Journal of Holistic Integrative Pharmacy 4 (2023) 199–209

Contents lists available at ScienceDirect

Journal of Holistic Integrative Pharmacy

journal homepage: www.keaipublishing.com/en/journals/journal-of-holistic-integrative-pharmacy

Genomics-based tools for drug discovery and development: From network maps to efficacy prediction

Junhao Fang [a,b], Qi Chen [a,b], Guoyu Wu [a,b,*]

[a] School of Pharmacy, Guangdong Pharmaceutical University, Guangzhou, 510006, China
[b] Center for Drug Research and Development, Guangdong Pharmaceutical University, Guangzhou, 510006, China

ARTICLE INFO

Keywords:
Genomics
Information technology
Drug discovery and development

ABSTRACT

Computing technology plays a crucial role in the field of drug discovery and development. With the rapid development of genomics and the improvement of databases, the application of genomics-based tools is important in drug discovery and development. These tools can deeply explore the information in gene expression profile databases, revealing the connections and interactions between drugs, diseases and genes, and providing strong support for drug discovery and development. This paper introduces various significant genomics-based tools for drug discovery and development, discusses the advantages of deep learning and artificial intelligence in utilizing large-scale genomic data, and reveals the development trends and future prospects of drug genomics tools. The continuous progress of these tools will provide more accurate and efficient support for drug discovery and development.

In the past few decades, the development of high-throughput sequencing technology has led to the improvement of gene expression profile databases, enabling us to measure and analyze large-scale gene expression data. The effective utilization of genomic data plays a crucial role in drug development and disease research. Traditional drug development focuses on studying pathological targets. However, for diseases with unknown targets and most small molecule drugs, we still lack a comprehensive understanding of their mechanisms and interconnections. Exploring the connections between diseases, genes, and drugs is currently a hotspot in scientific research.

Genomics-based tools for drug discovery and development can deeply explore the information in gene expression profile databases, revealing the connections and interactions between drugs, diseases, and genes. These tools assist in drug development by discovering potential drug candidates, thus greatly advancing research progress in the fields of drug development and disease research.

This paper discusses the development process of genomics-based tools for drug discovery and development (Fig. 1, Table 1) and explores how the combination of large-scale genomic data and computer technology will impact the future of drug development.

1. First-generation genomics-based tools for drug discovery and development

Before the advent of high-throughput sequencing technologies, scientists utilized traditional Sanger sequencing, which was slow and costly, making overall genomic studies challenging. The emergence of DNA microarrays provided a simple and natural tool for comprehensive and systematic genomics,[1] empowering high-throughput gene expression analysis and the establishment of expression profile databases.[2] This led to the establishment of various genomics databases, primarily expanding their data using DNA microarray technology (Fig. 2). It should be noted, however, that while this paper distinguishesS between first- and second-generation databases and analysis tools based on their establishment time and sequencing technology, it is important to recognize that databases have continually updated or merged with others. For instance, the first-generation database CMap1 merged into the second-generation database CMap2, while GEO, CCLE, and GDSC remain continuously updated.

1.1. Gene expression Omnibus (GEO)

GEO is a globally shared gene expression database created and maintained by the National Center for Biotechnology Information (NCBI)

* Corresponding author. School of Pharmacy, Guangdong Pharmaceutical University, Guangzhou, 510006, China.
E-mail address: wuguoyu@gdpu.edu.cn (G. Wu).

https://doi.org/10.1016/j.jhip.2023.11.001
Received 18 September 2023; Received in revised form 10 November 2023; Accepted 12 November 2023

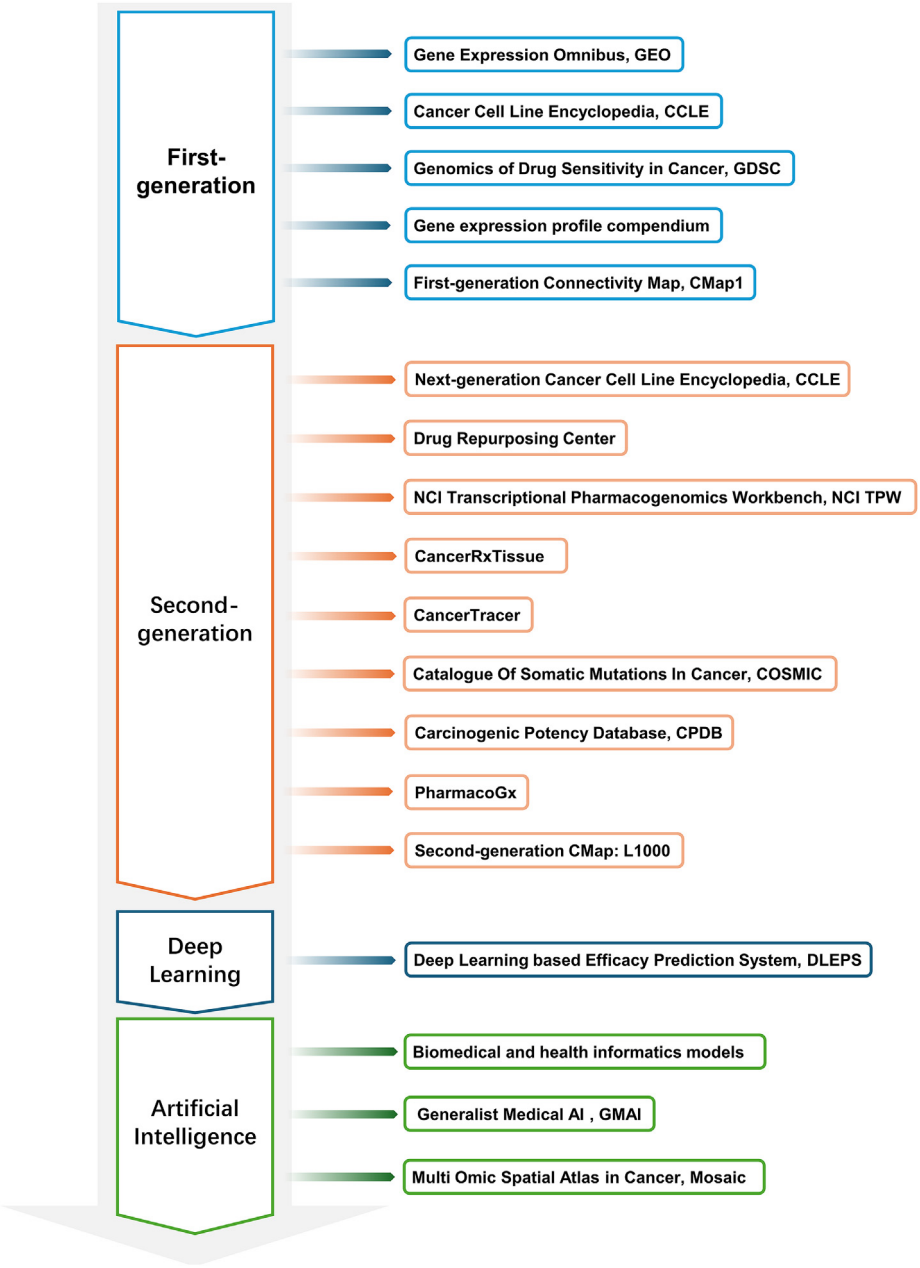

Fig. 1. Schematic showing the development of genomics-based tools.

in 2000, which includes microarray and other high-throughput data.[3] It can be accessed through http://www.ncbi.nlm.nih.gov/geo/and is used for effective and efficient drug repurposing and identification of drug targets/pathways. GEO provides a wide range of gene expression data, including samples from different cancer types and treatment conditions, enabling researchers to conduct large-scale bioinformatics analysis and identify potential biomarkers associated with radiation therapy sensitivity and patient prognosis.[4] Most importantly, GEO not only serves as a database but also simplifies the data analysis process, allowing all researchers to easily utilize the data in their own studies.[5–7] Researchers have developed various functions using GEO microarray datasets to reevaluate disease classification, identify potential drug side effects, and economically and efficiently analyze drug targets or pathways.[8]

1.2. Cancer Cell Line Encyclopedia (CCLE)

The first-generation Cancer Cell Line Encyclopedia (CCLE), completed in collaboration with the Broad Institute, Dana-Farber Cancer Institute, and other research institutes, can be accessed at https://sites.broadinstitute.org/ccle. It accurately characterizes the genetic features of cancer cell lines, including gene expression, chromosomal copy number, and large-scale parallel sequencing data from 947 human cancer cell lines.[9] Additionally, 24 anticancer drugs have been pharmacologically evaluated on 479 cell lines to determine predictors of drug sensitivity based on genetic, lineage, and gene expression factors. By considering gene predictions of drug response, a more personalized approach can be developed, accelerating the emergence of individualized treatment plans[10] and providing valuable insights for the development of new strategies for cancer treatment.[11]

1.3. Genomics of Drug Sensitivity in Cancer (GDSC)

Genomics of Drug Sensitivity in Cancer (GDSC) database (www.cancerRxgene.org) is a freely accessible public resource for studying

Table 1
An overview of databases.

