applied sciences

MDPI

*Article*

# Multi-Data Aspects of Protein Similarity with a Learning Technique to Identify Drug-Disease Associations

**Satanat Kitsiranuwat [1,2], Apichat Suratanee [3] and Kitiporn Plaimas [1,2,4,*]**

[1] Program in Bioinformatics and Computational Biology, Graduate School, Chulalongkorn University, Bangkok 10330, Thailand; 6187792520@student.chula.ac.th
[2] Advanced Virtual and Intelligent Computing (AVIC) Center, Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Bangkok 10330, Thailand
[3] Department of Mathematics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok 10800, Thailand; apichat.s@sci.kmutnb.ac.th
[4] Omics Sciences and Bioinformatics Center, Faculty of Science, Chulalongkorn University, Bangkok 10330, Thailand
* Correspondence: kitiporn.p@chula.ac.th; Tel.: +66-2-218-5220

**Abstract:** Drug repositioning has been proposed to develop drugs for diseases. However, the similarity in a single aspect may not be sufficient to reveal hidden information. Therefore, we established protein–protein similarity vectors (PPSVs) based on potential similarities in various types of biological information associated with proteins, including their network topology, proteomic data, functional analysis, and druggable property. Based on the proposed PPSVs, a separate drug–disease matrix was constructed for individual to prevent characteristics from being obscured between diseases. The classification technique was employed for prediction. The results showed that more than half of the tested disease models exhibited high performance, with overall F1 scores of more than 80%. Furthermore, comparing all diseases using traditional methods in one run, we obtained an (area under the curve) AUC of 98.9%. All candidate drugs were then tested in clinical trials ($p$-value $< 2.2 \times 10^{-16}$) and were known drugs based on their functions ($p$-value $< 0.05$). An analysis revealed that, in the functional aspect, the confidence value of an interaction in the protein–protein interaction network and the functional pathway score were the best descriptors for prediction. Based on the learning processes of PPSVs with an isolated disease, the classifier exhibited high performance in predicting and identifying new potential drugs for that disease.

**Keywords:** biological network; drug repositioning; drug repurposing; protein's interaction network; machine learning



**Citation:** Kitsiranuwat, S.; Suratanee, A.; Plaimas, K. Multi-Data Aspects of Protein Similarity with a Learning Technique to Identify Drug-Disease Associations. *Appl. Sci.* **2021**, *11*, 2914. https://doi.org/10.3390/app11072914

Academic Editor: Je-Keun Rhee

Received: 26 February 2021
Accepted: 22 March 2021
Published: 24 March 2021



## 1. Introduction

On average, it takes at least 10 years from the discovery and development of a candidate drug to commercialize it for the treatment of a disease. Drug discovery requires significant investment, and the probability of success remains low at less than 10% [1]. The process of drug discovery consists of (i) discovery and preclinical screening of compounds that affect a disease in the laboratory, (ii) safety review to confirm the safe usage, (iii) clinical research phase I, (iv) clinical research phase II, (v) clinical research phase III, in which the drug is tested on human subjects, (vi) Food and Drug Administration (FDA) review of the drug and subsequent approval, and (vii) FDA post-market safety monitoring of the drug [2]. Given this long process, alternative routes for the discovery of new drugs for a disease are required.

Drug repositioning, also known as drug repurposing, is a well-known strategy for finding new indications for an existing drug [2,3]. The process of drug repositioning consists of (i) compound identification to screen the candidate drug for use, (ii) compound acquisition to optimize candidate compounds, (iii) compound development to ensure the

safety and efficiency of the drug in preclinical research phase I or phase II by using existing data of the drug and (iv) FDA post-market safety monitoring [2]. Drug repositioning can be employed to reduce investment costs, resource use, and time [2,3]. For example, in the late 1980s, Pfizer discovered sildenafil, as a treatment for coronary artery disease, acting as a muscle relaxant to increase blood flow within the body. Later, it was discovered that the drug could be used to treat erectile dysfunction, and it is now sold under the brand name Viagra. Similarly, thalidomide, developed by the West German pharmaceutical company Grünenthal in the 1950s, was first used as a sedative. These days, it can also be used to treat several conditions such as the flu, nausea, and morning sickness in pregnant women [3,4].

To identify the new indications for an existing drug, a wide range of biological information needs to be known, particularly the interaction between proteins. In principle, a protein interacts with other proteins to regulate certain functions and activities within a biological system. By identifying relationships among proteins based on physical and functional interactions [5], a protein–protein interaction (PPI) network can be constructed. This information is a great useful resource for a further pipeline analysis in drug repositioning and can be used to infer drug–disease associations. Several computational approaches have been developed to predict drug–disease associations. Wu, Liu, and Yue proposed ensemble meta-paths and singular value decomposition (EMP-SVD) based on a heterogeneous network with three types of node: drugs, target proteins, and diseases [6]. The EMP-SVD requires a significant volume of interaction data for the drugs and target proteins, the target proteins and diseases, and the drugs and disease associations, including reliable negative, to construct suitable meta-paths. A reliable negative set contains information on drugs that cannot be used to treat a disease. However, a problem remains regarding how to obtain reliable negative sets. A concept that is widely used to support drug repositioning based on a similarity approach is guilt-by-association [7]. Several strategies have been developed based on this concept. For example, similar drugs may have common targets, side effects, or indications, and similar targets may have a common drug, and similar drugs may correspond to similar targets [7,8].

The most common and straightforward approach is to identify similarities between targets that share or correspond to the same drug [9–11]. One method that has been proposed for the wide-scale prediction of drug–disease indications is PREDICT [12], which is useful to observe similar drugs or drug-related proteins that are indicated for similar diseases. The PREDICT method combines various drug–drug and disease–disease similarities including chemical structure, predicted side effects, gene ontology, sequences, and phenotypes. The PREDICT [12] approach has been reported to obtain an area under the curve (AUC) for the receiver operating characteristic (ROC) of 0.92. In this strategy, the maximum score of the similarity scores was calculated for each drug–disease candidate pair.

Another approach is similarity-based large-margin learning of multiple sources (SLAMS), which uses multiple data sources for candidate drugs, including chemical structures, protein targets, and side-effect profiles, to identify novel drugs for the treatment of diseases [13]. SLAMS can be used to demonstrate various characteristics of drugs and protein targets that play an important role in drug repositioning models, with a reported AUC of 0.89. This compares favorably to PREDICT, which can achieve an AUC of 0.87 when using integrated multiple data sources, while the prediction scores of the SLAMS model were higher than the PREDICT model by 32% and 59% for precision and recall, respectively.

A similarity scheme that predicts approved and novel drug targets with new disease associations (SPANTD) has been used to investigate various interesting features, combining them in a scoring matrix. This scoring matrix includes protein similarities, common pathway sharing, binding sites, and disease similarities [14]. A genetic algorithm is employed on this scoring matrix to generate the prediction model that exhibits high performance (AUC = 0.97). The SPANTD model provided information on common pathways that are expressed as a connection between a drug and its targets. Therefore, the observation of the same action mechanisms can be used to suggest new drug targets for the repurposing of a drug.

Most computational methods for predicting the drug repositioning are based on guilt-by-association using observations of similarities between drug-related proteins or genes in order to indicate for similar diseases. In particular, the type of relationship among genes, i.e., whether they are positive (e.g., activation) or negative (e.g., inhibition), is important information that has been proposed for use in modeling drug repositioning. This information calculates the relationship between genes based on their gene expression profiles, particularly the relationship between target genes and disease genes for drug repositioning [15]. However, the gene expression data for the interaction between target genes and disease genes are not always available, and these missing data can hamper model construction for drug repositioning.

Lee and Yoon [16] integrated existing gene networks from several databases, including BioCarta [17], Reactome [18], the Pathway Interaction Database (PID) [19], and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway [20], to calculate the shortest path score between drug-related genes (or target genes) and disease genes for both positive or negative interaction types [16]. Because several target genes can be influenced by a drug, the interaction types between drugs and target genes were determined to combine the shortest path scores for target genes that were activated by a disease gene for each drug. If the interaction type was positive (e.g., an activator, stimulator, or inducer), the shortest path score was multiplied by +1; otherwise, it was multiplied by $-1$. Summing all of the shortest path scores expressed the value of a drug when associated with a disease gene. This technique suggested the assessment of model's performance in treating a disease with the drug.

