Knowledge-based artificial intelligence system for drug prioritization

Yinchun Su[1#], Jiashuo Wu[2#], Xilong Zhao[2#], Yue Hao[1], Ziyi Wang[2], Yongbao Zhang[2], Yujie Tang[2], Bingyue Pan[2], Guangyou Wang[1*], Qingfei Kong[1,3*], Junwei Han[2*]

[1] Department of Neurobiology, Harbin Medical University, Heilongjiang Provincial Key Laboratory of Neurobiology, Harbin, 150081, China

[2] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, 150081, PR China

[3] The Heilongjiang Provincial Joint Laboratory of Basic Medicine and Multiple Organ System Diseases (International Cooperation), Harbin, 150081, China

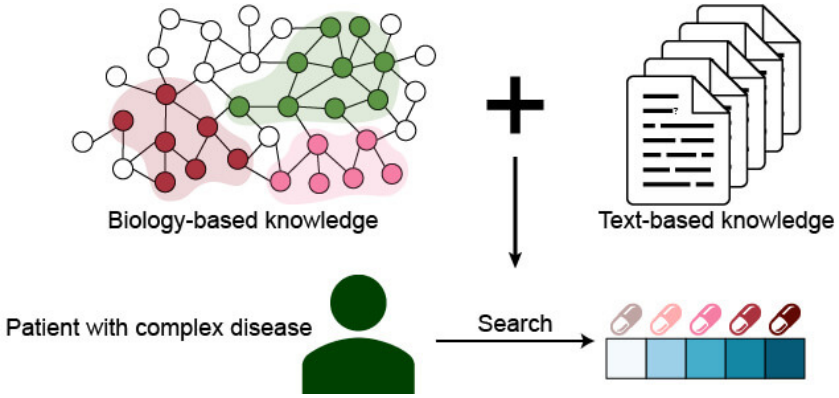
These authors contributed equally: Yinchun Su, Jiashuo Wu, Xilong Zhao.

Corresponding authors: Junwei Han (hanjunwei@ems.hrbmu.edu.cn; ORCID: 0000-0002-3276-0819), Qingfei Kong (kqfangel@hrbmu.edu.cn; ORCID: 0000-0002-6014-7632), or Guangyou Wang (wangguangyou@ems.hrbmu.edu.cn; ORCID: 0000-0002-9490-0637).

Abstract

In silico drug prioritization may be a promising and time-saving strategy to identify potential drugs, standing as a faster and more cost-effective approach than *de novo* approaches. In recent years, artificial intelligence has greatly evolved the drug development process. Here, we present a novel computational framework for drug prioritization, *labyrinth*, designed to simulate human knowledge retrieval and inference to identify potential drug candidates for each disease. With the integration of up-to-date clinical trials, literature co-occurrences, drug–target interactions, and disease similarities, our framework achieves over 90% predictive accuracy across clinical trial phases and strong alignment with clinical practice in TCGA cohorts. We have demonstrated effectiveness across 20 different disease categories with robust ROC-AUC metrics and the balance between predictive accuracy and model interpretability. We further demonstrate its effectiveness at both the population and the individual levels. This study not only demonstrates the capacity for its drug prioritization but underscores the importance of aligning computational models with intuitive human reasoning. We have wrapped the core function into an R package named *labyrinth*, which is freely available on GitHub under the GPL-v2 license (https://github.com/hanjunwei-lab/labyrinth).

Keywords: artificial intelligence, drug repositioning, cognitive simulation, language models.



1 Introduction

Drug prioritization, or drug repositioning, has emerged as a promising strategy in drug development, standing as a faster and more cost-effective approach than *de novo* approaches[1]. By identifying novel indications for existing drugs, computational drug repositioning relies on large-scale biological data to yield a more robust and reliable result[2], thus reducing the costs and enabling the potential for large-scale drug screening. Knowledge graphs have been used extensively in the recent few years in drug discovery for rare diseases as they integrate diverse biological and medical data and offer a structured approach to researchers in life sciences[3].

Several studies have successfully employed knowledge graphs in this area. Knowledge graphs are a large-scale, graph-structured databases integrating various types of data to represent entities, their relationships, and their semantic attributes. For instance, PrimeKG is a multimodal knowledge graph for precision medicine analyses to integrate 20 high-quality resources to describe 17,080 diseases with 4,050,249 relationships representing ten major biological scales[4]. Similarly, a comprehensive drug knowledge graph for the knowledge-driven drug repurposing method showed a promising knowledge graph[5]. Furthermore, SMR constructs a high-quality heterogeneous graph, integrating electronic medical records and medical knowledge graphs to avoid adverse drug reactions[6], also highlighting the importance of embedding techniques in enhancing graph-based predictions. However, these methods often lack the flexibility and context-aware reasoning that characterize human expertise in this domain.

Simulations of human cognitive abilities have shown great potential in learning and prediction. Human memory, especially long-term memory (LTM), serves as an inspiration for organizing diverse data types. It includes explicit memory and implicit memory, which enables humans to retrieve relevant information, identify patterns, and make optimal decisions, abilities that are usually lacking in computational drug discovery approaches. While our understanding of human memory is limited, the spreading activation network has proven to be a model with high explanatory power in explaining phenomena in human LTM[7]. In spreading activation networks, concepts and memories are represented as nodes, with their associated elements connected by edges[8]. It can be schematically represented with shorter or more substantial edges indicating a closer relationship between two nodes, typically resulting in a higher rate of recall[9].

Humans excel at considering problems from various perspectives, displaying greater flexibility in knowledge association compared to computers. Our work extends these perspectives by simulating cognitive processes in drug prioritization, uniquely integrating diverse medical knowledge sources into a human-like reasoning model. Here, we developed *labyrinth*, a computational framework that simulates human knowledge retrieval, specifically designed for drug prioritization in clinical settings. It addresses the main objective by identifying and prioritizing existing drugs for potential prioritization through the simulation of human cognitive processes, while aligning these predictions with real-world clinical outcomes. By integrating multiple sources of prior medical knowledge including clinical trials, literature cooccurrences, drug–target interactions, and disease similarities, *labyrinth* identifies potential drug candidates using a human-like knowledge retrieval approach. Our validation of *labyrinth* through several case studies illuminates its potential to offer unique insights by integrating biologically meaningful information with text-based data into a comprehensive model for addressing human diseases. We have wrapped the core function into an R package named

labyrinth, which is freely available on GitHub under the GPL-v2 license (https://github.com/hanjunwei-lab/labyrinth).

2 Materials and methods

2.1 Main components of *labyrinth*

We introduce *labyrinth*, a novel computational framework that simulates human cognition and decision-making processes to prioritize drugs for complex disease treatment. **Figure 1** depicts the simplified schema of *labyrinth*. It integrates two major knowledge sources: text-based information from medical corpora and biological knowledge from function interaction networks.

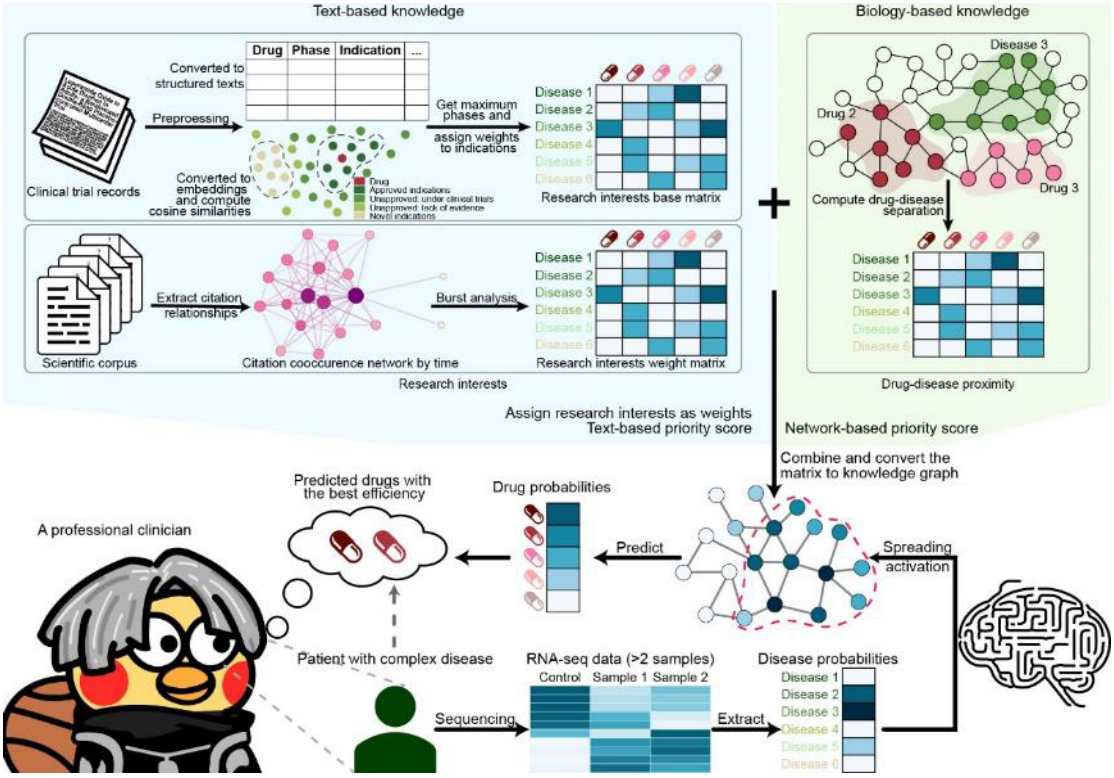


Figure 1. A simple schema of the *labyrinth*. We trained *labyrinth* through the integration of dual knowledge sources: text-based and biology-based. It calculates drug-disease proximity by analyzing the separation within a biologically meaningful functional interactome network. Next, clinical trial information is transformed into structured information to assign weights to drug-disease pairs, while literature from the Scientific Index (SCI) collection from the Web of Science is processed to extract drug–disease relationships, which are then represented in an *n*-dimensional vector space. Cosine similarities between drugs and diseases generate a matrix enriched with citation network analysis to capture the temporal influence of research papers, using citation burst ranges as weights. These processes culminate in a matrix that reflects research interests, which is combined with a biological knowledge matrix through probabilistic computation to simulate human knowledge retrieval for drug prioritization with the best efficiency. This simulation aims to mimic a professional clinician's decision-making process by mapping patients to potential treatments based on disease relevance and treatment efficacy, ultimately identifying candidate drugs with the highest potential for the patient's benefit.

The text-based component innovatively incorporates large volumes of textual data that are distilled into a knowledge network suitable for efficient retrieval. Clinical trial information is analyzed to extract the maximum phases reached for specific drug–disease pairs, which are then assigned weights accordingly. For drug–disease pairs without clinical phase information, *labyrinth* employs the word2vec algorithm to generate *n*-dimensional embeddings and calculate cosine similarities as proximity scores. Additionally, citation burst analysis is performed on the medical literature to quantify research interests for each drug–disease pair, resulting in a comprehensive matrix that reflects research interests. All of the parameters are default in this paper unless explicit explanation is given.

In the biological component, *labyrinth* evaluates the network proximity between drug target modules and disease gene modules within the functional interactome network. This proximity metric captures the biological relevance between drugs and diseases.

The final knowledge network integrates the textual and biological matrices through probabilistic computations. This simulates the human process of storing relevant knowledge in long-term memory for decision making. The reasoning process inside *labyrinth* is inspired by human cognitive principles with enhanced interpretability. *Labyrinth* then applies the random walk with restarts (RwR) algorithm on the integrated knowledge network to prioritize drugs, mimicking how clinicians abstract patients to several potential diseases and their severities when making treatment decisions. After making inferences, drugs with higher scores are considered more promising candidates for treatment.

2.2 Data collection and processing

Labyrinth leveraged multiple authoritative data sources to construct a comprehensive knowledge base for *labyrinth*. Drug information, including nomenclature, targets, and indications, was extracted from DrugBank, CTD, and ChEMBL databases (**Figure 2A**) so that drugs with identical chemical formulas were considered identical compounds and deduplicated information was obtained (**Figure 2B**). Clinical trial data was sourced from the Cochrane Library, while co-occurrence patterns in published literature were mined from the Web of Science corpus (**Figure 2C**). Functional relationships between proteins were compiled by integrating seven interaction databases into a unified network.

Text data from over 10 million publications underwent thorough preprocessing, including stop word removal using the approach proposed by Gerlach et al[10] and term vectorization via Skip-gram models[11]. In parallel, structured drug–disease relationships were quantified based on clinical trial phases, citation analysis, and network proximity between gene sets [**Figures 1**, **2C**(II), (III))]. These heterogeneous data streams were probabilistically combined into an integrated knowledge graph.

