Computational and Structural Biotechnology Journal 20 (2022) 3639–3652

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

An accurate prediction model of digenic interaction for estimating pathogenic gene pairs of human diseases

Yangyang Yuan [a,b,c], Liubin Zhang [a,b,c], Qihan Long [a,b,c], Hui Jiang [a,b,c], Miaoxin Li [a,b,c,d,e,*]

[a] Program in Bioinformatics, Zhongshan School of Medicine and The Fifth Affiliated Hospital, Sun Yat-sen University, Guangzhou 510080, China
[b] Center for Precision Medicine, Sun Yat-sen University, Guangzhou 510080, China
[c] Center for Disease Genome Research, Sun Yat-sen University, Guangzhou 510080, China
[d] Key Laboratory of Tropical Disease Control (SYSU), Ministry of Education, Guangzhou 510080, China
[e] Guangdong Provincial Key Laboratory of Biomedical Imaging and Guangdong Provincial Engineering Research Center of Molecular Imaging, The Fifth Affiliated Hospital, Sun Yat-sen University, Zhuhai, 519000, China

ARTICLE INFO

Article history:
Received 17 February 2022
Received in revised form 4 July 2022
Accepted 4 July 2022
Available online 07 July 2022

Keywords:
Genetic interaction
Digenic interaction effect
Machine learning
Pathogenic gene pairs
Enrichment
Genomic mapping

ABSTRACT

Increasing evidence shows that genetic interaction across the entire genome may explain a non-trivial fraction of genetic diseases. Digenic interaction is the simplest manifestation of genetic interaction among genes. However, systematic exploration of digenic interactive effects on the whole genome is often discouraged by the high dimension burden. Thus, numerous digenic interactions are yet to be identified for many diseases. Here, we propose a Digenic Interaction Effect Predictor (DIEP), an accurate machine-learning approach to identify the genome-wide pathogenic coding gene pairs with digenic interaction effects. This approach achieved high accuracy and sensitivity in independent testing datasets, outperforming another gene-level digenic predictor (DiGePred). DIEP was also able to discriminate digenic interaction effect from bi-locus effects dual molecular diagnosis (pseudo-digenic). Using DIEP, we provided a valuable resource of genome-wide digenic interactions and demonstrated the enrichment of the digenic interaction effect in Mendelian and Oligogenic diseases. Therefore, DIEP will play a useful role in facilitating the genomic mapping of interactive causal genes for human diseases.

1. Introduction

Digenic Interaction (DI) is the simplest manifestation of genetic interaction among genes [1], which means one gene with one or more pathogenic variants may alter (aggravate or relieve) the impact of the other gene on a phenotype, including the true digenic and the genetic modifier referred in [2]. DI is different from the widely used monogenic (or Mendelian) inheritance pattern [3] or polygenic model [4]. An in-depth study of the DI effect may enable us to better understand the gene interaction networks and elucidate the potential relationship between genes and phenotypes, thereby making up the missing heritability [5] to some extent.

There is recognition gradually that the inheritance model of many diseases is more complex than originally thought. Researchers indicated that the genetic mechanism of monogenic traits might not be simple as we thought [6], and the final phenotype of a monogenic disorder can be an amalgamation of multiple factors to some extent [7]. So far, the pathogenesis remains unknown even in many well-studied Mendelian disorders [8,9], and one of the main reasons is that the genetic interaction of several genes may cause a large portion of the phenotypic variation sometimes [10,11]. Since the gene-gene interactions may play a larger role in disease susceptibility than a single gene [12], more and more research studies have jumped out of the scope of monogenic pathogenicity. They have discovered relatively more pathogenic gene pairs with DI effect by various methods such as conditional hybrid experiments [13], family-based association studies [14], genome-wide association studies [15] and multi-omics joint analysis [16] (Text S1). Nevertheless, only finite progress has been made because such methods still have limitations on screening the pathogenic gene pairs on a genome-wide scale (Text S2).

DIDA (DIgenic diseases DAtabase) is the first database providing the mutated variants (small-scale mutations) and the corresponding genes involved in digenic diseases [17], which is a valuable resource for researchers to identify the pathogenic gene pairs with the digenic interaction effect. So far, many studies have already

[*] Corresponding author at: Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510080, China.

E-mail address: limiaoxin@mail.sysu.edu.cn (M. Li).

https://doi.org/10.1016/j.csbj.2022.07.011
2001-0370/

been conducted based on this valuable database. For example, Gazzo et al. used a machine learning algorithm to automatically differentiate true digenic and composite classes based on the variant, gene and pathway-oriented features [2]. Besides, Papadimitriou et al. developed a variant combinations pathogenicity predictor (VarCoPP [18]) mainly for predicting disease-causing variant combinations in gene pairs. Mukherjee et al. constructed the DiGePred [19] aimed to identify digenic disease-causing gene pairs, which addressed the same problem as our DIEP, but with limitations on the selection of negative training sets and gene-level features for model training.

Although functionally related genes are more likely to have an interactive effect, it is not always true because the factors such as the dimensions and degree of similarity of two genes may also influence the gene interaction. Studies have already indicated that if two genes have similar functions or structures, or are on the same pathogenic pathway, or have protein–protein interaction effect [20], or form the same protein complex [7], they will be more likely to have a digenic inheritance pattern. For example, Wong et al. integrated gene localization, mRNA expression and other data types to predict genetic interaction [21]. Another study indicated that MYBPH might regulate the disease phenotypes of cMYBPC carriers, because these two genes are similar not only in functions but also in genome sequences and structures [14]. However, which types of similarities are critical for the digenic potential remains unclear. Here, taking the above characteristics of digenic pattern into account, we assumed that it is more efficient to predict the interaction potential between two genes at the gene level. Therefore, an advanced machine learning model integrated comprehensive features demonstrating the relationship or similarity of two genes was adopted.

In this study, we hypothesize that the gene level features (e.g., the biological relatedness) may be important for DI potential and can be used to build a more accurate model for predicting pathogenic digenic interactions. Therefore, we collected five categories of gene-level biological characteristics, including gene-based, protein-based, structure-based, expression-based and phenotype-based categories. The whole procedure of this study is shown in Fig. 1. We then developed a Digenic Interaction Effect Predictor (DIEP) to estimate gene pairs with pathogenic digenic interaction effect, which may help the researchers further dissect the complex mechanism of diverse diseases.

2. Materials and methods

2.1. Training dataset construction

The positive set was collected from the DIDA database [17] directly, which contains 140 digenic gene pairs, and the detailed information of the positive gene pairs (named "DIDA") was available in Dataset S1. As for the negative set, we hypothesized that the probability of being digenic of the randomly combined gene pair is very low since the digenic interaction is a rare biological phenomenon. Thus, the initial negative set was self-constructed under the above null hypothesis containing two sections. First, a series of gene pairs consisting of randomly selected protein-coding genes (downloaded from HGNC [22] website, excluding the genes already exist in DIDA) was regarded as negatives (Random set, 50,000 pairs). Second, 13,390 pairs obtained by randomly combining 165 unique genes in DIDA after excluding the positive gene pairs in DIDA were also considered as negatives (DIDA_NDI set). Moreover, to ensure the reliability of the training set and obtain a more convincing predictive result, we removed duplications, conducted quality control and data filtering (see details below) on the two datasets above. Finally, we obtained 16,156 negative pairs in total (1947 in the DIDA_NDI set and 14,209 in the Random set). For two negative sets, 1,400 and 7,000 gene pairs were randomly extracted from the DIDA_NDI set and Random set separately for training (8,400 in total), and the remaining was used as one test set (7,756 in total, Test set). In general terms, the training set consisted of 140 positive gene pairs and 8,400 negative pairs (Dataset S1, Fig. 1).

2.2. Feature determination and database selection

We initially selected 33 features in 5 categories for the prediction model, including gene-level, protein-level, structure-level, expression-level and phenotype-level (Table 1). The Pearson correlation coefficient was calculated to show the correlation between all 21 features (after deleting features with a high fraction of missing) based on the whole training set (Fig. S1). Detailed information on features is described as follows.

2.2.1. Gene-level features

The basic biological features and the fundamental characters for evaluating the gene relationship or similarity.

- Residual Variation Intolerance Score (RVIS): A measure of gene intolerance of mutational burden. Genes known to carry few common functional variants may be more likely to cause certain diseases than genes that have many such variants [23]. Thus, the tolerance of a gene's functional genetic variation may influence the interactions and even the phenotypes. The feature value was extracted from dbNSFP [24] based on EVS (ESP6500) and ExAC [25] data, respectively.
- Gene Damage Index score (GDI): A genome-wide, gene-level metric of the mutational damage [26], used to predict whether a given human protein-coding gene is likely to harbor disease-causing mutations.
- Gene Recessive score: Estimated probability of being a recessive disease gene [27].
- The Essential State of genes: Whether the genes are necessary for basic developmental functions [28,29].
- Gene indispensability score: A probability prediction of the gene being essential, used to estimate the global perturbation caused by deleterious mutations in each gene [30].
- Semantic similarity of gene GO annotations: GO [31] is divided into the biological process (BP), molecular function (MF) and cellular component (CC). We used the GO annotations within each category to calculate the semantic similarity between two genes by R package GOSemSim [32] as a feature because functionally similar genes tend to contribute to a similar phenotype [33].
- Gene Functional Correlation: GeneMANIA [34] helps find the functional-associated genes. The gene-gene functional-related weight of each gene pair provided in the network file was used as the feature value.
- Common Interactions: The number of common interactive genes of two different genes was calculated according to the database ConsensusPathDB [35]. Besides, we also calculated the similarity of two interactive gene sets by Jaccard similarity coefficient as follows based on ConsensusPathDB as another feature, where the N is the number of intersections or unions of two gene sets:

$$\mathrm{Jaccard}_{\mathrm{sim}}(\mathrm{GeneA},\ \mathrm{GeneB}) = \frac{N(Interactions_{GenesetA} \cap Interactions_{GenesetB})}{N(Interactions_{GenesetA} \cup Interactions_{GenesetB})} \quad (1)$$

- Haploinsufficiency: Estimated probability of haploinsufficiency [36] of the gene.
- Biological Distance: The HGC (Human Gene Connectome) [37] indicates biologically plausible routes, distances, and degrees of separation between all pairs of human genes, among which the biological distance was considered as a biological relatedness between two genes.

