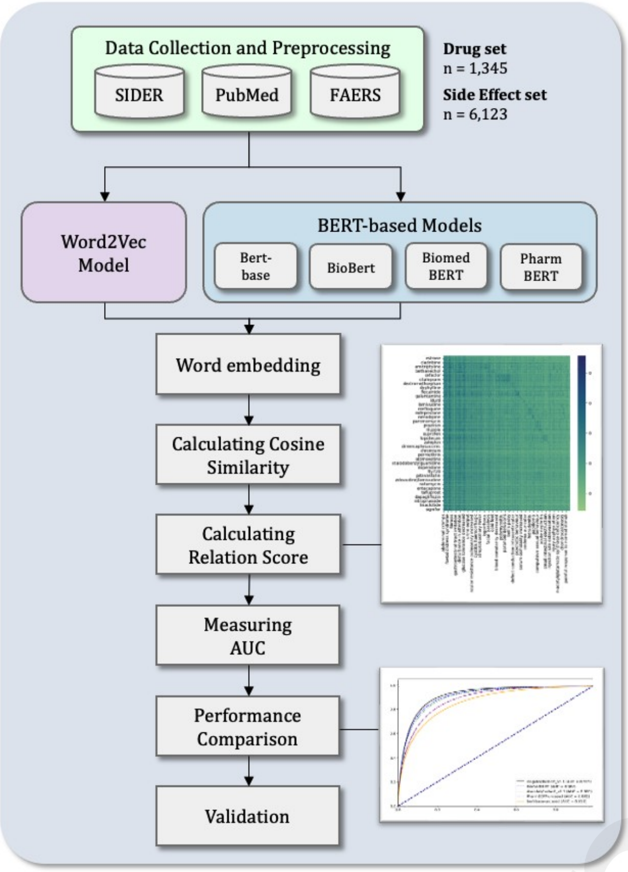


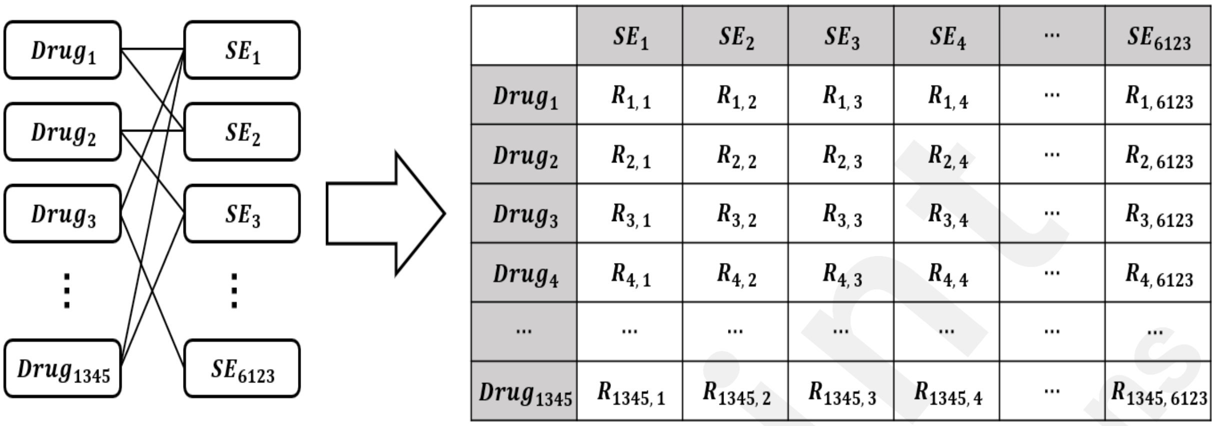
Figure 1. System overview: The predictive performance of drug-side effect relationships was evaluated using area under the curve.

First, we collected and refined the data for this study from the Side Effect Resource (SIDER) and PubMed. From the abstract sentences collected from PubMed, we selectively extracted only sentences containing drugs and side effects mentioned in SIDER. These extracted sentences were used to train a Word2Vec model, from which embedding vectors for drugs and side effects were derived. Based on the drug-side effect relationships in SIDER, BERT-based models were used to derive the embedding vectors of drugs and side effects.

Based on the derived embedding vectors, the cosine similarity between the drug and side effect pairs was calculated. Using the cosine similarity of drug and side effect pairs and existing known drug-side effect relationships, a relation score was calculated for all drug-side event combinations. Drug-side effect combinations with high relation scores were predicted to have a higher likelihood of being actually related [16]. To evaluate the accuracy of these predictions, we calculated area under the curve (AUC) values and compared the results of the Word2Vec model pipeline with those of the BERT-based model pipeline. The Fisher's exact test was used to verify the statistical significance of the predicted results.

Data Collection and Preprocessing

SIDER is a database that provides information on marketed drugs and their side effects [17]. The drug names recorded in SIDER followed those approved by the Food and Drug Administration (FDA), and side effect names used the Medical Dictionary for Regulatory Activities (MedDRA) terminology. Using version 4.1 of SIDER, we collected 158,096 unique pairs of drug-side effect relationships after removing duplicates. To use these 158,096 pairs as input values in the BERT-based models and for relation score calculations, we derived an adjacency matrix with drugs as rows and side effects as columns (Figure 2). In addition, 1,345 drug names and 6,123 side effect terms that appeared in the collected drug-side effect relationships were extracted and used as dictionaries for drugs and side effects.



	SE_1	SE_2	SE_3	SE_4	...	SE_{6123}
$Drug_1$	$R_{1,1}$	$R_{1,2}$	$R_{1,3}$	$R_{1,4}$	...	$R_{1,6123}$
$Drug_2$	$R_{2,1}$	$R_{2,2}$	$R_{2,3}$	$R_{2,4}$	...	$R_{2,6123}$
$Drug_3$	$R_{3,1}$	$R_{3,2}$	$R_{3,3}$	$R_{3,4}$	...	$R_{3,6123}$
$Drug_4$	$R_{4,1}$	$R_{4,2}$	$R_{4,3}$	$R_{4,4}$	...	$R_{4,6123}$
...	...	...	...	...	...	...
$Drug_{1345}$	$R_{1345,1}$	$R_{1345,2}$	$R_{1345,3}$	$R_{1345,4}$	...	$R_{1345,6123}$

Figure 2. Drug-SE adjacency matrix. This is a method to derive an adjacency matrix using drug-side effect relationships in SIDER. The relation R has the value of 1 if the drug-side effect relationship exists and 0 if it does not. SE: side effect.

We collected biomedical literature from PubMed, a biological literature database [18]. A total of 42,515,246 paper abstracts updated on December 8, 2022, were collected, and for training the Word2Vec model, only sentences in which the drugs and side effects mentioned in SIDER were mentioned at least once were extracted [16]. There were 14,289,160 sentences in which a drug was mentioned at least once and 32,107,327 sentences in which a side effect was mentioned at least once.

Calculating Cosine Similarity

For the 1345 drugs and 6123 side effects recorded in the adjacency matrix, we performed word embedding using BERT-based models and calculated the cosine similarity for all drug and side effect vector pairs. In this case, the cosine similarity is calculated using Equation (1).

$$Cosine\,Similarity=\frac{A\cdot B}{\|A\|\|B\|}=\frac{\sum_{i=1}^{n}A_i\times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\times\sqrt{\sum_{i=1}^{n}(B_i)^2}}\quad(1)$$

This process yielded 1,809,025 drug vector pairwise similarities and 37,491,129 side effect vector pairwise similarities.

