Cell & Bioscience

DATABASE

Open Access

Reproductive chemical database: a curated database of chemicals that modulate protein targets regulating important reproductive biological processes

Yuedi Cao[1†], Geng G. Tian[1,2†], Xiaokun Hong[3], Qing Lu[1], Ting Wei[3*], Hai‑Feng Chen[3*] and Ji Wu[1,4*]

Abstract

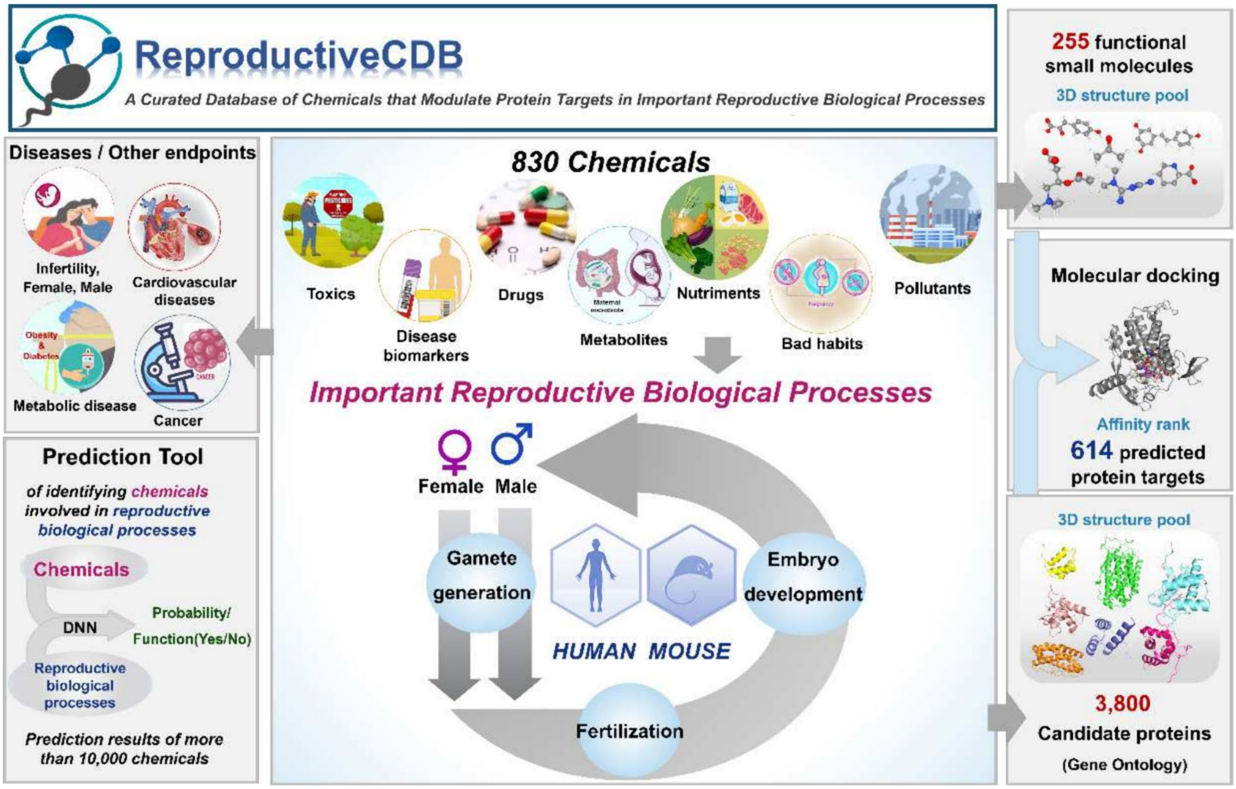
Recent studies have shifted the spotlight from adult disease to gametogenesis and embryo developmental events, and these are greatly affected by various environmental chemicals, such as drugs, metabolites, pollutants, and others. Growing research has highlighted the critical importance of identifying and understanding the roles of chemicals in reproductive biology. However, the functions and mechanisms of chemicals in reproductive processes remain incomplete. We developed a comprehensive database called the Reproductive Chemical Database (RCDB) (https://yu.life.sjtu.edu.cn/ChenLab/RCDB) to facilitate research on chemicals in reproductive biology. This resource is founded on rigorous manual literature extraction and precise protein target prediction methodologies. This database focuses on the delineation of chemicals associated with phenotypes, diseases, or endpoints intricately associated with four important reproductive processes: female and male gamete generation, fertilization, and embryo development in human and mouse. The RCDB encompasses 93 sub-GO processes, and it revealed 1447 intricate chemical–biological process interactions. To date, the RCDB has meticulously cataloged and annotated 830 distinct chemicals, while also predicting 614 target proteins from a selection of 3800 potential candidates. Additionally, the RCDB offers an online predictive tool that empowers researchers to ascertain whether specific chemicals play discernible functional roles in these reproductive processes. The RCDB is an exhaustive, cross-platform, manually curated database, which provides a user-friendly interface to search, browse, and use reproductive processes modulators and their comprehensive related information. The RCDB will help researchers to understand the whole reproductive process and related diseases and it has the potential to promote reproduction research in the pharmacological and pathophysiological areas.

Keywords Database, Chemical, Reproduction, Function prediction, Target protein screening

[†] Yuedi Cao and Geng G. Tian contributed equally to this work.

*Correspondence:
Ting Wei
weitinging@sjtu.edu.cn
Hai-Feng Chen
haifengchen@sjtu.edu.cn
Ji Wu
jiwu@sjtu.edu.cn
Full list of author information is available at the end of the article

Graphical Abstract


ReproductiveCDB
A Curated Database of Chemicals that Modulate Protein Targets in Important Reproductive Biological Processes
Diseases / Other endpoints
Infertility, Female, Male
Cardiovascular diseases
Obesity & Diabetes
Metabolic disease
Cancer
830 Chemicals
Toxics
Disease biomarkers
Drugs
Metabolites
Nutriments
Bad habits
Pollutants
Important Reproductive Biological Processes
Female Male
Gamete generation
Embryo development
HUMAN MOUSE
Fertilization
Prediction Tool
of identifying chemicals involved in reproductive biological processes
Chemicals
DNN
Probability/ Function(Yes/No)
Reproductive biological processes
Prediction results of more than 10,000 chemicals
255 functional small molecules
3D structure pool
Molecular docking
Affinity rank
614 predicted protein targets
3D structure pool
3,800 Candidate proteins
(Gene Ontology)

Background

In mammals, the reproductive process begins with the selection of the fittest gametes, which then follow a unique developmental pathway, participate in fertilization, and ultimately give rise to a viable embryo capable of developing into a new life [1]. These complex and highly organized activities require a range of factors deliberately selected to meet the dynamic needs of life [2–4], and disruptions by chemicals may cause severe outcomes, such as abnormal development [5–7], infertility [8], and even transgenerational or multigenerational effects [2, 9, 10]. The use of preconception antidiabetic drugs in men, such as metformin, is associated with birth defects in their offspring [10, 11]. Decades of research have highlighted the critical role of metabolism in preimplantation development [12–14]. In a recent study, Zhao et al. observed differences in 76 metabolites between two-cell embryos and blastocysts during mouse embryo development using ultra-sensitive metabolomics. L-2-hydroxyglutarate affects embryo development by inhibiting the erasure of H3K4me3 methylation, which is involved in metabolic pathways [15]. Besides direct chemical exposure or transformation, the maternal microbiome can also modulate fetal neurodevelopment in mice, probably through signaling by microbially modulated metabolites to neurons in the developing brain [16, 17]. Organisms regulate the functions of large macromolecules, such as DNA and protein, to cope with the constant changing environment, through interactions occurring between macromolecules and small molecules (i.e., small organic molecules). Small molecules are responsible for the essential processes of life, such as energy production, signaling, and pathogen eradication. Therapeutic drugs are often small molecules, which regulate the functions of macromolecules by interacting with them. Therefore, understanding whether a small molecule can interact with a macromolecule at the cellular and molecular levels, especially in the fields of medicine and molecular biology, is important [18].