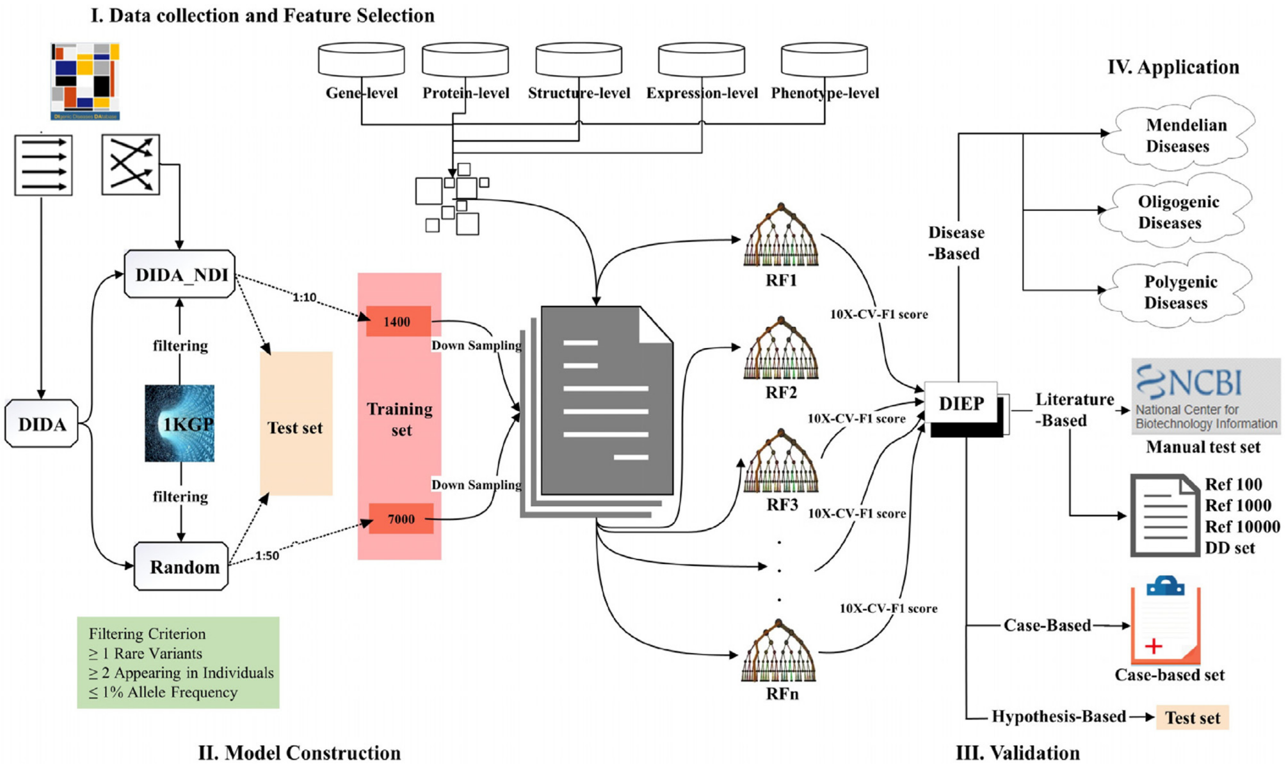


Fig. 1. The diagram of the framework for digenic interaction effect prediction, I. Data Collection and Feature Selection. Positive samples of the training set were collected from DIDA, which contains 140 positive gene pairs. Negative samples were obtained based on different theoretical assumptions, including two parts and 16,156 gene pairs in total, from which 8400 pairs were extracted for training and the remaining were used for testing. II. Model Construction. Down-sampling and bagging-based strategies were adopted to address the imbalanced issue for model construction. The weight of each classifier was assigned by the 10-fold cross-validation F1 score of each RF. III. Validation. Seven independent test sets were collected for model validation. One was self-constructed set under the null hypothesis, and another one was extracted from 7 trios cases with different rare diseases. Five were literature-based datasets, including one positive test set. The DD set was adopted for differentiating dual molecular diagnoses from the digenic interaction effect. IV. Application. Enrichment analysis was conducted in 15 different diseases to investigate the enrichment of the digenic interaction effect between disease-causing genes of the same disease.

● LOF intolerant: The probability of being loss-of-function intolerant [38].

● Functional Change Intolerance: A score for measuring the gene intolerance to functional change [39].

2.2.2. Protein-level features

● Biomedical interaction: We first adopted data from the database BioGRID (Biological General Repository for Interaction Datasets) [40] as one feature to see whether there is a biomedical interaction between two genes, and this feature was served as a binary variable.

● Protein-protein association: The STRING [41] was also considered as one type of feature. The "combined score" computed by combining the probabilities from the different evidence channels [42] was adopted and downloaded from https://string-db.org/ (9606.protein.links.v11.0.txt.gz). We considered that this database contains both known and predicted protein–protein associations. Here, the "association", from a functional perspective, can mean direct physical binding and indirect interaction, such as participation in the same metabolic pathway or cellular process as indicated in the publication [42].

● Functional interaction: The protein functional interaction effects such as catalyzing, regulating, activating, or acting as the protein complex supplied by REACTOME knowledgebase [43] were also considered.

2.2.3. Structure-level features

● Common structural domains: Two databases were adopted, including InterPro [44] and Pfam [45]. The information on structural domains was provided by UniProtKB. The feature values were also calculated using the Jaccard similarity coefficient.

2.2.4. Expression-level features

● Common highly expressed tissues: The gene expression data in different tissues of different genes from GTEx (Genotype-Tissue Expression) [46] were used. If the expression quantity of a gene in one tissue is 2.5 times more than the average level in all the tissues, this gene was considered highly expressed in this tissue. The feature value was calculated by the Jaccard similarity coefficient using the sets of highly expressed tissue for either gene in a pair.

● Protein abundance: The protein abundance data were obtained from PaxDb [47], and the feature values were calculated by adding and subtracting the abundance values of two genes separately.

● Gene coexpression: The degree of coexpression of two genes may reflect the inter-genes correlation. The gene coexpression data was obtained from COXPRESdb [48].

2.2.5. Phenotype-level features

● Semantic similarity of gene DO annotations: We used the disease ontology annotations to evaluate the semantic similarity with DOSE [49] between two genes, which was also adopted as an essential feature.

Table 1
Summary of the selected features and databases.

Category	Database/Method and Reference	Feature name	Fraction of missing (%)[1]	Significance (Wilcox/KS)[2]	Description[3]
Gene-level	RVIS [23]	RVIS_EVS.add / sub	59.14	.	The addition / subtraction of the residual variation intolerance score based on EVS (ESP6500) data.
		RVIS_ExAC.add / sub	58.25	.	The addition / subtraction of the residual variation intolerance score based on ExAC data.
	GDI [26]	GDI.add / sub	56.68	.	The addition / subtraction of the gene damage index score of two different genes.
	Gene Recessive score [27]	Recs.add / sub	35.04	<10E-3	The addition / subtraction of recessive scores of two different genes.
	The Essential State of genes [28,29]	EssgCom	11.58	.	The count of essential genes.
	Indispensability [30]	Indispensability.add / sub	58.57	.	The addition / subtraction of the gene indispensability score of two different genes.
	GOSemSim [32]	GOSemSim_MF GOSemSim_BP GOSemSim_CC	13.22	<10E-3	The Semantic similarity of Gene GO annotations for molecular function (MF), biological process (BP) and cellular component (CC).
	GeneMANIA [34]	GeneMANIAGG	4.9	<10E-3	The weight of the gene function relationship between two genes was calculated by integrating multiple functional association networks.
	ConsensusPathDB [35]	NumOfCommonInteraction commonInteractionJacSim	60.13	.	The number of common interactive genes of two different genes. The Jaccard similarity coefficient of two interactive gene sets.
	Haploinsufficiency [36]	HI.add / sub	23.83	<10E-3	The addition / subtraction of the probability of haploinsufficiency of two different genes.
	HGC [37]	BioDis	20.26		The biological distance between two genes.
	LOF [38]	LofIn.add / sub	13.48	add: 0.94 / 0.02 sub: 0.19 / 0.06	The addition / subtraction of the probability of being loss-of-function intolerant of two different genes.
	LoFtool [39]	FuncChangeInt.add / sub	66.02	.	The addition / subtraction of the probability of being functional change intolerance of two different genes.
Protein-level	BioGRID [40]	BioGRIDPP	19.9	<10E-3	Whether there is protein interaction between the two genes. “0″ indicates no protein interaction and “1” represents there is protein interaction.
	STRING [41]	STRINGPP	4.83	<10E-3	Protein-protein associations, including direct physical binding and indirect interaction from a functional perspective.
	REACTOME [43]	REAC_FI	33.9	<10E-3	Protein functional interaction effects
Structure-level	UniProtKB [50,51]	PS_2DbJacSim	5.4	.	The Jaccard similarity coefficient of two sets of structure domains, the maximum value by different databases for each gene pair was regarded as the feature value.
Expression-level	GTEx [46]	HighexpPer	6.68	<10E-3	The Jaccard similarity coefficient of two sets of high expressed tissues.
	PaxDb [47]	Abundance.add / sub	11.38	add: 0.001 / <10E-3	The addition / subtraction of abundance scores of two different genes.
	COXPRESdb [48]	COXPRESdbMRvalue	9.03	<10E-3	Gene co-expressed values.
Phenotype-level	DOSE [49]	DOSemSim	13.22	<10E-3	The Semantic similarity of disease ontology annotations.

Note: The value range of the features *.add / sub were [0,2], while the remaining were [0,1], except for the classification feature EssgCom and BioGRIDPP;

[1] The fraction of missing values was calculated based on each database independently. If there were several databases, the minimum miss rate was shown.

[2] The significance testing results of each feature (after feature selection) between positive set and negative set, p-values for deleted features were replaced by “.”. The significant test was conducted using two methods, including the Wilcoxon test and the Kolmogorov-Smirnov test.

[3] Features such as RVIS_EVS, GDI and Recessive score are specific to a single gene. Thus, we calculated the addition and the subtraction to indicate the correlation between two genes. The “.add” is the addition of two values, and the “.sub” is the subtraction.

2.3. Data preprocessing, imputation, quality control and filtering

2.3.1. Data preprocessing

To make the feature values in the range of 0 to 1, we conducted the normalization on all the retained features after feature selection as below.

$$Norm\_x_i = \frac{x_i}{\max(X)} \tag{2}$$

where x_i is the original feature value, X is the set of feature values in the corresponding database.

2.3.2. Imputation

Imputation for the features with high missing rates will lead to a great bias. Thus, before we conducted imputation, features with a high fraction of missing were deleted firstly (>40%, Table 1, see details in Feature selection section). The missing rates of the remaining features ranged from 4.9% to 35.04% (Table 1), and the imputation was then conducted on these features based on the whole-exome gene pair information. In our study, we used the multivariate feature imputation method with the IterativeImputer class provided in the scikit-learn implementation [52] to impute the missing values, and we used default arguments for all the parameters.

2.3.3. Quality control and data filtering

The original DIDA_NDI gene pairs consisted of a random combination of 165 unique genes from the DIDA (excluding the positive gene pairs in DIDA, 13,390 pairs). Thus, the label noise (false negative) will be generated and then harm the accuracy of the prediction model because some diseases in DIDA have several causal genes. Therefore, we used the 1000 Genomes Project Data [53] to purify the negative gene pairs. Most subjects in the 1000 Genomes Project were assumed to not suffer from severe diseases, and we considered that the mutated gene pair showing up in unaffected individuals was non-pathogenic. Thus, we searched each of the negative gene pairs ($13390 + 50000$) in each sample from the 1000 Genomes Project and defined that gene pairs with at least one rare (mutation frequency $\leq$ 1%) non-synonymous variant in both genes in 2 or more subjects as the true negative. The filtering process was conducted by KGGSeq [54], and only 1,948 and 14,209 gene pairs were retained for two negative training sets separately (Dataset S2). KGGSeq is a biological knowledge-based platform for genetic studies, and we have already implemented the above screening process in KGGSeq (by command "--digene-assoc") for other researchers (Details in Supplementary Information).

2.4. Feature selection and model construction

The random forest algorithm was used to construct the Digenic Interaction Effect Predictor (DIEP) based on the 140 positive gene pairs and 8,400 negative pairs. Then, given the unbalanced sample, down-sampling and bagging-based strategies were adopted for the random forest. Specifically, 150 pairs were obtained each time by down-sampling each negative set separately, and we then merged them with positive gene pairs to generate sub-samples. In this study, we generated 200 sub-samples in total by down-sampling for training.

The feature selection was conducted before model construction. First, features with a high fraction of missing (>40%, Table 1) or strongly correlated (r > 0.8, Fig. S1) were deleted. Then, we tried to adopt Recursive Feature Elimination (RFE) for further feature selection. A simple RF (Random Forest) classifier was built using the collected feature set with default parameters for each sub-sample. Next, the average Gini feature importance was calculated, and the least important feature was eliminated. Then, new classifiers were constructed using the updated feature set for each sub-sample, and the least important one was also deleted, and continued similarly until only one feature was retained. Finally, the best feature set was simply obtained after deleting the features with extremely low Gini feature importance (<0.01) (Fig. 2A) because of the small number of features in this study.