Calculating Relation Score

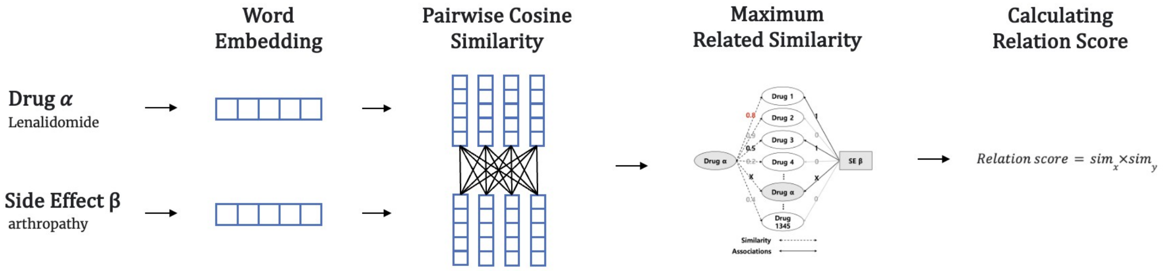


Figure 3. Illustration of the computation of the relation score between a specific $Drug_{\alpha}$ and a specific side effect SE_{β} .

Figure 3 shows the process of calculating the relation score. For all drug-side effect pairs embedded as vectors, the cosine similarity values obtained in the previous step were used to calculate the drug-side effect's relation score. The process of calculating the relation score between a specific $Drug_{\alpha}$ and a specific side effect SE_{β} was done using Equations (2) to (4).

$$sim_x = max\{\sim(D_{\alpha}, D_i) \vee D_i \in Related_{Drugs}\} \qquad (2\text{-}1)$$

$$Related_{Drugs} = \{D_i \vee Adjacency(SE_{\beta,} D_i) = 1 \, for \, i = 1, 2, \ldots, 1345\} \qquad (2\text{-}2)$$

$$sim_y = max\{\sim(SE_{\beta}, SE_i) \vee SE_i \in Related_{SE}\} \qquad (3\text{-}1)$$

$$Related_{SE} = \{SE_i \vee Adjacency(D_{\alpha}, SE_i) = 1 \, for \, i = 1, 2, \ldots, 6123\} \qquad (3\text{-}2)$$

$$Relation\,score = sim_x \times sim_y \qquad (4)$$

Similarity $¿_x$ takes the maximum of the similarity values of $Drug_{\alpha}$ with Drug D_i in Related_Drugs, the set of drugs known to be associated with side effect SE β, using Equation (2-1). The set Related_Drugs is obtained using Equation (2-2), and by referring to the values in the adjacency matrix consisting of 1345 drugs and 6123 side effects, we construct the set of drugs associated with that side effect by including in the set Related_Drugs those drugs that have side effect SE, and a value of 1 in the adjacency matrix, out of a total of 1345 drugs. In other words, the highest similarity value to drugs known to be associated with side effect SE is called similarity $¿_x$.

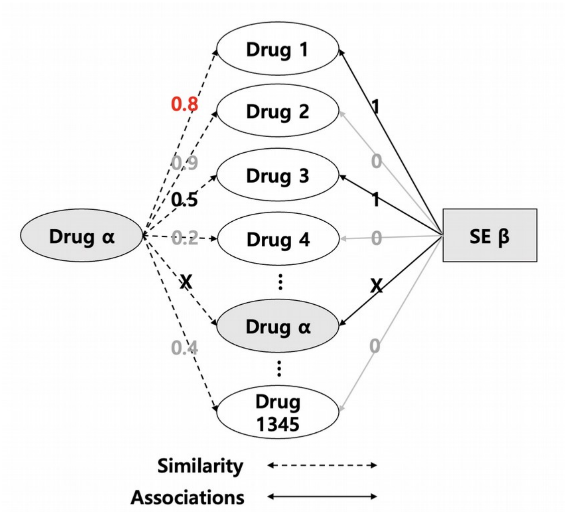


Figure 4. Process of calculating similarity sim_x.

Figure 4 shows the process of obtaining similarity $¿_x$ using Equation (2-1) and Equation (2-2). Based on the values in the adjacency matrix, the computational process was to maximize the similarity of only those drugs that were related to the side effect SE out of the total 1345 drugs. If $Drug_\alpha$ is in the Related_Drugs set, exclude $Drug_\alpha$ from the similarity calculation.

Similarity sim_y uses Equation (3-1) to take the maximum of the similarity values of side effects SE_β and SE_i in Related_SE, a set of side effects known to be related to the $Drug_\alpha$. The set Related_SE is obtained using Equation (3-2), and by referring to the values of the adjacency matrix mentioned above, we construct the set of side effects associated with the $Drug_\alpha$ out of the total 6123 side effects by including in the set Related_SE the drugs that have a value of 1 in the adjacency matrix with the $Drug_\alpha$. In other words, the highest similarity value to the side effects that are known to be related to the drug is called the similarity sim_y.

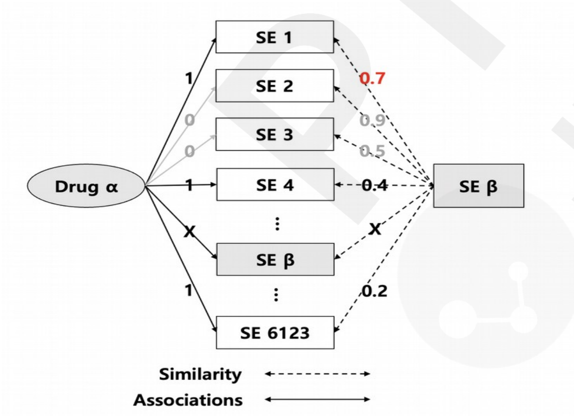


Figure 5. Process of calculating similarity sim_y.

Figure 5 shows the process of obtaining similarity sim_y using Equation (3-1) and Equation (3-2). Based on the values in the adjacency matrix, the computational process is to extract the similarity of only the side effects that are related to the $Drug_{\alpha,}$, out of the total 6123 side effects, and take the maximum value. If an SE_β belongs to the Related_SE set, exclude the SE_β from the similarity calculation.

Finally, the relation score between Drug and SE was obtained by multiplying sim_x and sim_y as shown in Equation (4).

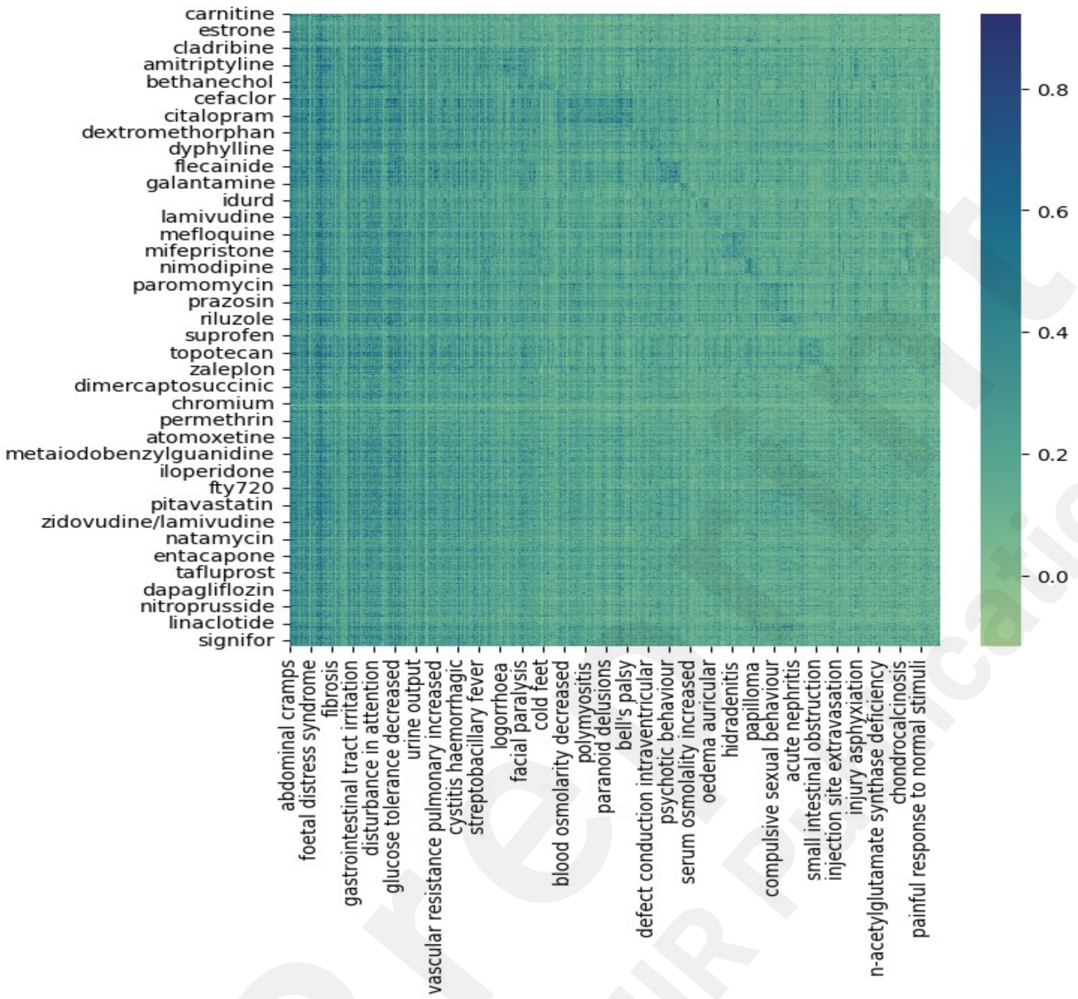


Figure 6. Heatmap of calculated relation score for 8,235,435 drug-side effect pairs.