Chemicals have been extensively used in many aspects on the basis of their effect on reproductive processes mentioned above. The detection of the metabolomics during important reproductive processes will help in further understanding the metabolic regulation of development [15]. Additionally, the microenvironment could contain metabolomics biomarkers for some diseases, providing new knowledge to improve the efficiency of critical techniques such as in assisted reproductive technology or to predict and improve outcomes in reproductive medicine [19]. In the field of drug development for the reproductive system, a research team developed an optimized TDI-10229 inhibitor for soluble adenylyl cyclase, which is crucial for mouse spermatogenesis [20]. This breakthrough demonstrates the possibility of two groundbreaking human contraception methods: non-hormonal male contraception and on-demand pharmacological contraception. Drug-induced or chemical-induced reproductive toxicity can lead to extremely serious outcomes, and reproductive toxicity accounts for 3% of cases of drug withdrawal/discontinuation [21]. Many studies used machine learning methods to identify and characterize compounds with reproductive toxicity from chemicals [22–24].

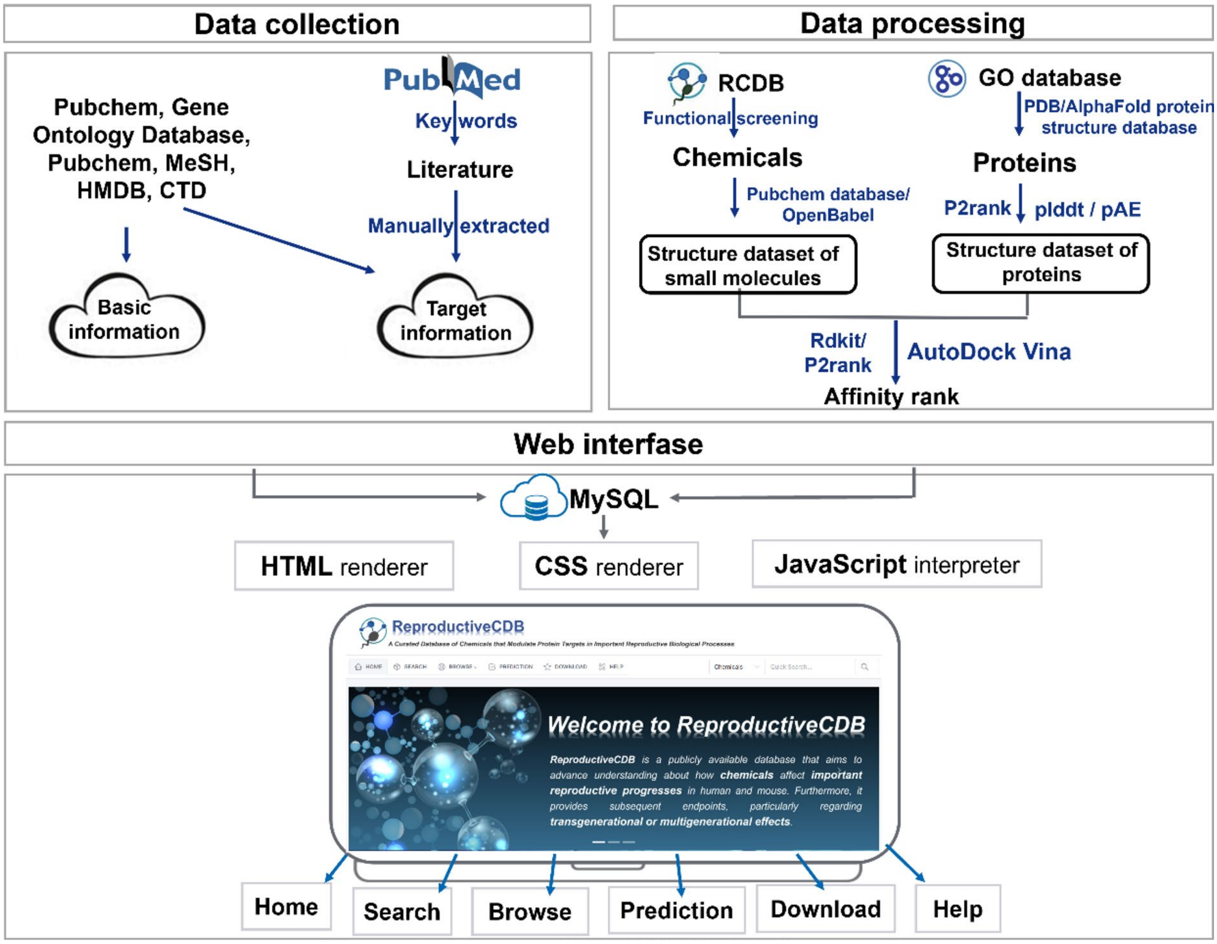
There are some large chemically-indexed databases, such as PubChem [25], Drugbank [26], ChEMBL [27], KEGG [28], and Comparative Toxicogenomics Database (CTD) [29], which do not focus on reproduction. Therefore, these databases have disadvantages, including incomplete information and difficulty in obtaining target information on reproduction. However, some databases include small molecules of reproductive toxicity, such as Toxin and Toxin-Target Database (T3DB) [30], TOXRIC [31], European Chemicals AgencyC&L Inventory, and the OECD eChemPortal. These data provide a single endpoint, such as positive (reproductive toxicants) or negative (non-reproductive toxicants), instead of the molecular mechanism/phenotype and do not reflect the diversity of phenotypes. Therefore, the application of these data for predicting toxicity is limited. As an integral part of biological processes in living organisms and the foundation of target identification and drug discovery, chemical–protein interactions are determined by computational methods. These methods circumvent the time and cost associated with experimental approaches. Information on these interactions is dispersed over many databases, texts, and prediction methods. However, researchers seeking a comprehensive understanding of important reproductive biological processes face a challenge in gathering relevant information because it is scattered across thousands of publications. Moreover, there is currently no database available that consolidates chemicals from various data sources that are pertinent to these biological processes.

To overcome these limitations, we developed the Reproductive Chemical Database (RCDB) (https://yu.life.sjtu.edu.cn/ChenLab/RCDB) of chemicals and their predicted protein targets that modulate corresponding processes. The RCDB has three key features. (i) This database is a comprehensive and integrative platform for reproductive chemicals. The RCDB is a versatile database that provides a multitude of annotations surrounding small molecules, including their basic information, phenotypes, related diseases, and predicted protein targets. The Gene Ontology (GO) biological process is used to define and standardize the functional periods of these phenotypes and diseases, with annotations that are restricted to the important reproductive processes. The RCDB is an extensive collection of annotations derived from over 600 literature sources, and it comprises more than 800 chemicals. These annotations provide a comprehensive view of their roles in the reproductive process and related diseases as reported in the literature. Additionally, this database offers excellent candidate molecules, such as metabolomics data that significantly change during the preferred reproductive biological process and markers for some diseases, and even new chemicals lacking functional evidence. (ii) The RCDB uses molecular docking to efficiently narrow the screening scope of target proteins. This database allows creative use of reverse docking to identify suitable protein targets for small molecules according to their functions in the GO biological processes. (iii) The RCDB provides datasets for researchers interested in investigating small molecules in important reproductive processes. The RCDB provides comprehensive, machine learning-ready sub-datasets for "how the endpoints occurred" instead of "endpoints that have already occurred", which can be downloaded and used as input/output for machine learning models.