Database/Tool	Description	Datasets	URL link
Gene Expression Omnibus (GEO)[6]	GEO is an international public repository that archives and freely distributes microarray, next-generation sequencing, and other forms of high-throughput functional genomics data. GEO2R is provided as a tool to compare two or more groups of Samples in order to identify genes that are differentially expressed across experimental conditions.	Genetic Data DataSets:4348 Series: 211,302 Platforms:25,476 Samples: 6,762,989	https://www.ncbi.nlm.nih.gov/geo/
Genomics of Drug Sensitivity in Cancer (GDSC)[12]	The GDSC database integrates heterogeneous cancer genomic datasets and anti-cancer drug responses on thousand cancer cell lines. GDSCTools is developed to identify clinically relevant genomic markers of drug response.	Compounds: 621 Dose-response curves: 576,758 Genomic associations tested: 722,057	http://www.cancerrxgene.org/
Gene expression profile compendium[17]	This method provides important tools for revealing cellular functions and interference mechanisms through genomics research, making it a reality to build a feature library of drugs based on expression profiles	–	–
Next-generation Cancer Cell Line Encyclopedia (CCLE) [40]	The Cancer Cell Line Encyclopedia (CCLE) project is an effort to conduct a detailed genetic characterization of a large panel of human cancer cell lines. The CCLE provides public access to genomic data, visualization and analysis for over 1100 cancer cell lines.	Genetic Data : 329 cell lines RNA Expression Data : 1019 cell lines Fusion calls : 1019 cell lines Epigenetic and Histone Modification Data DNA methylation by RRBS CCLE Global Chromatin Profiling data Proteomics Data Metabolomics Data	https://sites.broadinstitute.org/ccle/
Drug Repurposing Hub[41]	The Drug Repurposing Hub is a collection of FDA-approved drugs, drugs undergoing clinical trials, and pre-clinical tool compounds.	Samples: 16,826 Protein Targets: 2183 Unique Compounds: 7934 Drug Indications: 670	https://repo-hub.broadinstitute.org/repurposing
NCI Transcriptional Pharmacodynamics Workbench (NCI TPW)[42]	NCI TPW provides advanced computational and visualization tools for the genome-wide characterization of NCI-60 cell lines and response to 15 different anticancer drugs at different time points.	Genome: NCI-60 human cell lines Compounds: 15 different anticancer drugs	https://tpwb.nci.nih.gov
CancerRxTissue[45]	CancerRxTissue provides predictive models to predict drug sensitivity for both normal and tumor tissues.	Drug Sensitivity Predictions: 272	https://manticore.niehs.nih.gov/cancerRxTissue/
CancerTracer[48]	CancerTracer is a manually curated and integrated database for deciphering tumor heterogeneity at individual patient level.	Tumor Samples: 6000+	http://cailab.labshare.cn/cancertracer
Catalogue of Somatic Mutations in Cancer (COSMIC) [54]	COSMIC covers coding mutations, non-coding mutations, gene fusions, copy-number variants and drug-resistance mutations.	Total genomic variants: 2,385,4105 Genomic non-coding variants: 16,304,701 Genomic mutations within exons: 5,078,567 Samples: 1,520,321 Fusions: 19,428 Whole genome screen samples: 42,519 Copy number variants: 1,207,190 Gene expression variants: 9,215,470 Differentially methylated CpGs: 7,930,489	https://cancer.sanger.ac.uk/cosmic
Carcinogenic Potency Database (CPDB) [55]	CPDB collects results of 6540 chronic, long-term animal cancer tests on 1547 chemicals.	Animal Cancer tests: 6540	https://files.toxplanet.com/cpdb/cpdb.html
PharmacoGx[56]	PharmacoGx is an R package to analyze large-scale pharmacogenomic datasets.	–	https://github.com/bhklab/PharmacoGx
CMap: L1000 platform[60]	L1000 collects perturbation-driven gene expression datasets.	Profiles: 3.02 M Compounds: 33,609 Perturbagens: 81,979 Signatures: 1.16 M 240 cell contexts (12 primary)	https://clue.io/
Deep Learning based Efficacy Prediction System (DLEPS)[83]	DLEPS is a Python package that uses deep learning to predict the efficacy of drugs based on chemical structure, gene expression and target activity.	–	https://www.dleps.tech/
Generalist Medical AI (GMAI)[100]	GMAI is a type of medical AI model that can perform a variety of tasks with minimal or no task-specific labeled data. Through self-supervision on large and diverse datasets, GMAI could interpret various combinations of medical patterns.	–	–
Multi Omic Spatial Atlas in Cancer (Mosaic) [101]	MOSAIC generates multimodal data for a total of 7000 patients in seven cancer indications and develop AI-based analytical tools. Data modalities include spatial and single cell transcriptomics and proteomics, bulk molecular profiling, pathology images, and curated clinical information.	Cancer indications : 7 Patient samples : 7000 Data modalities : 6	https://www.mosaic-research.com/

the relationship between cancer drug sensitivity and genomic characteristics.[12] By analyzing genomic and drug sensitivity data from a large number of cancer cell lines, GDSC aims to facilitate the development of new cancer therapies through the preclinical identification of therapeutic biomarkers. In the GDSC project, researchers use high-throughput techniques to measure drug sensitivity in cancer cell lines for hundreds of drugs. At the same time, they analyze the genomic characteristics of these cell lines, including gene mutations, chromosomal rearrangements, and gene expression data. Through the integration and analysis of these large-scale datasets, GDSC can discover potential associations between

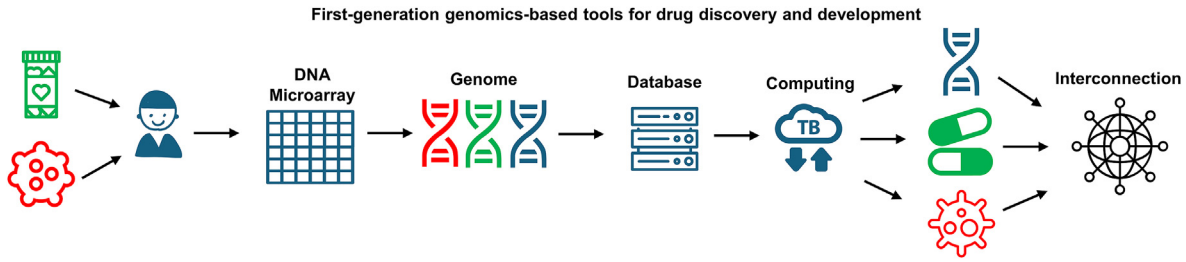

Fig. 2. Schematic depiction of 1st generation genomics-based tools.

the sensitivity of cancer cells to certain drugs and genomic features. These genomics-based data help to better understand the mechanism of action of cancer drugs and provide guidance for precision medicine[13,14] and personalized therapy.[15,16]

1.4. Gene expression profile compendium

In 2000, Hughes, Marton, and others proposed a method called "compendium" to detect the impact of non-characteristic disturbances caused by unknown factors on cell function.[17] They built a comprehensive analysis database by applying drugs to yeast and using DNA microarrays to construct a gene expression profile database. By comparing gene expression profiles, they identified the cellular pathways affected by the disturbance. This method does not rely on specific target proteins or pathways, but uses the pattern of overall transcriptional changes to "fingerprint" cellular processes. It can help understand the effects of various mutations and compounds on cells and has the potential to discover unknown cellular pathways or responses. This method provides important tools for revealing cellular functions and interference mechanisms through genomics research, making it a reality to build a feature library of drugs based on expression profiles.[18]

However, the compendium has its limitations.[19] It requires a large number of rich reference gene profiles and high-quality transcription profiles to support analysis. It has not been tested in mammalian cells, and the interpretation of interactions and associations between multiple transcription changes requires further validation and functional research. But this pioneering work has laid the foundation for the practical application of expression profiles.