We applied the above calculation method to 1345 drugs and 6123 side effects to calculate the relation scores for all drug-side effect pairs, resulting in a total of 8,235,435 drug-side effect pairs. Figure 6 shows the heatmap of the calculated relation scores.

Measuring AUC

In this study, AUC values were measured using 158,096 known drug-side-effect relationships provided by SIDER to evaluate the accuracy of predicting drug-side-event relationships based on scores calculated for a total of 8,235,435 drug-side-event pairs. Of the 8,235,435 calculated drug-side effect pairs, we assigned a class value of True to pairs that belonged to known drug-side effect relationships in SIDER and False to pairs that did not. All drug-side effect pairs were sorted by score, and a single receiver operating characteristic (ROC) curve was calculated. The generated ROC curves and AUC values were utilized to establish the optimal threshold for predicting whether a drug-side effect pair had a true relationship. If the drug-side effect relation score exceeds this diagnostic threshold, it is predicted that there is a relationship between the drug and the side effect [16].

Results

Performance Comparison

The AUCs of the BERT-based models using the proposed method are compared in Table 1 and Figure 7. The optimal threshold for prediction was set as the point at which the sum of the sensitivity and specificity was maximized. The models used in this study included clagator/biobert_v1.1 [19], BiomedBERT [14], dmis-lab/biobert_v1.1 [13], PharmBERT-uncased [15], and bert-base-uncased [11].

The clagator/biobert _v1.1 model achieved the highest AUC value of 0.915 at an optimal threshold of 0.289. In contrast, the bert-base-uncased model, a vanilla BERT model pre-trained on general corpora, showed the lowest performance with an AUC of 0.857 at an optimal threshold of 0.617. In other words, BERT pre-trained on the biomedical corpus outperformed vanilla BERT.

Table 1. Performance comparison of BERT-based models.

Model	AUC	Optimal threshold	Sensitivity	Specificity
clagator/biobert_v1.1	0.915[c]	0.289	0.870	0.830
BiomedBERT[a]	0.907	0.925	0.857	0.821
dmis-lab/biobert_v1.1	0.901	0.780	0.851	0.814
PharmBERT-uncased	0.882	0.460	0.817	0.796
bert-base-uncased[b]	0.857[d]	0.617	0.769	0.793

[a] The old model was named PubMedBERT

[b] Vanilla BERT model

[c] Highest value

[d] Lowest value

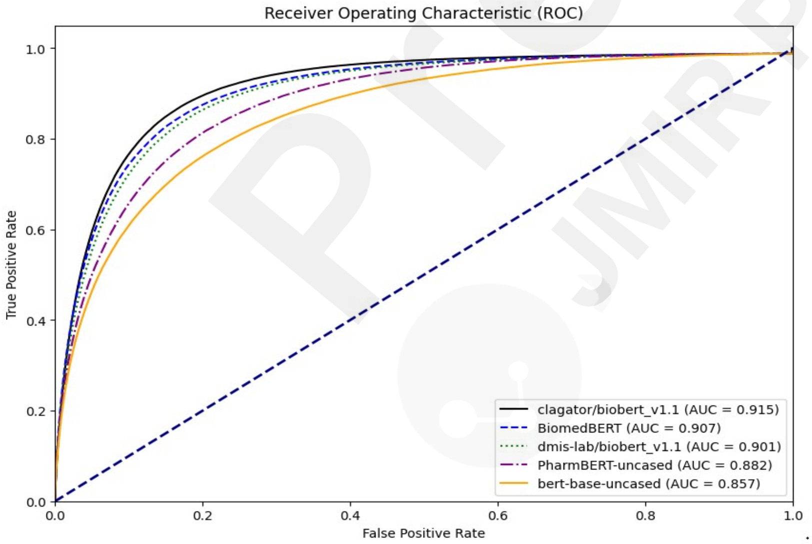


Figure 7. ROC curves for BERT-based models.

Table 2 and Figure 8 present the AUC obtained by reproducing the methodology of a previous study using the Word2vec model [16]. The AUC of the Word2Vec model was 0.85, with an optimal threshold of 0.11. The AUC of the Word2Vec model was lower than those of the BERT-based models, with the lowest value recorded among the models used in this study.

Table 2. Performance comparison of word2vec models.

Model	AUC	Optimal threshold	Sensitivity	Specificity
word2vec	0.848[a]	0.112	0.762	0.780

[a] Values obtained by reproducing previous research.

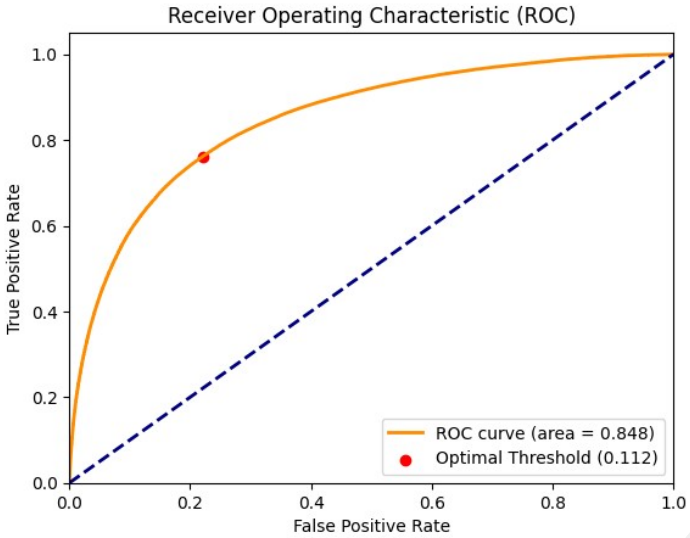


Figure 8. ROC curve for word2vec model.

Validation

For validation, we extracted drug-side-effect relationships from the FDA Adverse Event Reporting System (FAERS), a database not used in our methods. The FAERS database contains information on adverse drug events that are submitted to the FDA [20]. We used FAERS data from October 2012 to June 2023. After filtering the data with the 1,345 drugs and 6,123 side effects mentioned in SIDER, we extracted 901,361 known drug-side effect relationships from FAERS.

For validation, we used the results of clagator/biobert _v1.1 [19], which performed the best in this study. We conducted Fisher's exact test based on a contingency table using the prediction results from the relation scores of 8,235,435 drug-side effect pairs and 901,361 known drug-side effect relationships extracted from FAERS. The Fisher's exact test was repeated 100 times, and the average of all results was calculated. The results showed a p-value of *P*<.001 and an odds ratio of 4.830 (Table 3).

Table 3. *P*-value and odds ratio results from Fisher's exact test.

Average results from 100 repetitions of Fisher's exact test [a]	
P-value	Odds ratio
<.001	4.830

[a] clagator/biobert_v1.1 model

To validate the utility of our model's predictions, we analyzed case reports for a subset of drug-side effect relationships that were not present in the SIDER database but were predicted by our model with high association scores. Table 4 shows the cases of clobetasol-cushingoid, lenalidomide arthropathy, rosuvastatin-sleep disturbance, and gadolinium-acute pulmonary oedema. Upon conducting a literature review of the drug-side effect relationships shown in Table 4, we confirmed that these associations were reported in recent research findings published after 2016 [21-24].