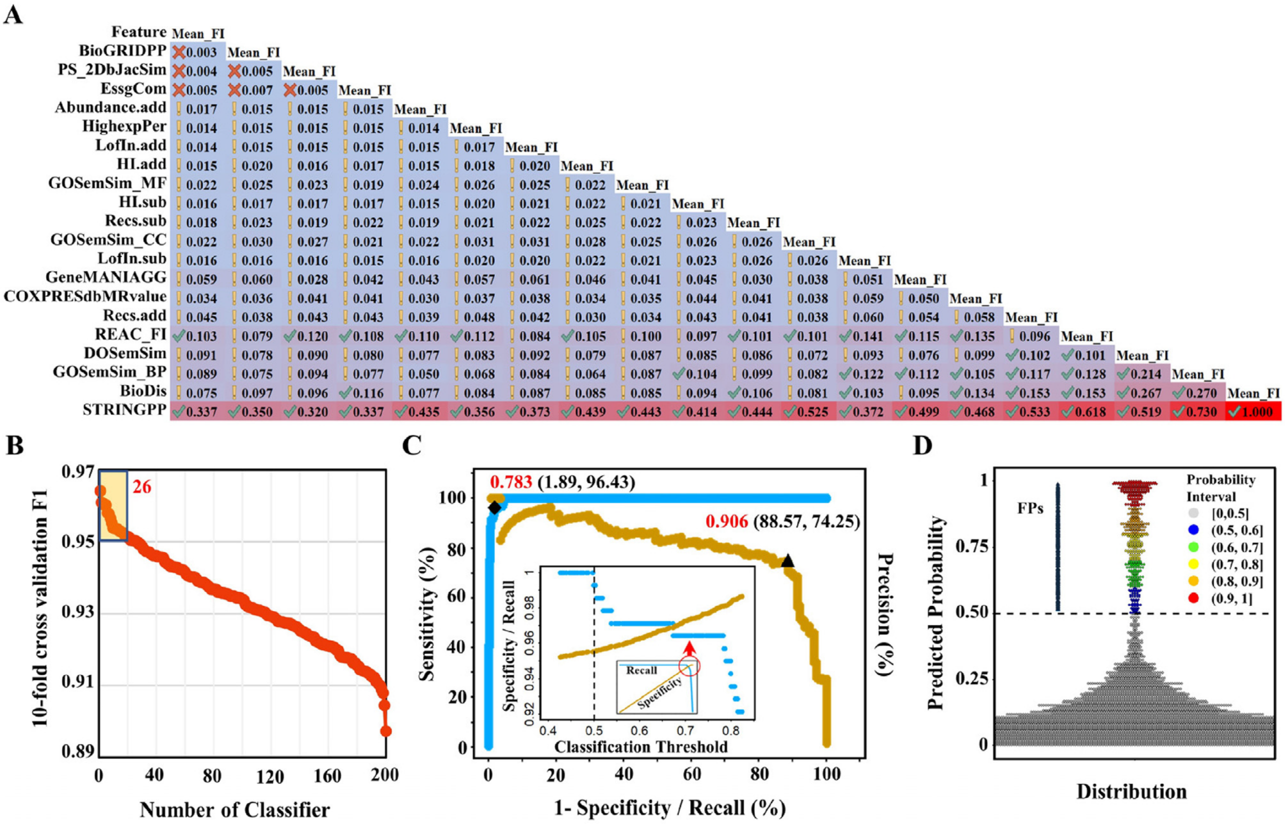
After the features were determined, we used the scikit-learn implementation [52] of the Random Forest (RF) algorithm (RandomForestClassifier) to train 200 single classifiers for each subsample. For each classifier, parameters including "n_estimators", "max_features", "max_depth", "min_samples_split" and "min_samples_leaf" were adjusted by GridSearchCV, using the "f1" as the evaluation criteria by 10-fold cross-validation (mean of the ten scores). F1 is the weighted harmonic average of Precision and Recall, which is widely used as an indicator to measure the accuracy of a binary model. Besides, feature importance for each feature in each classifier was calculated using Gini impurity. The top n classifiers with high mean 10-fold cross-validation F1 scores were then retained for the final prediction (n was determined according to the top numbers of 10x-CV-F1s), and the weight of each classifier was assigned using the corresponding 10x-CV-F1 score.

2.5. Validation of model effectiveness

To verify the performance of DIEP, we collected six independent negative sets and one positive test set. For negative sets, one was extracted from the original collected negative gene pairs (Test set) under the null hypothesis (see details above, Fig. 1). Then, we used our local samples to create a case-based test set to obtain more control datasets for validation. Seven trios (at least one healthy parent and affected child) with different rare genetic diseases were adopted (including an Asymmetric bone disease case, two Congenital microcephaly cases, an Epileptic encephalopathy case, a Hypertrophic cardiomyopathy case, a Pulmonary embolism case and a Charcot-Marie-Tooth atrophy case). We considered those variants only existing in unaffected of the same family to be neutral (non-pathogenic). Thus, such variants overlapping in at least three trios were extracted and mapped into genes separately. Then, we randomly combined every two genes and created this test set. Three literature-based sets provided by Papadimitriou [18] were 100, 1,000, and 10,000 neutral bi-locus unique combinations from the 1000 Genomes Project (1KGP) after we mapped the locus into genes using KGGSeq [54], and there were no overlapping gene pairs after checking all three sets. The last negative set was DD (Multiple (or dual) molecular diagnoses) set provided in [55]. Significantly, 64 probably digenic gene pairs (different from the positive gene pairs in DIDA) were manually collected from literature as a positive test set (Table S1) by searching with keywords such as "digenic disease", "digenic inheritance", "modifier gene", "genetic modifier".

2.6. Visualization of random forest and results interpretation

Many interpretation methods for trees in machine learning only summarize the impact of the input features as a whole (like Gini importance), which may not help explain the result of the single sample [56]. In order to explain the classification mechanism, dissect the results of our DIEP and make it readily understandable to readers, the visualization of the random forest classification results was conducted using the TreeExplainer [56] feature contribution proposed in SHAP (SHapley Additive exPlanations, https://github.com/slundberg/shap) [57]. In brief, the feature contribution of each input feature for each input gene pair was calculated by TreeExplainer, and then the most contributive features for a specific classification result were found. Finally, the plots were generated by ggplot2 [58] package in R [59].



Feature	Mean_FI																			
BioGRIDPP	✗0.003	Mean_FI																		
PS_2DbJacSim	✗0.004	✗0.005	Mean_FI																	
EssgCom	✗0.005	✗0.007	✗0.005	Mean_FI																
Abundance.add	! 0.017	! 0.015	! 0.015	! 0.015	Mean_FI															
HighexpPer	! 0.014	! 0.015	! 0.015	! 0.015	! 0.014	Mean_FI														
LofIn.add	! 0.014	! 0.015	! 0.015	! 0.015	! 0.015	! 0.017	Mean_FI													
HI.add	! 0.015	! 0.020	! 0.016	! 0.017	! 0.015	! 0.018	! 0.020	Mean_FI												
GOSemSim_MF	! 0.022	! 0.025	! 0.023	! 0.019	! 0.024	! 0.026	! 0.025	! 0.022	Mean_FI											
HI.sub	! 0.016	! 0.017	! 0.017	! 0.017	! 0.015	! 0.020	! 0.021	! 0.022	! 0.021	Mean_FI										
Recs.sub	! 0.018	! 0.023	! 0.019	! 0.022	! 0.019	! 0.021	! 0.022	! 0.025	! 0.022	! 0.023	Mean_FI									
GOSemSim_CC	! 0.022	! 0.030	! 0.027	! 0.021	! 0.022	! 0.031	! 0.031	! 0.028	! 0.025	! 0.026	! 0.026	Mean_FI								
LofIn.sub	! 0.016	! 0.016	! 0.016	! 0.015	! 0.016	! 0.020	! 0.020	! 0.022	! 0.021	! 0.023	! 0.026	! 0.026	Mean_FI							
GeneMANIAGG	! 0.059	! 0.060	! 0.028	! 0.042	! 0.043	! 0.057	! 0.061	! 0.046	! 0.041	! 0.045	! 0.030	! 0.038	! 0.051	Mean_FI						
COXPRESdbMRvalue	! 0.034	! 0.036	! 0.041	! 0.041	! 0.030	! 0.037	! 0.038	! 0.034	! 0.035	! 0.044	! 0.041	! 0.038	! 0.059	! 0.050	Mean_FI					
Recs.add	! 0.045	! 0.038	! 0.043	! 0.043	! 0.039	! 0.048	! 0.042	! 0.030	! 0.034	! 0.043	! 0.041	! 0.038	! 0.060	! 0.054	! 0.058	Mean_FI				
REAC_FI	✓0.103	! 0.079	✓0.120	✓0.108	✓0.110	✓0.112	! 0.084	✓0.105	! 0.100	! 0.097	✓0.101	✓0.101	✓0.141	✓0.115	✓0.135	! 0.096	Mean_FI			
DOSemSim	! 0.091	! 0.078	! 0.090	! 0.080	! 0.077	! 0.083	! 0.092	! 0.079	! 0.087	! 0.085	! 0.086	! 0.072	! 0.093	! 0.076	! 0.099	✓0.102	✓0.101	Mean_FI		
GOSemSim_BP	! 0.089	! 0.075	! 0.094	! 0.077	! 0.050	! 0.068	! 0.084	! 0.064	! 0.087	✓0.104	! 0.099	! 0.082	✓0.122	✓0.112	✓0.105	✓0.117	✓0.128	✓0.214	Mean_FI	
BioDis	! 0.075	! 0.097	! 0.096	✓0.116	! 0.077	! 0.084	! 0.087	! 0.085	! 0.085	! 0.094	✓0.106	! 0.081	✓0.103	! 0.095	✓0.134	✓0.153	✓0.153	✓0.267	✓0.270	Mean_FI
STRINGPP	✓0.337	✓0.350	✓0.320	✓0.337	✓0.435	✓0.356	✓0.373	✓0.439	✓0.443	✓0.414	✓0.444	✓0.525	✓0.372	✓0.499	✓0.468	✓0.533	✓0.618	✓0.519	✓0.730	✓1.000

Fig. 2. Parameter determination and model performance on the whole training set. (A) The change of the average feature importance of 200 RF classifiers when conducting the feature selection using the Recursive Feature Elimination (RFE) method. The initial number of features for REF was 20, and the average feature importance was calculated each time the least important feature was deleted. The value means the Gini feature importance (FI), the red cross indicates the features with FI < 0.01 (should be deleted), the yellow exclamation mark indicates the FI is in the range of [0.01, 0.1], and the green tick indicates the FI > 0.1. (B) The order of the 10-fold cross-validation F1 scores for the 200 single classifiers trained using sub-samples generated by the down-sampling method. The top 26 RF classifiers had 10x-CV-F1s ≥ 95%. (C) The ROC (blue) and PR (yellow) curves of the final predictor. The peak means the best classification threshold for each curve. The embedded plot showed the change of Recall and Specificity rate in the positive and negative training set with the increased threshold. The recall rate stabilized at 100% when the threshold was ≤0.5. (D) The probability distribution of the final predicted scores. The dark and light triangles represented the misclassified samples in the negative and positive gene pairs separately. The scatter plot showed the distribution of the predicted probability of digenic interaction for the whole training set (16156 pairs). One point represented one gene pair, and the height meant the probability of the digenic interaction effect, those gene pairs with the same probability would stay at the same height. Most of the gene pairs were at the upper and lower ends of the plot, which indicated a clear result by the classification model.

