scientific data

Check for updates

OPEN

DATA DESCRIPTOR

HCDT 2.0: A Highly Confident Drug-Target Database for Experimentally Validated Genes, RNAs, and Pathways

Xinying Liu[1,2], Dehua Feng[1,2], Jiaqi Chen[1], Tianyi Li[1], Xuefeng Wang[1], Ruijie Zhang[1], Jian Chen[1], Xingjun Cai[1], Huirui Han[1], Lei Yu[1], Xia Li[1], Bing Li[1]✉, Limei Wang[1]✉ & Jin Li[1]✉

Drug-target interactions constitute the fundamental basis for understanding drug action mechanisms and advancing therapeutic discovery. While existing drug-target databases have contributed valuable resources, they exhibit structural and functional fragmentation due to heterogeneous data sources and annotation standards. Building upon the high-confidence drug-gene interactions curated in HCDT 1.0, we present HCDT 2.0, a comprehensive and standardized resource that expands the scope through multiomics data integration. This update incorporates three-dimensional interactions including drug-gene, drug-RNA and drug-pathway interactions. The current version contains 1,284,353 curated interactions: 1,224,774 drug-gene pairs (678,564 drugs × 5,692 genes), 11,770 drug-RNA mappings (316 drugs × 6,430 RNAs), and 47,809 drug-pathway links (6,290 drugs × 3,143 pathways), alongside 16,317 drug-disease associations. To enhance biological interpretability, we further integrated pathway-gene and RNA-gene regulatory relationships. In addition, we integrated 38,653 negative DTIs covering 26,989 drugs and 1,575 genes. This integrative framework not only addresses critical gaps in cross-scale data representation but also establishes a robust foundation for systems pharmacology applications, including drug repurposing, adverse event prediction, and precision oncology strategies.

Background & Summary

The development of new drugs is a time-consuming and labor-intensive process[1], often hindered by the complexity of drug action mechanisms and the emergence of drug resistance[2]. The average cost of bringing a new drug to market is estimated to be around $2.6 billion, taking over a decade from discovery to launch[3]. Hence, it is urgent to find a new strategy to discover drugs[4]. Common drug interaction targets encompass genes, pathways, and RNA, with drugs capable of engaging in interactions with each of these components. Drug target research is crucial for drug development, helping us understand how drugs interact with specific targets for drug discovery and disease treatment[5]. At present, there are four mainstream methods for predicting drug-target interactions, including traditional neural network-based methods, graph neural network-based methods, knowledge graph embedding-based methods, and multimodal learning-based methods[6]. Scientists can use these methods to predict and validate more drug targets, but they depend on experimentally validated information about drug targets, encompassing key genes, pathways, and RNAs, among others. However, drug target research still faces some challenges. On the one hand, most drugs may only target a few targets, limiting the diversity of treatment options[7]. On the other hand, the complexity and diversity of targets also increase the difficulties and uncertainties in the drug development process[8]. Therefore, the integration of drug-target data is crucial for identifying potential drug targets and developing effective treatment strategies[9]. Despite the existence of various drug-focused databases such as ncDR[10], Lnc2Cancer[11] and SM2miR[12], a unified platform that integrates drug-gene, drug-pathway, and drug-RNA relationships remains a major gap in the current bioinformatics field.

[1]School of Biomedical Informatics and Engineering, Kidney disease research institute at the second affiliated hospital, Hainan Engineering Research Center for Health Big Data, Hainan Medical University, Haikou, Hainan, 571199, China. [2]These authors contributed equally: Xinying Liu, Dehua Feng. ✉e-mail: binglijpn2003@aliyun.com; wanglm@muhn.edu.cn; lijin@muhn.edu.cn

To fill this gap and provide a more holistic view of the complex drug-target interactions, we conducted HCDT 1.0 in 2022, a database focused on highly confident drug-gene relationships. In this study, we updated the drug-gene interaction in HCDT 2.0 and expanded the range of interactions about Drug-RNA and Drug-Pathway. HCDT 2.0 encompasses a wide spectrum of interactions, offering a rich resource for researchers in the field of bioinformatics.