Construction and content

Data collection and content

The general process of data collection is illustrated in Fig. 1. Data collection primarily falls into two major categories. The first category involves obtaining basic information and some target information by leveraging public databases, while the second involves the compilation of target information through a manual literature review. We searched PubMed using a list of keywords, such as 'chemicals', 'compound(s)', 'metabolite(s)' and 'small molecule(s)', and reproductive processes related keywords including but not limited to 'embryo development', 'oogenesis', 'spermatogenesis', 'fertilization' and their synonyms, as well as species information ('human' OR 'mouse' OR 'mice') and finally obtained more than 1000 related publications. To reliably collect the information, professional biocurators who major in reproductive biology engage in the manual curation of scientific literature. We employ a system



***ReproductiveCDB** is a publicly available database that aims to advance understanding about how **chemicals** affect **important reproductive progresses** in human and mouse. Furthermore, it provides subsequent endpoints, particularly regarding **transgenerational or multigenerational effects**.*

Fig. 1 Overview of data collection, processing, and database interface. Data collection involves obtaining target information through using publicly available databases. Data processing primarily involves the molecular docking analysis of chemicals and proteins, which is achieved by using AutoDock Vina software. The web interface is mainly divided into six parts, comprising "Home," "Search," "Browse," "Prediction," "Download," and "Help"

of peer review and regular discussions to guarantee the data's accuracy and careful to focus on collecting chemical-induced phenotypes to keep our resources focused on chemical-centric data. In terms of data collection, we conduct a professional manual curation of the acquired literature, aiming to extract target information that primarily includes chemical molecules, species, associated biological processes, relevant diseases and information, summaries derived from the original texts, species, and PMID. Furthermore, we employ standardized definitions of the relationships between small molecules and biological processes by utilizing terms such as 'increase', 'decrease', 'affect' and 'not'. Furthermore, to define phenotypes, we incorporated the GO database as a vocabulary source for non-disease biological outcomes. The GO resource is a widely recognized, community-accepted vocabulary with accession identifiers, ensuring that chemical–phenotype interactions are computable and interoperable with other databases. With regard to disease annotation, various diseases are standardized using the standard terminology in Medical Subject Headings. We referred to common methods of definition for the definition of generation [32]. Effects in the F1 and F2 generations are considered multigenerational, and effects in the F3 generation are considered transgenerational.

Data processing

Data generation of potential targets in humans and mice

Two lists of genes of important biological processes in humans and mice were created from the GO database [33]. Genes without a UniProtKB ID were discarded because the AlphaFold Protein Structure Database only contains models with an entry in UniProtKB [34]. All protein models in PDB format were downloaded from the Protein Data Bank database. If not available, the models will be downloaded from the AlphaFold Protein Structure Database downloads section on 12 October

2022. Selected genes without a model were a consequence of their products' length being larger than 2700 residues, and these were also discarded because AlphaFold models were not available for such lengths. Binding pocket prediction was performed using P2Rank [35] with standard settings. Pockets with a probability score (as provided by P2Rank) > 0.1 were considered as candidates for binding sites. In each model, the pocket with the highest probability score was selected as the binding site. Structures without predicted pockets or predicted pockets with a probability score < 0.1 were discarded. To further assess and shelter the AlphaFold source model quality, the predicted Local Distance Difference Test (pLDDT) [36] score of each α-carbon was extracted from the Protein Data Bank files. The proportion of residues with a pLDDT score > 70 (described as the threshold for good backbone prediction) [37] was then calculated. Only models with at least half of its total residues with a pLDDT score > 70 were considered for docking. Additionally, to assess the local quality of the binding pocket, the residues predicted to be part of the pocket by P2Rank were considered. Only models in which at least half of the residues in the pocket had a pLDDT score > 90 were kept. This stricter threshold was chosen because residues with a pLDDT score > 90 can be interpreted as having high quality and correct side-chain orientation [37]. Finally, the predicted aligned error (pAE) of the pocket residues was also analyzed. Therefore, the mean pAE of each residue of the pocket (as specified by P2Rank) with the rest of the residues of the pocket was calculated, and the overall mean pAE was obtained. Any model with a mean pocket pAE > 5 Å was discarded.

Generation of human and mouse datasets of functional ligands

To choose ligands, we conducted a meticulous search within the RCDB, and specifically focused on compounds associated with reproductive biological processes. Ligand structures were retrieved from PubChem [25] as three-dimensional (3D) SDF files whenever available. If 3D files were not accessible, we downloaded them as 2D SDF files and used Open Babel to generate the definitive 3D conformations [38]. The size of the binding box for each ligand was optimized according to Feinstein and Brylinski [39], using a radius of gyration to box side ratio of 0.35, and rounding up to the nearest integer. The radius of gyration for each ligand was calculated using the Python RDKit library Descriptors3D module.

Reverse docking for functional ligands

Docking simulations were performed using AutoDock Vina 1.1.2 [40]. The exhaustiveness parameter was set to 8 and the energy was set to 2. The search box center was chosen from the P2Rank predictions, and its size was calculated for each ligand as described above. Ten docking repetitions were performed with different random seeds for each receptor and ligand pair. The best mode for each pair was chosen from the lowest docking energy of all the repetitions. This resulted in a matrix of n ligands by m receptors, with the best possible energy for each pair. The results were visualized using Jsmol [41].

Website implementation

The RCDB website was implemented by using Hypertext Markup Language, Cascading Style Sheets, and JavaScript running on top of a MySQL database.

Identification of interactions between chemicals and important reproductive biological processes based on machine learning

Chemicals and their interactions with important reproductive biological processes were extracted from the RCDB. The RCDB provides positive and negative chemical–biological process interactions (CBIs). Any combination of chemical and biological processes constitutes a CBI. An example of this combination is that CBI (1,1) is constructed by chemical 1 and biological process 1. The features of CBI (1,1) are the combination of the features of biological process 1 and compound 1, and then we apply a z-score for normalization. Finally, we obtained 1020 CBIs in the mouse from the RCDB. Among these CBIs, 320 CBIs were positive and the others were negative. To show the convenience of the RCDB, we developed a deep learning-based method to predict CBIs. The workflow of our method is shown in Fig. 2.

There were three steps in our method for data preprocessing: construction of chemical features, construction of biological process features, and CBIs. To construct chemical features, the 2D SDF files of the chemicals were input into PaDEL-Descriptor software [42] to obtain 1444 dimensional 1D and 2D descriptors and 12 types of fingerprints for each chemical. We then calculated various compound–compound similarities by the Jaccard index [43, 44]. To construct biological process features, we assessed the semantic similarity between GO terms, which was based on an encoding method that quantitatively represents the semantic content (biological meanings) of GO terms. The encoding method integrates the semantic contributions from ancestral terms, including the specific term itself, within the GO graph [45]. To implement tenfold cross-validation, we randomly divided the positive and negative samples into 10 groups. One of the positive groups and one of the negative groups were then used as the testing set, and all of the other groups were used as the training set. This process

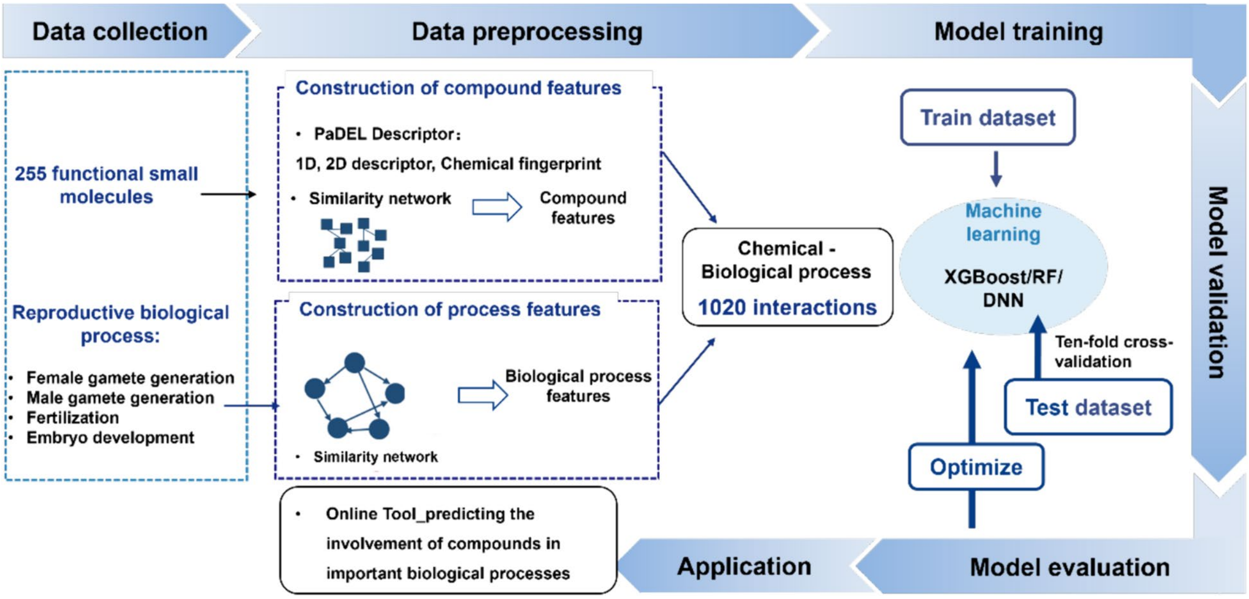

Data collection
Data preprocessing
Model training
255 functional small molecules
Reproductive biological process:
• Female gamete generation
• Male gamete generation
• Fertilization
• Embryo development
Construction of compound features
• PaDEL Descriptor:
1D, 2D descriptor, Chemical fingerprint
• Similarity network
Compound features
Construction of process features
Biological process features
• Similarity network
Chemical - Biological process
1020 interactions
Train dataset
Machine learning
XGBoost/RF/ DNN
Ten-fold cross-validation
Test dataset
Optimize
Model validation
• Online Tool_predicting the involvement of compounds in important biological processes
Application
Model evaluation

Fig. 2 Workflow for constructing machine learning (ML) models

ensured that the ratio of positive and negative samples was the same between the training set and the testing set.