2.7. Comparison with DiGePred

Since the DIEP and DiGePred [19] are relatively similar, we compared them systematically and comprehensively from three aspects. First, the PR AUCs (Area Under Precision-Recall Curve) were compared between DIEP and DiGePred based on two different test sets (unaffected-no-gene-overlap_non_digenic_pairs_held_out_test.csv and random-no-gene-overlap_non_digenic_pairs_held_out_test.csv) collected from DiGePred publication (https://github.com/CapraLab/DiGePred). To be consistent with the DiGePred paper, the predicted scores of DiGePred on these two sets were calculated using the best model (DiGePred_unaffected-no-gene-overlap_model.sav) provided in GitHub. Second, we compared the sensitivity between DIEP and DiGePred on our manually collected digenic gene pairs (Table S1), and the predicted scores of DiGePred were calculated using DiGePred Server (https://www.meilerlab.org/index.php/servers/show?s_id=28). Finally, we used our Manual test set (64 pairs) and Test set (7756 pairs) to generate 100 sub-test sets in a 1:1 ratio (128 gene pairs for each sub-test set), and we then conducted the McNemar's test on these 100 sub-test sets to show the difference in the performance of DIEP and DiGePred. The scores of DiGePred were also calculated using DiGePred Server.

2.8. Enrichment analysis

Enrichment analysis was conducted with 15 randomly selected diseases, including six Mendelian (Retinoblastoma, Neurofibromatosis, Hemophilia, Huntington, Phenylketonuria, Dilated cardiomyopathy), five oligogenic (Alport syndrome, Long QT syndrome, Megacolon, Amyotrophic lateral sclerosis, Bardet-Biedl syndrome) and four polygenic (Dementia, Diabetes, Heart disease, Hypertension) diseases. Digenic interaction effect of gene pairs in which both genes were from the same disease and two genes were from two different diseases were predicted respectively. The Fisher test and Chi-square test in R were used to evaluate the statistical significance of the enrichment [60]. Meanwhile, the multiple test correction was conducted by Benjamini and Hochberg FDR (BH) [61].

2.9. Compression of whole-genome predicted results

The whole-genome predicted results were stored in an upper triangular matrix (Fig. S2D), and we transformed it into a triple table (GeneA, GeneB, Digene_Score) (Fig. S2C) for the convenience of readers. There are 19,616 unique coding genes extracted from HGNC (HUGO Gene Nomenclature Committee) [22] in total, which made up 192,383,920 gene pairs. We extracted the first two columns of the triple table and stored the gene names in a dictionary file (Fig. S2A). The predicted digenic scores (Digene_Score) were stored in bytes to save space. We also provided a high-efficiency java package for searching the digenic scores for specific gene pairs (https://github.com/pmglab/DIEP). A detailed method was provided in Supplementary Information.

2.10. Data and software availability

All detailed supplementary information is shown in Supplementary Information (including the description of main abbreviations), and other supplementary data are available in our GitHub (https://github.com/pmglab/DIEP), including Supplementary Datasets S1-S8, Supplementary Tables S4–S8, training sets and test sets for machine learning, the single classifiers that made up the final DIEP, public codes and the high-efficiency java package for quick searching. The whole GitHub repository can be downloaded by "git clone https://github.com/pmglab/DIEP". However, due to the limited storage of Git LFS Data, Datasets S4, S5 and S7 with large file sizes were moved to our OneDrive and Google Driver (more stable for downloading), and see details in the corresponding README.md file in GitHub. The whole-genome predicted result (DIEP final database) was also hosted on our web server, and users can download it by "wget" according to the User manual on GitHub. Besides, the website of KGGSeq is https://pmglab.top/kggseq/.

3. Results

3.1. Performance of the proposed digenic interaction effect predictor

We proposed a framework to predict pathogenic gene pairs with digenic interaction effect based on the biological relatedness or similarity of the genes, named Digenic Interaction Effect Predictor (DIEP). The conventional Principal Component Analysis (PCA) result indicated that simple feature dimension reduction had limited classification power on the digenic effects of genes (Fig. S3). Hence, DIEP consisted of several Random Forest (RF) classifiers with different weights according to the bagging strategy. The RF algorithm (RandomForestClassifier) implemented by the Python package, scikit-learn, was trained and evaluated by 10-fold cross-validation. The gene pairs in the training set of DIEP were collected from DIDA and other resources. In total, 17 features from four categories were retained for the prediction after feature selection from 33 features (Fig. 2A and Table 1). Moreover, we also validated the trained model in 7 different datasets. Finally, the established prediction model was applied to investigate pathogenic gene pairs with digenic interaction effects for various diseases (Fig. 1).

To reduce the influence of the imbalanced sample (140 positives vs. 8,400 negatives) on the prediction performance, we adopted a down-sampling strategy to train the RF model. We randomly sampled 50 and 100 samples from each negative training set (DIDA_NDI and Random set) separately (150 pairs in total) and merged them with 140 positive gene pairs to generate subsamples (290 pairs) for training. We trained 200 RF classifiers using 200 different down-sampling sets. Finally, the top 26 RF classifiers with higher 10-fold cross-validation F1 scores ($\geq$95%) were singled out for ensembling (Fig. 2B), and the weight of each RF classifier was drawn on its 10-fold cross-validation F1 score.

The 26 RF ensemble model achieved an Area Under Curve (AUC) of Receiver Operating Characteristic Curve (ROC curve) as high as 0.996 (Fig. 2C). The ROC peak indicated a true positive (TP) rate of 96.43% and a false positive (FP) rate of 1.89% with the classification threshold of 0.783 on the whole training set. As for the Precision-Recall curve, the peak of the PR (Fig. 2C) showed a relatively low TP(Recall) rate (88.57%) and the precision (74.25%) at the threshold of 0.906. The figure embedded in Fig. 2C showed the change of Recall and Specificity rate in the positive and negative training set with the increased threshold. We found that the Recall rate stabilized at nearly 100% when the threshold was < 0.5. The threshold in our study was assigned as the default classification threshold (0.5) by RF. According to a predicted probability distribution of the training set in the scatter plot in Fig. 2D, we could easily find that most of the gene pairs were at the upper and lower ends of this plot, which indicated a clear classification result by the trained model.

3.2. Investigation of feature importance to the prediction

We then investigated the importance of each feature to the prediction. Fig. 3A showed the importance score of each feature from RF calculated by Gini impurity. The "STRINGPP" was the most crucial feature for distinguishing the digenic gene pairs in our predictor. Together with the "REAC_FI" and "BioDis", were the top 3 important features with relatively high feature importance scores. Besides, features including the "DOSemSim", "GOSemSim_BP" and "GeneMANIAGG" also played a role in prediction because a further reduction in the number of features will result in a less efficient predictor (Fig. 3B). Besides, the statistical test showed that most of the individual features were significantly different between the gene pairs with and without digenic effects (Table 1).

Moreover, to clarify the necessity and illustrate the significance of the selected features in our DIEP, we trained some other predictors with different subsets of input features (Fig. 3B). The Manual set and the Test set referred to in Table S2 were combined as one integrated test set. The results showed that those predictors only using subsets of all features did not perform well according to the five different indicators (Dataset S8). Specifically, for 10-fold cross-validation F1 scores, predictors including "Only_STRINGPP" (10x-CV-F1 = 0.938) and "Top3_Features" (0.943) performed worse than the Final predictor (0.954) in the training set. Moreover, the deletion of feature "STRINGPP" brought the 10x-CV-F1 down to 0.858 ("Without_STRINGPP"). Besides, DIEP had higher recall and specificity on the integrated test set with a competitive AUC score (0.95). The results showed that although the performance of "Only_STRINGPP" looked good, other features helped increase the recall on actual cases from 0.875 to 0.891 (spider plot in Fig. 3B and Dataset S8), and improved PR AUC by 6.6 percent (bar plot in Fig. 3B). Generally speaking, the Final-DIEP had a better performance based on five metrics for performance evaluation of predictors.

3.3. Detailed distribution of the feature values in DIEP

The Gini importance only summarized the impact of input features on the model as a whole, which paid no attention to local explanations for the classification of single gene pair. Thus, we used TreeExplainer [56] to calculate the contribution of input features on individual predictions (single gene pair), and reveal the impact of each feature sequentially to better understand the classification mechanism. We then looked into the detailed distributions of feature contributions in five datasets by DIEP. In the true positive sets, almost all gene pairs had positive contribution values (Fig. 4A and 4C, Table S3), and nearly all gene pairs in the true negative set had contribution values numerically less than or close to 0 at the features (Fig. 4D). Meanwhile, we found that the "STRINGPP" was the most contributive feature, with other features including "REAC_FI", "BioDis", "DOSemSim" and "GOSemSim_BP" following. This phenomenon suggested that if there is a strong protein–protein association between two genes in the STRING database [41],

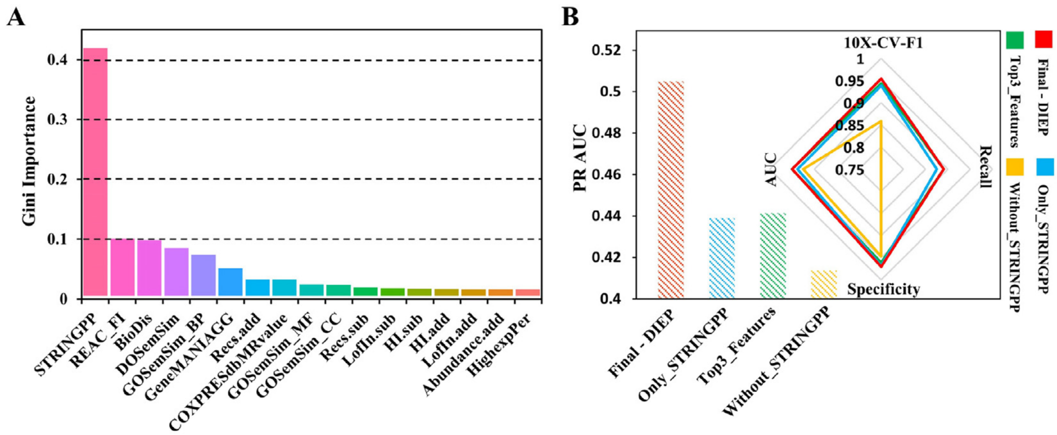


Fig. 3. The feature importance of DIEP and multiple predictors comparison. (A) The bar plot of the feature importance of DIEP. (B) Five metrics for the comparison of four predictors trained with different subsets of input features. The mean 10-fold cross-validation F1 scores were calculated based on the training set. Other indicators, including the PR AUCs, AUCs, recall and specificity were calculated based on the Manual set and the Test set (Table S2). "Only_STRINGPP" means that the predictor was trained with only one feature. And "Top3_Features" indicates that the predictor was trained with the top three features. "Without_STRINGPP" suggests that only the feature "STRINGPP" was deleted. The "Final" represents our DIEP.