Methods

Data collection. When constructing the HCDT 2.0 database, we adhered to a stringent methodology for data collection, curation, and integration to guarantee the precision and dependability of the dataset.

The HCDT2.0 database consists of three relationships, drug-gene, drug-RNA and drug-pathway. The data pertaining to drugs within three relational databases are uniform, including the name of the drug, multiple relationships corresponding to the same drug, simplified molecular input line entry system (SMILES), International Union of Pure and Applied Chemists (IUPAC) name, International Chemical Identifier (INCH), drug type, Molecular Formula and Molecular weight[4]. Among them, the most important is SMILE, because it is a unique identifier that distinguishes one DRUG from another[13].

Our HCDT 2.0 database contains 9 specialized databases for studying drug-gene interactions. In terms of genetic data, we ensured the inclusion of at least one of the following identifiers: gene symbol, Entrez ID, Ensemble ID, or UniProt ID, which can be mapped with the gene information in the HGNC database[4,14].

There are 6 databases dedicated to studying drug-RNA interactions. For RNA data, the dataset incorporates the RNA name, Ensemble ID, Transcript stable ID, Chromosome/scaffold name, GENCODE basic annotation, Phenotype description, Gene % GC content, Gene type, Transcript type, and Gene Synonym. The most important of these is the Ensemble ID, as this is usually a unique identifier that distinguishes an RNA from other RNAs.

There are 5 databases centered on drug-pathway interactions. Regarding pathway data, the information comprises the pathway name, REACTOME_ID, KEGG_HSAID, SMPDB_ID, ChEBI_ID, KEGG_ID, and GENEIDS. These data represent the ID of the pathway in the corresponding database, that is, the unique identifier that distinguishes the pathway from other pathways.

Data filtering. GENE data filtering: We still follow the HCDT 1.0 version, and the criteria are Ki, Kd, IC50, and EC50 with at least one $\leq$10 micromoles. Based on this, we have updated the database content.

RNA data filtering: Four databases were excluded in the following: lnc2cancer3.0[11] mainly focuses on the relationship between lncRNA and cancer. Although the database records drug information, it does not mention a direct relationship between the drug and the corresponding RNA; in LncMAP[15] and LNCmap[16], the majority of drug-target interactions are based on computational prediction, but not confirmed by biological experiments, which do not meet the highly confident purpose of this study; NoncoRNA[17] is a database of experimentally supported non-coding RNAs and drug targets in cancer. But there is a lot of predictive data in it, and we only filter the verified data among them. The remaining 6 databases were selected as the original data sources for drug-RNA relationships for the HCDT database. To ensure the high confidence of drug-target interactions, we used the following criteria:(i) the data are experimentally validated; (ii) the data must be of human origin. HCDT 2.0 consists of multiple databases integrated. We screened 9 drug-target interaction databases and excluded two drug-target prediction databases and one database with no direct drug-target relationship. The remaining 6 databases all met our criteria for high confidence. All drug-target interactions are validated by *in vivo* experiments and are guaranteed to be of human rather than other species origin.

Pathway data filtering: When filtering the drug pathway relational database, in order to ensure high confidence in drug-target interactions, we used the standard: (i) the data in the database must be able to find the corresponding pathway relationship with the drug; (ii) The data for these drug pathways have been experimentally validated rather than predicted. The reason why five databases were included in this study is that they can provide information on drug-corresponding signaling pathways, and the interaction data of these drug pathways have been validated. On the contrary, certain databases such as TTD (Therapeutic Target Database) are excluded. This is because the TTD database only infers the drug's action pathway based on the consistency between target genes and genes in specific pathways, and does not directly provide specific information on the drug's action pathway. Therefore, it does not meet the screening criteria of this study.

Drug-target classification. In HCDT 2.0, the data includes drug-genes, drug-RNAs and drug-pathways relationships. In this paragraph, we perform a categorical analysis of the data. As we all know, we have already had HCDT 1.0, and the data at that time only included drug-genes relationships.