Collection and treatment of mouse early embryos

All animal experiments followed procedures approved by the Institutional Animal Care and Use Committee of Shanghai Jiao Tong University. Six- to eight-week-old ICR/CD1 mice were collected and intraperitoneally injected with 10-IU pregnant mare serum gonadotropin (PMSG), followed by a human chorionic gonadotropin (hCG; 10 IU) injection 48 h later. Then, we mated those superovulated females with ICR/CD1 males and collected the zygotes from the oviducts of female mice at 20 h post-hCG injection and cultured in KSOM medium as control group. The treatment group of octyl octanoate dissolved in 0.1% ethanol were added to reach a final concentration of 0.5 μM in the culture medium. Early 2-cell and blastocyst stage embryos were harvested at 34 h, and 4 days post-hCG injection, respectively.

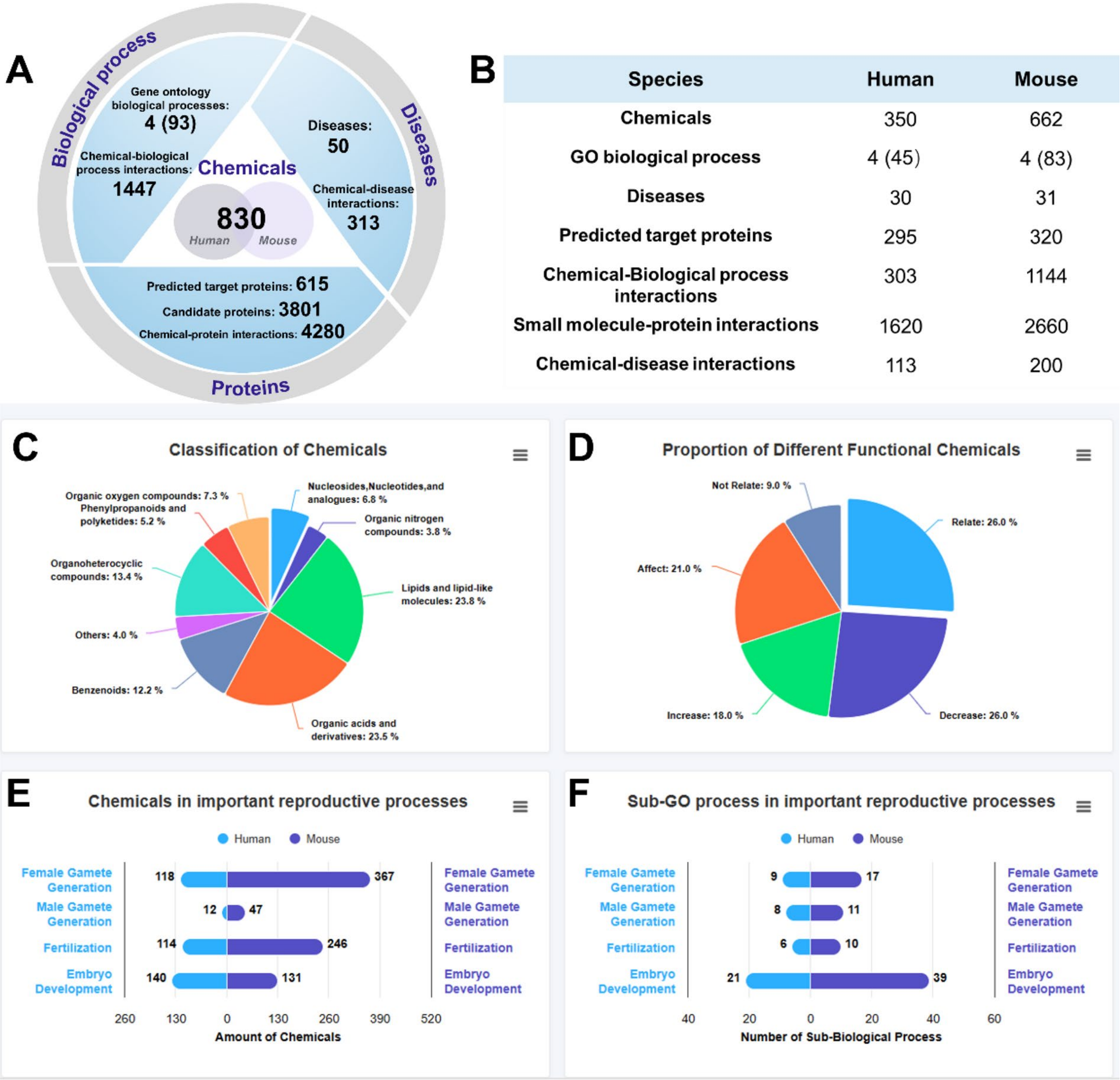
Results

Overview of the RCDB

The RCDB contains a map of 350 and 662 related chemicals in humans and mice, respectively, involving 30 diseases and 45 sub-GO process in human, and 31 diseases and 83 sub-GO process in mouse (Fig. 3B). These markers were derived from 4124 annotation entries collected from more than 600 published reports. On the basis of the focus on chemical properties and functionality, we classified compounds using pie charts (Fig. 3C). These graphs showed that the chemicals collected in the database exhibited a rich diversity. This finding suggests that the occurrence and regulation of reproductive processes require the participation of multiple metabolites, and various compounds in the environment may have a considerable effect on reproductive processes. Additionally, the diverse definitions of functionality and the relatively balanced distribution can meet the different requirements of a wider range of users. The amount of chemicals in important reproductive biological processes and the sub-GO process in important reproductive processes are shown in bar plots (Fig. 3F). The associated compounds and sub-GO processes were more abundant than the other three processes owing to the complexity of embryo development.

Web interface

The RCDB's public user interface allows intuitive browsing and searching of any data in the database. On the "Home" page, users can quickly explore reproductive processes by clicking on hyperlinks embedded in the web images—Important reproductive processes in the human or mouse (Fig. 4A). After clicking on the icons of different GO biological processes, users can further select relevant sub-GO processes in a pop-up window to retrieve the corresponding chemicals. In addition, the home page provides a quick search utility, which can be used to query the database for chemicals, proteins, biological processes, or diseases. In the "Browse" page, users can easily access chemicals, proteins, biological processes, and diseases by clicking on a dropdown



Species	Human	Mouse
Chemicals	350	662
GO biological process	4 (45)	4 (83)
Diseases	30	31
Predicted target proteins	295	320
Chemical-Biological process interactions	303	1144
Small molecule-protein interactions	1620	2660
Chemical-disease interactions	113	200