1.5. First-generation connectivity map (CMap1)

In 2006, Lamb, Crawford, and others constructed the first-generation CMap, which contains a gene expression profile database of 164 drugs and non-drugs.[20] The first-generation CMap uses microarray technology to measure the gene expression profile of mammalian cell lines after drug treatment, covering hundreds of genes, and adopts a non-parametric, rank-based pattern matching strategy for gene set enrichment analysis (GSEA).[21] CMap can use GSEA analysis to compare the similarity of gene expression profiles between different samples as a whole, revealing biological changes under different conditions and the corresponding gene set enrichment situation.[22] It does not require precise optimization of cell type, concentration, and treatment time, providing a truly universal, systematic, and biologically relevant method.

Researchers can submit a gene profile related to a specific disease, and CMap will compare the submitted gene profile with the expression profile database.[20] Researchers will receive a list of drugs, some of which may have a presumed therapeutic effect on the disease or a known mechanism of action, thereby enhancing the biological understanding of the disease. Submitting the expression profile after drug action can also clarify the mechanism of new drugs.[23,24] This method can reveal the connection between drugs, genes, and diseases. These analysis results are integrated into the Connectivity Map database, becoming a public resource for researchers to query and use.

Drug repurposing is an expensive and challenging method in drug development,[25] yet the first-generation CMap holds significant potential in drug development and drug repurposing.[26,27] Specific examples include a small molecule for alleviating muscular atrophy[28]; parbendazole used for treating osteoporosis[29]; identifying existing drugs for treating colorectal cancer[30]; multiple drugs targeting COX2 and ADRA2A repurposed for treating diabetes[31]; identifying agonists and antagonists of estrogen[32]; personalized treatment for clinical cancer therapy.[33] Common research methods for drug repurposing also include repurposing based on the side effects of existing drugs.[34]

However, the first-generation CMap has some limitations and drawbacks: the database is small in scale and lacks necessary richness; the high cost of commercial gene expression microarrays and RNA sequencing hinders the scale of CMap; in mammalian cell culture, traditional hierarchical clustering methods mainly detect structures related to cell type and similarity, and mask more subtle signals from short-term treatment with small molecules; hierarchical clustering methods require all gene expression profile data to be generated on the same gene chip platform, limiting its future practicality. An analysis method is needed which can detect multiple components of cell response under a given disturbance.

2. Second-generation genomics-based tools for drug discovery and development

Compared to traditional methods, DNA microarrays enable the simultaneous detection of thousands of gene expressions, saving time and cost. However, DNA microarrays are less sensitive to genes with significant expression differences, resulting in a narrower dynamic range.[35,36] In contrast, high-throughput sequencing technology can detect various aspects of gene expression levels, such as RNA expression level, Single Nucleotide Variants (SNVs), Copy Number Variations (CNVs), and Chromatin Immunoprecipitation Sequencing (ChIP-Seq), providing a more comprehensive genomic information.[37] NGS technology surpasses DNA microarrays in sensitivity and dynamic range,[38] bringing comprehensive, high-throughput, and accurate gene expression information, accelerating research on gene functions and regulatory mechanisms, and providing new tools and perspectives for disease diagnosis, drug development, and personalized medicine.[39] Based on NGS technology, second-generation databases emerged by incorporating NGS data onto the existing DNA microarray data, thus initiating a new era (Fig. 3). Second-generation databases are collaborative efforts combining DNA microarrays and NGS technology. Nevertheless, the proportion of NGS data in databases has been increasing over time due to its ability to characterize more information, driven by decreasing prices and continuous technological advancements.

2.1. Next-generation Cancer Cell Line Encyclopedia

The release of the next-generation Cancer Cell Line Encyclopedia has made a significant contribution to cancer research.[40] This version provides more comprehensive and detailed information about the genetic characteristics of cancer cell lines, and explores the associations between these characteristics and phenotypes such as the dependence on specific

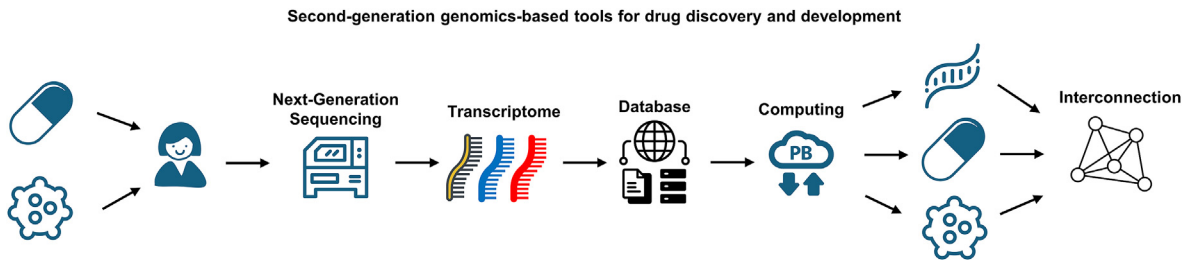

Fig. 3. Schematic depiction of 2nd generation genomics-based tools.

genes and response to drug treatment. Researchers also use the Cancer Dependency Map database to study the effects of individual genes on cancer cell proliferation to determine gene necessity. In addition, the updated encyclopedia covers multi-omics data such as genetic mutations, RNA splicing, DNA methylation, histone modifications, microRNA expression, and protein expression for over a thousand cell lines. This updated information helps to uncover new mechanisms of cancer suppression and may serve as targets for precision therapy. Overall, this version of the Cancer Cell Line Encyclopedia provides a deeper understanding and new research directions for cancer research.

2.2. Drug repurposing center: next-generation drug repository and information resource

The online Drug Repurposing Center (https://repo-hub.broadinstitute.org/repurposing) collects a large number of approved drugs with clinical research and known safety information, and provides detailed information about these drugs, such as drug structure, target interactions, pharmacological parameters, etc.[41] Drug development requires a significant amount of cost and time, and many potential drugs are eliminated in preclinical stages, with some drugs even being recalled after market release. These drugs are valuable resources, and reusing them could save a substantial amount of resources. Previously, large-scale evaluation of drug effects was very challenging, but the development of genomics, such as high-throughput expression profile databases like CMap, has facilitated drug repurposing. This forms the basis for the establishment of the Drug Repurposing Center.

2.3. NCI transcriptional pharmacogenomics Workbench (NCI TPW)

The NCI TPW is a platform (https://tpwb.nci.nih.gov) created by the National Cancer Institute.[42] By measuring the gene expression changes in the NCI-60 cell line panel after exposure to 15 anticancer drugs, common transcriptional responses across drugs and cell types have been identified, along with gene expression changes associated with drug sensitivity. Furthermore, the NCI TPW demonstrates its value in studying clinically relevant molecular hypotheses and identifying candidate biomarkers for drug activity.[43] The NCI TPW allows researchers to explore gene expression regulation through the associations between molecular pathways, drug targets, and drug sensitivity.[44] It provides us with a comprehensive resource that helps deepen our understanding of the sensitivity of common anticancer drugs to tumor cell properties.

2.4. CancerRxTissue

CancerRxTissue (https://manticore.niehs.nih.gov/cancerRxTissue) is a database constructed by researchers using data of gene expression and drug sensitivity in cancer cell lines to build predictive models.[45] The model identifies known interactions between drugs and genes then discovers potential novel drug-gene associations. The predictive models are applied to approximately 17,000 samples from the TCGA and GTEx databases to predict the sensitivity of normal and tumor tissues to drugs. The researchers have also created a website for users to visualize and download their prediction data for research purposes.

Although CancerRxTissue is relatively new, it has already demonstrated its utility, such as predicting pancreatic ductal adenocarcinoma patients who are more sensitive to paclitaxel[46] and predicting drug sensitivity for temozolomide (TMZ).[47] The tool may have significant implications for preclinical drug testing and phase I clinical trial design in future.

2.5. CancerTracer

CancerTracer (http://cailab.labshare.cn/cancertracer) is a manually curated database designed to track and characterize the evolutionary trajectories of tumor growth in individual patients.[48] Researchers have collected over 6000 tumor samples from 1548 patients, covering 45 different types of cancer. CancerTracer integrates clinical and genomic alteration data related to tumor heterogeneity, aiding researchers in better understanding the extent of tumor heterogeneity and its evolution during disease progression. This platform allows users to quickly explore the spatial composition and evolution trajectories of tumor subclones, enabling the identification of major and minor driver gene mutations, such as in the case of the TRACERx project wherein an attempt is made to track the evolution of non-small cell lung cancer and explore the impact of tumor heterogeneity on treatment outcomes.[49]

2.6. Catalogue of somatic mutations in cancer (COSMIC)

The Catalogue of Somatic Mutations in Cancer, COSMIC (https://cancer.sanger.ac.uk/cosmic), was launched in 2004 as a free resource initially focused on curating and displaying somatic mutation data for four genes: BRAF, HRAS, KRAS, and NRAS.[50] COSMIC focuses on curated genes, the preservation of somatic mutation data, and community sharing.[51] The COSMIC database provides researchers with vital information about somatic gene mutations, helping them assess the functional effects of different mutations and gain a deeper understanding of their impact on protein activities.[52] Understanding the functional effects of mutations in specific tumors can reveal insights into the mechanisms of tumor development and potential therapeutic targets.