2.2.1 Normalizing drug names

To acquire drug records from the Web of Science (WOS), it is essential to search drugs by their names. Drugs that are chemically equivalent can possess multiple synonyms, even though they are identified by one generic name. **Figure 2B** demonstrates the various names and identifiers for the three specific drugs. Our study incorporates three main sources of drug information: DrugBank, ChEMBL, and CTD to extract synonyms for the same drug, with detailed counts depicted in **Figure 2A**.

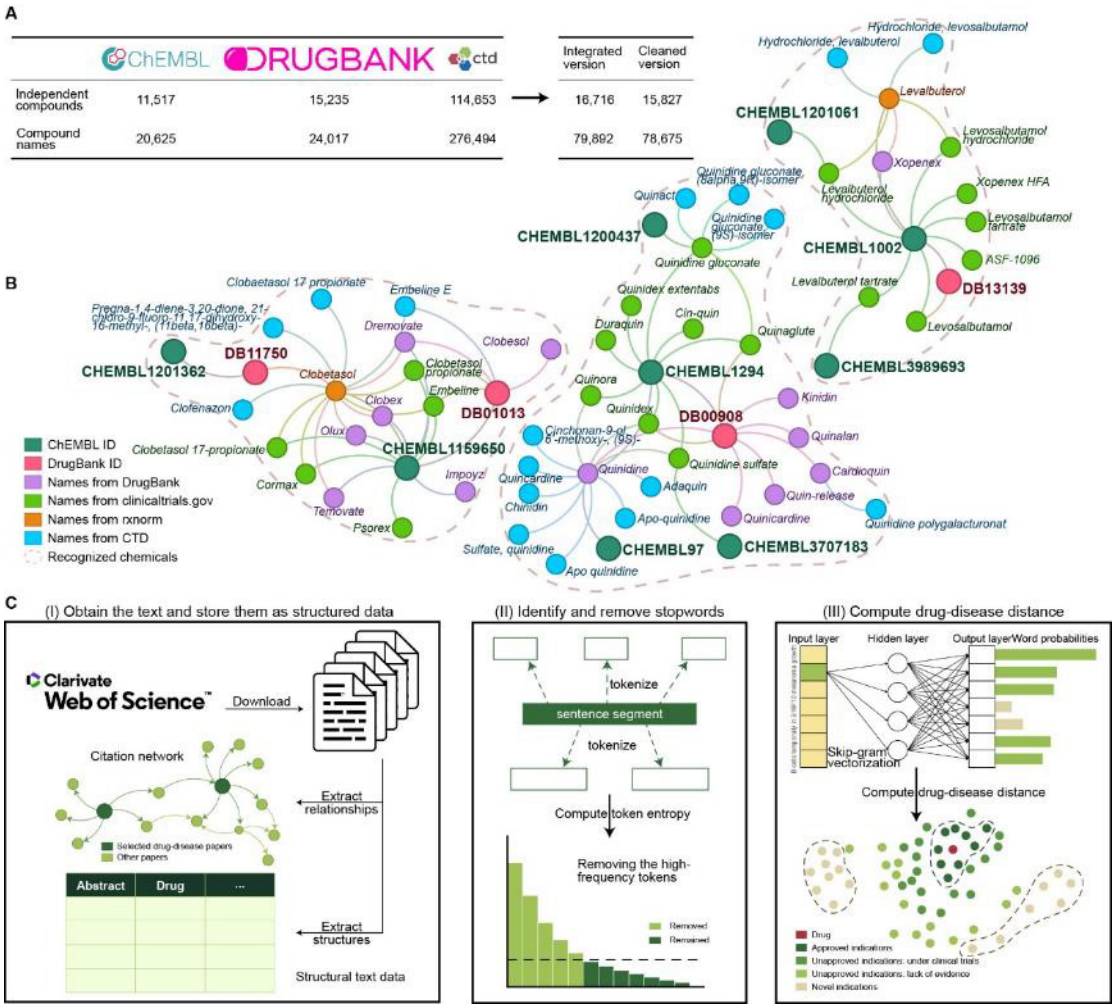


Figure 2. Data preparation process of *labyrinth*. (**A**) Integration of drug information from three distinct databases. (**B**) Network linking of drugs by name or chemical formula; we consider the drugs in each disjoint subfigure to have the same chemical structure. (**C**) Literature text preprocessing workflow. First, texts from the Web of Science are structured and analyzed for paper relationships. Second, text tokenization identifies and removes high-frequency stop words. Finally, we adopted Skip-gram to vectorize the tokens into embeddings, with drug–disease distances calculated via cosine similarity.

DrugBank is a comprehensive network-based repository containing extensive molecular information on drugs, including details on targets, labels, chemical properties, indications, and clinical trials[12]. In this study, we extracted 15,235 independent drugs and 24,017 drug names from DrugBank (version 5.1.10, January 2023).

ChEMBL is a large-scale bioactivity database designed to support drug discovery by providing extensive open data on the bioactivity of drug-like compounds[13,14]. It facilitates the answer to key scientific questions, including those related to health. We retrieved information on 11,907 drug-target interactions and 51,582 unique drug indications from ChEMBL.

The Comparative Toxicogenomic Database (CTD) is a robust and expansive public resource aimed at elucidating the impacts of environmental chemical exposures on human health[15]. Documenting over 30.5 million toxicogenomic relationships among chemicals, genes, and diseases, CTD is indispensable for research in toxicology, environmental health, biology, and pharmacogenomics. Our study incorporated 114,653 independent drug entries from the CTD.

We constructed an undirected network featuring drug names and assumed connections between names or identifiers of the same drug, while ensuring no links existed between different drugs. As **Figure 2B** shows, the network was initiated using DrugBank IDs as source nodes, which were then connected to drug names, brand names, and external links of ChEMBL ID found in DrugBank. This process was repeated in UniChem (updated on April 10, 2023), a comprehensive, non-redundant cross-reference database linking ChEMBL IDs to other external databases[16]. Due to the absence of names in the local dump in ChEMBL, we linked to DailyMed, ClinicalTrials, Rxnorm, and Expression Atlas to gather drug names. Subsequently, drug names were downloaded from the CTD and integrated into the original network.

Finally, this network construction resulted in the identification of 16,716 unique drugs and 79,892 names (**Figure 2A**). Prior to batch searching on WOS, we scrutinized the query terms and excluded compounds not typically distinguished from common foods, with removal criteria outlined in **Table S4**. This led to the removal of 889 compounds, leaving 15,827 unique drugs and 78,675 synonyms (**Figure 2A**). In this network, each chemical entity is assigned to a unique cluster ID without considering dosages and administration methods (as shown in **Figure 2B**). Different formulations of the same active chemical compound are grouped under one ID. For multicomponent drugs, we treated them as distinct entities with cross-references to individual components.

2.2.2 Obtaining the text and construct drug-wise citation networks

Figure 2C depicts the comprehensive data preprocessing flow utilized in this research. Initially, we retrieved the papers from WOS to establish a vast citation network, capturing the relationships among these papers with the details stored as structured textual data.

Bibliographical and citation data were meticulously searched for and acquired from the WOS database, which integrates multiple databases to furnish access to reference and citation information spanning various academic disciplines, with coverage extending from 1900 to the present. By integrating the varied compound names of drugs, we formulated query statements through the concatenation of the different names of a certain drug using “OR”. Subsequently, we manually downloaded all the records related to 15,827 drugs from the WOS Core Collection (akin to the Scientific Index, SCI) before April 2023, chosen for its inclusion of both high-quality papers and citation data[17]. Next, we then constructed a citation network culminating in a total of 10,535 drugs with at least one record in WOS core collection, culminating in a total of 10,535 citation networks.

2.2.3 Identifying stop words using information entropy method

Stop words are commonly used words in any language, aimed at omitting uninformative words and phrases to conserve storage spaces and enhance search efficiency. Essentially, the exclusion of stop words generally does not adversely affect the outcomes. Stop words typically include terms like “the”, “a”, “an”, and “and” that are frequent, noncontributory usage words in text.

There exists no standardized approach for identifying stop words. Currently, the standard strategy involves employing a manually curated list of words considered to be uninformative[18]. While several widely recognized stop word lists exist, their applicability is limited due to the omission of domain-specific terminology. For instance, words like “abstract”, “keyword”, “method”, and “acknowledgement” may be rarely used in everyday language but are prevalent

across academic papers, indicating the need for a more nuanced approach to stop word selection.

Traditional mainstream methods for identifying stop words, such as word frequency in documents and term frequency and inverse document frequency (TF-IDF), are less reliable across studies, thus making them less reliable for identifying stop words uniformly. In our research, we extracted stop words by randomly sampling 1% of the documents, segmenting these documents into word tokens, and then applying a reliable and state-of-the-art technique proposed by Gerlach et al[10]. Finally, we defined stop words as those with an absolute information content less than a threshold of $I^* < 0.2$.

2.2.4 Acquiring high-quality human diseases and clinical information using Cochrane Library

The Cochrane Library is a collection of high-quality, independent medical and healthcare evidence, including systematic reviews, actionable clinical answers, and reported controlled clinical trials with either randomized or quasi-randomized[19]. By utilizing its search API, we acquired synonyms for all human diseases and information on clinical trials from the Cochrane Library. In this study, we retrieved an average of 2.33 synonyms across 2,333 diseases with MeSH ID.

2.2.5 Analyzing drug-disease distance using Skip-gram

After removing the stop words with $I^* < 0.2$, we then employed the Skip-gram algorithm to vectorize the drug and disease mentioned in the medical corpus. Skip-gram is a semisupervised machine learning technique designed to identify contextually relevant words for a given target word by representing them as *n*-dimensional vectors, known as word embeddings[11]. The vector representation process begins with comprehensive text processing of our medical corpus, where both drug and disease terms are tokenized and vectorized together. The Skip-gram model learns contextual relationships by analyzing how terms co-occur and relate to each other within the literature. Upon completion of the model training on the corpus associated with each drug, these 300-dimension word embeddings successfully captured the semantic meanings of all included texts.

2.3 Model construction and relationship quantification

2.3.1 Compute clinical status indicator as drug-indication relationships

The first part of our model focuses on the drug-indication relationships. We sourced these relationships from the ChEMBL database[13], which catalogs indications for drugs approved worldwide by authorities such as the FDA, WHO, EMA, and BNF, along with clinical candidate drugs undergoing clinical trials evidenced by USAN, INN, or ClinicalTrials.gov. The value of the max phase attribute reflects the furthest stage reached in clinical development for a particular drug, aligned with clinical trial status: approved (4), phase 3 (3), phase 2 (2), phase 1 (1), and preliminary clinical investigation (0.5). We use the max phase incremented by one as an indicator of clinical status.

However, the ChEMBL database lacks comprehensive coverage of all current indications, particularly missing preclinical data on drugs prior to Phase 1 clinical trials. To address this gap

in our study, instead of merely substituting missing values with zeroes, we employed cosine similarities between specific drugs and diseases to fill these missings. To calculate similarities between drugs and diseases, we employ a unified vector space representation obtained from the Skip-gram model[20], ranging from -1 (exact opposite) to 1 (identical), with intermediate values indicating varying levels of similarity or dissimilarity.

Our approach processes the entire medical corpus simultaneously, ensuring all terms are embedded within the same *n*-dimensional semantic space. Since all vectors share the same dimensionality and are trained within the same semantic context, the application of cosine similarity is mathematically valid. This relationship measure is computed using the Euclidean dot product formula for two vectors, v_1 and v_2, expressed as:

$$\text{cosine similarity} = \cos\theta = \frac{v_1 \cdot v_2}{\|v_1\|\|v_2\|}. \quad (1)$$

The distribution of these cosine similarities across different clinical stages is shown in **Figure S3A**. We put our focus on clinical trial records, since they provide standardized, structured information about drug–disease relationships with precise terminology and validated therapeutic associations. Instead, all of the scientific corpus is huge and sometimes contradictory discussions.

2.3.2 Link drug indication and clinical trial information

Using the Cochrane Library, we extracted all clinical trial information and linked them with relevant publications, resulting in 1,588,863 entries. We then associated each clinical trial record with the corresponding disease by performing keyword-based matching. This process enabled us to make clinical trial IDs with the associated literature and categorize the literature into groups, distinguishing between those with and those without clinical trials.

