PNAS Nexus, 2023, 2, 1–11
https://doi.org/10.1093/pnasnexus/pgad147
Advance access publication 26 April 2023
Research Report

Evolution-strengthened knowledge graph enables predicting the targetability and druggability of genes

Yuan Quan †, Zhan-Kun Xiong †, Ke-Xin Zhang, Qing-Ye Zhang, Wen Zhang* and Hong-Yu Zhang *

Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan, Hubei 430070, P. R. China
*To whom correspondence should be addressed: Email: zhy630@mail.hzau.edu.cn; zhangwen@mail.hzau.edu.cn
†Y.Q. and Z.-K.X. contributed equally to this work.
Edited By: Shibu Yooseph

Abstract

Identifying promising targets is a critical step in modern drug discovery, with causative genes of diseases that are an important source of successful targets. Previous studies have found that the pathogeneses of various diseases are closely related to the evolutionary events of organisms. Accordingly, evolutionary knowledge can facilitate the prediction of causative genes and further accelerate target identification. With the development of modern biotechnology, massive biomedical data have been accumulated, and knowledge graphs (KGs) have emerged as a powerful approach for integrating and utilizing vast amounts of data. In this study, we constructed an evolution-strengthened knowledge graph (ESKG) and validated applications of ESKG in the identification of causative genes. More importantly, we developed an ESKG-based machine learning model named GraphEvo, which can effectively predict the targetability and the druggability of genes. We further investigated the explainability of the ESKG in druggability prediction by dissecting the evolutionary hallmarks of successful targets. Our study highlights the importance of evolutionary knowledge in biomedical research and demonstrates the potential power of ESKG in promising target identification. The data set of ESKG and the code of GraphEvo can be downloaded from https://github.com/Zhankun-Xiong/GraphEvo.

Keywords: targetability, druggability, evolution, knowledge graph, prediction model construction

Significance Statement

Identifying promising drug targets from tens of thousands of human genes is a critical step in modern drug discovery. To be a promising drug target, a gene must have both targetability and high druggability. Our analyses found that existing successful targets share some critical evolutionary hallmarks, and evolutionary information can facilitate the target prediction. Here, we proposed a concept "evolution-strengthened knowledge graph (ESKG)" and materialized this concept by establishing a data set containing more than 4 million triplets. Furthermore, we developed an ESKG-based machine learning model named GraphEvo, which can effectively predict the targetability and the druggability of genes. Our approach can provide some ideas for target research and help to improve the efficiency of drug research and development.

Introduction

Identifying promising drug targets with high clinical efficacy from tens of thousands of human genes (i.e. predicting the targetability of genes) is a critical step in modern drug discovery (1). It is well known that disease-causing genes are an important source of successful targets. In recent years, a large number of disease-associated genes have been identified based on biological experiments or in silico approaches (2). However, most of these disease-associated genes are not causative genes, and agents that target noncausative genes will lead to clinical inefficiency (3), wasting human and material resources in clinical trials. Therefore, the rational selection of drug targets is one of the effective ways to mitigate risk in preclinical drug discovery (4).

From the perspective of evolutionary medicine, the pathogenesis and development of various diseases (including cancers, neurological diseases, and cardiovascular diseases) are closely related to the evolutionary events of organisms (5–10). Indeed, a series of studies have revealed that evolutionary knowledge can facilitate the interpretation of disease pathogenesis and thus help predict causative genes (11–16). In evolutionary biology, whole-genome duplication (WGD) is generally regarded as an important evolutionary event for vertebrates (17). Genomic studies have found that the ancestral genome of humans experienced two WGD events during the early vertebrate period, and the production of ~30% of protein-coding genes in the human genome is involved in these two events (18–22). The genes generated in these two WGD events were named Ohnologs (22). Due to the

Competing Interest: The authors declare no competing interest.
Received: January 29, 2023. **Accepted:** April 21, 2023

high dosage sensitivity of Ohnolog genes (that is, the variation of gene copy number will have an impact on the phenotype), Ohnolog genes are closely related to the occurrence and development of human diseases (13, 18). In addition to Ohnologs, the evolutionary stages of genes are also associated with some human diseases (8). For example, cancer driver genes are significantly enriched in those originating from cellular organisms, Opisthokonta, and Eumetazoa (14). And drug targets that originated in Eumetazoa are significantly related to the function of neurological disease therapy (23). Previous studies have shown that existing successful drug targets share some common evolutionary hallmarks (i.e. Ohnologs and evolutionary stages) (3, 23–26). Our analyses found that the proportion of targets with direct evolutionary support increases significantly across the drug development pipeline, and selecting evolutionary information–supported targets could double the success rate in clinical development (23). Therefore, evolutionary information could facilitate the targetability prediction of genes.

In addition to the targetability of genes, the druggability is equally important for the identification of a promising drug target. The druggability of existing successful targets varies widely, namely, a few targets are extremely successful and cover 10 or more approved drugs (defined as highly druggable targets in this study), but most targets give birth to three or fewer approved drugs (2). For example, nuclear receptor subfamily 3 group C member 1 (NR3C1) is a representative drug target with high druggability. According to the records of the drug target database SCG-Drug (http://zhanglab.hzau.edu.cn/scgdrug) (3), there are currently more than 60 approved drugs for the treatment of inflammatory diseases targeting NR3C1. Another known highly druggable target is histamine receptor H1 (HRH1), which is targeted by ~50 approved drugs associated with allergic rhinitis and chronic urticaria (3). According to records of SCG-Drug (3), highly druggable targets, which account for 13% of human successful targets, cover more than 60% of approved drugs. The efficient prediction of gene druggability and further identification of highly druggable targets will greatly improve the efficiency of drug development.

Thanks to the development of modern biological technology, the amount of biomedical data has grown exponentially in recent years. As a consequence, determining how to integrate biomedical big data and extract useful knowledge has become an important topic in the field of modern biology. The concept of knowledge graph (KG) was formally proposed by Google in 2012 (27), and its initial roles were to optimize search results, enhance search quality, and improve Google user experiences (27). The KG is a multi-relational graph composed of entities (nodes) and relations (different types of edges). Each edge is represented as a triplet (head entity–relation–tail entity), which enables massive data to be gathered to achieve rapid response and reasoning based on knowledge. Furthermore, the KG facilitates the integration of data from multiple sources, improving the performance of data analysis to a level that cannot be delivered by a single data source or traditional databases.

At present, based on accumulated biomedical big data, several KGs have been successfully constructed and are widely used in the prediction of drug activities, drug side effects, drug–target interactions (DTIs), drug–drug interactions (DDIs), etc. (28–38). For example, Zhang et al. (34) successfully applied an advanced graph neural network (GNN) model to the discovery of anti-COVID-19 drug candidates through the KG construction. Hsieh et al. (35) established a COVID-19–related KG based on interactions among host genes, biological pathways, drugs, and phenotypes and further used the GNN algorithm to predict the drug combinations that may synergically treat COVID-19. Yu et al. (36) proposed a new model, SumGNN, which integrates KG with GNN, and can be used to predict DDIs. Wang et al. (37) constructed a KG4SL model that can be used to the prediction of synthetic lethality genes in human cancers through KG and GNN algorithm, which are an important source of anticancer drug targets. Xiong et al. (38) developed a novel multimodal framework called GraphPK, which integrates information from KG, drug–disease bipartite graph, and biological domain features for improving in silico drug repositioning. The above research indicated that KG construction has a variety of application scenarios and important application value in the field of drug research and development.

In this study, we first constructed an evolutionary knowledge–containing KG named the evolution-strengthened knowledge graph (ESKG). To validate the biological significance of the ESKG, we predicted the causative genes of representative diseases using ESKG and KG embedding models. We further developed a machine learning model named GraphEvo to predict the targetability and the druggability of genes using the ESKG-derived embedding as features. Next, we investigated the explainability of the ESKG in druggability prediction through dissecting the biological hallmarks of targets. Our study highlights the potential of evolutionary knowledge in target research.

