BioMedGraphica: An All-in-One Platform for Biomedical Prior Knowledge and Omic Signaling Graph Generation

Heming Zhang[1*], Shunning Liang[1*], Tim Xu[1,2*], Wenyu Li[2], Di Huang[1,2], Yuhan Dong[1], Guangfu Li[3], J. Philip Miller[1], S. Peter Goedegebuure[4,5], Marco Sardiello[10], Jonathan Cooper[10], William Buchser[9], Patricia Dickson[10], Ryan C. Fields[4,5], Carlos Cruchaga[6,7], Yixin Chen[2], Michael Province[8,9], Philip Payne[1], Fuhai Li[1,10#]

[1]Institute for Informatics, Data Science and Biostatistics (I2DB), [2]Department of Computer Science and Engineering, [4]Department of Surgery, [5]Siteman Cancer Center, [6]Department of Psychiatry, [7]NeuroGenomics and Informatics, [8]Division of Statistical Genomics, [9]Department of Genetics, [10]Department of Pediatrics, Washington University School of Medicine, Washington University in St. Louis, St. Louis, MO, USA. [3]Department of Surgery, School of Medicine, University of Connecticut, CT, 06032, USA. *Co-first authors; [#]Correspondence: Fuhai.Li@wustl.edu

Abstract

Artificial intelligence (AI) is revolutionizing scientific discovery because of its super capability, following the neural scaling laws, to integrate and analyze large-scale datasets to mine knowledge. Foundation models, large language models (LLMs) and large vision models (LVMs), are among the most important foundations paving the way for general AI by pre-training on massive domain-specific datasets. Different from the well annotated, formatted and integrated large textual and image datasets for LLMs and LVMs, biomedical knowledge and datasets are fragmented with data scattered across publications and inconsistent databases that often use diverse nomenclature systems in the field of AI for Precision Health and Medicine (AI4PHM). These discrepancies, spanning different levels of biomedical organization from genes to clinical traits, present major challenges for data integration and alignment. To facilitate foundation AI model development and applications in AI4PHM, herein, we developed *BioMedGraphica*, an all-in-one platform and unified text-attributed knowledge graph (TAKG), consists of 3,131,788 entities and 56,817,063 relations, which are obtained from 11 distinct entity types and harmonizes 29 relations/edge types using data from 43 biomedical databases. All entities and relations are labeled a unique ID and associated with textual descriptions (textual features). Since covers most of research entities in AI4PHM, BioMedGraphica supports the zero-shot or few-shot knowledge discoveries via new relation prediction on the graph. Via a graphical user interface (GUI), researchers can access the knowledge graph with prior knowledge of target functional annotations, drugs, phenotypes and diseases (drug-protein-disease-phenotype), in the graph AI ready format. It also supports the generation of knowledge-multi-omic signaling graphs to facilitate the development and applications of novel AI models, like LLMs, graph AI, for AI4PHM science discovery, like discovering novel disease pathogenesis, signaling pathways, therapeutic targets, drugs and synergistic cocktails.

Keywords: biomedical knowledge graph; precision medicine; graph AI models, knowledge graph integration and generation

1 Background and Summary

In recent years, the exponential growth of biomedical data has created unprecedented opportunities to advance research, improve clinical decision-making, and accelerate drug discovery. However, the landscape of biomedical knowledge remains highly fragmented, with essential information dispersed across a multitude of publications, databases, and proprietary datasets. This fragmentation presents significant challenges as different sources often employ inconsistent nomenclature and terminologies, hindering effective data integration[1]. The vast scope of biomedical data—from genes and proteins to clinical phenotypes and diseases—complicates the development of unified solutions, particularly in terms of entity matching and data harmonization. As a result, several knowledge graph systems have been developed to integrate this extensive array of data resources, aiming to merge information across the spectrum of biomedical domains[2–7]. However, existing knowledge graphs struggle to reconcile entities across these diverse, heterogeneous datasets, which are often noisy, inconsistent, and formatted in various ways. Without harmonizing nomenclature systems over multiple resources in biomedical domains, most current systems lack comprehensive coverage of biomedical resources, which leads to failure on implementing efficient matching algorithms, and rely heavily on manual work, making them less efficient for large-scale data integration. As a result, converting unstructured data into formats suitable for graph-based artificial intelligence (AI) models becomes an onerous task, and existing tools (e.g., mosGraphGen[8], IntergAO[9]) require extensive manual curation, limiting their scalability.

To address these challenges, ***BioMedGraphica*** was developed as an advanced platform that transforms the integration and utilization of biomedical data. By integrating data from **43** high-quality biomedical databases, we unify **11** key biomedical entity types—ranging from promoters, genes, transcripts, proteins, signaling pathways, metabolites and microbiotas to clinical phenotypes, diseases, and drugs—and **29** relations / edge types into a cohesive knowledge graph, resulting in **3,131,788** entities and **56,817,063** relations. With harmonizing over multiple knowledge bases, this study provides one of the most comprehensive biomedical knowledge graphs available today, enabling large-scale exploration of biological and clinical relations. Another core innovation is the use of language models, such as BioBERT[10], to generate high-quality embeddings that facilitate soft matching of phenotype, disease and drug entities across datasets. This approach enhances entity recognition by ranking potential matches based on similarity scores, allowing for more accurate and flexible integration compared to traditional rule-based methods. This machine learning-driven algorithm is particularly effective in addressing the variability and subtle differences inherent in biomedical data, providing greater precision in data harmonization.

The platform also distinguishes itself by offering a user-friendly, Windows-based client software with an intuitive graphical user interface (GUI). This interface allows researchers, clinicians, and data scientists to input heterogeneous biomedical datasets and receive integrated, structured outputs that are ready for graph-based AI applications. By lowering the barrier to entry, it fosters the widespread adoption of

knowledge graph technologies in both research and clinical settings, significantly accelerating translational research. Its ability to generate AI-ready datasets from complex inputs facilitates discoveries in areas such as biomarker identification, drug target exploration, disease etiology, and the development of personalized therapeutic strategies. Moreover, the platform is designed with scalability in mind. Its low coupling in entity space and streamlined software pipeline allow for continuous updates and integration of new data sources, ensuring that it remains up-to-date and highly relevant as biomedical knowledge expands. By addressing critical issues such as data fragmentation, inconsistent terminologies, and entity recognition, it provides a powerful tool for exploring complex biological systems and deriving new insights into disease mechanisms, treatment responses, and personalized medicine. Through its innovative application of machine learning and natural language processing (NLP) techniques, the platform not only pushes the boundaries of precision healthcare but also contributes to the next generation of biomedical research. By tackling key obstacles to data harmonization and entity recognition, this research empowers researchers and clinicians to better understand complex biological systems and accelerates the development of new therapeutic strategies. With its comprehensive and scalable architecture, ***BioMedGraphica*** stands at the forefront of efforts to integrate and utilize the vast and growing body of biomedical data.

2 Methods

2.1 Overview of Data Resources Used in *BioMedGraphica*

2.1.1 Entity Databases Introduction and Collection

A wide range of reputable biomedical databases were utilized to gather and integrate various types of data related to genes, transcripts, proteins, and other biomedical entities. This comprehensive integration ensured data consistency and accuracy, creating a unified framework essential for research. As shown in **Table 1**, the total number of entries in the original data file from each respective database were listed. For the ChEBI (Chemical Entities of Biological Interest) database, we utilized two primary datasets: one provided the mapping between ChEBI IDs and their corresponding InChI, while the other contained the mapping between ChEBI IDs and another database. For UNII (Unique Ingredient Identifier) database, our source data was obtained from two origins: one from PubChem, and the other provided by the FDA. For SILVA, we selected the LSU and SSU datasets. Similarly, for GTDB (Genome Taxonomy Database), we selected the data for both Archaea and Bacteria. The total number of entries after merging is indicated, with the individual row counts for each dataset provided in parentheses. Below is an expanded description of the databases used and the extracted data (see data overall details in **Table 1** and details of data resources in supplementary section A).

Table 1. Overview of Entity Databases

Database Names	Full Names	Entity Types	Total Number of Rows
Ensembl[11]	Ensembl	Gene	70,614
		Transcript	318,178
		Protein	150,346

OMIM[12]	Online Mendelian Inheritance in Man	Gene	28,833
HGNC[13]	HUGO Gene Nomenclature Committee	Gene	43,861
NCBI[14]	National Center for Biotechnology Information	Gene	68,932
		Microbiota	2,614,590
RefSeq[15]	NCBI - Reference Sequence Database	Gene	844,996
		Transcript	19,352
		Protein	320,530
RNACentral[16]	RNACentral	Transcript	66,789
UniProt[17]	Universal Protein Resource	Protein	20,428
Reactome[18]	Reactome	Pathway	2,711
KEGG[19]	Kyoto Encyclopedia of Genes and Genomes	Pathway	359
WikiPathways[20]	WikiPathways	Pathway	1,510
Pathway Ontology[21]	Pathway Ontology	Pathway	2,677
ComPath[22]	Comparative Pathology Platform of the University of Bern	Pathway	1,592
HMDB[23]	Human Metabolome Database	Metabolite	217,920
ChEBI[24]	Chemical Entities of Biological Interest	Metabolite	5,750
			393,172
		Drug	44,924
			225,102
SILVA[25]	SILVA	Microbiota	2,214,227(227318; 2,224,690)
Greengenes[26]	Greengenes	Microbiota	1,144,866
RDP[27]	Ribosomal Database Project	Microbiota	10,302
GTDB[28]	Genome Taxonomy Database	Microbiota	596,859(12,477; 584,382)
CTD[29]	The Comparative Toxicogenomics Database	Exposure	179,179
ToxCast[30]	Toxicity Forecasting	Exposure	9,403
ChemIDplus[31]	Chemical Identification Plus Database	Exposure	409,325
HPO[32]	Human Phenotype Ontology	Phenotype	22,789
ICD10 / ICD11[33,34]	International Classification of Diseases	Disease	12,597 / 36,044
DO[35]	Disease Ontology	Disease	38,037
MeSH[36]	Medical Subject Headings	Disease	5,032
UMLS[37]	Unified Medical Language System	Disease	14,036,386
SNOMED-CT[38]	Systematized Nomenclature of Medicine Clinical Terms	Disease	204,464
Mondo[39]	Mondo	Disease	134,406
PubChem[40]	Public Chemical Databases	Drug	123,155
CAS[41]	Chemical Abstracts Service	Drug	447,699
NDC[42]	National Drug Code	Drug	113,850
UNII[43]	Unique Ingredient Identifier	Drug	154,561
			152,870
DrugBank[44]	DrugBank	Drug	16,581

Not only the total number of entries after merging the two files were presented, but also the total number of rows were indicated for each dataset in parentheses.

2.1.2 Relation Database Introduction and Collection

This study integrates not only entity datasets but also a comprehensive range of relational datasets, facilitating the exploration of various biological and chemical interactions. These relational datasets capture complex relations between genes, transcripts, proteins, drugs, diseases, phenotypes, pathways, metabolites, and microbiotas, supporting advanced analyses in precision health (check data overall details in **Table 2** and **Table S1** for data collection details in supplementary section).

Table 2. General Information about Relation Databases

Database Names	Full Names	From	To	Edge Types	Number of Rows
Ensembl[11]	Ensembl	Gene	Transcript	Gene-Transcript	278,220
		Transcript	Protein	Transcript-Protein	123,540
RefSeq[15]	Reference Sequence Database	Gene	Transcript	Gene-Transcript	33,401
		Transcript	Protein	Transcript-Protein	30,193
UniProt[17]	Universal Protein Database	Transcript	Protein	Transcript-Protein	81,757
		Protein	Disease	Protein-Disease	204,411
BioGrid[45,46]	Biological General Repository for Interaction Datasets	Protein	Protein	Protein-Protein	925,035
STRING[47,48]	Search Tool for the Retrieval of Interacting Genes/Proteins	Protein	Protein	Protein-Protein	13,715,404
KEGG[19]	Kyoto Encyclopedia of Genes and Genomes	Protein	Protein	Protein-Protein	52,155
		Protein	Pathway	Protein-Pathway	21,051
		Drug	Pathway	Drug-Pathway	3,922
		Pathway	Protein	Pathway- Protein	24,475
		Pathway	Drug	Pathway-Drug	2,334
HPO[32]	Human Phenotype Ontology	Protein	Phenotype	Gene-Phenotype	312,812
		Protein	Disease	Gene-Disease	154,16
		Phenotype	Phenotype	Phenotype-Phenotype	19,434
		Phenotype	Disease	Phenotype-Disease	154,431
		Disease	Phenotype	Disease-Phenotype	154,431
DisGeNet[49]	DisGeNet	Protein	Disease	Protein-Disease	91,484
DISEASES[50]	DISEASES	Protein	Disease	Protein-Disease	346,173
HMDB[23]	Human Metabolome Database	Metabolite	Protein	Metabolite-Protein	863,759
		Metabolite	Disease	Metabolite-Disease	24,755
		Drug	Metabolite	Drug-Metabolite	3,258
MetaNetX[51]	MetaNetX	Metabolite	Metabolite	Metabolite-Metabolite	11,723
DisBiome[52]	DisBiome	Microbiota	Disease	Microbiota-Disease	616,390
MDAD[53]	Microbe-Drug Association Database	Microbiota	Drug	Microbiota-Drug	5,055
		Drug	Microbiota	Drug-Microbiota	5,055
PharmacoMicrobiomics[54]	PharmacoMicrobiomics	Microbiota	Drug	Microbiota-Drug	69
		Drug	Microbiota	Drug-Microbiota	69
CTD[29]	The Comparative Toxicogenomics Database	Exposure	Gene	Exposure-Gene	766,820
		Exposure	Pathway	Exposure-Pathway	1,591,611
		Exposure	Disease	Exposure-Disease	9,114,984
DO[35]	Disease Ontology	Disease	Disease	Disease-Disease	14,172
DrugBank[44]	DrugBank	Drug	Protein	Drug-Protein	5,350
		Drug	Drug	Drug-Drug	2,839,610
BindingDB[55]	Binding Database	Drug	Protein	Drug-Protein	1,541,006
DrugCentral[56]	DrugCentral	Drug	Protein	Drug-Protein	14,301
		Drug	Disease	Drug-Disease	42,307
SIDER[57]	Side Effect Resource	Drug	Phenotype	Drug-Phenotype	309,849

The last column represents the total number of rows in the original dataset from each database.

2.2 Harmonizing Resources

As shown in **Figure 1**, the integrated biomedical knowledge graph system, ***BioMedGraphica***, has been proposed. With collected datasets from various sources, the knowledge graph will integrate **11** different types of entities from **33** databases into a universal knowledge database. And The promoter entities were directly derived from the gene entities because ***BioMedGraphica***, by default, assumes that promoters influence genes. In addition, the relationship between those entities were included by harmonizing the **20** relational databases with **29** edge types. The details of merging and harmonizing can be found in the following descriptions.

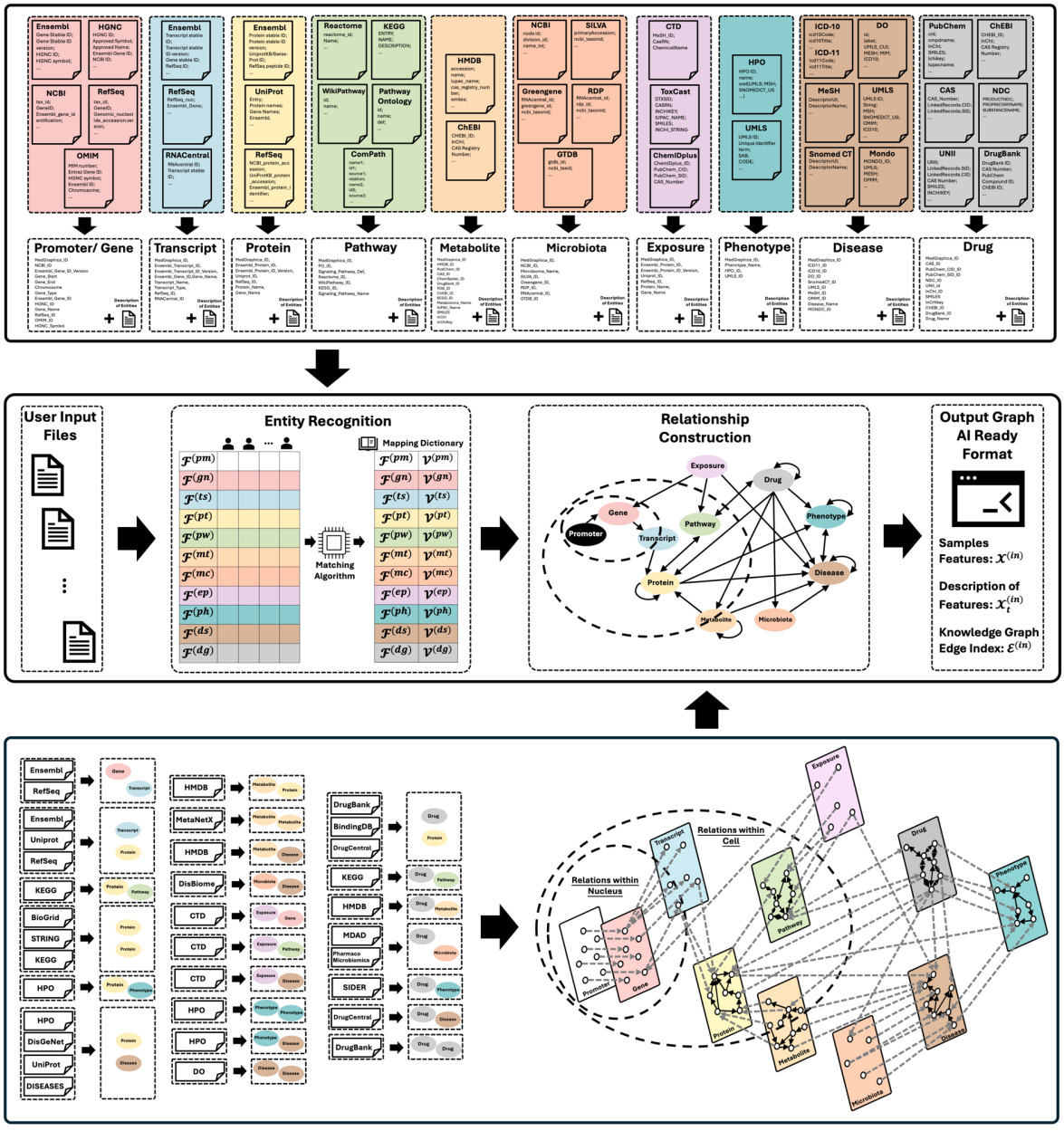


Figure 1. Overview of ***BioMedGraphica.*** Upper panel shows integration of the entities from various databases. Lower panel demonstrate the relation harmonization process and constructed knowledge graph.