The objective of our research was to develop a drug repurposing approach that could be used to predict novel drugs for the treatment of a target disease using multiple biological characteristics summarized in protein–protein similarity vectors (PPSVs), which identify similarities between drug-related proteins and disease-related proteins. The PPSVs were based on four types of biological information: the topological network, proteomic data, functional analysis, and druggable property. The hypothesis of this research was that if a target protein is similar to a disease-related protein, then the drug that relates to the target protein may interact with the disease-related protein. In other words, the similarities between target proteins and disease-related proteins in several biological characteristics can be used to indicate whether has the potential to treat a specific disease. Moreover, every disease affects the structure or function of specific characteristics, such as causes and symptoms; thus, designing a model specific to the disease may be a more effective approach. This research thus employed a classification technique based on a random forest algorithm to predict candidate drugs for specific diseases.

The rest of the paper is structured as follows. Section 2 describes the proposed method for predicting drugs used to treat individual diseases based on four biological characteristics included in the PPSVs. Section 3 summarizes the performance score of the model and the results from the validation of the candidate drug–disease pairs using information from clinical trials and in terms of functional similarity and compares its performance with other existing methods. Gold features, i.e., those that are most important for predicting whether a drug can be used to treat a specific disease, and novel drug–disease pairs are also discussed in this section. Section 4 describes the results and the limitations of this research. A conclusion is provided in Section 5.

## 2. Materials and Methods

In the present study, the workflow for the prediction of promising drugs for the treatment of a disease consisted of seven steps (Figure 1). First, three types of association including (1) the drug and target protein, (2) the disease and disease-related protein, and (3) the drug and disease, were collected from curated databases. Second, PPSVs were generated to determine similarities between target proteins and disease-related proteins. Third, a drug–disease matrix was constructed for each disease based on the PPSVs features. Fourth, a classification model was generated to predict candidate drugs for each disease.

Fifth, the candidate drugs were validated based on experimental knowledge from clinical trials and previous literature and an analysis of their functional similarities. Next, the performance of the proposed model was compared to existing methods. Finally, the gold features, i.e., the most important descriptors for the prediction, were determined to indicate the characteristics that are most relevant to the drug repositioning approach.

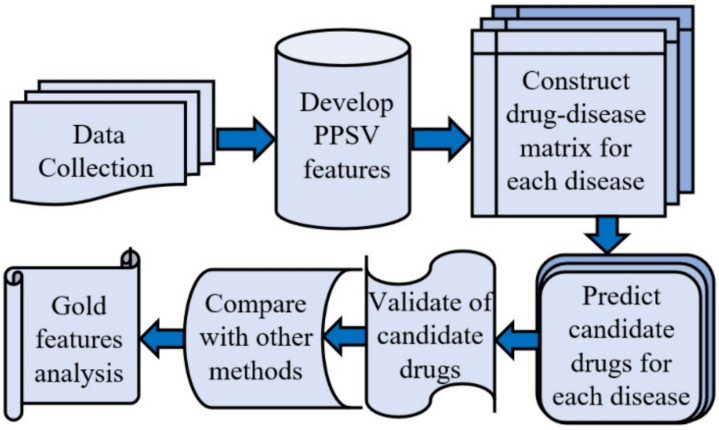


**Figure 1.** The workflow for identifying drugs for a disease.

### *2.1. Dataset*

The human protein–protein interaction (PPI) network was constructed using interaction information from the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) v.11database [21]. Only interactions with confidence scores higher than 0.8 were selected to construct the network. The associations between approved drugs and diseases were taken from the Comparative Toxicogenomics Database (CTD) [22] which collects the chemical and phenotypic characteristics of drugs. Only drug–disease pairs supported by therapeutic evidence were used for our analysis. The DrugBank database was used to map approved drugs, their target proteins, and their genes [23]. The group of genes or proteins associated with specific diseases were extracted from a database of gene-disease associations (DisGeNET) [24]. Overall, we obtained 14,264 known drug–disease associations involving 1317 approved drugs and 478 diseases and employed these associations as positive samples. The remaining 615,262 drug–disease associations from a combination of the approved drugs and diseases were classified as negative samples.

### *2.2. Protein–Protein Similarity Vectors (PPSVs)*

A PPSV is a feature vector of similarities between target proteins and disease-related proteins. The PPSVs were generated based on four types of biological information: the topological network, proteomic data (protein sequencing), functional analysis, and druggable property. An overview of a PPSV in four information types is summarized in Figure 2.

#### 2.2.1. Topological Network Information

We investigated the human PPI networks in two ways:

1. Neighboring similarity score: $Nei(P_j, P_k)$ The $Nei(P_j, P_k)$ score represents the cosine similarity between two proteins in terms of their common neighboring proteins or partners. This similarity score is calculated using the dot product of two vectors of proteins that are partners of the observed proteins. Then, it is divided by the magnitudes of each vector as shown in Equation (1):

$$Nei(P_j, P_k) = \frac{\vec{N(P_j)} \cdot \vec{N(P_k)}}{||\vec{N(P_j)}||\,||\vec{N(P_k)}||} = \frac{|N(P_j) \cap N(P_k)|}{\sqrt{|N(P_j)| \cdot |N(P_k)|}}, \quad (1)$$

where $P_j$ and $P_k$ are proteins $j$ and $k$, respectively. $N(P_j)$ and $N(P_k)$ are vectors for neighboring proteins of protein $j$ and $k$, respectively. For all $P_u$ in the network and $P_u \neq P_j$, $N(P_j) = 1$ if $P_u$ is a neighboring protein of $P_j$, otherwise, $N(P_j) = 0$. The $Nei(P_j, P_k)$ score ranges between 0 and 1. A score of 0 indicates that there are no common partners between two proteins while the score of 1 indicates that all proteins are partners of both proteins. A high neighboring similarity score indicates that the two proteins have high common neighbors in the human PPI network. This score is a good indication that these two proteins may cooperate within the same module to regulate the same functional task(s).

2. Closeness score: $Closer(P_j, P_k)$ $Closer(P_j, P_k)$ represents the closeness between two proteins based on the length of the shortest path in the PPI network. The score can be calculated as shown in Equation (2):

$$Closer(P_j, P_k) = \frac{1}{D(P_j, P_k)}, \quad (2)$$

where $D(P_j, P_k)$ is the length of the shortest path between proteins $j$ and $k$ in the human PPI network. This closeness score also ranges between 0 and 1. The score for a self-protein is assigned to 1 and the score for any two disjoint proteins is assigned to 0.

