molecules

MDPI

*Article*

# The Discovery of New Drug-Target Interactions for Breast Cancer Treatment

**Jiali Song [1,†], Zhenyi Xu [2,†], Lei Cao [2], Meng Wang [2], Yan Hou [1,*] and Kang Li [2,*]**

[1] Department of Biostatistics, School of Public Health, Peking University, Beijing 100191, China; songjl@bjmu.edu.cn
[2] Department of Epidemiology and Biostatistics, School of Public Health, Harbin Medical University, Harbin 150086, China; xuzy@hrbmu.edu.cn (Z.X.); caolei@hrbmu.edu.cn (L.C.); wangm@hrbmu.edu.cn (M.W.)
* Correspondence: houyan@bjmu.edu.cn (Y.H.); likang@ems.hrbmu.edu.cn (K.L.); Tel.: +86-010-8280-5952 (Y.H.); +86-0451-8750-2679 (K.L.)
† These authors contributed equally to this work.

**Abstract:** Drug–target interaction (DTIs) prediction plays a vital role in probing new targets for breast cancer research. Considering the multifaceted challenges associated with experimental methods identifying DTIs, the in silico prediction of such interactions merits exploration. In this study, we develop a feature-based method to infer unknown DTIs, called PsePDC-DTIs, which fuses information regarding protein sequences extracted by pseudo-position specific scoring matrix (PsePSSM), detrended cross-correlation analysis coefficient (DCCA coefficient), and an FP2 format molecular fingerprint descriptor of drug compounds. In addition, the synthetic minority oversampling technique (SMOTE) is employed for dealing with the imbalanced data after Lasso dimensionality reduction. Then, the processed feature vectors are put into a random forest classifier to perform DTIs predictions on four gold standard datasets, including nuclear receptors (NR), G-protein-coupled receptors (GPCR), ion channels (IC), and enzymes (E). Furthermore, we explore new targets for breast cancer treatment using its risk genes identified from large-scale genome-wide genetic studies using PsePDC-DTIs. Through five-fold cross-validation, the average values of accuracy in NR, GPCR, IC, and E datasets are 95.28%, 96.19%, 96.74%, and 98.22%, respectively. The PsePDC-DTIs model provides us with 10 potential DTIs for breast cancer treatment, among which erlotinib (DB00530) and FGFR2 (hsa2263), caffeine (DB00201) and KCNN4 (hsa3783), as well as afatinib (DB08916) and FGFR2 (hsa2263) are found with direct or inferred evidence. The PsePDC-DTIs model has achieved good prediction results, establishing the validity and superiority of the proposed method.

**Keywords:** DTIs prediction; breast cancer; drug repurposing; machine learning; PsePSSM; DCCA coefficient



**Citation:** Song, J.; Xu, Z.; Cao, L.; Wang, M.; Hou, Y.; Li, K. The Discovery of New Drug-Target Interactions for Breast Cancer Treatment. *Molecules* **2021**, *26*, 7474. https://doi.org/10.3390/molecules26247474

Academic Editors: Aurora Costales and Fernando Cortés-Guzmán

Received: 22 November 2021
Accepted: 7 December 2021
Published: 10 December 2021



## 1. Introduction

Breast cancer is the most common gynecological malignant tumor in the world [1], with incidence rates that outdistance other cancers in both transitioned and transitioning countries [2]. It is reported that the global incidence of breast cancer has increased at a rate of 0.5% annually [3]. Actually, hereditary and genetic factors can account for 5% to 10% of breast cancer cases [2]. So far, approximately 100 breast cancer risk loci have been identified in a genome-wide association study (GWAS) [4]. However, only a few of targets are specifically for the development of new drugs for breast cancer. For example, in the ChEMBL dataset, there are 13 targets corresponding to 348 compounds, among which 209 compounds' max phase is phase 4 in terms of breast cancer. Therefore, with the purpose of exploring new targets for drugs of breast cancer treatment, predicting new drug–target interactions (DTIs) is a good solution. The cost and time factors associated with the development of new drugs on a commercial scale [5–7] warrant the need for examining

the already approved drugs. So, we explore new DTIs via drugs approved by FDA and the breast cancer risk genes identified from large-scale genome-wide genetic studies [8–11].

Wet lab experiments can infer the DTIs by using various techniques of classical and reverse pharmacology [12], while experimental methods identifying DTIs are expensive, time-consuming, and challenging. Therefore, for complementing experimental results, the in silico prediction of interactions between drugs and their targets is desirable [13]. The computational (in silico) methods for predicting drug–target interactions can be broadly classified into three categories: ligand-based approaches, docking-based approaches, and chemogenomic approaches [13].

For the first category, the main idea of ligand-based approaches is that similar molecules usually bind to similar protein targets and show similar properties [14,15]. However, the disadvantages of these approaches are that disregarding the information on the protein domain limits novel interactions to the link between known ligands and protein families, and hence, insufficient known ligands of target proteins may lead to poor performance [16]. As to the second category, docking approaches are powerful molecular modeling methods, which apply molecular dynamics using the 3D structures of the proteins as well as drugs to predict DTIs [17,18]. However, they cannot be applied in some cases where the 3D structures of proteins are not known. Most of the membrane proteins, for instance, have no three-dimensional structures in the freely protein databases. The third category, i.e., chemogenomic approaches, integrates the chemical information of the compounds and genomic information of proteins into a unified framework to predict DTIs. The preponderance of chemogenomic approaches is due to the fact that they overcome the disadvantages of ligand-based and docking-based approaches that have been discussed previously [19]. One of the chemogenomic methods categories, i.e., feature based methods, represents the drug-target pair with a vector of descriptors that may be produced by combining the properties of drug and targets, and can be put into various machine learning models to predict novel interactions [17].

In this study, we develop a feature-based method to infer unknown DTIs, called PsePDC-DTIs. The process of this method is described as follows. First, fusing protein features are generated by the pseudo-position specific scoring matrix (PsePSSM) algorithm, detrended cross-correlation analysis coefficient (DCCA coefficient), and FP2 fingerprint features of drug molecules under four benchmark datasets. Secondly, the least absolute shrinkage and selection operator (Lasso) method is used to reduce the dimension and noise information in the original high-dimensional space. Thirdly, the synthetic minority oversampling technique (SMOTE) is employed with Lasso feature-selected data for dealing with a high degree of imbalance in the samples used in this study. Finally, an ensemble classifier, random forest, is adopted to perform DTI predictions on four gold standard datasets, including nuclear receptors, G-protein-coupled receptors, ion channels, and enzymes. In the experiment, we make predictions concerning the gold standard DTI datasets by 5-fold cross-validation. Moreover, we can predict new DTIs for breast cancer with its risk genes by using PsePDC-DTIs.

## 2. Results

### *2.1. Performance Evaluation*

In this study, the five-fold cross-validation approach is used to evaluate the performance of the prediction model. For each data set, all the DTIs are randomly divided into five parts of roughly equal size. Each part is taken in turn as the test set, while the remaining four parts serve as the training set to establish a prediction model.

The following parameters, Accuracy (ACC), Specificity (SP), and Sensitivity (SE), F score are calculated to assess the performance of the prediction model proposed in the experiment. The definition is as follows:

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

$$SE = \frac{TP}{TP + FN}$$

$$SP = \frac{TN}{TN + FP}$$

$$F = \frac{2TP}{2TP + FP + FN}$$

where true positive (TP) represents the number of positive pairs that are predicted to be interacting, whereas false positive (FP) is the count of negative pairs that are predicted to be interacting. Similarly, true negative (TN) is the total of negative pairs that are predicted to be non-interacting and false negative (FN) represents the number of positive pairs that are predicted to be non-interacting.

In addition, the receiver operating characteristic (ROC) is another important tool to assess the generalization performance of the model. ROC curve is a plot of the true positive rate (TPR) and false positive rate (FPR) which depicts the performance of a predictor at various threshold values. To compare these curves, area under the curve (AUC) is computed by summing the areas under the ROC curve.

A similar metric, the precision-recall curve (PR curve), can be obtained by using precision and recall at multiple threshold settings. The precision and recall ratio are defined as:

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

Area under the precision-recall (AUPR) can also be obtained by summing the areas under the PR curve. For skewed datasets like the DTIs datasets in this paper, AUPR is of more significance because it penalizes the false positives more as compared to AUC, and is thus more suitable for imbalanced datasets. The higher the value of AUPR, the better [20,21]. The general framework of the PsePDC-DTIs model is shown in Figure 1 for an intuitive understanding.

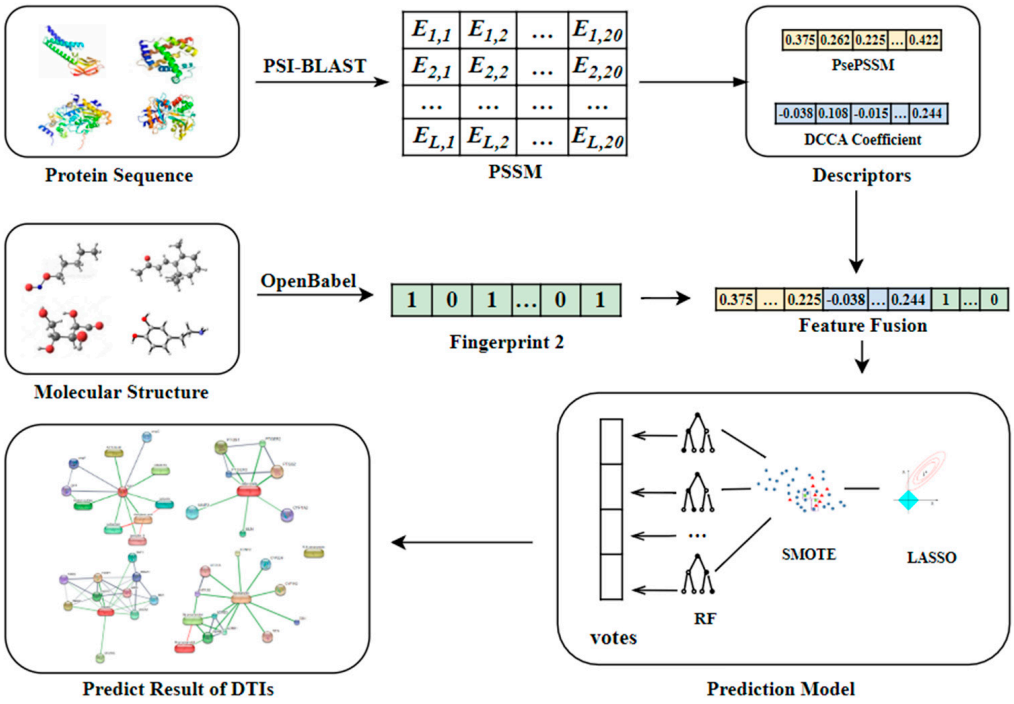


**Figure 1.** The workflow of PsePDC-DTIs.

### *2.2. Features Generation*

#### 2.2.1. Parameter Setting for PsePSSM and DCCA Coefficient

Both the PsePSSM and the DCCA coefficient algorithms can control the validity of the algorithm to extract the feature information of the protein sequences by adjusting some of the parameters in the algorithm. The selection of parameters $\lambda$ and $s$ is very important for the accuracy of a protein-target interactions prediction model. In order to discover the merits of the feature parameters, we use the benchmark datasets as the research object, while the optimal values of $\lambda$ and $s$ are selected by the prediction accuracy and average prediction accuracy of the four datasets under different parameters [22,23].