The classification of genes is consistent with version 1.0. That means genes are classified into four groups according to function[4]: genes that encode proteins, genes that do not encode ribonucleic acid (RNA), pseudogenes that have no actual function and the remaining genes whose function is not yet clear.

As for RNAs, the classification is based on the RNA types provided in the source database. Currently, they are categorized into five distinct groups: miRNA (microRNA), lncRNA (long non-coding RNA), RNA (general RNA), circRNA (circular RNA), and piRNA (PIWI-interacting RNA). Each of these categories represents a different class of RNA with specific biological functions and roles in gene regulation, cellular processes, and disease mechanisms (Fig. 1).

In terms of Pathways, they are generally not categorized because they describe continuity and interconnectedness in biological processes rather than discrete entities. Therefore, we classify them based on the sources of their different databases (Fig. 2).

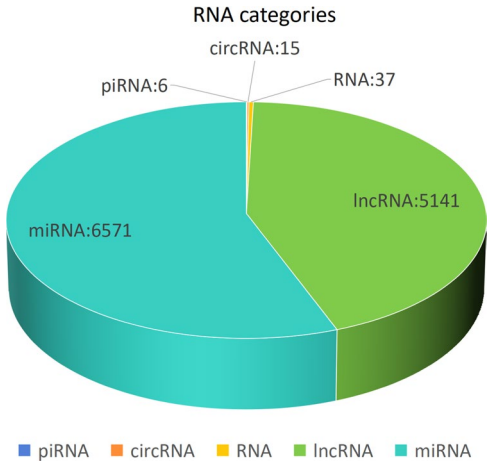

Fig. 1 RNA categories in HCDT 2.0.

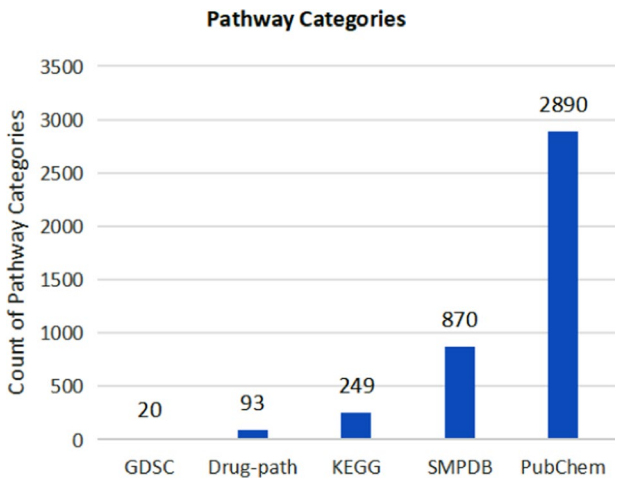

Fig. 2 Pathway categories in HCDT 2.0.

Drug-genes update. In the updated HCDT 2.0, the number of interaction relationships is 1,224,774 (Table 1), which has been expanded to a certain extent compared to the previous HCDT 1.0 version. This indicates that our HCDT 2.0 database is becoming a more comprehensive drug target interaction data resource. In HCDT 2.0, the newly added DSigDB[18] database is a new resource that relates drugs and their target genes. It contains 23,325 interaction data, supplementing the existing database content.

Compared with other commonly used databases such as BindingDB[19], GtoPdb[20], PharmGKB[21], and TTD[22], the unique contribution of DSigDB lies in its focus on drug signature information, which is of great significance for exploring drug reuse and its mechanism of action. The other databases have been updated to the latest version. BindingDB contains 353,167 interaction records, while GtoPdb and PharmGKB have 14,605 and 4,831 interaction records, respectively. TTD contains 530,553 interaction records.

Negative drug-target interactions. To comprehensively characterize drug-target relationships, we integrated negative Drug-Target Interactions (DTIs) in HCDT 2.0. The negative DTIs candidates were derived from BindingDB, ChEMBL, GtoPdb, PubChem, and TTD. Experimental binding affinity measurements (Ki/Kd/IC50/EC50/AC50/Potency $>100\,\mu M$)[23] were used to define these non-active interactions. We systematically integrated 38,653 negative DTIs across 26,989 drugs and 1,575 target genes (Table 2).

