Research Article

www.molinf.com

molecular informatics

doi.org/10.1002/minf.202200001

Drug Repurposing for Newly Emerged Diseases via Network-based Inference on a Gene-disease-drug Network

Li Qin,[a] Jiye Wang,[a] Zengrui Wu,*[a] Weihua Li,[a] Guixia Liu,[a] and Yun Tang*[a]

Abstract: Identification of disease-drug associations is an effective strategy for drug repurposing, especially in searching old drugs for newly emerged diseases like COVID-19. In this study, we put forward a network-based method named NEDNBI to predict disease-drug associations based on a gene-disease-drug tripartite network, which could be applied in drug repurposing. The novelty of our method lies in the fact that no negative data are required, and new disease could be added into the disease-drug network with gene as the bridge. The comprehensive evaluation results showed that the proposed method had good performance, with AUC value 0.948 ± 0.009 for 10-fold cross validation. In a case study, 8 of the 20 predicted old drugs have been tested clinically for the treatment of COVID-19, which illustrated the usefulness of our method in drug repurposing. The source code and data of the method are available at https://github.com/Qli97/NEDNBI.

Keywords: Drug repurposing · Disease-drug associations · Network-based inference

1 Introduction

From scratch to develop a successful drug, it usually takes an average of 10–15 years and consumes about 2.6 billion USD.[1] Therefore, it is wise to find new indications for approved drugs, namely drug repurposing, which is especially helpful for orphan diseases or newly emerged diseases.[2]

Since the outbreak of Coronavirus Disease 2019 (COVID-19) in December 2019, there have been hundreds of millions of confirmed cases and even millions of people lost their lives.[3] Although many kinds of vaccines have been developed for prevention, there is a huge demand to find therapeutic agents as soon as possible.[4,5] Among various approaches, computational drug repurposing has shown a great advantage compared to others.[6–11] Remdesvir[12] is an repurposing example that was approved for the treatment of COVID-19 by U.S. FDA on Oct. 22, 2020. Therefore, it is urgent to develop new computational drug repurposing method for such a newly emerged disease.

At present, there are many computational methods developed for drug repurposing, which can be roughly divided into machine learning methods and network-based methods. Many classification models with high accuracy were built by machine learning. For example, Gottlieb *et al.* followed a widely disseminated guilt-by-association hypothesis,[13] which suggested that different classes of drugs could be identified for the same disease. In their research,[14] a large number of drug-drug and disease-disease similarities were calculated and used to predict new indications for drugs or new compounds by logistic regression algorithm. Jiang *et al.*[15] performed sigmoid kernel technology to obtain similar characteristics of drugs and diseases, and Convolutional Neural Network was used to extract the similarity symbols as final feature descriptors. Then, these feature descriptors were used as inputs to train the random forest classifier, which could be used to predict the associations between drugs and diseases. At the same year, Jiang *et al.*[16] reported another machine learning method, named GIPAE, to predict associations between drugs and diseases. It shared the same dataset with the above model, but using different feature processing strategy - autoencoder and Gaussian interaction strategy.

Computational evaluation of molecular similarity is a method to relocate drugs and disease states.[17,18] However, due to the high correlation of high-dimensional features, these methods often lead to model overfitting.[19] Network-based analysis is also a widely used *in silico* approach.[18,20] Previous studies have confirmed that it is a good strategy to achieve the purpose of drug repurposing.

Currently, some network-based methods also employed similarity strategies to realize drug repurposing. They constructed similarity networks by calculating the chemical similarity of drugs and the semantic similarity of diseases, and combined different algorithms to predict new drug-disease associations. For example, Martinez *et al.*[21] performed information propagation on drugs, diseases and

[a] *L. Qin, J. Wang, Z. Wu, W. Li, G. Liu, Y. Tang*
Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism,
School of Pharmacy,
East China University of Science and Technology,
Shanghai 200237, China
E-mail: ytang234@ecust.edu.cn
zengruiwu@ecust.edu.cn

Supporting information for this article is available on the WWW under https://doi.org/10.1002/minf.202200001

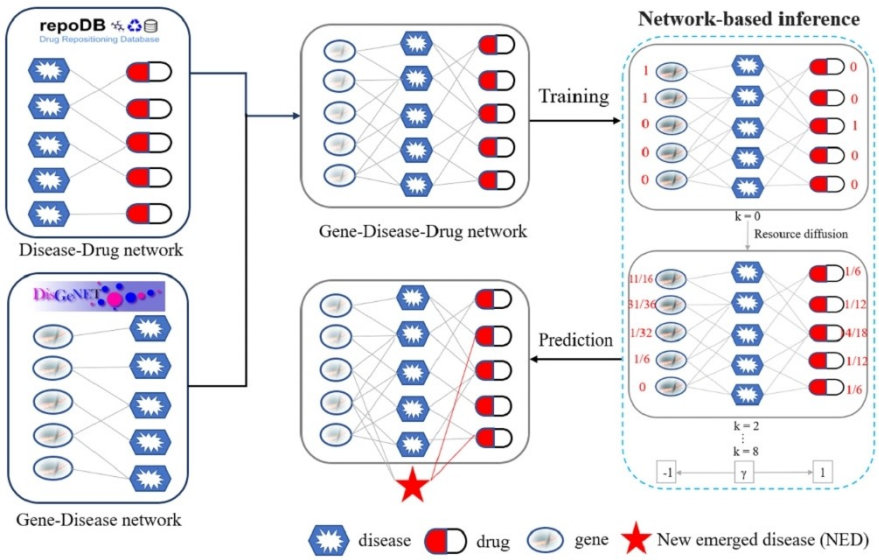

Figure 1. The workflow of computational systems for predicting novel disease-drug associations.

targets with ProphNet strategy to prioritize drug-disease associations, which was called DrugNet. Wang *et al.*[22] proposed a new computational framework named TL-HGBI to integrate three nodes of drugs, diseases and targets, and build a three-layer heterogeneous network for drug repurposing. Luo *et al.*[23] proposed another *in silico* method named MBiRW. In this research, they constructed a drug-disease bipartite network with bi-random walk algorithm and inferred the potential drug-disease associations by the classifier model. Yang *et al.*[24] presented HED method, which combined network embedding strategy to predict potential drug-disease associations on the drug-disease network. Chen *et al.*[25] introduced two inference ways of ProbS and HeatS to predict direct associations between drugs and diseases.