In this paper, the $\lambda$ value of PsePSSM algorithm indicates the sequence-order information of the amino acid residues in the protein sequence. To find the optimal $\lambda$ value, we set the $\lambda$ values from 0 to 15 in order. For the different $\lambda$ values, the gold standard datasets enzymes, ion channels, GPCRs, and nuclear receptors are classified by RF and tested by 5-fold cross-validation respectively. The results can be seen in Supplementary Table S2. To find the best $\lambda$ value more intuitively, the prediction accuracy and average accuracy with different $\lambda$ values for the four datasets is shown in Supplementary Figure S1. In order to unify the parameters in the model, we take the $\lambda$ value corresponding to the highest average accuracy for the optimal parameter which is up to 97.425% with $\lambda$ = 3. Therefore, an 80-dimensional feature vector is acquired when using the PsePSSM method with the optimal parameter $\lambda$ value of 3 to extract features of each target protein.

The $s$ value of DCCA coefficient algorithm determines the length of each overlapping segment in which the covariance and variance of the residuals are calculated. In the gold standard datasets E, IC, GPCRs, and NR, the length of the shortest protein sequence is 83, therefore, the maximum $s$ value is allowed for 82. To find the optimal $s$ value, we set $s$ values from 9 to 81 in turn. For the different $s$ values, benchmark datasets are classified by RF and tested by 5-fold cross-validation respectively. In order to unify the parameters in the model, we take the $s$ value corresponding to the highest average accuracy for the optimal parameter, which is up to 97.2925% when $s$ = 36. The prediction accuracy and average prediction accuracy with different values of the four datasets is shown in Supplementary Figure S2 and Table S3.

#### 2.2.2. The Dimensionality of the Generated Features

We can obtain a 526-dimension feature vector which is composed of an 80-dimension vector generated by PsePSSM, 190-dimension vector generated by DCCA coefficient, and 256-dimension vector of the FP2 format molecular fingerprint.

### *2.3. Predictive Performance of Lasso for Dimensionality Reduction*

As mentioned above, there are 526-dimension features for prediction, and the Lasso dimensionality reduction algorithm can extract useful information and discard redundancy from the complex information in the feature vector, which can improve the prediction process to some extent. The performance evaluation parameters for Lasso are shown in Supplementary Table S4. As we can see from Supplementary Table S4, the values of different indicators are comparable before and after using Lasso, which illuminates the ability of Lasso for extracting useful information.

### *2.4. Predictive Performance of SMOTE for Imbalanced Datasets*

The classification of data with imbalanced class presents a significant drawback of the performance attainable using most standard classifier learning algorithms, which assume a relatively balanced class distribution and equal misclassification costs [24]. For this reason, as mentioned above, the SMOTE method has been used to convert the Lasso feature-selected data from imbalanced to balanced form, which is implemented in the DMwR R package where the oversampling parameter, the undersampling parameter, and the nearest neighbor algorithm parameter are set to 500, 120, and 5, respectively.

Due to the number of positive examples is much smaller than the number of negative examples, the indicators SE and SP are proportional to the correct proportion of positive and negative examples in the sample, and the indicator ACC has no significance in measuring the merits of the algorithm [23]. Therefore, the indicators that can reasonably measure the evaluation performance of the prediction model are AUC and AUPR among the above-mentioned indicators. To reflect the effect of data balance on the prediction performance of the model more directly, the visual display of the AUC and AUPR comparison under NR, GPCR, IC, and E datasets on unbalanced datasets and balanced datasets is shown in Figure 2. The evaluation indicators mentioned above are shown in Supplementary Table S5.

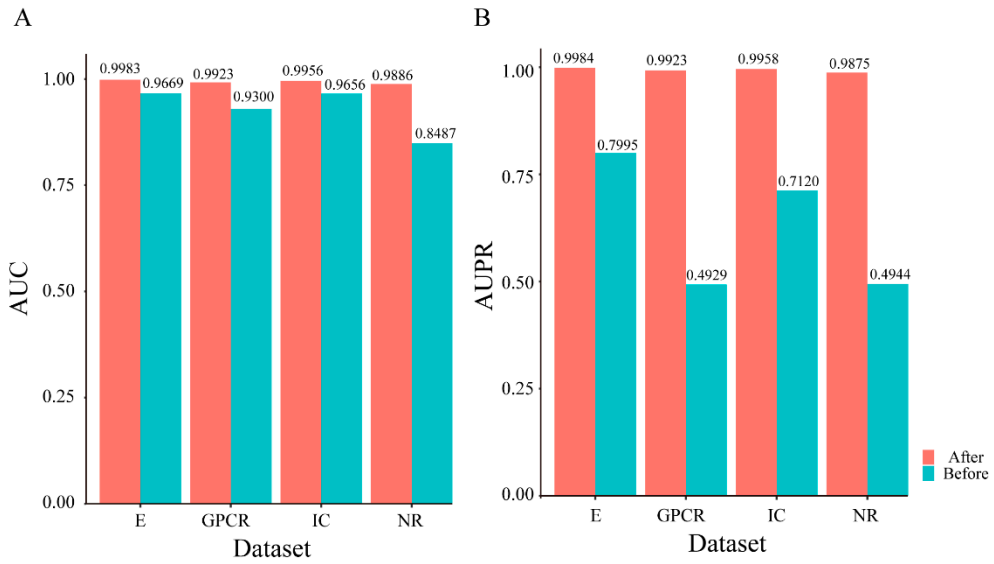


**Figure 2.** Predictive performance comparison of PsePDC-DTIs on benchmark datasets before and after SMOTE optimization in terms of AUC (**A**) and AUPR (**B**).

Figure 2 illuminates that the datasets after SMOTE processing have been vastly improved as far as AUC and AUPR are concerned. For the increase of AUC value after balancing, the highest is in the NR dataset with 0.1399, followed by GPCR dataset with 0.0623, E dataset with 0.0314, and IC dataset with 0.0300. As to AUPR, the highest is in GPCR dataset with 0.4994, followed by NR dataset with 0.4931, IC dataset with 0.2838, and E dataset with 0.1989. So, we can conclude that SMOTE processing lead to a greater improvement in the prediction performance.

### *2.5. Predictive Performance of RF for DTIs Prediction*

A classifier plays an important role in the quality of a prediction model, and thus might influence the prediction performance. In order to explore the machine learning (ML) methods which are used frequently, we investigate seven common classification algorithms of ML (i.e., random forest, naïve Bayes, decision tree, support vector machine, oneR, k-nearest neighbors, repeated incremental pruning to produce error reduction).

To ensure fairness, the target protein sequences are extracted by PsePSSM and DCCA coefficient for the four datasets constructed, and the drug compounds are expressed by FP2 format molecular fingerprint descriptor. After fusing the features, the Lasso method for dimensionality reduction and SMOTE method for skewed datasets are used. To obtain robust results and accurate comparison, we keep the same experimental conditions, where the same training drugs-target interaction pairs and test drugs-target interaction pairs are used across the seven classifiers in each cross-validation [25]. The prediction results on four datasets of seven classifiers are shown in Supplementary Table S6.

From the boldfaced fonts in Supplementary Table S6, we observe that RF significantly outperformed the other machine learning methods under four datasets in terms of ACC,

SP, F, AUC, and AUPR metrics. However, for the SE metric, SVM secured the first position with SE values of 96.98%, 97.50%, and 97.81% which is 1.15%, 1.96%, and 0.20% higher than RF in GPCR, IC, and E datasets, respectively. However, each of the SE values in these three datasets is over 95%, which means that more than 95% of actual DTIs can be correctly identified.

Figure 3 shows one of the ROC curves of seven different classification algorithms under the NR, GPCR, IC, and E datasets in five-fold cross-validation, while other ROC curves are shown in Supplementary Figure S3. Figure 4 reveals one of the PR curves of seven different classifiers under four datasets in five-fold cross-validation, while other PR curves can be found in Supplementary Figure S4.

According to Figures 3 and 4, the ROC and PR curves of the four datasets almost surround others with random forest as the classifier, and the corresponding AUC and AUPR values are also larger. Therefore, we choose random forest as the classification algorithm of the prediction model.