2.3.3 Assessing the proximity between disease genes and drug targets in the network

We also downloaded drug target information from DrugBank and ChEMBL. After removing duplicated items, we employed the network separation metric s_{TD} to validate the extent of overlap between a potential drug target set (module T) and disease gene sets (module D) in a biologically meaningful network such as a protein–protein interaction network. Let $G(V, E)$ represent the protein–protein interaction network; for any protein set, $A, B \in V$, s_{AB} represents the shortest path length between these two protein sets. The network separation metric compares the mean shortest distances between two modules and is defined as:

$$s_{TD} = d - \frac{w_T + w_D}{2}. \quad (2)$$

where d is the between-module distance between T and D, and either w_T or w_D is the within-module distance within each module A and B. All the indexes (d, w_T, w_D) are calculated by the mean shortest distance of s_{AB}. Accordingly, the separation metric $s_{TD} < 0$ indicates network overlap, whereas $s_{TD} > 0$ indicates nonoverlap[21]. The distribution of all the separation metrics is shown in **Figure S4A**.

2.3.4 Compute the importance of indications over time spans

Citation networks capture broader research impact and knowledge evolution. Sigma (Σ) is a widely used metric in identifying pivotal literature within a specific domain and gives insights into the evolution of scientific thought over time[22]. It combines the structural significance of nodes (as measured by betweenness centrality) with their temporal prominence (denoted by citation burst) properties of a node[23] into a singular metric computed as $\Sigma = (\text{centrality} + 1)^{\text{burstiness}}$, where higher values signify works of greater influential potential.

Betweenness centrality quantifies the degree to which a node serves as a bridge along the shortest paths between other nodes, reflecting its strategic position within the network[24]. A node with high betweenness typically links diverse clusters, facilitating the flow of information[23]. In this context, the betweenness centrality gauges the prominence of papers within the cocitation network related to a drug.

Burstiness in bibliometrics measures the frequency and intensity of citation spikes for a node in the cocitation network, indicating periods of heightened interest or activity[25]. Generally, it serves as a tool to outline and foresee research trajectories and thematic shifts within a specific field[26]. The burstiness index for individual papers can be determined using Joh Kleinberg's algorithm[27].

Our approach extends beyond the immediate influence of papers by focusing on their historical significance. Consequently, we diverge from CiteSpace methodology by aggregating the sum of burst durations identified by Kleinberg's algorithm for all papers and then normalized against a base of 19,500 to estimate burst range coverage. The distribution of the computed Sigma is shown in **Figure S4B**.

2.3.5 Construct the priority score by integrating proximity and a prior information

Having determined the Sigma for each piece of literature and the proximity distance for each disease–drug pair, we devised a priority score for these pairs. The distribution of this priority score is illustrated in **Figure S4C**,**D**.

We calculated two Sigma values for each pair, the average Sigma value for the literature with a clinical trial and the average Sigma value for the literature without any clinical trial. We then computed these two Sigma values to derive the disease–drug Sigma score. This score was subsequently adjusted using Yeo-Johnson normalization[28] and computed the robust *z* scores[29]. Finally, the priority score for each disease–drug pair is determined using the formula $q = \sqrt{(p_\Sigma + 1) \cdot (p_s + 1)} - 1$, where p_Σ and p_s represent the cumulative probabilities of the Sigma and separation metric, respectively. These metrics gauge the likelihood of encountering values less than or equal to each element within the data set. Overall, a higher priority score suggests a stronger potential for significant treatment effects.

2.4 Technical details in making predictions

2.4.1 Assign disease weights in RNA-seq data

In earlier steps, we trained *labyrinth* on a large medical corpus, intentionally omitting specific gene details. So, we can use *labyrinth* by inputting diseases and associated weights. We need to identify both the gene perturbations caused by the diseases, as well as the specific diseases, in order to provide treatment information from the sequencing data.

Identifying disease-related genes can be achieved through various methods, including gene coexpression modules and differential expression analysis. A prevalent method involves the analysis of gene coexpression networks to determine disease-associated modules[30]. However, this technique lacks a definitive cutoff value for determining significant values. Alternatively, differential expression analysis offers a robust means of identifying genes with notable changes in expression between different biological states, employing a significance threshold (i.e., $p < 0.05$ or $p_{\text{adj}} < 0.05$).

In our study, we employed the DESeq2 method for the differential expression analysis of count data derived from high-throughput sequencing assays[31]. DESeq2 calculates fold changes and dispersions in gene expression across varying experimental setups, leveraging generalized linear models alongside empirical Bayes shrinkage. Following the application of DESeq2 to our expression matrix, we processed the Wald statistics generated by DESeq2 through a random walk algorithm within the network. Subsequently, the identification of diseases was accomplished via Gene Set Enrichment Analysis (GSEA) employing the DisGeNET database resources.

2.4.2 Identify disease-related genes in individuals

In contrast to bulk RNA-seq data collected across multiple patients, the challenge of limited sample sizes for individual patients is significant. To address this issue, we have implemented robust principal component analysis (RPCA) as an alternative to DESeq2 for scenarios involving small-sample sizes[32]. RPCA effectively decomposes gene expression data into a low-rank matrix A, representing nondifferentially expressed genes and a sparse perturbation matrix S. This approach has proven to be highly accurate and biologically relevant in the analysis of small-sized samples.

2.4.3 Random walk with restarts (RWR)

In this study, we incorporate random walk with restart to mimic knowledge retrieval. It abstracts the real world for intelligence systems to enable them to solve complex tasks and reason about the world[33]. In simulating memory retrieval, the random walk technique mimics human semantic cognition[34,35], consistent with the process of human memory retrieval. In theory, the RWR extends the classic random walk model by adding a restart mechanism, which allows the walker to return to the starting node with a certain probability γ at each step. It operates by repeatedly moving from a current node to neighboring nodes in a graph with a transition probability of $1-\gamma$, or returning to the source node with a restart probability γ. This process iterates until the visiting probability distribution p converges, satisfying the value $p^{t+1} = (1-\gamma)Mp^t + \gamma p^0$. In the converged p^{t+1}, M is the column-normalized adjacency matrix with the network, $p^t = (p_1^t, p_2^t, \dots, p_n^t)'$ is the visiting probability of each node at time step t, and $p^0 = (p_1^0, p_2^0, \dots, p_n^0)'$ represents the initial probability distribution of nodes. The converged p provides a measure of each node's importance or similarity relative to the starting node in the certain network M.

2.4.4 Estimation of the target responsiveness using medical records

Estimating drug target responsiveness poses a greater challenge than assessing single drug effects, particularly when relying on incomplete medical records. To evaluate the responsiveness of drug targets within the TCGA database, we first classified outcomes as either

a response (incorporating both partial response and completed response) or no response. Subsequently, we directly associated drugs with their respective targets, quantifying the instances of response and no response for each. Lastly, we employed exact binomial tests rather than relying on the response-to-no-response ratio to minimize the impact of random fluctuations inherent in small samples.

2.4.5 Type 1 error rates

We evaluated the Type 1 error rates for identifying disease-related genes. We established a null matrix comprising 400 elements as a null distribution, arranged in 20 rows and 20 columns. The first 10 columns were designated as representing diseases, and the remaining 10 columns were treated as controls. After that, analyses involved conducting Welch *t* test, Student *t* test, and RPCA with different variants, with the two-sided *p* values being recorded for each. This procedure was repeated 5,000 times. At predetermined alpha levels of 0.10 (**Figure S5A**), 0.05 (**Figure S5B**), 0.01 (**Figure S5C**), and 0.001 (**Figure S5D**), our findings indicated that four out of six tests maintained the test-wise alpha without inflation. Additionally, these methods were validated against experiment-wise alpha levels of 0.05 (**Figure S6A**,**C**) and 0.01 (**Figure S6B**,**D**), both with (**Figure S6C**,**D**) and without the application of the Benjamini–Hochberg correction (**Figure S6A**,**B**). The comparative analysis demonstrated that RPCA effectively manages both experiment-wise and test-wise error rates, especially when integrated with sum squares controls.

2.4.6 Calculation of ROC-AUC for model evaluation

We evaluated the prediction performance by the receiver operating characteristic area under the curve (ROC-AUC). For each clinical trial phase (preclinical, phases 1–3, and approved treatments), we defined positive cases as drugs that had reached or passed that particular phase, while negative cases were drugs that had not reached that phase. The ROC-AUC combines true positive, true negative, false positive, and false negative rates into a single metric, with values ranging from 0 to 1, where 1.0 indicates perfect classification. It is defined as:

$$AUC = \int_{x=0}^{1} TPR\left(FPR^{-1}(x)\right) dx \tag{3}$$

where $TPR = \frac{\text{True Positive}}{\text{True Positive+False Negative}}$, and the $FPR = \frac{\text{False Positive}}{\text{False Positive+True Negative}}$.

3 Results

Our evaluation focuses on three key objectives: predictive accuracy in drug-disease association identification, model robustness under extreme perturbations, and clinical relevance of the predictions.

3.1 *Labyrinth* yields promising results in alignment with medical standards

To validate our approach, we initially evaluated the predictive accuracy across various diseases. This involved assessing the Spearman correlations between the priority scores assigned by *labyrinth* and the established weights in clinical trials alongside proximity metrics for each

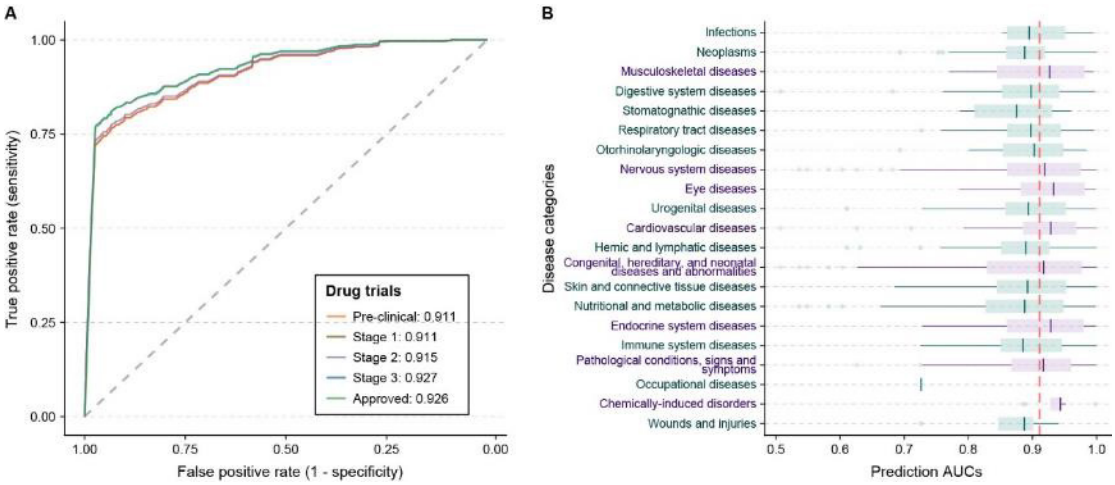


Figure 3. *Labyrinth* predicts well across all disease categories. (**A**) ROC curves across five clinical trial stages, demonstrating ROC-AUC values exceeding 0.90, indicating strong predictive performance. (**B**) Box plot of ROC-AUC values for stage 3 drug usability across disease categories, with dots indicating individual diseases and a red dashed line for the overall ROC-AUC value. *Labyrinth* showed a high accuracy is observed in all categories except for occupational and stomatognathic diseases.

drug–disease pair. *Labyrinth* exhibited moderate to high correlations, with coefficients of 0.60 for clinical trials and 0.80 for proximity, respectively.

Subsequently, we extended our analysis to encompass all human diseases, aiming to assess the predictive performance across five clinical trial phases, including preclinical, phases 1 to 3, and approved treatments. As illustrated in **Figure 3A**, the ROC-AUC values for all stages surpassed 0.90 in the entire data set, indicating a predictive success rate of over 90% in distinguishing between drugs classified for clinical trials or non-clinical trials.

Notably, *labyrinth* exhibited high predictive accuracy in determining drug usability for Stage 3 across all disease categories except for occupational and stomatognathic diseases (**Figure 3B**). Also, cardiovascular, endocrine system diseases, and neoplasms garnered the most significant benefits from *labyrinth*. Detailed ROC-AUC predictions for all diseases are provided in the Supporting Information.

3.2 *Labyrinth* learns implicit mechanisms that mediate specific drug response

Fatty-liver contributes to a 26% increase in overall health costs over five years[36] and exhibits disparities across racial and ethnic groups[37]. To demonstrate the predictive capability of our framework in addressing actual human diseases, we chose fatty liver as a case study. We inputted genes associated with fatty liver from DisGeNET into *labyrinth* and then identified the top ten therapeutic candidates (**Table S1**). Analysis revealed that three targets were common across four to six repurposed drugs (**Figure 4A**). These targets were then visualized within a functional interactome network, highlighted in **Figure 4B** with fatty liver-related genes emphasized. Notably, the top three most prevalent genes within this network (*PPARG*, *PPARA*, and *ESR1*) are directly implicated in fatty liver pathogenesis. *PPARG* and *PPARA* are members of the peroxisome proliferator-activated receptor (*PPAR*) nuclear receptor subfamily, while *ESR1* encodes the estrogen receptor nuclear receptor subfamily.