As mentioned earlier, model performance tends to get worsen due to similarity redundancy.[19] It has been proven that the analysis of network topology is superior to the strategy based on similarity in the reliability of model prediction as stated in our previous study.[26] Otherwise, some network models do not add molecular profiles related to diseases, which makes the prediction results less biologically meaningful.[24,25]

In recent years, we have developed a series of network-based methods, including network-based inference (NBI),[26] substructure-drug-target NBI (SDTNBI),[27] and balanced SDTNBI (bSDTNBI),[28] which are widely used in predictions of drug-target interactions,[26,27] drug-microRNA associations,[29] and drug-pathway associations.[30] These methods have two advantages compared with machine learning models, one is that the negative data are not required, the other is that chemical substructures are used as the bridge to link new compounds with known targets.

In this study, we proposed a new model named as NEDNBI (Newly Emerged Disease via Network-Based Inference) for drug repurposing based on our NBI series methods. The NEDNBI model took gene as the bridge to link new disease with known disease-drug network via NBI method on a gene-disease-drug heterogeneous network. As shown in Figure 1, at first, we constructed the known disease-drug network by collecting the associations of approved drug indications from the public database, then searched for the genes associated with these diseases to construct the disease-gene network. After that, we built a heterogeneous network based on the above two networks with three types of nodes: genes, diseases and drugs. The performance of our model was evaluated by 10-fold cross-validation and external validation. Finally, COVID-19 was used as a case study to illustrate the practicability of NEDNBI.

2 Materials and Methods

2.1 Data Collection and Preparation

In order to consider both drugs approved by FDA and those failed in clinical trials, we chose the repoDB[31] as our comprehensive pharmacopeia resource. RepoDB is a standard drug repurposing database which could provide drug-disease associations. There are four categories about the associations. The first category is the approved indication retrievals, which are based on DrugCentral database.[32] The others are failed indication retrievals collected from the Clinical Trials Transformation Initiative website and divided into three categories: withdrawn, suspended, and terminated. In this study, we only extracted

the approved drug indications category served as dataset 1. In the dataset, drugs were labeled as DrugBank[33] identifiers, and Unified Medical Language System[34] (UMLS) Metathesaurus Concept Unique Identifiers were mapped to indication terms.

Then, we searched the largest disease-related gene database DisGeNET (version 7.0, https://www.disgenet.org/)[35] according to disease UMLS identifiers, and selected the curated gene-disease associations to obtain all possible disease-gene associations. After that, the related genes were standardized by Entrez gene identifiers.

Dataset 2 was our external validation dataset derived from a commonly used gold standard dataset.[14] The same as dataset 1, in dataset 2 diseases were matched to the UMLS identifiers, drugs were represented by the DrugBank identifiers, and the relevant genes were collected from the DisGeNET database based on the UMLS identifiers of diseases. Obviously, the disease-gene associations present in dataset 1 were removed from dataset 2.

2.2 Construction of NEDNBI Model

To develop the NEDNBI model, we used the state-of-the-art network-based inference as described previously.[26–28] In this method, three networks were constructed: disease-drug network, disease-gene network, and combined gene-disease-drug network. Briefly, the combined gene-disease-drug network-based inference algorithm follows the hypothesis that diseases with similar genes will be associated with similar drugs. It performs resource diffusion process in the network of known relationships and predicts the potential interactions with the object.

Described mathematically, the disease-drug network was represented as a matrix A_{DC}:

$$A_{DC}(i,j)=\begin{cases}1 & \text{if disease } D_i \text{ links with drug } C_j\\ 0 & \text{otherwise}\end{cases} \quad (1)$$

Where $i\in(0, N_D]$, $j\in(0, N_C]$ are positive integers.

The disease-gene linkages were also represented as a matrix A_{DG}:

$$A_{DG}(i,j)=\begin{cases}1 & \text{if disease } D_i \text{ links with gene } G_j\\ 0 & \text{otherwise}\end{cases} \quad (2)$$

Where $i\in(0, N_D]$, $j\in(0, N_G]$ are positive integers.

Then, in order to eliminate the effects of new disease nodes without known drug associations in the process of resource diffusion, we defined two matrices that were similar to A_{DC} and A_{DG}, as mentioned above:

$$B_{DC}(i,j)=\begin{cases}A_{DC}(i,j) & \text{if } \sum_{l=1}^{N_C} A_{DC}(i,l)\neq 0\\ 0 & \text{otherwise}\end{cases} \quad (3)$$

Where $i\in(0, N_D]$, $j\in(0, N_C]$ are positive integers.

$$B_{DG}(i,j)=\begin{cases}A_{DG}(i,j) & \text{if } \sum_{l=1}^{N_C} A_{DC}(i,l)\neq 0\\ 0 & \text{otherwise}\end{cases} \quad (4)$$

Where $i\in(0, N_D]$, $j\in(0, N_G]$ are positive integers.

Based on these defined matrices, NEDNBI predicts potential disease-drug associations by performing resource diffusion in the gene-disease-drug network. Here, we import three adjustable parameters $\alpha\in[0, 1)$, $\beta\in[0, 1)$, and $\gamma\in[-1, 1]$ into the model to adjust the performance.

Firstly, two matrices were defined, the parameter α was used to adjust the initial resources allocation of different node types. The formula was adjusted as follows:

$$A'_{DC}(i,j)=\frac{\alpha\cdot A_{DC}(i,j)}{\sum_{l=1}^{N_C} A_{DC}(i,l)} \quad (5)$$

Where $i\in(0, N_D]$, $j\in(0, N_C]$ are positive integers.

$$A'_{DG}(i,j)=\frac{(1-\alpha)\cdot A_{DG}(i,j)}{\sum_{l=1}^{N_G} A_{DG}(i,l)} \quad (6)$$

Where $i\in(0, N_D]$, $j\in(0, N_G]$ are positive integers.