Drug-RNAs. In HCDT 2.0, we have added drug target information about drug-RNA. We collected drug-RNA information from six databases and found a total of 11,770 high-confidence interactions between 316 drugs and 6,430 RNAs (Table 3). Compared to single databases, HCDT 2.0 offers a significant expansion in interactions. Among these databases, DRmiRNA is the largest data provider, accounting for 37.03% of drugs, 11.84% of targets, and 46.21% of drug-target interactions.

We constructed a drug-RNA interaction network to reveal potential interactions between drugs and RNAs. A subnetwork for hub RNAs with a degree equal to or larger than 10 was illustrated in Fig. 3. It involves 20 hub RNAs and 56 drugs. For instance, miR-99b may be a target for ten drugs, which can be categorized into four

DATABASE	Number of Drugs	Number of Genes	Number of Interactions
BindingDB[19]	246573	1655	353167
ChEMBL[34]	424712	2465	628118
DGIdb[35]	10441	2863	39147
Drugbank[36]	5660	2900	530553
DSigDB[18]	3312	1016	23325
GtoPdb[20]	7476	1706	14605
PharmGKB[37]	1042	1384	4831
Pubchem[38]	4049	1920	25770
TTD[22]	345948	1358	530553
HCDT 2.0	678564	5692	1224774

Table 1. Statistics on the updated Drug-Gene data source in HCDT 2.0.

DATABASE	Number of Drugs	Number of Genes	Number of Interactions
BindingDB[19]	16710	991	19615
ChEMBL[34]	806	144	854
GtoPdb[20]	88	73	108
Pubchem[38]	2254	874	6962
TTD[22]	15715	829	21852
HCDT 2.0	26989	1575	38653

Table 2. Statistics on the Negative Drug-Target Interactions data source in HCDT 2.0.

Database	Number of drugs	Number of RNAs	Number of interactions
NoncoRNA[17]	120	631	2284
Dlnc[39]	42	4643	4884
DRlncRNA[10]	29	163	205
DRmiRNA[10]	117	761	5429
SM2miR[12]	147	798	2473
HARIBOSS[40]	16	37	37
HCDT 2.0	316	6430	11770

Table 3. Statistics on Drug-RNA in HCDT 2.0.

categories: monoclonal antibodies (Cetuximab for inhibiting cancer growth), corticosteroids (Dexamethasone for managing inflammation and immune response), hormone drugs (Tamoxifen for breast cancer by affecting hormones) and chemotherapy drugs (Carboplatin, Cisplatin, Doxorubicin, Mitomycin C, Vincristine, Gemcitabine for killing cancer cells in different ways). Our study on drug-RNA interactions shows that 6,822 interactions (57.97%) are sourced from a single database, 551 interactions (4.68% of the total) are derived from two databases, 68 interactions (0.58% of the total) are sourced from three databases. The abundant content within our HCDT 2.0 database was significantly underlined. The support of these interactions from multiple databases not only reinforces their credibility but also highlights their crucial role in drug-target research and the development of therapies.

Drug-pathways. In HCDT 2.0, we added new interactions between drugs and pathways. We collected drug-pathway information from 5 databases and obtained a total of 47,809 high-confidence interactions between 6,290 drugs and 3,143 pathways (Table 4). Among them, Pubchem is the largest data provider, accounting for 13.61% of drugs, 91.95% of targets, and 31.08% of drug-target interactions.