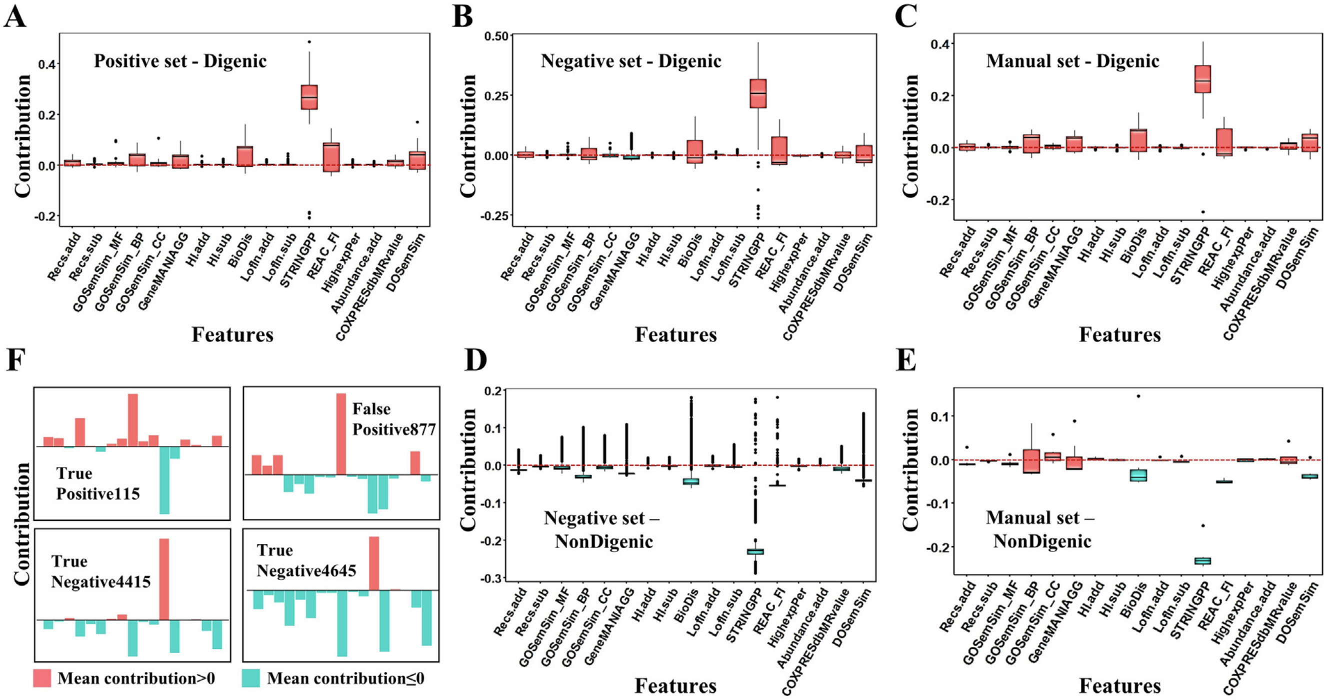


Fig. 4. The boxplot of feature contribution in the Positive and Negative set. In each dataset, the boxplot of predicted digenic gene pairs and non-digenic gene pairs were plotted separately in (A), (B), (C), (D) and (E). The red box indicates that the mean feature contribution of all the gene pairs is positive (≥ 0), and the blue is negative (<0). (F) The feature contribution pattern of an exceptional gene pair in the different datasets. The red color indicates the positive contribution (+) while blue is the negative (−). The order of the features is the same as (A)-(E).

the feature "STRINGPP" will drive the classification towards the digenic class with a positive contribution value. Besides, two genes will have a higher chance of belonging to the digenic class if they cause the diseases with high semantic similarity, have functional interaction, participate in the same biological process or have a close connection based on biological distance. However, for the gene pairs in the false positive or false negative sets, the votes of the features for the classification were contradictory. Some features drove the prediction to the digenic class, while others drove to the opposite in either the positive set (Fig. 4E) or the negative set (Fig. 4B).

To more insightfully understand feature contribution to the prediction, we further investigated the detailed prediction results of each gene pair (Tables S4–S6). Although the "STRINGPP" was the primary driving factor for classification (Fig. 4), the combined effects of other features cannot be overlooked. For instance, "STRINGPP" had negative contributions on 6 of the 140 positive gene pairs, among which gene pair *TEK/CYP1B1* (0.672) (115th, labeled as Positive115) was classified correctly but with a high negative contribution of "STRINGPP" (Fig. 4F). Similar in the negative set, Negative877 ($S_{NPHS1:CDH23}$ = 0.503) was misclassified as digenic, but the top two features had negative contributions, which may be caused by the high contribution of “BioDis”. Besides, Negative4415 ($S_{ZFAND6:XIRP1}$ = 0.39) and Negative4645 ($S_{HAUS3:TMEM150}$ = 0.277) were assigned as non-digenic, but the “STRINGPP” had a high positive contribution while most of the remaining features had relatively low negative contributions.

3.4. Further evaluation of DIEP in 7 independent datasets

We further assessed the DIEP in 7 additional independent test sets (Table S2). One was self-constructed and separated from the original collected negative gene pairs generated under the null hypothesis (Test set, Fig. 1, Dataset S3-Sheet3). One was extracted from 7 trios with different rare diseases (Case-based test set, Dataset S3-Sheet6). Four literature-based test sets were also adopted, from which three sets were also under the null hypothesis (including ref_100random set, ref_1000random set and ref_10000random set, S3 Dataset-Sheet9-11), and DD set was described in the section below (Dataset S3-Sheet5). The remaining set was the only positive test set manually curated from disease research articles (Manual test set, Dataset S3-Sheet4).

● **DIEP had low false-positive rates in the hypothesis-based and the case-based dataset.** The FP rate was only 2.95% in the self-constructed null hypothesis dataset (Test set), and only 24 (0.3%) false positives had relatively high probabilities (>0.9, also see the distribution of the predicted scores in Fig. 5A (Dataset S3)). In the Test set, the digenic interaction probability of the gene pair (*MKKS/CEP290*) was 0.967. The study indicated that the association of CEP290 and MKKS would affect the integrity of multiprotein complexes at the cilia transition zone and basal body in mice and zebrafish [62]. Besides, in the case-based test set, we got the FP rate of 1.54%, and only five false-positive gene pairs had predicted probabilities > 0.9 (Fig. 5B).

● **DIEP performed well in literature-based datasets.** In terms of the three literature-based test sets under the null hypothesis from another research group which contained 100, 1,000 and 10,000 random gene pairs, the FP rates were only 1%, 1.3% and 2.62% respectively by DIEP (Fig. 5C-5E). That is, our predictor achieved a specificity of close to 98% in these 3 test sets, and 68%, 65.4%, 62.7% of the samples had low predicted probabilities $\leq$ 0.1 (Dataset S3).

● **DIEP made more accurate predictions on the digenic effects of genetic diseases**. Significantly, we had manually curated 64 probably digenic gene pairs from various research studies, in which 14 pairs also served as the positive test set by another predictor for identifying digenic disease genes (DiGePred) [19]. As a result, our DIEP correctly predicted all 14 digenic pairs (100%) with relatively high probabilities (0.54–0.97), while DiGePred only had the TP rate of 57.14% at a suggested threshold of 0.496, and the predicted score only ranged from 0.528 to 0.804. For the remaining 50 gene pairs, DIEP had a TP rate of 86% (43/50), whereas DiGePred only predicted 19 of them correctly (38%). In general, DIEP had correctly predicted 89.06% of the digenic pairs in the Manual test set curated from literature (Fig. 5-F-5G), which outperformed DiGePred (42.19%) (Dataset S3, see the detailed comparison below).

● **Discriminate dual molecular diagnoses from digenic interaction effect.** Multiple (or dual) molecular diagnoses (DD) refers to the conjunction of two independent diseases caused by different mutated genes that show simultaneously on one patient [63]. DD is a type of bi-locus effect but not the digenic interaction effect, which is often confounded with digenic interaction in practice. We used gene pairs with pathogenic variants for 97 patients with Dual Molecular Diagnoses in [55]. Papadimitriou et al. [18] predicted 67 gene pairs (88%) out of 76 gene pairs in the dataset to have pathogenic variant combinations by VarCoPP, a tool for predicting the potential pathogenicity of variant combinations in gene pairs. On the contrary, DIEP only predicted 24 digenic interaction pairs (26.09%) out of 92 gene pairs in the 97 patients with DD (Fig. S4). Of the 24 pairs, only seven pairs had relatively high probabilities (>0.75, such as $S_{KCNQ2:SCN8A}$ = 0.914, $S_{KCNQ2:PRRT2}$ = 0.777). Although there is no direct evidence showing the digenic interaction between these genes, mutations in KCNQ2, PRRT2 and SCN8A were already identified in different types of epilepsy [64,65]. Besides, 27.17% of the gene pairs had a digenic score lower than 0.1, which indicated a non-digenic pattern or very low digenic interaction effect (Dataset S3). Meanwhile, we also used DiGePred to predict the above 92 gene pairs, and the result showed that only four pairs had the digenic effect (Dataset S3). In other words, DIEP and DiGePred will help differentiate DD from the digenic interaction effect efficiently.

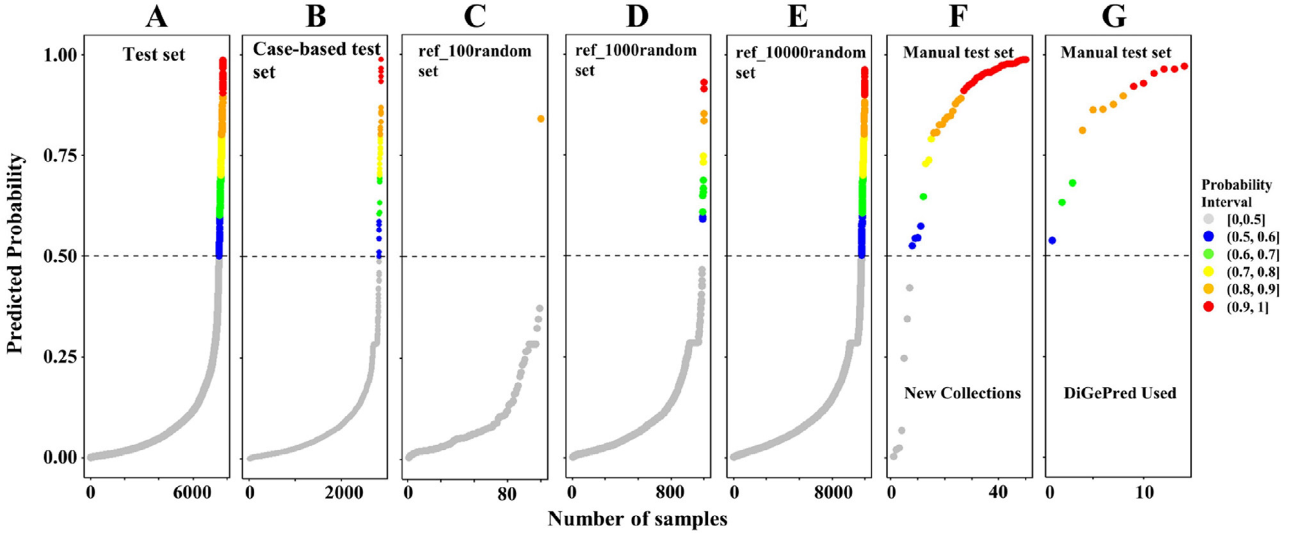


Fig. 5. The distribution of the predicted probability of digenic interaction effect in 6 additional test sets. (A) The distribution of the predicted probability of digenic interaction effect in the hypothesis-based set extracted from the original negative set (Fig. 1). (B) The test set was generated from healthy people of 7 trios with different rare diseases. (C)-(E) The predicted probability of digenic interaction effect in 3 literature-based sets used in VarCoPP. (F)-(G) The only positive test set from which the digenic gene pairs were manually curated from the disease research studies. Samples in (G) were also served as positive test set by DiGePred.