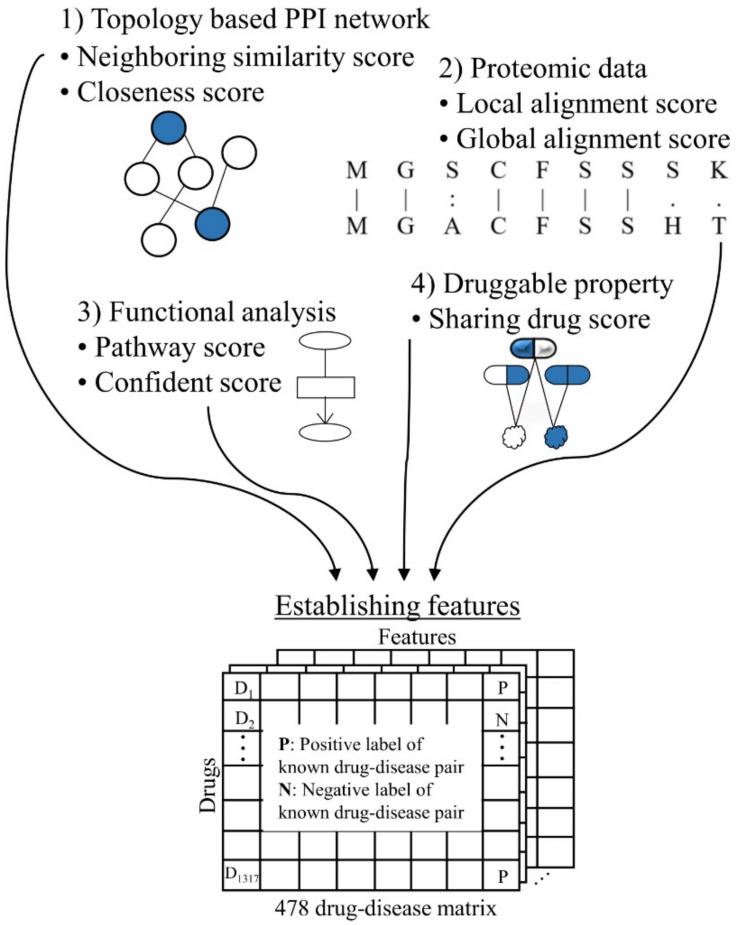


**Figure 2.** Overview of protein–protein similarity vectors (PPSVs) approach.

#### 2.2.2. Protein Sequencing

We conducted both local and global alignment to compare any two protein sequences.

1. Local alignment score, $Loc(P_j, P_k)$, indicates the most similarity alignment of regions within two protein sequences. The local alignment score represents the similarity in

the structure, function, and evolution of two protein sequences in some regions. The most similar regions of the sequences for proteins $j$ and $k$ are aligned. A drug that can bind to a certain region of a protein could also bind to the other protein if that protein has a similar region [25].

2. Global alignment score, $Glo(P_j, P_k)$, represents the score of the alignment of the entire protein sequences for proteins $j$ and $k$. The global alignment score may also reflect the similarities in the protein structure, function, or evolution of the two protein sequences. Both local and global alignment techniques were conducted using the Biostrings package in R language based on the BLOSUM62 substitution matrix with a gap opening of 10 and a gap extension of 0.5. These parameters are the same as the default values in the EMBOSS water tool option from the European Bioinformatics Institute.

#### 2.2.3. Pathway and Functional Analysis

We initially counted the number of common pathways for any two proteins as the pathway score $PW(P_j, P_k)$. We investigated any two proteins in the KEGG database [20,26,27]. The pathway score can be calculated using Equation (3):

$$PW(P_j, P_k) = \left|PW(P_j) \cap PW(P_k)\right|, \quad (3)$$

where $PW(P_j)$ and $PW(P_k)$ are the pathways found for proteins $j$ and $k$, respectively. A high pathway score indicates that two proteins operate in the same functional modules. In the same manner, a certain drug might disturb the pathway of proteins within the same functional modules during the treatment of a disease.

To determine the co-functions of any two proteins, we directly used the confidence score $Conf(P_j, P_k)$, which represents the approximate probability that there exists an interaction between two proteins if both proteins are in the same metabolic module within the KEGG database [26]. The $Conf(P_j, P_k)$ score between proteins $j$ and $k$ was retrieved from the STRING v.11 database [21]. High confidence scores represent a high possibility of the association between two proteins. Thus, a drug that can bind to one target protein may also disturb the functional modules of the other target protein related to other diseases. This provides evidence for the potential of repurposing a certain drug to treat other diseases. This confidence score is multiplied by 1000, giving a range of 0 to 1000. A self-protein is assigned a score of 1000, and any two disjoint proteins are assigned a score of 0.

#### 2.2.4. Druggable Property

We investigated the drugs associated with two target proteins by recording the number of common drugs that bind to both proteins. This score is defined as a shared drug score, $ShareDr(P_j, P_k)$ where $P_j$ and $P_k$ are any two proteins. A high score indicates that the two proteins share a high number of approved drugs that bind to them.

A PPSV feature between the proteins $j$ and $k$ is denoted as $PPSV(P_j, P_k)$. This feature is a vector containing the seven similarity scores: $Nei(P_j, P_k)$, $Closer(P_j, P_k)$, $Loc(P_j, P_k)$, $Glo(P_j, P_k)$, $PW(P_j, P_k)$, $Conf(P_j, P_k)$, and $ShareDr(P_j, P_k)$. Subsequently, these features were used to represent the similarities between a target protein and a disease-related protein. The R source code used to calculate these PPSV features is available at https://github.com/ksatanat/PPSV (accessed on 13 March 2021). All of these features were used to construct a drug–disease matrix in which a comparison of the PPSV values for all corresponding proteins was conducted, and the final values for each feature were then rescaled to a relative standard value between 0 and 1.

### 2.3. Constructing Drug–Disease Matrix

The three steps used to construct the drug–disease matrix in this study are summarized in Figure 3. First, PPSV table containing the PPSV features for a drug target protein and a disease-related protein was constructed. For a drug–disease pair, the features for all combinations between the drug target proteins and disease related proteins were calculated.

An example of a PPSV table for a drug is presented in Figure 3a. Second, the maximum PPSV score of all target proteins for a drug was identified (Figure 3b). The maximum PPSV score was selected as the similarity vector score for a drug and a disease-related protein. This vector for the drug and disease-related protein represents the similarities between all target proteins and a disease-related protein based on their topological PPI networks, proteomic data for sequence alignments, functional analysis, and druggable properties of proteins. Third, as shown in Figure 3c, the same process was employed for all combinations of drugs and disease-related proteins. Therefore, a drug–disease matrix for each single disease was constructed. All of the values in the column vectors of the matrix were rescaled to a range of 0 and 1. Eventually, we obtained a total of 478 matrices (for 478 diseases) with a size of 1315 drugs (representing the number of drugs) × the size of its PPSV, which depends on the number of disease proteins.

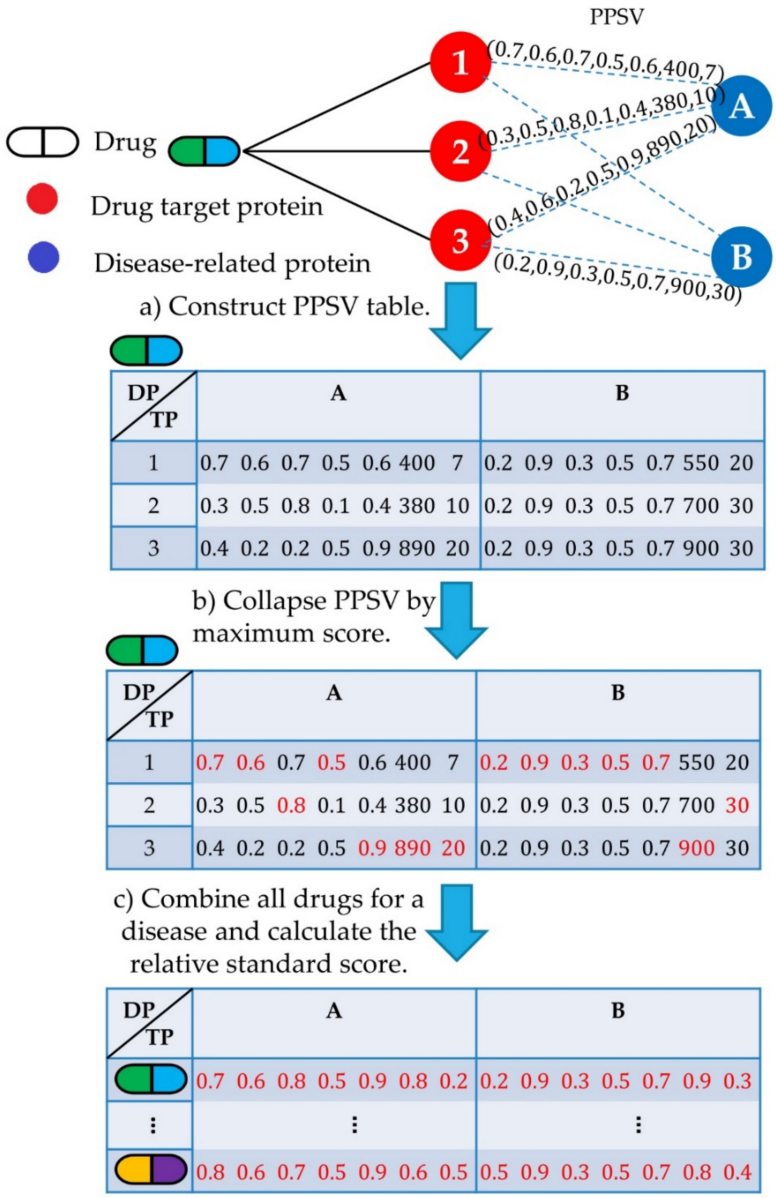


**Figure 3.** Construction of the drug–disease matrix for a disease. (**a**) Construct a PPSV table, (**b**) collapse PPSV using the maximum score, (**c**) combine all drugs for a disease and calculate the relative standard score.