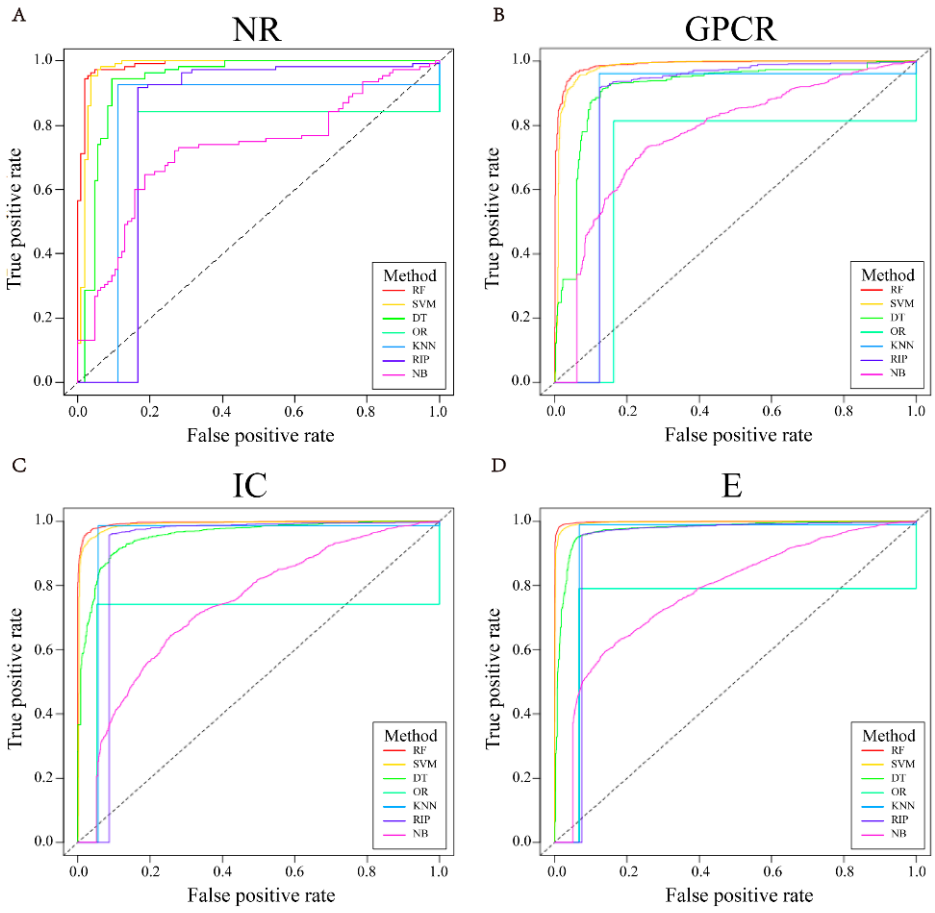


**Figure 3.** The representative ROC curves of different classifiers in 5-fold cross-validation of four datasets, i.e., NR (**A**), GPCR (**B**), IC (**C**), E (**D**).

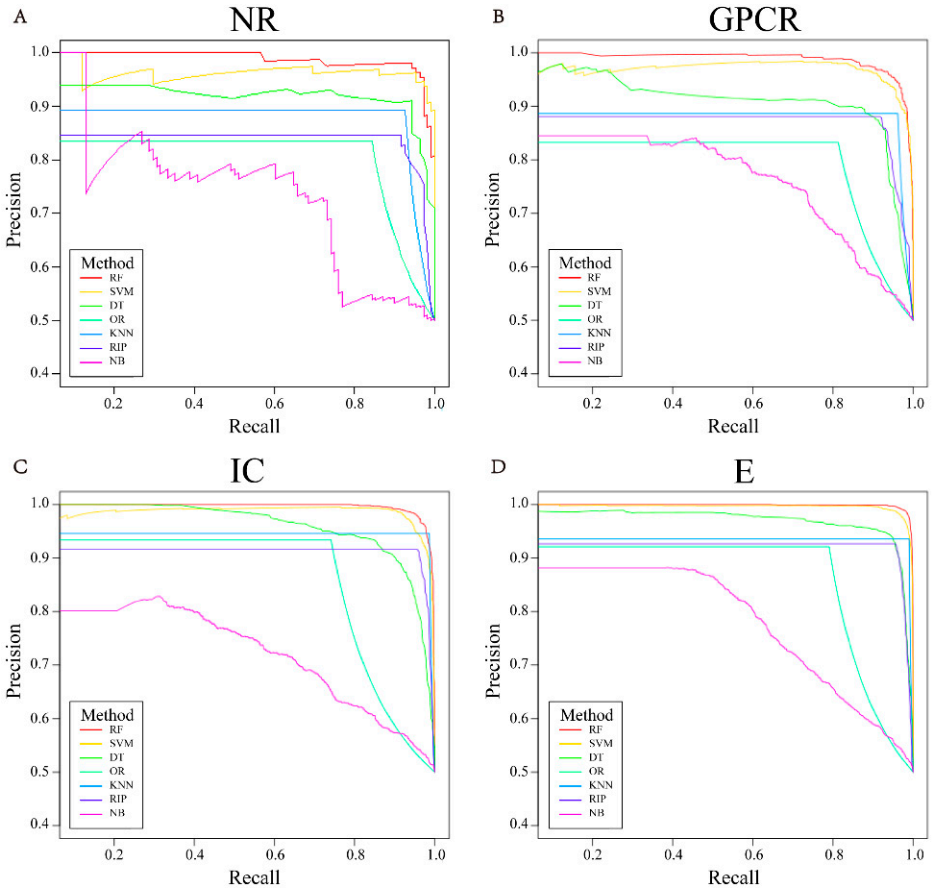


**Figure 4.** The representative PR curves of different classifiers in 5-fold cross-validation of four datasets, i.e., NR (**A**), GPCR (**B**), IC (**C**), E (**D**).

### *2.6. Predictive Performance of PsePDC-DTIs Compared with State-of-the-Art Methods*

There are a variety of prediction models proposed for detecting DTIs. Our method applies LASSO to select features and SMOTE to balance data for DTIs under gold standard datasets and evaluates prediction performance based on five-fold cross-validation. To further expound the efficiency of the predictor in this study, we compared our prediction performance with other methods which also used the same benchmark datasets and tested by five-fold cross-validation [16]. Table 1 lists the comparison results of other models, including NetCBP [26], Huang et al. [27], Bigram-PSSM [21], iDTI-ESBoost [20], Li et al. [28], KBMF2K [29], and NRLMF [30]. It can be seen that our predictor PsePDC-DTIs achieves AUC values of 0.9886, 0.9923, 0.9956, and 0.9983 on the NR, GPCR, IC, and E datasets, respectively, which significantly outperforms other methods for all datasets.

**Table 1.** Performance comparison of different approaches on benchmark datasets in terms of AUC.

| Models | NR | GPCR | IC | E |
|---|---|---|---|---|
| NetCBP [26] | 0.8394 | 0.8235 | 0.8034 | 0.8251 |
| Huang et al. [27] | 0.9634 | 0.9053 | 0.9382 | 0.9601 |
| Bigram-PSSM [21] | 0.8690 | 0.8720 | 0.8890 | 0.9480 |
| iDTI-ESBoost [20] | 0.9285 | 0.9322 | 0.9369 | 0.9689 |
| Li et al. [28] | 0.9300 | 0.9171 | 0.8856 | 0.9288 |
| KBMF2K [29] | 0.8240 | 0.8570 | 0.7990 | 0.8320 |
| NRLMF [30] | 0.9500 | 0.9690 | 0.9890 | 0.9870 |
| **PsePDC-DTIs** | **0.9886** | **0.9923** | **0.9956** | **0.9983** |

Moreover, Mousavian et al. [21] argued that the AUPR is a more accurate measure for evaluating performance in dealing with highly imbalanced datasets compared to the AUC for the reason that the highly ranked false positive samples are punished by the AUPR much more than the AUC. To compare the performance in terms of AUPR among Bigram-PSSM [21], iDTI-ESBoost [20], and NRLMF [30], we reported the AUPR values of the three predictors in Table 2. The AUPR values of our model PsePDC-DTIs are 0.9875, 0.9923, 0.9958, and 0.9984 on the NR, GPCR, IC, and E datasets, respectively. This clearly shows that our method PsePDC-DTIs outperforms other methods in terms of AUPR as well.

**Table 2.** Performance comparison of different approaches on benchmark datasets in terms of AUPR.

| Models | NR | GPCR | IC | E |
|---|---|---|---|---|
| Bigram-PSSM [21] | 0.4110 | 0.2820 | 0.3900 | 0.5460 |
| iDTI-ESBoost [20] | 0.7900 | 0.5000 | 0.4800 | 0.6800 |
| NRLMF [30] | 0.7280 | 0.7490 | 0.9060 | 0.8920 |
| **PsePDC-DTIs** | **0.9875** | **0.9923** | **0.9958** | **0.9984** |

The values of AUC and AUPR demonstrated above indicate the effectiveness of the extracted feature information, dimensionality reduction of features, balancing methods, and classifier proposed in this research.

### *2.7. Predictive Performance of PsePDC-DTIs Compared with State-of-the-Art Methods*

According to the information introduced above, we can confirm the reliability of our proposed model. In the inference process, we use all the known drugs and target proteins in the gold standard datasets as training data, and predict potential interactions between 52 human proteins and 1556 FDA approved drugs as mentioned in the datasets section.

For the 52 breast cancer target proteins and the PsePDC-DTIs model trained in gold standard datasets, we predict all the DTIs mentioned in Section 4.1.2 and rank them by their probability. There are 383 predicted DTIs with a probability greater than 0.5 reported in Supplementary Table S7, which means 0.47% pairs were predicted as interaction. This is in line with the fact that the number of non-interacting pairs is far more than the number of interaction pairs [21]. We extract the top 10 drug–target pairs ranked by their prediction probability values, as listed in Table 3, and present the potential mechanism of predicted DTIs in Figure 5. Figure 5A shows that IP3R, the target of caffeine (DB00201), regulates KCNN4 via $Ca^{2+}$ in the salivary secretion pathway. Figure 5B demonstrates that GF, the target of afatinib (DB08916), regulates RTK directly.