Results

Construction and validation of ESKG

In this study, we constructed an evolutionary knowledge–containing KG called ESKG (Fig. 1A). The ESKG not only contained various types of common biological data (such as gene–disease associations, gene–gene interactions, biological processes, subcellular localization of proteins, drug–target associations, and drug–disease associations) but also integrated the evolutionary data (Ohnologs and evolutionary stages) of genes (Fig. 1A). The constructed ESKG involved more than 4 million triplets and 16 kinds of relations, such as disease–disease associations, gene–gene associations, gene–disease associations, and disease–disease associations (Table S1).

Then, we utilized a classical KG embedding model named TransE (39) to learn low-dimensional vector representations (i.e. embeddings) of entities and relations in the ESKG. The embedding representations of the entities and relations learned by the TransE model were visualized by the t-distributed stochastic neighbor embedding (t-SNE) algorithm in our study. The t-SNE reduces high-dimensional vectors to graphical representations in a 2D space, and nearby points have similar embedding representations. Figure 1B–E shows that the embedding representations of the same type of entities or relations in the ESKG have relatively good colocalization in a 2D space, which reflects the effectiveness of the ESKG to a certain extent.

To further validate the biological significance of the ESKG, we applied the ESKG to the identification of causative genes for 19 *kinds* of complex diseases based on TransE (39) (see Materials and methods). Causative genes were collected from DisGeNET (https://www.disgenet.org/) (40). DisGeNET has developed a reliable scoring system for gene–disease associations with scores that range from 0 to 1, where higher scores represent higher confidence in gene–disease associations. The reliable scoring system of DisGeNET has been supported by extensive literature evidence with high confidence (40). In order to ensure the strong correlation and reliability of gene–disease associations, we selected the genes

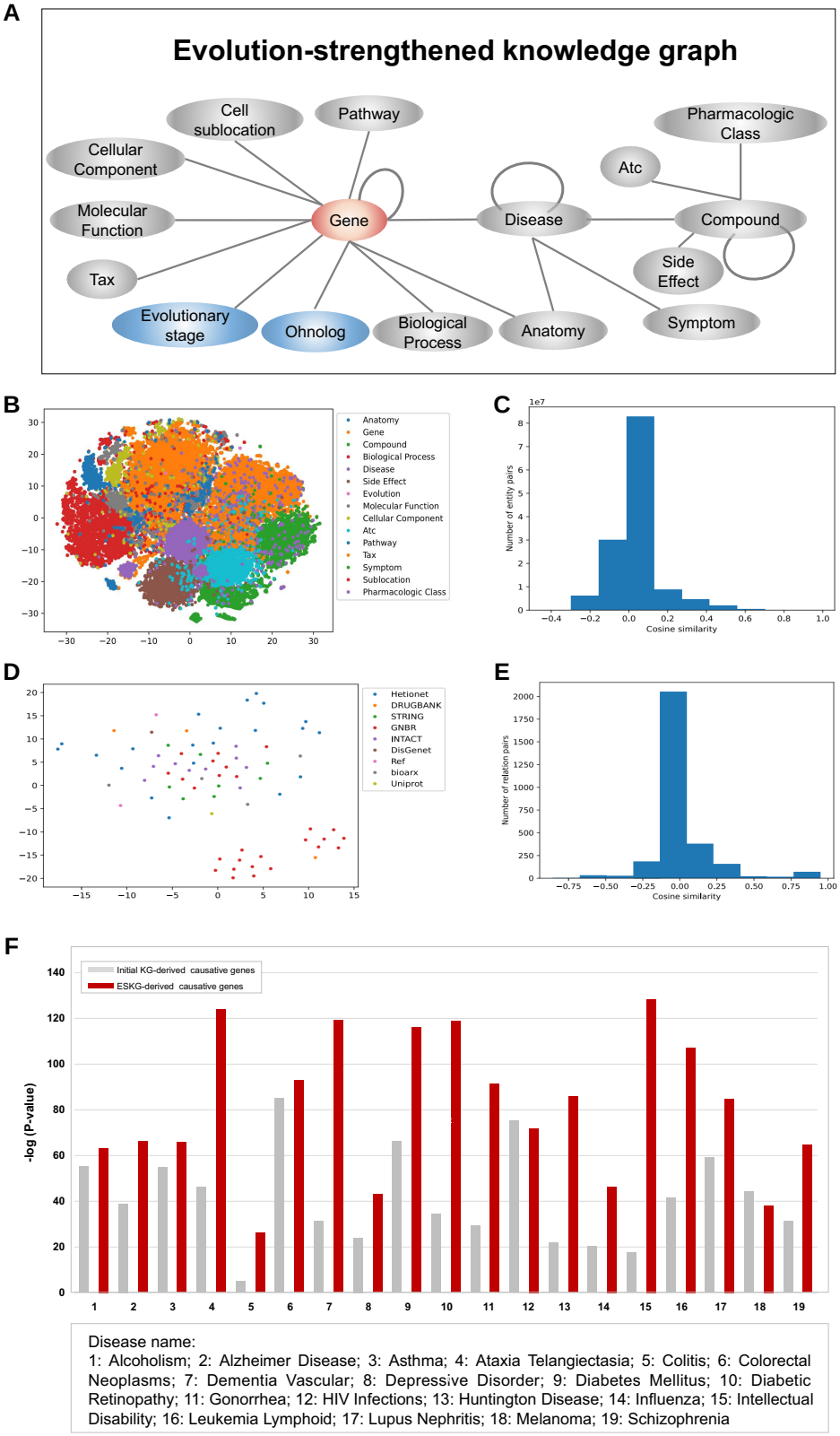

Fig. 1. Construction and validation of the ESKG. A) The types of entities and relations contained in the ESKG. The ESKG not only contains various types of common biological data (such as genes, diseases, biological processes, and drugs) but also integrates the evolutionary data (Ohnologs and evolutionary stages) of genes. B–E) TransE-learned embedding visualization of the entities and relations in the ESKG based on the t-SNE algorithm. Embeddings of the same type of entities or relations in the ESKG have relatively good colocalization in a 2D space, which validates the effectiveness of the ESKG. F) Performance comparison of the ESKG and initial KG in the prediction of causative genes for complex diseases. The results demonstrate that compared with the DRKG-derived initial KG, the ESKG showed superior power in the task of causative gene prediction for 17 of 19 kinds of diseases.

with the highest 10% of DisGeNET scores for each disease as causative genes for the corresponding disease, covering 72,266 gene–disease associations. Of these causative gene–disease associations, 10.41% (7,520 of 72,266) were existing successful targets in SCG-Drug database. This ratio was significantly higher than that of noncausative gene–disease associations included in the DisGeNET (4.26%, 30,942/725,588) ($P = 0$, hypergeometric distribution test). These results indicated that the prediction of causative genes can facilitate the identification of successful drug targets.

In this paper, to demonstrate the efficacy of the ESKG, we filtered out the existing triplets of DisGeNET-derived causative genes and the corresponding diseases in the initial KG and ESKG for each disease. As a result, compared with the initial KG which does not contain evolutionary knowledge (34), the ESKG was more effective in causative gene prediction in 17 *kinds* of these diseases (i.e. alcoholism, Alzheimer's disease [AD], asthma, ataxia telangiectasia, colitis, colorectal neoplasms, dementia vascular, depressive disorder, diabetes mellitus, diabetic retinopathy, gonorrhea, Huntington disease, influenza, intellectual disability, leukemia lymphoid, lupus nephritis, and schizophrenia) (Fig. 1F and Table S2). It is worth noting that the above diseases include two infectious diseases (gonorrhea and influenza). Although the classical strategy against infectious diseases in the past mainly targeted pathogen proteins, host molecule–targeted therapies have gradually become the focus of antiinfectious drug development due to their low drug resistance and broad spectrum (26). Therefore, the ESKG-predicted causative genes related to infectious diseases may be potential targets of antiinfective drugs. We found that several ESKG-predicted causative genes, which received high gene–disease association scores in DisGeNET, were confirmed by previous studies but were overlooked in the initial KG-based prediction.