Mid panel display the general procedures of BioMedGraphica, with entity recognition and relation construction based on the user input files, outputting the graph AI ready format files.

2.2.1 Entity Integration

Gene Entity Merging The Ensembl database was utilized as the primary basis for data integration. Initially, Ensembl, HGNC, and NCBI were merged based on matching Ensembl IDs. Subsequently, data from RefSeq and OMIM were incorporated, with NCBI IDs serving as the common identifier. The NCBI ID was chosen as the minimal unit for unifying the data, and a final integration of gene information was conducted according to NCBI IDs (refer to **Table 3** and **Figure S1** in the supplementary section for details). The columns highlighted in bold within the table denote those used for merging across databases, with the IDs in these columns being unique. Additionally, a textual description of each gene entity was appended. Using the free Perl script geneDocSum.pl provided by NCBI (download link: https://ftp.ncbi.nih.gov/gene/tools/geneDocSum.pl), all human records marked as current (alive) and containing summaries were retrieved. By mapping NCBI Gene IDs to corresponding entries in BioMedGraphica_Gene, the descriptions associated with BioMedGraphica_Gene IDs (BMG_GN) were obtained.

Table 3. Gene Entity Information

Database Names		Raw Data		After Data Cleaning		Total Number of BioMedGraphica ID
		Total Number of Rows	Unique	Total Number of Rows	Unique	
Ensembl	**Ensembl Gene ID**	70,614	70,611	70,611	70,611	215,608
	Ensembl Gene ID version	70,614	70,611	70,611	70,611	
	HGNC ID	46,525	40,982	46,522	40,982*	
	Total Number of Rows	70,614		70,611		
HGNC	HGNC ID	43,861	43,861	43,860	43,861*	
	Ensembl Gene ID	41,222	41,221	41,221	41,221	
	NCBI ID	43,807	43,807	43,806	43,807*	
	Total Number of Rows	43,861		43,860		
NCBI	**NCBI ID**	36,998	36,967	36,867	36,967*	
	Ensembl Gene ID	36,998	36,867	36,867	36,867	
	Total Number of Rows	36,998		36,867		
RefSeq	RefSeq ID	844,996	279,097	160,882	134,965*	
	NCBI ID	844,996	191,101	160,882	160,887	
	Total Number of Rows	844,996		160,882		
OMIM	OMIM ID	28,833	28,833	17,193	17,205*	
	NCBI ID	18,315	18,303	17,193	17,193	
	Total Number of Rows	28,833		17,193		

*this column contains multiple IDs in one row

Transcript Entity Merging The integration of the three databases utilized the Ensembl ID as the standard reference. For transcript entities, the Ensembl Transcript Stable ID was adopted as the smallest unit of data granularity. The integration process is illustrated in **Figure S2** of the supplementary section, and the merged results are detailed in **Table 4**. Bolded entries in the table identify the columns used for database merging, where the IDs in these columns are unique. Transcript descriptions were extracted from the Ensembl database using the BioMart API. By mapping the Transcript Stable ID to corresponding transcripts in BioMedGraphica, transcript descriptions were successfully assigned to the majority of BioMedGraphica transcripts.

Table 4. Transcript Entity Information

Database		Raw Data		After Data Cleaning		Total Number of BioMedGraphica ID
		Total Number of Rows	Unique	Total Number of Rows	Unique	
Ensembl	Ensembl Transcript ID	318,178	278,220	278,220	278,220	278,326
	Ensembl Transcript ID version	318,178	278,220	278,220	278,220	
	Ensembl Gene ID	318,178	70,611	278,220	70,611	
	RefSeq mRNA ID	85,578	66,801	47,717	66,801*	
	RefSeq ncRNA ID	34,813	19,091	16,758	19,091*	
	RefSeq MANE Select ID	38,479	19,288	19,288	19,288	
	Total Number of Rows	318,178		278,220		
RefSeq	**Ensembl Transcript ID version**	19,352	19,352	19,352	19,352	
	RefSeq ID	19,352	19,352	19,352	19,352	
	Total Number of Rows	19,352		19,352		
RNAcentral	RNAcentral ID	66,789	62,925	66,789	62,925	
	Ensembl Transcript ID	66,789	66,789	66,789	66,789	
	Total Number of Rows	66,789		66,789		

*this column contains multiple IDs in one row

Protein Entity Merging The integration process began by merging data from Ensembl and UniProt based on the Protein Stable ID Version. Subsequently, RefSeq data was incorporated by leveraging mapping relationships between RefSeq and the two databases. The Ensembl Protein ID Version was established as the minimal unit of data granularity for protein entities (refer to **Figure S3** in the supplementary section for the merging workflow and **Table 5** for detailed results). Bolded entries in the table highlight the columns used for cross-database merging, where the IDs are uniquely assigned. Protein descriptions were retrieved from the UniProt database using the UniProt API. By mapping UniProt IDs to corresponding proteins in BioMedGraphica, descriptive information was successfully provided for BioMedGraphica proteins.

Table 5. Protein Entity Information

Database			Raw Data		After Data Cleaning		Total Number of BioMedGraphica ID
			Total Number of Rows	Unique	Total Number of Rows	Unique	
Ensembl		Ensembl Protein ID	150,346	123,495	123,495	123,495	204,835
		Ensembl Protein ID version	150,346	123,495	123,495	123,495	
		UniProt ID	70,537	19,336	49,656	19,336*	
		RefSeq ID	74,348	66,631	47,503	66,631*	
		Total Number of Rows	150,346		123,495		
UniProt		UniProt ID	20,428	20,428	51714	20428	
		Ensembl Protein ID version	19,320	19,320*	50,606	50,606	
		Total Number of Rows	20,428		51,714		
RefSeq	RefSeq-Uniprot	RefSeq ID	320,530	105,808	99,482	81,855*	
		UniProt ID	320,530	116,043	99,482	99,482	
		Total Number of Rows	320,530		99,482		
	RefSeq-Ensembl	RefSeq ID	47,223	47,183	47,223	47,183	
		Ensembl Protein ID version	47,223	47,223	47,223	47,223	
		Total Number of Rows	68,932		47,223		

*this column contains multiple IDs in one row

Pathway Entities Integration The data integration process began with Pathway Ontology (PO) as the foundational framework, merging datasets from PO, KEGG, and Reactome. Missing data was subsequently addressed through equivalent mapping relations between KEGG and Reactome, as provided by ComPath. Finally, human pathway data from WikiPathway was integrated using equivalent mappings between KEGG and WikiPathway also facilitated by ComPath. Bolded columns in the table represent the fields used for merging with other databases, where the IDs in these columns are uniquely assigned (refer to **Figure S4** in the supplementary section for the detailed integration workflow and **Table 5** for results).

Table 5. Pathway Entity Information

Database		**Raw Data**		**After Data Cleaning**		**Total Number of BioMedGraphica ID**
		Total Number of Rows	**Unique**	**Total Number of Rows**	**Unique**	
Pathway Ontology	PO ID	2677	2677	2677	2677	6,724
	KEGG ID	237	211	201	179	
	Reactome ID	326	320	326	320	
	Total Number of Rows	2677		2677		
KEGG	**KEGG ID**	359	359	359	359	
	Total Number of Rows	359		359		
Reactome	**Reactome ID**	2711	2711	2711	2711	

	Total Number of Rows	2711		2711	
ComPath	**KEGG ID**	953	859	113	89
	Reactome ID	1,274	1,055	58	57
	WikiPathway ID	940	724	55	55
	Total Number of Rows	1,592		113	
WikiPathway	**WikiPathway ID**	1,510	1,510	1,510	1,510
	Total Number of Rows	1,510		1,510	

*this column contains multiple IDs in one row

Values separated by semicolons indicate data from multiple files

Metabolite Entities Integration The integration of the two databases utilized the ChEBI ID as the primary linking key. Subsequently, entries with identical HMDB IDs were consolidated, establishing the HMDB ID as the smallest unit of data granularity. Columns highlighted in bold within the table denote those used for database merging, ensuring the uniqueness of IDs in these columns (see **Figure S5** in the supplementary section for details on the merging process and **Table 6** for the results).

Table 6. Metabolite Entity Information

Database		Raw Data		After Data Cleaning		Total Number of BioMedGraphica ID
		Total Number of Rows	**Unique**	**Total Number of Rows**	**Unique**	
HMDB	HMDB ID	217,920	217,920	217,920	217,920	218,333
	CAS	15,672	15,647	15,672	15,647	
	ChemSpider ID	31,269	31,015	31,269	31,015	
	PubChem CID	104,230	103,682	104,230	103,682	
	ChEBI ID	13,701	13,562	13,701	13,562	
	PDB ID	522	520	522	520	
	KEGG ID	6,814	5,908	6,814	5,908	
	Total Number of Rows	217,920		217,920		
ChEBI	**ChEBI ID**	5,750; 393,172	5,750;161,158	2,931	2,931	
	CAS Number	NA; 28,867	NA; 28,695	1,698	1,670*	
	HMDB ID	NA; 19,619	NA; 19,203	2,201	2,169*	
	Total Number of Rows	5,750; 393,172		2,931		

*this column contains multiple IDs in one row

Values separated by semicolons indicate data from multiple files

Microbiota Entities Integration The NCBI Taxon ID was employed as the standard key for harmonizing data across all microbiota datasets. This identifier was chosen due to its widespread presence in the included databases, enabling data merging. Columns highlighted in bold within the accompanying table indicate those used for cross-database integration, ensuring the uniqueness of IDs in these fields. For a

detailed explanation of the integration methodology, refer to **Figure S6** in the supplementary section, with comprehensive results presented in **Table 7**.

Table 7. Microbiota Entities Information

Database		Raw Data		After Data Cleaning		Total Number of BioMedGraphica ID
		Total Number of Rows	Unique	Total Number of Rows	Unique	
NCBI	**NCBI ID**	2,614,590	2,614,590	532,783	532,783	616,390
	Total Number of Rows	2,614,590		532,783		
SILVA	SILVA ID	227,318; 2,224,690	157,873; 2,152,602	272,419	2,214,227*	
	NCBI ID	227,318; 2,224,690	51,697; 267,817	272,419	272,419	
	Total Number of Rows	227,318; 2,224,690		272,419		
Greengene	Greengene ID	1,144,866	1,144,866	92,684	1,144,866*	
	NCBI ID	1,144,866	92,684	92,684	92,684	
	RNAcentral ID	1,144,866	1,004,892	92,684	1,004,892*	
	Total Number of Rows	1,144,866		92,684		
RDP	RDP ID	10,302	10,302	2,487	10,302*	
	NCBI ID	10,302	2,487	2,487	2,487	
	RNAcentral ID	10,302	4,779	2,487	4,779*	
	Total Number of Rows	10,302		2,487		
GTDB	GTDB ID	12,477; 584,382	12,477; 584,382	92,444	596,859*	
	NCBI ID	12,477; 584,382	2,768; 89,701	92,444	92,444	
	Total Number of Rows	12,477; 584,382		92,444		

*this column contains multiple IDs in one row

Values separated by semicolons indicate data from multiple files

Exposure Entity Integration The data integration for this entity was based on the CAS number. Since all exposure databases contain the CAS ID, it was leveraged for integrating these databases. The bolded content in the table indicates the columns used for merging with other databases, where the IDs in those columns are unique (see **Figure S7** in the supplementary section for merging process and results in **Table 8**).

Table 8. Exposure Entities Information

Database		Raw Data		After Data Cleaning		Total Number of BioMedGraphica ID
		Total Number of Rows	Unique	Total Number of Rows	Unique	
CTD	MeSH ID	179,179	179,179	179,179	179,179	532,942

	CAS ID	56,661	56,661	56,661	56,661	
	Total Number of Rows	179,179		179,179		
ToxCast	ToxCast ID	9,403	9,403	9,403	9,403	
	CAS ID	9,403	9,403	9,403	9,403	
	Total Number of Rows	9,403		9,403		
ChemIDplus	ChemIDplus ID	409,325	409,325	409,354	409,325	
	CAS ID	409,325	409,325	409,354	409,325	
	PubChem CID	336,401*	327,151	336,430	327,151	
	PubChem SID	409,160*	409,160	409,160	409,160	
	Total Number of Rows	409,325		409,354		

*this column contains multiple IDs in one row

Phenotype Entity Merging The integration process began with importing data from the HPO database (version 2024-8-13), including HPO identifiers and their associated terms. Relevant rows containing valid phenotype labels were isolated by filtering out entries with raw HPO identifiers. To refine the phenotype labels, a predefined list of descriptive expressions deemed unnecessary for entity extraction, such as "obsolete," "increased," "decreased," and similar terms commonly used in phenotype descriptions, was systematically removed. This cleaning process employed regular expression patterns for precision. Afterward, duplicate entries were consolidated with the assistance of LLM, resulting in a refined dataset where each unique cleaned label (represented by a BioMedGraphica ID) was associated with one or more HPO IDs. Columns highlighted in bold within the table represent fields used for database merging, with IDs in these columns being unique (see **Figure S8** in the supplementary section for details on the merging workflow and **Table 9** for results).

Table 9. Phenotype entities database description

Database		**Raw Data**		**After Data Cleaning**		**Total Number of BioMedGraphica ID**
		Total Number of Rows	**Unique**	**Total Number of Rows**	**Unique**	
HPO	HPO ID	22,879	22,879	17,711	17,711	17,711
	Total Number of Rows	22,879		17,711		
UMLS	UMLS ID	14,036,386	3,211,875	14,914	15,958*	
	HPO ID	40,197	17,664	14,914	14,914	
	Total Number of Rows	14,036,386		14,914		

*this column contains multiple IDs in one row

Disease Entity Integration The integration of disease entities began with the alignment of UMLS and MeSH datasets. Subsequently, SNOMED-CT data was incorporated, leveraging its comprehensive mappings to ICD-10. This was followed by the consolidation of mappings for ICD-10 and ICD-11. Using the

relations provided by Disease Ontology, UMLS, MeSH, and ICD-10 were mapped to append the corresponding Disease Ontology (DO) IDs to the dataset. Missing UMLS data was then supplemented through mappings between UMLS and SNOMED-CT. Finally, Mondo data was integrated using its mappings to UMLS and MeSH. Throughout the integration process, the UMLS ID was designated as the smallest unit of data granularity, ensuring unique identification across the entire dataset. Bolded columns in the table indicate the fields used for database merging, with IDs in these columns being uniquely assigned (refer to **Figure S9** in the supplementary section for a detailed workflow and **Table 10** for results).

Table 10. Disease Entity Information

Database		Raw Data		After Data Cleaning		Total Number of BioMedGraphica ID
		Total Number of Rows	Unique	Total Number of Rows	Unique	
UMLS	UMLS ID	14,036,386	3,211,875	62,479	62,479	198,730
	ICD10	110,875	97,615	12,805	11,359*	
	MeSH ID	1,010,573	354,269	10,173	7,829	
	OMIM ID	197,480	107,706	6,129	8,470*	
	SNOMEDCT ID	988,281	372,524	37,334	37,278*	
	Total Number of Rows	14,036,386		62,479		
MeSH	**MeSH ID**	5,032	5,032	5,032	5,032	
	Total Number of Rows	5,032		5,032		
SnomedCT	**SnomedCT ID**	204,464	147,056	126,961	126,961	
	ICD10 ID	170,993	11,033	86,353	9,662*	
	Total Number of Rows	204,464		126,961		
ICD11	**ICD11 ID**	34,663	34,663	34,663	34,663	
	Total Number of Rows	36,044		34,663		
ICD10	**ICD10 ID**	12,597	12,597	10,077	10,077	
	ICD11 ID	12,301	6,710	10,077	5,876*	
	Total Number of Rows	12,597		10,077		
Disease Ontology	DO ID	38,037	10,942	10,942	10,942	
	UMLS ID	6,976	6,931	6,367	6,931*	
	MeSH ID	4,044	3,679	3,967	3,679*	
	ICD10 ID	3,656	2,438	3,530	2,438*	
	OMIM ID	6,010	5,979	5,492	5,979*	
	Total Number of Rows	38,037		10,942		
Mondo	Mondo ID	134,406	26,466	29,270	29,270	
	UMLS ID	20,918	20,918	20,918	20,918	
	MeSH ID	8,354	8,188	8,230	8,188*	
	OMIM ID	9,931	9,931	9,837	9,837*	
	Total Number of Rows	134,406		29,270		

*this column contains multiple IDs in one row

Drug Entity Merging The integration process commenced by merging NDC and UNII datasets using the SUBSTANCENAME as the key identifier. PubChem data was then incorporated through its mapping with PubChem CIDs. CAS data integration followed, leveraging the relation between PubChem CIDs and CAS

numbers. ChEBI data was subsequently added using InChI as the common identifier. DrugBank data was integrated next, utilizing mappings between DrugBank IDs, CAS numbers, and SIDs. Finally, any missing data within the same row was supplemented using synonyms from both PubChem and DrugBank. The CAS number was designated as the minimal unit of data granularity for this entity. Bolded entries in the table indicate the columns used for merging across databases, where IDs in these columns are uniquely assigned (refer to **Figure S10** in the supplementary section for details on the merging process and **Table 11** for the results).