One of the strengths of COSMIC is its regular release of new versions every few months.[53] COSMIC ensures that newly added cancer genes are published after comprehensive curation of relevant literature. Although there is no recent publication introducing a new version since the last major release,[54] the official website continues to be regularly updated, with the latest version v98 (May 2023).

2.7. Carcinogenic potency database (CPDB)

The Carcinogenic Potency Database, CPDB (https://files.toxplanet.com/cpdb/cpdb.html), is a standardized resource that collects chronic carcinogenicity test results for 45 years.[55] This database aims to collect and compile data on the carcinogenic effects of compounds. Currently, it includes data from 6153 experiments reported in literatures and the technical reports of the National Cancer Institute (NCI)/National Toxicology Program (NTP). CPDB provides information on the strain, gender,

compound administration route, target organs, histopathology, and the authors' evaluation of carcinogenicity for each experiment. It also offers quantitative data on statistical significance, tumor incidence, dose-response curve morphology, experiment duration, duration of compound administration, and dosage rate. With CPDB, researchers can assess the carcinogenic potential of specific chemicals and evaluate their risks to human health.

2.8. PharmacoGx: a computational pharmacogenomics platform

The aforementioned databases contain thousands of expression profile data, but their analysis standards are vague, fragmented, and lack standardized methods of access and analysis, limiting the potential of pharmacogenomics. An open-source R package called PharmacoGx is available on GitHub.

PharmacoGx is a computational pharmacogenomics platform designed to integrate large-scale pharmacogenomics datasets and provide standardized annotation, storage, access, and analysis procedures to foster the development of pharmacogenomics research.[56] The platform's design aims to eliminate biases from different data sources, such as batch effects, differences between analysis platforms, and cell-type-specific variations, to best reveal drug-induced effects. PharmacoGx consists of two fundamental components. The first component is an efficient data structure for storing pharmacological and molecular data, as well as the experimental metadata provided by pharmacogenomics datasets. This storage scheme provides a universal interface, standardizes cell line and drug identifiers, and is easily accessible. It also allows for comparative analysis across different pharmacogenomics datasets. The second component is a set of functions for data manipulation and mining tasks. These functions include bias removal, creating signatures representing drug-induced changes in cell line gene expression, implementing connectivity mapping analysis, and computing connectivity scores to infer the associations between drug-induced features and phenotypes. These functions are not limited to specific datasets and can be performed on different drug perturbation datasets.

What sets PharmacoGx apart is its ability to compare query results from multiple datasets in a unified database. This functionality makes it a powerful tool to assist researchers in pharmacogenomics research,[57] such as screening sensitizers for cancer radiotherapy[58] and studying in vitro drug sensitivity,[59] as well as developing new methods and functionalities to better understand drug-induced effects.

2.9. Second-generation CMap: L1000

In 2017, Subramanian, Narayan et al. established a truly feasible and practical comprehensive CMap using the L1000 chip sequencing platform.[60] The expanded CMap can be used to discover mechanism of action of small molecules and identify novel disease-related compounds (Fig. 4). It has high reproducibility that matches RNA sequencing, with high throughput and low cost, making it suitable for large-scale CMap. The researchers compared L1000 with the standard method of gene expression profile analysis, RNA-seq, and found that L1000 has a high similarity to RNA-seq in gene expression profile analysis. Although they are different analysis platforms, researchers also believe that if there were lower-cost RNA-seq options, RNA-seq would be a better choice. This feature provides the basis for L1000 as a more economical and high-throughput gene expression analysis method and demonstrates its feasibility and potential.

Importantly, the LINCS consortium L1000FWD software application visualizes gene expression features as graphs and can be used to search for similar or opposite features.[61] This study created a network using L1000 and cell viability features, with nodes representing features and organization based on similarity. By changing colors and shapes, similar feature clusters are visually displayed. This provides a global view of gene expression space in six human cell lines.

L1000 has been used for discovering drug mechanisms, such as confirming the moa of two histone deacetylase inhibitors and a topoisomerase inhibitor[62]; drug repurposing, such as the small molecule CGP-60474 as a potent antiseptic[63]; a small molecule for treating cystic fibrosis[64]; and a small molecule for treating melanoma[65]; functional annotation of disease genes[66]; and providing information for clinical trials, such as predicting drug side effects.[67]

Despite its many advantages, L1000 also has some limitations, including lower signal resolution (using only 1000 landmark transcripts), limited coverage, dependence on probe selection, and limitations in robustness for specific cell types.[68] Researchers need to consider these factors in combination with their specific needs and experimental designs when selecting analysis platforms. However, if lower-cost and more accurate sequencing methods become available, their advantages over L1000 would become more apparent, such as TempO-Seq (an upgraded version of L1000) compared to S1500+,[69] as expected.[70]

Like all large community resources, the full potential of CMap can only be realized over time.[71] Its usefulness in elucidating small molecule mechanisms, providing functional readouts of genetic variations, or generating new therapeutic hypotheses remains to be observed. However, the emergence of L1000 has rebuilt a wall for expression profile research, and like GEO, L1000's data can complement other databases.[72]

3. Application of deep learning in drug discovery and development

Deep learning includes many frameworks, which is a branch of machine learning.[73] Deep learning plays a crucial role in interpreting various types of data (Fig. 5). It encompasses various model architectures

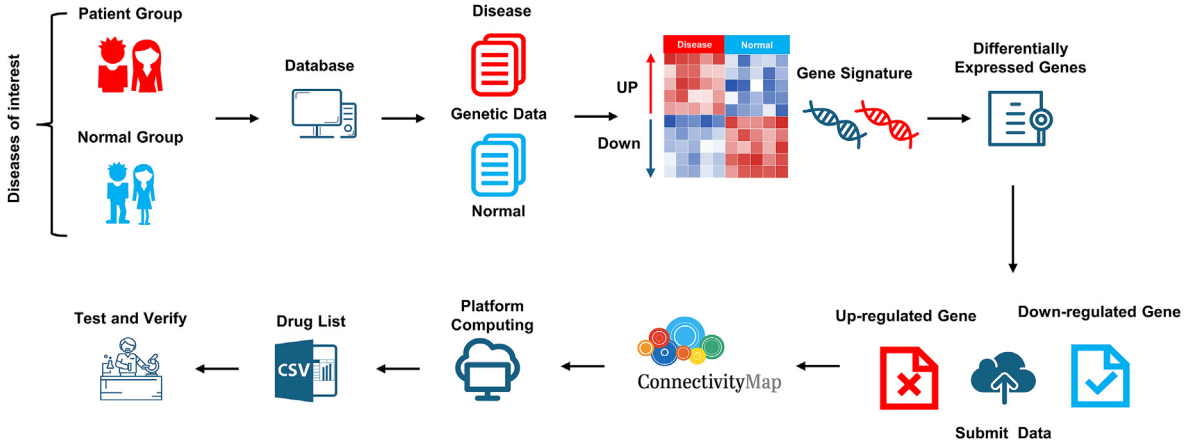

Fig. 4. A workflow to identify disease-related compounds using CMap.

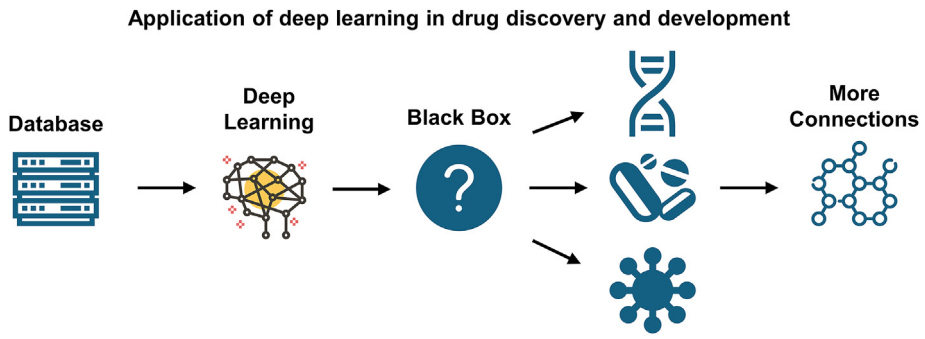

Fig. 5. Schematic depiction of the application of deep learning in drug discovery.

and algorithms, such as convolutional neural networks, recurrent neural networks, and autoencoders, among others. Deep learning has achieved significant advancements in search technology, natural language processing, data mining, data analysis, speech recognition, recommendation and personalization techniques, and other related fields.