To construct comprehensive multi-layered drug-target interaction networks, we systematically integrated heterogeneous data from multiple repositories. Pathway-gene associations were derived through aggregation of KEGG[24], Reactome[25], and Signaling Pathway Database (SMPDB)[26] annotations, resulting in 2,639 curated records. For RNA-gene regulatory relationships, two complementary approaches were employed: (1) a cis-regulatory element-based analysis where RNA splicing sites (RANcentral[27]) and gene splicing sites (Ensembl[28]) were mapped, and functional links were established for RNA-gene pairs with cis-distances $\leq$ 10 kb (11,509 records); (2) a direct evidence integration strategy that compiled RNA-target gene interactions from miRNA-target (miRTarBase[29]), lncRNA-target (LncTarD[30], LncRNA2Target[31]) databases, yielding 110,294 high-confidence interactions after rigorous curation and duplicate removal. This dual-method framework ensures both spatial proximity-based and direct evidence-based coverage of transcriptional regulatory mechanisms.

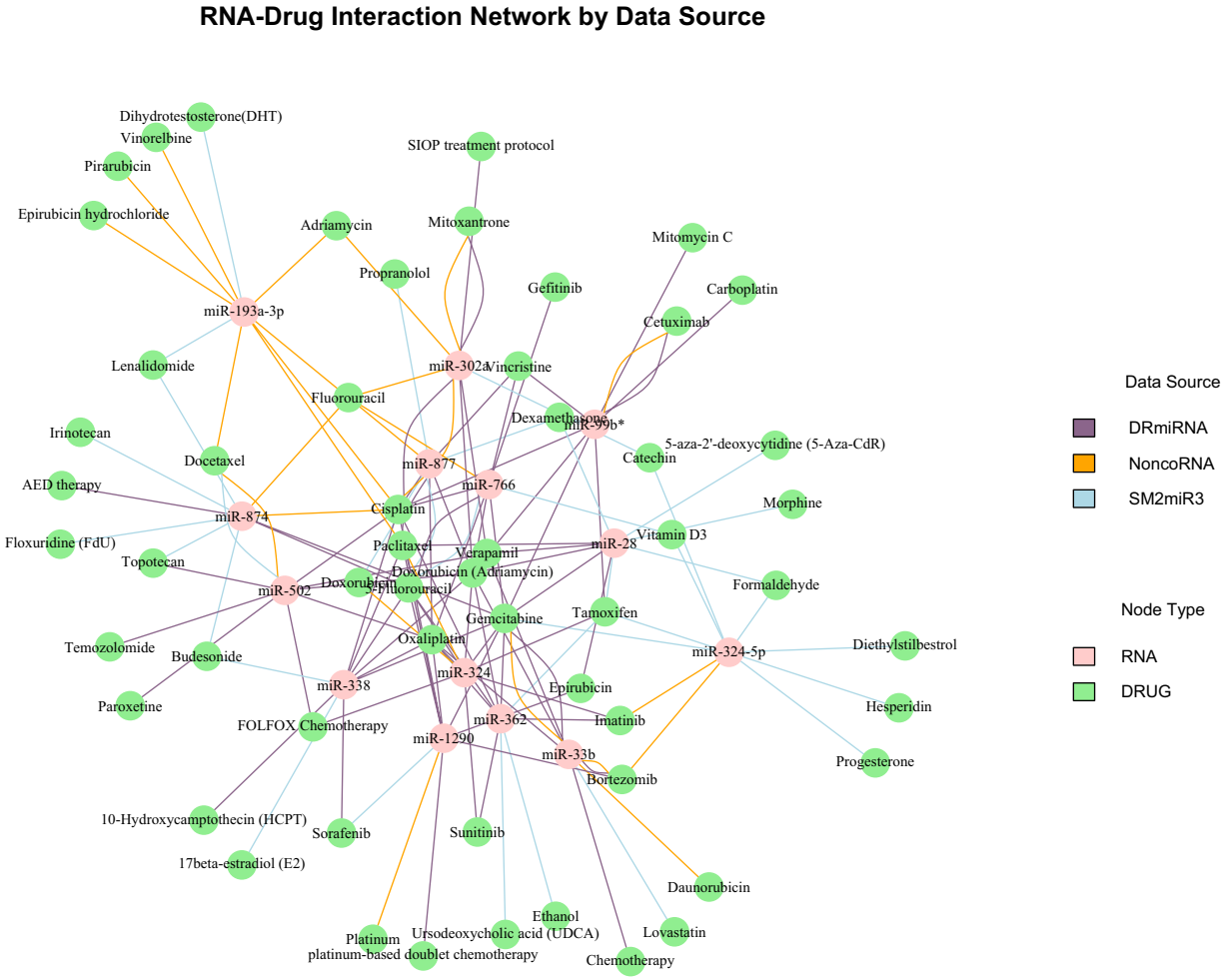

Fig. 3 Drug-RNA Interaction Subnetwork for hub RNAs.