Table 11. Drug Entity Information

Database		Raw Data		After Data Cleaning		Total Number of BioMedGraphica ID
		Total Number of Rows	Unique	Total Number of Rows	Unique	
NDC	**UNII Name**	111,491	8,809	7,859	7,859	626,581
	NDC ID	113,850	112,614	7,859	110,275*	
	Total Number of Rows	113,850		7,859		
UNII	UNII ID	152,870; 154,561	152,870; 154,561	157,960	154,569	
	UNII Name	152,870; 154,561	152,870; 154,561	156,082	154,568	
	PubChem CID	112,013; 112,245	110,750; 111,447	117,965	116,580	
	PubChem SID	152,727; NA	152,727; NA	156,082	152,727	
	CAS Number	NA; 116,361	NA; 114,937	119,699	114,937	
	Total Number of Rows	152,870; 154,561		157,960		
PubChem	**PubChem CID**	20,472; 114,969	20,472; 114,969	123,155	123,155	
	Total Number of Rows	20,472; 114,969		123,155		
CAS	CAS Number	447,699	447,699	402,378	386,907*	
	PubChem CID	389,855	402,378*	402,378	402,378	
	Total Number of Rows	447,699		402,378		
ChEBI	**ChEBI ID**	44,924; 225,102	44,924; 43,260	44,101; 17,691	44,924*; 17,691	
	CAS Number	NA; 27,735	NA; 17,601	NA; 17,691	NA; 17,601*	
	Total Number of Rows	44,924; 225,102		44,101; 17,691		
DrugBank	DrugBank ID	16,581	16,581	16,581	16,581	
	CAS Number	10,102	10,070	10,102	10,070	
	PubChem CID	8,724	8,724	8,724	8,724	
	PubChem SID	10,450	10,450	10,450	10,450	
	Total Number of Rows	16,581		16,581		

*this column contains multiple IDs in one row

Values separated by semicolons indicate data from multiple files

2.2.2 Relation Integration

The construction of edges utilized data from **22** distinct databases, mapping raw database IDs to their corresponding BioMedGraphica IDs to form relations. A notable challenge arose from one-to-many mappings, where a single database ID, such as A0PJY2 (UniProt ID), corresponded to multiple BioMedGraphica IDs (BMG_PT033926 and BMG_PT044226), due to the UniProt and Ensembl databases having one-to-many relations. Aside from this, all relations were directional and presented in a From-To format. To address bidirectional relations, two distinct methodologies were employed. The first involved reversing the direction of the relation. For instance, while protein-protein interactions are intrinsically bidirectional, the original dataset lacked explicit directionality. To resolve this, a reversed copy of the data was generated, merged with the original dataset, and duplicates were subsequently eliminated. The second approach entailed establishing new relations where reversal was inappropriate. For example, in disease-phenotype associations, reversing the data alone was insufficient; instead, a complementary phenotype-to-disease relation was created to accurately represent the connection. The edge structure was meticulously designed to conform to a one-to-one mapping framework, ensuring that each instance of a database ID mapping to multiple BioMedGraphica IDs resulted in the generation of distinct edges. This strategy significantly amplified the total number of edges, exceeding a straightforward summation of inter-database relations due to the one-to-many nature of the mappings.

Table 12. Harmonized Relations Information

Interaction Type	Database	Initial Edge Number	Matching		Total
			Unique	Total	
Gene-Transcript	Ensembl	278,220	274,774	277,924	278,352
	RefSeq	33,401	6,870	6,994	
Transcript-Protein	Ensembl	123,540	123,528	123,528	408,270
	Uniprot	122,941	58,831	306,748	
	RefSeq	30,193	6,865	47,823	
Protein-Protein	BioGrid	1,660,759	1,660,213	16,309,791	32,916,130
	STRING	13,715,404	13,287,544	13,287,544	
	KEGG	52,155	51,317	3,739,606	
Protein-Pathway	KEGG	21,051	20,866	207,492	207,492
Protein-Phenotype	HPO	255,891	255,703	2,598,140	2,598,140
Protein-Disease	UniProt	4,775	4,735	17,170	1,318,893
	DISEASES	346,173	283,986	608,291	
	HPO	7,336	7,139	8,256	
	DisGeNet	91,484	80,183	80,183	
Pathway-Protein	KEGG	24,475	24,367	239,199	239,199
Pathway-Drug	KEGG	2,334	2,011	3,227	3,227
Pathway-Exposure	CTD	1,591,611	1,537,730	1,532,389	1,532,389
Metabolite-Protein	HMDB	863,759	849,980	2,124,047	2,124,047
Metabolite-Metabolite	MetaNetX	23,711	886	931	931
Metabolite-Disease	HMDB	24,755	24,667	24,967	24,979
Microbiota-Disease	DisBiome	616,390	10,414	10,414	10,414
Microbiota-Drug	MDAD	5,055	668	1,770	1,837

	PharmacoMicrobiomics	69	67	67	
Exposure-Gene	CTD	766,820	763,150	763,150	763,150
Exposure-Pathway	CTD	1,591,611	1,537,730	1,532,389	1,532,389
Exposure-Disease	CTD	9,114,984	3,185,626	4,537,063	4,537,063
Phenotype-Phenotype	HPO	23,436	14,678	14,678	14,678
Phenotype-Disease	HPO	153,944	152,017	177,481	177,481
Disease-Phenotype	HPO	153,944	152,017	177,265	177,265
Disease-Disease	DO	11,683	9,654	121,806	121,806
Drug-Protein	DrugBank	21,941	3,208	18,989	269,972
	BindingDB	1,508,106	61,844	199,783	
	DrugCentral	14,301	13,412	72,391	
Drug-Pathway	KEGG	3,922	3,646	5,906	5,906
Drug-Metabolite	HMDB	3,258	2,431	3,577	3,577
Drug-Microbiota	MDAD	5,055	668	1,770	1,837
	PharmacoMicrobiomics	69	67	67	
Drug-Phenotype	SIDER	152,759	58,903	58,377	58,377
Drug-Disease	DrugCentral	58,013	45,669	117,234	117,234
Drug-Drug	DrugBank	2,839,610	2,679,244	7,156,432	7,156,432

2.3 Final Results

The database for ***BioMedGraphica*** includes **11** entity types and **27** edge types, contains **3,132,161** entities and **56,825,152** relations, composing the knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ (check number of each entity and edge type in **Table 13** and **Table 14**).

Table 13. Summarized Entity Information

Entity Type	Math Annotation	Count	Percentage
Promoter	$\mathcal{V}^{(pm)}$	215,608	6.8845%
Gene	$\mathcal{V}^{(gn)}$	215,608	6.8845%
Transcript	$\mathcal{V}^{(ts)}$	278,326	8.8871%
Protein	$\mathcal{V}^{(pt)}$	204,835	6.5405%
Pathway	$\mathcal{V}^{(pw)}$	6,724	0.2147%
Metabolite	$\mathcal{V}^{(mt)}$	218,333	6.9715%
Microbiota	$\mathcal{V}^{(mc)}$	616,390	19.6817%
Exposure	$\mathcal{V}^{(ep)}$	532,942	17.0172%
Phenotype	$\mathcal{V}^{(ph)}$	17,711	0.5655%
Disease	$\mathcal{V}^{(ds)}$	198,730	6.3456%
Drug	$\mathcal{V}^{(dg)}$	626,581	20.0071%
Total	$\mathcal{V}$	3,131,788	100%

Table 14. Summarized Information of Relation Types

Relation Type	Math Annotation	Count	Percentage
Promoter-Gene	$\mathcal{E}^{(pm-gn)}$	215,608	0.3795%
Gene-Transcript	$\mathcal{E}^{(gn-ts)}$	278,352	0.4899%
Transcript-Protein	$\mathcal{E}^{(ts-pt)}$	408,270	0.7186%
Protein-Protein	$\mathcal{E}^{(pt-pt)}$	32,916,130	57.9335%
Protein-Pathway	$\mathcal{E}^{(pt-pw)}$	207,492	0.3652%

Protein-Phenotype	$\mathcal{E}^{(pt-ph)}$	2,598,140	4.5728%
Protein-Disease	$\mathcal{E}^{(pt-ds)}$	1,318,893	2.3213%
Pathway- Protein	$\mathcal{E}^{(pw-pt)}$	239,199	0.4210%
Pathway-Drug	$\mathcal{E}^{(pw-dg)}$	3,227	0.0057%
Pathway-Exposure	$\mathcal{E}^{(pw-ep)}$	1,532,389	2.6971%
Metabolite-Protein	$\mathcal{E}^{(mt-pt)}$	2,124,047	3.7384%
Metabolite-Metabolite	$\mathcal{E}^{(mt-mt)}$	931	0.0016%
Metabolite-Disease	$\mathcal{E}^{(mt-ds)}$	24,967	0.0439%
Microbiota-Disease	$\mathcal{E}^{(mc-ds)}$	10,414	0.0183%
Microbiota-Drug	$\mathcal{E}^{(mc-dg)}$	1,837	0.0032%
Exposure-Gene	$\mathcal{E}^{(ep-gn)}$	763,150	1.3432%
Exposure-Pathway	$\mathcal{E}^{(ep-pw)}$	1,532,389	2.6971%
Exposure-Disease	$\mathcal{E}^{(ep-ds)}$	4,537,063	7.9854%
Phenotype-Phenotype	$\mathcal{E}^{(ph-ph)}$	14,678	0.0258%
Phenotype-Disease	$\mathcal{E}^{(ph-ds)}$	177,481	0.3124%
Disease-Phenotype	$\mathcal{E}^{(ds-ph)}$	177,265	0.3120%
Disease-Disease	$\mathcal{E}^{(ds-ds)}$	121,806	0.2144%
Drug-Protein	$\mathcal{E}^{(dg-pt)}$	269,972	0.4752%
Drug-Pathway	$\mathcal{E}^{(dg-pw)}$	5,906	0.0104%
Drug-Metabolite	$\mathcal{E}^{(dg-mt)}$	3,577	0.0063%
Drug-Microbiota	$\mathcal{E}^{(dg-mc)}$	1,837	0.0032%
Drug-Phenotype	$\mathcal{E}^{(dg-ph)}$	58,377	0.1027%
Drug-Disease	$\mathcal{E}^{(dg-ds)}$	117,234	0.2063%
Drug-Drug	$\mathcal{E}^{(dg-dg)}$	7,156,432	12.5956%
Total	$\mathcal{E}$	56,817,063	100%

2.4 Software for Improving Data Integration and Generation

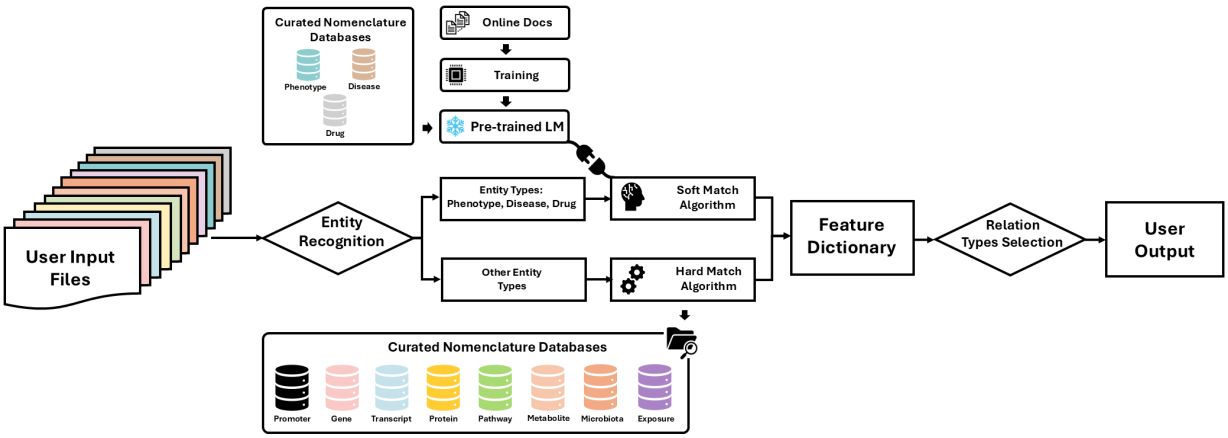


Figure 2. Workflow of software ***BioMedGraphica.***

As shown in **Figure 2**, user can input the files into ***BioMedGraphica*** software, which are denoted as $\mathcal{X} = \{\mathcal{X}^{(pm)}, \mathcal{X}^{(gn)}, \mathcal{X}^{(ts)}, \mathcal{X}^{(pt)}, \mathcal{X}^{(pw)}, \mathcal{X}^{(mt)}, \mathcal{X}^{(mc)}, \mathcal{X}^{(ep)}, \mathcal{X}^{(ph)}, \mathcal{X}^{(ds)}, \mathcal{X}^{(dg)}\}$, where $\mathcal{X}^{(e)} \in \mathbb{R}^{n^{(e)} \times |\mathcal{F}^{(e)}|}$ and e denotes one of the **11** entity types mentioned above, $n^{(e)}$ stands for the number of samples, and

$\mathcal{F}^{(e)}$represents features set of the entity type e. After inputting files into the software, the number of samples across the different entity types will be aligned into $n^{(in)}$, which is the intersection of all input files. In this way, the input files can be considered as a giant unified file with matrix $\mathcal{X}^{(in)} \in \mathbb{R}^{n^{(in)} \times \mathcal{F}^{(in)}}$, and $|\mathcal{F}^{(in)}| = \sum_e |\mathcal{F}^{(e)}|$. By matching the features $\mathcal{F}^{(in)}$ with entities $\mathcal{V}$ existing in ***BioMedGraphica*** knowledge graph $\mathcal{G}$, the entities will be formed with $\mathcal{V}^{(sub)}$ and mapping function $\mathcal{D}: \mathcal{F}^{(in)} \rightarrow \mathcal{V}^{(in)}$, which is curated in python dictionary format. Aside from this, users can choose the types of relations they would like to form in this process, resulting in $\mathcal{G}^{(in)} = (\mathcal{V}^{(in)}, \mathcal{E}^{(in)})$. Following we describe the details of how entities are matched, and relations are formed.

2.4.1 Entity Recognition

When matching the features inputted by users to the existing entities in the ***BioMedGraphica*** knowledge base, the specialized designed algorithm using a pre-trained BioBERT model was leveraged for phenotype, drug, disease entities, which allows for the comparison of phenotypic, drug and disease terms based on their semantic similarity for building the mapping function $\mathcal{D}$. Then, the similarity score will be calculated between a given queried feature name, f_{name} (f is the corresponded entity), and precomputed entity embeddings by scoring function S with

$$S(f) = \frac{\text{LM}(f_{\text{name}})^{\text{T}}\, \text{LM}(\mathcal{V}_{\text{name}})}{\|\text{LM}(f_{\text{name}})\| \cdot \|\text{LM}(\mathcal{V}_{\text{name}})\|} \tag{1}$$

, where f_{name} ($f_{\text{name}} \in \mathcal{F}_{\text{name}}^{(in)}$) is the queried feature name from the unified user input file, $\mathcal{F}_{\text{name}}^{(in)}$ ($\mathcal{F}_{\text{name}}^{(in)} \in \mathbb{R}^{|\mathcal{F}^{(in)}|}$) and $\mathcal{V}_{\text{name}}$ ($\mathcal{V}_{\text{name}} \in \mathbb{R}^{|\mathcal{V}|}$) is the corresponding entity names of $\mathcal{F}^{(in)}$ and $\mathcal{V}$, and pre-trained BioBERT language model (LM) model is denoted as LM. In detail, the model will process phenotype, drug and disease entities in ***BioMedGraphica*** by

$$z = \text{LM}(v_{\text{name}}) \tag{2}$$

, where v_{name} ($v_{\text{name}} \in \mathcal{V}_{\text{name}}$) is entity name and z ($z \in \mathbb{R}^d$) denotes the transformed embedding space for v_{name}. Similarly, the queried feature name will be embedded by

$$z' = \text{LM}(f_{\text{name}}) \tag{3}$$

, where z' ($z' \in \mathbb{R}^d$) denotes the transformed embedding space for f_{name}. Afterwards, the top k most similar entities will be extracted by

$$\mathcal{V}_k^{(f)} = I[\text{argmax}_k[S(f)]] \tag{4}$$

, where argmax_k can identify top k most similar entity names $\mathcal{V}_k^{(f)}$ ($\mathcal{V}^{(f)} \in \mathbb{R}^k$) and $I(\cdot)$ is the one-to-one mapping function which will map the entity names to entities in ***BioMedGraphica***. In these top k most similar entity, the user will define the only one entity, $\mathcal{V}^{(f)}$, to be matched for the queried feature name f_{name}. For other entity types, the hard match method was leveraged to search for exact entity name for the queried feature name f_{name} with $\mathcal{V}^{(f)}$. With this, the dictionary function $\mathcal{D}$ will be generated.

2.4.2 Relation / Knowledge Graph Construction

By extracting the corresponding entities $\mathcal{V}^{(sub)}$ of the input features $\mathcal{F}^{(in)}$ from the whole knowledge graph $\mathcal{G}$, users can select the edge types annotated in Table 14 to construct the $\mathcal{E}^{(in)}$.

2.4.3 Graphical User Interface (GUI) Design

The GUI for this workflow is designed to streamline the process of data input, recognition, and filtering before output. The interface begins with a user input section, where users can either upload their data file or manually input data. The system will automatically perform data recognition, displaying the detected data format in a preview pane for confirmation. Users are then presented with options to select the format recognition type from a dropdown or radio buttons (e.g., Entity Type A, Entity Type B, etc.), allowing them to specify the format more accurately if necessary. Once the format is confirmed, the GUI moves to the entity matching section, where users can match their input data to the BioMedGraphica ID system. Only the data that matches the BioMedgraphica ID will be kept for further processing. Users can then filter the matched data based on their choice of relational entities from a selection of databases (e.g., Relation Database 1, Relation Database 2). Finally, after applying the desired filters, the system will produce the data output, which users can download or view in a structured format, concluding the process. The GUI is designed to be user-friendly, guiding the user through each step with clear instructions and real-time feedback on their data processing choices.