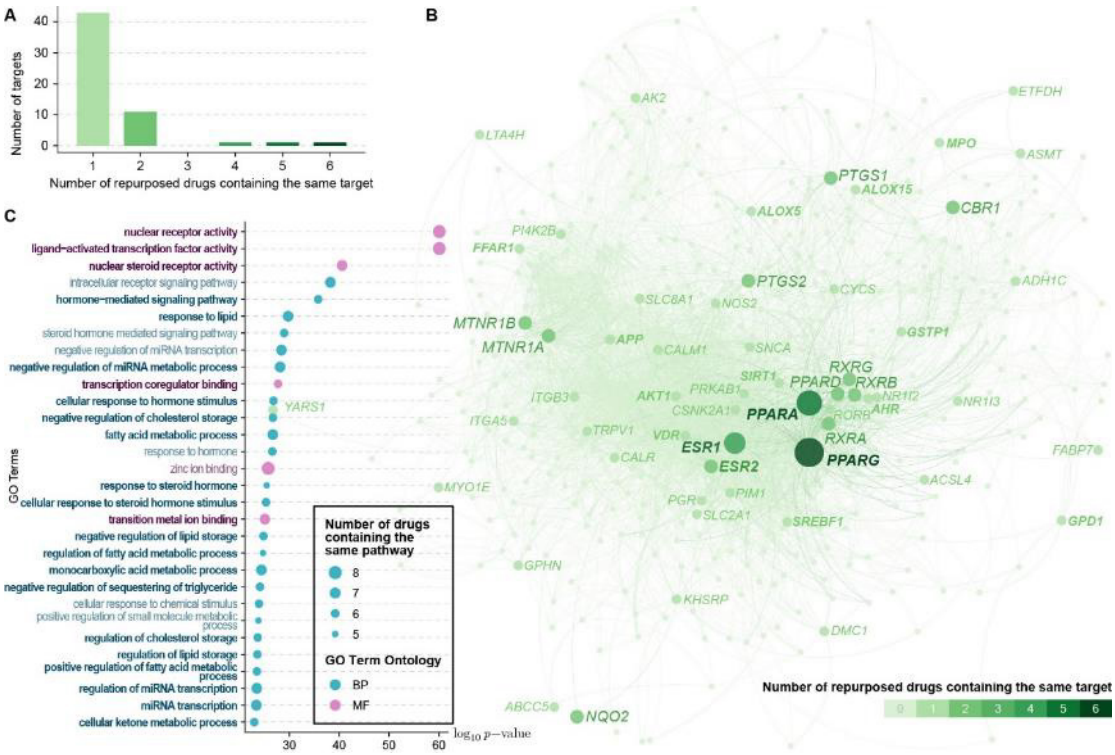


Figure 4. *Labyrinth* captures druggable targets and pathways: fatty liver as a case. (**A**) Bar plot highlighting the frequency of shared targets among repurposed drugs. (**B**) Protein–protein interaction network illustrating connections between fatty liver-associated proteins and repurposed drug targets, with node darkness and size indicating target frequency. The nodes with labels represent the top ten drug targets. The bolder text represents fatty liver related nodes. (**C**) Enrichment analysis of these top drug targets within GO terms. The color represents the subcategory of the GO term. The bolder texts are fatty liver related pathways. Pathways are sorted in descending order of log 10 transformed *p* value.

PPARG is targeted by six repurposed drugs. It plays a vital role in adipocyte differentiation, adipogenesis, and lipid metabolism, showing significantly elevated expression levels in patients with non-alcoholic fatty liver disease (NAFLD)[38]. *PPARA* is the second most common target. It is crucial for fatty acid oxidation regulation, with its hepatic expression often upregulated by high-fat diets[38]. *ESR1* is the third major target. *ESR1* predominantly influences the liver's response to estrogens, with *ESR1* knockout leading to increased weight and obesity in female rats[39],highlighting gender-specific effects. These findings underscore the significance of these targets in fatty liver pathogenesis, as demonstrated through animal studies. Despite the promise of *PPAR*-agonists in treatment, their clinical application is hindered by potential side effects such as idiosyncratic hepatotoxicity, fluid retention, and weight gain[40].

Additionally, we conducted a Gene Ontology (GO) enrichment analysis of these drug targets to assess the applicability of the repurposed drugs. As illustrated in **Figure 4C**, the analysis revealed that over two-thirds of the top 30 pathways are associated with fatty liver and fat metabolism, indicating the success of repurposing. Detailed references supporting this evidence are provided in **Table S2**.

3.3 *Labyrinth* aligns with medical practice

As mentioned above, we validated *labyrinth* using disease genes from DisGeNET. Further, we assessed its applicability to clinical practice with bulk RNA-seq data from The Cancer Genome

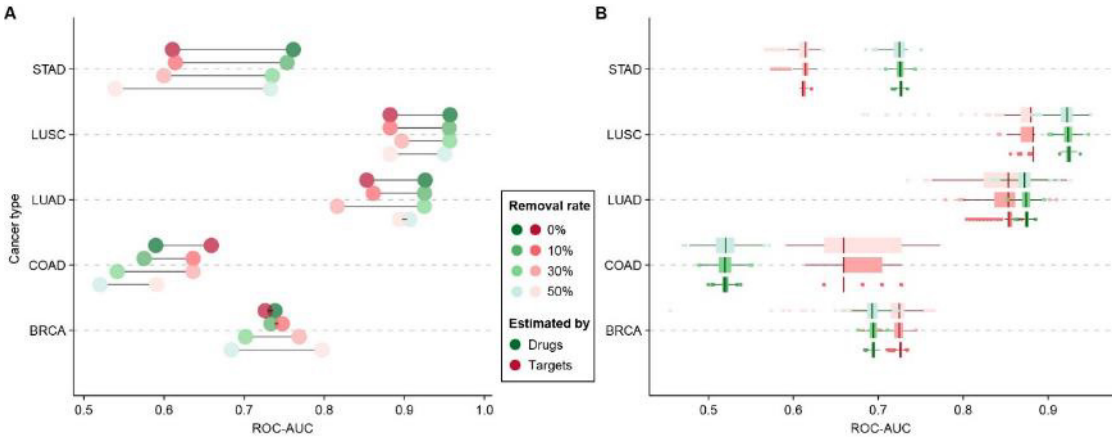


Figure 5. *Labyrinth* is aligned with medical practice and robustness in perturbations. (**A**) Dot plot predictive performance in five cancer types after edge deletions, with color coding for evaluation criteria and varying transparency indicating the last 10%, 30%, and 50% edge removal rates. (**B**) Box plot showing performance with random edge deletions, using the same color and transparency scheme to represent different criteria and deletion rates.

Atlas (TCGA), a comprehensive database documenting genomic characteristics across 33 cancer types that makes possible to compare characteristics among multiple types[41]. Utilizing treatment records and RNA-seq transcripts from TCGA, we aimed to confirm the alignment of *labyrinth* with medical practices.

Upon reviewing the TCGA data, we excluded patients lacking treatment records or with illogical treatment timelines (e.g., treatments ending before their start date). We further narrowed the data set by eliminating cancer types with uniform treatment responses, retaining only five cancer types for analysis.

Our evaluation focused on the usability of repurposed drugs for patients. As represented by the green dots in **Figure 5A**, the ROC-AUC values were high across these cancer types, indicating a strong alignment of *labyrinth* with clinical practice. Specifically, LUSC and LUAD showed over 90% coherence, and BRCA and STAD exhibited up to 70% coherence. The ROC-AUC result of COAD was less promising due to the lack of COAD samples, leading to a coherence of only around 60%. Detailed methodologies can be found in **Materials and Methods**.

Further analysis of the targets of repurposed drugs indicated a potential predictive value for patient response, as evidenced by similarly high ROC-AUC values shown in red dots in **Figure 5A**. This suggests that drug prioritization predictions can effectively predict patient responses.

3.4 *Labyrinth* is robust in corner cases and perturbations

Due to the prevailing risk-averse strategy in drug discovery, alongside a focus on previously validated drug targets by clinicians, researchers, and pharmacies has led to an oversight of potentially druggable proteins linked to diseases[42]. This trend results in a disproportionate focus on a limited set of widely studied drugs or treatments as visualized by the long-tail distribution of drugs in clinical trials. Specifically, a few drugs frequently appear in trials and papers, whereas the majority are seldom tested.

Beyond drug discovery, the broader fields of science and technology also exhibit a decline in innovation and a tendency toward conservatism despite an increase in research output[43] and the use of positive words[44]. This conservatism contributes to a knowledge distribution that is

heavily skewed toward widely studied topics. This bias toward the head of the distribution hinders addressing niches or corner cases. Compounding the issue, inconsistencies undermine the credibility of the publications with about half of the publications having found that at least one primary outcome was changed, introduced, or omitted compared to their written protocols[45]. Hence, it is crucial and necessary to examine the robustness when corner cases and uncertainties.

To assess robustness in facing corner cases, we evaluated the predictive accuracy after removing the last 10%, 30%, and 50% of the edges from the original model. Despite these slims, the predictive performance decreased by less than 10% in all cancer types (see **Figure 5A**). This indicates that nearly 10% of the predictions are impacted in the long run, thus demonstrating its resilience in corner cases.

Further analysis of clinical trial inconsistencies revealed a minor but significant portion of protocol discrepancies[46]. Treating these inconsistencies as random perturbations, we tested the performance of *labyrinth* under such conditions by randomly removing edges in its model[47]. As **Figure 5B** shows, the predictive accuracy remained consistent across all cancer types, even with up to 50% of the edges dropped. These validations indicate the robustness of *labyrinth* against both specific outlier scenarios and random perturbations.

3.5 *Labyrinth* can be utilized in personalized medicine

Few drug repurposing algorithms are adaptable for both cohort analyses and individual patient scenarios. We then applied *labyrinth* to a data set comprising samples from melanoma patients with clinically acquired resistance to *MAPK* inhibitor therapies (GSE65185)[48] to evaluate its utility in personalized medicine. This data set contains 67 samples from 24 patients, averaging approximately 2.7 samples per patient. Given the difficulty of performing differential analysis with such a matrix as above, we instead utilized robust principal component analysis (RPCA) for our analysis to avoid the scenario that the standard deviation calculation requires at least three samples per group. The main idea is illustrated in **Figure 6A** and described in **Materials and Methods**.

As for RPCA, having at least three samples per patient is preferable. By selecting patients with three or more samples, we focused on these nine patients for drug repurposing. The correlation map is shown in **Figure 6B** and **Table S3**, indicating the patient-wise correlation of 0.99, suggesting *labyrinth* repurposed similar drugs for these patients. Given their melanoma diagnosis, we analyzed the effectiveness of the repurposed drugs. The correlations exceed 0.80, which highlights distinct differences between patients with *MAPK* inhibitor-resistant melanoma and those with typical melanoma. Comparisons with skin cancer yielded correlations above 0.70. Furthermore, we depicted the principal component analysis (PCA) plot in **Figure S1**. Nine *MAPK* resilience patients are in a cluster, underscoring their shared characteristics among these patients.