Fig. 3 Statistical results of data stored in the RCDB. **A** Overview of the statistical situation of the database **B** Statistical results of humans and mice in the database. **C** Chemical classification in the database. **D** Proportion of different functional chemicals. **E** Number of chemicals in important reproductive processes. **F** Number of sub-GO processes of important reproductive processes

menu, and the complete list of matched entries can be returned (Fig. 4A). To perform data retrieval, the RCDB "Search" interface is comprised of several search options, including text search for chemicals, proteins, biological processes, and diseases (Fig. 4B). To perform a chemical search, the general name, synonyms, and CID number can be used as an input (e.g., in a search for spermidine, the following term can be used: "spermidine"). A specific protein can be queried using a protein name or gene name (e.g., a search for P07686 can be performed using "P07686" or "HCC7"). With regard to biological processes, the biological process name or GO ID can be used for searching (e.g., prostatic bud formation can be queried using "Prostatic Bud Formation" or the GO ID "0060513"). In the "disease search" section, users can search for the disease name (e.g., a search for infertility can be performed using "infertility"). In a search for the biological process and disease, all of the biological processes and diseases annotated in the database are below the search boxes for references. With clicking on the catalogue, it will automatically populate the content into the search box. And after clicking the "Go!" button,

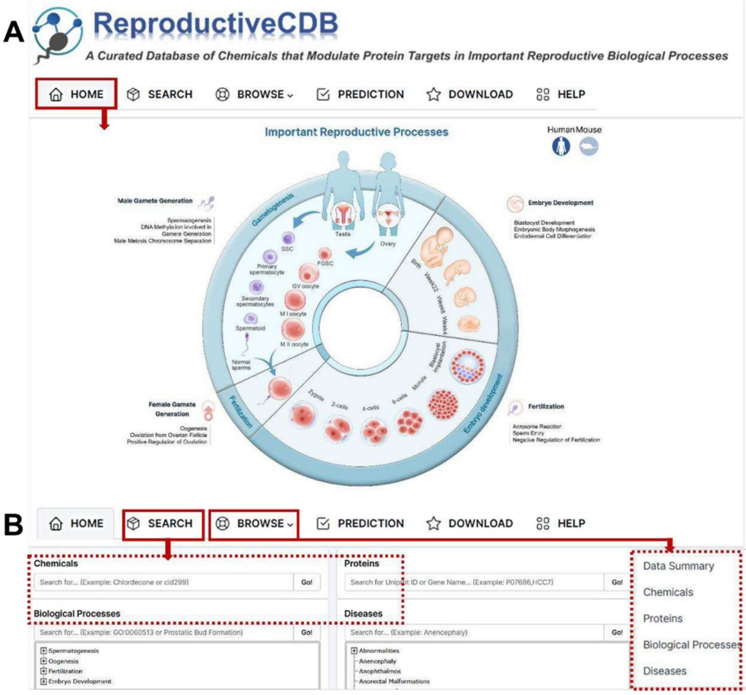


Fig. 4 Schematic workflow of the RCDB. **A** The web images in the home page allow quickly search for important reproductive biological processes. **B** The "Browse" and "Search" pages allow the users to browse and search chemicals

the search engine will return a result page showing comprehensive information.

In addition, the users can download all data via the "Download" page and also find a detailed tutorial on how to use the database on the "Help" page. If users have any questions, comments, or new research data and applications related to our database that have not yet been updated, they can give us feedback through the contact module.

Data retrieval from the RCDB

We developed convenient web-based modules to help users quickly retrieve target information and detailed annotations from the RCDB (Fig. 5A–D). As an example of conducting targeted searches for chemicals, such as glucose, the search results include four parts, basic information including a 2D structure image, the chemical name, molecular weight, and canonical SMILES. Additionally, links to other databases, such as Human Metabolome Database (HMDB) [46] and KEGG, have been implemented on the small molecule page, enabling users to access associated target pages. The reproductive function annotation in our database encompasses chemical-induced phenotypes, action qualifiers, function periods (including crucial reproductive processes, such as female and male gamete generation, fertilization, and embryo development), and specific phenotype details with a GO ID, short description, and corresponding PubMed reference. In the related disease annotation part, the database provides related diseases of chemicals with functional biological processes, species, and generations. We also collected some endpoints, such as low birth weight, preterm birth, and perinatal complications. Low birth weight is associated with an increased risk for type 2 diabetes [47], hypertension, and cardiovascular disease. The underlying cause of low birth weight remains obscured by multiple factors, such as the potential influence of chemicals. With regard to protein target prediction, a table of predicted proteins with further detailed information is provided according to the functions of chemicals (Fig. 5A). In the search results for others, the database provides the basic information for the biological process (GO name, definition, and synonyms), disease (disease name, synonyms, categories, and Medical Subject Headings ID), and protein (Entry, Protein name, Gene name, Organism). The related chemicals are listed accordingly for all of these targets.

Identification of chemicals within reproductive biological processes

In order for researchers to better identify chemicals associated with reproductive biological processes, we have predicted the involvement of 10,396 compounds that downloaded from PubChem database to find the

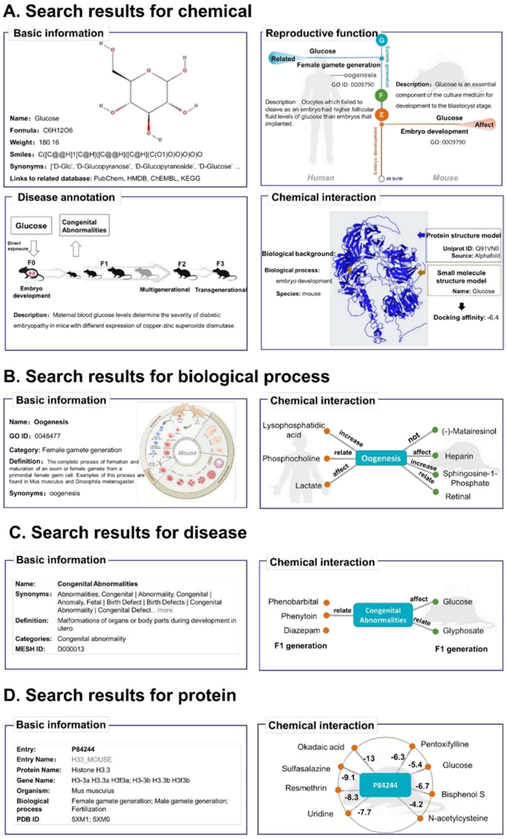


Fig. 5 Overview of search results of the RCDB. **A** Results of a search for glucose, including basic information, reproductive function, disease annotation, and predicted protein target prediction. **B** Results of a search for the biological process of oogenesis, including basic information and a related chemical information table. **C** Results of a search for disease of congenital abnormalities, including basic information and a related chemical information table. **D** Results of a search for the protein P84244, including basic information and chemical interactions

chemicals that play roles in the important reproductive processes in the mouse. An overview of prediction result of chemicals involved in reproductive biological processes is shown in Fig. 6. This information is presented in a table format, including the compound's CID, the biological process's GO ID, and the correlation probability between them. A final assessment of their relevance (Yes/No) is then made. The search functionality is also supported, and users can download the table containing all predicted results of compound analysis from the "Download" page.

We compared the performance of different methods, such as deep neural network (DNN), random forest, and extreme gradient boosting. We performed tenfold cross-validation and obtained the area under the curve, Accuracy, Recall, F1-score, and Precision for each method. The average and standard deviation of the area under the curve and the area under the precision-recall curve (AUPR) from tenfold cross-validation were calculated for each method (Table 1). DNN performed the best, with the highest area under the curve of 0.7052. We built a DNN model using all chemical–reproductive biological process associations and then predicted novel associations.

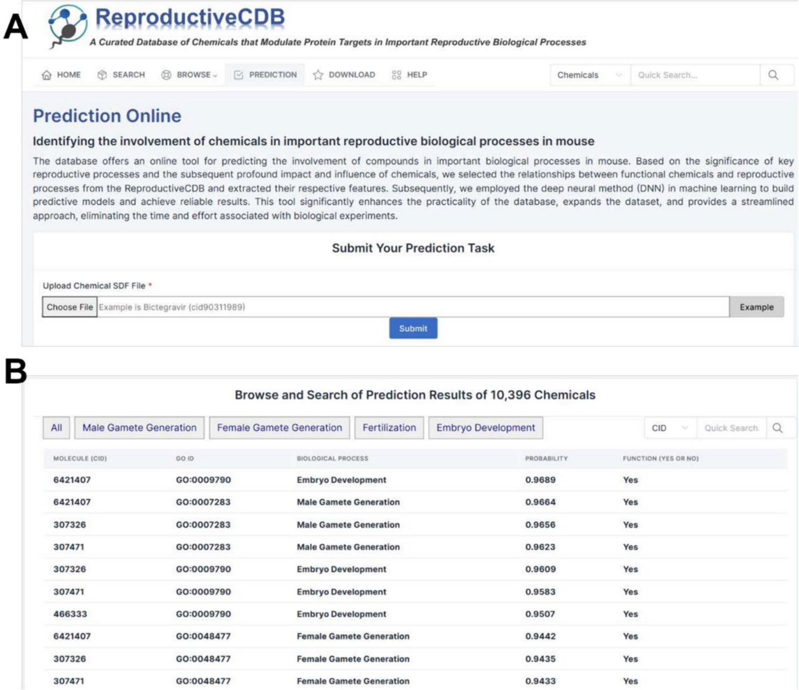
Because all known associations were used to construct the prediction model, the predicted associations require verification by published literature or other available sources. We are currently confirming the accuracy of our prediction model through two approaches. First, we select the top-ranked small molecules from the predicted results, and second, we validate them through cellular experiments.