3.5. A systematic comparison of the DIEP and DiGePred

We made a systematic comparison between DIEP and DiGePred in multiple aspects. First, we compared the PR AUCs of DIEP and DiGePred on two different test sets from DiGePred (unaffected- and random-no-gene-overlap_held_out_test). The results showed that DIEP performed much better than DiGePred on both test sets with higher PR AUCs (unaffected: 0.895 vs. 0.73; random: 0.683 vs. 0.555, Fig. 6A and Fig. 6B, Dataset S3). DiGePred performed even worse on our Manual test set with very low sensitivity (42.19%), and most of the probably digenic gene pairs were predicted as non-digenic with extremely low scores (Fig. 6C). Fig. 6E indicated that DIEP has correctly predicted nearly 89.06% of the digenic pairs from various research studies, while DiGePred gave wrong predictions on 57.81% of those gene pairs. Finally, McNemar's test was adopted to show the differences in model performance on 100 down-sampling sub-test sets. Fig. 6D showed that DIEP performed significantly better than DiGePred on all the sub-test sets (McNemar's p-value < 2.76E-04) with high prediction accuracy (Table S7).

3.6. Enrichment of digenic interaction effect in known disease genes

We then used DIEP to test the hypothesis that the pathogenic genes of the same disease are more likely to have the digenic interaction effect. We chose responsible genes of 15 different diseases randomly drawn from DisGeNET [66] for the hypothesis test. Diseases included six Mendelian, five oligogenic and four polygenic diseases (Table S10). First, 15 same disease gene pair sets were generated by randomly combining genes from the same disease, and 35 different diseases gene pair sets were created by randomly combining two genes from different diseases (Table S8). The digenic interaction effect was predicted respectively for each gene pair set. Then, gene pairs in the training set were excluded to avoid overfitting. Fig. 7A showed the statistical results of the digenic interaction effect between 15 same disease gene pair sets and 15 different representative diseases gene pair sets. Plenty of gene pairs from the same disease gene pair sets were predicted with the digenic interaction effect (Dataset S4). The percentage of digenic gene pairs ranged from 11.99% to 36.22%. Compared to gene pairs from different diseases gene pair sets, the enrichment of digenic interaction among gene pairs from the same disease was statistically significant for most diseases. The statistical analysis also confirmed the strong association between the same disease gene pairs and the digenic interaction effect (Fig. 7A, Table S8, Datasets S4-5).

The enrichment ratios (Odds ratios) of digenic interaction in Mendelian and oligogenic diseases were higher than those in polygenic diseases (Fig. 7B). In Mendelian diseases, for example, the probability of being digenic was 2.89 times higher in gene pairs from Retinoblastoma Disease (Retb, same disease gene pairs) compared with those from Alport syndrome (Alps) and Bardet-Biedl syndrome (Babi) (different diseases gene pairs) (Fig. 7C). All the enrichment ratios and the 95% confidence intervals for gene pairs from the same disease exposure were >1 (Table S8), showing a positive association between the same disease gene pairs and digenic interaction. In oligogenic diseases, gene pairs from Bardet-Biedl

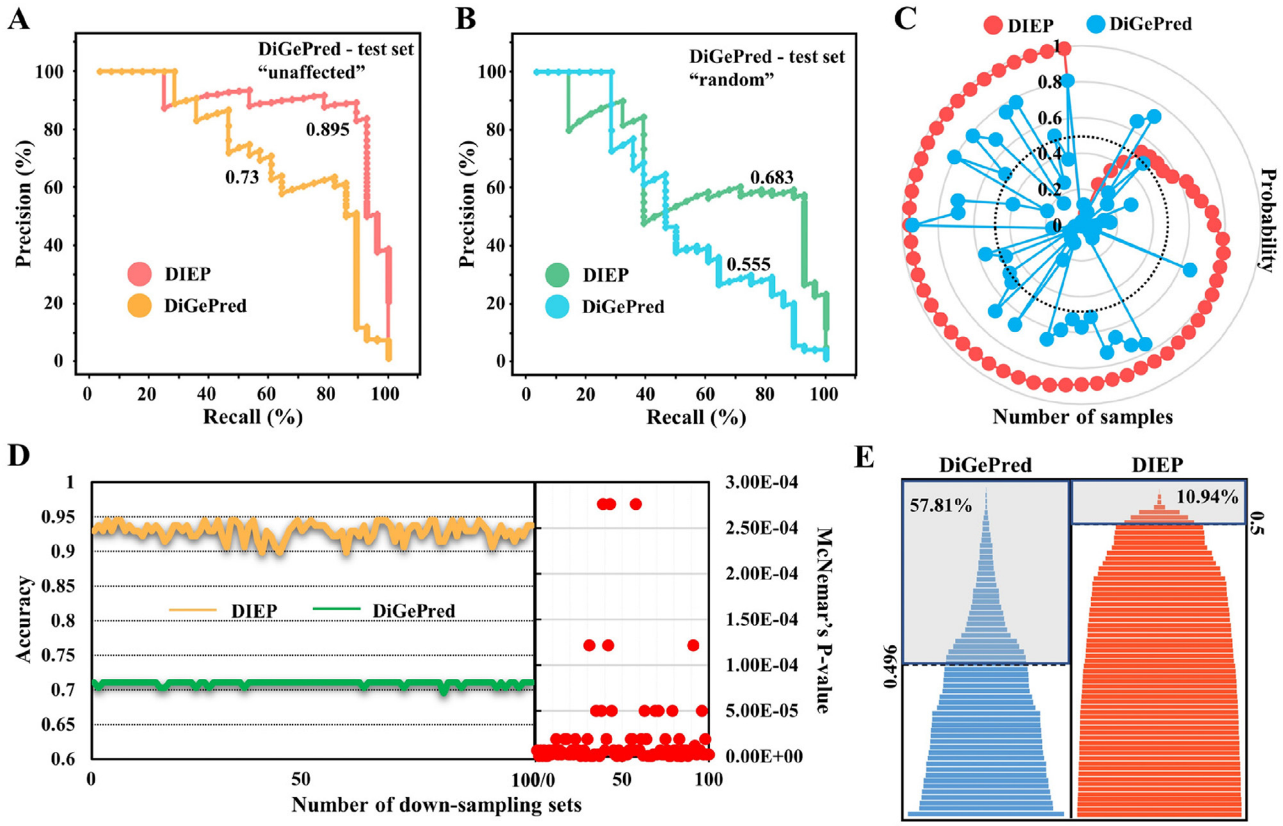


Fig. 6. The detailed comparison between DIEP and DiGePred. (A and B) The PR curves for DIEP and DiGePred on two held-out test sets from DiGePred paper (unaffected- and random-no-gene-overlap). The values near the curves indicate the PR AUCs of different curves. (C) The predicted results of DIEP and DiGePred on the Manual test set. (D) The comparison between DIEP and DiGePred on 100 down-sampling sub-test sets. Each sub-test set contains 128 gene pairs (64 positives from the Manual test set and 64 negatives down-sampled from the Test set). The scatter plot (red dots) indicates the McNemar's p-values of 100 sub-test sets, and the line chart shows the corresponding predicted accuracy of two methods in each set. (E) The distribution of predicted scores of 64 probably digenic gene pairs by DIEP and DiGePred. Each bar represents one gene pair, and the length of each line indicates the value size. The gray area shows the percentage of the wrong predictions of two methods under the best thresholds.

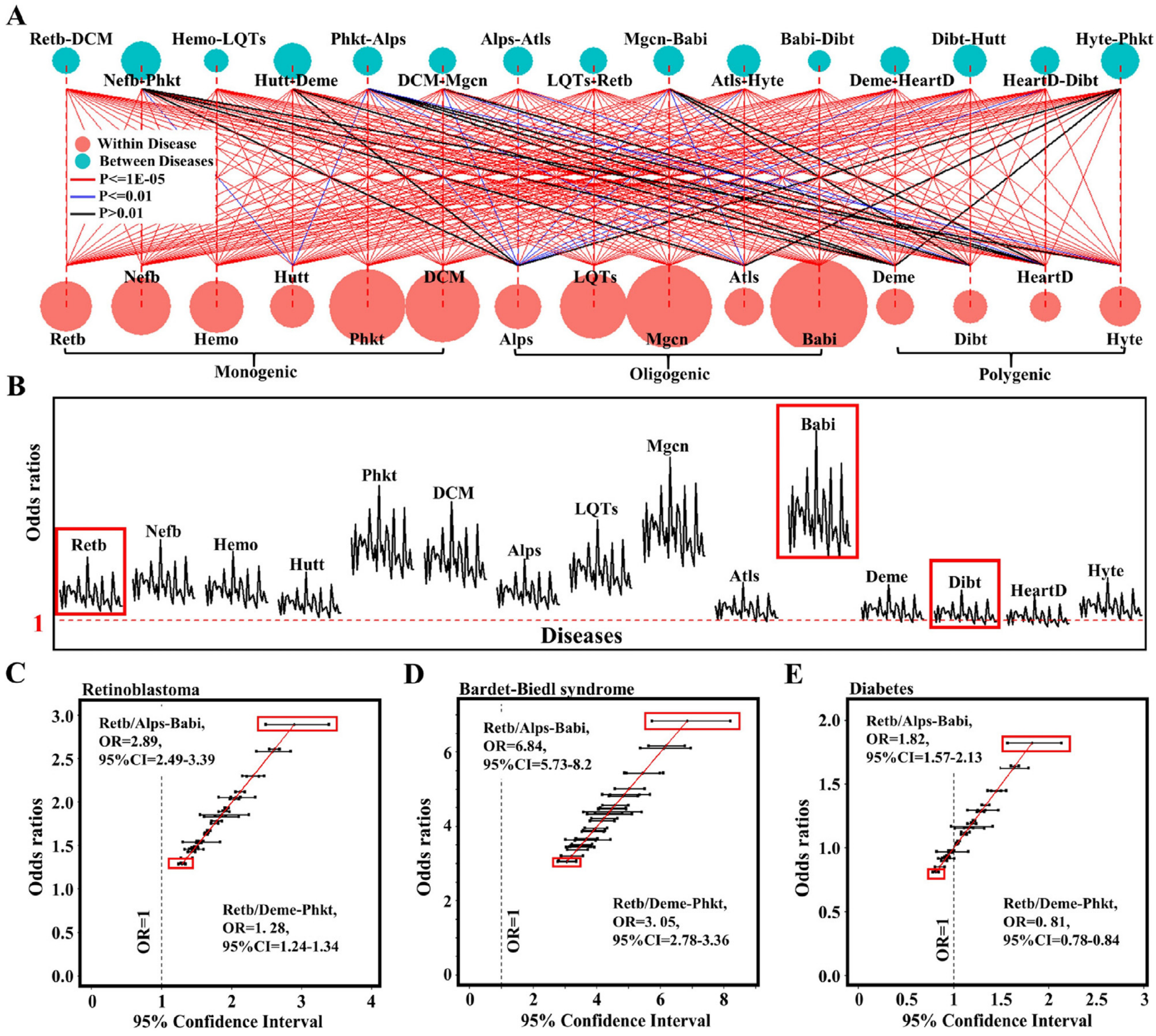


Fig. 7. The statistical result and the enrichment ratio plot for 15 same disease gene pair sets and 35 different diseases gene pair sets. (A) The statistic result between 15 same disease gene pair sets and 15 representative different diseases gene pair sets. The red circle indicates the proportion of digenic gene pairs in which both genes are from the same disease, and the blue one is the proportion of digenic gene pairs in which both genes are from different diseases. The size of the circle depended on the proportion. The red line indicates a significant result with P-value $\leq$ 1E-05, the black line indicates the P-value > 0.05, and the blue line indicates the P-value falls in the middle of two P values. (B) The scatter plot of odds ratios for comparisons between 15 same disease gene pair sets and 35 different diseases gene pair sets. The red dotted line shows the OR threshold of 1. (C) The detailed enrichment values of comparisons between monogenic disease Retinoblastoma Disease and 35 different diseases gene pair sets. Each line indicates one comparison, the length of the line means the 95% confidence interval and the red line link 35 enrichment values. The dotted line shows the Odds ratio = 1. (D) The detailed enrichment values of comparisons between oligogenic disease Bardet-Biedl syndrome and 35 different diseases gene pair sets. (E) The detailed enrichment values of comparisons between polygenic disease Diabetes and 35 different diseases gene pair sets.

syndrome (same disease gene pairs) were 6.84 times more likely to have the digenic interaction effect than those from Alps and Babi (different diseases gene pair) (OR = 6.84, 95%CI = 5.73–8.2) (Fig. 7D). However, the enrichment ratios in Amyotrophic lateral sclerosis and Alport syndrome were not always significantly > 1 (P.adjust-BH < 0.05), in which the underlying reason is unknown. In polygenic diseases, the largest enrichment ratio was only 2.23 ($OR_{Hyte:Alps\text{-}Babi}$ = 2.23) with the 95% CI within [1.91–2.61]. Gene pairs from Diabetes (same disease gene pair), for example, were only 1.82 times more likely with a digenic interaction effect than gene pairs from the Alps and Babi (different diseases gene pair) (OR = 1.82, 95%CI = 1.57–2.13) (Fig. 7E). 8.38% of the enrichment ratios in polygenic diseases were not significantly > 1 (P.adjust-BH < 0.05, Table S8).