Building on findings by Hugo et al.[48] regarding the role of c-Met upregulation in *MAPK* inhibitor resilience, we analyzed the top 100 drug targets against melanoma-related genes from DisGeNET, the *MAPK* signaling pathway, and the c-Met pathway from KEGG. As illustrated in **Figure 6C**, we focused on the largest connected component of the subgraph for readability. Our analysis revealed that repurposed drug targets are more distantly related to the *MAPK* signaling pathway ($s = 0.872$) than to the c-Met pathway ($s = 0.575$), with the *MAPK* pathway showing greater overlap with melanoma-related genes ($s = -0.398$) compared to

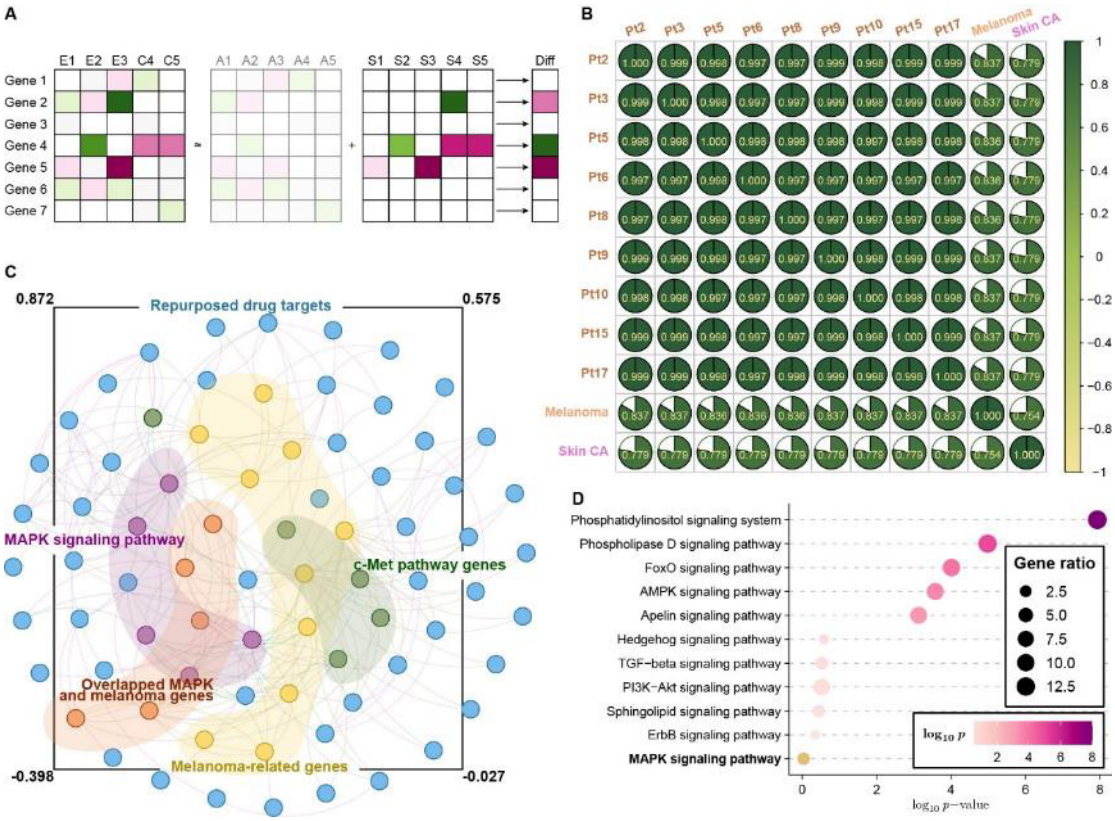


Figure 6. *Labyrinth* can be used in personalized medicine. (**A**) Overview of using RPCA as an alternative to differential analysis. (**B**) Correlation map for nine *MAPK*-resistance melanoma patients, compared to those with standard melanoma and skin cancer. (**C**) The protein–protein interaction network displaying interactions among repurposed drug targets, *MAPK* and c-Met pathways, and melanoma-related genes, focusing on the largest connected component for clarity. (**D**) Enrichment of repurposed targets in the KEGG signaling pathway subcategory. The pathways are sorted by log-transformed *p* values, with color transparency indicating gene ratios within each pathway.

the c-Met pathway ($s = -0.027$). Enrichment analysis in KEGG signaling subcategories identified the top nine pathways sorted by log-transformed *p* values, excluding the MAPK signaling pathway from the list of the top pathways (**Figure 6D**). This suggests the repurposed drugs might effectively circumvent the *MAPK* signaling pathway, potentially destabilizing c-Met gene expressions.

3.6 Benchmarking against other drug repurposing methods

To evaluate its performance against existing approaches, we utilized a recently published data set (GSE248619) documenting the phase 2 clinical trial of enzalutamide in metastatic castration-resistant prostate cancer (mCRPC) patients, which is an androgen receptor (*AR*) targeted drug[49]. This data set was used as a case study to test our model. We compared *labyrinth* with four state-of-the-art drug repurposing methods: DTSEA[50], DrugVsDisease[51], SubtypeDrug[52], and CMap[53]. To ensure fair comparison, all methods were evaluated using the same data set with their optimized parameters. In the comparison, we define successful repurposing as the ability to identify *AR*-targeted drugs. As illustrated in **Figure S7**, *labyrinth* achieved the highest ROC-AUC (0.773) among all methods tested.

4 Discussion

Drug repositioning is a critical challenge in contemporary pharmaceutical research. While existing computational approaches excel at pattern recognition[54], they often lack interpretability and reasoning processes. This study introduces *labyrinth* as a novel strategy for recommending existing drugs at both the population and individual levels. Our framework integrates explicit knowledge (clinical trials and drug–target interactions), implicit knowledge (literature-derived relationships), and experiential knowledge (expressions and historical treatment outcomes). This integration occurs within an associative network structure that mirrors human memory organization, where concepts are linked through meaningful relationships rather than simple co-occurrences. Our approach qualitatively and quantitatively validates the utility of *labyrinth* in identifying potential treatments for human diseases. Our model is designed as a knowledge-based system that simulates human cognitive processes in drug prioritization.

We tested *labyrinth* across various human diseases validated by the ROC-AUC metrics, and confirmed its predictions align closely with clinical trials. The robustness of the model was unexpectedly high particularly against random perturbations. However, the predictions of *labyrinth* are based on clinical trials, which raise significant concerns about their reliability and credibility. Approximately one-third of the researchers have admitted to engaging in practices potentially considered fraudulent, shedding light on the unsettling prevalence of questionable research integrity within the field. While some researchers justify their fraud practices under defensible reasons[55], the existence of published fraudulent studies is being verified in large multicentered random clinical trials, especially given the propensity of medical research to suffer from untrustworthy clinical trials[56] and results[57]. Thus, it is important to raise the danger of clinicians getting unreliable results from such a model.

Critically, evaluating the validity of a model solely on the basis of its efficiency in fitting can be misleading. Like in clinical settings, a clinician relies on their professional training from medical education, academic literature, reliable information sources from Google, and their own clinical experience to prescribe or evaluate drug efficacy for their patients. As an example, labyrinth integrates two independent knowledge sources to simulate human-like knowledge retrieval for drug prioritization, which limits the methods for validation.

Our methodology is inspired by the human cognitive process of memory retrieval rather than tuning and fitting an uninterpretable model. Consequently, simplicity often becomes a better solution when making interpretations[58]. However, simplifying the model does not mean overlooking the complexity of the data or the multifaceted nature of diseases and treatments. By aiming to model the process of human memory retrieval, we seek to capture this complexity in a manner that is both intuitive and scientifically sound. A core contribution of our work is the development of an interpretable model that aligns with human cognitive processes such as memory retrieval and reasoning. To this end, we aim to construct a model that not only predicts with high reliability but also aligns with intuitive, common-sense strategies for knowledge application and problem-solving.

Despite our efforts to perfect this balance of predictive accuracy and minimalism, integrating the processes for bulk RNA-seq data and individual patients into a unified model remains challenging. Preliminary experiments revealed a moderate correlation ($r_s > 0.4$) between disease weights derived using DESeq2 and RPCA methods across five cancer types in the TCGA database (**Figure S2**), potentially attributed to the extensive sample heterogeneity within the

TCGA cohorts. The current version of *labyrinth* operates at the main disease category level without considering the subtypes of diseases, which may limit its precision in cases in which treatment responses vary significantly among subtypes.

Moreover, patient costs may increase as a result of this approach extended to personalized treatment, particularly when multiple tissue samples are necessary for sequencing. Although these costs could be reduced by more streamlined methods, the current approach emphasizes the economic and logistical burdens of advanced sequencing techniques, especially for underprivileged populations. We consider the practicality and affordability of diagnostic methods, such as immunohistochemistry against the more sensitive but costlier next-generation sequencing in detecting biomarkers and isolating subtypes of cancer[59,60]. This consideration is crucial for ensuring equitable access to advanced medical diagnostics across diverse socioeconomic backgrounds.

In conclusion, *labyrinth* represents a pioneering approach to drug recommendation, distinguished by its simulation of cognitive processes and human knowledge retrieval. Its implications for future research are vast, promising a new direction in the integration of multidisciplinary efforts toward enhancing personalized medicine. This study not only demonstrates the potential of *labyrinth* in revolutionizing drug repurposing but also highlights the broader challenges of ensuring research integrity and accessibility in the journey toward advanced personalized medical solutions.

5 Conclusions

We present a computational framework that simulates the cognitive abilities of humans for drug repositioning and precision medicine applications. *Labyrinth* integrates multiple data sources like clinical trials, literature co-occurrences, drug–target interactions, and disease similarities to identify potential drug candidates in a human-like knowledge retrieval approach. Combining predictive accuracy with model interpretability, we demonstrate its robust performance across diverse diseases, underscoring the importance of aligning computational models with intuitive human reasoning for personalized medicine. By drawing inspiration from human memory and knowledge association, this interdisciplinary work advances drug prioritization while highlighting the value of biomimetic artificial intelligence in biomedical research.

Acknowledgment

We express our gratitude to Microsoft for providing the high-quality icons utilized in **Figure 1** and **Figure 2** during the preparation of the schematic diagrams.

Figure 2 includes logos from ChEMBL, DrugBank, CTD, and Web of Science. Web of Science is a trademark owned by Clarivate.

During the preparation of this manuscript and the documentation of *labyrinth*, we employed Claude for grammatical corrections and enhancements in readability. We thoroughly reviewed and took great care of the content to ensure its quality, making edits as necessary. Consequently, we take full responsibility for the content presented in this work.

Ethics approval and consent to participate

No need for approval and consent to disclose.

Data and software availability statement

The original data from DrugBank, CTD, ChEMBL, Cochrane Library, DisGeNET, and TCGA data are publicly available via their official websites. Data for the MAPK resistance cohort were sourced from GEO database (GSE65185), as detailed in the cited manuscript.

The functional interactome network was compiled from seven databases, including KEGG, Reactome, Biocarta, NCI, SPIKE, HumanCyc, and Panther.

Trained model files are hosted on GitHub repository (https://github.com/hanjunwei-lab/labyrinth).

All relevant data is available from the authors. To adhere to legal compliance requirements, unprocessed source data is provided upon reasonable request.

We have wrapped the core function into an R package named *labyrinth*, which is freely available on GitHub under the GPL-v2 license (https://github.com/hanjunwei-lab/labyrinth).

Funding sources

This work was supported by National Natural Science Foundation of China (grant no. 62372143 and 62072145) and the Natural Science Foundation of Heilongjiang Province (grant no. LH2019C042).

Competing interests

All authors claim that they have no competing interests.

Consent for publication

All authors contributed to the article and approved the submitted version.

Abbreviations

CTD, the Comparative Toxicogenomic Database

GO, Gene Ontology

GSEA, Gene Set Enrichment Analysis

LLM, large language models

LTM, long-term memory

NAFLD, non-alcoholic fatty liver disease

PCA, principal component analysis

PPAR, peroxisome proliferator-activated receptor

ROC-AUC, receiver operating characteristics area under the curve

RPCA, robust principal component analysis

RWR, random walk with restarts

STM, short-term memory

TCGA, The Cancer Genome Atlas

TF-IDF, term frequency and inverse document frequency

WOS, Web of Science

References

(1) Parvathaneni, V.; Kulkarni, N. S.; Muth, A.; Gupta, V. Drug Repurposing: A Promising Tool to Accelerate the Drug Discovery Process. *Drug Discov. Today* **2019**, *24* (10), 2076–2085. https://doi.org/10.1016/j.drudis.2019.06.014.

(2) Fernández-Torras, A.; Duran-Frigola, M.; Aloy, P. Encircling the Regions of the Pharmacogenomic Landscape That Determine Drug Response. *Genome Med.* **2019**, *11* (1), 17. https://doi.org/10.1186/s13073-019-0626-x.

(3) Sudhahar, S.; Ozer, B.; Chang, J.; Chadwick, W.; O'Donovan, D.; Campbell, A.; Tulip, E.; Thompson, N.; Roberts, I. An Experimentally Validated Approach to Automated Biological Evidence Generation in Drug Discovery Using Knowledge Graphs. *Nat. Commun.* **2024**, *15* (1), 5703. https://doi.org/10.1038/s41467-024-50024-6.

(4) Chandak, P.; Huang, K.; Zitnik, M. Building a Knowledge Graph to Enable Precision Medicine. *Sci. Data* **2023**, *10* (1), 67. https://doi.org/10.1038/s41597-023-01960-3.

(5) Zhu, Y.; Che, C.; Jin, B.; Zhang, N.; Su, C.; Wang, F. Knowledge-Driven Drug Repurposing Using a Comprehensive Drug Knowledge Graph. *Health Informatics J.* **2020**, *26* (4), 2737–2750. https://doi.org/10.1177/1460458220937101.

(6) Gong, F.; Wang, M.; Wang, H.; Wang, S.; Liu, M. S. M. R: Medical Knowledge Graph Embedding for Safe Medicine Recommendation. *Big Data Res.* **2021**, *23*, 100174. https://doi.org/10.1016/j.bdr.2020.100174.