Deep learning has demonstrated powerful capabilities in various fields, including image[74] and speech recognition,[75] among others. It has also made promising progress in natural language understanding,[76] such as topic classification, sentiment analysis, question answering, and language translation. The advantage of deep learning is in discovering complex structures in high-dimensional data without the need for manual feature engineering. With further development of learning algorithms and architectures, deep learning is expected to achieve more successes in the future.

3.1. Deep learning in the field of drug research

Thanks to the development of deep neural networks, they can effectively utilize the parallel computing capabilities of modern GPUs.[73] With significant advancements in GPU hardware and increased availability of GPU computing resources, deep learning has great potential in the field of drug research. It can leverage large-scale drug databases to predict various aspects of molecular pharmacology, selectivity, and toxicity in an automated manner. It can help accelerate the screening and optimization process of drug candidates and bring higher efficiency and success rates to drug discovery. Examples include drug structure-activity prediction,[77] computational models for accurately predicting splicing patterns based on genomic features and cellular context to investigate the impact of genetic variations on splicing,[78] predicting the effects of DNA mutations on gene expression and diseases,[79] predicting drug-induced liver injury,[80] rapid identification of effective DDR1 kinase inhibitors,[81] and discovering novel antibiotics using drug repurposing center combined with deep learning,[82] among others, which are changing the traditional drug development process.

3.2. Deep learning based efficacy prediction system (DLEPS)

Traditional drug research focuses on single targets or disease features, while most diseases, drugs, and genes have interrelationships. Combining deep learning with pharmacogenomics enables more accurate and comprehensive drug research and design. In 2021, Zhu J and others trained a new generation of deep learning-based efficacy prediction system (DLEPS) using the transcription profiles of over 20,000 small molecules from the L1000 project.[83] This system uses the feature changes in gene expression profiles under disease states as input to identify candidate drugs. It constructs a universal model through two-stage training, applicable to various diseases, especially for diseases without specific targets. By selecting the feature transcriptomes of disease genes for study and using the input of GSEA gene signatures, DLEPS calculates scores based on epigenomics. Then, corresponding compounds with positive and negative scores are visualized. However, the disadvantages of DLEPS are evident. It relies on the limited data volume of L1000, which is insufficient for deep learning, and DLEPS lacks continuous updates and has a small community. However, it has opened a new direction for combining transcriptional profile databases with deep learning.

3.3. Deep learning throughout the drug development process

Thanks to the development of next-generation sequencing,[84] "omics" technologies such as genomics,[85] epigenomics,[86] and pharmacogenomics have flourished,[87] resulting in exponentially growing databases. The progress in information technology, computer science, and computational biology, combined with the advancement of deep learning, has created a fertile ground for large AI models. Deep learning has already been applied in the healthcare system and can play a role throughout the entire process of drug development, from research to clinical application. For example, deep learning combined with various data types can identify cancer subtypes,[88] predict drug response and synergies,[89] facilitate co-administration,[90] and advance clinical pharmacology.[91] It also has a significant impact in drug repurposing, like the Drug Repurposing Center, which can save a substantial amount of resources for society. Combining drug repurposing with deep learning has also seen new developments.[92,93]

However, despite the remarkable achievements of deep learning in various aspects of drug development,[94] the application of deep learning-driven methods still needs to be translated into standard clinical practice, and the integration of computer simulations throughout the personalized medicine process remains a challenge.[95] Even though deep learning is now applied in various fields, there are still notable "disadvantages",[96] such as the "black-box" nature and the issue of trust in results.

4. Application of AI in drug discovery and development

Various genomics tools discussed earlier have inherited a massive amount of data, and when combined with deep learning, they can revolutionize drug development. These tools contain extensive information on drug compounds, biological activities, pharmacological characteristics, gene characteristics, etc. Through training and learning with deep learning models, researchers can predict the effects, side effects, and safety attributes of drugs. The advantage of this approach is that it can autonomously discover potential patterns and correlations in large-scale data, unveiling knowledge that humans may overlook or find difficult to perceive. By leveraging the computational power of computers and the advantages of deep learning, the drug development process can be accelerated, reducing reliance on human resources and helping discover more effective and safer drugs. However, it should be noted that, in drug development, deep learning models still require empirical data as input, and they inevitably face challenges in data quality and scarcity.[97] Additionally, the results obtained from deep learning models should be considered as reference and assistance, and final decisions still require clinical validation and evaluation by

professionals.

Deep learning is the key technology for establishing foundational models, which were limited in their impact due to the limitations of computer processing power. Today, with the continuous improvement in computational power, AI models have emerged as a logical consequence, such as GPT,[98] which has undergone a qualitative change due to the accumulation of massive data, surpassing our imagination regarding its scale and scope.

AI models refer to algorithms or systems designed to simulate human intelligence and solve specific problems. These models can employ various techniques and methods. Merely learning from individual datasets is insufficient for revealing more connections. By establishing AI models and utilizing multiple larger datasets and more advanced algorithms, more intrinsic connections could be revealed, and novel medicines might be achievable (Fig. 6).

4.1. Biomedical and health informatics models

Models in the field of healthcare and biomedical informatics process multimodal data from multiple sources through training. These sources include healthcare professionals, payment organizations, institutions (such as universities, non-profit organizations, and governments), pharmaceutical companies, wearable devices, and medical publications/forums.[99] Data can take various forms, including images (e.g., chest X-rays), videos (e.g., ultrasound), charts of chemical substances, tables of electronic health records, clinical notes, textual data, time series (e.g., electrocardiograms), and genetic data.

After training, these foundational models can be utilized for various tasks in the healthcare and biomedical fields,[99] such as diagnosis, prognosis, treatment recommendations, drug development, and disease risk prediction. Interestingly, applying models to these tasks generates new data, thereby improving the models and the tasks themselves.

4.2. Generalist Medical AI (GMAI)

In the clinical field, Generalist Medical AI (GMAI) is a type of medical AI model that can perform a variety of tasks with minimal or no task-specific labeled data.[100] Through self-supervision on large and diverse datasets, GMAI flexibly interprets various combinations of medical patterns, including data from imaging, electronic health records, laboratory results, genomics, charts, or medical texts. The model generates expressive outputs in the form of free-text explanations, verbal suggestions, or image annotations, highlighting its advanced medical reasoning capabilities.

4.3. Multi omic spatial atlas in cancer (mosaic)

Mosaic is a collaboration among AI precision medicine companies, spatial biology companies, and leading cancer research institutions including the University of Pittsburgh, Gustave Roussy Cancer Campus (France), and the Lausanne University Hospital (Switzerland).[101] It applies spatial omics technology and artificial intelligence techniques to advanced cancer research (https://www.mosaic-research.com/).[102] This project covers seven types of cancer: non-small cell lung cancer, triple-negative breast cancer, diffuse large B-cell lymphoma, ovarian cancer, glioblastoma, mesothelioma, and bladder cancer, with a total of 7000 patient samples, with 1000 samples for each cancer type. This collaboration leverages the strengths of clinical practice, drug research, and artificial intelligence, providing patient samples, generating high-quality data, developing AI-based analysis tools, and ultimately compiling a resource widely used in medical research.

Spatial omics, benefiting from the development of high-throughput sequencing technology.[103] reveals deeper and more comprehensive decoding of the system. Many studies have used spatial omics to do investigations in deeper levels, such as deciphering spatial organization and intercellular signaling in the microenvironment of skin cancer invasion through the integration of single-cell and spatial transcriptomics.[104]

4.4. Concerns and challenges regarding AI models

Experts in artificial intelligence focus on their respective AI domains, and there are significant gaps between different fields. For non-AI practitioners, it is challenging to establish models that meet their specific needs. For healthcare professionals, building their own models seems almost impossible, especially personalized models for individuals.

Deep learning is a key technology in building the foundational models, but the availability of foundational models is limited. It is challenging for experts in the field of artificial intelligence to collaborate with professionals in different domains, especially for medical researchers who find it almost impossible to establish their own models. The difficulty lies in how experts in artificial intelligence can collaborate with professionals in different fields to establish models that meet their specific needs. Good news is that there are already artificial intelligence experts who have collaborated with professionals in different disciplines, such as models for generating functional protein sequences across different families,[105] models for predicting drug-target interactions,[106] models for predicting drug toxicity,[107] models for compound-protein interaction prediction,[108] DeepConv-DTI model,[109] and MONN model,[110] which is a good start.