4. Discussion

In this study, we demonstrated that the gene level features (e.g., the biological relatedness or similarity) of different genes are highly effective for predicting the digenic interaction effect potential on disease phenotypes. Our DIEP, an accurate knowledge-based prediction model for the digenic interaction effect, achieved excellent performance with a 4.43% FP rate and 99.29% TP rate (Recall) on the whole training set. In addition, DIEP also had low FP rates in 6 independent negative testing datasets and high sensitivity (89.06%) in the manually curated gene pairs with digenic interaction effect in human diseases. The high performance of DIEP was attributed to three factors at least. First, we constructed a reliable and comprehensive benchmark dataset. Except for the acknowledged positive samples in DIDA, the negative datasets were generated by different theoretical assumptions. Among them, the DIDA_NDI set was directly generated by permuting gene pairs in the positive training set. The well-matched positive and negative datasets may reduce many confounding factors, including study-bias and literature-bias factors, and ensure high discrimination at key predictors. Besides, we adopted rare mutation profiles from the 1000 Genomes Project to purify the negative training set further. Second, the down-sampling and bagging-based methods were applied to solve imbalanced samples effectively [67,68] and help alleviate the overfitting problem because the predictor used the whole training set performed poorly, which only had a 42.19% accuracy in the Manual test set. In comparison, our final predictor DIEP ensembled multiple RFs worked better with an 89.06% sensitivity in the Manual test set and relatively high PR in the imbalanced training set (Fig. 3B). Third, we also collected a comprehensive set of predictors to enhance the performance, although some features individually had relatively low importance scores.

It should be noted that the digenic interaction estimated by DIEP is based on the gene-level functional relation *per se*. A positive prediction by DIEP depicts that the two genes in a pair may be functionally alternative, and distortions at both genes are needed to distort a biological function and subsequently cause a disease. Two genes with a digenic interaction effect may not necessarily be pathogenic in a person unless both genes had deleterious mutations. On the other hand, deleterious mutations will not manifest interactive pathogenic effects unless they occur in genes with digenic effects. Recently, a method named VarCoPP [18] has been proposed to estimate combined pathogenicity in two or more genes based on deleteriousness scores (e.g., CADD [69]) of mutations. It showed that different variant combinations in a gene pair could have various scores. We also quickly checked the prediction results by VarCoPP and DIEP (Dataset S3 - "ref-23bi"). In the original test set of VarCoPP, DIEP accurately predicted all the 23 positive gene pairs while VarCoPP gave wrong predictions at three variant pairs. Although it is hard to judge which level of information is more deterministic, this investigation suggested it is important to consider the gene level functions to predict digenic interaction. And we believe that effective integration of variant-level pathogenicity and gene-level interaction information will contribute to a more convincing prediction, and facilitate the genetic mapping of causal variants for real cases.

DIEP outperformed another alternative software, DiGePred, regarding predictive performance and usage efficiency, which addressed the same problem at the gene level. DIEP consistently outperformed DiGePred on both DiGePred's and DIEP's test sets. In the comparison based on the positive set (Manual test set), our DIEP showed much higher sensitivity than DiGePred (89.06% vs. 42.19%). Here are the possible reasons why DIEP is better than DiGePred. First, DIEP had a broader range of input features. The protein–protein association, the functional interaction, and the semantic similarity of gene DO and GO annotations played an important role in the prediction (Fig. 3A). Second, we curated a comprehensive and probably more representative training set and effective data filtering strategies to train the prediction model. However, DiGePred only used the unaffected non-digenic set created from relatives of UDN (UNDIAGNOSED DISEASES NETWORK) [19,70]. Third, we selected feature selection for more reliable input features; besides, down-sampling and bagging-based strategies were adopted for a more robust predictor in the random forest. In addition, we have designed an efficient compression algorithm for the predictive result of whole-exome gene pairs, and provided a high-efficiency java package for searching digenic potential for input genes.

DIEP also had the advantage of being robust to the confusion of gene pairs with bi-locus effects dual molecular diagnosis (DD). The pathogenic digenic interaction effect means that two genes may lead to the same disease phenotype by interaction effect. However, DD is a completely different concept, which means the coincidental independent segregation of two separate disease entities, and each one is caused by variants in separate linked or unlinked genes/loci [7]. Namely, DD refers to the conjunction of two independent monogenic diseases that show simultaneously in one patient [63], indicating that two distinct genes lead to different disease phenotypes. So genetic loci involved in DD segregate independently in most instances [55]. In other research, Versbraegen et al. aimed to classify three different types of bi-locus effects (including true digenic, modifier and dual molecular diagnosis) [63], but they could not distinguish between bi-locus effects and non-bi-locus effects. In contrast, DIEP helped differentiate DD (Fig. S4) from digenic interaction effects, mainly because the predicted model was constructed based on comprehensive gene-based similarity levels in multiple databases. As shown in Table S9, DIEP distinguished DD mainly because the "STRINGPP" votes for the classification of gene pairs in the DD set as a non-digenic class, which means that there is no very strong association between two genes with the DD effect. This is also consistent with the fact that DD indicates two independent genes, as discussed above. However, other features also play a role because some gene pairs with a positive contribution of "STRINGPP" were classified as non-digenic (e.g., DD69 in Table S9).

The feature "STRINGPP" in DIEP plays an important role in the classification. This is expected because the STRING database essentially is a "combined score" computed by combining the probabilities from the different evidence channels to indicate the protein–protein associations. Here, the "association", from a functional perspective, includes both direct physical binding and indirect interaction such as participation in the same metabolic pathway or cellular process as indicated in the publication [42]. Thus, we considered that "STRINGPP" is a comprehensive indicator for the association between genes. However, it should be noted that predictors only using the "STRINGPP" and without "STRINGPP" performed more poorly than DIEP. So there are also other features independent of "STRINGPP" for digenic interaction. Importantly, DIEP also estimated 82,099 gene pairs with STRINGPP values equal to 0 or missing in the original database as digenic pairs, and 2.31% of which even had the predicted digenic potential over 0.9 (Dataset S6).

There are several reasons why enrichment of digenic interaction effects was more conspicuous in Mendelian and oligogenic diseases than in polygenic diseases. First, polygenic diseases have different genetic spectrums from Mendelian and oligogenic disorders. The number of responsible genes for polygenic diseases is usually large, and the effect sizes of genes are small [71]. It is less likely that there are many strong interactions among susceptibility genes for polygenic or complex diseases. On the contrary, most responsible genes for Mendelian or oligogenic diseases have large effects, and the synergy of different genes may have greater impacts on final phenotypes. In addition, our positive gene pairs were collected from severe diseases, which may not represent well for common diseases. Finally, some tested genes may not be true susceptibility genes for complex diseases. It has been noted that it is very difficult to identify genuine causal genes of complex diseases, and many reported genes are indirectly associated with the diseases due to linkage disequilibrium [72].

DIEP will provide a valuable resource of digenic interaction effect for genetic mapping of human diseases. Traditional genetic mapping analyses such as genome-wide linkage and association studies often focus on individual locus or genes. That is because genome-wide interaction will substantially enlarge the number of hypothesis tests and lead to many false-positive findings by chance and an extensive computing burden. Therefore, most gene interaction studies were carried out based on existing knowledge about the diseases [14,73–76], which is rather limited. For example, methods identifying digenic interaction effects such as family-based association studies and Genome-wide association studies focus more on the validating level required to know the candidate gene pairs. However, in practice, especially for rare diseases or polygenic diseases, the disease genes are unknown. Besides, those genes with invisible influence are difficult to be identified alone. DIEP can be used to explore all potential gene pairs with digenic interaction effects. On the one hand, gene pairs with interaction effects will be considered exhaustively without limitation from the knowledge. On the other hand, most majorities of gene pairs without digenic interaction effect will be pre-excluded according to the prediction of DIEP (Dataset S7), which will substantially relieve false positives and computing burden. Remarkably, the whole-genome predicted results were stored as a triple table, which was compressed from ~3.56G to ~366 M with an efficient compression algorithm, providing convenience for researchers. Besides, an effective and efficient method was also accessed for searching the digenic scores for specific gene pairs (https://github.com/pmglab/DIEP).

A limitation of the present study is that the number of positive gene pairs was not large for both training and validation. The positive training gene pairs were only identified in a limited number of diseases, e.g., Bardet-Biedl syndrome, Familial long QT syndrome and Alport syndrome from DIDA. Such training sets may provide biased information for machine learning (or overfitting to some extent), which will make the final predictor inadequate in mapping digenic pairs in polygenic or more complex diseases. The positive test set contained only 64 probably digenic gene pairs, although the difference between the other software (VarCoPP and DiGePred) and DIEP was significant. Besides, Mikhael et al. had successfully applied DiGePred to Mayer–Rokitansky–Küster–Hauser Syndrome and found a likely digenic combination of LAMC1 and MMP14 [77], and our DIEP also discovered it (0.978). Furthermore, Iafusco et al. reported a new case of digenic GCK/HNF1A variants (DIEP predicted probability = 0.946) identified in a hyperglycemic subject. They indicated that identifying mutations in more than one gene will help researchers better understand the genetic cause of the diseases [78]. Therefore, despite the limited number of digenic pairs, DIEP did provide more reliable predicting results with a better performance. Certainly, the upcoming more gene pairs with digenic interaction effect in diseases will help renew the training set and enhance the robustness of the predictor.

Our work has three promising future usages. First, the DIEP can be used with genetic data in local samples to jointly prioritize interactive gene pairs in genetic studies of human diseases. The bioinformatics supported by accurate prediction of DIEP will increase the confidence of declaring a potential pathogenic gene-gene interaction detected in genetic samples. Second, the pre-calculated digenic interaction scores of all the coding gene pairs across the human genome can be used to narrow down the exploratory range in a genome-wide gene-interaction scan of human diseases among genetic samples. For instance, the scan can be carried out only at gene pairs with digenic interaction scores $\geq$ 0.5. As a result, the interactive pairs will decrease from 192,383,920 to 3,940,174 (2.05%), substantially reducing the computing and multiple testing burden. Finally, our genome-wide pre-calculated digenic interaction scores can be used to construct gene-gene interaction networks in the future. Most available genetic interaction networks nowadays are based on systematic screens conducted in animal models, e.g., yeast and Caenorhabditis elegans [1,79], while our scores relied on human diseases and will lead to a genetic interaction network more relevant to abnormal phenotypes in human diseases [80]. A comparative understanding of genetic interaction networks of different species may get insights into some long-standing genetic problems [81].