According to the newly defined matrices, the initial resource matrix (A) was represented as:

$$A=\begin{bmatrix}O & A'_{DG} & A'_{DC}\\ (A'_{DG})^T & O & O\\ (A'_{DC})^T & O & O\end{bmatrix} \quad (7)$$

After that, the transfer matrix after one-step resource diffusion was also calculated, where the parameters β and γ were used to adjust the weighted values of different edge types and the influence of hub nodes, respectively. The formula was adjusted as follows:

$$B=\begin{bmatrix}O & \beta\cdot B_{DG} & (1-\beta)\cdot B_{DC}\\ \beta\cdot (B_{DG})^T & O & O\\ (1-\beta)\cdot (B_{DC})^T & O & O\end{bmatrix} \quad (8)$$

$$C(i,j)=B(i,j)\cdot\Big[\sum_{l=1}^{N_D+N_G+N_C} B(l,j)\Big]^{\gamma} \quad (9)$$

$$W(i,j) = \begin{cases} \frac{C(i,j)}{\sum_{l=1}^{N_D+N_G+N_C} C(i,l)} & \text{if } C(i,j) \neq 0 \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

Where i, j∈(0, $N_D + N_G + N_C$) are positive integers.

The final resource matrix was calculated by:

$$F = A \times W^k \quad (11)$$

Where W^k is transfer matrix and k is the number of resource-diffusion process. The value of F(i, $N_D + N_G + N_C$) is the predicted score of D_i-C_j interaction.

2.3 Evaluation of NEDNBI Model

2.3.1 Parameter Optimization

In the process of model calculation, resource diffusion times including k = 2, 4, 6, and 8 and parameters (α, β, γ) are considered to adjust the model performance under different conditions. Grid search and 10-fold cross validation were conducted to obtain the optimal parameters. First, let γ = 0, we set the adjustment range of parameters α and β between [0,1), and the search step is 0.1. Then, with the optimal values of α and β, we set hub nodes parameter γ to range from −1 to 1, and obtain the optimal value of parameter γ within the search scope according to a search step of 0.1. At last, by comparing the performance under different conditions, the best model was obtained.

2.3.2 Cross Validation

In our network, the disease-drug associations were randomly divided into ten equal parts. In each training session, one equal portion was taken as the test set, and the rest, including the disease-drug associations and disease-gene associations, were taken as the training set. To eliminate contingency, the process was repeated 100 times.

2.3.3 Calculation of Evaluation Metrics

After 10-fold cross validation, the known disease D_i-drug C_j associations in the test set were used to compare with disease-drug associations predicted in the training set to obtain a set of evaluation metrics. Here, several metrics including precision (P), recall (R), and recall enhancement (e_R) were calculated, which could be used to quantitatively demonstrate the performance of the predictive model. These three indicators had been used in our previous studies.[26–28,36,37] Three metrics were defined as below:

$$P(L) = \frac{1}{M} \cdot \sum_{i=1}^{M} \frac{X_i(L)}{L} \quad (12)$$

$$R(L) = \frac{1}{M} \cdot \sum_{i=1}^{M} \frac{X_i(L)}{X_i} \quad (13)$$

$$e_R(L) = R(L) \cdot \frac{N}{L} \quad (14)$$

Where M and N are the numbers of diseases and drugs participated in calculation, respectively. For each disease, L represents the number of drugs evaluated in each cross-validation result. X_i means the missing disease-drug pairs for disease D_i. $X_i(L)$ means the number of correctly predicted drugs in the top L of the results predicted for disease D_i. According to the L value we set, through the verification of the prediction results of the training set, the disease-drug associations that can be correctly verified in the test set are called true positives. On the contrary, if the associations are unknown, it is considered false positives. The area under the receiver operating characteristic curve (AUC) was computed to show the overall performance of the model. For a model, the higher is the AUC value, the higher the model's ability to predict the correct outcome from all possibilities is. Finally, the evaluation indicators of the model were expressed in the form of mean value and standard deviation (mean ± SD).

2.3.4 External Validation

External validation set (dataset 2) was used to evaluate the generalization ability of the model. Zhang *et al.*[38] constructed a drug-disease prediction model with the similarity constrained matrix factorization method, and developed a user-friendly web server (http://www.bioinfotech.cn/SCMFDD/). We used the web server and NEDNBI model to make predictions for diseases.

In order to use the Zhang's model, firstly, we converted the UMLS identifiers of the diseases in dataset 2 into MeSH descriptors because diseases were represented by MeSH descriptors in Zhang's model. Then, the MeSH descriptors of the diseases were inputted into the public web server SCMFDD one-by-one. After all new disease-drug associations were predicted, we calculated the evaluation metrics, including R and e_R, and used these metrics to measure the generalization performance of the two models.

2.4 Case Study: Drug Repurposing for COVID-19

The NEDNBI model was used to predict the potential effective drugs for COVID-19. COVID-19 related genes were collected from the Comparative Toxicogenomics Database (CTD, https://

ctdbase.org/)[39] with "Direct Evidence" and Therapeutic Target Database (TTD, http://db.idrblab.net/ttd/),[40] we checked the two databases in May 2021, then mapped the collected genes to Entrez gene identifiers uniformly.

Through 10-fold cross validation and external validation set, we obtained the best model, which was then used to predict potentially effective drugs for COVID-19, with a prediction length of 20 ($L = 20$). To verify our prediction, we visited a web-based resource that provides publicly and privately supported clinical studies on widespread diseases and conditions (https://clinicaltrials.gov/). If the predicted drugs have been tested in clinical studies, our results are considered as validated.

3 Results

3.1 Data Collection and Analysis

In the NEDNBI model, gene-disease-drug tripartite network was constructed based on repoDB and DisGeNet databases. The built model was evaluated by an external validation set. The detailed data of disease-drug network and disease-gene network were listed in Table 1.

Table 1. Details of Disease-Drug Networks.

	Dataset 1	Dataset 2
Number of diseases	514	120
Number of drugs	1262	365
Number of genes	5675	1463
Number of DC associations	2981	664
Number of DG associations	15187	1925
Sparsity (%)*	0.46	1.52

*Sparsity (%): the ratio of No. of DC associations and the number of all possible disease-drug associations.