(7) Sharifian, F.; Samani, R. Hierarchical Spreading of Activation. In *Proc. of the Conference on Language, Cognition, and Interpretation*; IAU Press Isfahan, 1997; pp 1–10.

(8) Collins, A. M.; Loftus, E. F. A Spreading-Activation Theory of Semantic Processing. *Psychol. Rev.* **1975**, *82* (6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407.

(9) Anderson, J. R. A Spreading Activation Theory of Memory. *J. Verbal Learn. Verbal Behav.* **1983**, *22* (3), 261–295. https://doi.org/10.1016/S0022-5371(83)90201-3.

(10) Gerlach, M.; Shi, H.; Amaral, L. A. N. A Universal Information Theoretic Approach to the Identification of Stopwords. *Nat. Mach. Intell.* **2019**, *1* (12), 606–612. https://doi.org/10.1038/s42256-019-0112-6.

(11) Li, C.; Li, Z.; Wang, S.; Yang, Y.; Zhang, X.; Zhou, J. Semi-Supervised Network Embedding. In *Database Systems for Advanced Applications*; Candan, S., Chen, L., Pedersen, T. B., Chang, L., Hua, W., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2017; Vol. 10177, pp 131–147. https://doi.org/10.1007/978-3-319-55753-3_9.

(12) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018**, *46* (D1), D1074–D1082. https://doi.org/10.1093/nar/gkx1037.

(13) Zdrazil, B.; Felix, E.; Hunter, F.; Manners, E. J.; Blackshaw, J.; Corbett, S.; de Veij, M.; Ioannidis, H.; Lopez, D. M.; Mosquera, J. F.; Magarinos, M. P.; Bosc, N.; Arcila, R.; Kizilören, T.; Gaulton, A.; Bento, A. P.; Adasme, M. F.; Monecke, P.; Landrum, G. A.; Leach, A. R. The ChEMBL Database in 2023: A Drug Discovery Platform Spanning Multiple Bioactivity Data Types and Time Periods. *Nucleic Acids Res.* **2023**, gkad1004. https://doi.org/10.1093/nar/gkad1004.

(14) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40* (D1), D1100–D1107. https://doi.org/10.1093/nar/gkr777.

(15) Davis, A. P.; Wiegers, T. C.; Johnson, R. J.; Sciaky, D.; Wiegers, J.; Mattingly, C. J. Comparative Toxicogenomics Database (CTD): Update 2023. *Nucleic Acids Res.* **2023**, *51* (D1), D1257–D1262. https://doi.org/10.1093/nar/gkac833.

(16) Chambers, J.; Davies, M.; Gaulton, A.; Hersey, A.; Velankar, S.; Petryszak, R.; Hastings, J.; Bellis, L.; McGlinchey, S.; Overington, J. P. UniChem: A Unified Chemical Structure Cross-Referencing and Identifier Tracking System. *J. Cheminform.* **2013**, *5* (1), 3. https://doi.org/10.1186/1758-2946-5-3.

(17) Liu, W. A Matter of Time: Publication Dates in Web of Science Core Collection. *Scientometrics* **2021**, *126* (1), 849–857. https://doi.org/10.1007/s11192-020-03697-x.

(18) Fan, Y.; Arora, C.; Treude, C. Stop Words for Processing Software Engineering Documents: Do They Matter? arXiv **2023**. https://doi.org/10.48550/arXiv.2303.10439.

(19) Saldanha, I. J.; Adam, G. P.; Schmid, C. H.; Trikalinos, T. A.; Konnyu, K. J. Modernizing Evidence Synthesis for Evidence-Based Medicine. In *Clinical Decision Support and Beyond*; Elsevier, 2023; pp 257–278. https://doi.org/10.1016/B978-0-323-91200-6.00006-1.

(20) Lahitani, A. R.; Permanasari, A. E.; Setiawan, N. A. Cosine Similarity to Determine Similarity Measure: Study Case in Online Essay Assessment. In *2016 4th International Conference on Cyber and IT Service Management*; IEEE: Bandung, Indonesia, 2016; pp 1–6. https://doi.org/10.1109/CITSM.2016.7577578.

(21) Menche, J.; Sharma, A.; Kitsak, M.; Ghiassian, S. D.; Vidal, M.; Loscalzo, J.; Barabasi, A.-L. Uncovering Disease-Disease Relationships through the Incomplete Interactome. *Science* **2015**, *347* (6224), 1257601. https://doi.org/10.1126/science.1257601.

(22) Tattershall, E.; Nenadic, G.; Stevens, R. D. Detecting Bursty Terms in Computer Science Research. *Scientometrics* **2020**, *122* (1), 681–699. https://doi.org/10.1007/s11192-019-03307-5.

(23) Gaggero, G.; Bonassi, A.; Dellantonio, S.; Pastore, L.; Aryadoust, V.; Esposito, G. A Scientometric Review of Alexithymia: Mapping Thematic and Disciplinary Shifts in Half a Century of Research. *Front. Psychiatry* **2020**, *11*, 611489. https://doi.org/10.3389/fpsyt.2020.611489.

(24) Perez, C.; Germon, R. Graph Creation and Analysis for Linking Actors: Application to Social Data. In *Automating open source intelligence*; Syngress: Oxford, UK, 2016; pp 103–129.

(25) Amjad, T.; Shahid, N.; Daud, A.; Khatoon, A. Citation Burst Prediction in a Bibliometric Network. *Scientometrics* **2022**, *127* (5), 2773–2790. https://doi.org/10.1007/s11192-022-04344-3.

(26) Shen, L.; Xiong, B.; Li, W.; Lan, F.; Evans, R.; Zhang, W. Visualizing Collaboration Characteristics and Topic Burst on International Mobile Health Research: Bibliometric Analysis. *JMIR MHealth UHealth* **2018**, *6* (6), e135. https://doi.org/10.2196/mhealth.9581.

(27) Kleinberg, J. Bursty and Hierarchical Structure in Streams. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*; ACM: Edmonton Alberta Canada, 2002; pp 91–101. https://doi.org/10.1145/775047.775061.

(28) Yeo, I.-K. A New Family of Power Transformations to Improve Normality or Symmetry. *Biometrika* **2000**, *87* (4), 954–959. https://doi.org/10.1093/biomet/87.4.954.

(29) Gel, Y. R.; Miao, W.; Gastwirth, J. L. Robust Directed Tests of Normality against Heavy-Tailed Alternatives. *Comput. Stat. Data Anal.* **2007**, *51* (5), 2734–2746. https://doi.org/10.1016/j.csda.2006.08.022.

(30) Ye, H.; Sun, M.; Su, M.; Chen, D.; Liu, H.; Ma, Y.; Luo, W.; Li, H.; Xu, F. Identification of Disease-Related Genes and Construction of a Gene Co-Expression Database in Non-Alcoholic Fatty Liver Disease. *Front. Genet.* **2023**, *14*, 1070605. https://doi.org/10.3389/fgene.2023.1070605.

(31) Love, M. I.; Huber, W.; Anders, S. Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2. *Genome Biol.* **2014**, *15* (12), 550. https://doi.org/10.1186/s13059-014-0550-8.

(32) Liu, J.-X.; Wang, Y.-T.; Zheng, C.-H.; Sha, W.; Mi, J.-X.; Xu, Y. Robust PCA Based Method for Discovering Differentially Expressed Genes. *BMC Bioinformatics* **2013**, *14* (S8), S3. https://doi.org/10.1186/1471-2105-14-S8-S3.

(33) Ji, S.; Pan, S.; Cambria, E.; Marttinen, P.; Yu, P. S. A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33* (2), 494–514. https://doi.org/10.1109/TNNLS.2021.3070843.

(34) Kumar, A. A.; Steyvers, M.; Balota, D. A. Semantic Memory Search and Retrieval in a Novel Cooperative Word Game: A Comparison of Associative and Distributional Semantic Models. *Cogn. Sci.* **2021**, *45* (10), e13053. https://doi.org/10.1111/cogs.13053.

(35) Fathan, M. I.; Renfro, E. K.; Austerweil, J. L.; Beckage, N. M. Do Humans Navigate via Random Walks? Modeling Navigation in a Semantic Word Game. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*; Cognitive Science Society: Austin, TX, 2018.

(36) Baumeister, S. E.; Völzke, H.; Marschall, P.; John, U.; Schmidt, C.; Flessa, S.; Alte, D. Impact of Fatty Liver Disease on Health Care Utilization and Costs in a General Population: A 5-Year Observation. *Gastroenterology* **2008**, *134* (1), 85–94. https://doi.org/10.1053/j.gastro.2007.10.024.

(37) Rich, N. E.; Oji, S.; Mufti, A. R.; Browning, J. D.; Parikh, N. D.; Odewole, M.; Mayo, H.; Singal, A. G. Racial and Ethnic Disparities in Nonalcoholic Fatty Liver Disease Prevalence, Severity, and Outcomes in the United States: A Systematic Review and Meta-Analysis. *Clin. Gastroenterol. Hepatol.* **2018**, *16* (2), 198-210.e2. https://doi.org/10.1016/j.cgh.2017.09.041.

(38) Liss, K. H. H.; Finck, B. N. PPARs and Nonalcoholic Fatty Liver Disease. *Biochimie* **2017**, *136*, 65–74. https://doi.org/10.1016/j.biochi.2016.11.009.

(39) Khristi, V.; Ratri, A.; Ghosh, S.; Pathak, D.; Borosha, S.; Dai, E.; Roy, R.; Chakravarthi, V. P.; Wolfe, M. W.; Karim Rumi, M. A. Disruption of ESR1 Alters the Expression of Genes Regulating Hepatic Lipid and Carbohydrate Metabolism in Male Rats. *Mol. Cell. Endocrinol.* **2019**, *490*, 47–56. https://doi.org/10.1016/j.mce.2019.04.005.

(40) Nanjan, M. J.; Mohammed, M.; Prashantha Kumar, B. R.; Chandrasekar, M. J. N. Thiazolidinediones as Antidiabetic Agents: A Critical Review. *Bioorganic Chem.* **2018**, *77*, 548–567. https://doi.org/10.1016/j.bioorg.2018.02.009.

(41) Liu, J.; Lichtenberg, T.; Hoadley, K. A.; Poisson, L. M.; Lazar, A. J.; Cherniack, A. D.; Kovatich, A. J.; Benz, C. C.; Levine, D. A.; Lee, A. V.; Omberg, L.; Wolf, D. M.; Shriver, C. D.; Thorsson, V. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **2018**, *173* (2), 400. https://doi.org/10.1016/j.cell.2018.02.052.

(42) Vasan, K.; Gysi, D. M.; Barabási, A.-L. The Clinical Trials Puzzle: How Network Effects Limit Drug Discovery. *iScience* **2023**, *26* (12), 108361. https://doi.org/10.1016/j.isci.2023.108361.

(43) Park, M.; Leahey, E.; Funk, R. J. Papers and Patents Are Becoming Less Disruptive over Time. *Nature* **2023**, *613* (7942), 138–144. https://doi.org/10.1038/s41586-022-05543-x.

(44) Vinkers, C. H.; Tijdink, J. K.; Otte, W. M. Use of Positive and Negative Words in Scientific PubMed Abstracts between 1974 and 2014: Retrospective Analysis. *BMJ* **2015**, *351*, h6467. https://doi.org/10.1136/bmj.h6467.

(45) Dwan, K.; Gamble, C.; Williamson, P. R.; Kirkham, J. J.; the Reporting Bias Group. Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias — An Updated Review. *PLoS One* **2013**, *8* (7), e66844. https://doi.org/10.1371/journal.pone.0066844.

(46) Liu, M.; Gao, Y.; Yuan, Y.; Shi, S.; Yang, K.; Lu, C.; Wu, J.; Zhang, J.; Tian, J. Inconsistency and Low Transparency Were Found between Core Outcome Set Protocol and Full Text Publication: A Comparative Study. *J. Clin. Epidemiol.* **2021**, *131*, 59–69. https://doi.org/10.1016/j.jclinepi.2020.11.009.

(47) Stelling, J.; Sauer, U.; Szallasi, Z.; Doyle, F. J.; Doyle, J. Robustness of Cellular Functions. *Cell* **2004**, *118* (6), 675–685. https://doi.org/10.1016/j.cell.2004.09.008.

(48) Hugo, W.; Shi, H.; Sun, L.; Piva, M.; Song, C.; Kong, X.; Moriceau, G.; Hong, A.; Dahlman, K. B.; Johnson, D. B.; Sosman, J. A.; Ribas, A.; Lo, R. S. Non-Genomic and Immune Evolution of Melanoma Acquiring MAPKi Resistance. *Cell* **2015**, *162* (6), 1271–1285. https://doi.org/10.1016/j.cell.2015.07.061.