### 2.4. Predicting Candidate Drugs for a Disease

For a drug–disease matrix, all of the values in each column vector were rescaled to a range of 0 and 1. Known drug–disease associations were given a positive label; otherwise, they were given a negative label. A random forest classifier was employed as the predictive model to identify candidate drugs for the treatment of a disease. In total, we obtained 478 models for all investigated diseases. The python source code for building the prediction model is provided at https://github.com/ksatanat/PPSV (accessed on 13 March 2021).

Each drug–disease matrix was split by randomly selecting 20% of all drug–disease pairs to be a test set, and the remaining 80% were assigned to be a training set. This division was conducted to ensure the same proportion of each label. To obtain a balanced data set for each machine, a bootstrap randomization technique was used to generate five sets of negative labels from training data into five machines. We performed a 5-fold cross validation to avoid overfitting of each machine. For each fold with the same number of trees, a grid search cross-validation technique [28] with 5-fold cross-validation was employed to identify the best hyperparameters in the forest. The parameters ranged from 50 to 300 in increments of 50. For each disease, a drug–disease pair from the test set was applied to the five machines to obtain five prediction scores. The average of the prediction scores for a pair was calculated. Subsequently, the area under the curve (AUC) of the receiver operating characteristic was employed to evaluate the performance of the disease model. Moreover, the F1 score and accuracy score (ACC) were also employed to assess the performance. We repeated these process five times to prevent any bias. The average of the performance scores from these five replications was calculated to represent the performance of the model.

Finally, the average prediction score was calculated based on the prediction scores from five machines and five experiments for a drug–disease matrix. If the average prediction score for a drug–disease pair was greater than or equal to 0.5, the pair was considered a potential candidate drug for treatment of the disease.

### 2.5. Evaluating Performance of Model

The performance metrics employed in this study were the AUC, F1, and ACC scores. The AUC value represents the area under the curve for the ROC which is the curve plotted between true positive rate ($TPR$) and false positive rate ($FPR$) in several thresholds. The $TPR$ and $FPR$ are calculated as follows:

$$TPR = \frac{TP}{TP + FN} \tag{4}$$

$$FPR = \frac{FP}{FP + TN} \tag{5}$$

where true positive ($TP$) outcomes correctly predict a positive label, true negative ($TN$) outcomes correctly predict a negative label, false negative ($FN$) outcomes incorrectly predict a negative label, and false positive ($FP$) outcomes incorrectly predict in positive label. The F1 score is the harmonic mean between precision (PRE) and recall (REC), described as follows:

$$PRE = \frac{TP}{TP + FP} \tag{6}$$

$$REC = \frac{TP}{TP + FN} \tag{7}$$

$$F1 = \frac{2 * PRE * REC}{(PRE + REC)} \tag{8}$$

The accuracy (ACC) is the overall percentage of correct predictions in both positive and negative labels. It can be calculated as follows:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

The F1 and ACC scores are determined based on the criteria of the maximum F1 score. All performance scores ranged between 0 and 1. Scores closer to 1, indicate a higher performance. The measurement metrics have been employed in previous studies and are described in more details which can be found in [29,30].

## 3. Results

### 3.1. Investigating Performance of Predictive Models

For each disease, our model generated using the random forest classifier was used to recognize known drugs and to predict possible new drugs from among existing drugs. The average prediction score was computed from the prediction score for each drug from five trained machines. Then, the metrics AUC, F1, and ACC were then calculated using the average prediction score over five iterations. The performance of the model for predicting candidate drugs was assessed based on the average AUC, F1, and ACC scores from five iterations. The numbers of diseases whose predictive models exhibited a higher performance in terms of their AUC, F1, and ACC scores than specific thresholds are listed in Table 1.

**Table 1.** Number of diseases with a model performance higher than various thresholds.

| Threshold Range | AUC | F1 | ACC |
|---|---|---|---|
| ≥0.9 | 50 | 63 | 62 |
| ≥0.8 | 171 | 280 | 212 |
| ≥0.7 | 331 | 466 | 395 |
| ≥0.6 | 420 | 478 | 462 |
| ≥0.5 | 458 | 478 | 478 |
| Total number of diseases | 478 | 478 | 478 |

Table 1 shows the number of diseases in our proposed model yielded the performance values above the certain thresholds. At a threshold of 0.6 for the AUC, F1, and ACC scores, 420, 478, and 462 (87.9%, 100%, and 96.7%) diseases, respectively, performed above the threshold. When the threshold for the AUC, F1, and ACC scores was 0.9, a total of 50, 63, and 62 diseases, respectively, surpassed this threshold. Overall, more than half of the tested diseases produced a high performance with more than 70% of the AUC, F1, and ACC scores.

### 3.2. Validation of Known Drugs from Clinical Trial Data and Past Literature

Drugs whose average prediction score were greater than or equal to 0.5 for a certain disease were identified as a candidate drug. To analyze the effectiveness of our predictive model, current experimental knowledge, such as known drugs from clinical trial data and those studied in previous literature, was used to verify the predictions.

#### 3.2.1. Validation Using Known Drugs from Clinical Trial Data

Using current experimental knowledge based on clinical trial data from the AACT database [31] in R package, we applied Fisher's exact test [32] with a confusion table of size $2 \times 2$ with the rows of candidate drug–disease pairs and non-candidate drug–disease pairs and columns of clinical trials and non-clinical trials. The confusion table of the proposed model is presented in Table 2. The null hypothesis was that there was no association between candidate drug–disease pairs and the likelihood of being observed in clinical trials. In another words, the alternative hypothesis was that the candidate drug–disease pairs were more likely to be observed in clinical trials than in non-clinical trials.

**Table 2.** Confusion table for validating the predictions with known drugs in clinical trial data.

| N = 629,526 Pairs | Found in the Database under Clinical Trial Data | Not Found in the Database |
|---|---|---|
| Candidate drug–disease pairs | 5900 | 14,975 |
| Non-candidate drug–disease pairs | 21,734 | 586,917 |

The confusion table in Table 2 shows that 20,875 of the predicted candidate drugs-disease pairs were found in the database for 5900 pairs. The proposed model predicted 608,651 non-candidate drug–disease pairs, none of which were found in the database of 586,917 drug–disease pairs. The results of Fisher's exact test produced a $p$-value less than $2.2e^{-16}$, indicating that the candidate drug–disease pairs were significantly more likely to be found in the AACT database.

#### 3.2.2. Validation Using Previous Literature on Candidate Drugs

The candidate drugs–disease pairs from the proposed model were also validated by counting candidate drug–disease pairs found in the PubMed database. We employed the "easyPubMed" package in R language to retrieve scientific publication records from the PubMed database [33]. In total, 19,030 of the 20,875 candidate drug–disease pairs (91.2%) predicted by our model were found in PubMed. We also searched the AACT database for the use of the remaining 1845 candidate drug–disease pairs in clinical trial data and found 15 in this database (Supplementary Table S2).