In general, dataset 1 contained 2981 clinically confirmed disease-drug associations among 514 diseases and 1262 drugs. In addition, 15187 disease-gene associations were matched by 514 diseases. Dataset 2 was collected from the disease-drug associations compiled by Gottelib *et al.*[14] A total of 664 disease-drug associations connecting 120 diseases and 365 drugs were included after matching and screening, and 1925 disease-gene relationships were obtained according to the UMLS identifiers of the 120 diseases. UMLS integrates identifiers and relationships from numerous biomedical terms and assigns each identifier to one or more of the 137 semantic types. The semantic type distributions of diseases in the training set and test set were shown in Table 2. From Table 2 it could be seen that the trend of the degree distributions in the two datasets was roughly the same, though the number of disease types differed greatly.

Table 2. Number of Disease Semantic Types in Dataset 1 and Dataset 2.

UMLS Semantic Type Name	Number of Dataset 1	Number of Dataset 2
Disease or Syndrome	337	80
Neoplastic Process	89	17
Mental or Behavioral Dysfunction	26	4
Pathologic Function	23	6
Sign or Symptom	22	5
Finding	7	4
Congenital Abnormality	6	2
Injury or Poisoning	3	0
Acquired Abnormality	1	2
Total number	514	120

The topological properties of the two networks were also analyzed. By observing the distribution of degrees in the two networks shown in Figure 2, the drug degree distributions were significantly different. The largest number of diseases to which a drug was linked in dataset 1 was 51, whereas in dataset 2 it was only 11. On the other hand, although the total number of drugs and diseases in the two networks differed greatly, the trends of degree distributions were similar.

3.2 Description of the NEDNBI Model

Here, we evaluated the performance of the constructed model by correct prediction of the disease-drug associations, mainly including two aspects: (i) performance assessment of our model by 10-fold cross validation and (ii) evaluation of the generalization ability of the model with the external validation set.

3.2.1 Performance of NEDNBI Model

In order to make the model perform better, 10-fold cross validation and grid search methods were combined to optimize the parameters of the model (α, β, γ). Firstly, we selected $k=2$ and assumed $\gamma=0$, the variation trend of average AUC value of the model was observed, when the parameters α and β were searched with the step length of 0.1. As shown in Figure 3A, AUC score reached the maxima when $\alpha=0.1$ and $\beta=0.1$ indicating that the better performance of the model tended to more inclined to allocate initial resources to drug nodes rather than gene nodes. Secondly, the relationship between γ and AUC value of the model in the case of $k=2$, $\alpha=0.1$, and $\beta=0.1$ was revealed with 10-fold cross validation. As shown in Figure 3B, when γ was setting as -0.5, the highest AUC value was 0.948 ± 0.009. This illustrated that the performance of the model

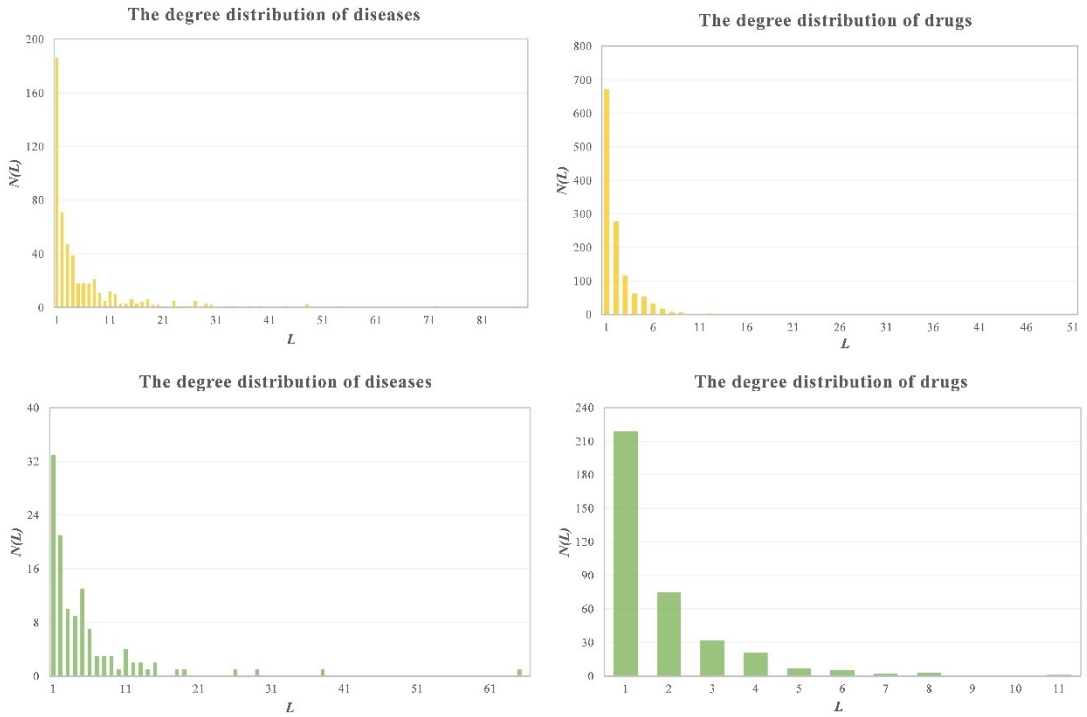

Figure 2. The degree distribution of diseases and drugs for networks constructed from the two datasets. The yellow represents dataset 1 and green represents dataset 2. L is the degree of nodes in the network and N(L) is the number of nodes with the degree of L.

