Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

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

Database Article

GETdb: A comprehensive database for genetic and evolutionary features of drug targets

Qi Zhang [a,1], Yang He [a,1], Ya-Ping Lu [b,c], Qi-Hao Wei [d], Hong-Yu Zhang [a], Yuan Quan [a,*]

[a] Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan 430070, PR China
[b] Sinopharm Genomics Technology Co., Ltd., Wuhan 430030, PR China
[c] Sinopharm Medical Laboratory (Wuhan) Co., Ltd., Wuhan 430030, PR China
[d] Sinopharm (Wuhan) Precision Medical Technology Co., Ltd., Wuhan 430030, PR China

ARTICLE INFO

Keywords:
Drug target
Genetic feature
Evolutionary feature
Comprehensive database
Prediction model

ABSTRACT

The development of an innovative drug is complex and time-consuming, and the drug target identification is one of the critical steps in drug discovery process. Effective and accurate identification of drug targets can accelerate the drug development process. According to previous research, evolutionary and genetic information of genes has been found to facilitate the identification of approved drug targets. In addition, allosteric proteins have great potential as targets due to their structural diversity. However, this information that could facilitate target identification has not been collated in existing drug target databases. Here, we construct a comprehensive drug target database named Genetic and Evolutionary features of drug Targets database (GETdb, http://zhanglab.hzau.edu.cn/GETdb/page/index.jsp). This database not only integrates and standardizes data from dozens of commonly used drug and target databases, but also innovatively includes the genetic and evolutionary information of targets. Moreover, this database features an effective allosteric protein prediction model. GETdb contains approximately 4000 targets and over 29,000 drugs, and is a user-friendly database for searching, browsing and downloading data to facilitate the development of novel targets.

1. Introduction

Drug discovery is a time-consuming, expensive and risky process [1]. According to an analysis of innovative drugs approved by the US Food and Drug Administration (FDA) from 2010 to 2020, the typical clinical development time for innovative drugs is 9.1 years [2]. The average cost of developing a new drug has been the subject of debate, with recent estimates ranging from $314 million to $2.8 billion [3]. Although substantial investment and time are devoted to the discovery of new drugs, clinical trials have a success rate of a mere 13% and a relatively elevated rate of drug attrition. [4]. The appropriate selection of a target is an effective way to reduce the risk and cost of clinical development of drugs [5]. The discovery of novel drug targets is one of the main focuses of biomedical research in the pharmaceutical industry and academia, providing the basis for the development of new drugs [6]. The number of marketed drugs for different targets (i.e. target's druggability) is non-uniformly distributed [7]. In 2015, it was observed that privileged target families, comprising 44% of the FDA-approved distinct human protein efficacy targets (667 in total), accounted for 70% of the therapeutic effects attributed to small molecule drugs [8]. This highlights the potential for identifying high potency targets that could lead to the discovery of multiple new drugs. Traditionally, the identification and validation of drug targets have been performed via three experimental techniques: nucleic acid microarrays, protein microarrays, and high-throughput Nuclear Magnetic Resonance (NMR)-based screening for drug-target interactions. However, these experimental approaches for drug target identification are characterized with both significantly high cost and time investments [9]. Thus, it is imperative to utilize target identification methodologies that are more accurate and efficient, due to their more reliable and effective support for drug development and individualized therapy [10].

Genetics is a method to explore the association between genotype and phenotype, which can help researchers identify the causative genes for diseases [11]. Causative genes underlying genetic disorders are frequently preferred targets for drug discovery and development, given their pivotal role in disease initiation and progression [12]. For example,

[*] Corresponding author.
E-mail address: quanyuan@mail.hzau.edu.cn (Y. Quan).
[1] These authors contributed equally to this work

https://doi.org/10.1016/j.csbj.2024.04.006
Received 7 November 2023; Received in revised form 25 March 2024; Accepted 1 April 2024
Available online 3 April 2024

in various cancers, mutations are observed in isocitrate dehydrogenase (IDH), an important enzyme involved in cellular metabolism. These mutations lead to the excessive production of 2-hydroxyglutarate, an oncogenic metabolite. Consequently, targeting IDH mutations has become an important strategy in tumor therapy [13]. Nuclear factor (erythroid-derived 2) 2 (*NRF2*) upregulation can counteract the increase of hemodynamic stress and protect the cardiovascular system, making it a potential target for cardiovascular disease treatment [14,15]. Superoxide dismutase 1 (*SOD1*) is an enzyme that removes harmful free radicals inside cells, however, *SOD1* mutations cause amyotrophic lateral sclerosis (ALS), a neurodegenerative disease, therefore, *SOD1* is regarded as a potential target for ALS treatment [16]. It has been reported that drug targets with human genetic support have twice the likelihood of being approved than those without support [17]. Among the 50 drug targets approved by FDA in 2021, two-thirds have human genetic evidence [18]. Based on genetic information, determining the function or role of gene products in normal physiology and pathogenic processes can help select appropriate and effective targets [19].

In addition to genetic information, our previous research has revealed that successful targets tend to share some similar evolutionary features, and evolutionary information can help identify drug targets with the greatest potential for therapeutic development [20]. According to a relatively consistent gene age data provided by Liebeskind et al. [21], genes can be categorized into eight evolutionary stages, which are the common ancestor of Cellular organisms, the common ancestor of Eukaryotes and Archaea (Euk_Archaea), Horizontal gene transfer from Bacteria (Euk + Bac), Eukaryota, Opisthokonta, Eumetazoa, Vertebrata and Mammals. By the year 2012, a noteworthy discovery was made regarding the identification of 498 successful targets. It was found that these targets significantly enriched in the common ancestor of cellular life and Euk + Bac [20], indicating that targets from these evolutionary stages are more likely to be successful compared to those from other stages. Enrichment of 581 cancer driver genes using consensus gene age data revealed significant enrichment of genes from eukaryotic, opisthokonta and eumetazoa [22]. Our previous study also analyzed the evolutionary origins of 36 human antiviral targets. Statistical results of the evolutionary information showed that the 36 targets were mainly distributed in eumetazoa ($p = 4.00 \times 10^{-2}$, hypergeometric test), and 21 of them were cellular membrane receptors significantly enriched in the eumetazoa ($p = 3.40 \times 10^{-4}$, hypergeometric test) [23]. DAZ interacting zinc finger protein 3 (*DZIP3*) 1st exon DNA methylation predicted the onset of early stage (AUC = 0.83, OR = 8.82) and all pathological Tumor-Node-Metastasis (pTNM) stages of colorectal cancer (AUC = 0.78, OR = 5.70), whereas *DZIP3* originates from the eumetazoa [24]. Genes from the eukaryotic were the most up-regulated in tumor samples [25], and enrichment of prognostic genes for three cancers, ovarian, breast and lung adenocarcinomas, found that they were all enriched in the eukaryota [26]. During whole-genome duplication (WGD) events, genes that undergo replication are commonly referred to as Ohnologs. Ohnologs have been proven to possess dosage sensitivity [27], making them a significant source of candidate drug targets [28]. Notably, our previous research findings have consistently indicated that Ohnologs exhibit a higher enrichment of successful drug targets compared to non-Ohnologs ($p < 1.51 \times 10^{-40}$, chi-square test) [20], suggesting that information on whether a gene is an Ohnolog can facilitate the drug target identification.

Allosteric proteins serve as a significant source of drug targets [29]. Allosteric proteins refer to proteins that can undergo conformational changes after binding to small molecule ligands [30]. They regulate various physiological activities of the body through the allosteric effect, which is a common phenomenon [31]. Orthosteric regulators reduce the possibility of substrate and enzyme reaction by competing with the substrate binding site, and this mechanism of action is prone to toxic side effects caused by homologous protein reactions [32]. Since the allosteric site and the orthosteric site do not overlap, allosteric regulators have higher selectivity and lower toxicity [33]. Allosteric proteins provide a promising, novel opportunity for the development of innovative therapeutics [34]. In this regard, we developed knowledge graph (KG)-based prediction models for allosteric proteins and integrated the results into GETdb. Unlike traditional networks that only display a single relationship type, KGs can integrate diverse heterogeneous information, including multiple entities and their complex relationships, and provide unstructured semantic relationships between entities, providing a richer way of expressing information for target research [35,36].