Regardless of traditional genomics tools or AI models, it is still necessary to experimentally validate the results obtained. However, the advantages of these tools are quite evident. By leveraging a large amount of data, they can reduce a significant amount of tedious workflow and greatly improve research efficiency.

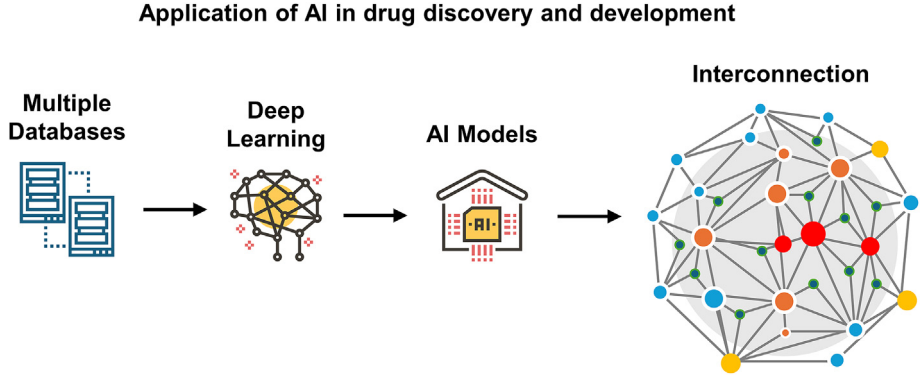

Fig. 6. Schematic depiction of the application of AI in drug discovery.

5. Conclusion and future perspectives

Since its inception, computers have been used to solve complex mathematical calculations and computational tasks. In the past, these calculations required a significant amount of time and manual labor, but the emergence of computers enabled rapid and accurate completion of these tasks. Computers can also store vast amounts of data and efficiently manage and retrieve it, a feat that is difficult for humans to accomplish. Today, the computing power of computers continues to grow exponentially, and technology is constantly advancing through innovative approaches. For example, three-dimensional stacked chips and new computing architectures like CUDA in graphics processing units (GPUs) have become important means of improving computing power.

The development of big data and computing power has made it relatively easier to establish pharmaceutical models compared to a few years ago. However, it remains an enormous challenge for non-artificial intelligence practitioners to build their own models. Perhaps in the future, AI self-generating models will become a reality, where providing data to the model will result in the automatic creation of an appropriate model. Personalized models will be within reach.

From the simple compilation to various transcriptomic and genomic databases, we can use these databases for various visual comparisons and matches to infer the attributes and features associated with diseases, drugs, and genes. This methodology provides important tools and approaches for drug development and disease research. Various genomics databases and tools have revolutionized drug discovery and development. Some novel therapeutic compounds or drugs have been obtained using those databases and tools (Table 2). However, from today's perspective, the data volume of these tools is still far from sufficient, and the databases are too scattered with inconsistent standards. This is also the reason for the emergence of PharmacoGx. Nowadays, we can aggregate a larger volume of data to build a unified model, which may enable analysis of all substances involved in biochemical reactions. Similar to GPT, when data reaches an unimaginable quantity, it triggers a qualitative change. Biochemical reactions may have subtle variations that are difficult for a human eye to discern, but AI can discover different patterns based on a large amount of data. Currently, AI does not need to transform complex structures like chemical formulas or proteins into simpler formats to understand them.

Although establishing AI models has become relatively simpler with the advancement of big data and computing power, building models still requires high costs. Apart from the cost aspect, there are many challenges in the pharmaceutical industry. Laboratories are often reluctant to share their data, including chemical compositions of drugs, pharmacological and toxicological experiments, animal experiments, and clinical trial results. These data are protected by patents, even though many classic drugs no longer have patents. Fortunately, some people advocate for community-based sharing, as seen in various genomics platforms, most of which are free to use. As productivity advances, we believe that information barriers will become lower. The greatest challenge in constructing models lies in data acquisition and processing. Finding better solutions to this problem is the challenge faced by leveraging information technology to assist drug development. Data-sharing systems enable researchers around the world to share genomic information. Traditionally, most of the data were collected and studied in silos and domain-specific. The type, format, content, or disciplinary focus of the data could be different. It is vital to developing international standards and principles for data sharing ensure a high data quality. And a faster, cheaper, and more accessible system is required to breaking down information barriers.

CRediT authorship contribution statement

Junhao Fang: Writing – original draft, Investigation. **Qi Chen:** Writing – original draft, Investigation. **Guoyu Wu:** Writing – review & editing, Supervision, Investigation, Conceptualization.

Table 2
Examples of drugs obtained using the database.

Database	Compounds	Disease	Reference
CMap	Tomatidine	Skeletal muscle atrophy	28
CMap	Parbendazole	Osteoporosis	29
GEO and CMap	10 potential compounds: trichostatin A, vorinostat, HC toxin, sodium phenylbutyrate, mycophenolic acid, irinotecan, etoposide, valproic acid, arachidonic acid, rifabutin	Colorectal cancer	30
The human metabolome database and CMap	4 drugs targeting COX2: diflunisal, nabumetone, niflumic acid and valdecoxib. 2 drugs targeting ADRA2A: phenoxybenzamine and Idazoxan.	Type 2 diabetes	31
GEO and CMap	KM-00927 and BRD-K75081836	Cancer	62
GEO and CMap	CGP-60474	Endotoxemia	63
GEO and CMap	An inhibitor of src-kinase: PP-2	Cystic fibrosis	64
GEO and CMap	Cinnarizine, digitoxigenin, and clofazimine	Metastatic uveal melanoma	65
The Drug Repurposing Hub and deep learning	Halicin	A wide phylogenetic spectrum of pathogens including Mycobacterium tuberculosis and carbapenem-resistant Enterobacteriaceae	82
The Drug Repurposing Hub	BRD4780	Mucin-1 kidney disease	111
The Drug Repurposing Hub and CCLE	Disulfiram, tepoxalin	Cancer	112
GEO and CMap	Oxytocin, fluoxetine, saquinavir and ribavirin	Sars-CoV-2	113
TCGA and PharmacoGx	Guanidine hydrochloride	Triple-negative breast cancer	114
The Drug Repurposing Hub	Obatoclax	Sars-CoV-2	115
The Drug Repurposing Hub and CCLE	CR8	CDK inhibitor: depletes cyclin k	116
DLEPS	Ataluren	Osteoporosis	117

Declaration of competing interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgements

This research was funded by National Natural Science Foundation of China (32200513); This study was also supported by Medical Scientific Research Foundation of Guangdong Province of China (A2022182). This study was also supported by Basic and Applied Basic Research Foundation of Guangzhou Municipal Science and Technology Bureau (2023A04J1140).