**Table 3.** Drug-target pairs ranked by prediction probability.

| Drug | Drug_Name | Target | Target_Name | Prob |
|---|---|---|---|---|
| **DB00201** | **Caffeine** | **hsa3783** | **KCNN4** | **0.988** |
| DB00277 | Theophylline | hsa3783 | KCNN4 | 0.982 |
| DB01412 | Theobromine | hsa3783 | KCNN4 | 0.93 |
| DB00530 | Erlotinib | hsa238 | ALK | 0.886 |
| DB00806 | Pentoxifylline | hsa3783 | KCNN4 | 0.884 |
| DB00824 | Enprofylline | hsa3783 | KCNN4 | 0.866 |
| **DB00530** | **Erlotinib** | **hsa2263** | **FGFR2** | **0.864** |
| DB00661 | Verapamil | hsa57719 | ANO8 | 0.846 |
| DB01303 | Oxtriphylline | hsa3783 | KCNN4 | 0.844 |
| **DB08916** | **Afatinib** | **hsa2263** | **FGFR2** | **0.806** |

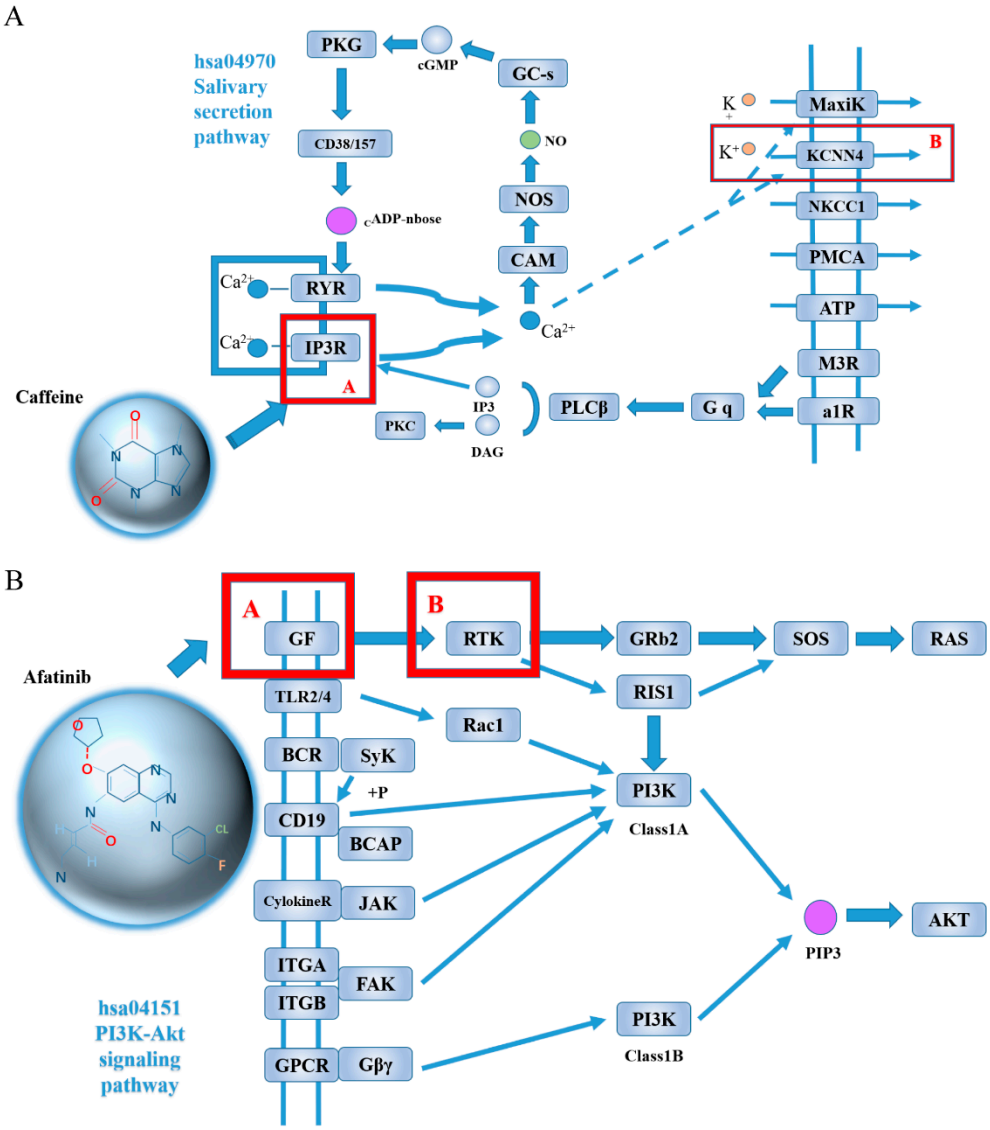


**Figure 5.** The potential mechanism illustration of predicted DTIs of Caffeine and KCNN4 (**A**), Afatinib and FGFR2 (**B**). In Figure 5B, GF contains EGF, RTK contains targets of EGFR, ERBB2, ERBB4, FGFR2.

## 3. Discussion

As shown above, the average values of accuracy in GPCR, IC, and E datasets are 95.28%, 96.19%, 96.74%, and 98.22%, respectively. The average AUC achieves 0.9886, 0.9923, 0.9956, and 0.9983 on the NR, GPCR, IC, and E datasets, respectively, which outperformed some methods reported elsewhere [20,21,26–29]. In the literature [21], it has been revealed that AUPR is the most appropriate metric for the comparison of such imbalanced datasets. Moreover, the average AUPR of PsePDC-DTIs outperformed other methods [20,21], reaching 0.9875, 0.9923, 0.9958, and 0.9984 under these four datasets, respectively.

During predicting DTIs between 52 human proteins of breast cancer and 1556 FDA approved drugs in the DrugBank database, our comprehensive model provides us with 10 potential DTIs, among which three DTIs are found with direct or inferred evidence. There is direct evidence about the DTIs of erlotinhas(DB00530) and FGFR2 (hsa2263) in SuperTarget. In addition, we obtain indirect evidence of predicted DTIs when the known target for a drug regulates the predicted target for this drug by using pathways from the KEGG database. Figure 5A shows that the target IP3R for caffeine (DB00201), which can be

found in DrugBank datasets, regulates KCNN4 via $Ca^{2+}$ in the salivary secretion pathway. This is important with respect to the fact that several studies demonstrate the relationship of salivary to breast cancer [31–38]. For example, Sawczuk et al. [36] indicated that salivary peroxidase may have particular clinical significance in non-invasive diagnostics of breast cancer. Liu et al.'s study [37] contributed to the screening of patients with early-stage breast cancer based on precise alterations of salivary glycopatterns.

Furthermore, we find six pathways to explain the relationship between afahasib (DB08916) and FGFR2 (hsa2263). Taking PI3K-Akt signaling pathway as an example, Figure 5B shows that the target GF (contains EGF) for afatinib (DB08916) which can be found in DrugBank datasets regulates RTK (contains targets of EGFR, ERBB2, ERBB4, FGFR2) directly. Again, this is significant as several studies suggest that PI3K-Akt signaling pathway is connected with breast cancer [39–61]. Chandarlapaty et al. [44] prospectively collected trastuzumab-refractory human breast cancers, and found that activation of the PI3K-Akt pathway through loss of PTEN or PIK3CA mutation was frequently observed. Other pathways about afatinib (DB08916) and FGFR2 (hsa2263) can be found in Supplementary Figure S5.

In the remaining predicted seven DTIs, although we could not find any evidence from databases, pathways, and literature, they still have the potentiality to be true positive DTIs [62]. For instance, some researches [63] propose that theophylline (DB00277) and caffeine (DB00201) are often regarded as a group which is related to breast cancer. Thus, it is possible that both drugs interact with the same target.

However, if the drug–target interactions dataset as training data is too large, the PsePDC-DTIs model cannot predict drug–target interactions rapidly because we use RF as classifier. Therefore, in order to improve the operating speed of the proposed model and keep the prediction accuracy, in the future, we will attempt to use a deep learning network as classifier. Moreover, to handle the class imbalance problem, our proposed model used SMOTE to generate artificial examples for the minority class. However, during the cross-validation process, the test dataset also contains the artificial examples generated by SMOTE, which may cause the current reported prediction performance exaggeration. Therefore, we will explore a more conservative and effective method for dealing with imbalanced data. In addition, further research into the new methods of the features will be essential because the algorithm of extracting the feature information of the protein sequences and drug compounds is very important for the performance of a protein–target interactions prediction model.

## 4. Materials and Methods

### *4.1. Datasets*

#### 4.1.1. Benchmark Datasets

The benchmark datasets are used for assessing the performance of PsePDC-DTIs by five-fold cross-validation. For this study, they are the gold standard datasets studied by Yamanishi et al. [64], obtained from http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/ (accessed on 9 December 2021). All data concerning DTIs pairs in the gold standard datasets are collected from the KEGG BRITE [65], BRENDA [66], SuperTarget [67], and Drug Bank databases [68]. The drug target links have been considered for four protein targets, namely enzymes(E), ion channels (IC), G-protein-coupled receptors (GPCR), and nuclear receptor (NR). As listed in Table 4, the number of known drugs target in these classes is 445, 210, 223, and 54, respectively, and the number of proteins known to be targeted by the drugs in these classes is 664, 204, 95, and 26 respectively. Among these drug-target pairs, 5127 pairs are known to interact with each other, and the number of interacting pairs in each class is 2926, 1476, 635, and 90 respectively.

**Table 4.** The benchmark data sets used in this study.

| Datasets | Drugs | Targets | Interactions | Positive Samples | Negative Samples | Sample Ratio |
|---|---|---|---|---|---|---|
| Enzyme | 445 | 664 | 2926 | 2926 | 292,554 | 99.98 |
| IC | 210 | 204 | 1476 | 1476 | 41,364 | 28.02 |
| GPCR | 223 | 95 | 635 | 635 | 20,550 | 32.36 |
| NR | 54 | 26 | 90 | 90 | 1314 | 14.60 |
| Total | 932 | 989 | 5127 | 5127 | 355,782 | - |

For a completely connected bipartite graph, there must be drugs × targets connections. Taking the enzyme dataset, for instance, there exist 455 × 664 = 302,120 drug-target pairs. In our study, 5172 pairs which are known to interact with each other are used as the positive samples while the rest of connections are considered negative samples. The number of samples for four datasets is listed in Table 4.