3 Data Records

Due to licensing restrictions or controlled access to certain portions of the data, raw data download links are provided in the supplementary materials. Tutorials for processing the raw data into harmonized entities and relations are available. After completing the procedures outlined in the tutorial, the ***BioMedGraphica*** database can serve as a comprehensive knowledge base, forming the foundation for initiating the ***BioMedGraphica*** software. A detailed software tutorial is also available for reference. Both tutorials can be found in GitHub link:
https://github.com/FuhaiLiAiLab/BioMedGraphica/blob/main/README.md

4 Technical Validation

4.1 Entities / Relations Matching Accuracy

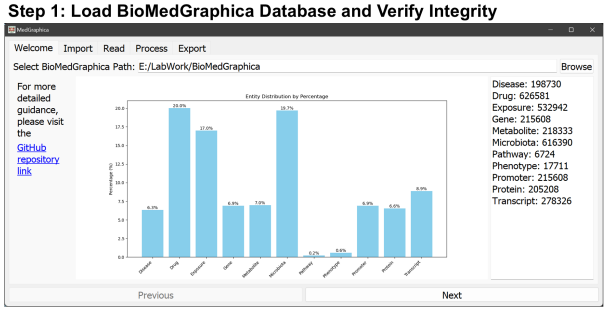
For entity recognition part, the hard match strategy can match most of the entities if such nomenclature system was collected in ***BioMedGraphica*** knowledge base. For soft matching strategy, the pre-trained language model significantly improved the matching accuracy and efficiency. Nevertheless, even current advanced large language models cannot provide solid support for named entity recognition in zero-shot or few-shots scenarios. For example, ChatGPT-4o will mistakenly match the NC_000019.10 (RefSeq ID) with an incorrect Ensembl ID ENSG00000272512. Aside from this, the relation identification is also very difficult, let alone the knowledge graph construction. For instance, when ChatGPT-4o tried to identify the relation between phenotype Leukocytosis and drug with CAS number 106-60-5, it will mistakenly match the drug

with incorrect name due to its poor performance on named entity recognition. Hence, it will make incorrect assertions about the relations.

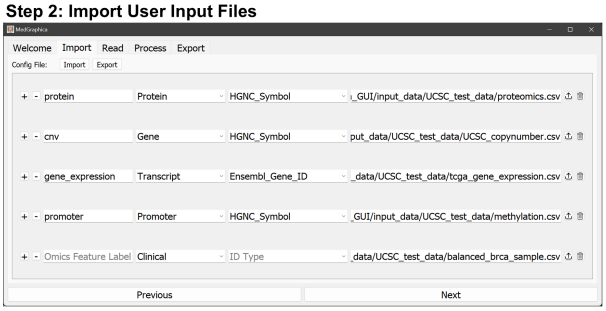
4.2 A Case Study of Software *BioMedGraphica*

Users are required to transform the raw data into formats such as csv, txt or tsv, which are compatible with conversion into two-dimensional dataframes. These formats allow the data to be easily structured and manipulated for further processing. By adhering to the software's prescribed workflow, users can ensure that the final output will be generated as numpy files, a widely used format for numerical data in scientific computing. For instance, a specific example using The Cancer Genome Atlas Program (TCGA) multi-omics and clinical cancer genomics dataset is provided. This dataset serves as the input, including selected features such as methylation, copy number variation (CNV), gene expression, proteomics, and breast cancer (BRCA) clinical data. After undergoing some basic data processing steps, the data can be transformed into the required $n \times F_p$ dimensions and imported into the software for integration within a graph-based AI model. For further technical details and step-by-step guidance, users are encouraged to refer to **Figure 3** and the associated GitHub repository link:
https://github.com/FuhaiLiAiLab/BioMedGraphica/blob/main/README.md

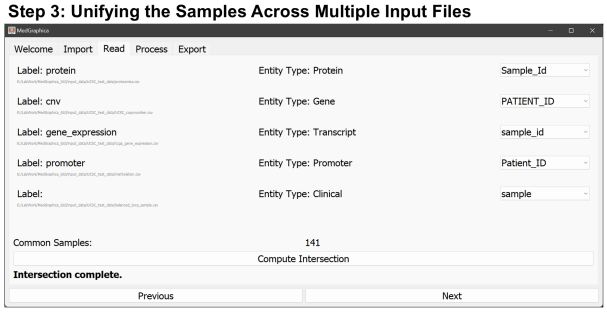
Step 1: Load BioMedGraphica Database and Verify Integrity



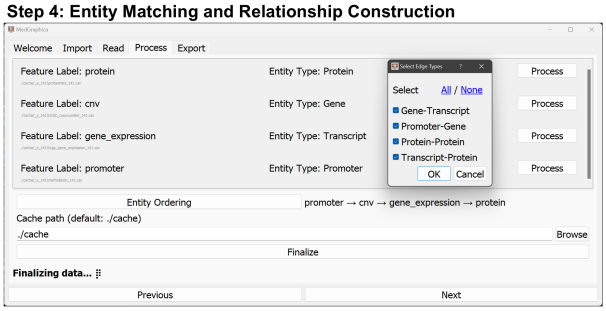
Step 2: Import User Input Files



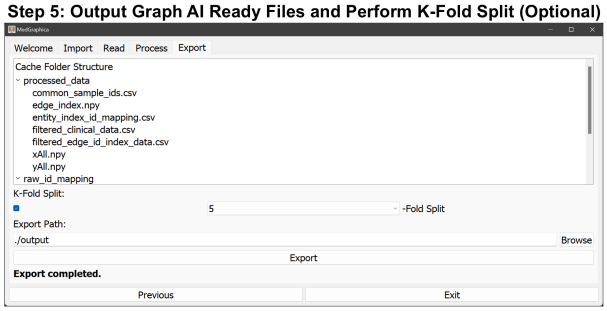
Step 3: Unifying the Samples Across Multiple Input Files



Step 4: Entity Matching and Relationship Construction



Step 5: Output Graph AI Ready Files and Perform K-Fold Split (Optional)



Final Output Files Structure

```
output/
  processed_data/
    common_sample_ids.csv
    edge_index.npy
    entity_index_id_mapping.csv
    filtered_clinical_data.csv
    filtered_edge_id_index_data.csv
    xAll.npy
    yAll.npy
    x_train_fold_1.npy
    ......
    y_test_fold_5.npy        } K-Fold Split (K=5)
  raw_id_mapping/
    cnv_id_map.csv
    gene_expression_id_map.csv
    promoter_id_map.csv
    protein_id_map.csv
    ......
```

Figure 3. *BioMedGraphica* Software Graphical User Interface and User Demonstration Using the TCGA Dataset. Step 1: Validate the integrity of the BioMedGraphica database by locating its path. This step includes assessing the distribution and count of various entities within the database. Step 2: Users input the file path and specify corresponding attributes for each file feature, such as entity labels, types, and data nomenclature. In the example using the TCGA dataset, *Ensembl_Gene_ID* and *HGNC_Symbol* were used to annotate four types of entities—protein, gene, transcript, and promoter. Clinical data served as output labels, requiring no additional label or entity type selection. Step 3: File attributes and the name of the first column in each file are displayed for user confirmation of the entity file format. After validation, users click 'Compute Intersection' to preprocess the data, reducing the TCGA BRCA demo dataset to identified 141 common samples. Step 4: After matching entities from the input files, relationship with subgraph from whole knowledge graph will be extracted. Users can choose to build all or specific relationships, such as Promoter-Gene, Gene-Transcript, Protein-Protein, and Transcript-Protein. Step 5: Review the output files and optionally perform a K-Fold Split for subsequent graph AI model training and evaluation. The final output directory includes processed graph AI-ready files and mappings from original entity IDs to BioMedGraphica-specific IDs.

5 Usage Notes

Database Preparation and Access Due to access control restrictions on the collected database, we provide users with data download links for convenient access. Additionally, we offer a Jupyter notebook that facilitates the merging of entities and the harmonization of relations into designated folders. Once the data has been curated and placed in the appropriate folder, users can run the software locally on their machine in a client-based environment to begin processing.

Required File Preparation Before running the software, it is essential to prepare the necessary files, including entity files and a clinical data file. Each file should use a standardized naming convention for sample IDs and a consistent database identifier format for features, like the Ensembl stable Gene ID. Files will be intersected in the final step to obtain a common sample set. For more detailed data formatting guidelines, please refer to the GitHub link:
https://github.com/FuhaiLiAiLab/BioMedGraphica/blob/main/README.md

Starting the GUI Once files are prepared, launch the GUI. On the Welcome tab, locate the BioMedGraphica database path and verify its integrity to ensure smooth processing. Next, go to the Import tab, where you'll provide each entity file with a unique label to identify it during subsequent processing. Select the appropriate entity type and ID type for each file from the dropdown menu. Then, use the file path button to select each file's path. When all inputs are complete, click the Export button at the top of the page to save inputs as config.csv, making future processing easier. Click Next to proceed to the Read tab.

File Reading and Validation In the Read tab, the software will read column names to identify the sample ID column in each file. For simplicity and readability, place the sample ID column as the first column in each file and label it as id. The software will also perform an intersection of sample IDs across all entity files to obtain a common set of samples, reducing storage requirements. After confirming that the data is read correctly, click Next to proceed to the Process tab.

Processing and Finalizing Files The Process tab includes individual processing for each entity file (saved to /cache in the root directory). Sort entity files by clicking and arranging them from top to bottom in the dialog box. Import clinical data and aggregate all individually processed entity files into a format required for GNN training with the Finalize function. Note that if files have many columns, the Finalize operation may need significant available RAM (20GB+). In case of issues or unexpected exits during individual file processing, reprocess the specific entity file, then use Finalize to aggregate it with the others. The final output heavily depends on the contents of the cache folder; if starting a new process, click the Clear Cache Folder button in the top-right corner to clear the cache. Once processing is complete, click Next to enter the Export tab. Here, preview all processed files in the cache folder. After confirming accuracy, set the save path and click Export. When all tasks are finished, click Exit to close the software.

Acknowledgements

Competing Interests

The authors have no competing interests.

Author contribution

Methodology and project was designed by HZ, SL, TX, WL, DH, YD, GL, PM, MS, JC, WB, PD, PG, RF, CC, YC, MP, PP, FL. The manuscript was written by HZ, SL, TX, WL, FL. FL conceptualized the project. SL, HZ, WL, YD contributed to data collection and analysis. TX, HZ developed software.

Funding Information

This study was partially supported by NIA 1R21AG078799-01A1, NLM 1R01LM013902-01A1, NIA R56AG065352, NINDS 1RM1NS132962-01.

Supplementary Materials

Section A. Details of Data Resources

A.1. Data Resources for Entities

Ensembl[11] Ensembl is a widely used resource for genome annotation and provides access to a wide variety of genomic data across numerous species, with a strong focus on vertebrate genomes. The Ensembl project integrates gene, transcript, and protein data, offering detailed genomic features. This database was accessed through the BioMart API, which allows flexible retrieval of large datasets based on specific criteria. For gene entities, Gene Stable IDs were selected, including versioned identifiers that ensure traceability across different releases. Important genomic features such as gene start and end positions, biotypes, and chromosomal coordinates were extracted. To maintain consistency in gene nomenclature, the mapping relations between Ensembl and the HUGO Gene Nomenclature Committee (HGNC) were preserved. This consistency is crucial for ensuring that gene annotations align across various databases. Additionally, transcript entities were obtained using Transcript Stable IDs and corresponding Gene IDs, while protein entities were extracted with Protein Stable IDs. Mapping relations between Ensembl, UniProt, and RefSeq were also preserved to ensure accurate and cross-compatible dataset integration.

OMIM[12] **(Online Mendelian Inheritance in Man)** OMIM is an essential resource for understanding the genetic basis of human diseases and provides detailed information on gene-phenotype relations. The database integrates clinical features with genetic data, offering insights into the hereditary nature of various conditions. Data from OMIM was retrieved to capture gene-related records, particularly focusing on mapping relations between OMIM IDs and NCBI gene IDs. This facilitated the standardization of gene-related data across different resources. Furthermore, HGNC symbols were retained to align OMIM gene identifiers with other databases used in this study. Chromosomal information was also supplemented, which aids in genomic localization and contextual understanding of the data. Ensuring the uniqueness of gene records was a priority, and merging was performed based on gene IDs to guarantee that each entry in the dataset remained unique and free from redundancy.

HGNC[13] **(HUGO Gene Nomenclature Committee)** The HGNC is the authoritative resource for assigning unique symbols and names to human genes. As gene nomenclature can vary across different databases, the HGNC serves as a standard for the human genome, providing approved gene symbols and names. Data was accessed via BioMart to extract HGNC-approved gene information, including attributes such as HGNC ID, gene symbol, gene name, and chromosomal location. The inclusion of HGNC data ensures that gene-related information in the dataset is standardized and consistent with official naming conventions. Mapping relations between HGNC, Ensembl, and NCBI IDs were retained to facilitate cross-referencing across these major databases. To ensure accuracy and prevent redundancy, the uniqueness of Ensembl IDs was verified during the merging process.

NCBI[14] **(National Center for Biotechnology Information) - Gene** The NCBI Gene database provides extensive information on genes and their functions, supporting a wide range of research in genetics, genomics, and bioinformatics. For this study, human gene data was extracted, retaining key attributes such as NCBI gene IDs, gene symbols, descriptive gene names, and chromosomal positions. The NCBI Gene database is an important resource for identifying gene sequences, gene structure, and gene functions, making it essential for the construction of a comprehensive gene dataset. Mapping relations between NCBI gene IDs and Ensembl IDs were preserved to ensure consistency across datasets, facilitating the integration of data from different sources. For microbiome-related data, entries from the NCBI Taxonomy Database were also included. This database provides authoritative taxonomic classifications, focusing on bacterial taxa, and corresponding NCBI Taxon IDs were retained to ensure accurate classification and integration with other microbiome datasets.

NCBI - RefSeq[15] **(Reference Sequence Database)** RefSeq is a curated collection of publicly available nucleotide sequences and their corresponding protein translations, which provides a critical reference standard for the annotation of genes, transcripts, and proteins. RefSeq data was retrieved for both gene and transcript entities in this study, focusing on entries with the status of either "REVIEWED" or "MODEL" to ensure high data quality. Essential attributes such as gene ID, RefSeq ID, and chromosomal information were retained to provide accurate gene annotations. Additionally, the MANE project, which provides a set of transcript alignments between RefSeq and Ensembl, was utilized to ensure that transcript mapping between these databases was consistent and high-quality. Protein entities were also integrated, with mapping relations between RefSeq, UniProt, and Ensembl retained to ensure cross-database compatibility. Uniqueness of the Ensembl IDs was verified throughout the data processing stages to ensure data integrity.

RNAcentral[16] RNAcentral is a comprehensive resource for non-coding RNA sequences, integrating data from over 40 specialist databases. RNAcentral provides access to a wide variety of RNA sequence information, including microRNAs, tRNAs, and other functional RNA molecules that play critical roles in gene regulation. Human-specific RNAcentral IDs and corresponding Ensembl IDs were retrieved for this study, ensuring that non-coding RNA entities could be accurately integrated with gene and protein data from other databases. The uniqueness of each Ensembl ID was verified to ensure the integrity of the dataset and to avoid duplications during the integration process.

UniProt[17] **(Universal Protein Resource)** UniProt is a globally recognized repository of protein sequences and functional information. It provides detailed annotations on protein sequences, structure, function, and interactions. Data from UniProt was accessed via its API, and UniProt IDs, along with protein names and their corresponding Ensembl IDs, were retrieved. This enabled the integration of protein-specific data with the broader dataset, ensuring that protein information was accurately cross-referenced with gene and

transcript data from Ensembl. The uniqueness of each Ensembl ID was verified during the data integration process to ensure consistency and to prevent errors in protein-related data.

Reactome[18] Reactome is a curated knowledgebase of biological pathways, and it is a key resource for understanding the molecular mechanisms underlying cellular processes. Human-specific pathway data was extracted from Reactome for this study, enabling the integration of pathway-related information with gene and protein data. The inclusion of Reactome pathways facilitates research into functional genomics and systems biology, where pathway analysis is critical for understanding complex biological processes.

KEGG[19] **(Kyoto Encyclopedia of Genes and Genomes)** KEGG is a comprehensive resource for understanding high-level functions and utilities of biological systems, such as cells, organisms, and ecosystems. Human pathway data was retrieved using the bioservices package, with a focus on integrating KEGG pathways with other biological pathways from Reactome and WikiPathways. The inclusion of KEGG enables the dataset to support metabolic and signaling pathway analysis, providing valuable insights into cellular functions and disease mechanisms.

WikiPathways[20] WikiPathways is an open, collaborative platform for the curation of biological pathways. Data from WikiPathways was converted to CSV format, and human-specific pathways were filtered for inclusion in this study. Mapping relations between WikiPathways, KEGG, and Reactome were maintained to ensure consistent integration of pathway-related data. The inclusion of WikiPathways supports research into a wide variety of biological pathways, complementing the curated data from Reactome and KEGG.

Pathway Ontology[21] Pathway Ontology provides a standardized framework for the classification of biological pathways and their relations. Preprocessing of the OBO-formatted file enabled the extraction of PO IDs, and mapping relations with KEGG and Reactome were preserved. This integration allows for comprehensive pathway analysis, ensuring that biological pathways from multiple sources can be consistently linked.

ComPath[22] ComPath is a database that integrates pathway mapping relations across KEGG, Reactome, and WikiPathways. All equivalent mappings were selected for this study, ensuring that pathway data from different sources could be cross-referenced. This comprehensive approach to pathway integration enables in-depth biological pathway analysis and facilitates the exploration of molecular mechanisms underlying diseases.

HMDB[23] **(Human Metabolome Database)** HMDB is the most comprehensive, freely accessible database of small molecule metabolites found in the human body. It provides extensive mapping relations related to metabolomics data. For this study, HMDB data was parsed from XML files, retaining key attributes such as

CAS number, SMILES, InChI, and mapping relations with other databases. The inclusion of HMDB supports research into human metabolism, drug interactions, and disease mechanisms, enabling detailed metabolomics analysis.

ChEBI[24] **(Chemical Entities of Biological Interest)** ChEBI is a database focused on 'small' chemical compounds and is used extensively for research in chemistry and biology. ChEBI provides manually annotated information about the structure, formula, and biological roles of chemical entities. In this study, ChEBI data was selected for drug entities, particularly those with a 3-star rating to ensure the highest data quality. Important attributes such as ChEBI ID, InChI, and the mapping relation to CAS Registry Numbers were retained. In addition to drug entities, metabolome data from ChEBI was included, focusing on human metabolites. Mapping relations with other databases, such as the Human Metabolome Database (HMDB), were preserved to enable cross-referencing of metabolite information.

SILVA[25] SILVA is a high-quality, curated database of ribosomal RNA (rRNA) sequences, widely used for taxonomic classification of microbial communities. Data from both the small subunit (SSU) and large subunit (LSU) ribosomal RNA sequences were included, along with corresponding NCBI Taxon IDs. The SILVA database provides valuable insights into the composition of microbiomes, supporting research into microbial diversity and ecology.