### *3.3. Verification of Functional Similarities*

The functional similarity of both candidate and non-candidate drug–disease pairs was investigated by comparing them with the functional modules of all known drugs for each disease. The steps for the verification of functional similarity are illustrated in Figure 4. First, enrichment analysis was conducted for the candidate drugs, non-candidate drugs, and known drugs based on functional modules or pathways for each disease using the KEGG database [26] (Figure 4a). Second, we compared the similarity of the functional modules for each candidate and non-candidate drug with that of a known drug for each disease (Figure 4b). The similarity score for the functional modules was computed using Jaccard similarity as follows:

$$Jaccard(M_d, M_{kd}) = \frac{|M_d \cap M_{kd}|}{|M_d \cup M_{kd}|}, \quad (10)$$

where $M_d$, $M_{kd}$ are the set of functional modules for a drug and a known drug, respectively.

We then computed the Jaccard similarity score for the functional modules of each candidate and non-candidate drug in comparison to all known drugs. The maximum Jaccard score was employed to represent the functional similarity score for each candidate and non-candidate drug in comparison with all known drugs for a disease (Figure 4c). Subsequently, the functional similarity scores for all candidate and non-candidate drugs were computed. Finally, we validated the functional similarity scores for the candidate and non-candidate drugs in comparison with known drugs using the Wilcoxon rank-sum test.

The null hypothesis was that there was no difference in the medians functional similarity scores between the candidate drugs and the non-candidate drugs when compared with the known drugs. In other words, the alternative hypothesis was that the median functional similarity scores for the candidate drugs was higher than that of the non-candidate drugs when compared with known drugs. This was applied to all diseases.

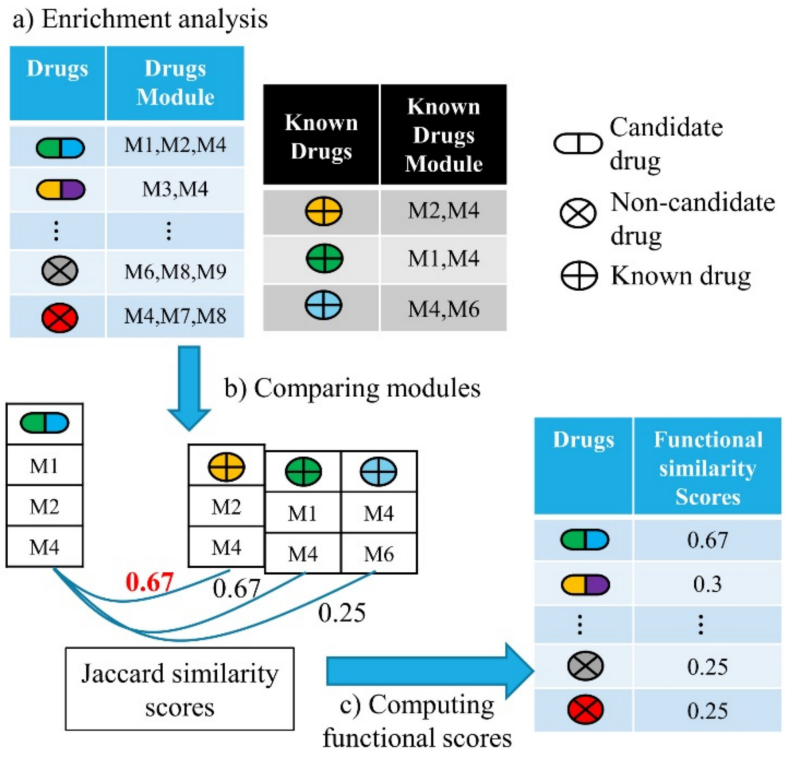


**Figure 4.** Verification of candidate drug–disease pairs in terms of their functional similarity using a Wilcoxon rank-sum test. (**a**) Enrichment analysis, (**b**) comparison of the functional modules with known drugs, and (**c**) functional score computations.

The results showed that there were 478 diseases with corresponding *p*-values that were lower than 0.05 in the Wilcoxon rank-sum test. Hence, the functional similarity scores of the candidate drugs were higher than the scores for non-candidate drugs when comparing them with known drugs for all 478 diseases. This indicates that the candidate drugs associated with each disease exhibited stronger functional similarity with known drugs than with non-candidate drugs.

### *3.4. Comparison of Our Method with Other Existing Methods*

We compared the performance of our proposed method with other existing methods. Lee and Yoon [16] described a method based on a directed gene network using the random forest classifier in 2018. They generated models corresponding to each disease with the weight of the out-degree in the gene directed network including the positive or negative associations between genes and between the drugs and their target genes. An assessment of Lee's method in predicting drug–disease associations revealed that the random forest classifier produced excellent prediction performance. To compare the performance of our proposed approach with Lee's method, the same set of drugs, diseases, and known drug–disease pairs were employed as inputs with the total number of diseases set at 460.

The percentage of diseases whose proposed method and Lee's method exhibited a higher performance in terms of their AUC, F1, and ACC scores than specific thresholds are listed in Table 3. For thresholds of 0.5 or more, our proposed model consistently produced acceptable AUC, F1, and ACC scores for 440 or more of the 460 diseases (95.9%), whereas Lee's method produced acceptable scores for fewer than 440 diseases. Additionally, the results show that higher AUC, F1, and ACC thresholds lead to a higher number of efficient models for the prediction of new drugs for a disease.

**Table 3.** Percentage of diseases meeting the indicated performance range for the proposed method and Lee's method.

| Threshold Range | AUC | | F1 | | ACC | |
|---|---|---|---|---|---|---|
| | Lee (2018) | Our Model | Lee (2018) | Our Model | Lee (2018) | Our Model |
| ≥0.9 | 0.9% | 10.2% | 0.2% | 12.6% | 0.4% | 12.4% |
| ≥0.8 | 9.6% | 35.9% | 2.6% | 59.1% | 2.0% | 44.8% |
| ≥0.7 | 34.6% | 67.4% | 11.1% | 97.6% | 15.2% | 75.0% |
| ≥0.6 | 72.4% | 85.7% | 38.3% | 100.0% | 57.8% | 95.0% |
| ≥0.5 | 94.8% | 95.9% | 54.1% | 100.0% | 94.3% | 100.0% |

Wu et al. [6] proposed a drug–disease associations model in 2019. They generated five meta-paths which are drug–disease matrices based on drug–disease, drug–protein, and disease–protein interaction data, that included reliable negative, i.e., a set of drugs that cannot be used to treat diseases. They applied the singular value decomposition (SVD) technique to extract the latent features for the drugs and diseases. Then, they combined these latent features to represent drug–disease pairs. Random forest classification was employed to generate five models from the five meta-paths. They subsequently applied the ensemble technique to combine all five random forest models. Wu's approach outputs all candidate drug–disease pairs but our method proposes candidate drugs for each disease. To compare the two approaches, we combined all candidate drugs for all diseases and then randomly selected the non-candidate drugs for each disease with the same number of those candidate drugs corresponding with disease. This produced a set of 1317 drugs, their targets, and 478 diseases for comparison.

Table 4 demonstrates that the performance values with 478 common diseases in our proposed model are higher than that of Wu's method. Thus, the results showed that our model has outperformed that other model. We also analyzed the performance of Wu's method, Lee's method, and our proposed method at the same time by combining candidate drugs for all of the diseases in Lee's method in order to evaluate the same set of drugs, diseases, and known drug–disease pairs.

**Table 4.** Performance of our model compared with Wu's method.

| Performance Values | Wu (2019) | Our Model |
|---|---|---|
| AUC | 0.913 | 0.988 |
| AUPR | 0.921 | 0.988 |
| PRE | 0.864 | 0.956 |
| REC | 0.82 | 0.939 |
| ACC | 0.834 | 0.947 |
| F1 | 0.841 | 0.948 |

Table 5 shows the performance of the three models for the 460 diseases and 1315 drugs common to all methods. Lee's method, which is based on a random forest classifier, and our method produced a higher performance, while the area under the precision-recall curve was greater for Wu's method than for Lee's method. The performance scores for our method were also higher than those of Lee's method. Overall, our method dominated the other models based on the common set of diseases and drugs. The ROC curve for the three approaches is presented in Figure 5.