Table 4. Case studies of model predictions

Drug α	Side Effect β	Similarity X	Similarity Y	Relation Score	Model Prediction
clobetasol	cushingoid	0.850	0.979	0.833	True
lenalidomide	arthropathy	0.953	0.837	0.799	True
rosuvastatin	sleep disturbance	0.952	0.825	0.786	True
gadolinium	acute pulmonary oedema	0.891	0.866	0.771	True

Discussion

In this study, we propose a method to extract information about drug-adverse effect relationships inherent in the pre-trained parameters of language models and predict relationship scores, indicating the possibility of unknown drug-adverse effect relationships. This is accomplished using known drug-adverse effect relationship data and embedding vectors from language models trained on biomedical corpora.

This study extends previous drug-side effect prediction research that used the Word2Vec model to predict the relationship between drugs and side effects more precisely [16]. Instead of the context-independent nature of Word2Vec, we applied context-based BERT-based models pre-trained on biomedical corpora to achieve more accurate predictions of drug-side effect relationships.

In the field of biomedical text mining, it is important to effectively model complex terms, concepts, and their interactions [13]. However, Word2Vec, as a context-independent model, has limited ability to fully capture this complexity [25-27]. To address this issue, we introduced BERT-based models instead of the traditional Word2Vec model to improve performance.

Our study confirmed that BERT-based models demonstrated superior performance in predicting drug-side-effect relationships. We evaluated the performance of BERT-based models using the relation score methodology proposed by Lim [16], and the clagator/biobert_v1.1 [19] model achieved the highest performance with an AUC of 0.915 at an optimal threshold of 0.289. This suggests that BERT-based models perform better in predicting drug-side effect relationships compared to the 0.85 AUC achieved by the Word2Vec model in a previous study. Therefore, our findings support the notion that context-aware BERT-based models outperform context-independent Word2Vec models in terms of embedding performance [26].

Additionally, our study demonstrates that BERT models pre-trained on biomedical corpora outperform vanilla BERT models pre-trained on general corpora. The vanilla BERT models trained on general corpora do not fully capture the specificity of the biomedical field [14]. To overcome this limitation, BERT-based models pre-trained on biomedical data were developed [13-15, 28, 29]. We compared the performance of BERT models pre-trained on biomedical data with that of vanilla BERT and found that the former showed superior performance. These results demonstrate that BERT models specialized for biomedical applications can provide more accurate drug-side effect relationship predictions based on a deeper understanding of the domain. This aligns with previous studies that emphasize the importance of domain-specific models in BERT model applications [13, 15, 30], and our research further highlights the necessity and importance of developing domain-specific BERT models for specific tasks.

We performed external validation using FAERS data and found a high statistical significance (*P*<.001) between 8,235,435 predicted drug-side effect relationships and 901,361 actual data extracted from FAERS. Additionally, to verify the real-world applicability of the model's predicted results, we conducted case studies on drug-side effect relationships that were not confirmed in the SIDER database. We found that these drug-side effect relationships have been reported in recent research findings published after 2016. This suggests that our methodology using the BERT-based model proposed in this study is applicable to the prediction of ADRs in practice. Considering this, we expect that our proposed methodology will allow for earlier detection of potential ADRs, increasing the likelihood of success in the drug development process and reducing the time and cost of ADR studies.

Limitations

One of the limitations of this study is that the drug side effect database used was not composed of the most up-to-date data. The most recent version of the SIDER database was updated in 2015. As a result, despite using BERT-based models trained on the latest biomedical corpus, the prediction process using embedding vectors from the language model may not fully reflect current drug-side effect relationships. If more recent drug-side effect data were incorporated, it is highly likely that the performance of the prediction model could be further improved.

Another limitation is the reliability of the case reports used to validate the predicted drug-side-effect relationships. Case reports rely on a single clinically reported case, making it difficult to establish clear causal relationships between drugs and adverse reactions, and their small sample sizes can limit their generalizability. Therefore, future research should utilize systematic clinical data or large-scale cohort studies to enhance the reliability of predictive models.

Conclusions

This study proposes a method to extract drug-side effect relationship information inherent in the pre-trained parameters of language models and predict the likelihood of unknown drug-side effect relationships based on that information. BERT-based models pre-trained with biomedical data have the advantage of being context-dependent, allowing them to deeply understand the multilayered meaning of complex biomedical data. However, traditional Word2vec models are limited in their ability to fully capture this complexity owing to their context-independent nature. In this study, we extended the methodology using Word2Vec from previous research to a methodology using BERT-based models for drug side-effect prediction, resulting in improved accuracy. Additionally, by comparing BERT-based models, we explored a model optimized for predicting drug-side effect relationships. Furthermore, external validation using FAERS data and a literature review of selected cases emphasized the practical applicability of the proposed methodology.

Author’s Contributions

MP and WJ designed the study and drafted the manuscript. WJ was responsible for data collection, preprocessing, and validation. MP was responsible for the relation score calculation methodology and implementation of the BERT model. WN conducted the reproduction study using Word2vec. DA and JS contributed to the manuscript revision. SL (Suehyun Lee) and SL (Seunghee Lee) contributed equally to and supervised the study. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

None declared.

Abbreviations

ADR: Adverse Drug Reaction
AUC: Area Under the Curve
BERT: Bidirectional Encoder Representations from Transformers
FAERS: The FDA Adverse Event Reporting System
FDA: Food and Drug Administration
MedDRA: Medical Dictionary for Regulatory Activities
NLP: Natural Language Processing
ROC: Receiver Operating Characteristic
SE: Side Effect
SIDER: Side Effect Resource

References

1. Edwards IR, Aronson JK. Adverse drug reactions: definitions, diagnosis, and management. The Lancet. 2000;356(9237):1255-9. doi: 10.1016/S0140-6736(00)02799-9.
2. Pirmohamed M, James S, Meakin S, Green C, Scott AK, Walley TJ, et al. Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. Bmj. 2004;329(7456):15-9. doi: 10.1136/bmj.329.7456.15.
3. Lazarou J, Pomeranz BH, Corey PN. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. Jama. 1998;279(15):1200-5. doi: 10.1001/jama.279.15.1200.
4. Huang L-C, Wu X, Chen JY. Predicting adverse side effects of drugs. BMC Genomics. 2011;12:1-10. doi: 10.1186/1471-2164-12-S5-S11.
5. Leaman R, Wojtulewicz L, Sullivan R, Skariah A, Yang J, Gonzalez G, editors. Towards internet-age pharmacovigilance: extracting adverse drug reactions from user posts in health-related social networks. Proceedings of the 2010 workshop on biomedical natural language processing; 2010.
6. Gurulingappa H, Mateen‐Rajpu A, Toldo L. Extraction of potential adverse drug events from medical case reports. Journal of Biomedical Semantics. 2012;3:1-10. doi: 10.1186/2041-1480-3-15.
7. Nikfarjam A, Sarker A, O'connor K, Ginn R, Gonzalez G. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. Journal of the American Medical Informatics Association. 2015;22(3):671-81. doi: 10.1093/jamia/ocu041.
8. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems. 2013;26. doi: 10.48550/arXiv.1310.4546 Focus to learn more.
9. Cohen KB, Demner-Fushman D. Biomedical natural language processing: John Benjamins; 2014. ISBN: 9027271062.
10. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word

embeddings with subword information and MeSH. Scientific Data. 2019;6(1):52. doi: 10.1038/s41597-019-0055-0.