Greengenes[26] Greengenes is a database of 16S ribosomal RNA gene sequences used for the identification of microbial species. Data was sourced from RNAcentral, and RNAcentral IDs, Greengenes IDs, and NCBI Taxon IDs were retained to ensure consistent taxonomic classification of microbiome-related data. The inclusion of Greengenes allows for the accurate classification of bacterial species, supporting research into microbiomes and their impact on human health.

RDP[27] **(Ribosomal Database Project)** The Ribosomal Database Project (RDP) provides quality-controlled ribosomal RNA gene sequence data. Similar to Greengenes, RDP data was sourced via RNAcentral, and mapping relations between RNAcentral IDs, RDP IDs, and NCBI Taxon IDs were preserved. This allows for the consistent classification of microbial entities, supporting microbiome research and analysis.

GTDB[28] **(Genome Taxonomy Database)** GTDB is a comprehensive resource for the classification of Archaea and Bacteria. Data from GTDB was retrieved for both archaeal and bacterial entities, with GTDB IDs and NCBI Taxon IDs retained to ensure accurate taxonomic classification. By verifying the uniqueness of NCBI Taxon IDs, the dataset provides reliable support for microbiome research, enabling the exploration of microbial diversity across various environments.

CTD[29] **(The Comparative Toxicogenomics Database)** CTD is a pivotal resource for integrating chemical, gene, disease, and exposure data, facilitating the study of toxicogenomics and environmental health. CTD serves as an entity-centric database where chemicals, genes, and diseases are interconnected through curated interaction data. For chemical entities, CTD uses standardized identifiers such as Chemical Abstracts Service (CAS) numbers to ensure consistent representation and integration with other chemical databases like PubChem.

ToxCast[30] **(Toxicity Forecasting)** ToxCast is a large-scale program developed by the U.S. Environmental Protection Agency (EPA) to assess and predict the potential toxicological effects of chemicals using high-throughput screening methods. ToxCast evaluates thousands of chemicals across hundreds of biological assays, providing a comprehensive dataset to analyze chemical interactions with various biological pathways. The database is focused on integrating chemical, biological, and toxicity data to improve hazard identification, prioritize chemicals for further testing, and reduce reliance on traditional animal-based toxicity studies.

ChemIDplus[31] **(Chemical Identification Plus Database)** ChemIDplus is a comprehensive resource developed by the U.S. National Library of Medicine (NLM) that provides detailed information on over 400,000 chemical substances, including small molecules, mixtures, and complex compounds. It offers access to chemical properties, structures, synonyms, and regulatory data, making it a vital tool for researchers in toxicology, pharmacology, and environmental sciences. ChemIDplus utilizes standardized identifiers such as CAS Registry Numbers to ensure consistency and interoperability with other databases like PubChem.

HPO[32] **(Human Phenotype Ontology)** The Human Phenotype Ontology (HPO) provides a standardized vocabulary for phenotypic abnormalities encountered in human disease. Data was imported from the HPO database, version 2024-8-13, and relevant phenotype labels were extracted. After filtering and cleaning unwanted descriptive expressions, mapping relations between HPO IDs were retained, ensuring that phenotypic data could be integrated with other disease and genomic datasets. This integration facilitates research into genotype-phenotype correlations, a key area in genetic and clinical research.

ICD[33,34] **(International Classification of Diseases)** The International Classification of Diseases (ICD), maintained by the World Health Organization, is the global standard for the coding and classification of diseases. Both ICD-10 and ICD-11 codes were included to ensure that the dataset could be used in various research and clinical contexts. Mapping relations between ICD versions were retained, allowing for compatibility across different healthcare systems and facilitating research on disease epidemiology and outcomes.

Disease Ontology[35] **(DO)** The Disease Ontology (DO) provides a standardized ontology for the classification of human diseases. DO includes cross-references to other medical ontologies, such as UMLS, MeSH, and ICD-10, which were retained in this study to ensure consistent disease classification across databases. The inclusion of DO enabled the dataset to capture detailed and structured information on diseases, supporting research in medical informatics and bioinformatics.

MeSH[36] **(Medical Subject Headings)** MeSH is a comprehensive controlled vocabulary for the purpose of indexing journal articles and books in the life sciences. It is widely used in medical and biomedical research for categorizing diseases, drugs, and other entities. In this study, MeSH terms were retrieved from the MeSH XML files, focusing on records under the Diseases category. Mapping relations with UMLS, ICD-10, and other disease ontologies were preserved to ensure consistency in terminology across datasets. This facilitated the integration of disease data and enabled the dataset to support detailed disease-related analyses.

UMLS[37] **(Unified Medical Language System)** The Unified Medical Language System (UMLS), developed by the National Library of Medicine (NLM), integrates multiple biomedical terminologies into a single framework. The Disease or Syndrome category of UMLS was selected for this study, with an emphasis on "Preferred" terms defined in English. Mapping relations between UMLS, MeSH, SNOMED-CT, and ICD-10 were maintained to ensure accurate classification and cross-referencing of disease entities. The inclusion of UMLS data ensures that disease-related data can be consistently linked across multiple terminological systems, facilitating research in clinical informatics and biomedical research.

SNOMED-CT[38] **(Systematized Nomenclature of Medicine Clinical Terms)** SNOMED-CT is an international clinical terminology that is used to code the entire scope of human medical practice, including diseases, symptoms, diagnoses, and treatments. Data from the Snapshot version of SNOMED-CT was used to extract active entries from the "Disorder" category, preserving mapping relations with ICD-10. This allowed for the integration of clinical disease information with other ontologies, enhancing the utility of the dataset for both clinical and research applications.

Mondo[39] The Mondo Disease Ontology integrates multiple disease ontologies and databases, offering comprehensive cross-references to UMLS, MeSH, and other classification systems. Data from Mondo was included in this study, with a focus on preserving mapping relations between Mondo IDs, UMLS, and MeSH. This integration enabled the consistent classification of disease entities, ensuring that disease-related data from different sources could be accurately linked.

PubChem[40] PubChem is a large database of chemical molecules and their biological activities, maintained by the National Center for Biotechnology Information (NCBI). It is widely used for retrieving chemical

information related to small molecules, including drugs, metabolites, and other compounds. For this study, data from the Drug and Medication Information and Pharmacology and Biochemistry categories within the PubChem compound catalog was extracted. Key chemical descriptors, such as InChI, SMILES, InChIKey, and IUPAC names, were selected to provide detailed chemical structure information. The PubChem CID (Compound Identifier) was used as a unique identifier to facilitate consistent cross-referencing of chemical compounds across datasets.

CAS[41] **(Chemical Abstracts Service)** CAS is a division of the American Chemical Society, and its Common Chemistry database provides information on chemical substances, including their molecular structure, properties, and nomenclature. Data from CAS was accessed via the PubChem platform, and mapping relations between CAS numbers and PubChem CIDs were retained. This ensures that chemical data can be accurately linked across multiple resources. Ensuring the uniqueness of CIDs was critical to maintaining reliable cross-referencing of chemical entities, particularly for pharmacological and biochemical data.

NDC[42] **(National Drug Code)** The National Drug Code (NDC) is a unique identifier for medications in the United States, maintained by the U.S. Food and Drug Administration (FDA). It is an essential resource for drug-related data integration. Data from the NDC was selected for inclusion, focusing on the NDC code and substance names, which correspond to the UNII (Unique Ingredient Identifier) code's preferred term. By ensuring the uniqueness of each substance name, accurate integration with UNII data was facilitated, allowing for comprehensive drug-related analyses.

UNII[43] **(Unique Ingredient Identifier)** The Unique Ingredient Identifier (UNII) system, maintained by the FDA, assigns unique identifiers to chemical substances, including active ingredients in drugs. UNII data was sourced from both the PubChem website and the FDA, with mapping relations between UNII codes, PubChem CIDs, and CAS numbers being preserved. Additionally, structural descriptors such as SMILES and InChIKeys were included, providing a detailed representation of the chemical substances. This ensures that UNII data can be integrated seamlessly with other chemical and pharmacological databases.

DrugBank[44] DrugBank is a unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug-target information. Data from DrugBank was included in this study to retain mapping relations between DrugBank IDs and other chemical identifiers, such as PubChem CID, SID (Substance ID), and CAS numbers. DrugBank's extensive annotation of drug targets and mechanisms of action made it a valuable resource for cross-referencing drugs with their molecular and clinical effects, enabling more in-depth pharmacological studies.

A.2 Data Resources for Relations

Ensembl[11] Ensembl is a comprehensive genome browser and database that provides a wealth of information on gene sequences, annotations, and relations across multiple species. It supports the analysis of gene-transcript interactions by linking genes to their corresponding transcripts. Ensembl also provides transcript-protein interaction, providing detailed annotations of how transcripts give rise to protein products. The dataset is essential for understanding gene structure, function, and the consequences of gene expression.

NCBI - RefSeq[15] **(Reference Sequence Database)** RefSeq is a well-curated collection of gene, transcript, and protein sequences, offering high-quality data for gene-transcript and transcript-protein relations. It provides standardized and curated sequences that ensure consistency in gene annotations. RefSeq is crucial for researchers needing reliable reference sequences for various biological analyses, particularly in understanding the relations between transcripts and their encoded proteins.

UniProt[17] **(Universal Protein Resource)** UniProt is a leading repository of protein sequence and functional information. It plays a dual role by linking transcripts to their corresponding protein products. In addition to capturing transcript-protein interactions, UniProt also includes annotations of protein-disease relations, making it essential for understanding how protein dysfunctions can lead to disease.

BioGrid[46] BioGrid is a key resource for protein-protein interaction data, curated from both high-throughput and small-scale experimental studies. This database is essential for exploring how proteins interact within cellular networks, facilitating the study of complex biological processes such as signaling pathways, metabolic networks, and structural assemblies. BioGrid data on protein-protein interactions supports a wide range of applications, from basic research to drug discovery.

STRING[48] STRING is a database of known and predicted protein-protein interactions, integrating data from various sources such as experimental studies, computational predictions, and publicly available text collections. It is essential for understanding the functional interactions between proteins and mapping protein interaction networks. STRING helps to identify potential interactions that play critical roles in biological processes and disease states, making it a valuable tool for systems biology research.

KEGG[19] **(Kyoto Encyclopedia of Genes and Genomes)** KEGG is a comprehensive database that integrates genomic, chemical, and systemic functional information, offering valuable insights into various biological interactions. It is essential for studying protein-protein interactions, illustrating how proteins cooperate in cellular processes, as well as gene-pathway interactions, showing how genes function within specific biological pathways. Furthermore, KEGG explores drug-pathway interactions, revealing how drugs influence these pathways, and facilitates the study of pathway-gene and pathway-drug interactions, providing a clear understanding of how pathways are regulated by genes and targeted by drugs.

HPO[32] **(Human Phenotype Ontology)** HPO provides a standardized vocabulary of phenotypic abnormalities associated with human diseases. It is invaluable for connecting genes to phenotypes (gene-phenotype interaction), linking diseases to their phenotypic presentations (disease-phenotype interaction), and mapping genes to diseases (gene-disease interaction). HPO also facilitates the study of phenotype-phenotype relations, enabling researchers to compare phenotypic similarities and differences across genetic conditions.

DisGeNet[49] DisGeNet is a comprehensive platform that integrates data on gene-disease associations from multiple sources, including expert-curated databases, scientific literature, and publicly available repositories. It plays a critical role in identifying gene-disease interactions, helping to elucidate the genetic basis of various diseases. DisGeNet supports research into disease mechanisms by providing insights into the complex genetic networks that underlie disease phenotypes.

DISEASES[50] The DISEASES database provides information on protein-disease associations, integrating data from literature mining and manually curated sources. It links proteins to the diseases they are associated with, offering a detailed view of how protein dysfunctions contribute to disease phenotypes. DISEASES is especially useful for identifying molecular mechanisms underlying diseases and for exploring potential therapeutic targets.

HMDB[23] **(Human Metabolome Database)** HMDB is an extensive resource that provides detailed information on human metabolites, including drugs, drug metabolites, and endogenous small molecules. It captures a wide range of interactions, including drug-metabolome, metabolome-disease, and metabolome-protein relations. HMDB supports research in metabolomics, systems biology, and pharmacology, providing data on metabolic pathways, metabolite-protein interactions, and the role of metabolites in health and disease.

MetaNetX[51] It is a comprehensive resource developed by the SIB Swiss Institute of Bioinformatics to facilitate the standardization, integration, and analysis of genome-scale metabolic networks (GSMNs) and biochemical pathways. MetaNetX allows users to construct, modify, and analyze metabolic models through tools for flux balance analysis (FBA), reaction knockout simulations, and network comparison. By integrating data from diverse sources and providing a standardized framework, MetaNetX is a valuable tool for researchers in systems biology and bioinformatics, enabling a deeper understanding of complex metabolic processes.

DisBiome[52] DisBiome is a database that focuses on the relations between microbiomes and diseases. It captures microbiome-disease interactions, providing insights into how microbial taxa are associated with

health and disease. DisBiome supports research into the role of the human microbiome in various disease conditions, facilitating the exploration of microbial communities as potential biomarkers or therapeutic targets.

MDAD[53] **(Microbe-Drug Association Database)** MDAD is a comprehensive resource that compiles clinically and experimentally validated associations between microbes and drugs. It contains 5,055 entries, encompassing 1,388 drugs and 180 microbes, sourced from multiple drug databases and scientific publications. Each record in MDAD includes detailed annotations, such as molecular forms of drugs, links to DrugBank, microbe target information from UniProt, and original reference citations. This database serves as a valuable tool for researchers aiming to understand microbe-drug interactions, facilitating advancements in drug discovery, disease therapy, and personalized medicine.

PharmacoMicrobiomics[54] It is a field that examines the interactions between the human microbiome and drugs, focusing on how microbial communities influence drug metabolism, efficacy, and toxicity. This bidirectional relation involves microbes activating, inactivating, or transforming drugs into metabolites with altered effects, while drugs, in turn, can reshape the composition and function of the microbiome. These interactions have profound implications for personalized medicine, as variations in the microbiome can affect individual drug responses, side effects, and therapeutic outcomes. By understanding these dynamics, PharmacoMicrobiomics aims to optimize drug therapies, reduce adverse effects, and pave the way for microbiome-targeted medical interventions.

CTD (The Comparative Toxicogenomics Database)[29] CTD is a publicly available, manually curated resource that provides insights into the complex relations between chemicals, genes, and diseases, with a specific emphasis on environmental exposures. CTD integrates data on chemical-gene interactions, chemical-disease associations, and gene-disease relations, offering researchers a unique platform to explore the molecular mechanisms underlying toxicological effects and exposure-related health outcomes. By including exposure-related information, CTD helps bridge the gap between environmental science and molecular biology, enabling studies on how environmental factors influence gene function and contribute to disease etiology. This resource is particularly valuable for advancing research in toxicogenomics, precision medicine, and environmental health.

DO (Disease Ontology)[35] DO is a standardized biomedical ontology that provides a structured vocabulary and hierarchical classification for human diseases, enabling consistent annotation and integration of disease-related data across research and clinical domains. Each disease entry is assigned a unique identifier and is cross-referenced with external resources such as OMIM, ICD, SNOMED CT, and MeSH, ensuring interoperability and facilitating data harmonization. By linking diseases to their etiology, molecular mechanisms, and clinical manifestations, DO supports applications in translational medicine, computational

biology, and precision medicine. Its integration with genomic and phenotypic datasets makes it a critical tool for advancing disease research, biomarker discovery, and therapeutic development.

DrugBank[44] DrugBank is a comprehensive resource that integrates detailed information on drugs and their targets. It captures multiple types of interactions, including protein-drug, drug-drug relations. DrugBank provides data on drug mechanisms, drug interactions, and the diseases they are used to treat, making it an essential tool for pharmacological research and drug development. It also supports studies on how drugs interact with biological systems at the molecular level.

BindingDB[55] BindingDB is a public repository of measured binding affinities between proteins (mainly drug targets) and small, drug-like molecules. It supports research into protein-drug interactions by providing experimental data on the binding affinities of drugs to their target proteins. BindingDB is a valuable resource for drug discovery and pharmacology, helping researchers identify potential drug candidates and understand the molecular mechanisms of drug action.

DrugCentral[56] DrugCentral is a centralized portal for drug information, offering data on drug-protein, drug-disease, interactions. It integrates information on drug indications, targets, and mechanisms of action, supporting the study of therapeutic interventions and pharmacodynamics. DrugCentral is an important resource for researchers exploring drug repurposing, drug development, and clinical applications.

SIDER[57] **(Side Effect Resource)** SIDER provides comprehensive data on the adverse effects of drugs, linking pharmaceutical compounds to their phenotypic side effects. This resource is essential for studying drug-phenotype interactions, helping researchers understand the unintended consequences of drug use. SIDER supports pharmacovigilance efforts and aids in optimizing drug safety profiles by highlighting potential risks associated with pharmaceutical compounds.