Database	Number of drugs	Number of Pathways	Number of interactions
Drug-path[41]	1284	93	11769
GDSC[42]	433	20	438
KEGG[43]	4899	249	10495
PubChem[38,44]	856	2890	14857
SMPDB[45]	1490	870	10250
HCDT 2.0	6290	3143	47809

Table 4. Statistics on Drug-Pathway in HCDT 2.0.

Database	Number of drugs	Number of Diseases	Number of interactions
TTD[22]	6449	1089	7823
CTD[32]	1425	380	6471
KEGG[24]	984	397	2023
HCDT 2.0	7728	1473	16317

Table 5. Statistics on Drug-Disease in HCDT 2.0.

Drug-diseases. In HCDT 2.0, we systematically integrated drug-disease associations from three complementary databases: the Comparative Toxicogenomics Database (CTD)[32], KEGG[24], and TTD[22]. This integration resulted in 16,317 curated records, encompassing 7,728 unique drugs and 1,473 distinct diseases (Table 5). The inclusion of these multi-source interactions not only strengthens the database's utility for drug repositioning and

precision medicine applications but also enables holistic analysis of molecular connectivity across drugs, genes, RNAs, pathways, and diseases through unified multi-omics data.

Data Records

The dataset described in HCDT 2.0 is publicly available via the following: https://doi.org/10.6084/m9.figshare.28098734[33].

All data were standardized: drugs were uniformly annotated with PubChem CID and name, genes with HGNC symbols, while RNAs and pathways retained their original database-specific identifiers and nomenclature to ensure cross-source consistency. The structure comprises five tables: (1) Drug-Genes (featuring DRUG_NAME, PUBCHEM_CID, GENE_SYMBOL, HGNC_ID) for validated molecular targets; (2) Drug-RNAs with RNA identifiers (DRUG_NAME, PUBCHEM_CID, RNA_NAME, RNA_ID); (3) Drug-Pathways (PATHWAY_NAME, REACTOME_ID, KEGG_ID); (4) Drug-Disease (Disease_Name, ICD-11, MESH, OMIM); and (5) Negative DTIs providing experimentally confirmed non-interacting pairs. All tables share PUBCHEM_CID as a universal drug identifier and include standardized annotation schemas (Supplementary Table S1), enabling systematic integration of multi-omics data and supporting applications ranging from drug repurposing to explainable target discovery.

Technical Validation

HCDT 2.0 ensures the accuracy and reliability of its data through several validation steps:

Experimental Validation: All interactions, whether drug-gene, drug-RNA, or drug-pathway, are validated through *in vivo* or experimental data. No predictions or computational models are used in the final dataset.

Data Consistency: To ensure the consistency and integrity of identifiers across datasets, all drug, gene, RNA, and pathway names have been standardized with widely accepted identifiers (e.g., PubChem CID, Ensemble ID, HGNC ID).

Cross-database Validation: The interactions in HCDT 2.0 are sourced from multiple databases, providing additional validation and reinforcing the credibility of the data. Cross-referencing between databases allows for the identification of interactions that are supported by multiple sources, enhancing the trustworthiness of the database.

Usage Notes

HCDT 2.0 database is accessible online at http://hainmu-biobigdata.com/hcdt2/index.php.

Code availability

No custom code was used in the curation or validation of this dataset.