Most of the currently available drug target databases have recorded basic biomedical information of protein targets (the three-dimensional structure, function and interactions with small molecule ligands and so on), such as DrugBank (https://go.drugbank.com/), Therapeutic Targets Database (TTD) (https://db.idrblab.net/ttd/), DGIdb (https://dgidb.org/), and the UniProt (https://www.uniprot.org/). However, the evolutionary and genetic information of these targets is not adequately collated and documented in these databases. This situation may result in some important evolutionary and genetic information being overlooked, thus affecting our overall understanding and application of these targets, which to some extent hinders the development of novel drugs. The features and functions of allosteric proteins are important for the development of novel drugs. In this study, we construct a comprehensive drug target database called Genetic and Evolutionary features of drug Targets database (GETdb, http://zhanglab.hzau.edu.cn/GETdb/page/index.jsp). This database not only integrates and standardizes data from dozens of commonly used drug and target databases, but also innovatively includes the genetic and evolutionary information of targets. Moreover, it features an effective allosteric protein prediction model. GETdb contains approximately 4000 targets and over 29,000 drugs, and is a user-friendly database for searching, browsing and downloading data to facilitate the development of novel targets (Fig. 1).

2. Materials and methods

2.1. Collection and processing of drug target information

We collected information on drugs and targets from three databases: DrugBank, TTD, and DGIdb. Initially, we obtained the latest versions of drug target information by downloading relevant files from these databases. The XML files of DrugBank were parsed using the Python ElementTree module to extract pertinent details, such as human-related drug names, drug types, drug groups, target actions, target names (gene names), and other relevant attributes. Basic drug target information was also extracted from the remaining two databases.

We merged the information from DrugBank and TTD based on drug-target pairs. The PyMeSHSim software package [37], an integrated Python toolkit for biomedical naming entity identification and standardization, was utilized in this study. It exhibits a semantic similarity ranging from 0.89 to 0.99 when compared to manually identified web terms. In this study, PyMeSHSim was utilized to extract the unique identifier (ID) terms of Medical Subject Headings (MeSH) from the drug description text present in both the TTD and DrugBank databases. Subsequently, drug indications from these databases were merged based on their respective MeSH IDs. Finally, the medical subject terms associated with these MeSH IDs were employed to facilitate accurate subject identification and description. Additionally, we integrated this merged dataset with DGIdb using drug-gene pairs as the matching criterion, considering the lack of target names in DGIdb. To provide transparency, we added a new column labeled "source" to the merged dataset, indicating the data source of drug-target (gene) pairs.

2.2. Collection and processing of genetic features for drug targets

The Disease Gene Network (DisGeNET, https://www.disgenet.org/) integrates gene-disease association information from multiple databases and a large number of literature sources. We obtained the SQLite database encompassing comprehensive information and extracted the gene-

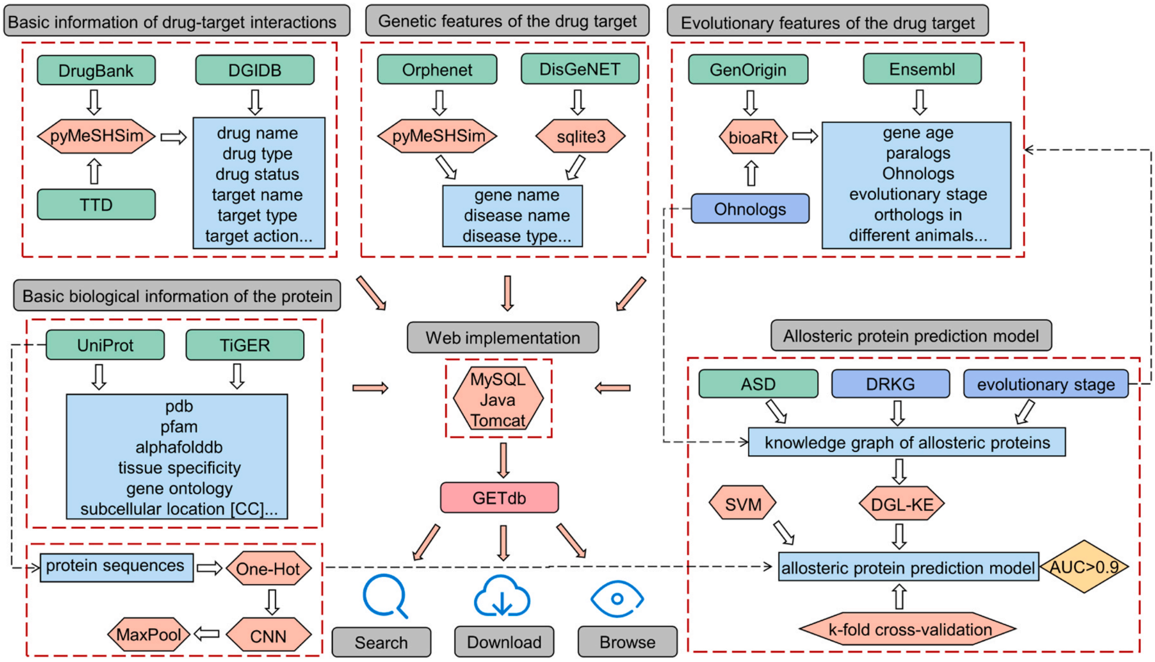


Fig. 1. Flow chart of GETdb. This study constructs a comprehensive drug target database called Genetic and Evolutionary features of drug Targets database (GETdb, http://zhanglab.hzau.edu.cn/GETdb/page/index.jsp). This database not only integrates and standardizes data from dozens of commonly used drug and target databases, but also innovatively includes the genetic and evolutionary information of targets. Moreover, it features an effective allosteric protein prediction model. GETdb contains approximately 4000 targets and over 29,000 drugs, and is a user-friendly database for searching, browsing and downloading data to facilitate the development of novel targets.

disease association data from the database. DisGeNET has developed a gene-disease relationship scoring model, with scores ranging from 0 to 1, where higher scores indicate higher confidence in the gene-disease associations. The mentioned information was additionally incorporated into our study. Rare disease drug development has always been a challenging field [38]. To better support the discovery and development of drugs for rare diseases, we utilized the Orphanet database (https://www.orpha.net/consor/cgi-bin/index.php) to collect information on the relationships between rare diseases and genes. Orphanet is a European collaborative network dedicated to improving the diagnosis, prevention, and treatment of rare diseases and contains information on over 6000 rare diseases worldwide.

The disease descriptions from diverse data sources were standardized utilizing the Unified Medical Language System (UMLS) to ensure consistency in disease names and to mitigate potential confusion arising from synonyms, abbreviations, and other variations [39]. The standardization process was performed using PyMeSHSim. Subsequently, the two gene-disease datasets were merged based on the CUI identifier (a unique concept identifier in the UMLS) and gene names, with an additional column labeled as "source" appended to indicate the respective data sources.

2.3. Collection of evolutionary features for drug targets

Information on the stage of origin of the target, the age of the target, Ohnologs, orthologs, paralogs, and phenotypic similarity of orthologous genes between humans and mice were collected as evolutionary features of drug targets. Liebeskind et al. have inferred the age of genes based on 13 popular homology inference algorithms [21] and classified human genes into eight major categories. This information on the stage of origin was added to our database to further understand the function and disease relevance of genes. The age of a gene is strongly correlated with its function and also with human disease [40,41], and to gain a clearer understanding of the evolutionary features of drug targets, we downloaded age information of human protein-coding genes from GenOrigin (http://genorigin.chenzxlab.cn/#/). Duplicate Ohnologs generated during WGD events have been shown to be metrologically sensitive and a potential source of drug candidates [27], and the information was therefore incorporated into our database construction. In addition, Orthologs are formed by species evolution and paralogs are often functionally similar. We collected information on paralogous genes and orthologous genes across multiple species including Alpaca (*Vicugna pacos*), Chimpanzee (*Pan troglodytes*), Dog (*Canis lupus familiaris*), Guinea Pig (*Cavia porcellus*), Macaque (*Macaca*), Mouse (*Mus musculus*), Pig (*Sus scrofa*), Rat (*Rattus norvegicus*), and Rabbit (*Oryctolagus cuniculus*) from Ensembl 108 (https://www.ensembl.org/index.html?redirect=no). Although the sequences of orthologous genes are highly conserved, functional divergence between orthologous gene products frequently occurs during evolution [42]. Doyeon Ha et al. conducted evolutionary rewiring of regulatory networks and identified 642 high-phenotype-similarity genes and 642 low-phenotype-similarity genes based on phenotype similarity (PS) scores [43]. These genes were used to explain the phenotypic differences between orthologous genes in humans and mice. The data have been integrated into our database for target discovery purposes.