From the prediction result, "phthalic acid, 2-diethylaminoethyl pentyl ester" (CID: 6421407) is predicted to be involved in all four reproductive biological processes, with probabilities > 0.9. To investigate the principle of structurally similar molecules showing similar biological activities, we conducted a comparative analysis of structure and function between phthalic acid, 2-diethylaminoethyl pentyl ester and butyl benzyl phthalate (CID: 2347) with known reproductive activity. The parameters describe the results of a chemical structure similarity comparison (Fig. 7) analyzed by ChemMine Tools [48]. These two molecules mentioned above showed a high degree of structural similarity, and butyl benzyl phthalate showed functional involvement across all four major reproductive processes within the RCDB. This finding is in agreement with the prediction results for phthalic acid, 2-diethylaminoethyl pentyl ester in the database.

We also identified octyl octanoate (CID: 61294) as a potential contributor to embryo development (Fig. 8A). To validate our prediction, we conducted experiments using mouse embryos. Initially, we introduced octyl octanoate into the culture medium and co-cultured it with fresh embryos collected from the pregnant mouse for 4 days. Subsequently, we assessed the developmental progress and found that octyl octanoate inhibited mouse embryo development, resulting in arrest at the two-cell to eight-cell stage (Fig. 8B). Collectively, our findings strongly support the effectiveness of our chemical function prediction model.

Discussion

We have constructed a comprehensive, manually curated, user-friendly database (RCDB) for users to search, browse, and explore chemicals of interest. We also introduced the reverse docking method to assist researchers investigating target proteins of small molecules. This method provides a convenient reference for researchers from non-bioinformatics backgrounds


A
ReproductiveCDB
A Curated Database of Chemicals that Modulate Protein Targets in Important Reproductive Biological Processes
HOME
SEARCH
BROWSE
PREDICTION
DOWNLOAD
HELP
Chemicals
Quick Search...

Prediction Online

Identifying the involvement of chemicals in important reproductive biological processes in mouse

The database offers an online tool for predicting the involvement of compounds in important biological processes in mouse. Based on the significance of key reproductive processes and the subsequent profound impact and influence of chemicals, we selected the relationships between functional chemicals and reproductive processes from the ReproductiveCDB and extracted their respective features. Subsequently, we employed the deep neural method (DNN) in machine learning to build predictive models and achieve reliable results. This tool significantly enhances the practicality of the database, expands the dataset, and provides a streamlined approach, eliminating the time and effort associated with biological experiments.

Submit Your Prediction Task

Upload Chemical SDF File *

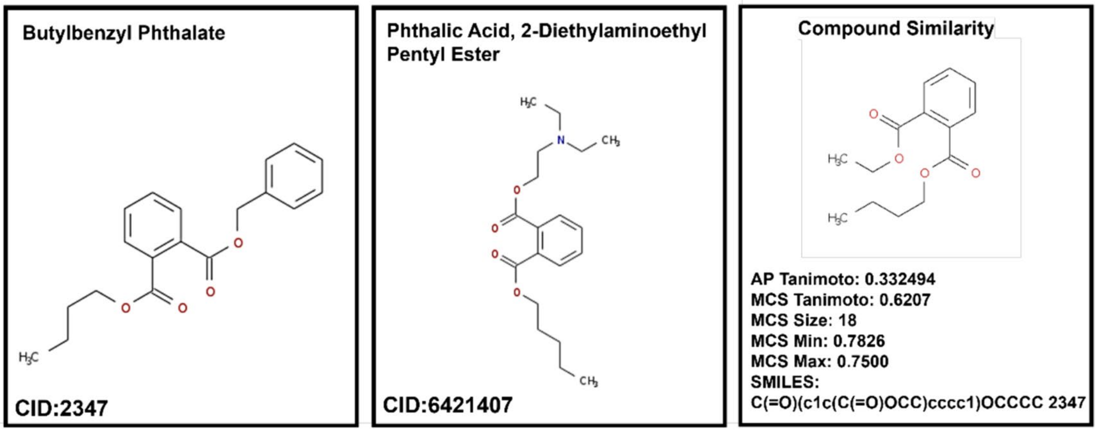
Browse and Search of Prediction Results of 10,396 Chemicals

MOLECULE (CID)	GO ID	BIOLOGICAL PROCESS	PROBABILITY	FUNCTION (YES OR NO)
6421407	GO:0009790	Embryo Development	0.9689	Yes
6421407	GO:0007283	Male Gamete Generation	0.9664	Yes
307326	GO:0007283	Male Gamete Generation	0.9656	Yes
307471	GO:0007283	Male Gamete Generation	0.9623	Yes
307326	GO:0009790	Embryo Development	0.9609	Yes
307471	GO:0009790	Embryo Development	0.9583	Yes
466333	GO:0009790	Embryo Development	0.9507	Yes
6421407	GO:0048477	Female Gamete Generation	0.9442	Yes
307326	GO:0048477	Female Gamete Generation	0.9435	Yes
307471	GO:0048477	Female Gamete Generation	0.9433	Yes

Fig. 6 Overview of a functional module for predicting chemicals involved in reproductive biological processes. **A** Web page of the online tool. **B** List of chemicals involved in important biological processes predicted by the DNN model

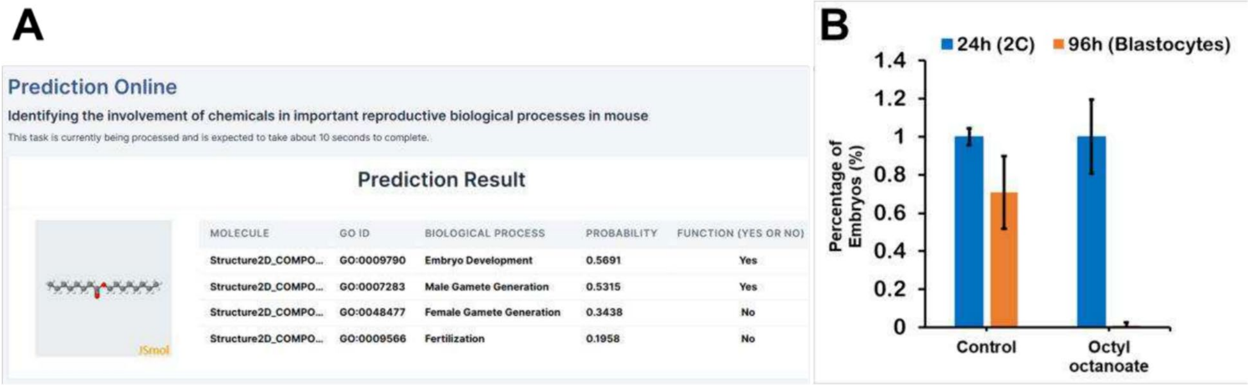
Table 1 Performance of SVM-based PMI identification

Method	AUC	Accuracy	Precision	Recall	F1-score
XGBoost	0.6918 (0.0323)	0.6990 (0.0353)	0.5581 (0.1028)	0.1789 (0.0910)	0.2620 (0.1022)
RF	0.5796 (0.0342)	0.7138 (0.0277)	0.6363 (0.1744)	0.2221 (0.0564)	0.3233 (0.0793)
DNN	0.7052 (0.0361)	0.7208 (0.0248)	0.6711 (0.1743)	0.5741 (0.2379)	0.4663 (0.1041)



AP Tanimoto: 0.332494
MCS Tanimoto: 0.6207
MCS Size: 18
MCS Min: 0.7826
MCS Max: 0.7500
SMILES:
C(=O)(c1c(C(=O)OCC)cccc1)OCCCC 2347

Fig. 7 Results of chemicals structure similarity comparison



Prediction Online

Identifying the involvement of chemicals in important reproductive biological processes in mouse

This task is currently being processed and is expected to take about 10 seconds to complete.