Table S1. Download Links and Access Control for Entity Databases

Database		Access	Download Link
Ensembl	Gene	public access	BioMart API
	Transcript	public access	BioMart API
	Protein	public access	BioMart API
OMIM	Gene	public access	https://omim.org/static/omim/data/mim2gene.txt
HGNC	Gene	public access	https://www.genenames.org/cgi-bin/download/custom?col=gd_hgnc_id&col=gd_app_sym&col=gd_app_name&col=gd_pub_eg_id&col=gd_pub_ensembl_id&status=Approved&hgnc_dbtag=on&order_by=gd_hgnc_id&format=text&submit=submit
NCBI	Gene	public access	https://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2ensembl.gz https://ftp.ncbi.nlm.nih.gov/gene/DATA/gene_info.gz
	Microbiota	public access	https://ftp.ncbi.nih.gov/pub/taxonomy/taxdmp.zip
RefSeq	Gene	public access	https://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2refseq.gz

	Transcript	public access	https://ftp.ncbi.nlm.nih.gov/refseq/MANE/MANE_human/current/MANE.GRCh38.v1.3.summary.txt.gz
	Protein	public access	https://ftp.ncbi.nlm.nih.gov/gene/DATA/gene_refseq_uniprotkb_collab.gz https://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2ensembl.gz
RNACentral	Transcript	public access	https://ftp.ebi.ac.uk/pub/databases/RNAcentral/current_release/id_mapping/database_mappings/ensembl.tsv
UniProt	Protein	public access	API
Reactome	Pathway	public access	https://reactome.org/download/current/ReactomePathways.txt
KEGG	Pathway	public access	Fetching data via R and Python
WikiPathways	Pathway	public access	Fetching data via Python
Pathway Ontology	Pathway	public access	https://download.rgd.mcw.edu/ontology/pathway/pathway.obo
ComPath	Pathway	public access	https://compath.scai.fraunhofer.de/export_mappings
HMDB	Metabolite	public access	https://hmdb.ca/downloads
ChEBI	Metabolite	public access	https://www.ebi.ac.uk/chebi/chebiOntology.do?chebiId=77746
	Drug	public access	https://ftp.ebi.ac.uk/pub/databases/chebi/Flat_file_tab_delimited/chebiId_inchi_3star.tsv
			https://ftp.ebi.ac.uk/pub/databases/chebi/Flat_file_tab_delimited/database_accession_3star.tsv
SILVA	Microbiota	public access	https://www.arb-silva.de/fileadmin/silva_databases/current/Exports/taxonomy/ncbi/taxmap_embl-ebi_ena_lsu_ref_138.2.txt.gz
			https://www.arb-silva.de/fileadmin/silva_databases/current/Exports/taxonomy/ncbi/taxmap_embl-ebi_ena_ssu_ref_138.2.txt.gz
Greengenes	Microbiota	public access	https://ftp.ebi.ac.uk/pub/databases/RNAcentral/current_release/id_mapping/database_mappings/greengenes.tsv
RDP	Microbiota	public access	https://ftp.ebi.ac.uk/pub/databases/RNAcentral/current_release/id_mapping/database_mappings/rdp.tsv
GTDB	Microbiota	public access	https://data.ace.uq.edu.au/public/gtdb/data/releases/latest/ar53_metadata.tsv.gz https://data.ace.uq.edu.au/public/gtdb/data/releases/latest/bac120_metadata.tsv.gz
CTD	Exposure	public access	https://ctdbase.org/reports/CTD_chemicals.csv.gz
ToxCast	Exposure	public access	https://clowder.edap-cluster.com/files/6114f600e4b0856fdc65865c
ChemIDplus	Exposure	public access	Fetching data via Python
HPO	Phenotype	public access	https://hpo.jax.org/data/ontology
UMLS	Phenotype	Registration required	https://download.nlm.nih.gov/umls/kss/2024AA/umls-2024AA-full.zip?_gl=1*14ig82q*_ga*MTA5NTI1Nzc2My4xNzEwOTU5NjM5*_ga_7147EPK006*MTcyMzU3NDM0NC41My4xLjE3MjM1NzUyNzYuMC4wLjA.*_ga_P1FPTH9PL4*MTcyMzU3NDM0NC41My4xLjE3MjM1NzUyNzYuMC4wLjA
	Disease	Registration required	
ICD10 / ICD11	Disease	public access	https://icdcdn.who.int/static/releasefiles/2024-01/SimpleTabulation-ICD-11-MMS-en.zip https://icdcdn.who.int/static/releasefiles/2024-01/mapping.zip
Disease Ontology	Disease	public access	https://github.com/DiseaseOntology/HumanDiseaseOntology/blob/main/DOreports/allXREFinDO.tsv
MeSH	Disease	public access	https://nlmpubs.nlm.nih.gov/projects/mesh/MESH_FILES/xmlmesh/desc2024.xml
SNOMED-CT	Disease	Registration required	https://download.nlm.nih.gov/umls/kss/IHTSDO2024/IHTSDO20240801/SnomedCT_InternationalRF2_PRODUCTION_20240801T120000Z.zip?_gl=1*xret7k*_ga*MTA5NTI1Nzc2My4xNzEwOTU5NjM5*_ga_7147EPK006*MTcyMzU4ODA3OC41NC4xLjE3MjM1ODgyNDYuMC4wLjA.*_ga_P1FPTH9PL4*MTcyMzU4ODA3OS41NC4xLjE3MjM1ODgyNDYuMC4wLjA
Mondo	Disease	public access	https://github.com/monarch-initiative/mondo/blob/master/reports/xrefs.tsv https://github.com/monarch-initiative/mondo/releases/latest/download/mondo.obo
PubChem	Drug	public access	https://pubchem.ncbi.nlm.nih.gov/#query=RqHgEbsv3pPpudyqXtiVibNCzyJqQo7z9NaVv-

			Hh77v3rs&alias=PubChem%20Compound%20TOC:%20Drug%20and%20Medication%20Information
			https://pubchem.ncbi.nlm.nih.gov/#query=fDvaK_Z8k8Ck7hv3mY9S3nQVH3UjzJXS7_eOnvTmnJ_0_6A&alias=PubChem%20Compound%20TOC:%20Pharmacology%20and%20Biochemistry
CAS	Drug	public access	Fetching data via Python
NDC	Drug	public access	https://www.accessdata.fda.gov/cder/ndctext.zip
UNII	Drug	public access	https://precision.fda.gov/uniisearch/archive/latest/UNII_Data.zip Fetching data via Python
DrugBank	Drug	public access	https://go.drugbank.com/releases/5-1-12/downloads/all-drug-links

Table S2. Download Links and Access Control for Relation Databases

Database		Access	Download Link
Ensembl	Gene-Transcript	public access	BioMart API
	Transcript-Protein		BioMart API
RefSeq	Gene-Transcript	public access	https://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/RefSeqGene/LRG_RefSeqGene
	Transcript-Protein		https://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/RefSeqGene/LRG_RefSeqGene
UniProt	Transcript-Protein	public access	API
	Protein-Disease		https://rest.uniprot.org/uniprotkb/stream?compressed=true&fields=accession%2Ccc_disease&format=tsv&query=%28*%29+AND+%28model_organism%3A9606%29
BioGrid	Protein-Protein	public access	https://downloads.thebiogrid.org/File/BioGRID/Release-Archive/BIOGRID-4.4.237/BIOGRID-ALL-4.4.237.mitab.zip
STRING	Protein-Protein	public access	https://stringdb-downloads.org/download/protein.links.full.v12.0/9606.protein.links.full.v12.0.txt.gz
KEGG	Protein-Protein	public access	Fetching data via R
	Protein-Pathway		
	Drug-Pathway		
	Pathway- Protein		
	Pathway-Drug		
HPO	Protein-Phenotype	public access	https://hpo.jax.org/data/annotations
	Protein-Disease		https://hpo.jax.org/data/annotations
	Phenotype-Phenotype		https://hpo.jax.org/data/ontology
	Phenotype-Disease		https://hpo.jax.org/data/annotations
	Disease-Phenotype		https://hpo.jax.org/data/annotations
DisGeNet	Protein-Disease	Registration required	API
DISEASES	Protein-Disease	Registration required	https://download.jensenlab.org/human_disease_benchmark.tsv
MetaNetX	Metabolite- Metabolite	public access	https://www.metanetx.org/cgi-bin/mnxget/mnxref/chem_xref.tsv
			https://www.metanetx.org/cgi-bin/mnxget/mnxref/chem_isom.tsv
DisBiome	Microbiota-Disease	public access	https://disbiome.ugent.be/export
MDAD	Microbiota-Drug	public access	https://github.com/Sun-Yazhou/MDAD/blob/master/MDAD.zip
	Drug-Microbiota		
PharmacoMicrobiomics	Microbiota-Drug	public access	http://pharmacomicrobiomics.com/view/relation/
	Drug-Microbiota		
HMDB	Metabolite-Protein	public access	https://hmdb.ca/downloads
	Metabolite-Disease		
	Drug-Metabolite		

CTD	Exposure-Gene	public access	https://ctdbase.org/reports/CTD_chem_gene_ixns.csv.gz
	Exposure-Pathway		https://ctdbase.org/reports/CTD_chem_pathways_enriched.csv.gz
	Exposure-Disease		https://ctdbase.org/reports/CTD_chemicals_diseases.csv.gz
Disease Ontology	Disease-Disease	public access	https://raw.githubusercontent.com/DiseaseOntology/HumanDiseaseOntology/refs/heads/main/src/ontology/HumanDO.obo
DrugBank	Drug-Protein	Registration required	https://go.drugbank.com/releases/5-1-12/downloads/target-all-polypeptide-ids
	Drug-Drug		https://go.drugbank.com/releases/5-1-12/downloads/all-full-database
BindingDB	Drug-Protein	public access	https://www.bindingdb.org/bind/downloads/BindingDB_All_202409_tsv.zip
DrugCentral	Drug-Protein	public access	https://drugcentral.org/ActiveDownload
	Drug-Disease		
SIDER	Drug-Phenotype	public access	http://sideeffects.embl.de/media/download/meddra_all_se.tsv.gz

Section B. Details of Entity and Relation Integration

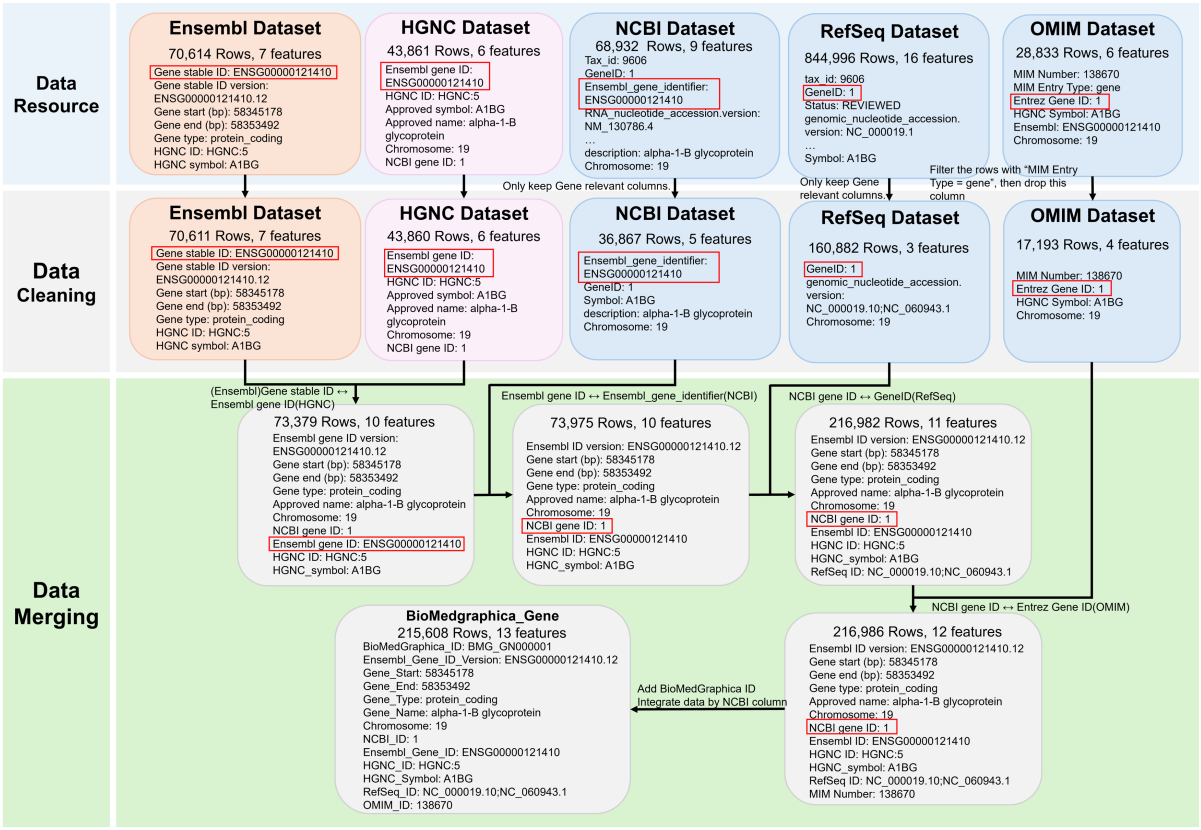


Figure S1. Details of Gene Entity Merging Process

Figure S1 provides a detailed overview of the integration process for BioMedGraphica Gene, using A1BG as an example. The "Data Resource" section depicts the original datasets sourced from various databases for gene entity integration. The "Data Cleaning" section presents the cleaned data format prepared for integration. Columns highlighted in red boxes indicate the key matching fields used during the merging process. In the "Data Merging" section, the gray boxes showcase the data format at each step of database integration. The overall gene entity integration employs an outer join approach: Ensembl and HGNC databases are merged first, followed by integration with the NCBI Gene database. Subsequently, RefSeq and OMIM are incorporated sequentially. The final unification is based on the NCBI Gene ID, ensuring that all entries with the same ID are consolidated.

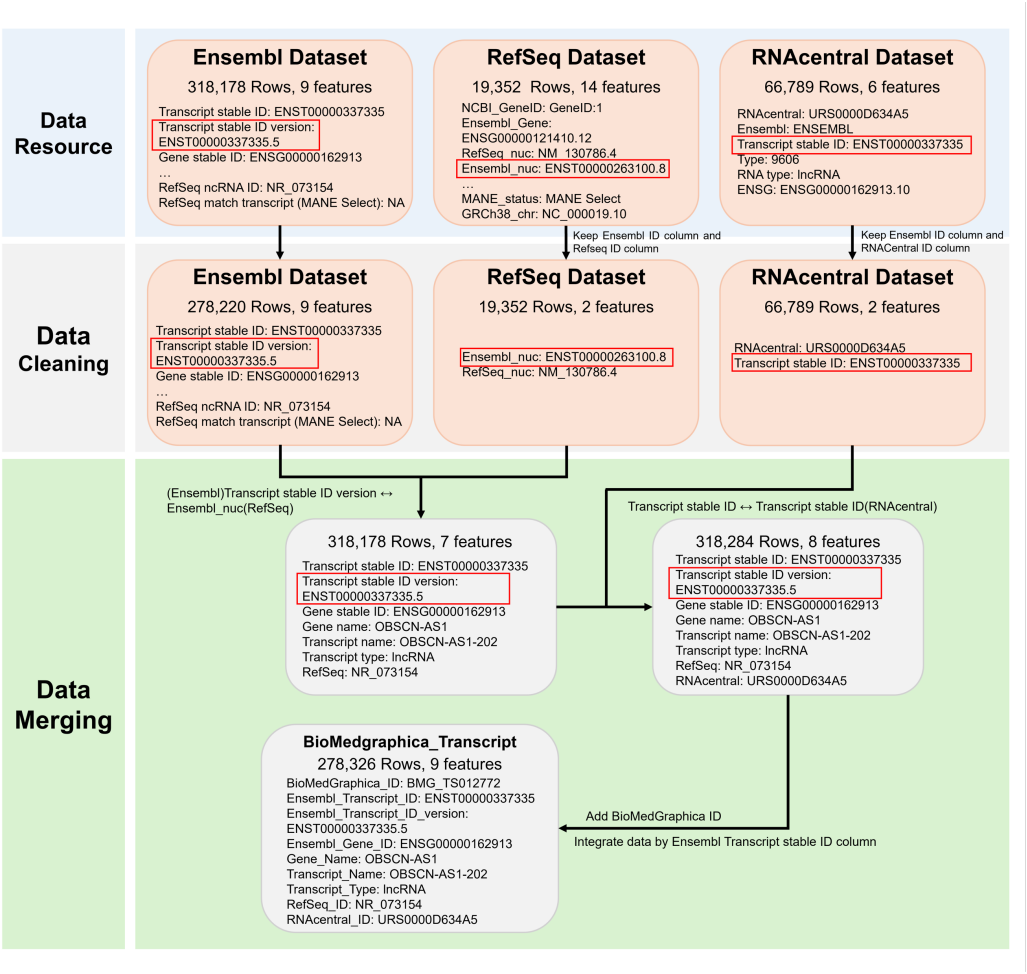


Figure S2. Details of Transcript Entity Merging Process

A detailed depiction of the integration process for BioMedGraphica transcript has been provided **in Figure S2**, using ENST00000337335.5 as an example. The "Data Resource" section shows the raw data sourced from databases used in transcript entity integration. Since the original RefSeq dataset lacked a corresponding RefSeq ID for ENST00000337335.5, an alternative transcript was selected as a supplementary example. The "Data Cleaning" section presents the cleaned data format prepared for integration. Columns highlighted in red boxes indicate the key matching fields used during the merging process. In the "Data Merging" section, gray boxes illustrate the data format after each step of database integration. The integration process for transcript entities employs an outer join approach: first, the Ensembl and RefSeq databases are merged, followed by integration with the RNAcentral database. The Ensembl stable ID serves as the primary unit for final data unification, consolidating all entries with the same Ensembl stable ID.