2.4. Collection of basic biological features for drug targets

It has been shown that genes with tissue specificity are twice as likely to be targets as common genes [44], which would facilitate the discovery of new therapeutic targets, and in this study, we utilized the tissue-specific gene data derived from 96 distinct human tissues as delineated by Lüleci and Yılmaz using the extended tau score methodology [45]. To gather essential biological features of drug targets, we collected information from UniProt database (accessed September 2022), including Gene Ontology (GO) annotations, single nucleotide polymorphism (SNP) data, motif information, subcellular localization details, as well as Protein Data Bank (PDB) and Pfam entry identifiers. The GO information provided insights into the functional annotations of the target proteins, categorizing them into molecular function, biological process, and cellular component. SNP data allowed us to analyze the impact of genetic variations on protein function and phenotype. Motif information aided in identifying specific sequence patterns associated with functional regions. Subcellular localization details shed light on the intracellular distribution of the target proteins. Furthermore, the inclusion of PDB and Pfam entry identifiers facilitated direct access to corresponding three-dimensional protein structures and conserved domains. These comprehensive datasets were integrated into our study. The AlloSteric Database (ASD 2023) (http://mdl.shsmu.edu.cn/ASD/) serves as a repository for a substantial collection of experimentally validated or reported allosteric proteins. Within our GETdb, we have incorporated information on 1052 allosteric proteins associated with humans, sourced from the ASD 2023.

2.5. Construction of allosteric protein prediction model

KGs have emerged as a powerful tool for integrating heterogeneous data sources such as chemical, genomic, and biomedical data, to facilitate drug discovery and development [46]. The Drug Repositioning Knowledge Graph (DRKG) is a widely utilized knowledge graph that aggregates diverse data sources related to drugs and biomedicine, providing researchers with a rich information resource. In this graph, nodes may include entities such as drugs, diseases, genes, and edges reflect the complex interactions between these entities, such as potential therapeutic effects of drugs on diseases or associations between drugs and genes. In our study, DRKG serves as the Primary Knowledge Graph (Primary KG) [47]. To construct an Evolutionary-enhanced Knowledge Graph (Evolutionary-enhanced KG), Ohnologs and evolutionary stage information were incorporated into the primary KG. In the processing of both the Primary KG and the Evolutionary-enhanced KG, we employed the DGL-KE framework to train the embeddings on our research team’s Linux server, equipped with multiple Intel E5–2697V4 CPUs, each featuring 18 cores, 36 threads, and 45 MB of cache space. The TransE_l2 algorithm was utilized to transform all entities and relationships within the KGs into a 400-dimensional vector.

There exists a profound correlation between the structure of proteins and their functions, thus, accurately identifying the structural characteristics of proteins is of paramount importance for a deeper understanding of their functionalities. In this study, we have employed Convolutional Neural Networks (CNN), a deep learning technique, for the effective extraction of feature information from allosteric proteins. Specifically, protein sequence data was initially sourced from the UniProt database. These sequences underwent a series of preprocessing steps to ensure data consistency and processability. This included the normalization of sequence lengths, as well as the implementation of one-hot encoding techniques, transforming each amino acid residue into a 20-dimensional vector. In these vectors, only one of the 20 elements is set to 1, with the rest being 0, thereby ensuring the uniqueness of each amino acid [48]. The CNN blocks consisting of three stacked 1D-CNNs in the HyperAttentionDTI model were utilized to extract the protein sequence features [49]. Convolutional kernels of different scales (32, 64, and 96) were utilized to capture the relationships between sequence segments of different lengths, and the max pooling was done to extract the most important features of each channel and reduce the dimensionality of the output vector. Finally, a feature matrix is obtained in which each protein is represented by a 160-dimensional vector.

In this study, we utilized data on 837 human-related allosteric proteins collected from the ASD 2019. These data revealed that 821 of these proteins were present in our KGs. Consequently, we selected these 821 proteins as positive samples for machine learning training. To create a balanced training dataset, an equal number of non-allosteric proteins were randomly selected from the protein entities in the KG to serve as negative samples. This approach led to the formation of a training dataset comprising 1642 samples, split evenly between 821 positive and 821 negative samples. We extracted protein entity features from both primary and evolutionary-enhanced KGs, and combined them with the sequence features of the proteins. Each protein was then represented as a 560-dimensional feature vector for the training dataset. Four different machine learning methods, Support Vector Machines (SVM), Random Forests (RF), Logistic Regression (LR), and K-Nearest Neighbors (KNN) algorithms, were employed to construct models for predicting allosteric proteins. These models were trained using protein feature vectors as input. Two types of KGs were used: the Primary KG and the Evolutionary-enhanced KG, leading to the creation of two models named the "Primary Model" and the "Evolutionary-Enhanced Model". To optimize the practical utility of the models and ensure the stability of the results, we employed a 3-fold cross-validation approach to train the classifiers, repeating the experiment 200 times. The performance of the classifiers was evaluated by calculating the area under the Receiver Operating Characteristic (ROC) curve (AUC), which generally reflects the overall efficacy of the model. Accuracy, error rate, recall rate, and ROC values were used as criteria to evaluate the different models. The optimal model was then employed to generate predictions for all proteins in KG, and these predictions, including the output decision values and variability status, were recorded in GETdb.

2.6. Implementation

GETdb is a database application based on the JavaWeb framework, designed to provide an online platform for convenient and efficient querying and analysis of drug target data. The application uses MySQL 8.0 as the database version, which contains a variety of approved and potential targets, and provides functions of query and browsing. The front-end of GETdb uses a bootstrap framework for page rendering to facilitate user interaction and usage. GETdb runs on a tomcat server, which is written in Java and is stable and scalable. The GETdb website runs on the Linux-based Apache Web server 7.0.108 (http://www.apache.org) and is developed using MySQL 8.0.29 (http://www.mysql.com), providing a convenient online data platform for researchers. We have posted the database source code on GitHub at https://github.com/Seay-7/GETdb/releases/tag/GETdbVersion1.0.

3. Results

3.1. Content statistics of the database

GETdb contains information on 4337 targets interacting with 29,116 drugs from DrugBank, TTD and DGIdb. There are 2927 approved drugs and 513 successful targets (Table 1). In collecting genetic features of targets, we standardized the disease names according to UMLS IDs and integrated two datasets from Orphanet and DisGeNET. 30,447 UMLS IDs were normalized by PyMeSHSim to 8100 disease names, significantly reducing the complexity of the dataset. We conducted an enrichment analysis on 513 successful targets within GETdb and found a significant enrichment of successful targets in four evolutionary stages (Cellular organisms, Euk+Bac, Eumetazoa, and Vertebrata) based on the hypergeometric test ($p < 0.05$) (Fig. 2). 12,778 pairs of Ohnologs and 1263,664 pairs of Paralog genes were included as evolutionary features in our drug target database. The top three organisms with the highest number of homologous genes were Mouse (*Mus musculus*), Chimpanzee (*Pan troglodytes*), and Rabbit (*Oryctolagus cuniculus*). Disease genes tend to be expressed in the tissues where disease occurs. Upon collecting and analyzing gene expression specificity across 96 different tissues, it was observed that the testes exhibit the highest number of tissue-specifically expressed genes, with a total of 1319. However, the liver stands out as the tissue with the highest number of successfully targeted genes, with 40 of its tissue-specifically expressed genes identified as successful targets. GETdb contains 1052 allosteric proteins supported by literature and patent evidence from ASD 2023. In addition, we have added 3D structural information to 7710 proteins in the GETdb database, including detailed data on resolution and 3D structure determination methods, a total of 133,158 data entries.

Table 1
Data summary of GETdb.

Data item	Number of drug/ gene/target	Data item	Number of drug/ gene/target
Drug type		**Evolutionary features of target**	
		Ohnolog[b]	12,778
Biotech	464	Paralog gene[c]	1,263,664
Small molecule	5496	*Ortholog gene*[c]	
Clinical phase of drug		With alpaca	8388
		With chimpanzee	21,121
Approved	2942	With dog	16,719
Clinical trial	23,184	With guinea pig	15,853
Preclinical	3081	With macaque	19,160
Discontinued	1045	With mouse	24,854
Withdrawn	56	With pig	16,883
Clinical phase of target		With rat	13,902
		With rabbit	20,245
		Phenotypic similarity with mouse[d]	
Successful	513	High	642
Clinical trial	891	Low	642
Preclinical	36	*Evolutionary stage of gene*[e]	
Discontinued	45	Cellular organisms	853
Literature-reported	293	Euk Archaea	203
Patented-recorded	119	Euk+Bac	1475
Genetic features of target[a]		Eukaryota	5439
		Eumetazoa	4810
		Mammalia	2216
Disease-associated genes	20,442	Opisthokonta	1066
		Vertebrata	2560

[a] derived from DisGeNET database (https://www.disgenet.org/) and Orphanet database (https://www.orpha.net/consor/cgi-bin/index.php).

[b] derived from the work by Takashi Makino et al.[55]

[c] derived from Ensembl database (https://www.ensembl.org/index.html?redirect=no).

[d] derived from the work by Doyeon Ha et al. [43].