For example, through the ESKG and TransE model, we predicted that *CCAAT enhancer binding protein alpha* (*CEBPA*) may act as a causative gene of lymphocytic leukemia. CEBPA is a transcription factor that plays a role in cell cycle regulation and granulocyte differentiation (41). A study has shown that the gene mutation and abnormal regulation of *CEBPA* at the transcriptional, translational, and post-translational levels can induce acute myeloid leukemia (42), which is an important branch of lymphocytic leukemia. In addition, we predicted that *caspase-3* (*CASP3*) is a causative gene of AD. Previous studies have demonstrated that the activation of proteins in the caspase family is associated with AD-related neurodegeneration (43). *CASP3*, an important member of the caspase family, has been shown to be a major effector in the apoptotic cascade leading to neuronal apoptosis (44). Type 1 diabetes mellitus (T1DM), also known as autoimmune diabetes, is a chronic disease characterized by the loss of pancreatic beta cells leading to insulin deficiency, which leads to hyperglycemia (45). Using the ESKG and the TransE model, we predicted that *uncoupling protein 2* (*UCP2*) could be a causative gene of T1DM. UCP2 is a transporter of the inner mitochondrial membrane, which is considered to be a key regulator of energy and glucose homeostasis (46). It has been reported that *UCP2* is negative for the regulation of insulin secretion, and its gene polymorphism is also associated with diabetes mellitus and other chronic complications of diabetes (47, 48). There is increasing evidence that drugs can treat diabetes by down-regulating the expression of *UCP2* (49). These results indicate that evolutionary knowledge–containing KG (ESKG) is more effective in predicting causative genes compared with the initial KG.

Prediction of gene targetability by ESKG

To verify the effectiveness of the ESKG in the targetability prediction for human genes, we first applied the ESKG-derived information of genes to construct the targetability prediction model named GraphEvo. In this study, we used ESKG-derived embeddings (learned from TransE) as input features of genes and adopted the ensemble learning algorithm boosting to develop the targetability prediction model (Fig. 2). In the modeling process, we took the target–disease pairs that were marketed by the Food and Drug Administration (FDA) before the year 2000 as positive samples and randomly generated a considerable number of gene–disease pairs without clinical trial records as negative samples. As a result, the area under the receiver operating characteristic (ROC) curve (AUC) of our targetability prediction model reached 0.82 (F1 score = 0.82).

Moreover, we found that the accuracy of gene targetability predicted by GraphEvo increased significantly across the target development pipeline (Table S3). For targets in the preclinical stage, only 24.59% were identified by GraphEvo. For targets in clinical stages, these ratios were 27.14% (phase I), 34.73% (phase II), and 46.95% (phase III). In the approved stage, up to 68.59% of currently successful targets were identified by GraphEvo, indicating that GraphEvo can effectively predict the targetability potential of in-research targets. In addition, based on the sequence, structural, physicochemical, and human system profile information of the targets, Zhu et al. (50) used an in silico method to evaluate the approval potential of 31 targets in phase III clinical trials in 2009, of which 16 targets were predicted to be promising candidates. By 2018, 10 of these 16 (62.5%) promising targets were approved by the FDA (51). In comparison, we used our targetability prediction model (GraphEvo) to analyze the approval potential of these 31 targets. The results of our model showed that seven targets had approval potential, and they all (100%) became approved targets by 2018. The above study demonstrated the application value of ESKG and GraphEvo in targetability prediction.

In addition to predicting the targetability of genes, we also calculated the associated scores between drug targets and diseases using ESKG. Similar to predictions of disease causative genes, we used the TransE model to calculate the associated potential of each successful target-involved triplet (i.e. the successful drug target, disease, and corresponding relation) in the ESKG and used it as the associated score between a drug target and a certain disease category. In this study, we used the three disease categories with the highest associated scores as potential therapeutic activities of successful targets. The results showed that for 468 successful targets, the approved drug activities of 77.78% (364) of the targets are consistent with ESKG-derived disease categories, reflecting the potential of ESKG in the activity prediction of drug targets (Table S4).

Prediction of gene druggability by ESKG

Application of ESKG in gene druggability prediction

GraphEvo could also predict the druggability of genes and further identify highly druggable targets by fusing the features derived from the ESKG and the target–disease graph (TDG) (Fig. 2). Based on King et al.'s (52) and Quan et al.'s (3) data, we obtained 1,536 approved target–disease pairs, covering 468 successful targets, and 114 of 468 (24.36%) had a number of approved drugs greater than or equal to 10, which were defined as highly druggable targets in our paper (Fig. 3A and B and Table S5). First, we used the KG embedding model TransE (39) to extract the embeddings of the ESKG as features of drug targets. In addition, it is obvious that target–disease associations could provide abundant information about the druggability of drug targets. Therefore, using a graph convolutional network (GCN) with the layer attention (53),

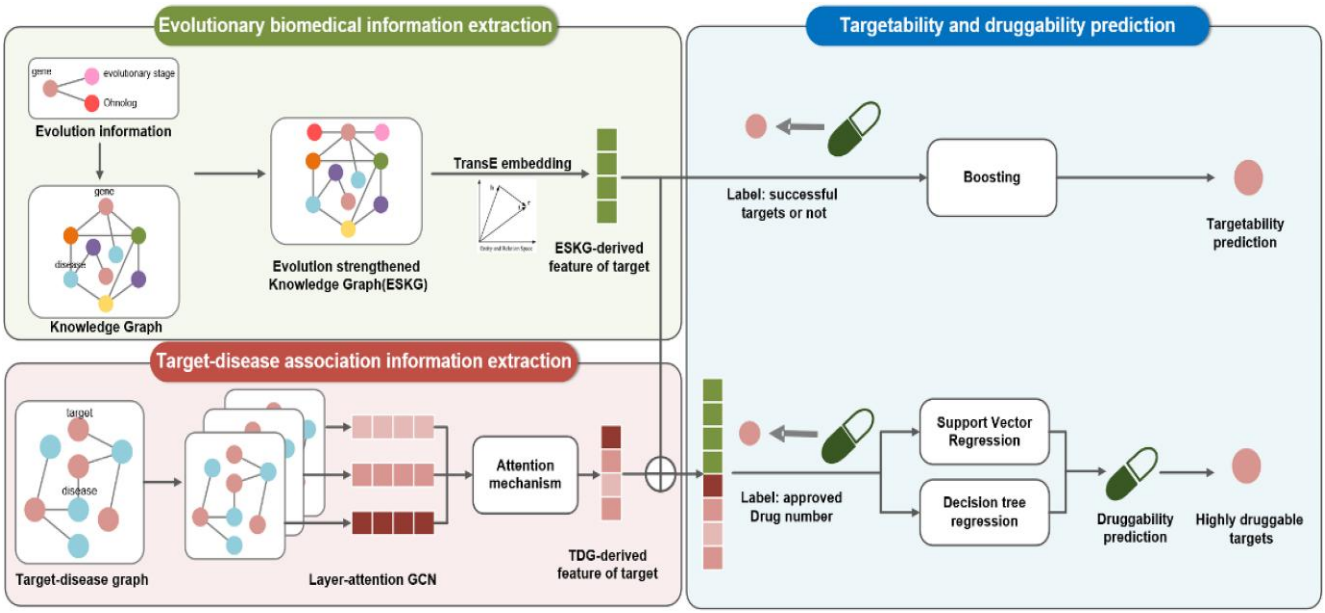

Fig. 2. Construction of the gene targetability and druggability prediction model (GraphEvo). In this study, we used ESKG-derived embeddings (learned from TransE) as input features of genes and adopted the ensemble learning algorithm boosting to develop the targetability prediction model. In the modeling process, we took the target–disease pairs that were marketed by the FDA before the year 2000 as positive samples and randomly generated a considerable number of gene–disease pairs without clinical trial records as negative samples. For a candidate highly druggable target, we concatenated the ESKG-derived features and TDG-derived features as the final features to predict the potential druggability of drug targets, and the druggability (number of approved drugs) of the target was used as the label of the training sample. Then, we utilized the machine learning model of support vector regression and decision tree regression to construct the identification model for highly druggable targets. To obtain robust prediction results, the final predicted scores are the averages of these two types of regression methods.