**Table 5.** Performance of our model compared with Lee's and Wu's methods.

| Performance Values | Lee (2018) | Wu (2019) | Our Model |
|---|---|---|---|
| AUC | 0.932 | 0.909 | 0.989 |
| AUPR | 0.911 | 0.919 | 0.988 |
| PRE | 0.91 | 0.844 | 0.956 |
| REC | 0.866 | 0.826 | 0.940 |
| ACC | 0.885 | 0.830 | 0.947 |
| F1 | 0.888 | 0.835 | 0.948 |

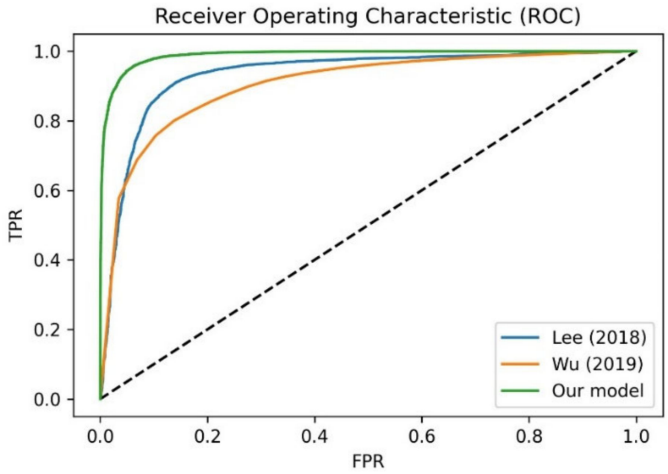


**Figure 5.** Receiver operating characteristic (ROC) of our model compared with Lee's and Wu's methods.

### *3.5. Gold Features Analysis*

This research proposed a method to investigate the relationship between target proteins and disease-related proteins based on several biological information, such as topological network, proteomic data, functional analysis, and druggable property, using PPSVs. The PPSVs consisted of seven features: the neighboring similarity score, the closeness score, the local alignment score, the global alignment score, the pathway score, the confident score, and the shared drug score. We investigated the gold features of our approach, which represent the best descriptors for the prediction results, using Pearson correlation coefficients (PCC). The PCC is used to measure the linear correlation between two variables. If a PPC is close to 1 or $-1$, the two variables have a highly positive or negative linear correlation, respectively. We calculated the PCC score for the seven features of each disease-related protein and the labels of drug–disease pairs that were added to the drug–disease matrix for each disease. The absolute PCC scores were then ranked in descending order for each disease to produce a highly linear correlation of the seven features for each disease-related protein. We then repeated this process for all diseases. The number of features that satisfied the top two of the absolute score in PCC for all 478 diseases is presented in Figure 6.

Figure 6 shows that the confidence score appeared in the highest and second highest PCC for 122 and 117 diseases, respectively, while the pathway score was found among the top two PPC for 113 and 85 diseases, respectively. The top two features for both the highest and second-highest PCC were the confidence and pathway scores, which were related to functional similarity in the PPSVs, thus these were identified as the gold features. The third feature most commonly appearing among the highest and second-highest PPC was the neighboring similarity score in the PPSVs. Therefore, two proteins that have a highly similar neighborhood might cooperate in the same module to regulate the same functional task. Figure 6 also reveals that the number of local alignment features ranked among the highest and second-highest PCC was higher than that of global alignment features. Hence, the local alignment score was more important for the model predictions than the global alignment score.

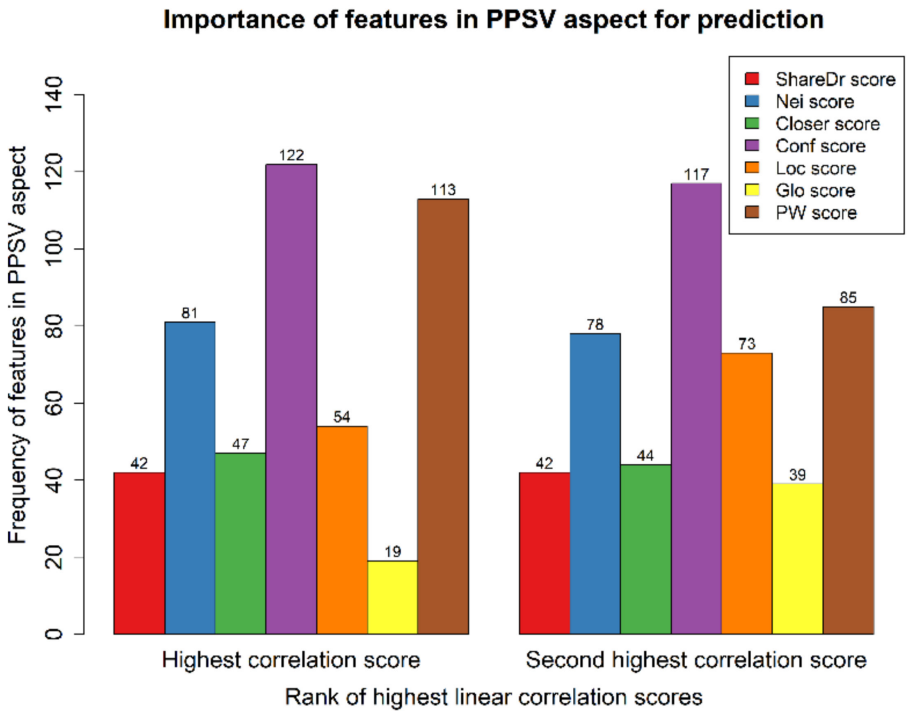


**Figure 6.** Frequency of features in PPSVs satisfied the highest and second-highest linear correlations for all 478 diseases.

Interestingly, we ran our model using only the important features identified above, i.e., the confidence, pathway, neighboring similarity, and local alignment scores, which lead to better performance for a higher performance threshold (Table 6). In particular, using the important features to generate the model, the number of diseases that satisfied the performance threshold of $\geq$0.90 increased. However, the number of diseases decreased for the performance scores of AUC and F1 more than 70% (the threshold $\geq$0.70). Based on this, it can be concluded that these important features play a crucial biological role in identifying drug–disease associations, but some diseases require other biological information in the PPSV to be predicted. Thus, the final model was trained on all features for all diseases.

**Table 6.** Percentage of diseases meeting the indicated performance range for model using PPSVs and important features.

| Threshold Range | AUC | | F1 | | ACC | |
|---|---|---|---|---|---|---|
| | Model Using PPSVs | Model Using Important Features | Model Using PPSVs | Model Using Important Features | Model Using PPSVs | Model Using Important Features |
| $\geq$0.9 | 10.2% | 23.8% | 12.6 | 22.2% | 0.4% | 22.2% |
| $\geq$0.8 | 35.9% | 39.1% | 59.1 | 55.0% | 2.0% | 45.8% |
| $\geq$0.7 | 67.4% | 58.6% | 97.6 | 84.3% | 15.2% | 67.4% |
| $\geq$0.6 | 85.7% | 77.2% | 100.0 | 100.0% | 57.8% | 83.3% |
| $\geq$0.5 | 95.9% | 88.7% | 100.0 | 100.0% | 94.3% | 100.0% |

### *3.6. Investigation of Novel Drug–Disease Pairs*

In this section, we reveal the novel of candidate drug–disease pairs based on false positives of our method with high performance of more than 90% of AUC score. The 50 diseases with AUCs of more than 90% are shown in Table 1. The candidate drugs corresponding to these 50 diseases were ranked in descending order based on the functional similarity score. The functional similarity score is the maximum of Jaccard similarity score for the functions of a drug and all known drugs (see Section 3.3). We investigate these novel candidate drug–disease pairs that were incorrect positive predictions in more detail in this section.