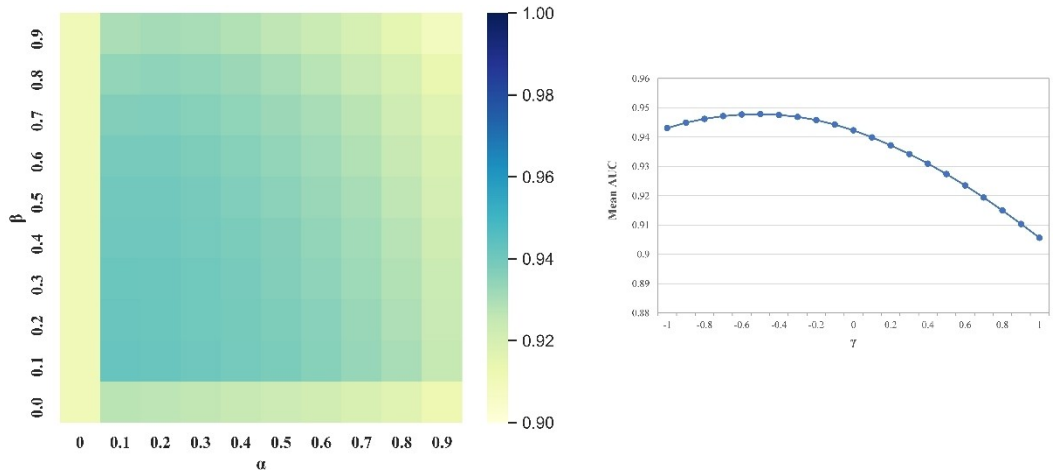

Figure 3. (**A**) The relationship between two parameters α, β and the average AUC value for the model of dataset 1 in 10-fold cross validation. (**B**) Relationships among parameter γ and the average AUC value for the model of dataset 1 in 10-fold cross validation.

could be better when the influence of hub nodes was appropriately weakened.

Following the previous processes, we also compared the influence of different resource diffusion times (k = 2, 4, 6, 8) on the model performance (Figure 4). The results showed that the model's performance increased slightly with the increase of resource diffusion times. Other evaluation indicators (P and R) of the model also showed the same trend (Table S1). From the table we could see that when k = 6 and k = 8, the model performed equally and optimally. However, in consideration of saving computing resources, we chose k = 6. Meanwhile, from Table S1 it could be seen that the values of P and R varied differently with the different prediction length (L = 5, 10, 15, 20). With the optimal value of k (k = 6), P value decreased while R value increased as the increase of L. Finally, the model obtained the best performance on dataset 1 where the parameter conditions were k = 6, α = 0.1, β = 0.1, and γ = −0.5.

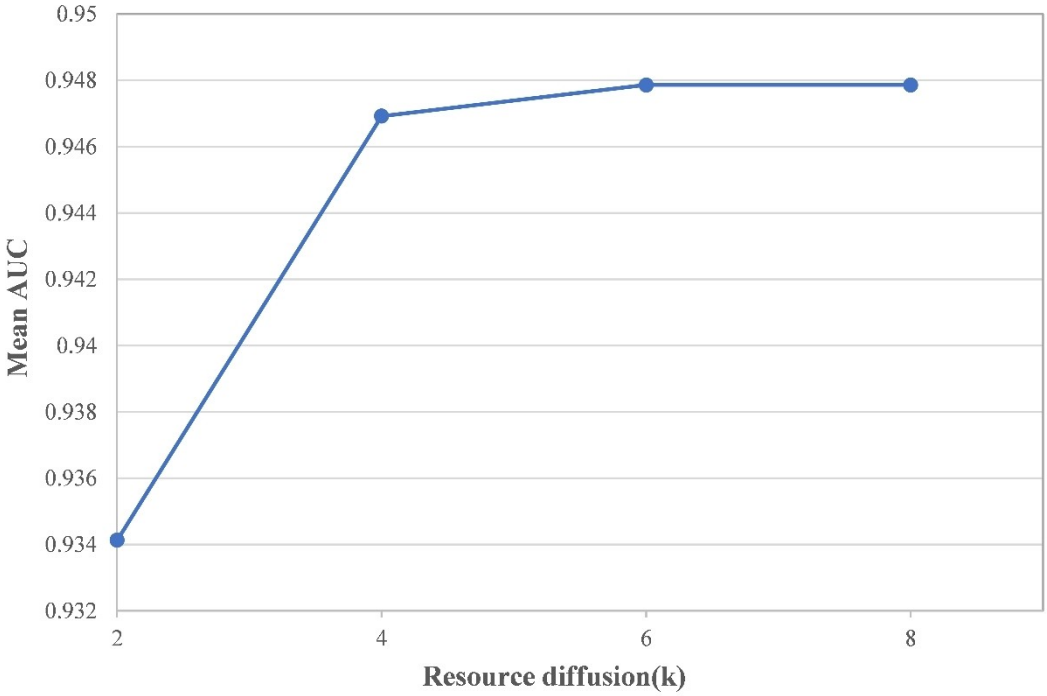

Figure 4. Relationships among parameter k and the average AUC value for the model of dataset1 in 10-fold cross validation.

3.2.2 Evaluation of Model Generalization Ability

Based on the external dataset 2, we also investigated the generalization ability of our model. The whole number of the diseases in external dataset was 120. We set the resource diffusion times and three adjustable paraments to their best value, that is, $k=6$, $\alpha=0.1$, $\beta=0.1$, and $\gamma=-0.5$. On the basis of the best performance of model, we tested it on another data set to evaluate the success rate of recalling the correct disease-drug associations.

In addition, we compared the NEDNBI model with a state-of-art model reported by Zhang *et al.*[38] Only 79 of the 120 diseases were successfully predicted with the MeSH descriptors. The results of measurement indicators of the two models were given in Table 3, including R value and e_R

Table 3. Performance of dataset 2 on our model when $k=6$, $L=20$.

Model	R	e_R
NEDNBI	0.204	12.870
Zhang *et al.*[38]	0.133	1.862

value. From Table 3, it is east to see that the values of R and e_R calculated from the top 20 prediction results ($L=20$), the generalization ability of our model was better than Zhang's model.

3.3 Case Study for COVID-19

To further evaluate the predictive capability of our NEDNBI model, a case study was performed on COVID-19. At first, a total of 112 genes related to COVID-19 were collected from CTD and TTD databases. There were 27 genes from CTD with "Direct Evidence" and 105 genes from TTD. After duplicates were removed, the 112 genes were uniformly matched into the Entrez gene identifiers, the potential disease-drug associations were then predicted by the organized disease-gene network on our best model. In order to measure if our predictions were consistent with the experimental knowledge, we checked the extent of results which are present at current clinical trials in the advanced clinical trial report website (http://clinicaltrials.gov). For each disease, the associated drugs were ranked based on the predictive score, and here we collected the top 20 drugs as the predictive results. A predicted disease-drug association is considered valid if it has ever been used in clinical trials. The predicted results for COVID-19 were listed in Table 4.