(49) Perez-Navarro, E.; Conteduca, V.; Funes, J. M.; Dominguez, J. I.; Martin-Serrano, M.; Cremaschi, P.; Fernandez-Perez, M. P.; Gordoa, T. A.; Font, A.; Vázquez-Estévez, S.; González-Del-Alba, A.; Wetterskog, D.; Mellado, B.; Fernandez-Calvo, O.; Méndez-Vidal, M. J.; Climent, M. A.; Duran, I.; Gallardo, E.; Rodriguez Sanchez, A.; Santander, C.; Sáez, M. I.; Puente, J.; Tudela, J.; Marinas, C.; López-Andreo, M. J.; Castellano, D.; Attard, G.; Grande, E.; Rosino, A.; Botia, J. A.; Palma-Mendez, J.; De Giorgi, U.; Gonzalez-Billalabeitia, E. Prognostic Implications of Blood Immune-Cell Composition in Metastatic Castration-Resistant Prostate Cancer. *Cancers* **2024**, *16* (14), 2535. https://doi.org/10.3390/cancers16142535.

(50) Su, Y.; Wu, J.; Li, X.; Li, J.; Zhao, X.; Pan, B.; Huang, J.; Kong, Q.; Han, J. DTSEA: A Network-Based Drug Target Set Enrichment Analysis Method for Drug Repurposing against COVID-19. *Comput. Biol. Med.* **2023**, *159*, 106969. https://doi.org/10.1016/j.compbiomed.2023.106969.

(51) Pacini, C.; Iorio, F.; Gonçalves, E.; Iskar, M.; Klabunde, T.; Bork, P.; Saez-Rodriguez, J. DvD: An R/Cytoscape Pipeline for Drug Repurposing Using Public Repositories of Gene Expression Data. *Bioinforma. Oxf. Engl.* **2013**, *29* (1), 132–134. https://doi.org/10.1093/bioinformatics/bts656.

(52) Han, X.; Kong, Q.; Liu, C.; Cheng, L.; Han, J. SubtypeDrug: A Software Package for Prioritization of Candidate Cancer Subtype-Specific Drugs. *Bioinforma. Oxf. Engl.* **2021**, *37* (16), 2491–2493. https://doi.org/10.1093/bioinformatics/btab011.

(53) Subramanian, A.; Narayan, R.; Corsello, S. M.; Peck, D. D.; Natoli, T. E.; Lu, X.; Gould, J.; Davis, J. F.; Tubelli, A. A.; Asiedu, J. K.; Lahr, D. L.; Hirschman, J. E.; Liu, Z.; Donahue, M.; Julian, B.; Khan, M.; Wadden, D.; Smith, I. C.; Lam, D.; Liberzon, A.; Toder, C.; Bagul, M.; Orzechowski, M.; Enache, O. M.; Piccioni, F.; Johnson, S. A.; Lyons, N. J.; Berger, A. H.; Shamji, A. F.; Brooks, A. N.; Vrcic, A.; Flynn, C.; Rosains, J.; Takeda, D. Y.; Hu, R.; Davison, D.; Lamb, J.; Ardlie, K.; Hogstrom, L.; Greenside, P.; Gray, N. S.; Clemons, P. A.; Silver, S.; Wu, X.; Zhao, W.-N.; Read-Button, W.; Wu, X.; Haggarty, S. J.; Ronco, L. V.; Boehm, J. S.; Schreiber, S. L.; Doench, J. G.; Bittker, J. A.; Root, D. E.; Wong, B.; Golub, T. R. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **2017**, *171* (6), 1437-1452.e17. https://doi.org/10.1016/j.cell.2017.10.049.

(54) Ren, Z.; Zeng, X.; Lao, Y.; Zheng, H.; You, Z.; Xiang, H.; Zou, Q. A Spatial Hierarchical Network Learning Framework for Drug Repositioning Allowing Interpretation from Macro to Micro Scale. *Commun. Biol.* **2024**, *7* (1), 1413. https://doi.org/10.1038/s42003-024-07107-3.

(55) John, L. K.; Loewenstein, G.; Prelec, D. Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychol. Sci.* **2012**, *23* (5), 524–532. https://doi.org/10.1177/0956797611430953.

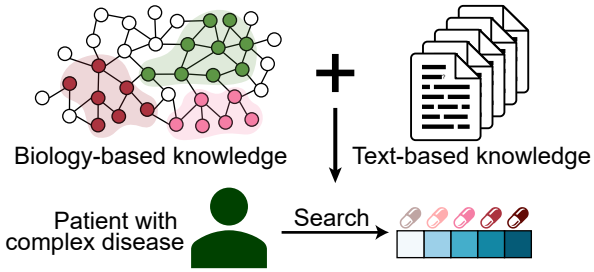
(56) Carlisle, J. B. False Individual Patient Data and Zombie Randomised Controlled Trials Submitted to Anaesthesia. *Anaesthesia* **2021**, *76* (4), 472–479. https://doi.org/10.1111/anae.15263.

(57) Avenell, A.; Bolland, M. J.; Gamble, G. D.; Grey, A. A Randomized Trial Alerting Authors, with or without Coauthors or Editors, That Research They Cited in Systematic Reviews and Guidelines Has Been Retracted. *Account. Res.* **2024**, *31* (1), 14–37. https://doi.org/10.1080/08989621.2022.2082290.

(58) Orozco-Sevilla, V.; Coselli, J. S. Commentary: Occam's Razor: The Simplest Solution Is Always the Best. *J. Thorac. Cardiovasc. Surg.* **2022**, *164* (4), 1053–1054. https://doi.org/10.1016/j.jtcvs.2020.10.087.

(59) Hondelink, L. M.; Schrader, A. M. R.; Asri Aghmuni, G.; Solleveld-Westerink, N.; Cleton-Jansen, A.-M.; van Egmond, D.; Boot, A.; Ouahoud, S.; Khalifa, M. N.; Wai Lam, S.; Morreau, H.; Bovee, J. V. M. G.; van Wezel, T.; Cohen, D. The Sensitivity of Pan-TRK Immunohistochemistry in Solid Tumours: A Meta-Analysis. *Eur. J. Cancer Oxf. Engl.* **2022**, *173*, 229–237. https://doi.org/10.1016/j.ejca.2022.06.030.

(60) Sukswai, N.; Khoury, J. D. Immunohistochemistry Innovations for Diagnosis and Tissue-Based Biomarker Detection. *Curr. Hematol. Malig. Rep.* **2019**, *14* (5), 368–375. https://doi.org/10.1007/s11899-019-00533-9.



Graphical abstract. We trained *labyrinth* with text-based and biology-based knowledge.

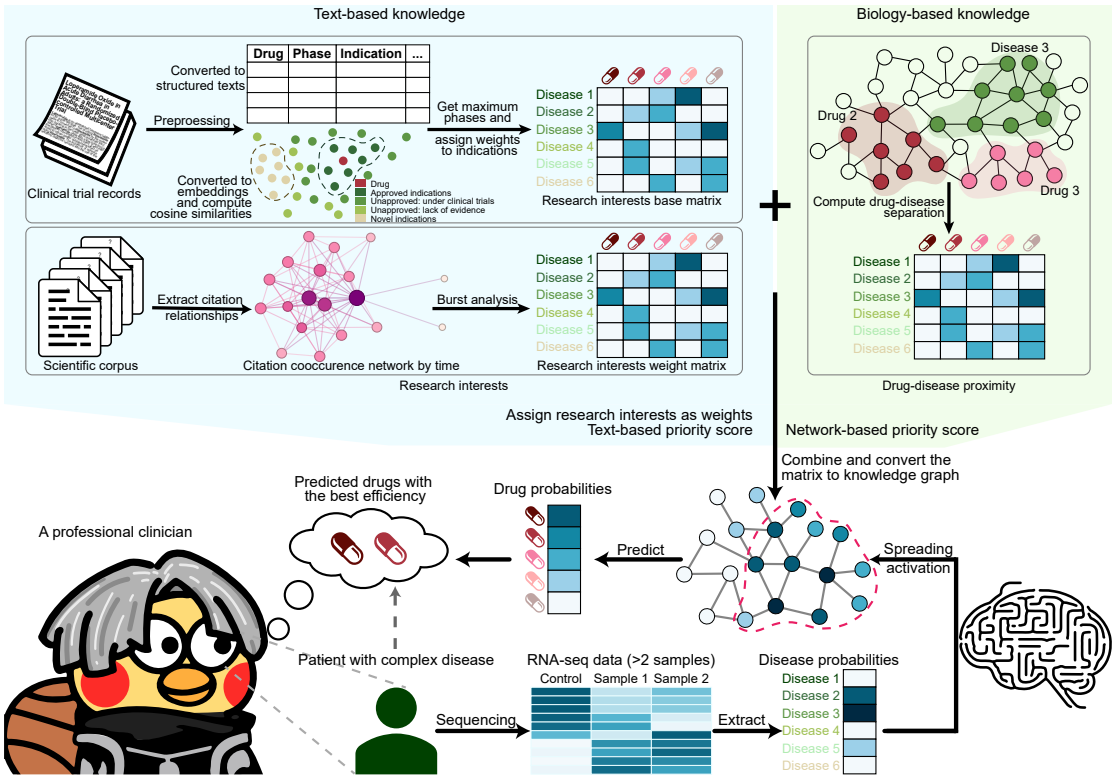


Figure 1. A simple schema of the *labyrinth*. We trained *labyrinth* through the integration of dual knowledge sources: text-based and biology-based. It calculates drug-disease proximity by analyzing the separation within a biologically meaningful functional interactome network. Next, clinical trial information is transformed into structured information to assign weights to drug-disease pairs, while literature from the Scientific Index (SCI) collection from the Web of Science is processed to extract drug–disease relationships, which are then represented in an *n*-dimensional vector space. Cosine similarities between drugs and diseases generate a matrix enriched with citation network analysis to capture the temporal influence of research papers, using citation burst ranges as weights. These processes culminate in a matrix that reflects research interests, which is combined with a biological knowledge matrix through probabilistic computation to simulate human knowledge retrieval for drug prioritization with the best efficiency. This simulation aims to mimic a professional clinician's decision-making process by mapping patients to potential treatments based on disease relevance and treatment efficacy, ultimately identifying candidate drugs with the highest potential for the patient's benefit.

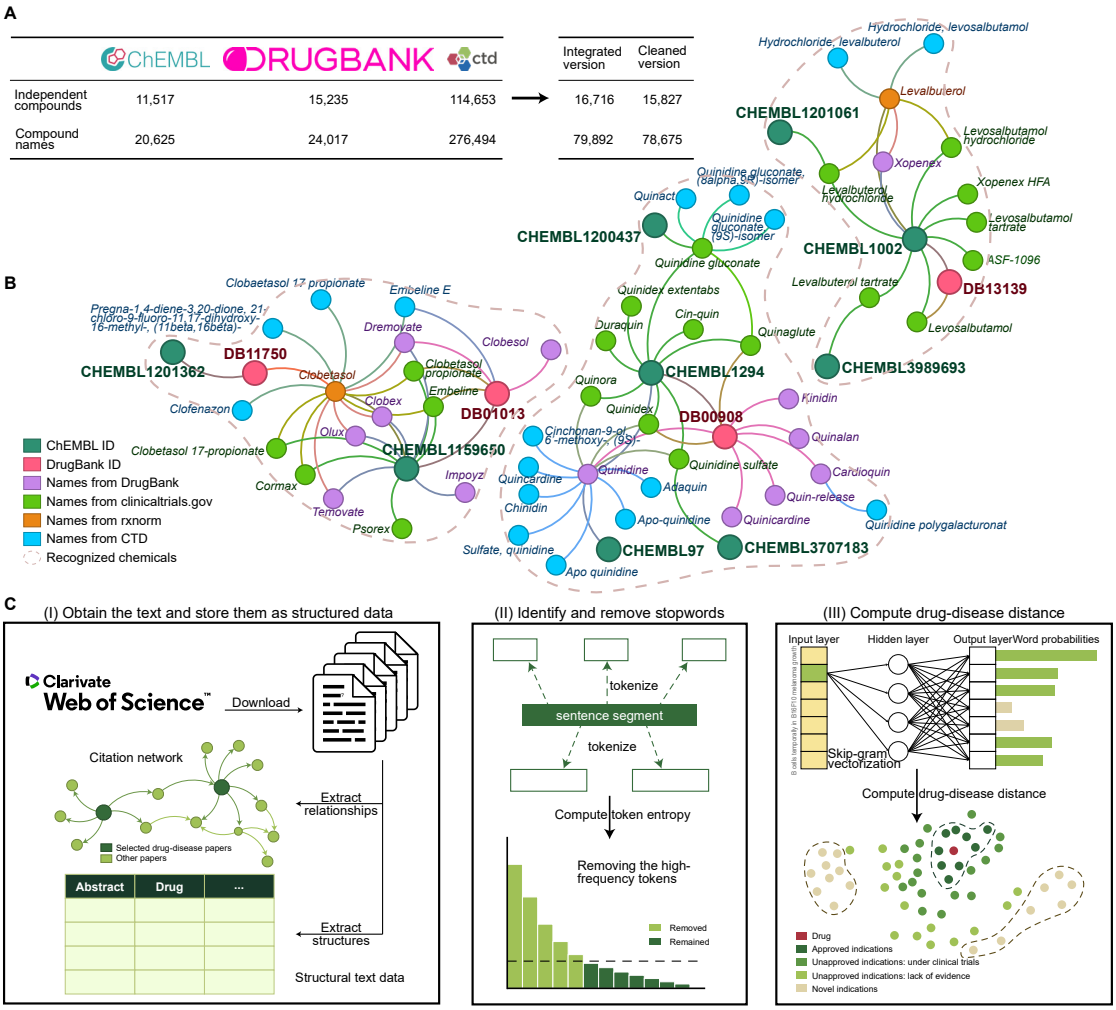


Figure 2. Data preparation process of *labyrinth*. (**A**) Integration of drug information from three distinct databases. (**B**) Network linking of drugs by name or chemical formula; we consider the drugs in each disjoint subfigure to have the same chemical structure. (**C**) Literature text preprocessing workflow. First, texts from the Web of Science are structured and analyzed for paper relationships. Second, text tokenization identifies and removes high-frequency stop words. Finally, we adopted Skip-gram to vectorize the tokens into embeddings, with drug–disease distances calculated via cosine similarity.