Received: 30 December 2024; Accepted: 9 April 2025;
Published online: 25 April 2025

References

1. Kang, H. *et al*. Drug-disease association prediction with literature based multi-feature fusion. *Front Pharmacol* **14**, 1205144 (2023).
2. Laufer, M. K. Monitoring antimalarial drug efficacy: current challenges. *Curr Infect Dis Rep* **11**, 59–65 (2009).
3. Yella, J.K., Yaddanapudi, S., Wang, Y. & Jegga, A.G. Changing Trends in Computational Drug Repositioning. *Pharmaceuticals (Basel)* **11** (2018).
4. Chen, J. *et al*. HCDT: an integrated highly confident drug-target resource. *Database (Oxford)* **2022** (2022).
5. Singh, R. S., Angra, V., Singh, A., Masih, G. D. & Medhi, B. Integrative omics - An arsenal for drug discovery. *Indian J Pharmacol* **54**, 1–6 (2022).
6. Li, X., Xiong, Z., Zhang, W. & Liu, S. Deep learning for drug-drug interaction prediction: A comprehensive review. *Quant. Biol.* **12**, 30–52 (2024).
7. Heaney, L. G. *et al*. Research in progress: Medical Research Council United Kingdom Refractory Asthma Stratification Programme (RASP-UK). *Thorax* **71**, 187–189 (2016).
8. Preskorn, S. H. CNS drug development: part III: future directions. *J Psychiatr Pract* **17**, 49–52 (2011).
9. Xia, F. *et al*. A Novel Computational Framework for Precision Diagnosis and Subtype Discovery of Plant With Lesion. *Front Plant Sci* **12**, 789630 (2021).
10. Dai, E. *et al*. ncDR: a comprehensive resource of non-coding RNAs involved in drug resistance. *Bioinformatics* **33**, 4010–4011 (2017).
11. Gao, Y. *et al*. Lnc2Cancer 3.0: an updated resource for experimentally supported lncRNA/circRNA cancer associations and web tools based on RNA-seq and scRNA-seq data. *Nucleic Acids Res* **49**, D1251–d1258 (2021).
12. Liu, X. *et al*. SM2miR: a database of the experimentally validated small molecules' effects on microRNA expression. *Bioinformatics* **29**, 409–411 (2013).
13. Kpanou, R., Osseni, M. A., Tossou, P., Laviolette, F. & Corbeil, J. On the robustness of generalization of drug-drug interaction models. *BMC Bioinformatics* **22**, 477 (2021).
14. Seal, R. L. *et al*. Genenames.org: the HGNC resources in 2023. *Nucleic Acids Res* **51**, D1003–d1009 (2023).
15. Li, Y. *et al*. LncMAP: Pan-cancer atlas of long noncoding RNA-mediated transcriptional network perturbations. *Nucleic Acids Res* **46**, 1113–1123 (2018).
16. Yang, H. *et al*. The LncRNA Connectivity Map: Using LncRNA Signatures to Connect Small Molecules, LncRNAs, and Diseases. *Sci Rep* **7**, 6655 (2017).
17. Li, L. *et al*. NoncoRNA: a database of experimentally supported non-coding RNAs and drug targets in cancer. *J Hematol Oncol* **13**, 15 (2020).
18. Yoo, M. *et al*. DSigDB: drug signatures database for gene set analysis. *Bioinformatics* **31**, 3069–3071 (2015).
19. Liu, T. *et al*. BindingDB in 2024: a FAIR knowledgebase of protein-small molecule binding data. *Nucleic Acids Res* (2024).
20. Harding, S. D. *et al*. The IUPHAR/BPS Guide to PHARMACOLOGY in 2024. *Nucleic Acids Res* **52**, D1438–d1449 (2024).
21. Barbarino, J. M., Whirl-Carrillo, M., Altman, R. B. & Klein, T. E. PharmGKB: A worldwide resource for pharmacogenomic information. *Wiley Interdiscip Rev Syst Biol Med* **10**, e1417 (2018).
22. Zhou, Y. *et al*. TTD: Therapeutic Target Database describing target druggability information. *Nucleic Acids Res* **52**, D1465–d1477 (2024).