From Table 4, we could see that 8 of the 20 potential therapeutic agents we predicted for COVID-19 have been tested clinically, and most of them were tested more than once. The relevant trial IDs of one drug are not completely listed.

Methylprednisolone is a prednisolone derivative glucocorticoid with higher potency than prednisone, it was used to treat inflammation or immune reactions across a variety of organ systems, endocrine conditions, and neoplastic

Table 4. Clinical trials of the top 20 drugs predicted for COVID-19.

DrugBank ID	Generic Name	Identity*
DB00787	Acyclovir	
DB00103	Agalsidase beta	
DB09031	Miltefosine	
DB09357	Dexpanthenol	
DB00860	Prednisolone	NCT03708718
DB00620	Triamcinolone	
DB01270	Ranibizumab	
DB00085	Pancrelipase	
DB00867	Ritodrine	
DB00959	Methylprednisolone	NCT04345445; NCT04673162; NCT04603729; NCT04485429; NCT04355247; NCT04909918; NCT04329650;
DB01234	Dexamethasone	NCT04707534; NCT04513184; NCT04640168; NCT04909918; NCT05004753; NCT04832880; NCT04834375
DB01380	Cortisone acetate	
DB00009	Alteplase	NCT04357730; NCT04640194
DB00443	Betamethasone	NCT04569825
DB00741	Hydrocortisone	NCT04359511; NCT04348305
DB00242	Cladribine	
DB00552	Pentostatin	
DB00635	Prednisone	NCT04551781; NCT04795583; NCT04451174; NCT04359511; NCT04492358
DB08877	Ruxolitinib	NCT04355793; NCT04348071; NCT04414098; NCT04366232; NCT04377620; NCT04354714; NCT04338958; NCT04424056; NCT04403243
DB01119	Diazoxide	

*Identity: the fact that the drug has been registered in Clinicaltrials.gov for investigational treatment of COVID-19, and the NCT number is the Trial ID.

diseases. Dexamethasone, is also a corticosteroid fluorinated at position 9 used to treat endocrine, rheumatic, collagen, and other conditions. Besides, it is structurally similar to other corticosteroids like hydrocortisone and prednisolone, which are also used in clinical trials. Then, we found that Methylprednisolone and Dexamethasone are currently one of 65 experimental unapproved treatments for COVID-19 after we checked the DrugBank COVID-19 Dashboard. Betamethasone is a long-acting corticosteroid with immunosuppressive and anti-inflammatory properties, and the study on its efficacy against COVID-19 began on August 1, 2020. Ruxolitinib is an anticancer drug and Janus kinase (JAK) inhibitor. It has been investigated to treat patients with COVID-19 accompanied by severe systemic hyperinflammation. Alteplase is a biotech drug used in the emergency treatment of myocardial infarction, ischemic stroke, and pulmonary emboli. There was a study to investigate whether different doses of Alteplase help people with severe breathing problems because of COVID-19. Meanwhile, Triamcinolone and Cortisone acetate are also glucocorticoids predicted by our model. Although they have not been clinically studied yet, there is a reason to believe that these two drugs could be considered in studies to look at whether older drugs have therapeutic effect on COVID-19.

4 Discussion

In this study, we proposed a new model named NEDNBI to predict disease-drug associations, in which a gene-disease-drug tripartite network was constructed by combining NBI algorithm and gene information together. The NEDNBI model can predict new associations by performing resource diffusion with known relationships between nodes in the network.

Through systematic evaluation, our prediction model was verified by comprehensive evaluation including 10-fold cross validation and external validation. Our model performs best under the following conditions: $k=6$, $\alpha=0.1$, $\beta=0.1$, and $\gamma=-0.5$, whose generalization ability has been verified by dataset 2. After that, the best model was used to infer drugs that might work for COVID-19, the most prevalent disease today. In particular, 8 of the 20 drugs we predicted have been studied clinically for COVID-19. Among them, Methylprednisolone and Dexamethasone are currently two of the 65 experimental treatments for COVID-19. Therefore, the NEDNBI model not only shows high accuracy of prediction, but also can serve as a guidance for clinical trials.

NEDNBI takes genes as the bridge to link diseases and drugs. Compared with other models without gene information,[23–25] our predicted disease-drug associations would be more instructive and might be helpful for discovery of other therapeutics from old drugs targeting specific diseases.[19] Moreover, due to gene linkage, NEDNBI can easily infer possible associations of drugs for diseases outside the network. Case study in the paper proves that our model could predict some effective results for the newly emerged disease.

Recently, there are many studies on prediction of disease-drug associations, which is also known as drug repurposing. In most studies, the goal is to improve the performance of predictive model by combining multiple characteristic information about diseases or drugs.[41–44] However, it might be easy to lead to overfitting if some highly correlated features are not removed in model construction, which was also mentioned in previous studies.[19,26] From another point of view, machine learning methods usually require a balance of positive data and negative data.[45] However, it is difficult to obtain negative data which mean that a drug has been validated to have no effect on a disease.[46] Some researches tried to find reliable negative data from unlabeled ones, and used them to build a reasonable prediction model.[16,24,46] Nevertheless, the

negative data are not required in our NEDNBI model. We only need positive data (the FDA approved drug indications) to construct the model. Potential disease-drug associations could be predicted by collecting indications for FDA-approved drugs retrieved by investigators from the DrugCentral Database.[32]

It is also worth noting that predictions of disease-drug associations can be made by utilizing a variety of data, including the use of bioinformatics elements associated with both drugs and diseases, such as gene expression data,[47,48] side effects of drug.[49,50] Taking common elements into account, it may make the predicted disease-drug associations more significant clinically and pharmacologically. However, there are many other studies only using common elements of diseases and drugs,[51–54] which might fail to make predictions when there is no known common element between drugs and diseases. Another method for predicting disease-drug associations is taking advantage of direct disease-drug associations and characteristics of drugs and diseases,[55,56] including these we mentioned earlier.[14,24,25,57,58]

Although our proposed model performs well, there are still some space for model improvement. Firstly, the values of the evaluation indicators of NEDNBI model on the external validation set were not satisfactory and need to be improved. Actually, we need collect more data for both training set and validation set, which could help to enhance the generalization ability of the model. Secondly, the type of gene-disease interactions was not considered when constructing the disease-gene network. Taekeon Lee *et al.*[59] divided the interaction type into "positive", "negative", and "neutral" based on the relationship between elements when constructing disease-drug network. Marina Sirota *et al.*[60] made use of gene expression microarray datasets. They believed that a drug may have therapeutic effect on a disease, if in a specific state that the whole genome expression of a certain group is opposite to the change caused by the disease and drug interference. Therefore, various relationships between diseases and genes should be considered in our future research.