References

1. Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays. *Nat Genet*. 1999;21(1):33–37.
2. Brazma A, Robinson A, Cameron G, Ashburner M. One-stop shop for microarray data. *Nature*. 2000;403(6771):699–700.
3. Barrett T, Troup DB, Wilhite SE, et al. NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Res*. 2007;35(Database issue):D760–D765.
4. Meng J, Li P, Zhang Q, Yang Z, Fu S. A radiosensitivity gene signature in predicting glioma prognostic via EMT pathway. *Oncotarget*. 2014;5(13):4683–4693.
5. Ni M, Ye F, Zhu J, et al. ExpTreeDB: web-based query and visualization of manually annotated gene expression profiling experiments of human and mouse from GEO. *Bioinformatics*. 2014;30(23):3379–3386.
6. Clough E, Barrett T. The gene expression omnibus database. *Methods Mol Biol*. 2016;1418:93–110.
7. Li J, Zhang Y, Gao Y, et al. Downregulation of HNF1 homeobox B is associated with drug resistance in ovarian cancer. *Oncol Rep*. 2014;32(3):979–988.
8. Hu G, Agarwal P. Human disease-drug network based on genomic expression profiles. *PLoS One*. 2009;4(8):e6536.
9. Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012; 483(7391):603–607.
10. Covell DG. Data mining approaches for genomic biomarker development: applications using drug screening data from the cancer genome project and the cancer cell line encyclopedia. *PLoS One*. 2015;10(7):e0127433.
11. Pavel AB, Korolev KS. Genetic load makes cancer cells more sensitive to common drugs: evidence from Cancer Cell Line Encyclopedia. *Sci Rep*. 2017;7(1):1938.
12. Yang W, Soares J, Greninger P, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res*. 2013;41(Database issue):D955–D961.
13. Garnett MJ, McDermott U. The evolving role of cancer cell line-based screens to define the impact of cancer genomes on drug response. *Curr Opin Genet Dev*. 2014; 24(100):114–119.
14. Wang L, McLeod HL, Weinshilboum RM. Genomics and drug response. *N Engl J Med*. 2011;364(12):1144–1153.
15. Weinshilboum RM, Wang L. Pharmacogenetics and pharmacogenomics: development, science, and translation. *Annu Rev Genom Hum Genet*. 2006;7: 223–245.
16. Garnett MJ, Edelman EJ, Heidorn SJ, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*. 2012;483(7391):570–575.
17. Hughes TR, Marton MJ, Jones AR, et al. Functional discovery via a compendium of expression profiles. *Cell*. 2000;102(1):109–126.
18. Gunther EC, Stone DJ, Gerwien RW, Bento P, Heyes MP. Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro. *Proc Natl Acad Sci U S A*. 2003;100(16):9608–9613.
19. Stoughton RB, Friend SH. How molecular profiling could revolutionize drug discovery. *Nat Rev Drug Discov*. 2005;4(4):345–350.
20. Lamb J, Crawford ED, Peck D, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*. 2006;313(5795): 1929–1935.
21. Mootha VK, Lindgren CM, Eriksson KF, et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*. 2003;34(3):267–273.
22. Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545–15550.
23. Iorio F, Bosotti R, Scacheri E, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc Natl Acad Sci USA*. 2010;107(33): 14621–14626.
24. Isik Z, Baldow C, Cannistraci CV, Schroeder M. Drug target prioritization by perturbed gene expression and network information. *Sci Rep*. 2015;5:17417.
25. Chong CR, Sullivan Jr DJ. New uses for old drugs. *Nature*. 2007;448(7154): 645–646.
26. Lussier YA, Chen JL. The emergence of genome-based drug repositioning. *Sci Transl Med*. 2011;3(96):96ps35.
27. Zhang L, Kang W, Lu X, Ma S, Dong L, Zou B. Weighted gene co-expression network analysis and connectivity map identifies lovastatin as a treatment option of gastric cancer by inhibiting HDAC2. *Gene*. 2019;681:15–25.
28. Dyle MC, Ebert SM, Cook DP, et al. Systems-based discovery of tomatidine as a natural small molecule inhibitor of skeletal muscle atrophy. *J Biol Chem*. 2014; 289(21):14913–14924.
29. Brum AM, van de Peppel J, van der Leije CS, et al. Connectivity Map-based discovery of parbendazole reveals targetable human osteogenic pathway. *Proc Natl Acad Sci U S A*. 2015;112(41):12711–12716.
30. Wen Q, O'Reilly P, Dunne PD, et al. Connectivity mapping using a combined gene signature from multiple colorectal cancer datasets identified candidate drugs including existing chemotherapies. *BMC Syst Biol*. 2015;9(Suppl 5):S4. Suppl 5.
31. Zhang M, Luo H, Xi Z, Rogaeva E. Drug repositioning for diabetes based on 'omics' data mining. *PLoS One*. 2015;10(5):e0126082.
32. Frasor J, Stossi F, Danes JM, Komm B, Lyttle CR, Katzenellenbogen BS. Selective estrogen receptor modulators: discrimination of agonistic versus antagonistic activities by gene expression profiling in breast cancer cells. *Cancer Res*. 2004; 64(4):1522–1533.
33. Lim SM, Lim JY, Cho JY. Targeted therapy in gastric cancer: personalizing cancer treatment based on patient genome. *World J Gastroenterol*. 2014;20(8):2042–2050.
34. Nugent T, Plachouras V, Leidner JL. Computational drug repositioning based on side-effects mined from social media. *PeerJ Computer Science*. 2016;2.
35. Okoniewski MJ, Miller CJ. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinf*. 2006;7:276.
36. Shi L, Jones WD, Jensen RV, et al. The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies. *BMC Bioinf*. 2008;9(Suppl 9):S10. Suppl 9.
37. Guttmacher AE, Collins FS. Realizing the promise of genomics in biomedical research. *JAMA*. 2005;294(11):1399–1402.
38. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10(1):57–63.
39. Marguerat S, Bahler J. RNA-seq: from technology to biology. *Cell Mol Life Sci*. 2010; 67(4):569–579.
40. Ghandi M, Huang FW, Jane-Valbuena J, et al. Next-generation characterization of the cancer cell line encyclopedia. *Nature*. 2019;569(7757):503–508.
41. Corsello SM, Bittker JA, Liu Z, et al. The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat Med*. 2017;23(4):405–408. eng.
42. Monks A, Zhao Y, Hose C, et al. The NCI transcriptional pharmacodynamics workbench: a tool to examine dynamic expression profiling of therapeutic response in the NCI-60 cell line panel. *Cancer Res*. 2018;78(24):6807–6817.
43. Min DJ, Zhao Y, Monks A, et al. Identification of pharmacodynamic biomarkers and common molecular mechanisms of response to genotoxic agents in cancer cell lines. *Cancer Chemother Pharmacol*. 2019;84(4):771–780.
44. Bell CR, Pelly VS, Moeini A, et al. Chemotherapy-induced COX-2 upregulation by cancer cells defines their inflammatory properties and limits the efficacy of chemoimmunotherapy combinations. *Nat Commun*. 2022;13(1):2063.
45. Li Y, Umbach DM, Krahn JM, Shats I, Li X, Li L. Predicting tumor response to drugs based on gene-expression biomarkers of sensitivity learned from cancer cell lines. *BMC Genom*. 2021;22(1):272.
46. Hamdan FH, Abdelrahman AM, Kutschat AP, et al. Interactive enhancer hubs (iHUBs) mediate transcriptional reprogramming and adaptive resistance in pancreatic cancer. *Gut*. 2023;72(6):1174–1185.
47. Kaluzinska-Kolat Z, Kolat D, Kosla K, Pluciennik E, Bednarek AK. Molecular landscapes of glioblastoma cell lines revealed a group of patients that do not benefit from WWOX tumor suppressor expression. *Front Neurosci*. 2023;17:1260409.
48. Wang C, Yang J, Luo H, et al. CancerTracer: a curated database for intrapatient tumor heterogeneity. *Nucleic Acids Res*. 2020;48(D1):D797–D806.
49. Jamal-Hanjani M, Hackshaw A, Ngai Y, et al. Tracking genomic cancer evolution for precision medicine: the lung TRACERx study. *PLoS Biol*. 2014;12(7):e1001906.
50. Bamford S, Dawson E, Forbes S, et al. The COSMIC (Catalogue of somatic mutations in cancer) database and website. *Br J Cancer*. 2004;91(2):355–358.
51. Forbes S, Clements J, Dawson E, et al. Cosmic 2005. *Br J Cancer*. 2006;94(2): 318–322.
52. Alvarez MJ, Shen Y, Giorgi FM, et al. Functional characterization of somatic mutations in cancer using network-based inference of protein activity. *Nat Genet*. 2016;48(8):838–847.
53. Forbes SA, Beare D, Bindal N, et al. COSMIC: high-resolution cancer genetics using the Catalogue of somatic mutations in cancer. *Curr Protoc Hum Genet*. 2016;91, 10 11 11-10 11 37.
54. Tate JG, Bamford S, Jubb HC, et al. COSMIC: the Catalogue of somatic mutations in cancer. *Nucleic Acids Res*. 2019;47(D1):D941–D947.
55. Fitzpatrick RB. CPDB: carcinogenic potency database. *Med Ref Serv Q*. 2008;27(3): 303–311.
56. Smirnov P, Safikhani Z, El-Hachem N, et al. PharmacoGx: an R package for analysis of large pharmacogenomic datasets. *Bioinformatics*. 2016;32(8):1244–1246.
57. Smirnov P, Kofia V, Maru A, et al. PharmacoDB: an integrative database for mining in vitro anticancer drug screening studies. *Nucleic Acids Res*. 2018;46(D1): D994–D1002.
58. Smith I, Bell R, Lambie M, Haibe-Kains B, Bratman S. Abstract PO-071: characterizing transcriptomic indicators of radiosensitivity in cancer and identifying sensitizing therapeutic agents. *Clin Cancer Res*. 2021;27(8_suppl ment). PO-071-PO-071.
59. Yao F, Madani Tonekaboni SA, Safikhani Z, et al. Tissue specificity of in vitro drug sensitivity. *J Am Med Inf Assoc*. 2018;25(2):158–166.
60. Subramanian A, Narayan R, Corsello SM, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*. 2017;171(6):1437–1452 e1417.
61. Wang Z, Lachmann A, Keenan AB, Ma'ayan A. L1000FWD: fireworks visualization of drug-induced transcriptomic signatures. *Bioinformatics*. 2018;34(12):2150–2152.
62. Liu TP, Hsieh YY, Chou CJ, Yang PM. Systematic polypharmacology and drug repurposing via an integrated L1000-based Connectivity Map database mining. *R Soc Open Sci*. 2018;5(11):181321.
63. Han HW, Hahn S, Jeong HY, et al. LINCS L1000 dataset-based repositioning of CGP-60474 as a highly potent anti-endotoxemic agent. *Sci Rep*. 2018;8(1):14969.
64. Wang Y, Arora K, Yang F, et al. PP-2, a src-kinase inhibitor, is a potential corrector for F508del-CFTR in cystic fibrosis. *bioRxiv*. 2018:288324.
65. Fagone P, Caltabiano R, Russo A, et al. Identification of novel chemotherapeutic strategies for metastatic uveal melanoma. *Sci Rep*. 2017;7:44564.
66. Berger AH, Brooks AN, Wu X, et al. High-throughput phenotyping of lung cancer somatic mutations. *Cancer Cell*. 2016;30(2):214–228.
67. Wang Z, Clark NR, Ma'ayan A. Drug-induced adverse events prediction with the LINCS L1000 data. *Bioinformatics*. 2016;32(15):2338–2345.
68. Musa A, Ghoraie LS, Zhang SD, et al. A review of connectivity map and computational approaches in pharmacogenomics. *Briefings Bioinf*. 2018;19(3): 506–523.