#### 4.1.2. DTIs Dataset Constructed by Drugs of FDA-Approved and Targets of Breast Cancer

In order to predict new DTIs of breast cancer treatment for drugs approved by the FDA, we propose a DTIs dataset whose drugs are from a dataset named DrugBank_approved [69] which contains 1556 FDA-approved drugs until 2016. As to the targets, we use the 110 putative target genes of breast cancer identified by Baxter et al. [10] and 179 genes whose predicted expression was associated with breast cancer risk [11] for drug repurposing. According to these 286 genes (removing 3 duplicates), we obtain 52 human proteins annotated as members of the four classes of target proteins (NR, GPCR, IC, and E) in KEGG GENES, which are listed in Supplementary Table S1. The DTIs which are generated by connecting each target protein with each drug molecule (only target protein and drug can be linked by aside) can be used for drug repurposing of breast cancer.

### *4.2. Methods for Features Generation*

#### 4.2.1. Pseudo-Position Specific Scoring Matrix (PsePSSM)

The PsePSSM algorithm employed in the study is proposed by K.C. Chou [70]. PsePSSM is the extraction of the features of protein sequences, which can be obtained by translating the position specific scoring matrix (PSSM) of different dimensions for different protein sequences into the same dimension. The uniform vector is convenient for our subsequent study. PSSM [71] represents the evolutionary information of the protein sequences, which needs to blast the protein FASTA file against the UniProt database for constructing through PSI-BLAST [72]. For this study, the parameters of PSI-BLAST are set with three iterations, E-value is equal to 0.001, while the rest of the parameters are set by default. The constructed PSSM format for a protein sequence P with L amino acid residues is shown as formula (1). The rows of PSSM inform the corresponding amino acid positions in the protein sequence P, and columns of PSSM indicate the 20 native amino acid types that may be mutated.

$$P_{PSSM} = \begin{bmatrix} E_{1,1} & E_{1,2} & \cdots & E_{1,j} & \cdots & E_{1,20} \\ E_{2,1} & E_{2,2} & \cdots & E_{2,j} & \cdots & E_{2,20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ E_{i,1} & E_{i,2} & \cdots & E_{i,j} & \cdots & E_{i,20} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ E_{L,1} & E_{L,2} & \cdots & E_{L,j} & \cdots & E_{L,20} \end{bmatrix} \quad (1)$$

where $E_{i,j}$ represents the value of the residue of the $i$-th position in the amino acid sequence being mutated to the $j$-th native amino acid residue.

The elements of PSSM are normalized by formula (2), whose PSSM value ranges from 0 to 1, while the value in the original PSSM matrix ranges from −9 to 11.

$$E'_{i,j} = \frac{1}{1 + e(-E_{i,j})} \tag{2}$$

Because proteins with different lengths will correspond to matrices with different numbers of rows, in order to make the PSSM descriptor a uniform representation, a protein sequence P is represented by formula (3):

$$\overline{P}_{PSSM} = \begin{bmatrix} \overline{E}_1 & \overline{E}_2 & \cdots & \overline{E}_j & \cdots & \overline{E}_{20} \end{bmatrix}^T \tag{3}$$

where $\overline{E}_j = \frac{1}{L}\sum_{i=1}^{L} E'_{i,j}$, $\overline{E}_j$ manifests the average score of the amino acid residue in protein $P$ being mutated to $j$ amino acid type during the process of evolution.

Next, we transform PSSM of a single protein into a feature vector PsePSSM, as formula (4) shown.

$$P_{PsePSSM} = \begin{bmatrix} \overline{P}_{PSSM}{}^T & \theta_1^1 & \theta_2^1 & \cdots & \theta_{20}^1 & \cdots & \theta_1^\lambda & \theta_2^\lambda & \cdots & \theta_{20}^\lambda \end{bmatrix}^T \tag{4}$$

where $\theta_j^\lambda = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda}(E'_{i,j} - E'_{(i+\lambda),j})^2 (\lambda < L; j = 1, 2, \cdots, 20)$, $\theta_j^\lambda$ is the correlation factor by coupling the $\lambda$-th-most contiguous PSSM scores along the protein chain for the amino acid type $j$. Therefore, a protein sequence generates a 20 + 20 × $\lambda$-dimensional feature vector using PsePSSM algorithm. PsePSSM matrix can be regarded as PSSM matrix when $\lambda$ = 0. For this study, the optimal parameter of $\lambda$ needs to be selected, so that the highest accuracy of a protein–target interactions prediction model is obtained.

4.2.2. Detrended Cross-Correlation Analysis Coefficient (DCCA Coefficient)

Using the detrended cross-correlation analysis coefficient method, more protein information that truly reflects protein samples' intrinsic correlation could be extracted from the PSSM matrix. DCCA coefficient was initially proposed by Podobnik and Stanley [73], which can be used to quantify the level of cross-correlation between two non-stationary time series [74]. Here, each amino acid is taken as one property and the PSSM including the evolutionary information expression is considered as the time series of all properties. The 20 columns in the PSSM matrix are considered to be 20 non-stationary time series [22,75].

For two arbitrary different columns of a normalized PSSM, $\{x_i\}$ and $\{y_i\}$ $(i = 1, 2, \cdots, L)$, new time series $X_k$ and $Y_k$ are calculated by using formula (5).

$$\begin{cases} X_k = \sum_{i=1}^{k} x_i & k = 1, 2, \cdots, L \\ Y_k = \sum_{i=1}^{k} y_i & k = 1, 2, \cdots, L \end{cases} \tag{5}$$

Then, the integrated time series $X_k$ and $Y_k$ are divided into $(L - s)$ overlapping segments, and each segment which starts at $i$ and ends at $I + s$ contains $(s + 1)$ values. For each segment of the data, the fitting values $\widetilde{X}_{i,k}$ and $\widetilde{Y}_{i,k} (i \le k \le i + s)$ can be obtained by the least squares linearly fitting. The covariance and variance of the residuals in each segment are calculated by formula (6)–(8):

$$f_{xy}^2(s,i) = \frac{1}{s+1}\sum_{k=i}^{i+s}(X_k - \widetilde{X}_{i,k})\left(Y_k - \widetilde{Y}_{i,k}\right) \tag{6}$$

$$f_{xx}^2(s,i) = \frac{1}{s+1}\sum_{k=i}^{i+s}(X_k - \widetilde{X}_{i,k})^2 \tag{7}$$

$$f_{yy}^2(s,i) = \frac{1}{s+1}\sum_{k=i}^{i+s}(Y_k - \widetilde{Y}_{i,k})^2 \tag{8}$$

Next, we average all $(L - s)$ overlapping segments and obtain the fluctuation function shown in formula (9)–(11):

$$f_{xy}^2(s) = \frac{1}{L-s}\sum_{i=1}^{L-s} f_{xy}^2(s,i) \tag{9}$$

$$f_{xx}^2(s) = \frac{1}{L-s}\sum_{i=1}^{L-s} f_{xx}^2(s,i) \tag{10}$$

$$f_{yy}^2(s) = \frac{1}{L-s}\sum_{i=1}^{L-s} f_{yy}^2(s,i) \tag{11}$$

Finally, the DCCA coefficient of two different time series $\{x_i\}$ and $\{y_i\}$ is defined as formula (12). Hence, for the 20 columns in the PSSM matrix considered to be 20 non-stationary time series, a 190-dimensional feature vector will be generated for a certain $s$ via the DCCA coefficient algorithm. We need to select the optimal parameter of $s$ to obtain the highest accuracy of a protein–target interactions prediction model.

$$\rho_{DCCA}(s) = \frac{f_{xy}^2(s)}{f_{xx}(s)f_{yy}(s)} \tag{12}$$

The value of $\rho_{DCCA}$ ranges from $-1 \leq \rho_{DCCA} \leq 1$. Logically, 1 means perfect cross-correlation, 0 represents no cross-correlation, and $-1$ indicates perfect anti-cross-correlation [76].

#### 4.2.3. FP2 Molecular Fingerprint

Drug compounds are expressed by FP2 format molecular fingerprint descriptor that can be converted to a decimal digit sequence between 0 and 15 as a drug molecule 256-dimensional vector using OpenBabel Software (available from http://openbabel.org, accessed on 9 December 2021) [23].

### *4.3. Lasso for Dimensionality Reduction of Features*

Shi et.al [23] proved that the least absolute shrinkage and selection operator (Lasso) method can effectively reduce information redundancy and delete some unimportant features compared with principal components analysis (PCA), ReliefF, and Elastic net. Therefore, we use Lasso as the dimensionality reduction algorithm for this paper. LASSO proposed by Tibshirani [77] is a compression estimation method with $l_1$ regularization implemented to achieve a sparse solution. LASSO is used to perform feature selection by forcing many parameters corresponding to the irrelevant and redundant features to zero value, and retaining the features corresponding to the non-zero coefficients for subsequent classification [78–80]. The aim of this approach is to minimize the cost function:

$$\sum_{n=1}^{N}\left(y_n - \sum_m x_{nm}\beta_m\right)^2 + \lambda\sum_{m=1}^{M}|\beta_m| \tag{13}$$

where $y_n$ represents the corresponding response vector of a DTI pair, that is, the class label of the sample, $N$ is the number of samples, $x_{nm}$ is the $m$-th feature of the $n$-th sample, $\lambda$ is the regularization parameter, and $\beta_m$ is the regression coefficients of $m$-th feature [78].

Therefore, through formula (13), we eliminate the noise and redundant information contained in the high-dimensional data obtained after the original drug and target feature extraction

### *4.4. SMOTE for High-Dimensional Class-Imbalanced Data*

As shown in Table 4, there are severe imbalance problems between the positive and negative samples of four gold standard datasets. The ratio of negative samples to positive samples (sample ratio) is used for measuring the degree of imbalance. There is a high degree of imbalance in the enzyme dataset with the sample ratio reaching 99.98. In contrast, the nuclear receptor dataset has a low degree of imbalance, with a sample ratio that

barely reaches 14.60. In order to deal with imbalanced data, some important techniques are proposed, such as random undersampling, random oversampling, and the synthetic minority oversampling technique (SMOTE). SMOTE overcomes imbalances by generating artificial data, while random undersampling and random oversampling replicate and add the observations from the minority class [80]. Therefore, this study uses SMOTE, which is a powerful method and creates artificial data based on feature space similarities from minority samples to handle the problems.