5 Conclusions

In this study, we proposed a new model named NEDNBI to predict disease-drug associations. The NEDNBI model integrated network-based inference and gene information, and achieved high average AUC value of 0.948 under the 10-fold cross validation. The performance of the model was further verified in the external validation set. As a case study, we predicted several old drugs with potential efficacy for COVID-19, which is extremely helpful for drug discovery targeting this newly emerged disease. Our model outperformed some previous methods that need to provide gene expression data, drug side effects, multiple properties of disease or drug to predict disease-drug associations. Especially, for a disease outside the disease-drug network, drugs with potential associations can be predicted simply by providing the relevant genes of the disease. The NEDNBI model does not need negative data, which are usually required by traditional machine learning algorithms. It will be a useful tool to predict potential disease-drug associations for newly emerged diseases. The source code and data of the model are available at https://github.com/Qli97/NEDNBI.

Acknowledgements

This work was supported by the National Key Research and Development Program of China (Grant 2019YFA0904800), the National Natural Science Foundation of China (Grants 81872800, 82173746 and 82104066), and Shanghai Frontiers Science Center of Optogenetic Techniques for Cell Metabolism (Shanghai Municipal Education Commission, Grant 2021 Sci & Tech 03–28).

Conflict of Interest

None declared.

Data Availability Statement

The data that support the findings of this study are openly available at https://github.com/Qli97/NEDNBI.

References

[1] H. C. S. Chan, H. Shan, T. Dahoun, H. Vogel, S. Yuan, *Trends Pharmacol. Sci.* **2019**, *40*, 592–604.
[2] D. Sardana, C. Zhu, M. Zhang, R. C. Gudivada, L. Yang, A. G. Jegga, *Briefings Bioinf.* **2011**, *12*, 346–356.
[3] J. Majumder, T. Minko, *AAPS J.* **2021**, *23*, 14.
[4] S. Gocer, C. Turk, S. V. Ozguven, M. Doganay, *North. clin. Istanb.* **2021**, *8*, 529–536.
[5] H. Zou, Y. Yang, H. Dai, Y. Xiong, J. Q. Wang, L. Lin, Z. S. Chen, *Front. Pharmacol.* **2021**, *12*, 732403.
[6] A. Mukherjee, A. Verma, S. Bihani, A. Burli, K. Mantri, S. Srivastava, *Drug Discovery Today Technol.* **2021**, *39*, 1–12.
[7] Y. Zhou, F. Wang, J. Tang, R. Nussinov, F. Cheng, *Lancet Digit Health* **2020**, *2*, e667-e676.
[8] G. Fiscon, P. Paci, *BMC Bioinf.* **2021**, *22*, 150.
[9] D. Morselli Gysi, I. do Valle, M. Zitnik, A. Ameli, X. Gan, O. Varol, S. D. Ghiassian, J. J. Patten, R. A. Davey, J. Loscalzo, A. L. Barabasi, *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2025581118.
[10] G. Fiscon, F. Conte, S. Amadio, C. Volonte, P. Paci, *Neurotherapeutics* **2021**, *18*, 1678–1691.
[11] P. Sibilio, S. Bini, G. Fiscon, M. Sponziello, F. Conte, V. Pecce, C. Durante, P. Paci, R. Falcone, G. D. Norata, L. Farina, A. Verrienti, *Biomed. Pharmacother.* **2021**, *142*, 111954.