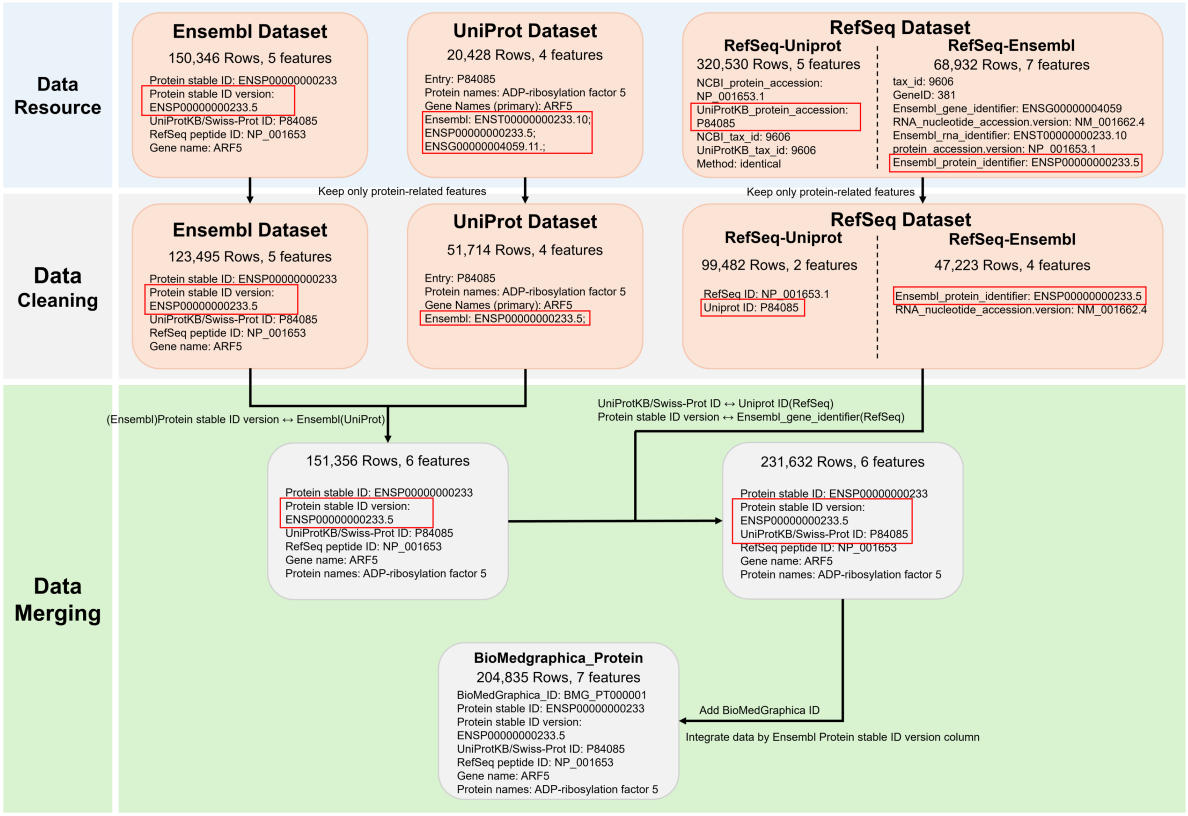


Figure S3. Details of Protein Entity Merging Process

Figure S3 illustrates the integration process for BioMedGraphica Protein, using ENSP00000000233.5 as a representative example. The "Data Resource" section outlines the raw datasets obtained from various databases utilized in protein entity integration. The "Data Cleaning" section highlights the standardized format of the data after preparation for integration. Key matching columns, marked in red boxes, were used to align data across sources. The "Data Merging" section visualizes the transformation of data formats through successive integration steps, represented by gray boxes. The integration process employs an outer join methodology, starting with the merging of Ensembl and UniProt databases. This combined dataset is then integrated with RefSeq. The final step uses the Ensembl stable ID version as the primary key to unify entries, ensuring that all records associated with the same Ensembl stable ID version are consolidated.

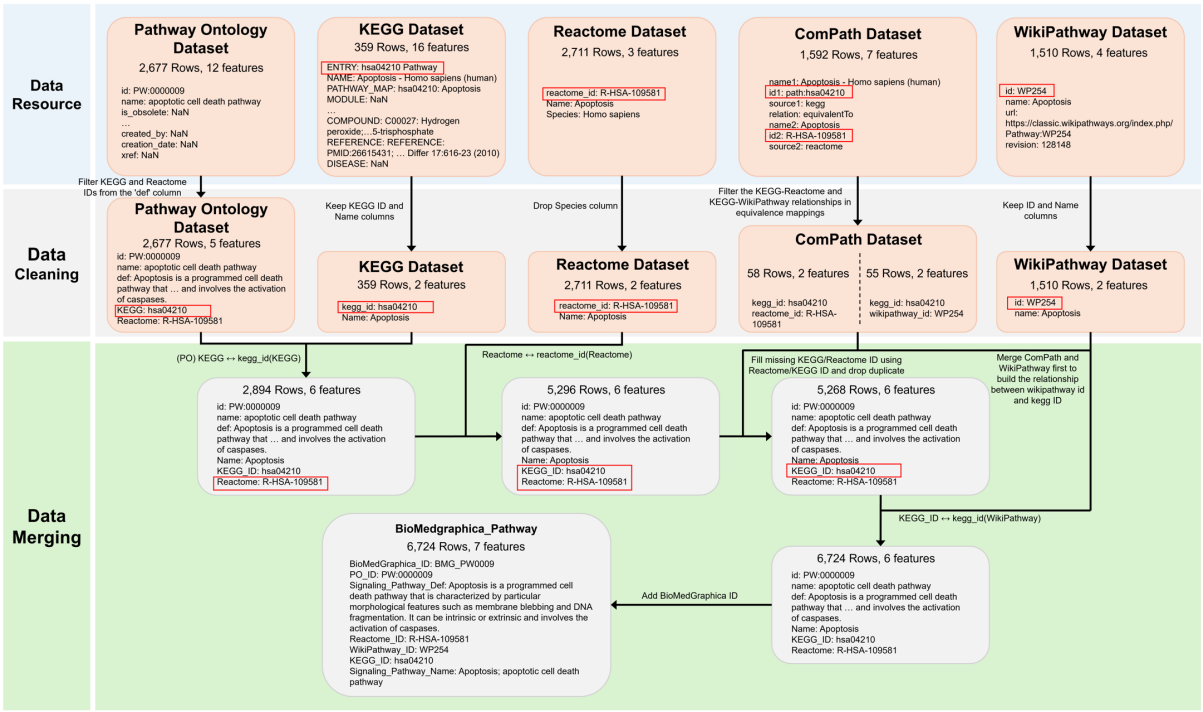


Figure S4. Details of Pathway Entity Merging Process

Figure S4 provides a detailed illustration of the integration process for BioMedGraphica Pathway, using PW:0000009 as an example. The "Data Resource" section represents the raw data from the databases used in the integration of the pathway entity. The "Data Cleaning" section displays the format of the cleaned data prepared for integration. The data highlighted in the red boxes indicates the key matching columns used for merging. In the "Data Merging" section, the gray boxes show the format of the data after each step of database integration. The pathway entity integration process follows an outer join method. First, the Pathway Ontology and KEGG databases are merged, followed by the integration of Reactome data into the combined dataset, and subsequently ComPath and WikiPathway are integrated in sequence.

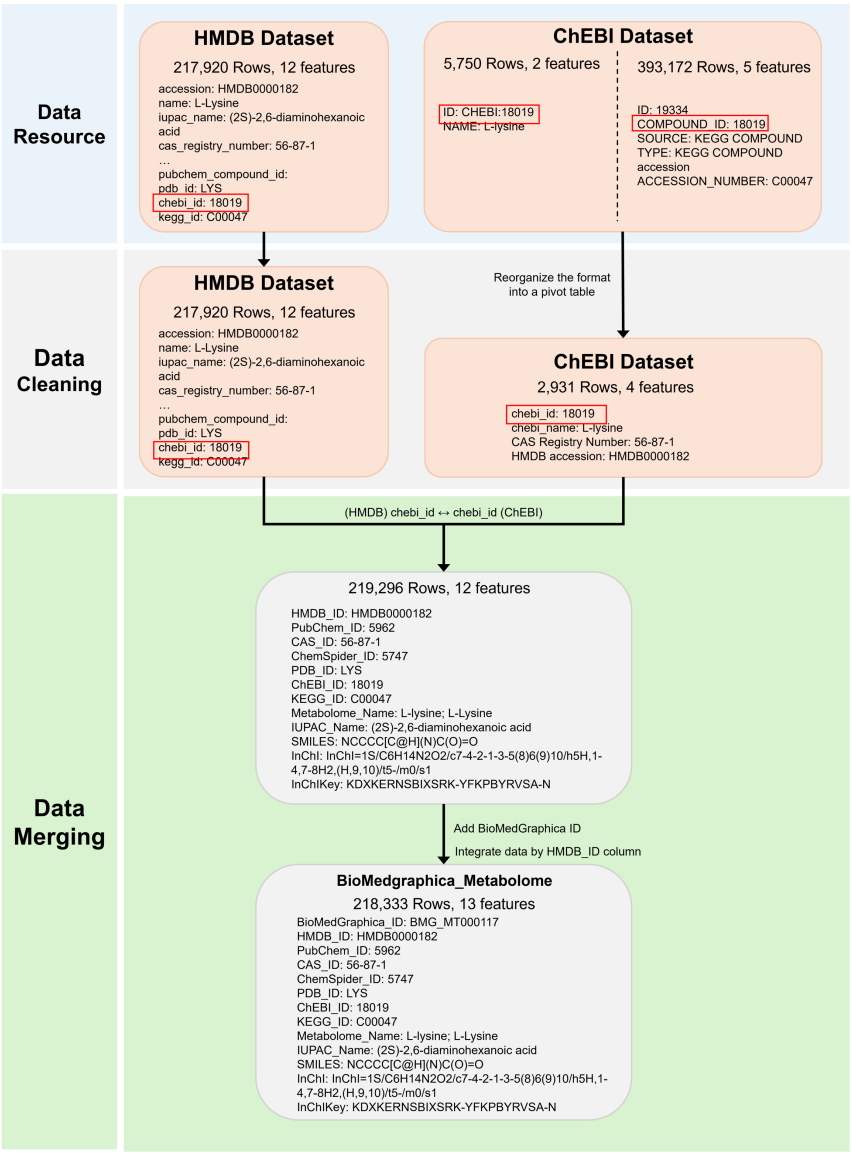


Figure S5. Details of Metabolite Entity Merging Process

Figure S5 provides a detailed illustration of the integration process for BioMedGraphica Metabolite, using ChEBI: 18019 as an example. The "Data Resource" section represents the raw data from the databases used in the integration of the metabolite entity. The "Data Cleaning" section displays the format of the cleaned data prepared for integration. The data highlighted in the red boxes indicates the key matching columns used for merging. In the "Data Merging" section, the gray boxes show the data format after each step of database integration. The metabolite entity integration process follows an outer join method, merging data from the HMDB and ChEBI databases. Finally, the HMDB ID is used as the minimal unit for data unification, consolidating all entries with the same HMDB ID.

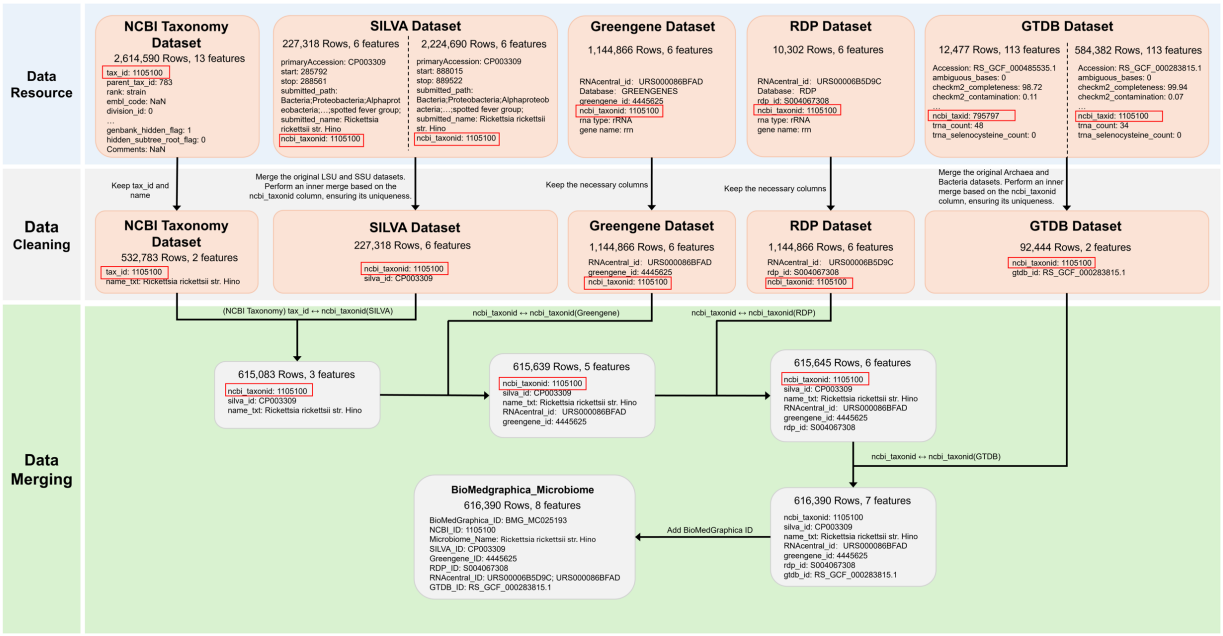


Figure S6. Details of Microbiota Entity Merging Process

Figure S6 provides a detailed illustration of the integration process for BioMedGraphica Microbiota, using NCBI Taxon ID: 1105100 as an example. The "Data Resource" section represents the raw data from the databases used in the integration of the microbiota entity. The "Data Cleaning" section shows the format of the cleaned data prepared for integration. The data highlighted in the red boxes indicates the key matching columns used for merging. In the "Data Merging" section, the gray boxes display the data format after each step of database integration. The microbiota entity integration process follows an outer join strategy, first merging data from the NCBI Taxonomy and SILVA databases, followed by integration with Greengenes, RDP, and GTDB in sequence. Finally, the NCBI Taxon ID is used as the primary unit for data unification, consolidating all entries with the same NCBI Taxon ID.

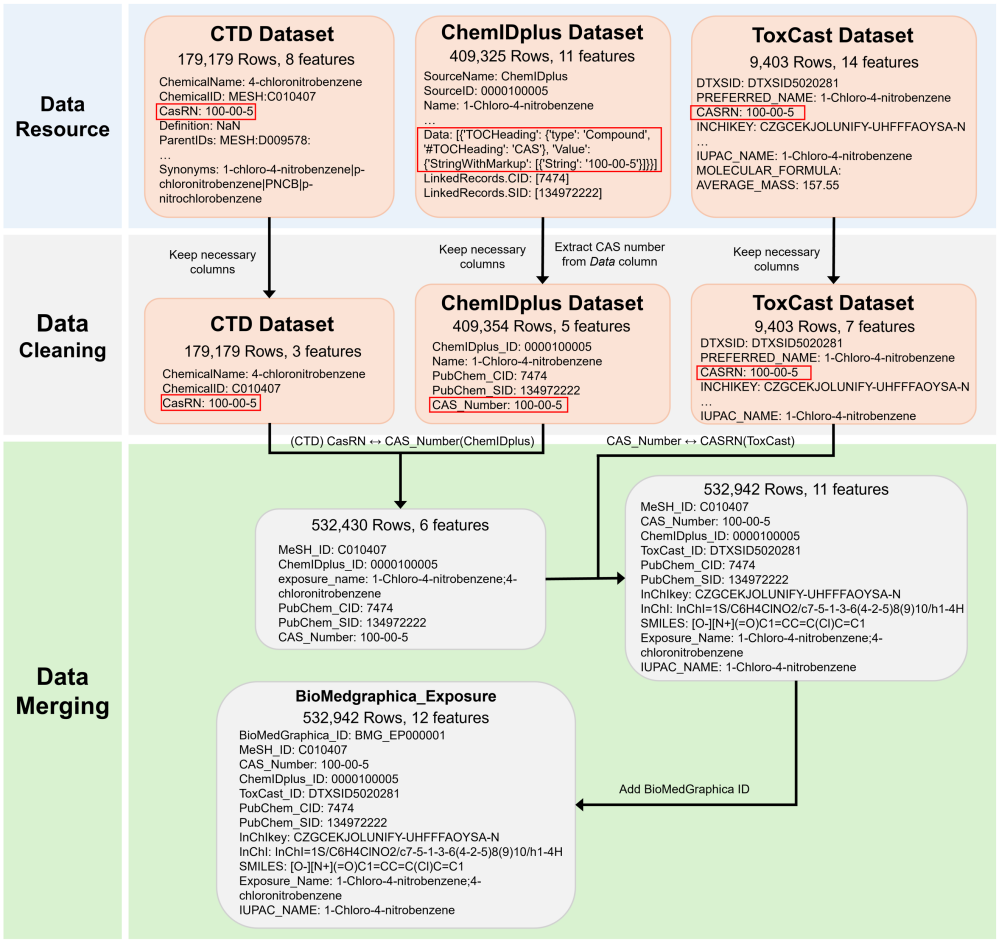


Figure S7. Details of Exposure Entity Merging Process

Figure S7 uses CAS number 100-00-05 as an example to illustrate the integration process for BioMedGraphica Exposure. The "Data Resource" section displays the raw data from databases used for the exposure entity. The "Data Cleaning" section shows the cleaned data format prepared for integration. The data highlighted in the red boxes indicates the key matching columns used for merging. In the "Data Merging" section, the gray boxes display the data format after each step of database integration. The exposure entity integration process follows an outer join strategy, first merging data from the CTD and ChemIDplus databases, followed by integration with ToxCast.

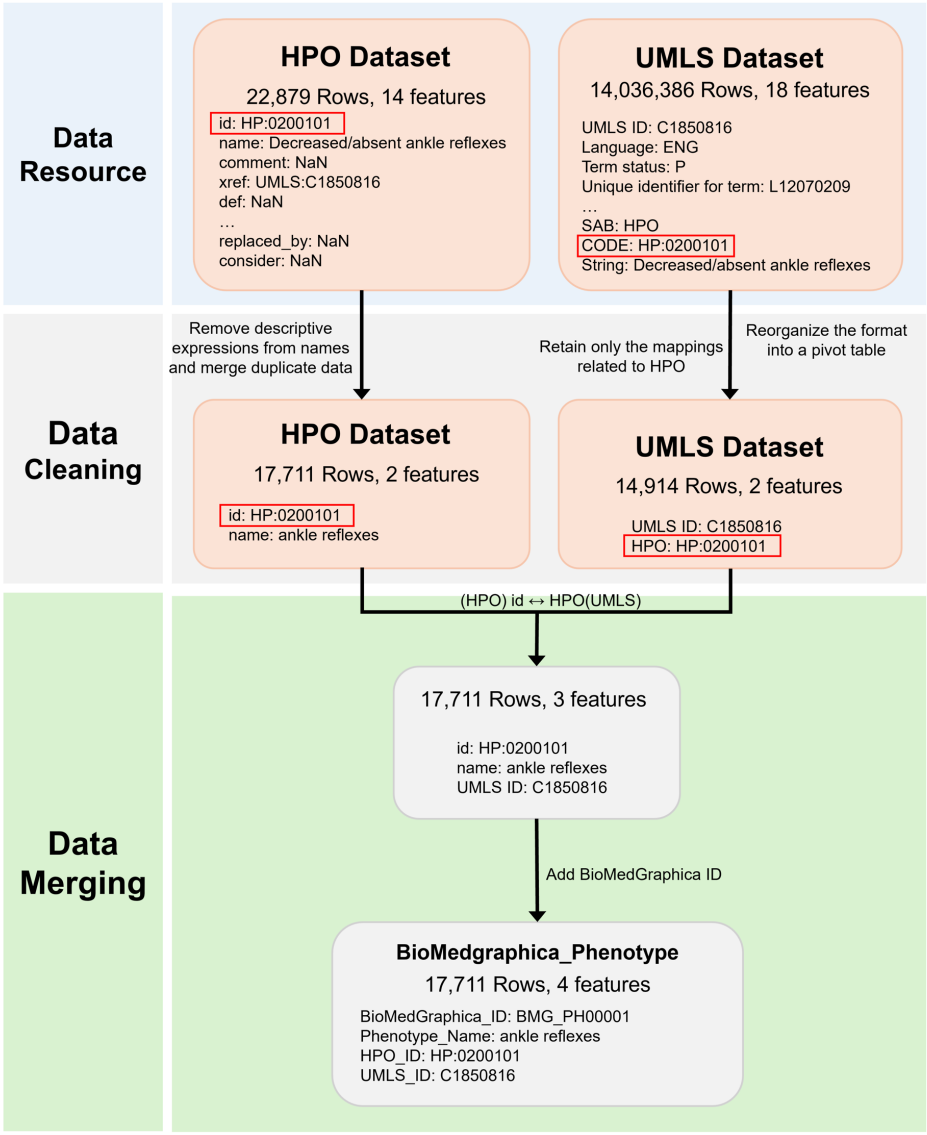


Figure S8. Details of Phenotype Entity Merging Process

Figure S8 provides a detailed illustration of the integration process for BioMedGraphica Phenotype, using HP: 0200101 as an example. The "Data Resource" section represents the raw data from the databases used in the integration of the phenotype entity. The "Data Cleaning" section displays the format of the cleaned data prepared for integration. For HPO, descriptive terms in the original names were removed. The data highlighted in the red boxes indicates the key matching columns used for merging. In the "Data Merging" section, the gray boxes show the format of the data after each step of database integration. The phenotype entity integration process follows an outer join approach, merging data from the HPO and UMLS databases.