SMOTE, proposed by Chawla et al. [81], is one of the most popular oversampling methods. Its main idea is to interpolate a new synthetic minority class sample on the line that connects a randomly chosen minority class sample and one of its k-nearest neighbors belonging to the minority class samples. Specifically speaking, for each positive sample $z$, one gets its k-nearest neighbors from other positive samples. Then, one chooses one positive sample $\overline{z}$ among the neighbors [82]. Finally, this generates the synthetic sample $z_{new}$ by inserting between $z$ and $\overline{z}$ as follows:

$$z_{new} = z + rand(0,1) \times (\overline{z} - z) \tag{14}$$

where $rand(0,1)$ refers to generate a random number between 0 and 1. Thus, a new, more balanced dataset is formed.

### *4.5. RF for DTIs Prediction*

Random forest (RF) [83] is one of the famous bagging techniques based on decision tree models which is fast, robust to noise, does not overfit, but provides possibilities for the explanation and visualization of its output. In this study, RF was applied as a classification method by constructing a multitude of decision trees at training time and outputting the number of votes cast of all the trees [84]. Supposing the number of training cases were $P$ and the total number of features in the classifier were $Q$. After making $p$ bootstrap sample sets from the original training sample set, set up an unpruned tree with each sample set. At each node of the tree, randomly choose $q$ features ($q < Q$) as a candidate variable on which to make the decision at that node [85]. With the generation of multiple classification trees, a random forest is built

## 5. Conclusions

In this paper, we develop a novel method for predicting and identifying DTIs, called PsePDC-DTIs. Specifically, the proposed method combines the pseudo-position specific scoring matrix (PsePSSM) and detrended cross-correlation analysis coefficient (DCCA coefficient) to extract the features of the protein sequences, for which PsePSSM feature extraction considers the sequence-order information of the protein sequence, and the DCCA coefficient uses the columns in the PSSM as the least squares fitting and the trend elimination as the non-stationary time series to remove the PSSM between the cross-correlation [22]. When using PsePSSM and DCCA coefficient, $\lambda$ = 3 and $s$ = 36 are selected, respectively. Drug compounds are expressed by FP2 format molecular fingerprint descriptor. The redundant information in the drug–target datasets is effectively removed by least absolute shrinkage and selection operator (Lasso). For dealing with the high degree of imbalance in the samples used in this study, the synthetic minority oversampling technique (SMOTE) is employed. The classification algorithm to predict DTIs is the random forest (RF) classifier. The five-fold cross-validation method is used in this work to assess the predictive performance of PsePDC-DTIs on four benchmark datasets. The PsePDC-DTIs model has achieved good prediction results, which shows that the proposed method is better than the state-of-art methods and appropriately designed.

**Supplementary Materials:** The following are available online, Figure S1: Prediction result of selecting different $\lambda$ on four datasets, Figure S2: Prediction result of selecting different $s$ on four datasets, Figure S3: the ROC curves of different classifiers in 5-fold cross-validation, Figure S4: the PR curves of different classifiers in 5-fold cross-validation, Figure S5: the inferred evidence for DTIs of afatinib

(DB08916) and FGFR2 (hsa2263). (a) RTK contains EGFR, FGFR2, GF contains EGF; (b) GF contains EGF, RTK contains FGFR2, EGFR; (c) GF contains EGF, GFR contains EGFR, ERBB2, FGFR2; (d) RTK contains EGFR, ERBB2, ERBB4, FGFR2, GF contains EGF; (e) RTK contains FGFR2, EGFR, GF contains EGF. Table S1: Targets for drug repurposing of breast cancer, Table S2: Prediction result of selecting different $\lambda$ on four datasets, Table S3: Prediction result of selecting different $s$ on four datasets, Table S4: Prediction results on four datasets before and after Lasso for dimensionality reduction, Table S5: Prediction results on four datasets before and after SMOTE optimization, Table S6: Prediction results on four datasets of seven classifiers, Table S7: Drug-target interaction pairs with a probability score no less than 0.5.

**Author Contributions:** Conceptualization, J.S. and Z.X.; data curation, M.W.; formal analysis, Z.X.; funding acquisition, Y.H.; investigation, K.L. and Y.H.; methodology, J.S.; project administration, Y.H.; resources, Z.X.; software, J.S.; supervision, K.L.; validation, L.C. and M.W.; visualization, J.S.; writing—original draft, J.S.; writing—review & editing, K.L. and Y.H. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Natural Science Foundation of China, grant number 81773550, 82173615.

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

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The datasets generated during and analysed during the current study are available in the github repository, https://github.com/SongJialiJiali/PsePDC-DTIs-model.git (accessed on 9 December 2021).

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

**Sample Availability:** Not applicable.

## References

1. Maurya, A.P.; Brahmachari, S. Current Status of Breast Cancer Management in India. *Indian J. Surg.* **2020**. [CrossRef]
2. Bray, F.; Ferlay, J.; Soerjomataram, I.; Siegel, R.L.; Torre, L.A.; Jemal, A. Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **2018**, *68*, 394–424. [CrossRef]
3. Green, M.; Raina, V. Epidemiology, Screening and Diagnosis of Breast Cancer in the Asia–Pacific Region: Current Perspectives and Important Considerations. *Asia Pac. J. Clin. Oncol.* **2008**, *4*, S5–S13. [CrossRef]
4. You, J.; McLeod, R.D.; Hu, P. Predicting Drug-Target Interaction Network Using Deep Learning Model. *Comput. Biol. Chem.* **2019**, *80*, 90–101. [CrossRef]
5. Chong, C.R.; Sullivan, D.J. New Uses for Old Drugs. *Nature* **2007**, *448*, 645–646. [CrossRef]
6. Kola, I.; Landis, J. Can the Pharmaceutical Industry Reduce Attrition Rates? *Nat. Rev. Drug Discov.* **2004**, *3*, 711–715. [CrossRef]
7. Kamb, A.; Wee, S.; Lengauer, C. Why Is Cancer Drug Discovery so Difficult? *Nat. Rev. Drug Discov.* **2007**, *6*, 115–120. [CrossRef]
8. Chen, H.-R.; Sherr, D.H.; Hu, Z.; DeLisi, C. A Network Based Approach to Drug Repositioning Identifies Plausible Candidates for Breast Cancer and Prostate Cancer. *BMC Med. Genom.* **2016**, *9*, 51. [CrossRef]
9. DiMasi, J.A.; Hansen, R.W.; Grabowski, H.G. The Price of Innovation: New Estimates of Drug Development Costs. *J. Health Econ.* **2003**, *22*, 151–185. [CrossRef]
10. Baxter, J.S.; Leavy, O.C.; Dryden, N.H.; Maguire, S.; Johnson, N.; Fedele, V.; Simigdala, N.; Martin, L.-A.; Andrews, S.; Wingett, S.W.; et al. Capture Hi-C Identifies Putative Target Genes at 33 Breast Cancer Risk Loci. *Nat. Commun.* **2018**, *9*, 1028. [CrossRef]
11. Wu, L.; Shi, W.; Long, J.; Guo, X.; Michailidou, K.; Beesley, J.; Bolla, M.K.; Shu, X.-O.; Lu, Y.; Cai, Q.; et al. A Transcriptome-Wide Association Study of 229,000 Women Identifies New Candidate Susceptibility Genes for Breast Cancer. *Nat. Genet.* **2018**, *50*, 968–978. [CrossRef]
12. Takenaka, T. Classical vs Reverse Pharmacology in Drug Discovery. *BJU Int.* **2008**, *88*, 7–10. [CrossRef]
13. Ezzat, A.; Wu, M.; Li, X.-L.; Kwoh, C.-K. Computational Prediction of Drug–Target Interactions Using Chemogenomic Approaches: An Empirical Survey. *Brief. Bioinform.* **2019**, *20*, 1337–1357. [CrossRef]
14. Mitchell, J.B. The Relationship between the Sequence Identities of Alpha Helical Proteins in the PDB and the Molecular Similarities of Their Ligands. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1617–1622. [CrossRef]
15. Keiser, M.J.; Roth, B.L.; Armbruster, B.N.; Ernsberger, P.; Irwin, J.J.; Shoichet, B.K. Relating Protein Pharmacology by Ligand Chemistry. *Nat. Biotechnol.* **2007**, *25*, 197–206. [CrossRef]
16. Mahmud, S.M.H.; Chen, W.; Meng, H.; Jahan, H.; Liu, Y.; Hasan, S.M.M. Prediction of Drug-Target Interaction Based on Protein Features Using Undersampling and Feature Selection Techniques with Boosting. *Anal. Biochem.* **2020**, *589*, 113507. [CrossRef]
17. Sachdev, K.; Gupta, M.K. A Comprehensive Review of Feature Based Methods for Drug Target Interaction Prediction. *J. Biomed. Inform.* **2019**, *93*, 103159. [CrossRef]