[e] derived from the work by Benjamin J Liebeskind et al. [21]

3.2. Web interface

GETdb offers users eight distinct functional modules, encompassing the Home, Genetic Features, Evolutionary Features, Allosteric Protein, Drug, Browse, Help, and Download. The home page module serves as the central page of the system, aiming to provide users with a brief introduction to the development background and significant significance of GETdb. At the top of the home page, users can gain insights into the advantages of the GETdb database. The middle section of the home page vividly presents the primary components of the database through five graphical representations, showcasing key data on evolutionary features of drug targets, genetic features, and allosteric protein prediction models. At the bottom of the page, GETdb provides users with statistical information, including the scale and content of the database, along with a concise description of the various operations and data queries available to users within GETdb.

The genetic features module offers two search boxes, allowing users to access relevant genetic feature information by utilizing the "Search by Gene Name" box. This information includes the disease names associated with the gene, the unique identifiers "CUI" of the diseases within the UMLS system, as well as the types or nature of the associations between the gene and diseases. Additionally, the module provides a "Search by Disease Name" box, enabling users to retrieve gene-related information associated with a specific disease. For example, by searching for "Alzheimer disease type 1", the user can obtain all the genes associated with the disease, such as amyloid beta precursor protein (*APP*), ETS proto-oncogene 2, transcription factor (*ETS2*) and *SOD1*, which are ranked in ascending order according to their associated scores with the disease (Fig. 3**a**).

The module of evolutionary features in GETdb includes three drop-down search boxes. The "Search by Gene Name" box allows users to input a gene name to determine the evolutionary stage to which the gene belongs. For example, a query for evolutionary information for three targets associated with Alzheimer disease type 1 (*APP*, *ETS2* and *SOD1*) shows that *APP* and *ETS2* originate from the Eumetazoa and *SOD1* from the Cellular organisms, both of which are evolutionary stages enriched for successful targets, although *APP* has been identified as a successful target and *ETS2* as well as *SOD1* are currently in clinical trials, but their evolutionary stages suggest that these two genes possess great potential

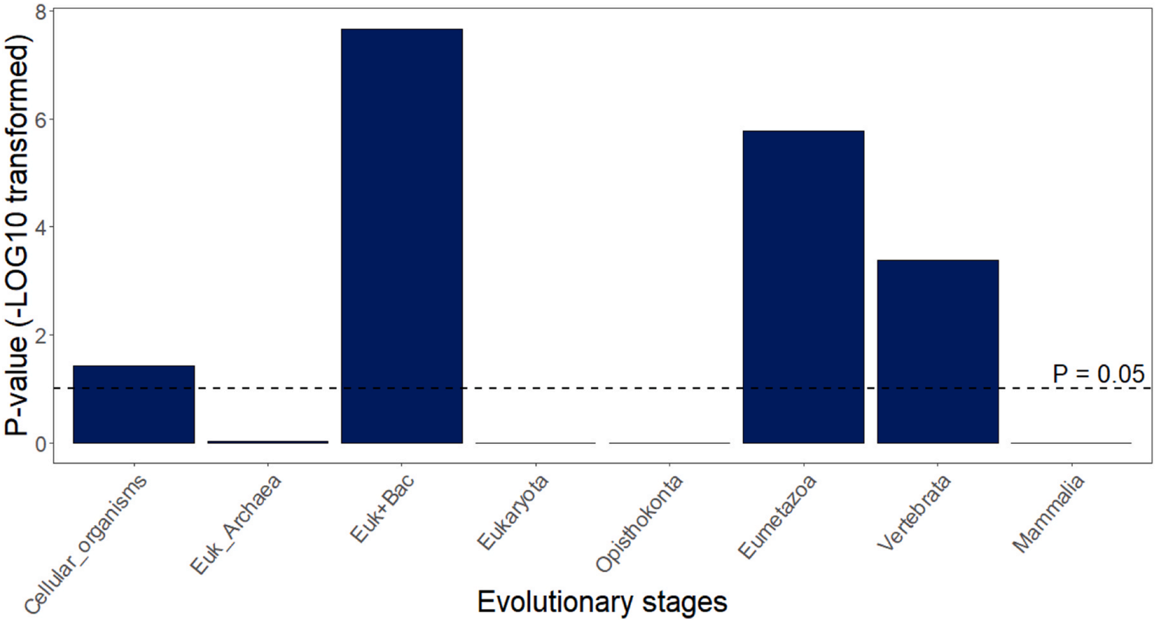
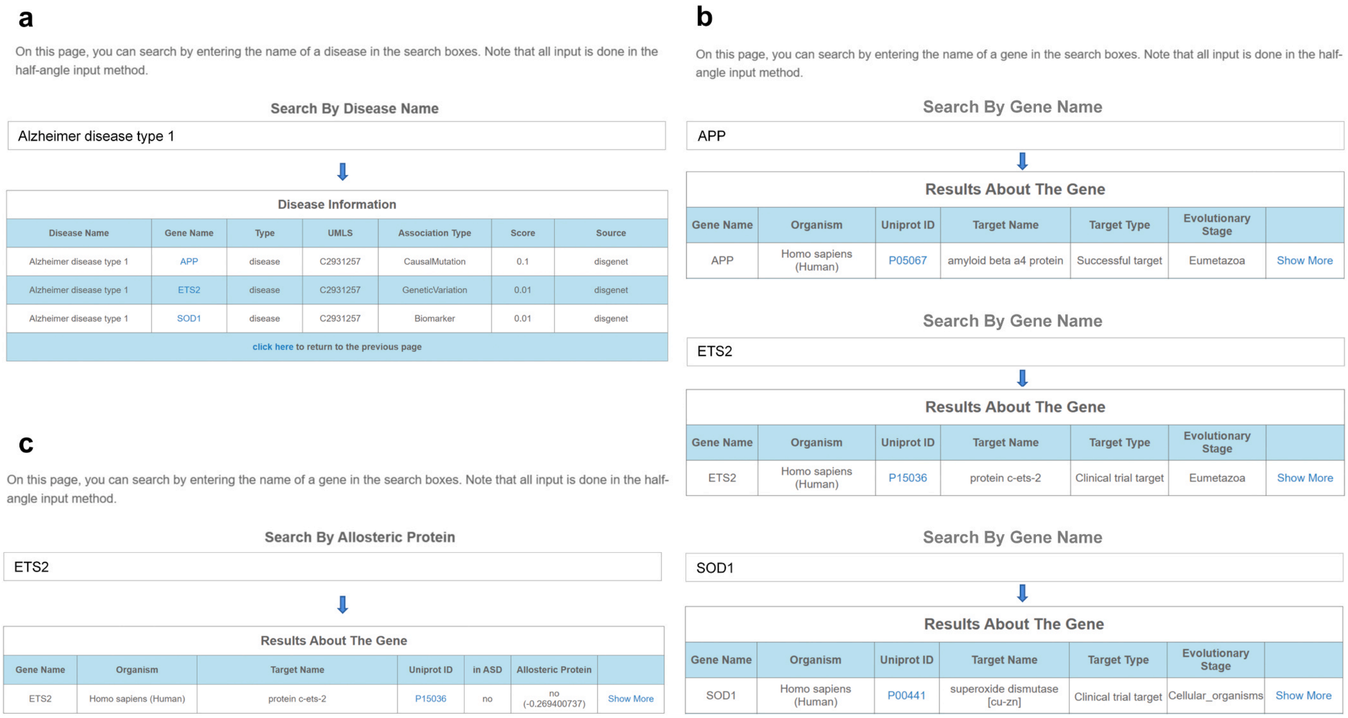


Fig. 2. Enrichment pattern of successful targets in GETdb. Results of enrichment analyses of successful targets from 8 different evolutionary stages. A successful target means that recorded in Therapeutic Target Database (TTD). Each bar represents the enrichment of successful targets at the corresponding evolutionary stage compared to all successful targets. Larger values on the vertical axis indicate higher enrichment of successful targets at this stage.



On this page, you can search by entering the name of a disease in the search boxes. Note that all input is done in the half-angle input method.

Search By Disease Name

Alzheimer disease type 1

Disease Information

Disease Name	Gene Name	Type	UMLS	Association Type	Score	Source
Alzheimer disease type 1	APP	disease	C2931257	CausalMutation	0.1	disgenet
Alzheimer disease type 1	ETS2	disease	C2931257	GeneticVariation	0.01	disgenet
Alzheimer disease type 1	SOD1	disease	C2931257	Biomarker	0.01	disgenet

click here to return to the previous page

On this page, you can search by entering the name of a gene in the search boxes. Note that all input is done in the half-angle input method.