we also obtained graph-derived features of drug targets from the TDG. Finally, we concatenated the ESKG-derived features and TDG-derived features of drug targets and utilized the machine learning model of support vector regression and decision tree regression to construct an identification model for highly druggable targets (Fig. 2). In detail, our research divided 468 successful targets into training and test sets according to their approval times; that is, drug targets approved before the year 2000 were the training set, and drug targets approved after the year 2000 were the test set. We mainly used two common machine learning evaluation indicators, the AUC and top-*K* recall, to evaluate the effectiveness of this model. GraphEvo achieved an AUC of 0.95, and the top-30 recall was 100% in the test set, which means that our model could recall all 9 highly druggable targets in the top 30 predictions (see Materials and methods).

Moreover, we found that our model could identify highly druggable targets that are likely to be excluded by previous studies. Recently, a risk gene predictive model called integrative RIsk Gene Selector (iRIGS) has been established by integrating multiomic data and gene networks, and ~100 high-confidence risk genes of schizophrenia were predicted based on iRIGS and considered as potential targets (54). However, some highly druggable targets associated with dozens of antischizophrenia drugs, such as dopamine receptor D2 (DRD2) and 5-hydroxytryptamine receptor 1A (HTR1A), were overlooked by iRIGS. It is interesting to notice that both targets (DRD2 and HTR1A) could be successfully predicted by GraphEvo (Table S6).

Explainability of ESKG in gene druggability prediction

The above results demonstrated the satisfying performance of the ESKG-derived model (GraphEvo) in the task of gene druggability prediction. We speculated that an important reason for these results is that highly druggable targets may have some evolutionary hallmarks, which could be extracted from the ESKG. It was found that 78.95% of highly druggable targets are Ohnologs (Fig. 3C and Table S5), which are generated by two WGD events in the early vertebrate lineage and are strongly associated with diseases because of their dosage sensitivity (13, 18). The Ohnolog ratio of highly druggable targets was significantly higher than that of nonhighly druggable targets ($P = 7.40 \times 10^{-8}$, χ^2 test) (Fig. 3C).

In addition to the evolutionary hallmark of Ohnologs, our analysis revealed that highly druggable targets were enriched in genes that originated in the Eumetazoa stage ($P = 1.86 \times 10^{-7}$, hypergeometric distribution test) (Fig. 3D and Table S5). We reasoned that this may be because the approved activities of a high percentage of highly druggable targets belong to the psychiatry and psychology category (Fig. 3E), and the nervous system of organisms first appeared in the Eumetazoa stage according to evolutionary common sense (55). Taken together, our results indicate that compared with targets with fewer successful drugs, highly druggable targets have more distinctive evolutionary hallmarks that have been integrated by the ESKG.

Discussion

The development of new drugs is an expensive, time-consuming, and complex process. The rational selection of drug targets is an effective strategy for reducing the risk in the clinical development of drugs. Therefore, identifying promising drug targets from tens of thousands of human genes is one of the critical and most challenging steps in the modern drug development process. To be a promising drug target, a gene must have both targetability (i.e. potential to be a drug target) and high druggability (i.e. potential to be targeted by a large number of drugs).