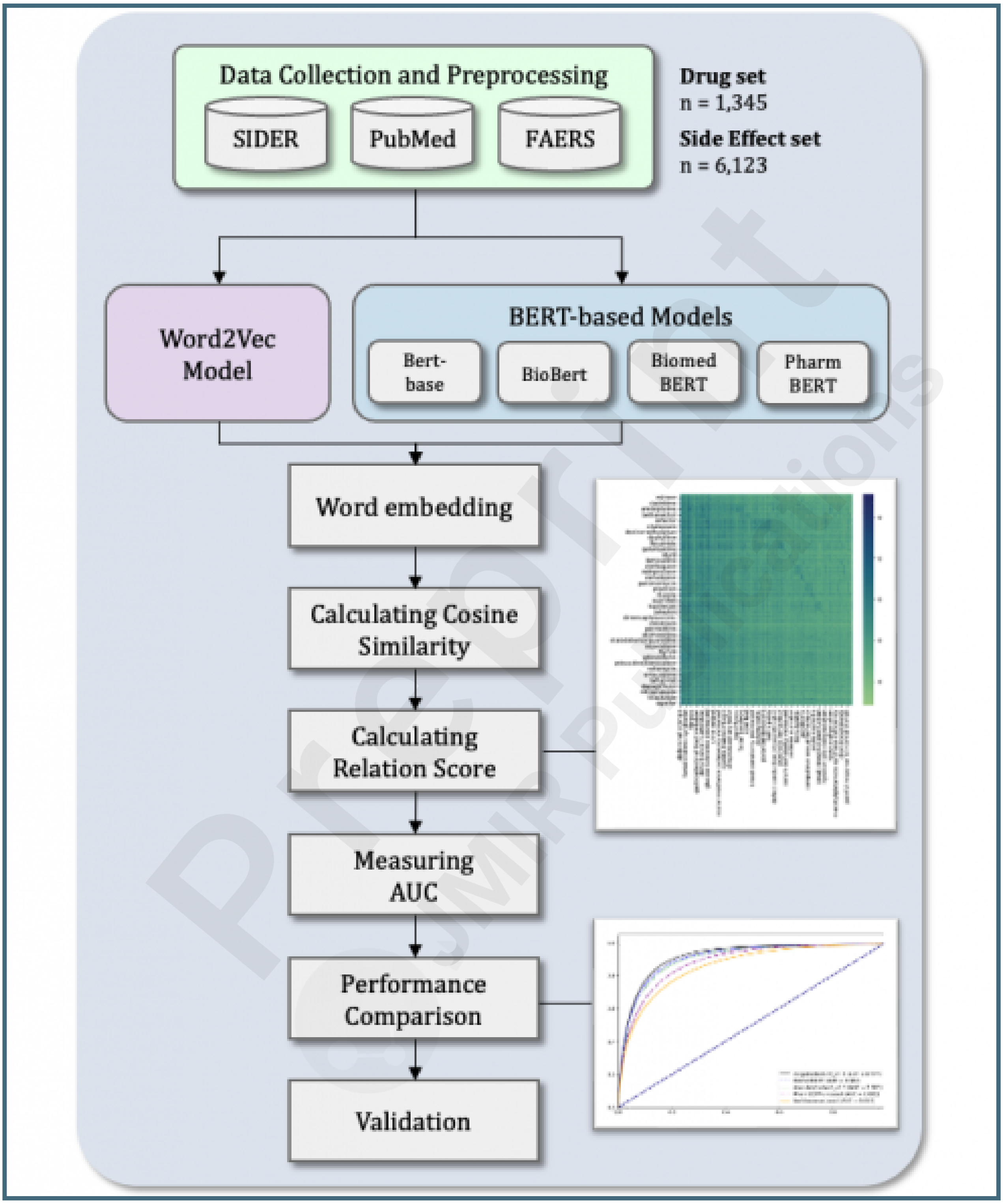
11. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805. 2018. doi: 10.48550/arXiv.1810.04805.
12. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30. doi: 10.48550/arXiv.1706.03762.
13. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234-40. doi: 10.1093/bioinformatics/btz682.
14. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH). 2021;3(1):1-23. doi: 10.1145/3458754.
15. ValizadehAslani T, Shi Y, Ren P, Wang J, Zhang Y, Hu M, et al. PharmBERT: a domain-specific BERT model for drug labels. Briefings in Bioinformatics. 2023;24(4):bbad226. doi: 10.1093/bib/bbad226.
16. Seungsoo L, Hayon L, Youngmi Y. Prediction of New Drug-Side Effect Relation using Word2Vec Model-based Word Similarity. The Journal of Korean Institute of Information Technology. 2020;18(11):25-33. doi: 10.14801/jkiit.2020.18.11.25.
17. Kuhn M, Letunic I, Jensen LJ, Bork P. The SIDER database of drugs and side effects. Nucleic Acids Research. 2016;44(D1):D1075-D9. doi: 10.1093/nar/gkv1075.
18. Canese K, Weis S. PubMed: the bibliographic database. The NCBI handbook. 2013;2(1).
19. biobert_v1.1_pubmed_nli_sts. Hugging Face Hub. URL: https://huggingface.co/clagator/biobert_v1.1_pubmed_nli_sts [accessed 2023-08-23].
20. US Food and Drug Administration. Questions and Answers on FDA's Adverse Event Reporting System (FAERS). URL: https://www.fda.gov/drugs/surveillance/questions-and-answers-fdas-adverse-event-reporting-system-faers [accessed 2023-08-23].
21. Ghirardo S, De Nardi L, Tommasini A, Barbi E, Tornese G. Topical clobetasol: an overlooked cause of Cushing syndrome. Endocrine, Metabolic & Immune Disorders-Drug Targets (Formerly Current Drug Targets-Immune, Endocrine & Metabolic Disorders)
22. Icard C, Mocquot P, Nogaro J-C, Despas F, Gauthier M. Lenalidomide-induced arthritis: A case report and review of literature and pharmacovigilance databases. Journal of Oncology Pharmacy Practice. 2022;28(2):453-6. doi: 10.1177/10781552211038001.. 2021;21(12):2300-2. doi: 10.2174/1871530321666210426131423.
23. Trambowicz K, Gorzelak-Pabiś P, Broncel M. Statins And Sleep–Clinical Effects. Atherosclerosis. 2019;287:e202. doi: 10.1016/j.atherosclerosis.2019.06.613.
24. Lucas A, Mohan G, Winkler A, Gardner K, Whalen M. Acute lung injury following gadolinium contrast: a case report. Journal of Emergency and Critical Care Medicine. 2021;5. doi: 10.21037/jeccm-20-117.
25. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L, editors. Deep Contextualized Word Representations. 2018 June; New Orleans, Louisiana: Association for Computational Linguistics.
26. Si Y, Wang J, Xu H, Roberts K. Enhancing clinical concept extraction with contextual embeddings. Journal of the American Medical Informatics Association. 2019;26(11):1297-304. doi: 10.1093/jamia/ocz096.
27. Di Gennaro G, Buonanno A, Palmieri FA. Considerations about learning Word2Vec. The Journal of Supercomputing. 2021:1-16. doi: 10.1007/S11227-021-03743-2.
28. Beltagy I, Lo K, Cohan A. SciBERT: A pretrained language model for scientific text. arXiv preprint arXiv:190310676. 2019. doi: 10.48550/arXiv.1903.10676.