Search By Gene Name

APP

Results About The Gene

Gene Name	Organism	Uniprot ID	Target Name	Target Type	Evolutionary Stage	
APP	Homo sapiens (Human)	P05067	amyloid beta a4 protein	Successful target	Eumetazoa	Show More

Search By Gene Name

ETS2

Results About The Gene

Gene Name	Organism	Uniprot ID	Target Name	Target Type	Evolutionary Stage	
ETS2	Homo sapiens (Human)	P15036	protein c-ets-2	Clinical trial target	Eumetazoa	Show More

Search By Gene Name

SOD1

Results About The Gene

Gene Name	Organism	Uniprot ID	Target Name	Target Type	Evolutionary Stage	
SOD1	Homo sapiens (Human)	P00441	superoxide dismutase [cu-zn]	Clinical trial target	Cellular_organisms	Show More

On this page, you can search by entering the name of a gene in the search boxes. Note that all input is done in the half-angle input method.

Search By Allosteric Protein

ETS2

Results About The Gene

Gene Name	Organism	Target Name	Uniprot ID	in ASD	Allosteric Protein	
ETS2	Homo sapiens (Human)	protein c-ets-2	P15036	no	no (-0.269400737)	Show More

Fig. 3. Search interface of GETdb. (a) The "Search by Disease Name" in genetic features module allows the user to retrieve information about the genes associated with a disease, which includes all the genes associated with the disease, the disease's unique identifier "CUI" in the UMLS system, and the types of associations. (b) The "Search by Gene Name" in evolutionary features module allows users enter the name of a gene and submit a query. The query results in a comprehensive summary of the gene, which contains key details such as gene name, species, UniProt ID, target name, target type and its evolutionary stage. (c) In the “Allosteric Protein” module, users can retrieve relevant information about the encoded proteins by entering a gene name. This information includes the protein name, UniProt ID, whether it is recorded in the ASD 2023, and the predictions from the allosteric protein modeling model, including the prediction outcome and corresponding decision value.

to become successful targets (Fig. 3**b**). Additionally, users may access further information regarding the gene of interest by clicking on the UniProt ID, which will provide access to the UniProt database. By selecting the "Show More" button, users can explore further detailed information about the target gene, leading them to a secondary screen with more comprehensive features. In the "Search by Ohnolog" box, users can search for Ohnolog genes. If the input gene is an Ohnolog, the system returns the duplicate copy that originated from WGD. In case the input gene is not an Ohnolog, "NA" is returned. Moreover, in the "Search by Ortholog" box, users have the ability to search for homologous genes in eight model organisms, namely Alpaca (*Vicugna pacos*), Chimpanzee (*Pan troglodytes*), Dog (*Canis lupus familiaris*), Guinea Pig (*Cavia porcellus*), Macaque (*Macaca*), Mouse (*Mus musculus*), Pig (*Sus scrofa*), Rat (*Rattus norvegicus*), and Rabbit (*Oryctolagus cuniculus*).

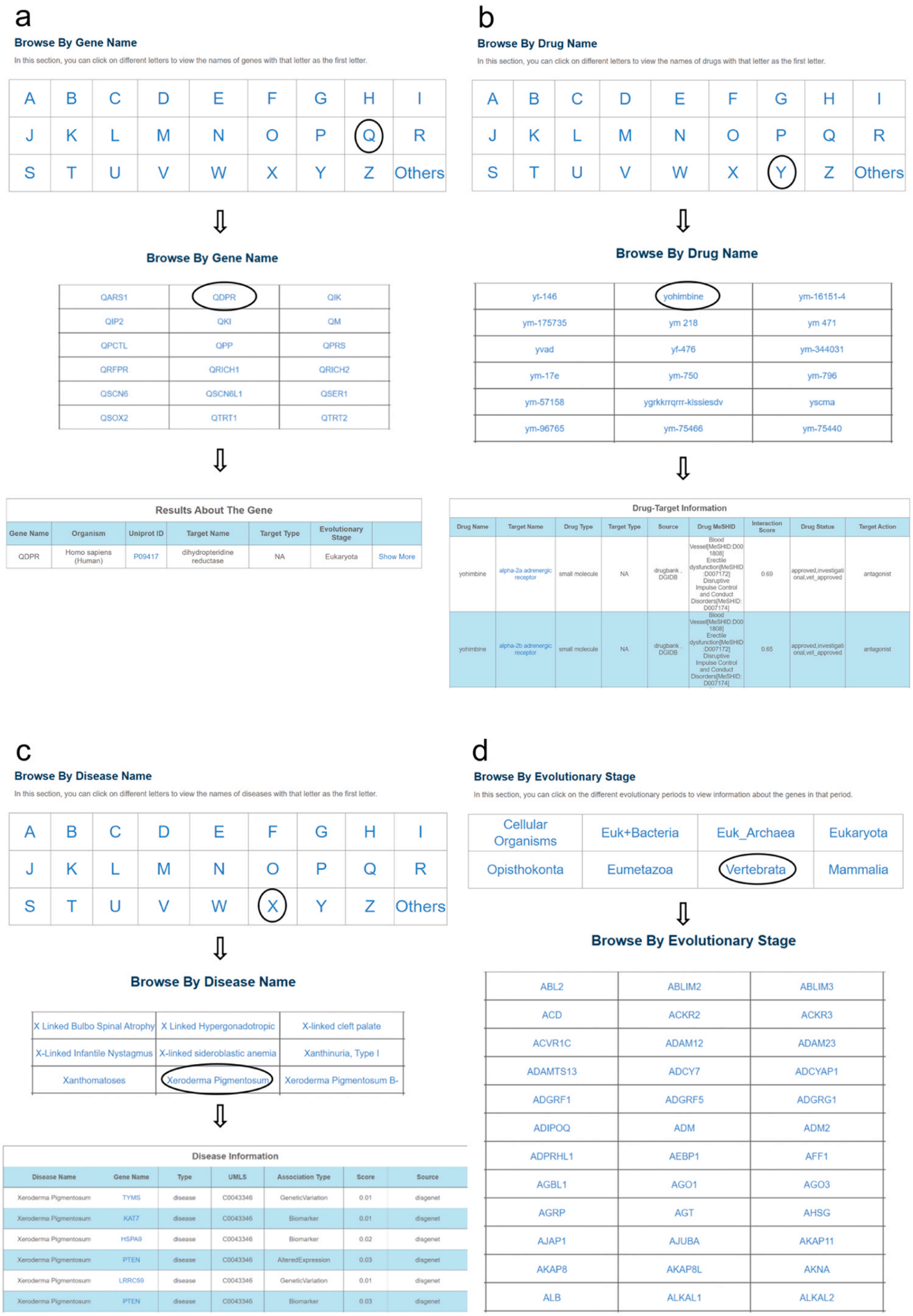
In the “Allosteric Protein” module, users are able to determine whether the corresponding protein is an allosteric protein or not by entering a gene name. As an example, the query result of *ETS2*, a clinical trial target of Alzheimer disease type 1, shows that this protein is not listed in ASD 2023 and it is predicted to be a non-allosteric protein according to the evolutionary-enhanced model, which implies that there is a low probability for *ETS2* to be an allosteric target (Fig. 3**c**). In the drug search section, users are prompted to enter the name of the drug of interest to obtain comprehensive information about the drug and its corresponding targets. The results page not only displays the names and types of the drug and targets but also provides additional details, including drug status, indications, interaction scores between the drug and genes, interaction types, and data sources. To obtain more information about a specific target, users can click on the corresponding target name.

On the other hand, the browsing module offers users access to information based on the initial letters of gene, drug, or disease names, with non-alphabetical fields categorized as "Other". It also allows users to access all genes specific to a particular evolutionary stage by selecting the relevant stage. Once the desired letter abbreviation or evolutionary stage is chosen, the results are displayed in a tabular format, enabling users to click on any gene, drug, or disease of interest to view more detailed information (Fig. 4). Furthermore, the interface provides several auxiliary features, including the "TOP" button located at the bottom right of the screen, allowing users to easily navigate back to the top of the page if the results page is lengthy. The help interface provides a comprehensive overview of the database construction process, detailed instructions on how to use GETdb, and a brief introduction to our team. Lastly, the download module enables users to access options for downloading drug target data, gene-disease information, and gene-related data. Please note that users are required to accept the relevant disclaimer before proceeding with downloads.

GETdb has completed its internal testing, garnering positive feedback from testers including software developers and bioinformatics researchers. They praised its systematic and comprehensive access to target information, its quick response times, and its provision of reliable support for drug target research. This feedback highlights the database's potential to streamline pharmaceutical research and aid in the efficient discovery of drug targets, marking a promising step forward for its future development and application.

3.3. Validation of allosteric protein models

We utilized support vector machines, random forests, logistic regression, and k-nearest neighbor algorithms to construct the primary model and evolutionary-enhanced model for predicting allosteric proteins. Following the evaluation of two models, we found that the performance of the evolutionary-enhanced model surpassed that of the primary model in terms of accuracy, precision, recall, and AUC (Table 2), indicating that the integration of evolutionary and sequence



Browse By Gene Name

In this section, you can click on different letters to view the names of genes with that letter as the first letter.