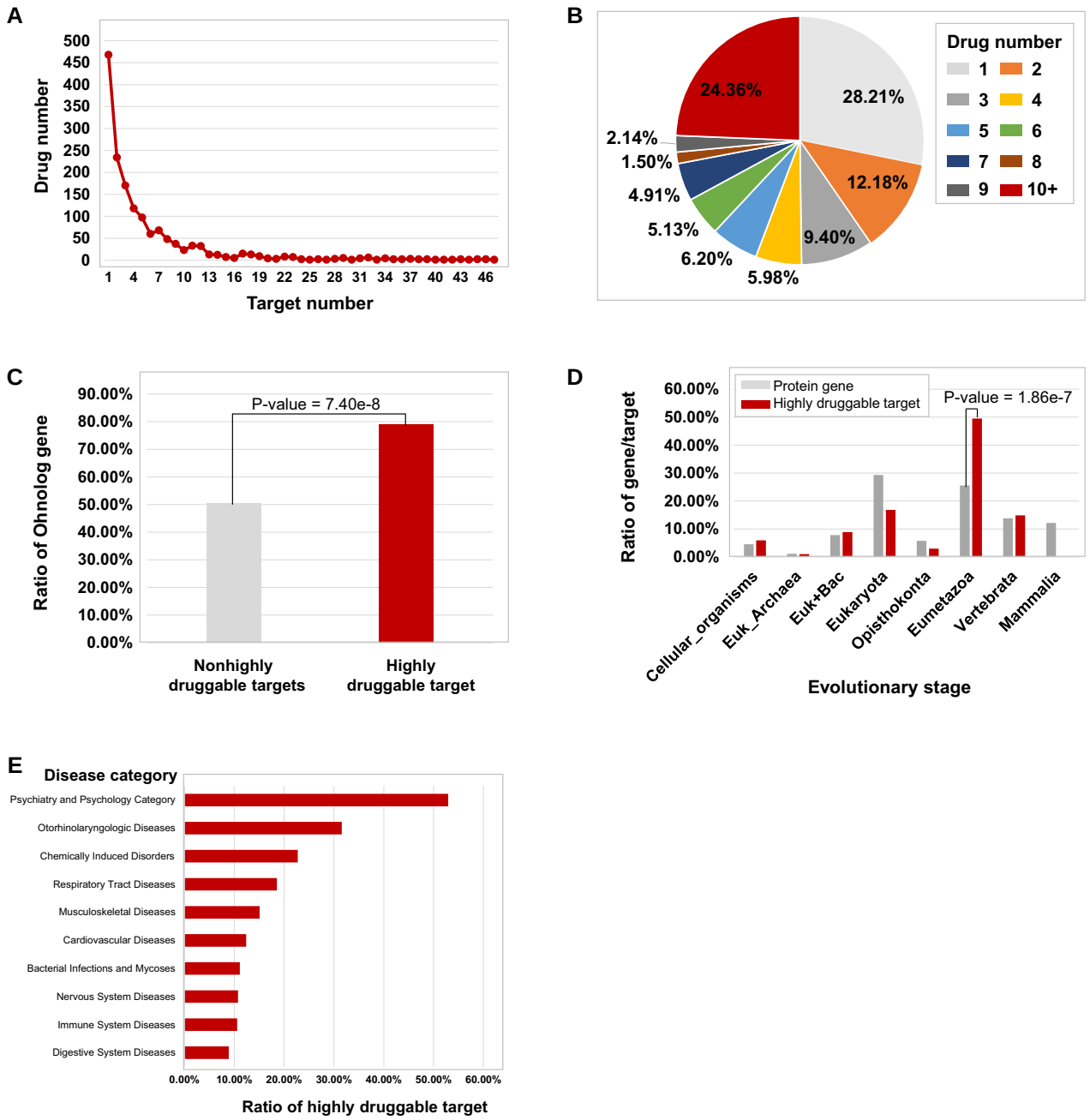

Fig. 3. Evolutionary hallmarks of highly druggable targets. A, B) Distribution of the number of approved drugs corresponding to successful drug targets. Based on King et al.'s (52) and Quan et al.'s (3) data, we obtained 1,536 approved target–disease pairs, covering 468 successful targets. A total of 114 of 468 (24.36%) targets had at least 10 approved drugs, and these were defined as highly druggable targets in this study. C) Comparison of the ratios of Ohnolog genes between highly druggable targets and nonhighly druggable targets. The Ohnolog ratio of highly druggable targets was significantly higher than that of nonhighly druggable targets ($P = 7.40 \times 10^{-8}$, χ^2 test). D) Comparison of the ratios of evolutionary stages between highly druggable targets and other protein-coding genes. The majority of highly druggable targets (49.50%) originated from the Eumetazoa stage ($P = 1.86 \times 10^{-7}$, hypergeometric distribution test). E) The druggability of different disease category–associated drug targets. The results show that more than 50% of psychiatry and psychology category–associated targets are highly druggable targets.

In recent years, benefiting from the development of high-throughput biotechnology, the biomedical field has accumulated massive data, such as genome, transcriptome, proteome, metabolome, and metagenome data. Traditionally, researchers have attempted to integrate these data using advanced computational, bioinformatic, or statistical strategies (including network-based method and machine learning–based method), aiming to predict DTIs or drug–pathway associations and thus discover promising drug targets and potential drugs (56–58). Indeed, rapid progress has been achieved in the biomedical field, such as the identification of thousands of disease-associated genes and dozens of innovative drug targets, enabling the discovery of hundreds of new drugs. However, we are still faced with the dilemma of missing targets and the possibility that no drugs will be available for a large number of complex diseases, resulting in many deaths every year. The traditional method of drug target mining seems to have entered a bottleneck stage. Therefore, the biomedical field urgently needs to incorporate a

new knowledge context in the hope of making breakthroughs in target research.

Several recent studies have established that evolutionary knowledge of genes facilitates the identification of drug targets (28–38). Therefore, we constructed an ESKG and validated the biological significance of this KG through the causative gene prediction of diseases. Then, based on the ESKG, we developed an in silico model (GraphEvo) that could predict the druggability of genes and further identify highly druggable targets. This model does not depend on the protein structures of genes, which greatly expands the range of druggable genes. Finally, we systematically dissected the evolutionary hallmarks of existing successful targets to explain the efficient performance of the ESKG in the druggability prediction of genes.

However, this study has some limitations. First, it is well known that identifying causative genes responsible for complex diseases is an enormous task, which needs to take into account variants contained in genes, patterns of linkage disequilibrium, allele frequencies in different populations, and regulatory effects of those variants, among others. Therefore, using causative genes predicted only at gene level as potential drug targets is biologically insufficient and may lead to undesired therapeutic effects. Nevertheless, our results indicate that ESKG-predicted causative genes are supported by multiple experimental researches, which can validate the biological applicability of ESKG to some extent. Second, the ESKG constructed in this paper is a static KG that only contains data that were accumulated up to a certain point in time. With the continuous updating of biomedical knowledge, the entities in the KG and the relations between entities will continue to change. Previous articles showed that a dynamic graph containing time series data, that is, a graph with a time dimension, can extract knowledge more accurately than a static graph (59). Besides, it is well known that the targetability and druggability of a gene are also related to the 3D structure of its coded protein. Thanks to the rapid progresses in structural biology, such as protein structure prediction by AlphaFold (60), our future research will incorporate the 3D structure information of proteins to bring more effective predictions. In addition to protein targets, a number of studies in recent years have shown that some microRNAs (miRNAs) play crucial roles in human diseases and can be targeted by small-molecule drugs, making the miRNA-based diagnosis and therapeutic target discovery become the focus of drug research and development (61–63). However, to the best of our knowledge, there is no successful miRNA target in the field of drug development. Considering that the construction of prediction model needs to be based on sufficient training data and the biomedical data related to miRNA targets are relatively lacking, the current KG has not yet integrated miRNA-related biomedical knowledge. Integrating the disease-related miRNA information into ESKG and further predicting the targetability and druggability of miRNA are important topics worthy of future research. Notably, it should be pointed out that whether a protein or a miRNA could become a highly druggable target is affected by a variety of conditions, including some nonacademic (such as commercial) factors. This study only investigated the potential of proteins to become highly druggable targets at an academic level.

In summary, we proposed a concept "ESKG" and materialized this concept by establishing a data set containing more than 4 million triplets. ESKG-based machine learning model (GraphEvo) can effectively predict the targetability and druggability of genes, which demonstrates the important value of evolutionary knowledge in biomedicine and is helpful to streamline the drug discovery pipeline. The data set of ESKG and the code of GraphEvo can be downloaded from https://github.com/Zhankun-Xiong/GraphEvo.

Materials and methods

Data sources and preprocessing

Standardization of disease terms

The collection of successful drug targets used in this study was obtained from King et al.'s research (52). The data of approved drugs and drug–target associations were downloaded from the SCG-Drug database (http://zhanglab.hzau.edu.cn/scgdrug) (3). Due to differences in the disease terms used in different sources, we used the Unified Medical Language System (UMLS) to standardize the disease terms of drug targets and indication annotations of drugs. In this study, Medical Subject Headings (MeSH) was selected as the vocabulary source of UMLS, and the MetaMap tool was used to process disease text descriptions to obtain a standardized disease vocabulary (64). This tool uses natural language processing (NLP) and computer linguistics techniques to process input biomedical text descriptions into standardized biomedical texts. The processing results of MetaMap were confirmed in previous studies, and it has been widely used in medical text processing (64). Next, we systematically analyzed the approved drug number distribution of successful drug targets. In this article, successful drug targets covering 10 or more approved drugs are defined as highly druggable targets.

Collection of causative genes of diseases

This study also used MetaMap to standardize disease descriptions of genes from DisGeNET (https://www.disgenet.org/) (40). DisGeNET has developed a reliable scoring system for gene–disease associations with scores that range from 0 to 1, where higher scores represent higher confidence in gene–disease associations. The reliable scoring system of DisGeNET takes into account the number and type of data sources (treatment level and model organism) and the number of publications supporting the gene–disease association (40). It has been supported by extensive literature evidence with high confidence. To further improve the reliability of disease-causative genes, we selected the genes with the highest 10% DisGeNET scores for each disease as causative genes for the corresponding disease. We then obtained 72,266 target–disease pairs that could be further used for KG construction.

Collection of evolutionary data

Information on the evolutionary stages of genes was collected from research published by Liebeskind et al. in 2016 (65). This study integrated predictions of 13 popular homology detection algorithms, greatly reducing the bias caused by single-algorithm prediction (64). In this work, human genes were classified into eight evolutionary stages according to their origin times, and a total of ~18,000 human genes were considered. The numbers of genes involved in the evolutionary stages are as follows: (i) common ancestor of cellular organisms: 812 genes; (ii) common ancestor of Eukaryota and Archaea (Euk_Archaea): 201 genes; (iii) horizontal gene transfer from Bacteria (Euk + Bac): 1,395 genes; (iv) Eukaryota: 5,240 genes; (v) Opisthokonta: 1,030 genes; (vi) Eumetazoa: 4,567 genes; (vii) Vertebrata: 2,469 genes; and (viii) Mammals: 2,180 genes. In addition, the information about Ohnologs was obtained from the study by Makino et al. (18). These data contain 9,057 pairs of Ohnologs involving 7,295 human genome genes.