29. Huang K, Altosaar J, Ranganath R. Clinicalbert: Modeling clinical notes and predicting hospital readmission. arXiv preprint arXiv:190405342. 2019. doi: 10.48550/arXiv.1904.05342.
30. Khadhraoui M, Bellaaj H, Ammar MB, Hamam H, Jmaiel M. Survey of BERT-base models for scientific text classification: COVID-19 case study. Applied Sciences. 2022;12(6):2891. doi: 10.3390/app12062891

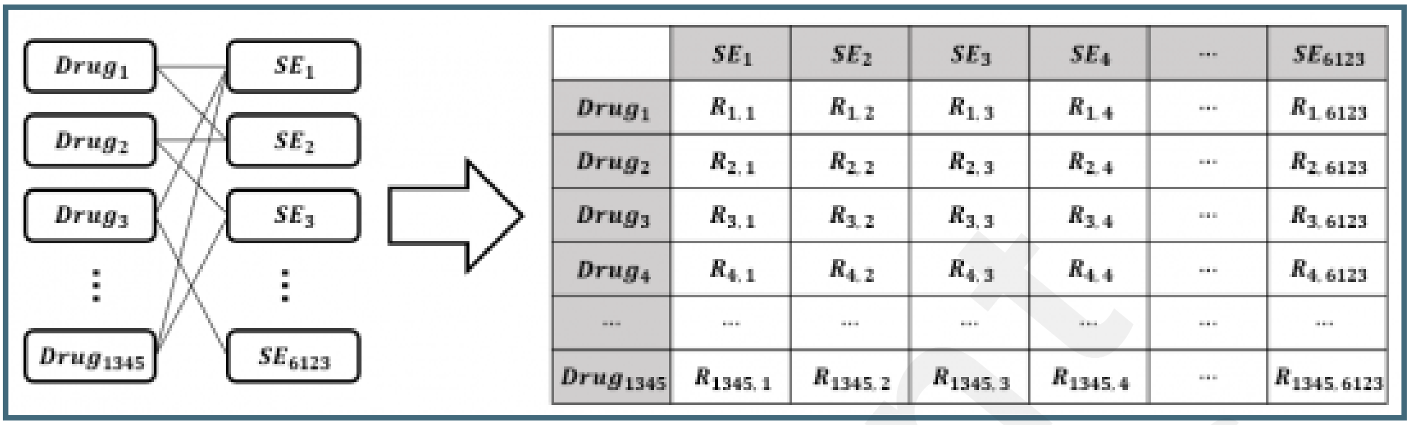
Supplementary Files

Figures

System overview: The predictive performance of drug-side effect relationships was evaluated using area under the curve.



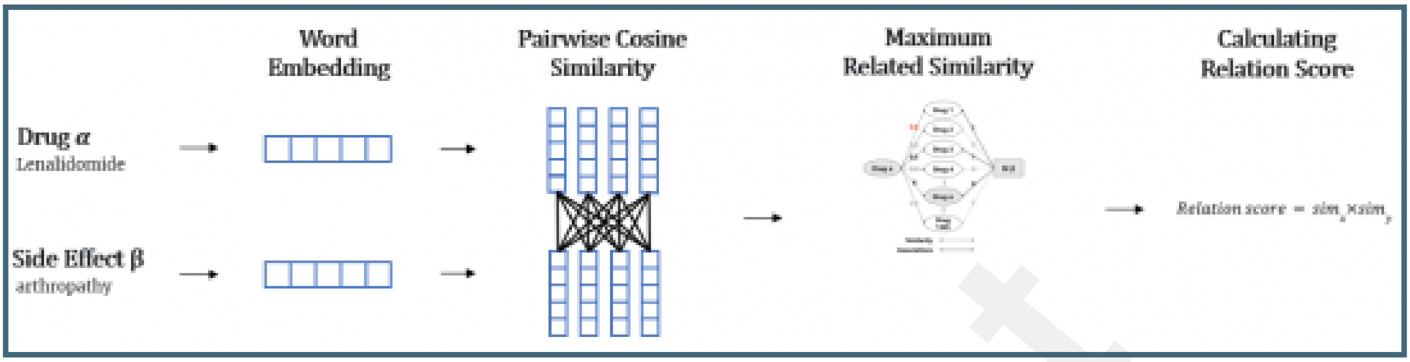
Drug-SE adjacency matrix. This is a method to derive an adjacency matrix using drug-side effect relationships in SIDER. The relation R has the value of 1 if the drug-side effect relationship exists and 0 if it does not. SE: side effect.



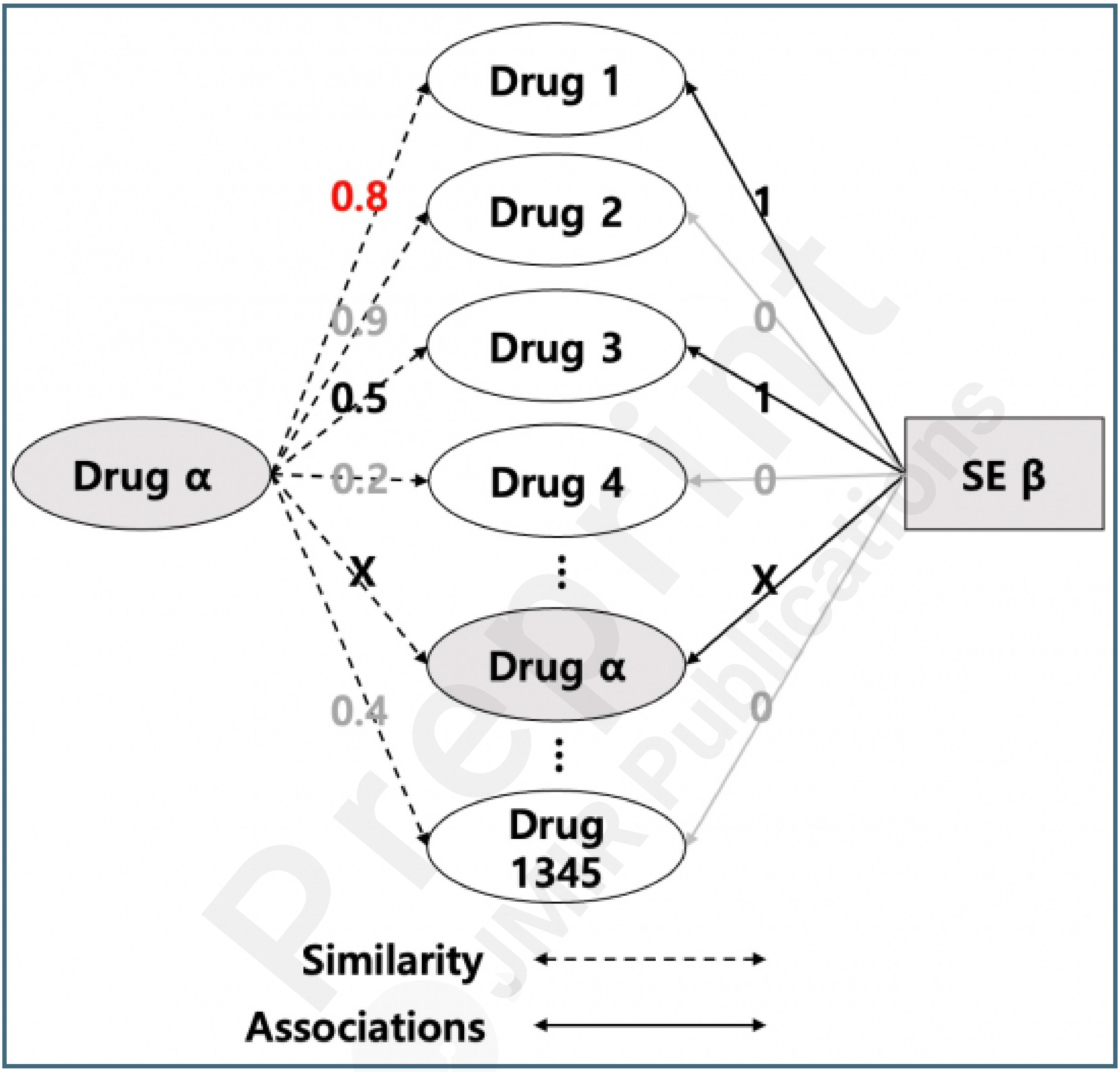
	SE_1	SE_2	SE_3	SE_4	...	SE_{6123}
$Drug_1$	$R_{1,1}$	$R_{1,2}$	$R_{1,3}$	$R_{1,4}$	...	$R_{1,6123}$
$Drug_2$	$R_{2,1}$	$R_{2,2}$	$R_{2,3}$	$R_{2,4}$	...	$R_{2,6123}$
$Drug_3$	$R_{3,1}$	$R_{3,2}$	$R_{3,3}$	$R_{3,4}$	...	$R_{3,6123}$
$Drug_4$	$R_{4,1}$	$R_{4,2}$	$R_{4,3}$	$R_{4,4}$	...	$R_{4,6123}$
...	...	...	...	...	...	...
$Drug_{1345}$	$R_{1345,1}$	$R_{1345,2}$	$R_{1345,3}$	$R_{1345,4}$	...	$R_{1345,6123}$