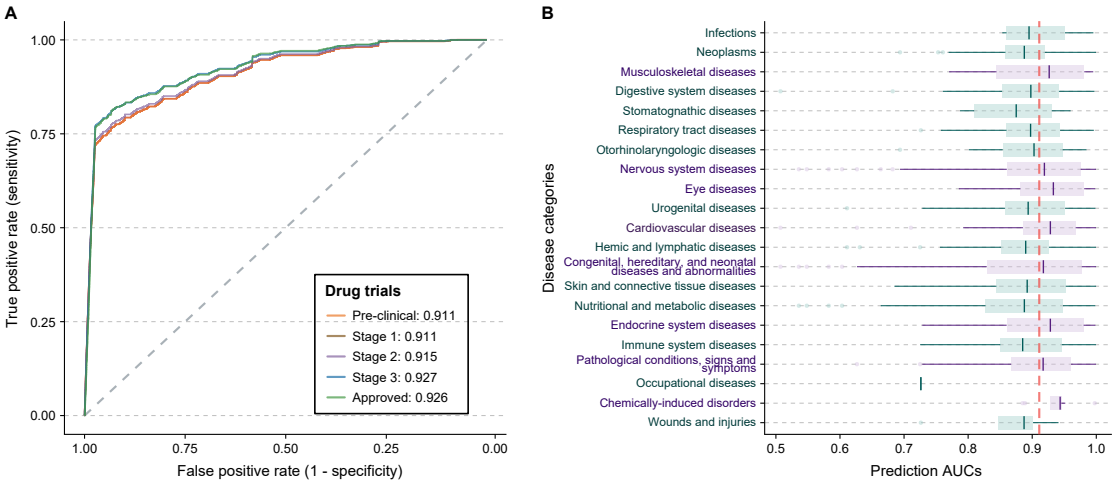


Figure 3. *Labyrinth* predicts well across all disease categories. (**A**) ROC curves across five clinical trial stages, demonstrating ROC-AUC values exceeding 0.90, indicating strong predictive performance. (**B**) Box plot of ROC-AUC values for stage 3 drug usability across disease categories, with dots indicating individual diseases and a red dashed line for the overall ROC-AUC value. *Labyrinth* showed a high accuracy is observed in all categories except for occupational and stomatognathic diseases.

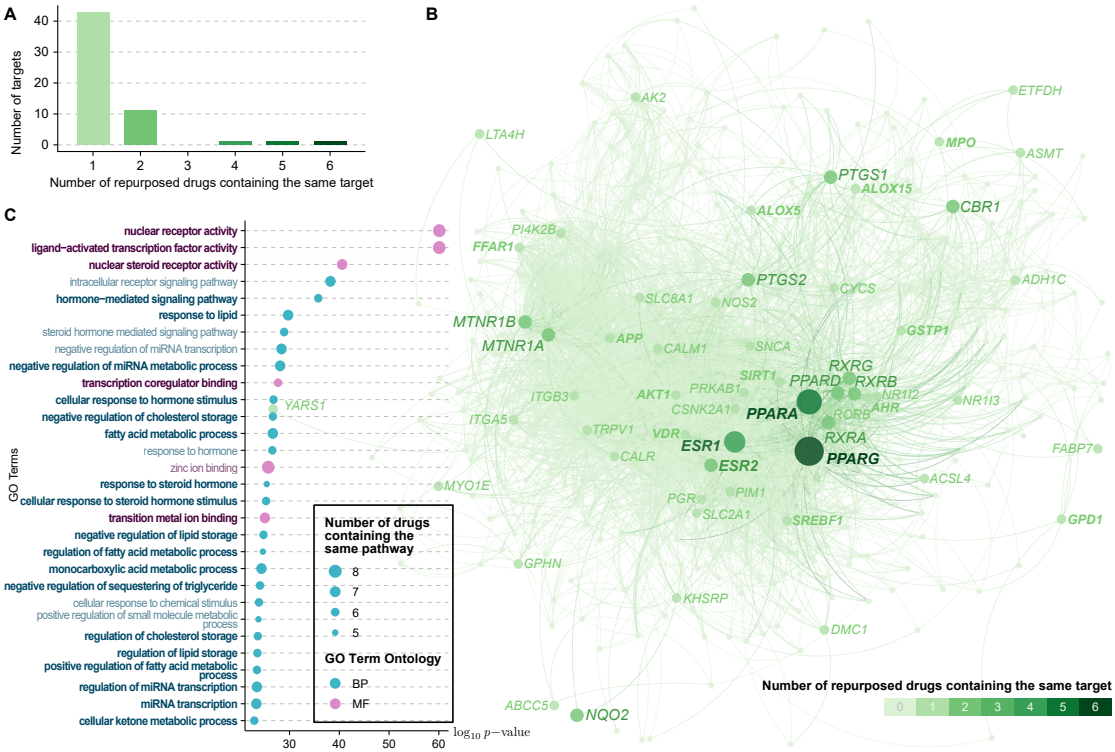


Figure 4. *Labyrinth* captures druggable targets and pathways: fatty liver as a case. (**A**) Bar plot highlighting the frequency of shared targets among repurposed drugs. (**B**) Protein–protein interaction network illustrating connections between fatty liver-associated proteins and repurposed drug targets, with node darkness and size indicating target frequency. The nodes with labels represent the top ten drug targets. The bolder text represents fatty liver related nodes. (**C**) Enrichment analysis of these top drug targets within GO terms. The color represents the subcategory of the GO term. The bolder texts are fatty liver related pathways. Pathways are sorted in descending order of log 10 transformed *p* value.

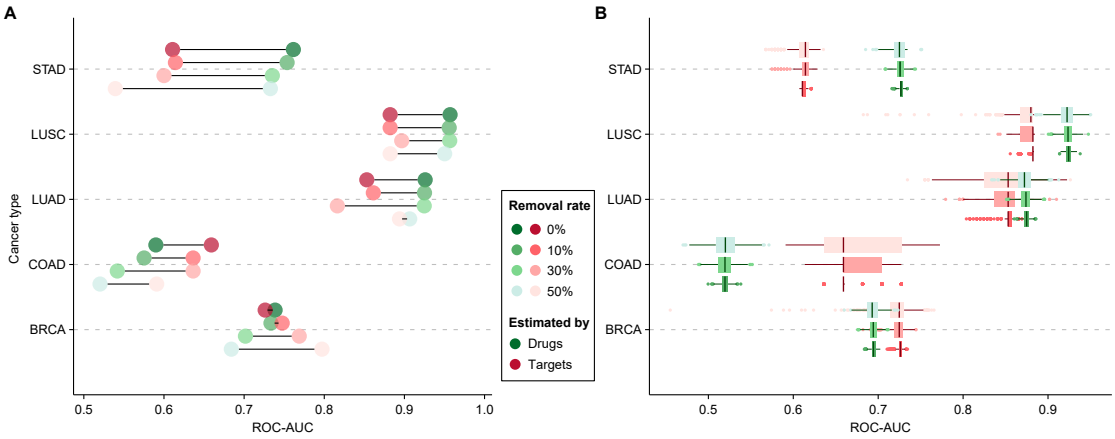


Figure 5. *Labyrinth* is aligned with medical practice and robustness in perturbations. (**A**) Dot plot predictive performance in five cancer types after edge deletions, with color coding for evaluation criteria and varying transparency indicating the last 10%, 30%, and 50% edge removal rates. (**B**) Box plot showing performance with random edge deletions, using the same color and transparency scheme to represent different criteria and deletion rates.

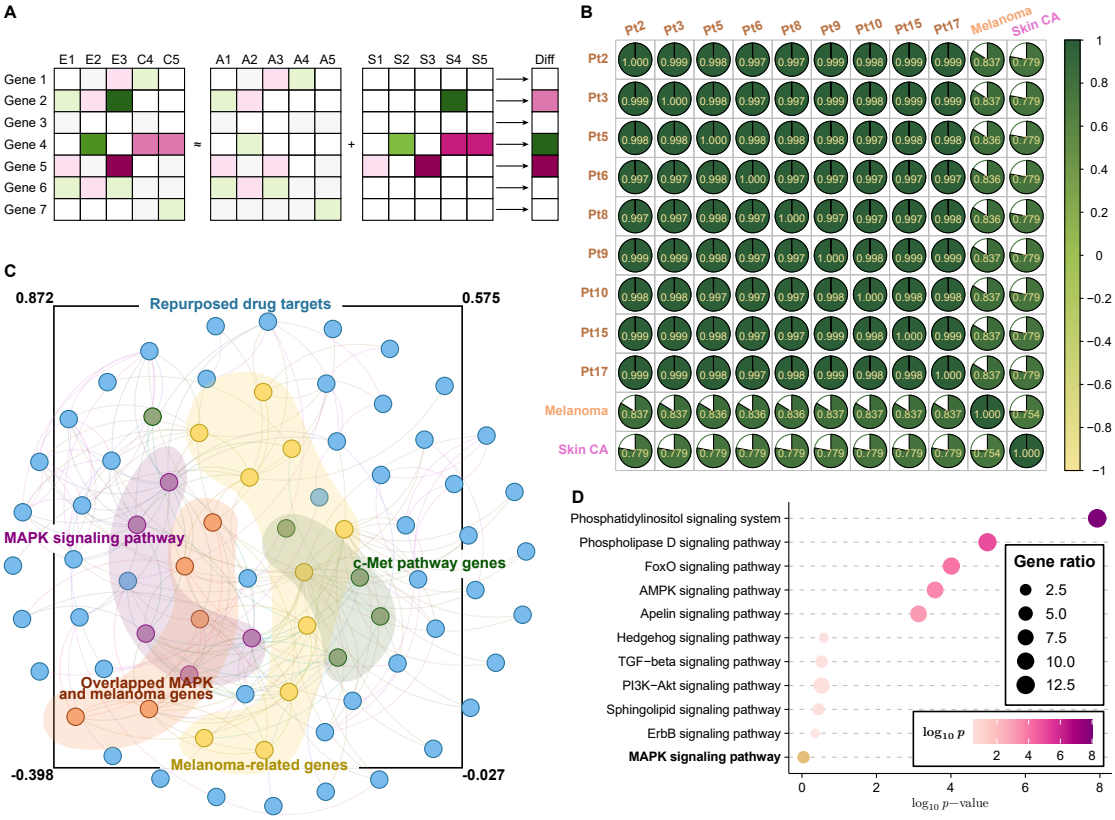


Figure 6. *Labyrinth* can be used in personalized medicine. (**A**) Overview of using RPCA as an alternative to differential analysis. (**B**) Correlation map for nine *MAPK*-resistance melanoma patients, compared to those with standard melanoma and skin cancer. (**C**) The protein–protein interaction network displaying interactions among repurposed drug targets, *MAPK* and c-Met pathways, and melanoma-related genes, focusing on the largest connected component for clarity. (**D**) Enrichment of repurposed targets in the KEGG signaling pathway subcategory. The pathways are sorted by log-transformed *p* values, with color transparency indicating gene ratios within each pathway.