23. Tomašič, T. *et al*. Selective DNA Gyrase Inhibitors: Multi-Target in Silico Profiling with 3D-Pharmacophores. *Pharmaceuticals (Basel)* **14** (2021).
24. Kanehisa, M., Furumichi, M., Sato, Y., Matsuura, Y. & Ishiguro-Watanabe, M. KEGG: biological systems database as a model of the real world. *Nucleic Acids Res* **53**, D672–d677 (2025).
25. Milacic, M. *et al*. The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Res* **52**, D672–d678 (2024).
26. Wishart, D. S. *et al*. PathBank 2.0-the pathway database for model organism metabolomics. *Nucleic Acids Res* **52**, D654–d662 (2024).
27. RNAcentral 2021. secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Res* **49**, D212–d220 (2021).
28. Harrison, P. W. *et al*. Ensembl 2024. *Nucleic Acids Res* **52**, D891–d899 (2024).
29. Cui, S. *et al*. miRTarBase 2025: updates to the collection of experimentally validated microRNA-target interactions. *Nucleic Acids Res* **53**, D147–d156 (2025).
30. Zhao, H. *et al*. LncTarD 2.0: an updated comprehensive database for experimentally-supported functional lncRNA-target regulations in human diseases. *Nucleic Acids Res* **51**, D199–d207 (2023).
31. Cheng, L. *et al*. LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. *Nucleic Acids Res* **47**, D140–d144 (2019).
32. Davis, A. P. *et al*. Comparative Toxicogenomics Database's 20th anniversary: update 2025. *Nucleic Acids Res* **53**, D1328–d1334 (2025).
33. Liu, X-Y *Figshare* https://doi.org/10.6084/m9.figshare.28098734.
34. Zdrazil, B. *et al*. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res* **52**, D1180–d1192 (2024).
35. Cannon, M. *et al*. DGIdb 5.0: rebuilding the drug-gene interaction database for precision medicine and drug discovery platforms. *Nucleic Acids Res* **52**, D1227–d1235 (2024).
36. Knox, C. *et al*. DrugBank 6.0: the DrugBank Knowledgebase for 2024. *Nucleic Acids Res* **52**, D1265–d1275 (2024).
37. Corpas, M. *et al*. Addressing Ancestry and Sex Bias in Pharmacogenomics. *Annu Rev Pharmacol Toxicol* **64**, 53–64 (2024).
38. Kim, S. *et al*. PubChem 2025 update. *Nucleic Acids Res* (2024).
39. Jiang, W. *et al*. D-lnc: a comprehensive database and analytical platform to dissect the modification of drugs on lncRNA expression. *RNA Biol* **16**, 1586–1591 (2019).
40. Panei, F. P., Torchet, R., Ménager, H., Gkeka, P. & Bonomi, M. HARIBOSS: a curated database of RNA-small molecules structures to aid rational drug design. *Bioinformatics* **38**, 4185–4193 (2022).
41. Zeng, H., Qiu, C. & Cui, Q. Drug-Path: a database for drug-induced pathways. *Database (Oxford)* **2015**, bav061 (2015).
42. Yang, W. *et al*. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res* **41**, D955–961 (2013).
43. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27–30 (2000).
44. Kim, S. *et al*. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res* **49**, D1388–d1395 (2021).
45. Jewison, T. *et al*. SMPDB 2.0: big improvements to the Small Molecule Pathway Database. *Nucleic Acids Res* **42**, D478–484 (2014).

Acknowledgements

This work was supported by the Natural Science Foundation of Hainan Province [Nos. 824RC514, 821QN0894, 621MS041]; National Natural Science Foundation of China [No.32260155].

Author contributions

Data collection: Xinying Liu and Dehua Feng. Data processing and analysis: Xinying Liu, Dehua Feng, Jiaqi Chen, Tianyi Li, Jian Chen. Technical validation support: Ruijie Zhang, Xingjun Cai, Huirui Han, Lei Yu, Xia Li, Bing Li, Limei Wang, and Jin Li. Writing and reviewing the manuscript: All authors contributed.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-04981-2.

Correspondence and requests for materials should be addressed to B.L., L.W. or J.L.