]

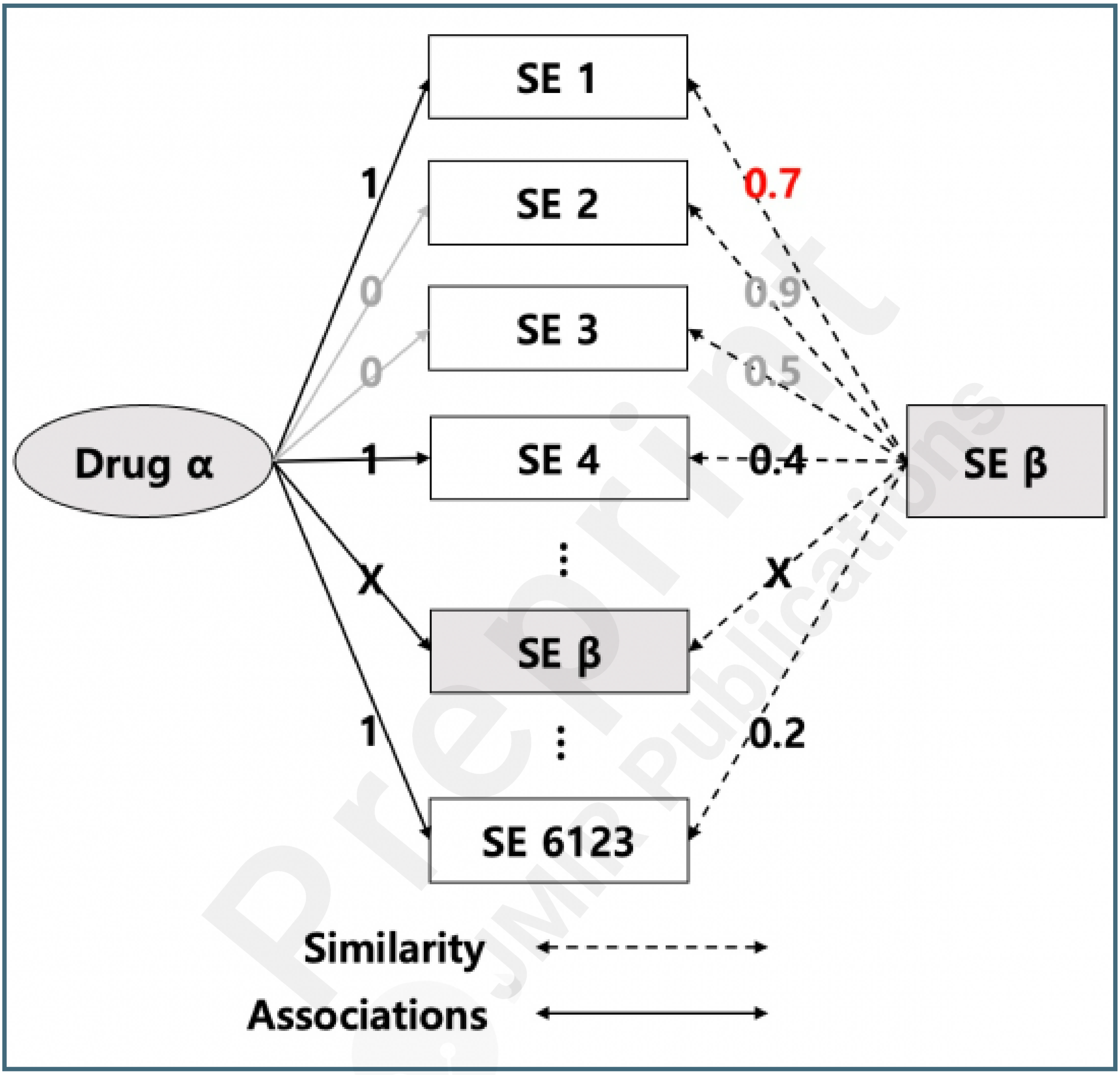
Illustration of the computation of the relation score between a specific Drug? and a specific side effect SE?.



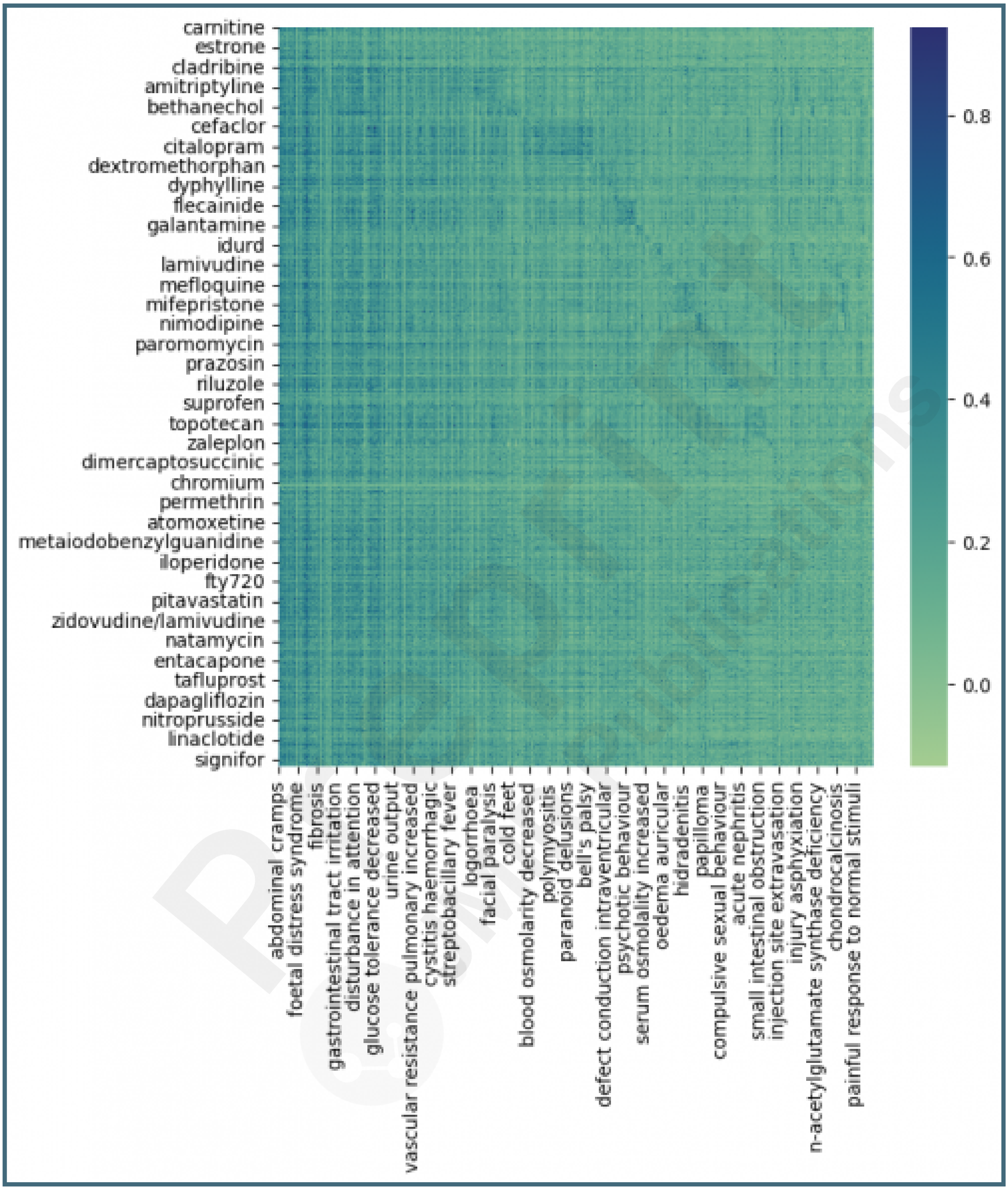
Process of calculating similarity sim?.