Construction of the ESKG

Here, we use the newly published drug repurposing knowledge graph (DRKG) (34) as the basis of the ESKG to build the KG. The DRKG is a large-scale comprehensive medical KG constructed by the Amazon Shanghai AI Lab and several scientific research institutions, involving a variety of drug and biomedical data sources (including drug–target associations, drug–disease associations, drug side effects, gene ontologies, gene–disease associations, biological pathways, and biological processes). The DRKG mines data from six public large-scale pharmaceutical databases (DrugBank, Hetionet, GNBR, STRING, IntAct, and DGIdb) and 22 million medical studies and then organizes and normalizes them. The DRKG currently contains ~97,000 entities belonging to 13 entity types and ~5,874,000 triplets of data belonging to 107 relation types. Since an important application of our ESKG is to construct an identification model for highly druggable targets, we removed all triplets of drug–target associations from the DRKG to form the initial KG to verify the performance of this model. Next, we extracted KG triplets from evolutionary data (including Ohnologs and evolutionary stages of genes) and manually corrected the triplets to ensure that each entity or relation was unique in the ESKG. Through the integration of the initial KG and evolutionary data, we obtained a comprehensive ESKG.

Prediction of causative genes

After constructing the ESKG, we counted 4,512 diseases in the graph and standardized the names of these diseases through the NCBI MeSH hierarchical system (https://www.ncbi.nlm.nih.gov/mesh). Then, all diseases can be divided into 24 disease categories, which covered common complex diseases: neoplasms, nervous system diseases, nutritional and metabolic disease categories, immune system diseases, etc. After that, for each disease category, one to two representative diseases (which have sufficient numbers of known causative genes) were selected for the prediction of causative genes, and a total of 19 kinds of diseases (alcoholism, AD, asthma, ataxia telangiectasia, colitis, colorectal neoplasms, dementia vascular, depressive disorder, diabetes mellitus, diabetic retinopathy, gonorrhea, HIV infections, Huntington disease, influenza, intellectual disability, leukemia lymphoid, lupus nephritis, melanoma, and schizophrenia) were obtained. To ensure the reliability and avoid circularity of prediction results, first, we needed to filter out the existing triplets of DisGeNET-derived causative genes and the corresponding diseases in the initial KG and ESKG for each disease. Then, we obtained 19 disease-specific initial KGs and ESKGs (a total of 38 KGs) to predict causative genes and randomly divided the triplets of each KG into a training set (90%), a validation set (5%), and a test set (5%).

Here, we used a classical KG embedding model named TransE (39) to learn the embeddings of genes and diseases. The basic idea of TransE is to describe a triplet (h, r, and t) in the KG as the translation of the head entity (h) and the tail entity (t) in a continuous vector space through the relation (r); that is, $\mathbf{h}+\mathbf{r}\approx\mathbf{t}$. Here, $\mathbf{h}$, $\mathbf{r}$, and $\mathbf{t}$ represent the embeddings of h, r, and t, respectively. To measure the plausibility of the relations, a distance-based scoring function is adopted by TransE:

$$f_r(h,\ t) = -\mathbf{h}+\mathbf{r}-\mathbf{t}_2. \tag{1}$$

$\|\bullet\|_2$ represents the L_2 norm, which was used as the norm of TransE in this study. After scoring, the output scores are then passed to the margin-based ranking loss, which is defined as follows:

$$\mathcal{L} = \sum_{\xi\in\Delta}\sum_{\xi\in\Delta'}[\gamma + f_r(\xi') - f_r(\xi)]_+, \tag{2}$$

where $[x]_+ \triangleq \max(0,\ x)$ and γ denote the margin separating positive triplets and negative triplets. Δ and Δ' denote the positive and negative triplets, respectively. Then, the loss is used by the Adam optimizer to update the embeddings of the entities. Finally, the embeddings for each entity and each relationship type were obtained by TransE. In this study, we implement TransE based on a high-performance KG embedding framework named Deep Graph Library-Knowledge Graph (DGL-KE) (66).

Causative gene prediction could be viewed as a KG completion problem. This problem can be represented as a ranking task, which is essentially the task of learning a prediction function that scores high on true triplets and low on false triplets. For each disease, we calculated edge scores between all genes in the corresponding KG and the disease based on their embeddings by using the following algorithm:

$$d = \gamma - \|\mathbf{h}+\mathbf{r}-\mathbf{t}\|_2\ . \tag{3}$$

$$\text{score} = \text{LogSigmoid}(d) = \log\left(\frac{1}{1+\exp(-d)}\right). \tag{4}$$

Note that here, we use LogSigmoid to make all scores <0, so the higher a score is, the stronger the association between entities. Then, we sorted edge scores in descending order, and the top 1% was predicted to be causative genes for each disease. To compare the performance of the causative gene prediction on ESKG with that on the initial KG, we calculated the number of intersections between the TransE-predicted genes and the DisGeNET-derived causative genes.

Construction of the targetability and druggability prediction model (GraphEvo)

GraphEvo consists of four main components (Fig. 2): (i) evolutionary biomedical feature extraction of genes from the ESKG; (ii) target–disease association feature extraction of genes from the TDG; (iii) construction of the targetability prediction model; and (iv) construction of the druggability prediction model.

Evolutionary biomedical feature extraction of genes from the ESKG

Considering that evolutionary information helps identify highly druggable targets, we constructed an ESKG and used TransE (39) to extract evolutionary information by calculating the embeddings of drug targets in the ESKG. After updating the embeddings of all entities in the ESKG through Eqs. (1) and (2), we obtained the KG embeddings $\mathbf{X}\in\mathbb{R}^{M\times k}$ for all targets, where M and k denote the numbers of targets and the dimensionality of the KG embeddings, respectively. We denote the embedding of the target i as ESKG-derived feature $\mathbf{x}_i$.

Target–disease association feature extraction of genes from the TDG

The known target–disease associations could provide abundant information about the druggability of drug targets, which can enhance the identification of highly druggable targets. We first denote target–disease associations as a binary matrix $\mathbf{A}\in\{0,\ 1\}^{M\times N}$, where M and N denote the numbers of targets and diseases, respectively; $\mathbf{A}_{ij}=1$ if a target t_i interacts with a

disease d_j and $\mathbf{A}_{ij} = 0$, otherwise. Then, we constructed the known TDG defined by the adjacency matrix $\mathbf{A}_H$:

$$\mathbf{A}_H = \begin{bmatrix} 0 & \mathbf{A} \\ \mathbf{A}^T & 0 \end{bmatrix} \in \mathbb{R}^{(M+N)\times(M+N)}. \quad (5)$$

Next, we utilized the GCN, a multilayer connected neural network architecture, to learn the low-dimensional representations of targets, i.e. TDG-derived features. Specifically, given the adjacency matrix $\mathbf{A}_H$ of the known target–disease association graph, the layerwise propagation rule of the GCN is formulated as:

$$\mathbf{H}^{(l+1)} = \sigma\left(\tilde{\mathbf{D}}^{-\frac{1}{2}}\,\tilde{\mathbf{A}}_H\,\tilde{\mathbf{D}}^{-\frac{1}{2}}\,\mathbf{H}^{(l)}\,\mathbf{W}^{(l)}\right), \quad (6)$$

with the initialized embeddings of the nodes as:

$$\mathbf{H}^{(0)} = \begin{bmatrix} 0 & \mathbf{A} \\ \mathbf{A}^T & 0 \end{bmatrix}, \quad (7)$$