18. Xie, L.; Evangelidis, T.; Xie, L.; Bourne, P.E. Drug Discovery Using Chemical Systems Biology: Weak Inhibition of Multiple Kinases May Contribute to the Anti-Cancer Effect of Nelfinavir. *PLoS Comput. Biol.* **2011**, *7*, e1002037. [CrossRef]
19. Mousavian, Z.; Masoudi-Nejad, A. Drug-Target Interaction Prediction via Chemogenomic Space: Learning-Based Methods. *Expert Opin. Drug Metab. Toxicol.* **2014**, *10*, 1273–1287. [CrossRef]
20. Rayhan, F.; Ahmed, S.; Shatabda, S.; Farid, D.M.; Mousavian, Z.; Dehzangi, A.; Rahman, M.S. IDTI-ESBoost: Identification of Drug Target Interaction Using Evolutionary and Structural Features with Boosting. *Sci. Rep.* **2017**, *7*, 17731. [CrossRef]
21. Mousavian, Z.; Khakabimamaghani, S.; Kavousi, K.; Masoudi-Nejad, A. Drug–Target Interaction Prediction from PSSM Based Evolutionary Information. *J. Pharmacol. Toxicol. Methods* **2016**, *78*, 42–51. [CrossRef]
22. Yu, B.; Li, S.; Qiu, W.; Wang, M.; Du, J.; Zhang, Y.; Chen, X. Prediction of Subcellular Location of Apoptosis Proteins by Incorporating PsePSSM and DCCA Coefficient Based on LFDA Dimensionality Reduction. *BMC Genom.* **2018**, *19*, 478. [CrossRef]
23. Shi, H.; Liu, S.; Chen, J.; Li, X.; Ma, Q.; Yu, B. Predicting Drug-Target Interactions Using Lasso with Random Forest Based on Evolutionary Information and Chemical Structure. *Genomics* **2019**, *111*, 1839–1852. [CrossRef]
24. Sun, Y.; Wong, A.K.C.; Kamel, M.S. Classification of Imbalanced Data: A Review. *Int. J. Pattern Recognit. Artif. Intell.* **2009**, *23*, 687–719. [CrossRef]
25. Yamanishi, Y.; Kotera, M.; Kanehisa, M.; Goto, S. Drug-Target Interaction Prediction from Chemical, Genomic and Pharmacological Data in an Integrated Framework. *Bioinformatics* **2010**, *26*, i246–i254. [CrossRef]
26. Chen, H.; Zhang, Z. A Semi-Supervised Method for Drug-Target Interaction Prediction with Consistency in Networks. *PLoS ONE* **2013**, *8*, e62975. [CrossRef]
27. Huang, Y.-A.; You, Z.-H.; Chen, X. A Systematic Prediction of Drug-Target Interactions Using Molecular Fingerprints and Protein Sequences. *Curr. Protein Pept. Sci.* **2018**, *19*, 468–478. [CrossRef]
28. Li, Z.; Han, P.; You, Z.-H.; Li, X.; Zhang, Y.; Yu, H.; Nie, R.; Chen, X. In Silico Prediction of Drug-Target Interaction Networks Based on Drug Chemical Structure and Protein Sequences. *Sci. Rep.* **2017**, *7*, 11174. [CrossRef]
29. Gönen, M. Predicting Drug-Target Interactions from Chemical and Genomic Kernels Using Bayesian Matrix Factorization. *Bioinformatics* **2012**, *28*, 2304–2310. [CrossRef]
30. Liu, Y.; Wu, M.; Miao, C.; Zhao, P.; Li, X.-L. Neighborhood Regularized Logistic Matrix Factorization for Drug-Target Interaction Prediction. *PLoS Comput. Biol.* **2016**, *12*, e1004760. [CrossRef]
31. Yamanishi, Y.; Araki, M.; Gutteridge, A.; Honda, W.; Kanehisa, M. Prediction of Drug–Target Interaction Networks from the Integration of Chemical and Genomic Spaces. *Bioinformatics* **2008**, *24*, i232–i240. [CrossRef]
32. Kanehisa, M. From Genomics to Chemical Genomics: New Developments in KEGG. *Nucleic Acids Res.* **2006**, *34*, D354–D357. [CrossRef] [PubMed]
33. Schomburg, I.; Chang, A.; Ebeling, C.; Gremse, M.; Heldt, C.; Huhn, G.; Schomburg, D. BRENDA, the Enzyme Database: Updates and Major New Developments. *Nucleic Acids Res.* **2004**, *32*, D431–D433. [CrossRef] [PubMed]
34. Günther, S.; Kuhn, M.; Dunkel, M.; Campillos, M.; Senger, C.; Petsalaki, E.; Ahmed, J.; Urdiales, E.G.; Gewiess, A.; Jensen, L.J.; et al. SuperTarget and Matador: Resources for Exploring Drug-Target Relationships. *Nucleic Acids Res.* **2008**, *36*, D919–D922. [CrossRef]
35. Wishart, D.S.; Knox, C.; Guo, A.C.; Cheng, D.; Shrivastava, S.; Tzur, D.; Gautam, B.; Hassanali, M. DrugBank: A Knowledgebase for Drugs, Drug Actions and Drug Targets. *Nucleic Acids Res.* **2008**, *36*, D901–D906. [CrossRef] [PubMed]
36. Ba-alawi, W.; Soufan, O.; Essack, M.; Kalnis, P.; Bajic, V.B. DASPfind: New Efficient Method to Predict Drug–Target Interactions. *J. Cheminform.* **2016**, *8*, 15. [CrossRef]
37. Chou, K.-C. Prediction of Protein Cellular Attributes Using Pseudo-Amino Acid Composition. *Proteins Struct. Funct. Genet.* **2001**, *43*, 246–255. [CrossRef] [PubMed]
38. Jones, D.T. Protein Secondary Structure Prediction Based on Position-Specific Scoring Matrices. *J. Mol. Biol.* **1999**, *292*, 195–202. [CrossRef]
39. Altschul, S. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef]
40. Podobnik, B.; Stanley, H.E. Detrended Cross-Correlation Analysis: A New Method for Analyzing Two Nonstationary Time Series. *Phys. Rev. Lett.* **2008**, *100*, 084102. [CrossRef]
41. Zebende, G.F. DCCA Cross-Correlation Coefficient: Quantifying Level of Cross-Correlation. *Phys. Stat. Mech. Its Appl.* **2011**, *390*, 614–618. [CrossRef]
42. Liang, Y.; Liu, S.; Zhang, S. Geary Autocorrelation and DCCA Coefficient: Application to Predict Apoptosis Protein Subcellular Localization via PSSM. *Phys. Stat. Mech. Its Appl.* **2017**, *467*, 296–306. [CrossRef]
43. Podobnik, B.; Jiang, Z.-Q.; Zhou, W.-X.; Stanley, H.E. Statistical Tests for Power-Law Cross-Correlated Processes. *Phys. Rev. E* **2011**, *84*, 066118. [CrossRef]
44. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B Methodol.* **1996**, *58*, 267–288. [CrossRef]
45. Friedman, J.; Hastie, T.; Tibshirani, R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J. Stat. Softw.* **2010**, *33*, 1–22. [CrossRef]
46. Tibshirani, R. Regression Shrinkage and Selection via the Lasso: A Retrospective. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2011**, *73*, 273–282. [CrossRef]