A	B	C	D	E	F	G	H	I
J	K	L	M	N	O	P	Q	R
S	T	U	V	W	X	Y	Z	Others

Browse By Gene Name

QARS1	QDPR	QIK
QIP2	QKI	QM
QPCTL	QPP	QPRS
QRFPR	QRICH1	QRICH2
QSCN6	QSCN6L1	QSER1
QSOX2	QTRT1	QTRT2

Results About The Gene

Gene Name	Organism	Uniprot ID	Target Name	Target Type	Evolutionary Stage	
QDPR	Homo sapiens (Human)	P09417	dihydropteridine reductase	NA	Eukaryota	Show More

Browse By Drug Name

In this section, you can click on different letters to view the names of drugs with that letter as the first letter.

A	B	C	D	E	F	G	H	I
J	K	L	M	N	O	P	Q	R
S	T	U	V	W	X	Y	Z	Others

Browse By Drug Name

yt-146	yohimbine	ym-16151-4
ym-175735	ym 218	ym 471
yvad	yf-476	ym-344031
ym-17e	ym-750	ym-796
ym-57158	ygrkkrrqrrr-klssiesdv	yscma
ym-96765	ym-75466	ym-75440

Drug-Target Information

Drug Name	Target Name	Drug Type	Target Type	Source	Drug MeSHID	Interaction Score	Drug Status	Target Action
yohimbine	alpha-2a adrenergic receptor	small molecule	NA	drugbank , DGIDB	Blood Vessel[MeSHID:D001808] Erectile dysfunction[MeSHID:D007172] Disruptive Impulse Control and Conduct Disorders[MeSHID:D007174]	0.69	approved,investigational,vet_approved	antagonist
yohimbine	alpha-2b adrenergic receptor	small molecule	NA	drugbank , DGIDB	Blood Vessel[MeSHID:D001808] Erectile dysfunction[MeSHID:D007172] Disruptive Impulse Control and Conduct Disorders[MeSHID:D007174]	0.65	approved,investigational,vet_approved	antagonist

Browse By Disease Name

In this section, you can click on different letters to view the names of diseases with that letter as the first letter.

A	B	C	D	E	F	G	H	I
J	K	L	M	N	O	P	Q	R
S	T	U	V	W	X	Y	Z	Others

Browse By Disease Name

X Linked Bulbo Spinal Atrophy	X Linked Hypergonadotropic	X-linked cleft palate
X-Linked Infantile Nystagmus	X-linked sideroblastic anemia	Xanthinuria, Type I
Xanthomatoses	Xeroderma Pigmentosum	Xeroderma Pigmentosum B-

Disease Information

Disease Name	Gene Name	Type	UMLS	Association Type	Score	Source
Xeroderma Pigmentosum	TYMS	disease	C0043346	GeneticVariation	0.01	disgenet
Xeroderma Pigmentosum	KAT7	disease	C0043346	Biomarker	0.01	disgenet
Xeroderma Pigmentosum	HSPA9	disease	C0043346	Biomarker	0.02	disgenet
Xeroderma Pigmentosum	PTEN	disease	C0043346	AlteredExpression	0.03	disgenet
Xeroderma Pigmentosum	LRRC59	disease	C0043346	GeneticVariation	0.01	disgenet
Xeroderma Pigmentosum	PTEN	disease	C0043346	Biomarker	0.03	disgenet

Browse By Evolutionary Stage

In this section, you can click on the different evolutionary periods to view information about the genes in that period.

Cellular Organisms	Euk+Bacteria	Euk_Archaea	Eukaryota
Opisthokonta	Eumetazoa	Vertebrata	Mammalia

Browse By Evolutionary Stage

ABL2	ABLIM2	ABLIM3
ACD	ACKR2	ACKR3
ACVR1C	ADAM12	ADAM23
ADAMTS13	ADCY7	ADCYAP1
ADGRF1	ADGRF5	ADGRG1
ADIPOQ	ADM	ADM2
ADPRHL1	AEBP1	AFF1
AGBL1	AGO1	AGO3
AGRP	AGT	AHSG
AJAP1	AJUBA	AKAP11
AKAP8	AKAP8L	AKNA
ALB	ALKAL1	ALKAL2

Fig. 4. Browse interface of GETdb. (a) Search by the first letter of the gene and obtain information about it. (b) Search by drug initials for drug-related information. (b) Search by disease initials for disease-related information. (d) Search for all genes in the Vertebrata period.

information effectively enhances the accuracy and reliability of identifying allosteric proteins. In particular, the AUC of the evolutionary enhanced model (0.914) was significantly higher than that of the primary model (0.588) when using the support vector machine. Therefore, the evolutionary-enhanced model constructed using the support vector machine was selected as the preferred allosteric protein prediction model. To verify the effectiveness of the model, we conducted an analysis of the mechanism of action for ten potential allosteric proteins, with the aim of comparing their three-dimensional structures before and after binding to ligands to ascertain the presence of allosteric effects. We searched their crystal structures and reviewed the relevant literature and eventually found evidence for conformational changes in nine proteins (except for G protein beta subunit 3) (Table 3). Overall, our model is highly reliable and practical.

The SVM classifies each entity according to its features and calculates a decision value. The decision value's sign determines whether the predicted protein is an allosteric protein or not: a positive value means it is an allosteric protein and a negative value means it is not. A higher absolute value indicates a higher probability of classification into that category. For our model predictions on allosteric proteins, we first excluded the confirmed allosteric proteins listed in the ASD 2019. Subsequently, we selected the top ten potential allosteric proteins based on the highest model scores. Remarkably, four of these proteins served as successful drug targets (Table 3). For the remaining six genes, we

Table 2
Evaluation results of two allosteric protein prediction models.

	Evaluation indicators	Primary model[a]	Evolutionary-enhanced model[b]
Support Vector Machines	Accuracy	0.5644516	0.8289771
	Precision	0.5623055	0.8293212
	Recall	0.5841126	0.8415918
	AUC	0.5884506	0.9144113
Random Forest	Accuracy	0.6299890	0.8292773
	Precision	0.6340979	0.8229830
	Recall	0.6213205	0.8401114
	AUC	0.6308832	0.8296594
Logistic Regression	Accuracy	0.5453957	0.6775801
	Precision	0.5456479	0.6860938
	Recall	0.5420708	0.6568874
	AUC	0.5491645	0.7132236
K-Nearest Neighbor Classification	Accuracy	0.5409907	0.7116161
	Precision	0.5574970	0.8619031
	Recall	0.4045184	0.5048360
	AUC	0.5415975	0.7118271

[a] Primary model is the allosteric protein prediction model constructed based on the original knowledge graph.

[b] Evolutionary-enhanced model is the allosteric protein prediction model constructed using evolutionary-enhanced knowledge graphs and protein sequence features.

Table 3
Potential allosteric proteins predicted by evolutionary-enhanced model (Top 10).

Gene Name	Decision Value	Evidence (PMID[a])	Successful target
NOS1	2.304509569	29772550	No
KNG1	2.096214024	10627460	No
ANXA1	2.094845521	11178908	Yes
GNB3	1.942980884	21772288	No
NOS2	1.904724304	19737939	No
HTR2A	1.849801463	30723326	Yes
SRC	1.83821705	9024657	Yes
PRKAA1	1.837083033	17851531	No
PTK2B	1.815861602	18031286	No
SCN5A	1.804870786	22473783	Yes

[a] PubMed Unique Identifier (PMID) is the number of literatures in the fields of life sciences and medicine included in the PubMed search engine (https://pubmed.ncbi.nlm.nih.gov/).

assessed their potential as drug targets by comparing the consistency between diseases associated with these genes and the therapeutic effects of related unapproved drugs. Further investigation was conducted only for candidate targets associated with diseases that scored above 0.5 in DisGeNET.

Through data collection and analysis, we have identified the *NOS2* as a promising drug target. The *NOS2*, also known as inducible nitric oxide synthase (*iNOS*) or nitric oxide synthase 2 (*NOS2*), exhibits structural components that include an N-terminal oxidase domain and a C-terminal reductase domain, which are interconnected through a binding region for calmodulin (CaM). The study has provided evidence of dynamic conformational changes within *NOS2*. Notably, the crystal structure analysis of the Ca2 + -bound protein complex CaCaM·FMN has revealed the existence of four distinct conformations, with pronounced disparities observed in the CaM binding peptide (Leu515-Ser535 residues) and the rotational movement of CaM around Arg536/Glu47 (CaM) pairs. These rotations induce significant perturbations in the CaM domain, playing a critical role in facilitating efficient electron transfer between different redox centers within *iNOS* [50].