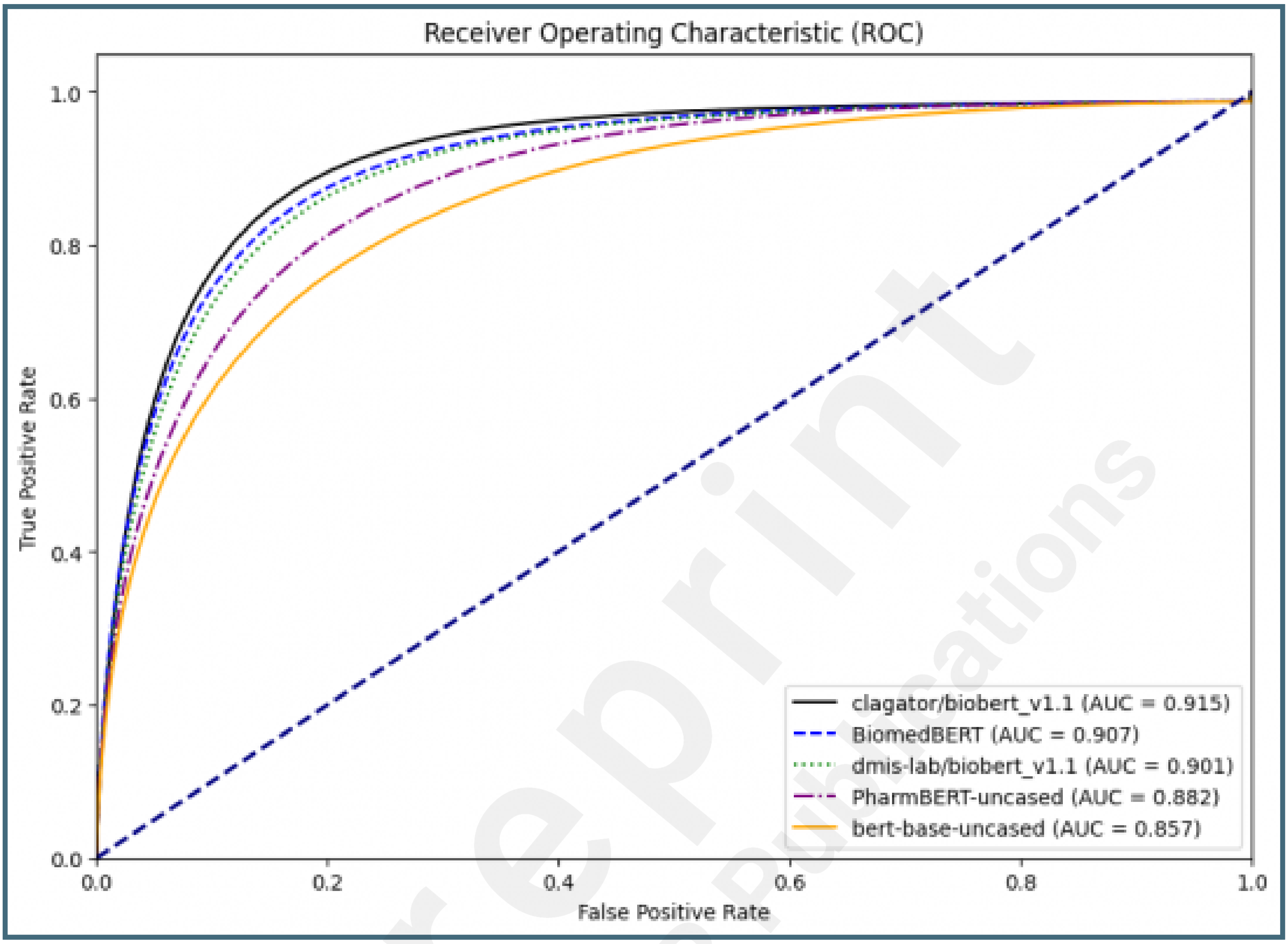
Process of calculating similarity sim?.



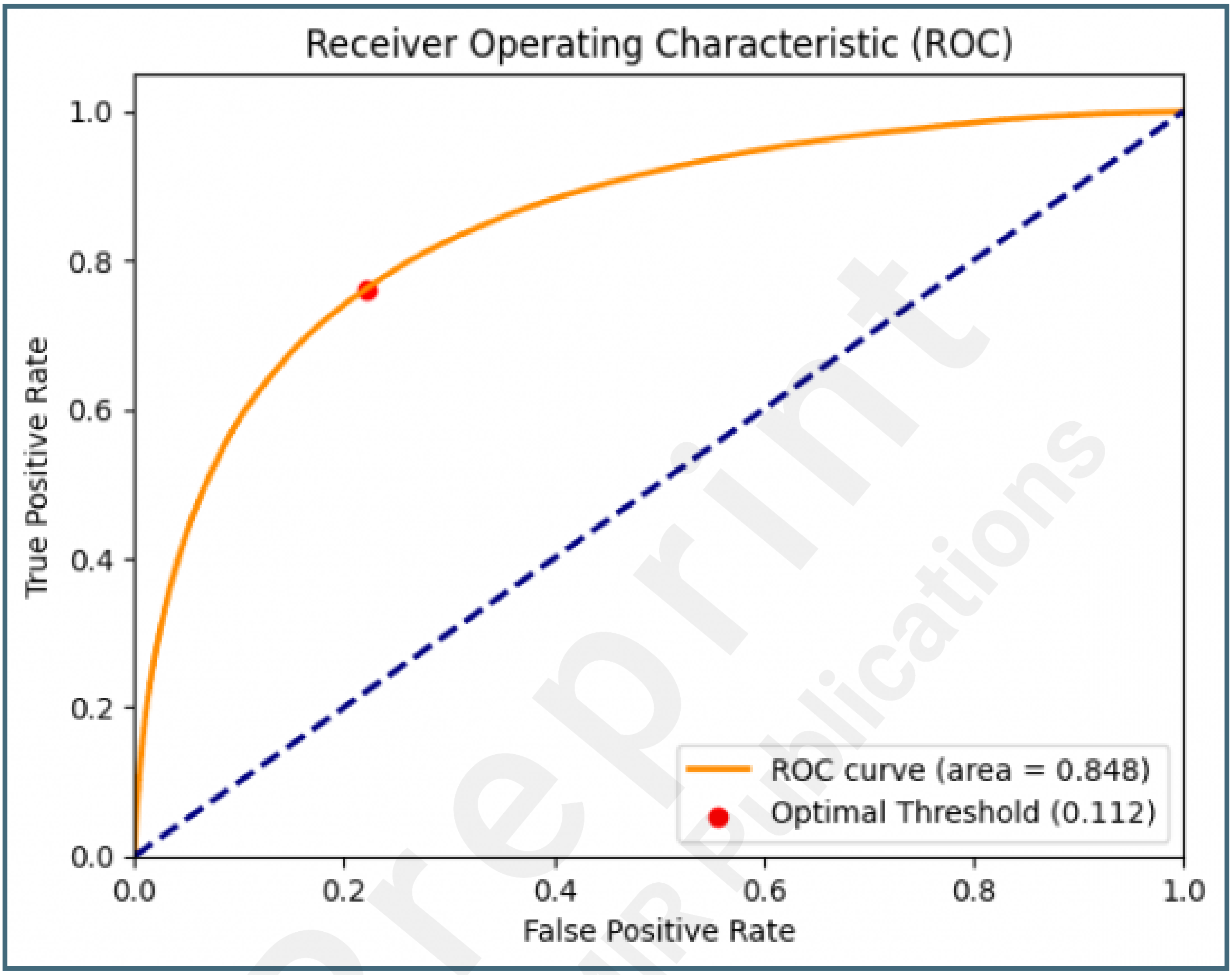
Heatmap of calculated relation score for 8,235,435 drug-side effect pairs.



ROC curves for BERT-based models.



ROC curve for word2vec model.



TOC/Feature image for homepages

In this study, we propose a novel approach for predicting unknown drug-side effect relationships using embedding vectors from language models pre-trained on biomedical corpora.