The top 20 candidate drug–disease pairs were identified by ranking the functional similarity score of the candidate drugs and known drugs in descending order, as shown in Table 7. It was found that the average functional similarity score for the top 20 candidate drug–disease pairs was about 0.98. In other words, 20 candidate drugs for diseases with false positive labels were 98% similar to their known drugs. However, correct positive prediction of candidate drug–disease pairs (i.e., true positives) were 100% similar to their known drugs. This indicates that these 20 novel candidate drugs, with their high functional similarity score of 98% could be used to treat the corresponding disease.

**Table 7.** The top 20 candidate drug–disease pairs.

| Chemical Name | Drug (Drug Bank IDs) | Disease Name | Disease (MESH IDs) | Functional Similarity Score |
|---|---|---|---|---|
| Anileridine | DB00913 | Nausea | D009325 | 1.00 |
| Antipyrine | DB01435 | Osteoarthritis | D010003 | 1.00 |
| Bromocriptine | DB01200 | Depressive Disorder, Major | D003865 | 1.00 |
| Levallorphan | DB00504 | Nausea | D009325 | 1.00 |
| Oxprenolol | DB01580 | Tachycardia, Supraventricular | D013617 | 1.00 |
| Paliperidone Palmitate | DB01267 | Depressive Disorder, Major | D003865 | 1.00 |
| Pergolide | DB01186 | Bipolar Disorder | D001714 | 1.00 |
| Salicylsalicylic Acid | DB01399 | Osteoarthritis | D010003 | 1.00 |
| Testosterone | DB00624 | Testicular Diseases | D013733 | 1.00 |
| Tocopherols | DB11251 | Testicular Diseases | D013733 | 1.00 |
| Triazolam | DB00897 | Depressive Disorder, Major | D003865 | 1.00 |
| Tropicamide | DB00809 | Basal Ganglia Diseases | D001480 | 1.00 |
| Zinc Chloride | DB14533 | Depressive Disorder, Major | D003865 | 1.00 |
| Ergotamine | DB00696 | Basal Ganglia Diseases | D001480 | 0.96 |
| Ergotamine | DB00696 | Delirium | D003693 | 0.96 |
| Ergotamine | DB00696 | Schizophrenia, Paranoid | D012563 | 0.96 |
| Loratadine | DB00455 | Panic Disorder | D016584 | 0.95 |
| Zopiclone | DB01198 | Epilepsy, Temporal Lobe | D004833 | 0.95 |
| Lornoxicam | DB06725 | Osteoarthritis | D010003 | 0.94 |
| Isradipine | DB00270 | Bipolar Disorder | D001714 | 0.94 |

Novel of candidate drug–disease associations are useful for proposing alternative indications of existing drugs, assessing side effects, and determining potential drug resistance. Pergolide (DrugBank ID: DB01186) is dopamine receptor agonist used for the treatment of Parkinson disease [34]. In our investigation, we found that pergolide has a high functional similarity score with the known drug zuclopenthixol (DrugBank ID: DB01624), which is used to treat bipolar disorder, as shown in the first highlight. Some studies have investigated the possibility of the pergolide being indicated for the treatment of bipolar disease (MESH ID: D001714). Bouckoms and Mangini explored pergolide as a supplement to antidepressant therapy with tricyclic antidepressants and monoamine oxidase inhibitors. Pergolide successfully adjusted the mood, interest, and energy of 11 out to 20 bipolar patients within a week [35]. However, pergolide has been withdrawn from several markets, the US and Canadian markets, because it was found to increase the risk of cardiac valvulopathy [36].

Further, one of the top 20 drug–disease associations identified in this study suggested that isradipine (DrugBank ID: DB00270) is also related to bipolar disease (MESH ID:

D001714), as shown in the second highlight. Ostacher et al. investigated the potential use of isradipine in the treatment of bipolar depression [37]. Clinical trial information of isradipine in the treatment of bipolar depression can be found at www.clinicaltrials.gov (NCT01784666). Therefore, our identified candidate drug–disease associations might be useful in further pharmaceutical analysis of drug repositioning. The list of all predicted drug–disease associations identified using our method is presented in Supplementary Table S1.

## 4. Discussion

Our method presented the topological network, proteomic data, functional analysis, and druggable property as important information for target proteins and disease-related proteins for use in PPSVs. The human PPI network was reconstructed using interactions with a high confidence score of more than 80%. This network provides high confidence for the relationship between proteins in terms of the neighboring similarity score and the closeness score in the PPSVs. Selecting maximum scores from the PPSVs to construct a drug–disease matrix allowed the similarity in various characteristics between proteins to be identified. Thus, a high score for PPSV features represented high confidence in the similarity between all of the target proteins and a disease protein.

A limitation of this research is that the investigated disease ideally should have had 10 or more approved drugs. We had a restriction in obtaining enough data of drug–disease pairs for fitting the model in 5-fold cross validation to predict candidate drugs for each disease. Our method is significantly different from Wu's and Lee's methods. Wu's method required many known drug–disease pairs to generate suitable meta-paths for the prediction of novel drug–disease associations from among all diseases. Lee's method is based on a directed gene network and positive or negative types of gene interactions to predict drug–disease pairs for each disease. However, our technique is based on various biological information within a PPSV to predict candidate drugs for each disease.

A comparison of the performance of our proposed method, Wu's, and Lee's methods demonstrated that both Lee's and our proposed methods produced a strong performance because the two models employed a similar classifier and a similar approach to generate a model for each disease when predicting candidate drugs. However, further analysis indicated that our method outperformed both Wu's and Lee's methods. The reason for this may be that our developed PPSVs contains various sources of biological information, including the topological network, proteomic data, functional analysis, and druggable property. Note that, the side effects were not considered by the methodology. The side effects of existing approved drugs have been thoroughly examined during clinical trials. It is supposed that there are not many opportunities to use this method. The best descriptors for drug–disease associations in our study were neighboring score, confident score, local alignment score, and pathway score. They exhibited good performance as indicators in the search of existing drugs for use with other diseases.

Our gold features analysis revealed that the best descriptors for predicting drug–disease associations were the confidence score and pathway score, which were related to functional similarity in the PPSVs. The results also indicated that the local alignment score was more important than the global alignment score for prediction. Hence, our finding suggests that the interaction among proteins in the PPI network is crucial for assessing whether a drug can bind to an alternative protein based on functional interactions.

## 5. Conclusions

This research proposed the use of multiple biological types of information, including proteomic data, functional analysis, and druggable property, for target proteins and disease-related proteins to produce PPSVs for an investigation into the similarity between proteins in the human PPI network. The proposed method predicted approved and novel drug–disease associations for individual diseases based on random forest classification. Further, the experimental knowledge from clinical trial data and the PubMed database were used to

verify the predicted candidate drugs. It was found that the predicted candidate drugs were significantly more functionally similar to known drugs. Our proposed model was also found to outperform other existing techniques, with the confidence and pathway scores proving to the best descriptors for prediction process in terms of functional similarity in PPSVs. The novel candidate drug–disease pairs identified in this study can be investigated further using pharmaceutical analysis in the laboratory.

**Supplementary Materials:** The following are available online at https://www.mdpi.com/2076-3417/11/7/2914/s1, Table S1: List of all predicted drug–disease associations. Table S2: Candidate drug–disease pairs that found in clinical trials even if not found in PubMed database.

**Author Contributions:** Conceptualization, S.K., A.S. and K.P.; methodology, S.K., A.S. and K.P.; funding acquisition, A.S.; formal analysis, S.K.; validation, S.K., A.S. and K.P.; writing—original draft preparation, S.K.; writing—review and editing, S.K., A.S. and K.P.; supervision K.P. and A.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by King Mongkut's University of Technology North Bangkok. Contract no. KMUTNB-63-KNOW-027.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data sharing is not applicable to this article.