Prediction Result

MOLECULE	GO ID	BIOLOGICAL PROCESS	PROBABILITY	FUNCTION (YES OR NO)
Structure2D_COMPO...	GO:0009790	Embryo Development	0.5691	Yes
Structure2D_COMPO...	GO:0007283	Male Gamete Generation	0.5315	Yes
Structure2D_COMPO...	GO:0048477	Female Gamete Generation	0.3438	No
Structure2D_COMPO...	GO:0009566	Fertilization	0.1958	No

Fig. 8 Functional roles of octyl octanoate in mouse embryos. **A** Prediction results of octyl octanoate. **B** Octyl octanoate inhibits mouse embryo development

and serves as a valuable supplement to current prediction databases. The RCDB is a useful resource for researchers interested in comprehending the role of chemicals in reproductive biology. This database provides a solid foundation and platform for studying research and translational applications in reproductive development.

Considering the current research status and challenges in the field of reproduction, the RCDB has the potential to be used in the following aspects. (i) The RCDB can provide a holistic view of the role of chemicals in reproductive biological processes. The users of this database can investigate basic information, relevant functions, diseases, and predicted target proteins corresponding to chemicals' functional periods. (ii) The RCDB can provide ideas for new modulators of biological processes and potential drug candidates. Based on the predicted protein targets and biological process information, a series of new modulators (e.g., activators or inhibitors for oogenesis) might be designed and synthesized according to the known chemicals through some indirect mechanisms. (iii) The RCDB provides detailed information about related diseases. In this database, we have taken a novel approach to gathering related disease information and other endpoints, with a particular emphasis on transgenerational or multigenerational effects, which has not been reported in other databases. This information facilitates the association of diseases with relevant proteins and chemicals, which enables further studies of the underlying pathological mechanisms. Overall, in creating a comprehensive database of chemicals and their biological activities, this platform provides insights that could help guide the development of new therapeutic strategies to enhance reproductive health.

Some databases focus on important reproductive biological processes or diseases, such as the database "SpermatogenesisOnline 1.0", which is dedicated to genes related to spermatogenesis [49], or the database "IDDB" [50], which focuses on genes related to infertility. However, these databases are limited to single biological processes or diseases and do not emphasize the study of chemicals, which may limit the diversity of their applications and the joint analysis of data. Compared with databases that focus on compounds, general databases typically have larger datasets and address broader scientific questions, which may hinder the specialization of the database's application in the field of reproduction. An example of these differences between databases is that our target data are also reflected in other databases, such as T3DB, which aims to advance understanding on how environmental exposures affect human health [30]. In contrast to this database, we place more emphasis on metabolites that are non-toxic, but play major roles, and our annotations related to reproduction are more comprehensive. Additionally, we have added features, such as target protein prediction and molecular reproductive function prediction. Therefore, the RCDB serves as a unique resource in summarizing reproductive-related chemical data.

Further extensions of the following aspects will be conducted. First, with the advancement of technology and experimental techniques, more functional or related small molecules involved in important reproductive processes will be identified. Therefore, we will continue to track new studies and frequently update the RCDB with the addition of new chemicals' annotations. Second, chemicals of other species, such as zebrafish, *Drosophila*, and *Caenorhabditis elegans*, which are important model animals for reproduction research, will be added in a future version of the RCDB to provide more comprehensive information for the users. Moreover, comprehensive insights on livestock will be provided to recognize the pivotal role of reproductive efficiency as a paramount determinant of their substantial economic significance. Third, we will work to optimize the "biochemical interactions" part of chemicals. The protein target prediction of the database will be improved by adding a more promising and comprehensive protein target pool. With the advent of single-cell RNA sequencing technology, researchers can systematically and accurately characterize more proteins in various biological processes or cell populations. Additionally, small molecules not only interact with proteins to function, but also act with other macromolecules, such as DNA and RNA. Therefore, we will add more interaction information in the next release. Finally, with regard to the prediction tool, we will incorporate new data or data features and optimize machine learning methods. This will further optimize the prediction methods to enhance their predictive performance, and we also explore new prediction tasks, such as predicting more specific sub-GO biological processes. Additionally, we will validate more prediction results using cellular experiments and add them to the database. Furthermore, we will add other types of analysis and applications to the database to further enhance its use and expand its scope.

Conclusions

To our understanding, RCDB is the first uniquely dedicated database for the interpretation of the roles of chemicals in reproductive biological processes and the first database with prediction of the involvement of chemicals in GO reproductive biological processes. Taken together, our unique database holds significant importance in both biological and practical perspectives. Firstly, it facilitates the exploration and elucidation of regulatory blueprints governing important reproductive processes and pinpoint causes of many health issues. Secondly, it offers novel insights that support drug development, promote the development of the field of clinical applications in reproduction and so on.

Abbreviations

CTD	Comparative Toxicogenomics Database
T3DB	Toxin and Toxin-Target Database
pLDDT	Predicted Local Distance Difference Test
3D	Three-dimensional
CBIs	Chemical–biological process interactions
PMSG	Pregnant mare serum gonadotropin
hCG	Human chorionic gonadotropin
HMDB	Human Metabolome Database
DNN	Deep neural network
AUPR	Area under the precision-recall curve

Acknowledgements
Not applicable.

Author contributions

Y.C. and G.G.T. made the construct of database, T.W. performed the construct of prediction model. Y.C. did biological experiments and drafted the manuscript. X.H. and Q.L. reviewed the manuscript with concern. J.W. and H.C. supervised the experiment work and devised this study.

Funding
This work was supported by the National Key Research and Development Program of China (2022YFA0806301 and 2020YFA0907701), Center for HPC at Shanghai Jiao Tong University, the National Natural Science Foundation of China (32171242 and 21977068 and 22Z033003297 and 82201780), the Fundamental Research Funds for the Central Universities (YG2023LC03) and Key Research and Development Program of Ningxia Hui Autonomous Region (2022BFH02004).

Availability of data and materials

The datasets generated and analysed during the current study are all available at the website: https://yu.life.sjtu.edu.cn/ChenLab/RCDB. The implementation of prediction method and the preprocessed data is available at https://github.com/skytguuu/RCDB/tree/main.

Declarations

Ethics approval and consent to participate

Animal experimentation was approved by the Institutional Animal Care and Use Committee of Shanghai and performed in accordance with the National Research Council Guide for Care and Use of Laboratory Animals. The ethical approval number for our research is A2016084.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details
[1]Key Laboratory for the Genetics of Development & Neuropsychiatric Disorders (Ministry of Education), Bio-X Institutes, Shanghai Jiao Tong University, Shanghai 200240, China. [2]School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai 200240, China. [3]State Key Laboratory of Microbial Metabolism and Joint International Research Laboratory of Metabolic & Developmental Sciences, National Experimental Teaching Center for Life Sciences and Biotechnology, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China. [4]Key Laboratory of Fertility Preservation and Maintenance of Ministry of Education, Ningxia Medical University, Yinchuan 750004, China.

Received: 4 December 2023 Accepted: 31 May 2024
Published online: 06 June 2024

References

1. Wasserzug-Pash P, Klutstein M. Epigenetic changes in mammalian gametes throughout their lifetime: the four seasons metaphor. Chromosoma. 2019;128(3):423–41.
2. Chason RJ, Csokmay J, Segars JH, DeCherney AH, Armant DR. Environmental and epigenetic effects upon preimplantation embryo metabolism and development. Trends Endocrinol Metab. 2011;22(10):412–20.
3. Tian GG, Li J, Wu J. Alternative splicing signatures in preimplantation embryo development. Cell Biosci. 2020;10:33.
4. Tian GG, Hou C, Li J, Wu J. Three-dimensional genome structure shapes the recombination landscape of chromatin features during female germline stem cell development. Clin Transl Med. 2022;12(6): e927.
5. Porterfield SP. Thyroidal dysfunction and environmental chemicals–potential impact on brain development. Environ Health Perspect. 2000;108(suppl 3):433–8.
6. Grandjean P, Landrigan PJ. Developmental neurotoxicity of industrial chemicals. Lancet. 2006;368(9553):2167–78.