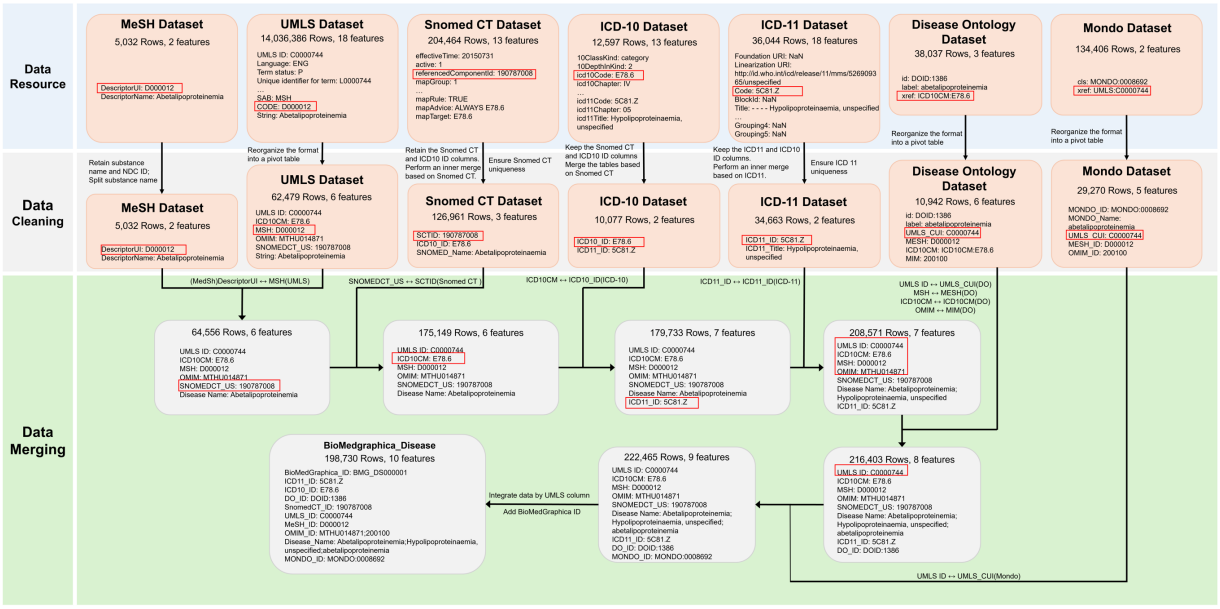


Figure S9. Details of Disease Entity Merging Process

Figure S9 provides a detailed illustration of the integration process for BioMedGraphica Disease, using C0000744 as an example. The "Data Resource" section represents the raw data from the databases used in the integration of the disease entity. The "Data Cleaning" section shows the format of the cleaned data prepared for integration. The data highlighted in the red boxes indicates the key matching columns used for merging. In the "Data Merging" section, the gray boxes display the data format after each step of database integration. The disease entity integration process follows an outer join methodology. First, the MeSH and UMLS databases are merged, followed by the integration of SNOMED CT data with the combined dataset. Subsequently, ICD-10, ICD-11, Disease Ontology, and Mondo are integrated in sequence. Finally, the UMLS ID serves as the primary identifier for data unification, consolidating all entries with the same UMLS ID.

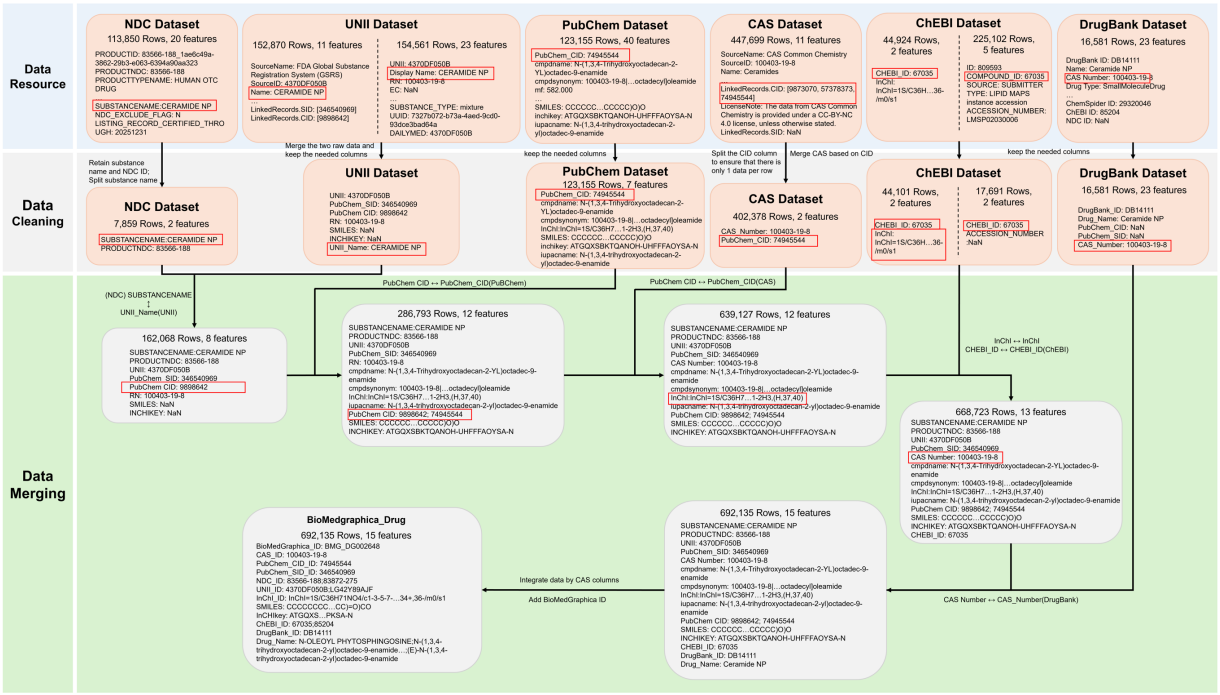


Figure S10. Details of Drug Entity Merging Process

Figure S10 provides a detailed illustration of the integration process for BioMedGraphica Drug, using 100403-19-8 as an example. The "Data Resource" section represents the raw data from the databases used in the integration of the drug entity. The "Data Cleaning" section displays the format of the cleaned data prepared for integration. The data highlighted in the red boxes indicate the key matching columns used for merging. In the "Data Merging" section, the gray boxes show the format of the data after each database integration step. The drug entity integration process follows an outer join approach. First, the NDC and UNII databases are merged, followed by integrating the combined data with the PubChem database, and subsequently with CAS, ChEBI, and DrugBank. Finally, the CAS number is used as the primary identifier for data unification, ensuring all entries with the same CAS number are consolidated.

References

1. Hulsen T, Jamuar SS, Moody AR, et al. From big data to precision medicine. *Front Med (Lausanne)*. 2019;6(MAR). doi:10.3389/fmed.2019.00034

2. Kendall TJ, Jimenez-Ramos M, Turner F, et al. An integrated gene-to-outcome multimodal database for metabolic dysfunction-associated steatotic liver disease. *Nat Med*. 2023;29(11):2939-2953. doi:10.1038/s41591-023-02602-2

3. Fernández-Torras A, Duran-Frigola M, Bertoni M, Locatelli M, Aloy P. Integrating and formatting biomedical data as pre-calculated knowledge graph embeddings in the Bioteque. *Nat Commun*. 2022;13(1). doi:10.1038/s41467-022-33026-0

4. Königs C, Friedrichs M, Dietrich T. The heterogeneous pharmacological medical biochemical network PharMeBINet. *Sci Data*. 2022;9(1). doi:10.1038/s41597-022-01510-3

5. Ma T, Zhang A. Integrate multi-omics data with biological interaction networks using Multi-view Factorization AutoEncoder (MAE). *BMC Genomics*. 2019;20. doi:10.1186/s12864-019-6285-x

6. Cavalleri E, Cabri A, Soto-Gomez M, et al. An ontology-based knowledge graph for representing interactions involving RNA molecules. *Sci Data*. 2024;11(1). doi:10.1038/s41597-024-03673-7

7. MacNamara A, Nakic N, Amin Al Olama A, et al. Network and pathway expansion of genetic disease associations identifies successful drug targets. *Sci Rep*. 2020;10(1). doi:10.1038/s41598-020-77847-9

8. Zhang H, Cao D, Chen Z, et al. mosGraphGen: a novel tool to generate multi-omic signaling graphs to facilitate integrative and interpretable graph AI model development. doi:10.1101/2024.05.15.594360

9. Ma S, Zeng AG, Haibe-Kains B, Goldenberg A, Dick JE, Wang B. *Integrate Any Omics: Towards Genome-Wide Data Integration for Patient Stratification*.; 2024.

10. Lee J, Yoon W, Kim S, et al. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234-1240. doi:10.1093/bioinformatics/btz682

11. Howe KL, Achuthan P, Allen J, et al. Ensembl 2021. *Nucleic Acids Res*. 2021;49(D1):D884-D891. doi:10.1093/nar/gkaa942

12. Amberger J, Bocchini CA, Scott AF, Hamosh A. McKusick's Online Mendelian Inheritance in Man (OMIM®). *Nucleic Acids Res*. 2009;37(SUPPL. 1). doi:10.1093/nar/gkn665

13. Povey S, Lovering R, Bruford E, Wright M, Lush M, Wain H. The HUGO Gene Nomenclature Committee (HGNC). *Hum Genet*. 2001;109(6):678-680. doi:10.1007/s00439-001-0615-0

14. Schoch CL, Ciufo S, Domrachev M, et al. NCBI Taxonomy: A comprehensive update on curation, resources and tools. *Database*. 2020;2020. doi:10.1093/database/baaa062

15. O'Leary NA, Wright MW, Brister JR, et al. Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016;44(D1):D733-D745. doi:10.1093/nar/gkv1189

16. Sweeney BA, Petrov AI, Burkov B, et al. RNAcentral: A hub of information for non-coding RNA sequences. *Nucleic Acids Res*. 2019;47(D1):D221-D229. doi:10.1093/nar/gky1034

17. Wu CH, Apweiler R, Bairoch A, et al. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*. 2006;34(Database issue). doi:10.1093/nar/gkj161

18. Fabregat A, Jupe S, Matthews L, et al. The reactome pathway knowledgebase. *Nucleic Acids Res*. 2018;46(D1):D649-D655.

19. Kanehisa MGS, Goto S. KEGG: kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 2000;28:27-30. doi:10.1093/nar/28.1.27

20. Kelder T, Van Iersel MP, Hanspers K, et al. WikiPathways: building research communities on biological pathways. *Nucleic Acids Res*. 2012;40(D1):D1301-D1307.

21. Petri V, Jayaraman P, Tutaj M, et al. The pathway ontology–updates and applications. *J Biomed Semantics*. 2014;5:1-12.

22. Domingo-Fernández D, Hoyt CT, Bobis-Álvarez C, Marín-Llaó J, Hofmann-Apitius M. ComPath: an ecosystem for exploring, analyzing, and curating mappings across pathway databases. *NPJ Syst Biol Appl*. 2018;4(1):43.

23. Wishart DS, Guo A, Oler E, et al. HMDB 5.0: the human metabolome database for 2022. *Nucleic Acids Res*. 2022;50(D1):D622-D631.

24. Degtyarenko K, De matos P, Ennis M, et al. ChEBI: A database and ontology for chemical entities of biological interest. *Nucleic Acids Res*. 2008;36(SUPPL. 1). doi:10.1093/nar/gkm791

25. Quast C, Pruesse E, Yilmaz P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*. 2012;41(D1):D590-D596.

26. DeSantis TZ, Hugenholtz P, Larsen N, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol*. 2006;72(7):5069-5072.

27. Cole JR, Wang Q, Fish JA, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res*. 2014;42(D1):D633-D642.

28. Parks DH, Chuvochina M, Waite DW, et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol*. 2018;36(10):996-1004.

29. Davis AP, Grondin CJ, Johnson RJ, et al. Comparative Toxicogenomics Database (CTD): Update 2021. *Nucleic Acids Res*. 2021;49(D1):D1138-D1143. doi:10.1093/nar/gkaa891

30. Dix DJ, Houck KA, Martin MT, Richard AM, Setzer RW, Kavlock RJ. The ToxCast program for prioritizing toxicity testing of environmental chemicals. *Toxicological sciences*. 2007;95(1):5-12.

31. Tomasulo P. ChemIDplus-super source for chemical and drug information. *Med Ref Serv Q*. 2002;21(1):53-59.

32. Köhler S, Gargano M, Matentzoglu N, et al. The human phenotype ontology in 2021. *Nucleic Acids Res*. 2021;49(D1):D1207-D1217.

33. Organization WH. *International Statistical Classification of Diseases and Related Health Problems: Alphabetical Index*. Vol 3. World Health Organization; 2004.

34. Organization WH. International classification of diseases for mortality and morbidity statistics (11th Revision). Published online 2018.

35. Schriml LM, Arze C, Nadendla S, et al. Disease ontology: A backbone for disease semantic integration. *Nucleic Acids Res*. 2012;40(D1). doi:10.1093/nar/gkr972

36. Lipscomb CE. Medical subject headings (MeSH). *Bull Med Libr Assoc*. 2000;88(3):265.

37. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res*. 2004;32(suppl_1):D267-D270.

38. Donnelly K. SNOMED-CT: The advanced terminology and coding system for eHealth. *Stud Health Technol Inform*. 2006;121:279.

39. Vasilevsky N, Essaid S, Matentzoglu N, et al. Mondo Disease Ontology: harmonizing disease concepts across the world. In: *CEUR Workshop Proceedings, CEUR-WS*. Vol 2807. ; 2020.

40. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH. PubChem: A public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res*. 2009;37(SUPPL. 2). doi:10.1093/nar/gkp456

41. Peryea T, Southall N, Miller M, et al. Global Substance Registration System: Consistent scientific descriptions for substances related to health. *Nucleic Acids Res*. 2021;49(D1):D1179-D1185. doi:10.1093/nar/gkaa962

42. Tribble DA. The National Drug Code explained. *American Journal of Health-System Pharmacy*. Published online 2024:zxae274.

43. Weisgerber DW. Chemical abstracts service chemical registry system: history, scope, and impacts. *Journal of the American Society for Information Science*. 1997;48(4):349-360.

44. Knox C, Wilson M, Klinger CM, et al. DrugBank 6.0: the DrugBank Knowledgebase for 2024. *Nucleic Acids Res*. 2024;52(D1):D1265-D1275. doi:10.1093/nar/gkad976

45. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res*. 2006;34(suppl_1):D535-D539.

46. Oughtred R, Stark C, Breitkreutz BJ, et al. The BioGRID interaction database: 2019 update. *Nucleic Acids Res*. 2019;47(D1):D529-D541.

47. Szklarczyk D, Franceschini A, Wyder S, et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*. 2015;43(D1):D447-D452.

48. Szklarczyk D, Gable AL, Lyon D, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res*. 2019;47(D1):D607-D613.

49. Piñero J, Bravo À, Queralt-Rosinach N, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res*. Published online 2016:gkw943.

50. Pletscher-Frankild S, Pallejà A, Tsafou K, Binder JX, Jensen LJ. DISEASES: Text mining and data integration of disease-gene associations. *Methods*. 2015;74:83-89. doi:10.1016/j.ymeth.2014.11.020

51. Moretti S, Tran VDT, Mehl F, Ibberson M, Pagni M. MetaNetX/MNXref: unified namespace for metabolites and biochemical reactions in the context of metabolic models. *Nucleic Acids Res*. 2021;49(D1):D570-D574.

52. Janssens Y, Nielandt J, Bronselaer A, et al. Disbiome database: linking the microbiome to disease. *BMC Microbiol*. 2018;18:1-6.

53. Sun YZ, Zhang DH, Cai SB, Ming Z, Li JQ, Chen X. MDAD: a special resource for microbe-drug associations. *Front Cell Infect Microbiol*. 2018;8:424.

54. Doestzada M, Vila AV, Zhernakova A, et al. Pharmacomicrobiomics: a novel route towards personalized medicine? *Protein Cell*. 2018;9(5):432-445.

55. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res*. 2016;44(D1):D1045-D1053.

56. Ursu O, Holmes J, Knockel J, et al. DrugCentral: online drug compendium. *Nucleic Acids Res*. Published online 2016:gkw993.

57. Kuhn M, Campillos M, Letunic I, Jensen LJ, Bork P. A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol*. 2010;6(1):343.