[12] https://www.fda.gov/drugs/news-events-human-drugs/fdas-approval-veklury-remdesivir-treatment-covid-19-science-safety-and-effectiveness.
[13] A. P. Chiang, A. J. Butte, *Clin. Pharmacol. Ther.* **2009**, *86*, 507–510.
[14] A. Gottlieb, G. Y. Stein, E. Ruppin, R. Sharan, *Mol. Syst. Biol.* **2011**, *7*, 496.
[15] H. J. Jiang, Z. H. You, Y. A. Huang, *J. Transl. Med.* **2019**, *17*, 382.
[16] H. J. Jiang, Y. A. Huang, Z. H. You, *BioMed Res. Int.* **2019**, *2019*, 2426958.
[17] J. T. Dudley, T. Deshpande, A. J. Butte, *Briefings Bioinf.* **2011**, *12*, 303–311.
[18] J. Li, S. Zheng, B. Chen, A. J. Butte, S. J. Swamidass, Z. Y. Lu, *Briefings Bioinf.* **2016**, *17*, 2–12.
[19] F. Cheng, H. Hong, S. Yang, Y. Wei, *Briefings Bioinf.* **2017**, *18*, 682–697.
[20] M. L. Shahreza, N. Ghadiri, S. R. Mousavi, J. Varshosaz, J. R. Green, *Briefings Bioinf.* **2018**, *19*, 878–892.
[21] V. Martinez, C. Navarro, C. Cano, W. Fajardo, A. Blanco, *Artificial Intelligence in Medicine* **2015**, *63*, 41–49.
[22] W. H. Wang, S. Yang, X. Zhang, J. Li, *Bioinformatics* **2014**, *30*, 2923–2930.
[23] H. Luo, J. Wang, M. Li, J. Luo, X. Peng, F. X. Wu, Y. Pan, *Bioinformatics* **2016**, *32*, 2664–2671.
[24] K. Yang, X. Zhao, D. Waxman, X. M. Zhao, *Chaos* **2019**, *29*, 123109.
[25] H. Chen, H. Zhang, Z. Zhang, Y. Cao, W. Tang, *Comput Math Methods Med* **2015**, *2015*, 130620.
[26] F. Cheng, C. Liu, J. Jiang, W. Lu, W. Li, G. Liu, W. Zhou, J. Huang, Y. Tang, *PLoS Comput. Biol.* **2012**, *8*, e1002503.
[27] Z. R. Wu, F. X. Cheng, J. Li, W. H. Li, G. X. Liu, Y. Tang, *Briefings Bioinf.* **2017**, *18*, 333–347.
[28] Z. R. Wu, W. Q. Lu, D. Wu, A. Q. Luo, H. P. Bian, J. Li, W. H. Li, G. X. Liu, J. Huang, F. X. Cheng, Y. Tang, *Br. J. Pharmacol.* **2016**, *173*, 3372–3385.
[29] J. Li, K. C. Lei, Z. R. Wu, W. H. Li, G. X. Liu, J. W. Liu, F. X. Cheng, Y. Tang, *Oncotarget* **2016**, *7*, 45584–45596.
[30] J. Wang, Z. Wu, Y. Peng, W. Li, G. Liu, Y. Tang, *J. Chem. Inf. Model.* **2021**, *61*, 2475–2485.
[31] A. S. Brown, C. J. Patel, *Sci. Data* **2017**, *4*, 170029.
[32] O. Ursu, J. Holmes, J. Knockel, C. G. Bologa, J. J. Yang, S. L. Mathias, S. J. Nelson, T. I. Oprea, *Nucleic Acids Res.* **2017**, *45*, D932-D939.
[33] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, M. Wilson, *Nucleic Acids Res.* **2018**, *46*, D1074-D1082.
[34] O. Bodenreider, *Nucleic Acids Res.* **2004**, *32*, D267–270.
[35] J. Pinero, J. Sauch, F. Sanz, L. I. Furlong, *Comput. Struct. Biotechnol. J.* **2021**, *19*, 2960–2967.
[36] F. X. Cheng, Y. D. Zhou, W. H. Li, G. X. Liu, Y. Tang, *PLoS One* **2012**, *7*, e41064.
[37] J. Fang, Z. Wu, C. Cai, Q. Wang, Y. Tang, F. Cheng, *J. Chem. Inf. Model.* **2017**, *57*, 2657–2671.
[38] W. Zhang, X. Yue, W. R. Lin, W. J. Wu, R. Q. Liu, F. Huang, F. Liu, *BMC Bioinf.* **2018**, *19*, 233.
[39] C. J. Grondin, A. P. Davis, J. A. Wiegers, T. C. Wiegers, D. Sciaky, R. J. Johnson, C. J. Mattingly, *Current research in toxicology* **2021**, *2*, 272–281.
[40] Y. Wang, S. Zhang, F. Li, Y. Zhou, Y. Zhang, Z. Wang, R. Zhang, J. Zhu, Y. Ren, Y. Tan, C. Qin, Y. Li, X. Li, Y. Chen, F. Zhu, *Nucleic Acids Res.* **2020**, *48*, D1031-D1041.
[41] W. Zhang, Y. Chen, F. Liu, F. Luo, G. Tian, X. Li, *BMC Bioinf.* **2017**, *18*, 18.
[42] W. Zhang, F. Liu, L. Luo, J. Zhang, *BMC Bioinf.* **2015**, *16*, 365.
[43] H. Iwata, R. Sawada, S. Mizutani, Y. Yamanishi, *J. Chem. Inf. Model.* **2015**, *55*, 446–459.
[44] P. Xuan, Y. K. Cao, T. G. Zhang, X. Wang, S. X. Pan, T. H. Shen, *Bioinformatics* **2019**, *35*, 4108–4119.
[45] Z. Wu, W. Li, G. Liu, Y. Tang, *Front. Pharmacol.* **2018**, *9*, 1134.
[46] J. Liu, Z. Zuo, G. Wu, *IEEE Trans. Nanobioscience* **2020**, *19*, 547–555.
[47] H. Cui, M. H. Zhang, Q. M. Yang, X. Y. Li, M. Liebman, Y. Yu, L. Xie, *BioMed Res. Int.* **2018**, *2018*, 4028473.
[48] J. Zhu, J. Wang, X. Wang, M. Gao, B. Guo, M. Gao, J. Liu, Y. Yu, L. Wang, W. Kong, Y. An, Z. Liu, X. Sun, Z. Huang, H. Zhou, N. Zhang, R. Zheng, Z. Xie, *Nat. Biotechnol.* **2021**, *39*, 1444–1452.
[49] L. Yang, P. Agarwal, *PLoS One* **2011**, *6*, e28025.
[50] Y. Wang, S. Chen, N. Deng, Y. Wang, *PLoS One* **2013**, *8*, e78518.
[51] J. von Eichborn, M. S. Murgueitio, M. Dunkel, S. Koerner, P. E. Bourne, R. Preissner, *Nucleic Acids Res.* **2011**, *39*, D1060–1066.
[52] L. Wang, Y. Wang, Q. Hu, S. Li, *CPT Pharmacometrics Syst Pharmacol* **2014**, *3*, e146.
[53] T. C. Wiegers, A. P. Davis, K. B. Cohen, L. Hirschman, C. J. Mattingly, *BMC Bioinf.* **2009**, *10*, 326.
[54] L. Yu, J. Huang, Z. Ma, J. Zhang, Y. Zou, L. Gao, *BMC Med. Genomics* **2015**, *8*, S2.
[55] M. Oh, J. Ahn, Y. Yoon, *PLoS One* **2014**, *9*, e111668.
[56] H. Iwata, R. Sawada, S. Mizutani, Y. Yamanishi, *J. Chem. Inf. Model.* **2015**, *55*, 446–459.
[57] Z. Yu, F. Huang, X. Zhao, W. Xiao, W. Zhang, *Briefings Bioinf.* **2021**, *22*, bbaa24.
[58] Z. Li, Q. Huang, X. Chen, Y. Wang, J. Li, Y. Xie, Z. Dai, X. Zou, *Front. Chem.* **2020**, *7*, 924.
[59] T. Lee, Y. Yoon, *BMC Bioinf.* **2018**, *19*, 446.
[60] M. Sirota, J. T. Dudley, J. Kim, A. P. Chiang, A. A. Morgan, A. Sweet-Cordero, J. Sage, A. J. Butte, *Sci. Transl. Med.* **2011**, *3*, 96ra7.

Received: January 5, 2022
Accepted: March 25, 2022
Published online on April 7, 2022