7. Zhu X, Tian GG, Yu B, Yang Y, Wu J. Effects of bisphenol A on ovarian follicular development and female germline stem cells. Arch Toxicol. 2018;92(4):1581–91.
8. Juul A, Almstrup K, Andersson A-M, Jensen TK, Jørgensen N, Main KM, et al. Possible fetal determinants of male infertility. Nat Rev Endocrinol. 2014;10(9):553–62.
9. Zou K, Ding G, Huang H. Advances in research into gamete and embryo-fetal origins of adult diseases. Sci China Life Sci. 2019;62(3):360–8.
10. Wensink MJ, Lu Y, Tian L, Shaw GM, Rizzi S, Jensen TK, et al. Preconception antidiabetic drugs in men and birth defects in offspring: a nationwide cohort study. Ann Intern Med. 2022;175(5):665–73.
11. Chen J, Wang L, Tian GG, Wang X, Li X, Wu J. Metformin promotes proliferation of mouse female germline stem cells by histone acetylation modification of Traf2. Stem Cell Rev Rep. 2023;19(7):2329–40.
12. Sharpley MS, Chi F, Hoeve J, Banerjee U. Metabolic plasticity drives development during mammalian embryogenesis. Dev Cell. 2021;56(16):2329–47.
13. Biggers JD, Whittingham DG, Donahue RP. The pattern of energy metabolism in the mouse oöcyte and zygote. Proc Natl Acad Sci USA. 1967;58(2):560–7.
14. Yuan X, Tian GG, Pei X, Hu X, Wu J. Spermidine induces cytoprotective autophagy of female germline stem cells in vitro and ameliorates aging caused by oxidative stress through upregulated sequestosome-1/p62 expression. Cell Biosci. 2021;11(1):107.
15. Zhao J, Yao K, Yu H, Zhang L, Xu Y, Chen L, et al. Metabolic remodelling during early mouse embryo development. Nat Metab. 2021;3(10):1372–84.
16. Vuong HE, Pronovost GN, Williams DW, Coley EJL, Siegler EL, Qiu A, et al. The maternal microbiome modulates fetal neurodevelopment in mice. Nature. 2020;586(7828):281–6.
17. Banerjee S, Suter MA, Aagaard KM. Interactions between environmental exposures and the microbiome: implications for fetal programming. Curr Opin Endocr Metab Res. 2020;13:39–48.
18. Chen L, Qian Z, Fen K, Cai Y. Prediction of interactiveness between small molecules and enzymes by combining gene ontology and compound similarity. J Comput Chem. 2010;31(8):1766–76.
19. Bracewell-Milnes T, Saso S, Abdalla H, Nikolau D, Norman-Taylor J, Johnson M, et al. Metabolomics as a tool to identify biomarkers to predict and improve outcomes in reproductive medicine: a systematic review. Hum Reprod Update. 2017;23(6):723–36.
20. Balbach M, Rossetti T, Ferreira J, Ghanem L, Ritagliati C, Myers RW, et al. On-demand male contraception via acute inhibition of soluble adenylyl cyclase. Nat Commun. 2023;14(1):637.
21. Siramshetty VB, Nickel J, Omieczynski C, Gohlke B-O, Drwal MN, Preissner R. WITHDRAWN—a resource for withdrawn and discontinued drugs. Nucleic Acids Res. 2015;44(D1):D1080–6.
22. Feng H, Zhang L, Li S, Liu L, Yang T, Yang P, et al. Predicting the reproductive toxicity of chemicals using ensemble learning methods and molecular fingerprints. Toxicol Lett. 2021;340:4–14.
23. Aljarf R, Tang S, Pires DEV, Ascher DB. embryoTox: using graph-based signatures to predict the teratogenicity of small molecules. J Chem Inf Model. 2023;63(2):432–41.
24. Jiang C, Yang H, Di P, Li W, Tang Y, Liu G. In silico prediction of chemical reproductive toxicity using machine learning. J Appl Toxicol. 2019;39(6):844–54.
25. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res. 2021;49(D1):D1388–95.
26. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2017;46(D1):D1074–82.
27. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, et al. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res. 2018;47(D1):D930–40.
28. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, et al. KEGG for linking genomes to life and the environment. Nucleic Acids Res. 2007;36(suppl_1):D480–4.
29. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, Wiegers J, Wiegers TC, et al. Comparative toxicogenomics database (CTD): update 2021. Nucleic Acids Res. 2020;49(D1):D1138–43.
30. Wishart D, Arndt D, Pon A, Sajed T, Guo AC, Djoumbou Y, et al. T3DB: the toxic exposome database. Nucleic Acids Res. 2015;43(Database issue):D928–34.
31. Wu L, Yan B, Han J, Li R, Xiao J, He S, et al. TOXRIC: a comprehensive database of toxicological data and benchmarks. Nucleic Acids Res. 2022;51(D1):D1432–45.
32. Brehm E, Flaws JA. Transgenerational effects of endocrine-disrupting chemicals on male and female reproduction. Endocrinology. 2019;160(6):1421–35.
33. Consortium TGO. Gene Ontology Consortium: going forward. Nucleic Acids Res. 2014;43(D1):D1049–56.
34. Consortium TU. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2020;49(D1):D480–9.
35. Krivák R, Hoksza D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. J Cheminform. 2018;10(1):39.
36. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics. 2013;29(21):2722–8.
37. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. Highly accurate protein structure prediction for the human proteome. Nature. 2021;596(7873):590–6.
38. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: an open chemical toolbox. J Cheminform. 2011;3(1):33.
39. Feinstein WP, Brylinski M. Calculating an optimal box size for ligand docking and virtual screening against experimental and predicted binding pockets. J Cheminform. 2015;7(1):18.
40. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31(2):455–61.
41. Rego N, Koes D. 3Dmol.js: molecular visualization with WebGL. Bioinformatics. 2014;31(8):1322–4.
42. Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem. 2011;32(7):1466–74.
43. Zeng X, Zhu S, Liu X, Zhou Y, Nussinov R, Cheng F. deepDR: a network-based deep learning approach to in silico drug repositioning. Bioinformatics. 2019;35(24):5191–8.
44. Deng Y, Xu X, Qiu Y, Xia J, Zhang W, Liu S. A multimodal deep learning framework for predicting drug–drug interaction events. Bioinformatics. 2020;36(15):4316–22.
45. Wang JZ, Du Z, Payattakool R, Yu PS, Chen C-F. A new method to measure the semantic similarity of GO terms. Bioinformatics. 2007;23(10):1274–81.
46. Wishart DS, Guo A, Oler E, Wang F, Anjum A, Peters H, et al. HMDB 5.0: the human metabolome database for 2022. Nucleic Acids Res. 2022;50(D1):D622–31.
47. Whincup PH, Kaye SJ, Owen CG, Huxley R, Cook DG, Anazawa S, et al. Birth weight and risk of type 2 diabetes: a systematic review. JAMA. 2008;300(24):2886–97.
48. Backman TW, Cao Y, Girke T. ChemMine tools: an online service for analyzing and clustering small molecules. Nucleic Acids Res. 2011;39(Web Server issue):W486–91.
49. Zhang Y, Zhong L, Xu B, Yang Y, Ban R, Zhu J, et al. SpermatogenesisOnline 1.0: a resource for spermatogenesis based on manual literature curation and genome-wide data mining. Nucleic Acids Res. 2013;41(Database issue):D1055–62.
50. Wu J, Li D, Liu X, Li Q, He X, Wei J, et al. IDDB: a comprehensive resource featuring genes, variants and characteristics associated with infertility. Nucleic Acids Res. 2021;49(D1):D1218–24.

Publisher's Note