In conclusion, the DIEP is an accurate and superior model to predict the digenic interaction effect of genes for diseases, which may effectively relieve the dimension burden in genetic mapping to reveal more gene interactions.

CRediT authorship contribution statement

Yangyang Yuan: Data curation, Formal analysis, Methodology, Resources, Investigation, Validation, Visualization, Writing – original draft, Writing – review & editing. **Liubin Zhang:** Formal analysis, Software, Validation, Writing – review & editing. **Qihan Long:** Data curation, Visualization, Writing – review & editing. **Hui Jiang:** Data curation, Writing – review & editing. **Miaoxin Li:** Conceptualization, Funding acquisition, Project administration, Supervision, Validation, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by grants from National Natural Science Foundation of China (31970650 and 32170637), and Guangdong project (2017GC010644). Besides, we would like to thank Prof. Yanni Zeng for the valuable suggestions. We thank the GTEx Consortium and 1000 Genomes Projects for providing access to expression and sequence variant data. And we also want to thank the provider of the DIDA and all the public databases adopted in our manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2022.07.011.

References

[1] Costanzo M et al. A global genetic interaction network maps a wiring diagram of cellular function. Science 2016;353(6306).
[2] Gazzo A et al. Understanding mutational effects in digenic diseases. Nucl Acids Res 2017;45(15):e140.
[3] Nussbaum RL, et al., Patterns of Single-Gene Inheritance, in Thompson &amp Thompson Genetics in Medicine. 2007. p. 115-149.
[4] Comings DE. Polygenic inheritance and micro/minisatellites. Mol Psychiatry 1998;3(1):21–31.
[5] Kuzmin E et al. Systematic analysis of complex genetic interactions. Science 2018;360(6386).
[6] Scriver CR, Waters PJ. Monogenic traits are not simple lessons from phenylketonuria. Trends Genet 1999;15(7):267–72.
[7] Deltas C. Digenic inheritance and genetic modifiers. Clin Genet 2018;93 (3):429–38.
[8] Babar U. Monogenic disorders: an overview. Int J Adv Res 2017;5(2):1398–424.
[9] J.F., R. and K. N, Oligogenic disease, in Vogel and Motulsky's Human Genetics. 2010. p. 211-241.
[10] Badano JL, Katsanis N. Beyond Mendel: an evolving view of human genetic disease transmission. Nat Rev Genet 2002;3(10):779–89.
[11] Gormley P et al. Common variant burden contributes to the familial aggregation of migraine in 1,589 families. Neuron 2018;99(5):1098.
[12] McKinney BA et al. Machine learning for detecting gene-gene interactions: a review. Appl Bioinformatics 2006;5(2):77–88.
[13] Mouton J et al. Ascribing novel functions to the sarcomeric protein, myosin binding protein H (MyBPH) in cardiac sarcomere contraction. Exp Cell Res 2015;331(2):338–51.
[14] Mouton JM et al. MYBPH acts as modifier of cardiac hypertrophy in hypertrophic cardiomyopathy (HCM) patients. Hum Genet 2016;135 (5):477–83.

[15] Corvol H et al. Genome-wide association meta-analysis identifies five modifier loci of lung disease severity in cystic fibrosis. Nat Commun 2015;6:8382.

[16] Oprea GE et al. Plastin 3 is a protective modifier of autosomal recessive spinal muscular atrophy. Science 2008;320(5875):524–7.

[17] Gazzo AM et al. DIDA: a curated and annotated digenic diseases database. Nucl Acids Res 2016;44(D1):D900–7.

[18] Papadimitriou S et al. Predicting disease-causing variant combinations. Proc Natl Acad Sci U S A 2019;116(24):11878–87.

[19] Mukherjee S et al. Identifying digenic disease genes via machine learning in the Undiagnosed Diseases Network. Am J Hum Genet 2021;108(10):1946–63.

[20] Schaffer AA. Digenic inheritance in medical genetics. J Med Genet 2013;50(10):641–52.

[21] Wong SL et al. Combining biological networks to predict genetic interactions. PNAS 2004;101(44):15682–7.

[22] Tweedie S et al. Genenames.org: the HGNC and VGNC resources in 2021. Nucl Acids Res 2021;49(D1):D939–46.

[23] Petrovski S et al. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet 2013;9(8):e1003709.

[24] Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat 2011;32(8):894–9.

[25] Karczewski KJ et al. The ExAC browser: displaying reference data information from over 60 000 exomes. Nucl Acids Res 2017;45(D1):D840–5.

[26] Itan Y et al. The human gene damage index as a gene-level approach to prioritizing exome variants. PNAS 2015;112(44):13615–20.

[27] MacArthur DG et al. A systematic survey of loss-of-function variants in human protein-coding genes. Science 2012;335(6070):823–8.

[28] Georgi B, Voight BF, Bucan M. From mouse to human: evolutionary genomics analysis of human orthologs of essential genes. PLoS Genet 2013;9(5): e1003484.

[29] Zhang W et al. New genes drive the evolution of gene interaction networks in the human and mouse genomes. Genome Biol 2015;16:202.

[30] Khurana E et al. Interpretation of genomic variants using a unified biological network approach. PLoS Comput Biol 2013;9(3).

[31] Ashburner M et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 2000;25(1):25–9.

[32] Yu G et al. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. Bioinformatics 2010;26(7):976–8.

[33] Asif M et al. Identifying disease genes using machine learning and gene functional similarities, assessed through Gene Ontology. PLoS ONE 2018;13(12):e0208626.

[34] Warde-Farley, D., et al., The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucl Acids Res, 2010. 38(Web Server issue): p. W214-20.

[35] Kamburov A et al. ConsensusPathDB–a database for integrating human functional interaction networks. Nucl Acids Res 2009;37(Database issue): D623–8.

[36] Huang N et al. Characterising and predicting haploinsufficiency in the human genome. PLoS Genet 2010;6(10):e1001154.

[37] Itan Y et al. HGCS: an online tool for prioritizing disease-causing gene variants by biological distance. BMC Genomics 2014;15:256.

[38] Lek M et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature 2016;536(7616):285–91.

[39] Fadista J et al. LoFtool: a gene intolerance score based on loss-of-function variants in 60 706 individuals. Bioinformatics 2017;33(4):471–4.

[40] Oughtred R et al. The BioGRID interaction database: 2019 update. Nucl Acids Res 2019;47(D1):D529–41.

[41] Szklarczyk D et al. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucl Acids Res 2019;47(D1):D607–13.

[42] von Mering C et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucl Acids Res 2005;33(Database issue):D433–7.

[43] Croft D et al. The Reactome pathway knowledgebase. Nucl Acids Res 2014;42(Database issue):D472–7.

[44] Mitchell AL et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucl Acids Res 2019;47(D1):D351–60.

[45] El-Gebali S et al. The Pfam protein families database in 2019. Nucl Acids Res 2019;47(D1):D427–32.

[46] Carithers LJ, Moore HM. The Genotype-Tissue Expression (GTEx) Project. Biopreserv Biobank 2015;13(5):307–8.

[47] Wang M et al. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. Proteomics 2015;15(18):3163–8.

[48] Obayashi T et al. COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. Nucl Acids Res 2019;47(D1):D55–62.

[49] Yu G et al. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics 2015;31(4):608–9.

[50] The UniProt C. UniProt: the universal protein knowledgebase. Nucl Acids Res 2017;45(D1):D158–69.

[51] UniProt C. UniProt: a worldwide hub of protein knowledge. Nucl Acids Res 2019;47(D1):D506–15.

[52] Pedregosa F et al. Scikit-learn: machine learning in Python. J Mach Learn Res 2011;12:2825–30.

[53] Genomes Project C et al. A map of human genome variation from population-scale sequencing. Nature 2010;467(7319):1061–73.

[54] Li MX et al. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucl Acids Res 2012;40(7):e53.

[55] Posey JE et al. Resolution of disease phenotypes resulting from multilocus genomic variation. N Engl J Med 2017;376(1):21–31.

[56] Lundberg SM et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2020;2(1):56–67.

[57] Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30 (Nips 2017), 2017. 30.

[58] Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2009.

[59] Team RC. R: A language and environment for statistical computing. msor connections, 2014. 1(1).

[60] Bland JM, Altman DG. Statistics notes – The odds ratio. Br Med J 2000;320(7247):1468.

[61] Benjamini Y, Y.J.J.o.t.R.S.S.S.B.M. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing. 1995. 57(1): p. 289-300.

[62] Rachel RA et al. Combining Cep290 and Mkks ciliopathy alleles in mice rescues sensory defects and restores ciliogenesis. J Clin Invest 2012;122(4):1233–45.

[63] Versbraegen N et al. Using game theory and decision decomposition to effectively discern and characterise bi-locus diseases. Artif Intell Med 2019;99:101690.

[64] Zara F et al. Genetic testing in benign familial epilepsies of the first year of life: clinical and diagnostic significance. Epilepsia 2013;54(3):425–36.

[65] Lindy AS et al. Diagnostic outcomes for genetic testing of 70 genes in 8565 patients with epilepsy and neurodevelopmental disorders. Epilepsia 2018;59(5):1062–71.

[66] Piñero J et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucl Acids Res 2017;45(D1): D833–9.

[67] Mirza B et al. Machine learning and integrative analysis of biomedical big data. Genes (Basel) 2019;10(2).

[68] Chawla NV. Data mining for imbalanced datasets. An Overview 2005:875–86.

[69] Kircher M et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 2014;46(3):310–5.

[70] Gahl WA, Wise AL, Ashley EA. The Undiagnosed Diseases Network of the National Institutes of Health: A National Extension. JAMA 2015;314(17):1797–8.

[71] International Schizophrenia, C., et al., Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature, 2009. 460(7256): p. 748-52.

[72] Shen X, Carlborg O. Beware of risk for increased false positive rates in genome-wide association studies for phenotypic variability. Front Genet 2013;4:93.

[73] Stanke F et al. The CF-modifying gene EHF promotes p.Phe508del-CFTR residual function by altering protein glycosylation and trafficking in epithelial cells. Eur J Hum Genet 2014;22(5):660–6.

[74] Yamamura T et al. Functional splicing analysis in an infantile case of atypical hemolytic uremic syndrome caused by digenic mutations in C3 and MCP genes. J Hum Genet 2018;63(6):755–9.

[75] Timberlake AT et al. Two locus inheritance of non-syndromic midline craniosynostosis via rare SMAD6 and common BMP2 alleles. Elife 2016;5.

[76] Dhungel N et al. Parkinson's disease genes VPS35 and EIF4G1 interact genetically and converge on alpha-synuclein. Neuron 2015;85(1):76–87.

[77] Mikhael S et al. Genetics of agenesis/hypoplasia of the uterus and vagina: narrowing down the number of candidate genes for Mayer-Rokitansky-Kuster-Hauser Syndrome. Hum Genet 2021.

[78] Iafusco F et al. NGS analysis revealed digenic heterozygous GCK and HNF1A variants in a child with mild hyperglycemia: a case report. Diagnostics (Basel) 2021;11(7).

[79] Dixon SJ et al. Systematic mapping of genetic interaction networks. Annu Rev Genet 2009;43:601–25.

[80] Baryshnikova A et al. Genetic interaction networks: toward an understanding of heritability. Annu Rev Genomics Hum Genet 2013;14:111–33.

[81] Boone C, Bussey H, Andrews BJ. Exploring genetic interactions and networks with yeast. Nat Rev Genet 2007;8(6):437–49.