47. Shrivastava, S.; Jeyanthi, P.M.; Singh, S. Failure Prediction of Indian Banks Using SMOTE, Lasso Regression, Bagging and Boosting. *Cogent Econ. Financ.* **2020**, *8*, 1729569. [CrossRef]
48. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
49. Raghuwanshi, B.S.; Shukla, S. SMOTE Based Class-Specific Extreme Learning Machine for Imbalanced Learning. *Knowl.-Based Syst.* **2020**, *187*, 104814. [CrossRef]
50. Breiman, L. Random Forests. *Mach. Learn* **2001**, *45*, 5–32. [CrossRef]
51. Shar, P.A.; Tao, W.; Gao, S.; Huang, C.; Li, B.; Zhang, W.; Shahen, M.; Zheng, C.; Bai, Y.; Wang, Y. Pred-Binding: Large-Scale Protein–Ligand Binding Affinity Prediction. *J. Enzyme Inhib. Med. Chem.* **2016**, *31*, 1443–1450. [CrossRef]
52. Cao, D.-S.; Zhang, L.-X.; Tan, G.-S.; Xiang, Z.; Zeng, W.-B.; Xu, Q.-S.; Chen, A.F. Computational Prediction of Drug Target Interactions Using Chemical, Biological, and Network Features. *Mol. Inform.* **2014**, *33*, 669–681. [CrossRef]
53. Gornitsky, M.; Velly, A.M.; Mohit, S.; Almajed, M.; Su, H.; Panasci, L.; Schipper, H.M. Altered Levels of Salivary 8-Oxo-7-Hydrodeoxyguanosine in Breast Cancer. *JDR Clin. Transl. Res.* **2016**, *1*, 171–177. [CrossRef] [PubMed]
54. Jeschke, U. Can We Find Breast Cancer via Salivary Fluid Glycosylation Analyses? *EBioMedicine* **2018**, *28*, 4. [CrossRef]
55. Zeitzer, J.M.; Nouriani, B.; Neri, E.; Spiegel, D. Correspondence of Plasma and Salivary Cortisol Patterns in Women with Breast Cancer. *Neuroendocrinology* **2014**, *100*, 153–161. [CrossRef]
56. Bretschneider, N.; Brand, H.; Miller, N.; Lowery, A.J.; Kerin, M.J.; Gannon, F.; Denger, S. Estrogen Induces Repression of the Breast Cancer and Salivary Gland Expression Gene in an Estrogen Receptor Dependent Manner. *Cancer Res.* **2008**, *68*, 106–114. [CrossRef] [PubMed]
57. In der Maur, C.D.; Klokman, W.J.; van Leeuwen, F.E.; Tan, I.B.; Rutgers, E.J.T.; Balm, A.J.M. Increased Risk of Breast Cancer Development after Diagnosis of Salivary Gland Tumour. *Eur. J. Cancer* **2005**, *41*, 1311–1315. [CrossRef]
58. Sawczuk, B.; Maciejczyk, M.; Sawczuk-Siemieniuk, M.; Posmyk, R.; Zalewska, A.; Car, H. Salivary Gland Function, Antioxidant Defence and Oxidative Damage in the Saliva of Patients with Breast Cancer: Does the BRCA1 Mutation Disturb the Salivary Redox Profile? *Cancers* **2019**, *11*, 1501. [CrossRef] [PubMed]
59. Liu, X.; Yu, H.; Qiao, Y.; Yang, J.; Shu, J.; Zhang, J.; Zhang, Z.; He, J.; Li, Z. Salivary Glycopatterns as Potential Biomarkers for Screening of Early-Stage Breast Cancer. *EBioMedicine* **2018**, *28*, 70–79. [CrossRef]
60. Streckfus, C.F.; Arreola, D.; Edwards, C.; Bigler, L. Salivary Protein Profiles among HER2/Neu-Receptor-Positive and -Negative Breast Cancer Patients: Support for Using Salivary Protein Profiles for Modeling Breast Cancer Progression. *J. Oncol.* **2012**, *2012*, 1–9. [CrossRef]
61. Chang, P.-H.; Hwang-Verslues, W.W.; Chang, Y.-C.; Chen, C.-C.; Hsiao, M.; Jeng, Y.-M.; Chang, K.-J.; Lee, E.Y.-H.P.; Shew, J.-Y.; Lee, W.-H. Activation of Robo1 Signaling of Breast Cancer Cells by Slit2 from Stromal Fibroblast Restrains Tumorigenesis via Blocking PI3K/Akt/ -Catenin Pathway. *Cancer Res.* **2012**, *72*, 4652–4661. [CrossRef]
62. Smit, L.; Berns, K.; Spence, K.; Ryder, W.D.; Zeps, N.; Madiredjo, M.; Beijersbergen, R.; Bernards, R.; Clarke, R.B. An Integrated Genomic Approach Identifies That the PI3K/AKT/FOXO Pathway Is Involved in Breast Cancer Tumor Initiation. *Oncotarget* **2016**, *7*, 2596–2610. [CrossRef]
63. Tao, J.J.; Castel, P.; Radosevic-Robin, N.; Elkabets, M.; Auricchio, N.; Aceto, N.; Weitsman, G.; Barber, P.; Vojnovic, B.; Ellis, H.; et al. Antagonism of EGFR and HER3 Enhances the Response to Inhibitors of the PI3K-Akt Pathway in Triple-Negative Breast Cancer. *Sci. Signal.* **2014**, *7*, ra29. [CrossRef]
64. Gonzalez-Angulo, A.M.; Blumenschein, G.R. Defining Biomarkers to Predict Sensitivity to PI3K/Akt/MTOR Pathway Inhibitors in Breast Cancer. *Cancer Treat. Rev.* **2013**, *39*, 313–320. [CrossRef] [PubMed]
65. Pierobon, M.; Ramos, C.; Wong, S.; Hodge, K.A.; Aldrich, J.; Byron, S.; Anthony, S.P.; Robert, N.J.; Northfelt, D.W.; Jahanzeb, M.; et al. Enrichment of PI3K-AKT–MTOR Pathway Activation in Hepatic Metastases from Breast Cancer. *Clin. Cancer Res.* **2017**, *23*, 4919–4928. [CrossRef] [PubMed]
66. Chandarlapaty, S.; Sakr, R.A.; Giri, D.; Patil, S.; Heguy, A.; Morrow, M.; Modi, S.; Norton, L.; Rosen, N.; Hudis, C.; et al. Frequent Mutational Activation of the PI3K-AKT Pathway in Trastuzumab-Resistant Breast Cancer. *Clin. Cancer Res.* **2012**, *18*, 6784–6791. [CrossRef] [PubMed]
67. Gris-Oliver, A.; Palafox, M.; Monserrat, L.; Brasó-Maristany, F.; Òdena, A.; Sánchez-Guixé, M.; Ibrahim, Y.H.; Villacampa, G.; Grueso, J.; Parés, M.; et al. Genetic Alterations in the PI3K/AKT Pathway and Baseline AKT Activity Define AKT Inhibitor Sensitivity in Breast Cancer Patient-Derived Xenografts. *Clin. Cancer Res.* **2020**, *26*, 3720–3731. [CrossRef] [PubMed]
68. Ramaswamy, B.; Lu, Y.; Teng, K.-Y.; Nuovo, G.; Li, X.; Shapiro, C.L.; Majumder, S. Hedgehog Signaling Is a Novel Therapeutic Target in Tamoxifen-Resistant Breast Cancer Aberrantly Activated by PI3K/AKT Pathway. *Cancer Res.* **2012**, *72*, 5048–5059. [CrossRef]
69. Le Rhun, E.; Bertrand, N.; Dumont, A.; Tresch, E.; Le Deley, M.-C.; Mailliez, A.; Preusser, M.; Weller, M.; Revillion, F.; Bonneterre, J. Identification of Single Nucleotide Polymorphisms of the PI3K-AKT-MTOR Pathway as a Risk Factor of Central Nervous System Metastasis in Metastatic Breast Cancer. *Eur. J. Cancer* **2017**, *87*, 189–198. [CrossRef]
70. Yi, Y.W.; Hong, W.; Kang, H.J.; Kim, H.J.; Zhao, W.; Wang, A.; Seong, Y.-S.; Bae, I. Inhibition of the PI3K/AKT Pathway Potentiates Cytotoxicity of EGFR Kinase Inhibitors in Triple-Negative Breast Cancer Cells. *J. Cell. Mol. Med.* **2013**, *17*, 648–656. [CrossRef]
71. Yang, S.X.; Polley, E.; Lipkowitz, S. New Insights on PI3K/AKT Pathway Alterations and Clinical Outcomes in Breast Cancer. *Cancer Treat. Rev.* **2016**, *45*, 87–96. [CrossRef] [PubMed]

72. Cavazzoni, A.; Bonelli, M.A.; Fumarola, C.; La Monica, S.; Airoud, K.; Bertoni, R.; Alfieri, R.R.; Galetti, M.; Tramonti, S.; Galvani, E.; et al. Overcoming Acquired Resistance to Letrozole by Targeting the PI3K/AKT/MTOR Pathway in Breast Cancer Cell Clones. *Cancer Lett.* **2012**, *323*, 77–87. [CrossRef] [PubMed]
73. Riggio, M.; Polo, M.L.; Blaustein, M.; Colman-Lerner, A.; Lüthy, I.; Lanari, C.; Novaro, V. PI3K/AKT Pathway Regulates Phosphorylation of Steroid Receptors, Hormone Independence and Tumor Differentiation in Breast Cancer. *Carcinogenesis* **2012**, *33*, 509–518. [CrossRef]
74. Khan, M.A.; Jain, V.K.; Rizwanullah, M.; Ahmad, J.; Jain, K. PI3K/AKT/MTOR Pathway Inhibitors in Triple-Negative Breast Cancer: A Review on Drug Discovery and Future Challenges. *Drug Discov. Today* **2019**, *24*, 2181–2191. [CrossRef] [PubMed]
75. Sharma, V.; Sharma, A.K.; Punj, V.; Priya, P. Recent Nanotechnological Interventions Targeting PI3K/Akt/MTOR Pathway: A Focus on Breast Cancer. *Semin. Cancer Biol.* **2019**, *59*, 133–146. [CrossRef] [PubMed]
76. Delaloge, S.; DeForceville, L. Targeting PI3K/AKT Pathway in Triple-Negative Breast Cancer. *Lancet Oncol.* **2017**, *18*, 1293–1294. [CrossRef]
77. Basho, R.K.; Gilcrease, M.; Murthy, R.K.; Helgason, T.; Karp, D.D.; Meric-Bernstam, F.; Hess, K.R.; Herbrich, S.M.; Valero, V.; Albarracin, C.; et al. Targeting the PI3K/AKT/MTOR Pathway for the Treatment of Mesenchymal Triple-Negative Breast Cancer: Evidence From a Phase 1 Trial of MTOR Inhibition in Combination With Liposomal Doxorubicin and Bevacizumab. *JAMA Oncol.* **2017**, *3*, 509. [CrossRef]
78. Ciruelos Gil, E.M. Targeting the PI3K/AKT/MTOR Pathway in Estrogen Receptor-Positive Breast Cancer. *Cancer Treat. Rev.* **2014**, *40*, 862–871. [CrossRef]
79. Costa, R.L.B.; Han, H.S.; Gradishar, W.J. Targeting the PI3K/AKT/MTOR Pathway in Triple-Negative Breast Cancer: A Review. *Breast Cancer Res. Treat.* **2018**, *169*, 397–406. [CrossRef]
80. Fengjiao, J.; Zhaozhen, W.; Xiao, H.; Jiahui, Z.; Zihe, G.; Xiao, H.; Junfang, Q.; Chen, L.; Yue, W. The PI3K/Akt/GSK-3β/ROS/EIF2B Pathway Promotes Breast Cancer Growth and Metastasis via Suppression of NK Cell Cytotoxicity and Tumor Cell Susceptibility. *Cancer Biol. Med.* **2019**, *16*, 38. [CrossRef]
81. Paplomata, E.; O'Regan, R. The PI3K/AKT/MTOR Pathway in Breast Cancer: Targets, Trials and Biomarkers. *Ther. Adv. Med. Oncol.* **2014**, *6*, 154–166. [CrossRef] [PubMed]
82. Massihnia, D.; Galvano, A.; Fanale, D.; Perez, A.; Castiglia, M.; Incorvaia, L.; Listì, A.; Rizzo, S.; Cicero, G.; Bazan, V.; et al. Triple Negative Breast Cancer: Shedding Light onto the Role of Pi3k/Akt/Mtor Pathway. *Oncotarget* **2016**, *7*, 60712–60722. [CrossRef] [PubMed]
83. Woo, S.-U.; Sangai, T.; Akcakanat, A.; Chen, H.; Wei, C.; Meric-Bernstam, F. Vertical Inhibition of the PI3K/Akt/MTOR Pathway Is Synergistic in Breast Cancer. *Oncogenesis* **2017**, *6*, e385. [CrossRef] [PubMed]
84. Wen, M.; Zhang, Z.; Niu, S.; Sha, H.; Yang, R.; Yun, Y.; Lu, H. Deep-Learning-Based Drug–Target Interaction Prediction. *J. Proteome Res.* **2017**, *16*, 1401–1409. [CrossRef]
85. Ding, R.; Shi, J.; Pabon, K.; Scotto, K.W. Xanthines Down-Regulate the Drug Transporter ABCG2 and Reverse Multidrug Resistance. *Mol. Pharmacol.* **2012**, *81*, 328–337. [CrossRef] [PubMed]