The *NOS2* gene variants, rs2779248 and rs1137933, have demonstrated significant associations with type 2 diabetes mellitus (T2DM), indicating a potential role in increasing susceptibility to this condition. Specifically, the presence of the *NOS2* rs2779248 variant allele C and genotype TC, as well as the *NOS2* rs1137933 variant allele A and genotype GA, may contribute to an elevated risk of developing T2DM [51]. Furthermore, a query in the DrugBank database revealed pimagedine, a drug targeting *NOS2*, primarily used for the study and treatment of diabetic nephropathy. This finding further validates the potential of *NOS2* as a therapeutic target for T2DM. Additionally, our study identifies *NOS2* as an Ohnolog, providing additional support for its potential as a target. Moreover, we explored the origin of *NOS2* and found that it originated in Euk+Bac. This finding aligns with the results from the previous successful target enrichment stage, strengthening the possibility of *NOS2* as a therapeutic target. These findings offer valuable clues and guidance for further exploration and development of diabetes treatment drugs targeting *NOS2*.

3.4. Application of GETdb in drug target identification

The GETdb is dedicated to providing genetic and evolutionary information aimed at optimizing the process of drug target identification and validation. In order to validate its application, four targets were selected for case analysis in this study, which included two targets identified as successful along with two targets that failed to achieve the expected success (Table 4). These targets were selected on the basis of their genetic scores, evolutionary stages and whether they exhibited allosteric properties.

The comparative analysis between Group A and Group B reveals that targets with higher genetic scores, such as *ABCA1* for Tangier disease, demonstrate a higher probability of success. This finding underscores the importance of genetic relevance in the target selection process. Moreover, the comparison between Group B and Group C indicates that targets with allosteric properties (e.g., *LDLR*) also tend to be successful, further highlighting the significance of the target's allosteric properties in its success.

Further analysis between Group A and Group D indicates that despite the genetic association of *HRAS* with Syndrome, Costello disease, it was not considered a successful target because it did not fall within the "enriched evolutionary stages of successful targets" identified (Cellular organisms, Euk+Bac, Eumetazoa, Vertebrata). This comparison not only confirms the role of evolutionary stages in determining the probability of target success but also highlights the importance of integrating genetic and evolutionary information in the target selection process.

Access to GETdb and the utilization of this database by drug development researchers could have facilitated the rapid identification of these key factors, potentially leading to the adjustment of their target selection strategies. This approach could enhance the efficiency of drug development and reduce the risk of failure during the development process. In summary, GETdb provides strong theoretical support and demonstrates its significant added value in the practical drug development process through empirical analyses.

Table 4
The comparison of successful and discontinued drug targets based on genetic scores, evolutionary stages and allosteric properties.

Group	Successful target		Discontinued target	
	A	B	C	D
Target (Gene)	ABCA1	LDLR	HSPG2	HRAS
Disease	Tangier Disease	Tangier Disease	Tangier Disease	Syndrome, Costello
Genetic Score[a]	1	0.01	0.01	1
Evolutionary Stage[b]	Euk+Bac	Eumetazoa	Eumetazoa	Eukaryota
Allosteric Protein[c]	yes	yes	no	yes

[a] derived from DisGeNET database (https://www.disgenet.org/).

[b] derived from the work by Benjamin J Liebeskind et al. [21]

[c] predicted by the SVM-trained evolutionary-enhanced model.

4. Conclusion and discussion

The innovative drug development process is a multifaceted and time-consuming undertaking, with target discovery representing a critical step in the process [1,6], thus rapid and accurate identification of drug targets is crucial for expediting drug development. Genetics is crucial in understanding the links between genotype and phenotype, which enables researchers to pinpoint key causative genes for various diseases [11]. Given the strong correlation between key causative genes and disease, a number of studies have demonstrated that these genes represent a critical source of modern drug targets. Indeed, genetically supported targets are twice as likely to receive approval [17] and the proportion of drug mechanisms with direct genetic support increases significantly throughout the drug development pipeline [12]. Benefiting from the development of evolutionary biomedicine, there is growing evidence that the pathogenesis and development of multiple diseases (including cancer, neurological diseases, cardiovascular diseases and drug resistance) are closely related to the evolutionary history of humans [52,53]. Therefore, evolutionary information of genes can also help in the identification of drug targets.

However, genetics and evolution are neglected in the current drug target databases, and to fill this gap, we merged information from three current highly recognized drug target databases, they are DrugBank, TTD, and DGIdb, to obtain more comprehensive drug target information. On this basis, we collected gene-disease association information from DisGeNET and Orphanet as genetic features. The evolutionary stages, Ohnologs, orthologs, and paralogs of the targets, as well as the phenotypic similarity information of orthologous genes between humans and mice, were collected as evolutionary features. We standardized and merged these data to construct a large-scale comprehensive drug target database, GETdb, with the aim of providing a comprehensive and user-friendly platform to search for target information. In addition, allosteric regulators offer several significant advantages over orthosteric regulators, including higher selectivity, lower toxicity, and better inhibition of drug resistance [32,33,54], pointing the way to the development of novel drugs. In our study, we created a primary model and an evolutionary-enhanced model for predicting allosteric proteins. The first, we defined two KGs. The DRKG serves as the primary KG. By adding gene age and evolutionary stage information to DRKG, we created an evolutionary-enhanced KG. The primary model is built on the primary KG, and utilizes basic entity relationship features from the primary KG and protein sequence features as inputs for the machine learning model. In contrast, the evolutionary-enhanced model is based on the evolutionary-enhanced KG, and incorporates both the evolutionary-enhanced KG's basic entity relationship features and protein sequence features. Notably, we explored the use of support vector machine, random forest, logistic regression, and k-nearest neighbor algorithms to construct prediction models for allosteric proteins. To assess the effectiveness of these methods, we calculated the accuracy, precision, recall, and AUC of the models as key performance metrics. Comparing the models built using the four different machine learning methods, we found that the performance of the evolutionary-enhanced model consistently outperformed that of the Primary model. So, we decided to use this model to predict allosteric proteins, and the predictions were considered as useful resources to be added to our database.

To effectively use the information provided by GETdb to select drug targets, researchers can adopt a systematic approach that combines genetic and evolutionary data. For example, when selecting potential drug targets, genetic associations between genes and specific diseases need to be considered, with a focus on those genes that are strongly associated with disease. Such genes are supported by genetic evidence and have a higher likelihood of being successful drug development targets [17–19]. Also, the inclusion of evolutionary information can significantly improve the ability to identify targets. An assessment of the evolutionary stage of genes revealed that genes located at key evolutionary stages (Cellular organisms, Euk+Bac, Eumetazoa, and Vertebrata) were more likely to be successful targets, while Ohnologs were also more likely to be successful targets than non-Ohnologs [20]. This evolutionary-based perspective helps to screen for promising targets. Understanding gene conservation across species is also critical for model organism selection and predicting drug efficacy and safety in humans. In addition, our prediction models for allosteric proteins, especially the evolutionary-enhanced model, provide a novel pathway for the discovery of new targets. By focusing on potential targets that play a key role in cellular function, researchers can identify unique targets with advantages such as higher specificity and lower drug resistance based on this model. Besides, the tissue specificity of the target can be used to select targets that act at the right site, thus optimizing efficacy and reducing adverse effects. In summary, by utilizing the comprehensive data in GETdb, researchers can simplify the process of selecting promising drug targets based on criteria such as genetic relatedness, evolutionary conservation and allosteric regulatory potential. This strategy not only helps to rapidly identify targets, but also facilitates the development of more effective and safer drugs.

To keep its information accurate and useful, we will continue to update GETdb. This includes adding newly discovered drug target data and removing outdated information, as well as introducing new features that are closely related to target information. We expect GETdb to play an increasing role in drug target discovery as well as novel drug development.

CRediT authorship contribution statement

Yuan Quan: Data curation, Visualization, Funding acquisition, Methodology, Supervision, Writing – review & editing. **Hong-Yu Zhang:** Methodology, Writing – review & editing. **Qi Zhang:** Data curation, Writing – original draft, Writing – review & editing, Visualization. **Yang He:** Software, Writing – review & editing. **Ya-Ping Lu:** Data curation, Writing – review & editing. **Qi-Hao Wei:** Data curation, Writing – review & editing.

Declaration of Competing Interest

The authors declare no competing interests.

Data Availability

All data and resources of GETdb are freely available at http://zhanglab.hzau.edu.cn/GETdb/page/index.jsp.

Acknowledgments

This work was supported by National Natural Science Foundation of China (grant 32300545) and Young Elite Scientists Sponsorship Program by CAST (2023QNRC001).