**Acknowledgments:** We acknowledge National e-Science Infrastructure Consortium (http://www.e-science.in.th) (accessed on 11/05/2020) for providing computing resources that have contributed to the research results reported within this paper.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Roses, A.D. Pharmacogenetics in drug discovery and development: A translational perspective. *Nat. Rev. Drug Discov.* **2008**, *7*, 807–817. [CrossRef] [PubMed]
2. Ashburn, T.T.; Thor, K.B. Drug repositioning: Identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* **2004**, *3*, 673–683. [CrossRef]
3. Sleigh, S.H.; Barton, C.L. Repurposing Strategies for Therapeutics. *Pharm. Med.* **2010**, *24*, 151–159. [CrossRef]
4. Shim, J.S.; Liu, J.O. Recent Advances in Drug Repositioning for the Discovery of New Anticancer Drugs. *Int. J. Biol. Sci.* **2014**, *10*, 654–663. [CrossRef]
5. Safari-Alighiarloo, N.; Taghizadeh, M.; Rezaei-Tavirani, M.; Goliaei, B.; Peyvandi, A.A. Protein-protein interaction networks (PPI) and complex diseases. *Gastroenterol. Hepatol. Bed Bench* **2014**, *7*, 17–31.
6. Wu, G.; Liu, J.; Yue, X. Prediction of drug-disease associations based on ensemble meta paths and singular value decomposition. *BMC Bioinform.* **2019**, *20*, 134. [CrossRef]
7. Hodos, R.A.; Kidd, B.A. Computational Approaches to Drug Repurposing and Pharmacology. *HHS Public Access* **2016**, *8*, 1–46.
8. Hodos, R.A.; Kidd, B.A.; Shameer, K.; Readhead, B.P.; Dudley, J.T. In silico methods for drug repurposing and pharmacology. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **2016**, *8*, 186–210. [CrossRef]
9. Cheng, F.; Liu, C.; Jiang, J.; Lu, W.; Li, W.; Liu, G.; Zhou, W.; Huang, J.; Tang, Y. Prediction of Drug-Target Interactions and Drug Repositioning via Network-Based Inference. *PLoS Comput. Biol.* **2012**, *8*, e1002503. [CrossRef] [PubMed]
10. Bleakley, K.; Yamanishi, Y. Supervised prediction of drug-target interactions using bipartite local models. *Bioinformatics* **2009**, *25*, 2397–2403. [CrossRef]
11. Ding, H.; Mamitsuka, H.; Zhu, S. Similarity-basedmachine learning methods for predicting drug-target interactions: A brief review. *Brief. Bioinform.* **2013**, *15*, 734–747. [CrossRef]
12. Gottlieb, A.; Stein, G.Y.; Ruppin, E.; Sharan, R. Predict: A method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.* **2011**, *7*, 496. [CrossRef] [PubMed]
13. Blockeel, H.; Kersting, K.; Nijssen, S.; Zelezny, F. Machine Learning and Knowledge Discovery in Databases. *arXiv* **2012**, arXiv:1207.6324.
14. Khalid, Z.; Sezerman, O.U. Computational drug repurposing to predict approved and novel drug-disease associations. *J. Mol. Graph. Model.* **2018**, *85*, 91–96. [CrossRef]
15. Yu, H.; Choo, S.; Park, J.; Jung, J.; Kang, Y.; Lee, D. Prediction of drugs having opposite effects on disease genes in a directed network. *BMC Syst. Biol.* **2016**, *10*, S2. [CrossRef]
16. Lee, T.; Yoon, Y. Drug repositioning using drug-disease vectors based on an integrated network. *BMC Bioinform.* **2018**, *19*, 1–12. [CrossRef]

17. Nishimura, D. BioCarta. *Biotech Softw. Internet Rep.* **2001**, *2*, 117–120. [CrossRef]
18. Jassal, B.; Matthews, L.; Viteri, G.; Gong, C.; Lorente, P.; Fabregat, A.; Sidiropoulos, K.; Cook, J.; Gillespie, M.; Haw, R.; et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **2020**, *48*, D498–D503. [CrossRef] [PubMed]
19. Schaefer, C.; Anthony, K.; Krupa, S.; Buchoff, J.; Day, M.; Hannay, T.; Buetow, K. PID: The pathway interaction database. *Nucleic Acids Res.* **2009**, *37*, 674–679. [CrossRef] [PubMed]
20. Kanehisa, M.; Sato, Y.; Furumichi, M.; Morishima, K.; Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res.* **2019**, *47*, D590–D595. [CrossRef] [PubMed]
21. Szklarczyk, D.; Gable, A.L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N.T.; Morris, J.H.; Bork, P.; et al. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **2019**, *47*, D607–D613. [CrossRef]
22. Davis, A.P.; Grondin, C.J.; Johnson, R.J.; Sciaky, D.; McMorran, R.; Wiegers, J.; Wiegers, T.C.; Mattingly, C.J. The Comparative Toxicogenomics Database: Update 2019. *Nucleic Acids Res.* **2019**, *47*, D948–D954. [CrossRef] [PubMed]
23. Wishart, D.S.; Feunang, Y.D.; Guo, A.C.; Lo, E.J.; Marcu, A.; Grant, J.R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074–D1082. [CrossRef] [PubMed]
24. Pinero, J.; Ramirez-Anguita, J.M.; Sauch-Pitarch, J.; Ronzano, F.; Centeno, E.; Sanz, F.; Furlong, L.I. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res.* **2020**, *48*, 845–855.
25. Haupt, V.J.; Daminelli, S.; Schroeder, M. Drug Promiscuity in PDB: Protein Binding Site Similarity Is Key. *PLoS ONE* **2013**, *8*, e5894. [CrossRef]
26. Yi, Y.; Fang, Y.; Wu, K.; Liu, Y.; Zhang, W. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **2020**, *19*, 3316–3332.
27. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* **2019**, *28*, 1947–1951. [CrossRef]
28. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python Fabian. *J. Mach. Learn. Res.* **2011**, *11*, 2825–2830.
29. Le, N.Q.K.; Do, D.T.; Chiu, F.Y.; Yapp, E.K.Y.; Yeh, H.Y.; Chen, C.Y. XGBoost improves classification of MGMT promoter methylation status in IDH1 wildtype glioblastoma. *J. Pers. Med.* **2020**, *10*, 128. [CrossRef]
30. Le, N.Q.K.; Do, D.T.; Hung, T.N.K.; Lam, L.H.T.; Huynh, T.-T.; Nguyen, N.T.K. A Computational Framework Based on Ensemble Deep Neural Networks for Essential Genes Identification. *Int. J. Mol. Sci.* **2020**, *21*, 9070. [CrossRef] [PubMed]
31. Department of Health and Human Services; National Institutes of Health (NIH). NIH Policy on the Dissemination of NIH-Funded Clinical Trial Information. US Government Publishing Office (GPO): Washington, DC, USA, 2016; Volume 81, p. 26. Available online: https://federalregister.gov/d/2016-22379 (accessed on 9 February 2020).
32. Sprent, P. Fisher Exact Test. In *International Encyclopedia of Statistical Science*; Metzler, J.B., Ed.; Springer: Berlin/Heidelberg, Germany, 2011; pp. 524–525.
33. Canese, K.; Weis, S. PubMed: The Bibliographic Database. The NCBI Handbook. 2013. Available online: https://pubmed.ncbi.nlm.nih.gov/ (accessed on 9 February 2020).
34. Francis Lam, Y.W. Clinical pharmacology of dopamine agonists. *Pharmacotherapy* **2000**, *20*, 17–25. [CrossRef] [PubMed]
35. Bouckoms, A.; Mangini, L. Pergolide: An antidepressant adjuvant for mood disorders? *Psychopharmacol. Bull.* **1993**, *29*, 207–211. [PubMed]
36. National Center for Biotechnology Information. "PubChem Compound Summary for CID 47811, Pergolide" PubChem. Available online: https://pubchem.ncbi.nlm.nih.gov/compound/Pergolide (accessed on 23 December 2020).
37. Ostacher, M.J.; Iosifescu, D.V.; Hay, A.; Blumenthal, S.R.; Sklar, P.; Perlis, R.H. Pilot investigation of isradipine in the treatment of bipolar depression motivated by genome-wide association. *Bipolar Disord.* **2014**, *16*, 199–203. [CrossRef] [PubMed]