where $\tilde{\mathbf{A}}_H = \mathbf{A}_H + \mathbf{I}$ is the adjacency matrix of the target–disease association graph with added self-connections, and $\mathbf{I}$ is the identity matrix; $\tilde{\mathbf{D}}_{ii} = \sum_j \tilde{\mathbf{A}}_{H,ij}$ is the degree matrix of $\tilde{\mathbf{A}}_H$, $\mathbf{W}^{(l)}$ is a layer-specific trainable weight matrix, and $\sigma(\cdot)$ denotes an activation function; $\mathbf{H}^{(l)} \in \mathbb{R}^{(M+N)\times k}$ is the node embeddings at the lth layer, and k is the dimensionality of the embeddings. After L layers, we obtained L k-dimensional embeddings from different graph convolution layers. Since the embeddings at different layers capture different structural information in the association graph, we utilized an attention mechanism to combine the embeddings of different layers and obtained the final embeddings of targets and diseases as $\begin{bmatrix} \mathbf{H}_T \\ \mathbf{H}_D \end{bmatrix} = \sum a_l \mathbf{H}^l$, where $\mathbf{H}_T \in \mathbb{R}^{M\times k}$ and $\mathbf{H}_D \in \mathbb{R}^{N\times k}$ are the final embeddings of targets and diseases, respectively; a_l is autolearned by the neural networks and initialized as $\frac{1}{(l+1)}$, $l = 1, 2, \ldots, L$. The learning objective of the feature extraction is to reconstruct the known target–disease associations. Concretely, the reconstruction of the known target–disease associations is defined as the generalized inner product of the target and disease final representations:

$$\tilde{\mathbf{Y}} = \mathbf{H}_T\,\mathbf{W}\,\mathbf{H}_D^T, \quad (8)$$

where $\mathbf{W} \in \mathbb{R}^{k\times k}$ is a trainable matrix. Then, we adopted the focal loss (67) as the loss function to calculate the loss between $\tilde{\mathbf{Y}}$ and $\mathbf{A}$. Next, we utilized the Adam optimizer to minimize the loss function and updated the embeddings of drugs and targets. Finally, $\mathbf{H}_T$ is used as the TDG-derived feature. Specifically, we denote the TDG-derived feature for the target i as $\mathbf{h}_i$.

Construction of the targetability prediction model

In this study, we used the evolutionary biomedical features extracted from the ESKG as input features of genes and adopted the ensemble learning algorithm *boosting* to develop the targetability prediction model. In the modeling process, we took the target–disease pairs that were marketed by the FDA before the year 2000 as positive samples and randomly generated a considerable number of gene–disease pairs without clinical trial records as negative samples.

Construction of the druggability prediction model

For a candidate highly druggable target i, we integrated its evolutionary biomedical information and target–disease information by concatenating the ESKG-derived features $\mathbf{x}_i$ and TDG-derived features $\mathbf{h}_i$ as the final features $\mathbf{h}_{\text{final}}$ to predict the potential druggability of the drug target (Fig. 2).

$$\mathbf{h}_{\text{final}} = \mathbf{x}_i || \mathbf{h}_i, \quad (9)$$

where || represents vector concatenation operation. In the process of model construction, the druggability (number of approved drugs) of the target was used as the label of the training sample, and we utilized the machine learning model of support vector regression and decision tree regression to construct the identification model for the highly druggable target. To obtain robust prediction results, the final predicted scores are the averages of these two types of regression methods. For the target i, we obtained the druggability score as follows:

$$\text{druggability\_score} = \frac{\text{SVR}(\mathbf{h}_{\text{final}}) + \text{DTR}(\mathbf{h}_{\text{final}})}{2}. \quad (10)$$

Finally, the bool score of whether the target is a highly druggable target is defined as:

$$\text{priviledged\_score} = \begin{cases} 0 \text{ if druggability\_score} < \text{threshold} \\ 1 \text{ if druggability\_score} \geq \text{threshold} \end{cases}.$$

Supplementary material

Supplementary material is available at *PNAS Nexus* online.

Funding

This work was funded by the National Natural Science Foundation of China (grant 31870837 and grant 62072206) and the Fundamental Research Funds for the Central Universities (2662021JC008 and 2662021XXQD001).

Author contributions

Conceptualization: H.Y.-Z. and Y.Q.; methodology: Y.Q., Z.-K.X., K.-X.Z., Q.Y.-Z., W.Z., and H.Y.-Z.; writing (original draft): Y.Q., Z.-K.X., and K.-X.Z.; and writing (review and editing): W.Z. and H.Y.-Z.

Data availability

All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary material. The code is available at https://github.com/Zhankun-Xiong/GraphEvo.

References

1 Emmerich C-H, *et al.* 2021. Improving target assessment in biomedical research: the GOT-IT recommendations. *Nat Rev Drug Discov*. 20:64–81.
2 Gates A-J, Gysi D-M, Kellis M, Barabási A-L. 2021. A wealth of discovery built on the human genome project—by the numbers. *Nature* 590:212–215.
3 Quan Y, *et al.* 2019. Systems chemical genetics-based drug discovery: prioritizing agents targeting multiple/reliable disease-associated genes as drug candidates. *Front Genet*. 10:474.
4 Dahlin J-L, Inglese J, Walters MA. 2015. Mitigating risk in academic preclinical drug discovery. *Nat Rev Drug Discov*. 14:279–294.
5 Benton M-L, *et al.* 2021. The influence of evolutionary history on human health and disease. *Nat Rev Genet*. 22:269–283.
6 Perry G-H. 2021. Evolutionary medicine. *eLife* 10:e69398.