References

[1] Xue H, Li J, Xie H, et al. Review of drug repositioning approaches and resources. Int J Biol Sci 2018;14:1232–44.
[2] Brown DG, Wobst HJ, Kapoor A, et al. Clinical development times for innovative drugs. Nat Rev Drug Discov 2022;21:793–4.
[3] Wouters OJ, McKee M, Luyten J. Estimated research and development investment needed to bring a new medicine to market, 2009-2018. JAMA 2020;323:844–53.
[4] Zhong F, Xing J, Li X, et al. Artificial intelligence in drug design. Sci China Life Sci 2018;61:1191–204.
[5] Dahlin JL, Inglese J, Walters MA. Mitigating risk in academic preclinical drug discovery. Nat Rev Drug Discov 2015;14:279–94.
[6] Lee A, Lee K, Kim D. Using reverse docking for target identification and its applications for drug discovery. Expert Opin Drug Discov 2016;11:707–15.
[7] Gates AJ, Gysi DM, Kellis M, et al. A wealth of discovery built on the Human Genome Project - by the numbers. Nature 2021;590:212–5.
[8] Santos R, Ursu O, Gaulton A, et al. A comprehensive map of molecular drug targets. Nat Rev Drug Discov 2017;16:19–34.

[9] Nath A, Kumari P, Chaube R. Prediction of human drug targets and their interactions using machine learning methods: current and future perspectives. Methods Mol Biol 2018;1762:21–30.
[10] Guo X, Chitale P, Sanjana NE. Target discovery for precision medicine using high-throughput genome engineering. Adv Exp Med Biol 2017;1016:123–45.
[11] Tam V, Patel N, Turcotte M, et al. Benefits and limitations of genome-wide association studies. Nat Rev Genet 2019;20:467–84.
[12] Nelson MR, Tipney H, Painter JL, et al. The support of human genetic evidence for approved drug indications. Nat Genet 2015;47:856–60.
[13] Pirozzi CJ, Yan H. The implications of IDH mutations for cancer development and therapy. Nat Rev Clin Oncol 2021;18:645–61.
[14] He F, Ru X, Wen T. NRF2, a transcription factor for stress response and beyond. Int J Mol Sci 2020;21.
[15] Vashi R, Patel BM. NRF2 in cardiovascular diseases: A Ray of Hope! J Cardiovasc Transl Res 2021;14:573–86.
[16] Abati E, Bresolin N, Comi G, et al. Silence superoxide dismutase 1 (SOD1): a promising therapeutic target for amyotrophic lateral sclerosis (ALS). Expert Opin Ther Targets 2020;24:295–310.
[17] King EA, Davis JW, Degner JF. 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 2019;15:e1008489.
[18] Ochoa D, Karim M, Ghoussaini M, et al. Human genetics evidence supports two-thirds of the 2021 FDA-approved drugs. Nat Rev Drug Discov 2022;21:551.
[19] Cully M. Target validation: Genetic information adds supporting weight. Nat Rev Drug Discov 2015;14:525.
[20] Quan Y, Wang ZY, Chu XY, et al. Evolutionary and genetic features of drug targets. Med Res Rev 2018;38:1536–49.
[21] Liebeskind BJ, McWhite CD, Marcotte EM. Towards Consensus Gene Ages. Genome Biol Evol 2016;8:1812–23.
[22] Chu XY, Jiang LH, Zhou XH, et al. Evolutionary origins of cancer driver genes and implications for cancer prognosis. Genes (Basel) 2017;8.
[23] Xu X, Zhang QY, Chu XY, et al. Facilitating Antiviral Drug Discovery Using Genetic and Evolutionary Knowledge. Viruses 2021;13.
[24] Quan Y, Liang F, Wu D, et al. Blood cell DNA methylation of aging-related ubiquitination gene DZIP3 can predict the onset of early stage colorectal cancer. Front Oncol 2020;10:544330.
[25] Trigos AS, Pearson RB, Papenfuss AT, et al. Altered interactions between unicellular and multicellular genes drive hallmarks of transformation in a diverse range of solid tumors. Proc Natl Acad Sci USA 2017;114:6406–11.
[26] Zhou XH, Chu XY, Xue G, et al. Identifying cancer prognostic modules by module network analysis. BMC Bioinforma 2019;20:85.
[27] Xie T, Yang QY, Wang XT, et al. Spatial colocalization of human ohnolog pairs acts to maintain dosage-balance. Mol Biol Evol 2016;33:2368–75.
[28] Quan Y, Luo ZH, Yang QY, et al. Systems chemical genetics-based drug discovery: prioritizing agents targeting multiple/reliable disease-associated genes as drug candidates. Front Genet 2019;10:474.
[29] Lu S, Shen Q, Zhang J. Allosteric methods and their applications: facilitating the discovery of allosteric drugs and the investigation of allosteric mechanisms. Acc Chem Res 2019;52:492–500.
[30] Lu S, He X, Ni D, et al. Allosteric modulator discovery: from serendipity to structure-based design. J Med Chem 2019;62:6405–21.
[31] Nussinov R, Tsai CJ. Allostery in disease and in drug discovery. Cell 2013;153:293–305.
[32] Guarnera E, Berezovsky IN. Allosteric sites: remote control in regulation of protein activity. Curr Opin Struct Biol 2016;37:1–8.
[33] Wenthur CJ, Gentry PR, Mathews TP, et al. Drugs for allosteric sites on receptors. Annu Rev Pharm Toxicol 2014;54:165–84.
[34] Lu S, Ji M, Ni D, et al. Discovery of hidden allosteric sites as novel targets for allosteric drug design. Drug Discov Today 2018;23:359–65.
[35] Ji S, Pan S, Cambria E, et al. A survey on knowledge graphs: representation, acquisition, and applications. IEEE Trans Neural Netw Learn Syst 2022;33:494–514.
[36] Zeng X, Tu X, Liu Y, et al. Toward better drug discovery with knowledge graph. Curr Opin Struct Biol 2022;72:114–26.
[37] Luo ZH, Shi MW, Yang Z, et al. pyMeSHSim: an integrative python package for biomedical named entity recognition, normalization, and comparison of MeSH terms. BMC Bioinforma 2020;21:252.
[38] Simoens S, Picavet E, Dooms M, et al. Cost-effectiveness assessment of orphan drugs: a scientific and political conundrum. Appl Health Econ Health Policy 2013;11:1–3.
[39] Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Yearb Med Inf 1993:41–51.
[40] Capra JA, Stolzer M, Durand D, et al. How old is my gene? Trends Genet 2013;29:659–68.
[41] Maxwell EK, Schnitzler CE, Havlak P, et al. Evolutionary profiling reveals the heterogeneous origins of classes of human disease genes: implications for modeling disease genetics in animals. BMC Evol Biol 2014;14:212.
[42] King MC, Wilson AC. Evolution at two levels in humans and chimpanzees. Science 1975;188:107–16.
[43] Ha D, Kim D, Kim I, et al. Evolutionary rewiring of regulatory networks contributes to phenotypic differences between human and mouse orthologous genes. Nucleic Acids Res 2022;50:1849–63.
[44] Ryaboshapkina M, Hammar M. Tissue-specific genes as an underutilized resource in drug discovery. Sci Rep 2019;9:7233.
[45] Luleci HB, Yilmaz A. Robust and rigorous identification of tissue-specific genes by statistically extending tau score. BioData Min 2022;15:31.
[46] MacLean F. Knowledge graphs and their applications in drug discovery. Expert Opin Drug Discov 2021;16:1057–69.
[47] Zeng X, Song X, Ma T, et al. Repurpose Open Data to Discover Therapeutics for COVID-19 Using Deep Learning. J Proteome Res 2020;19:4624–36.
[48] Wang D, Zeng S, Xu C, et al. MusiteDeep: a deep-learning framework for general and kinase-specific phosphorylation site prediction. Bioinformatics 2017;33:3909–16.
[49] Zhao Q, Zhao H, Zheng K, et al. HyperAttentionDTI: improving drug-protein interaction prediction by sequence-based deep learning with attention mechanism. Bioinformatics 2022;38:655–62.
[50] Xia C, Misra I, Iyanagi T, et al. Regulation of interdomain interactions by calmodulin in inducible nitric-oxide synthase. J Biol Chem 2009;284:30708–17.
[51] Nikkari ST, Maatta KM, Kunnas TA. Functional Inducible Nitric Oxide Synthase Gene Variants Associate With Hypertension: A Case-Control Study in a Finnish Population-The TAMRISK Study. Med (Baltim) 2015;94:e1958.
[52] Benton ML, Abraham A, LaBella AL, et al. The influence of evolutionary history on human health and disease. Nat Rev Genet 2021;22:269–83.
[53] Perry GH. Evolutionary medicine. Elife 2021;10.
[54] Wu P, Clausen MH, Nielsen TE. Allosteric small-molecule kinase inhibitors. Pharm Ther 2015;156:59–68.
[55] Makino T, McLysaght A. Ohnologs in the human genome are dosage balanced and frequently associated with disease. Proc Natl Acad Sci USA 2010;107:9270–4.