7 Bull J-J, Barrick J-E. 2017. Arresting evolution. *Trends Genet*. 33: 910–920.
8 Domazet-Loso T, Tautz D. 2010. Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC Biol*. 8:66.
9 Chen H, Lin F, Xing K, He X. 2015. The reverse evolution from multicellularity to unicellularity during carcinogenesis. *Nat Commun*. 6:6367.
10 Stearns S-C, Nesse R-M, Govindaraju D-R, Ellison P-T. 2010. Evolution in health and medicine Sackler colloquium: evolutionary perspectives on health and medicine. *Proc Natl Acad Sci U S A*. 107:1691–1695.
11 Wu H, Ma B-G, Zhao J-T, Zhang H-Y. 2007. How similar are amino acid mutations in human genetic diseases and evolution. *Biochem Biophys Res Commun*. 362:233–237.
12 Zhang H-Y, Chen L-L, Li X-J, Zhang J. 2010. Evolutionary inspirations for drug discovery. *Trends Pharmacol Sci*. 31:443–448.
13 Xie T, Yang Q-Y, Wang X-T, McLysaght A, Zhang HY. 2016. Spatial colocalization of human Ohnolog pairs acts to maintain dosage-balance. *Mol Biol Evol*. 33:2368–2375.
14 Chu X-Y, Jiang L-H, Zhou X-H, Cui Z-J, Zhang H-Y. 2017. Evolutionary origins of cancer driver genes and implications for cancer prognosis. *Genes (Basel)* 8:182.
15 Chu X-Y, Quan Y, Zhang H-Y. 2020. Human accelerated genome regions with value in medical genetics and drug discovery. *Drug Discov Today*. 25:821–827.
16 Quan Y, Zhang K-X, Zhang H-Y. 2023. The gut microbiota links disease to human genome evolution. *Trends Genet*. S0168–9525: 00032–X.
17 Altenhoff A-M, *et al.* 2016. Standardized benchmarking in the quest for orthologs. *Nat Methods*. 13:425–430.
18 Makino T, McLysaght A. 2010. Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci U S A*. 107:9270–9274.
19 Dehal P, Boore J-L. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol*. 3:e314.
20 McLysaght A, Hokamp K, Wolfe K-H. 2002. Extensive genomic duplication during early chordate evolution. *Nat Genet*. 31: 200–204.
21 Nakatani Y, Takeda H, Kohara Y, Morishita S. 2007. Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res*. 17:1254–1265.
22 Ohno S, Wolf U, Atkin N-B. 1968. Evolution from fish to mammals by gene duplication. *Hereditas* 59:169–187.
23 Quan Y, Wang Z-Y, Chu X-Y, Zhang H-Y. 2018. Evolutionary and genetic features of drug targets. *Med Res Rev*. 38:1536–1549.
24 Wang Z-Y, Fu L-Y, Zhang H-Y. 2012. Can medical genetics and evolutionary biology inspire drug target identification? *Trends Mol Med*. 18:69–71.
25 Tong X-Y, Quan Y, Zhang H-Y. 2021. NUDT5 as a novel drug target and prognostic biomarker for ER-positive breast cancer. *Drug Discov Today*. 26:620–625.
26 Xu X, *et al.* 2021. Facilitating antiviral drug discovery using genetic and evolutionary knowledge. *Viruses* 13:2117.
27 Singhal A. 2012. Introducing the knowledge graph: things, not strings. *Official Google Blog*.
28 Nicholson D-N, Greene C-S. 2020. Constructing knowledge graphs and their biomedical applications. *Comput Struct Biotechnol J*. 18:1414–1428.
29 Zeng X, *et al.* 2020. Repurpose open data to discover therapeutics for COVID-19 using deep learning. *J Proteome Res*. 19:4624–4636.
30 Zheng S, *et al.* 2021. PharmKG: a dedicated knowledge graph benchmark for bomedical data mining. *Brief Bioinform*. 22: bbaa344.
31 MacLean F. 2021. Knowledge graphs and their applications in drug discovery. *Expert Opin Drug Discov*. 16:1057–1069.
32 Geleta D, *et al.* 2021. Biological insights knowledge graph: an integrated knowledge graph to support drug development. *bioRxiv* 10.28.466262.
33 Wang H, Huang F, Xiong Z, Zhang W. 2022. A heterogeneous network-based method with attentive meta-path extraction for predicting drug–target interactions. *Brief Bioinform*. 23:bbac184.
34 Zhang R, *et al.* 2021. Drug repurposing for COVID-19 via knowledge graph completion. *J Biomed Inform*. 115:103696.
35 Hsieh K, *et al.* 2021. Drug repurposing for COVID-19 using graph neural network and harmonizing multiple evidence. *Sci Rep*. 11: 23179.
36 Yu Y, *et al.* 2021. SumGNN: multi-typed drug interaction prediction via efficient knowledge graph summarization. *Bioinformatics* 37:2988–2995.
37 Wang S, *et al.* 2021. KG4SL: knowledge graph neural network for synthetic lethality prediction in human cancers. *Bioinformatics* 37:i418–i425.
38 Xiong Z, Huang F, Wang Z, Liu S, Zhang W. 2022. A multimodal framework for improving *in silico* drug repositioning with the prior knowledge from knowledge graphs. *IEEE/ACM Trans Comput Biol Bioinform*. 19:2623–2631.
39 Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. In Proceedings of the 26th International Conference on Neural Information Processing Systems 2, 2787–2795 (2013).
40 Piñero J, *et al.* 2020. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Res*. 48:D845–D855.
41 Szmajda D, Krygier A, Jamroziak K, Zebrowska-Nawrocka M, Balcerczak E. 2019. Expression level of CEBPA gene in acute lymphoblastic leukemia individuals. *Sci Rep*. 9:15640.
42 Green C-L, *et al.* 2010. Prognostic significance of *CEBPA* mutations in a large cohort of younger adult patients with acute myeloid leukemia: impact of double *CEBPA* mutations and the interaction with *FLT3* and *NPM1* mutations. *J Clin Oncol*. 28:2739–2747.
43 Chu J, Lauretti E, Praticò D. 2017. Caspase-3-dependent cleavage of Akt modulates tau phosphorylation via GSK3β kinase: implications for Alzheimer's disease. *Mol Psychiatry*. 22:1002–1008.
44 Louneva N, *et al.* 2008. Caspase-3 is enriched in postsynaptic densities and increased in Alzheimer's disease. *Am J Pathol*. 173: 1488–1495.
45 Katsarou A, *et al.* 2017. Type 1 diabetes mellitus. *Nat Rev Dis Primers*. 3:17016.
46 Lee H-J, *et al.* 2008. Associations between polymorphisms in the mitochondrial uncoupling proteins (UCPs) with T2DM. *Clin Chim Acta*. 398:27–33.
47 Liu J, Li J, Li WJ, Wang CM. 2013. The role of uncoupling proteins in diabetes mellitus. *JDiabetes Res*. 2013:585897.
48 Rudofsky G Jr, *et al.* 2006. Functional polymorphisms of *UCP2* and *UCP3* are associated with a reduced prevalence of diabetic neuropathy in patients with type 1 diabetes. *Diabetes Care* 29:89–94.
49 Sun L-L, *et al.* 2011. MicroRNA-15a positively regulates insulin synthesis by inhibiting uncoupling protein-2 expression. *Diabetes Res Clin Pract*. 91:94–100.
50 Zhu F, *et al.* 2009. What are next generation innovative therapeutic targets? Clues from genetic, structural, physicochemical, and systems profiles of successful targets. *J Pharmacol Exp Ther*. 330:304–315.

51 Zhu F, Li X-X, Yang S-Y, Chen Y-Z. 2018. Clinical success of drug targets prospectively predicted by in silico study. *Trends Pharmacol Sci*. 39:229–231.

52 King E-A, Davis J-W, Degner J-F. 2019. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS Genet*. 15:e1008489.

53 Yu Z, Huang F, Zhao X, Xiao W, Zhang W. 2021. Predicting drug–disease associations through layer attention graph convolutional network. *Brief Bioinform*. 22:bbaa243.

54 Wang Q, *et al*. 2019. A Bayesian framework that integrates multi-omics data and gene networks predicts risk genes from schizophrenia GWAS data. *Nat Neurosci*. 22:691–699.

55 Miller G. 2009. On the origin of the nervous system. *Science* 325-(5936):24–26.

56 Chen X, *et al*. 2016. Drug–target interaction prediction: databases, web servers and computational models. *Brief Bioinform*. 17: 696–712.

57 Wang C-C, Zhao Y, Chen X. 2021. Drug-pathway association prediction: from experimental results to computational models. *Brief Bioinform*. 22:bbaa061.

58 Chu Z, *et al*. 2022. Hierarchical graph representation learning for the prediction of drug–target binding affinity. *Inf Sci (Ny)*. 613: 507–523.

59 Pareja A, *et al*. 2020. EvolveGCN: evolving graph convolutional networks for dynamic graphs. *Proceedings of the AAAI Conference on Artificial Intelligence* 34:5363–5370.

60 Jumper J, *et al*. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596:583–589.

61 Chen X, Guan N-N, Sun Y-Z, Li J-Q, Qu J. 2020. MicroRNA-small molecule association identification: from experimental results to computational models. *Brief Bioinform*. 21(1):47-61.

62 Chen X, Zhou C, Wang C-C, Zhao Y. 2021. Predicting potential small molecule-miRNA associations based on bounded nuclear norm regularization. *Brief Bioinform*. 22:bbab328.

63 Wang C-C, Zhu C-C, Chen X. 2022. Ensemble of kernel ridge regression-based small molecule-miRNA association prediction in human disease. *Brief Bioinform*. 23:bbab431.

64 Aronson A-R. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*. 2001:17–21.

65 Liebeskind B-J, McWhite C-D, Marcotte E-M. 2016. Towards consensus gene ages. *Genome Biol Evol*. 8:1812–1823.

66 Wang M-J, *et al*. 2019. Deep graph library: a graph-centric, highly-performant package for graph neural networks. *arXiv* 1909.01315.

67 Lin T-Y, Goyal P, Girshick R, He K, Dollar P. Focal loss for dense object detection. In IEEE International Conference on Computer Vision (ICCV), 2999–3007 (2017).