69. Mav D, Shah RR, Howard BE, et al. A hybrid gene selection approach to create the S1500+ targeted gene sets for use in high-throughput transcriptomics. *PLoS One*. 2018;13(2):e0191105.
70. Bushel PR, Paules RS, Auerbach SS. A comparison of the TempO-seq S1500+ platform to RNA-seq and microarray using rat liver mode of action samples. *Front Genet*. 2018;9:485.
71. Keenan AB, Wojciechowicz ML, Wang Z, et al. Connectivity mapping: methods and applications. *Annual Rev Biomed Data Sci*. 2019;2(1):69–92.
72. Duan Q, Reid SP, Clark NR, et al. L1000CDS(2): LINCS L1000 characteristic direction signatures search engine. *NPJ Syst Biol Appl*. 2016;2:16015.
73. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436–444.
74. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM*. 2017;60(6):84–90.
75. Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process Mag*. 2012;29(6):82–97.
76. 2018. Improving Language Understanding by Generative Pre-training. In: 2018.
77. Ma J, Sheridan RP, Liaw A, Dahl GE, Svetnik V. Deep neural nets as a method for quantitative structure-activity relationships. *J Chem Inf Model*. 2015;55(2):263–274.
78. Leung MK, Xiong HY, Lee LJ, Frey BJ. Deep learning of the tissue-regulated splicing code. *Bioinformatics*. 2014;30(12):i121–i129.
79. Xiong HY, Alipanahi B, Lee LJ, et al. RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science*. 2015;347(6218):1254806.
80. Xu Y, Dai Z, Chen F, Gao S, Pei J, Lai L. Deep learning for drug-induced liver injury. *J Chem Inf Model*. 2015;55(10):2085–2093.
81. Zhavoronkov A, Ivanenkov YA, Aliper A, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat Biotechnol*. 2019;37(9):1038–1040.
82. Stokes JM, Yang K, Swanson K, et al. A deep learning approach to antibiotic discovery. *Cell*. 2020;180(4):688–702 e613.
83. Zhu J, Wang J, Wang X, et al. Prediction of drug efficacy from transcriptional profiles with deep learning. *Nat Biotechnol*. 2021;39(11):1444–1452.
84. Mutz KO, Heilkenbrinker A, Lonne M, Walter JG, Stahl F. Transcriptome analysis using next-generation sequencing. *Curr Opin Biotechnol*. 2013;24(1):22–30.
85. Malone ER, Oliva M, Sabatini PJB, Stockley TL, Siu LL. Molecular profiling for precision cancer therapies. *Genome Med*. 2020;12(1):8.
86. Mensaert K, Denil S, Trooskens G, Van Criekinge W, Thas O, De Meyer T. Next-generation technologies and data analytical approaches for epigenomics. *Environ Mol Mutagen*. 2014;55(3):155–170.
87. Relling MV, Evans WE. Pharmacogenomics in the clinic. *Nature*. 2015;526(7573):343–350.
88. Guinney J, Dienstmann R, Wang X, et al. The consensus molecular subtypes of colorectal cancer. *Nat Med*. 2015;21(11):1350–1356.
89. Kuenzi BM, Park J, Fong SH, et al. Predicting drug response and synergy using a deep learning model of human cancer cells. *Cancer Cell*. 2020;38(5):672–684 e676.
90. Menden MP, Wang D, Mason MJ, et al. Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat Commun*. 2019;10(1):2674.
91. Zhavoronkov A, Vanhaelen Q, Oprea TI. Will artificial intelligence for drug discovery impact clinical pharmacology? *Clin Pharmacol Ther*. 2020;107(4):780–785.
92. Yella JK, Yaddanapudi S, Wang Y, Jegga AG. Changing trends in computational drug repositioning. *Pharmaceuticals*. 2018;11(2).
93. Lazarczyk M, Duda K, Mickael ME, et al. Adera2.0: a drug repurposing workflow for neuroimmunological investigations using neural networks. *Molecules*. 2022;27(19).
94. Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development. *Drug Discov Today*. 2021;26(1):80–93.
95. Boniolo F, Dorigatti E, Ohnmacht AJ, Saur D, Schubert B, Menden MP. Artificial intelligence in early drug discovery enabling precision medicine. *Expet Opin Drug Discov*. 2021;16(9):991–1007.
96. Marcus G. *Deep Learning: A Critical Appraisal*. 2018. *arXiv preprint arXiv:180100631*.
97. Bommasani R, Hudson DA, Adeli E, et al. *On the Opportunities and Risks of Foundation Models*. 2021. *arXiv preprint arXiv:210807258*.
98. Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. 2020;33:1877–1901.
99. Dash S, Acharya BR, Mittal M, Abraham A, Kelemen A. *Deep Learning Techniques for Biomedical and Health Informatics*. Springer; 2020.
100. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. *Nature*. 2023;616(7956):259–265.
101. Lehar J, Madissoon E, Chevallier J, et al. MOSAIC: multi-omic spatial atlas in cancer, effect on precision oncology. *J Clin Oncol*. 2023;41(16_suppl l):e15076-e15076.
102. Lewis SM, Asselin-Labat ML, Nguyen Q, et al. Spatial omics and multiplexed imaging to explore cancer biology. *Nat Methods*. 2021;18(9):997–1012.
103. Wu Y, Cheng Y, Wang X, Fan J, Gao Q. Spatial omics: navigating to the golden era of cancer research. *Clin Transl Med*. 2022;12(1):e696.
104. Yerly L, Pich-Bavastro C, Di Domizio J, et al. Integrated multi-omics reveals cellular and molecular interactions governing the invasive niche of basal cell carcinoma. *Nat Commun*. 2022;13(1):4897.
105. Madani A, Krause B, Greene ER, et al. Large language models generate functional protein sequences across diverse families. *Nat Biotechnol*. 2023;41(8):1099–1106.
106. Chen W, Chen G, Zhao L, Chen CY. Predicting drug-target interactions with deep-embedding learning of graphs and sequences. *J Phys Chem A*. 2021;125(25):5633–5642.
107. Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: toxicity prediction using deep learning. *Front Environ Sci*. 2016;3.
108. Tsubaki M, Tomii K, Sese J. Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*. 2019;35(2):309–318.
109. Lee I, Keum J, Nam H. DeepConv-DTI: prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS Comput Biol*. 2019;15(6):e1007129.
110. Li S, Wan F, Shu H, Jiang T, Zhao D, Zeng J. MONN: a multi-objective neural network for predicting compound-protein interactions and affinities. *Cell Sys*. 2020;10(4):308–322.e311.
111. Dvela-Levitt M, Kost-Alimova M, Emani M, et al. Small molecule targets TMED9 and promotes lysosomal degradation to reverse proteinopathy. *Cell*. 2019;178(3):521–535. e523. eng.
112. Corsello SM, Nagari RT, Spangler RD, et al. Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nat Can (Ott)*. 2020;1(2):235–248.
113. Imami AS, McCullumsmith RE, O'Donovan SM. Strategies to identify candidate repurposable drugs: COVID-19 treatment as a case example. *Transl Psychiatry*. 2021;11(1):591.
114. Bakhsh MR, Rouhi L, Ghaedi K, Hashemi M, Peymani M, Samarghandian S. Therapeutic effects of guanidine hydrochloride on breast cancer through targeting KCNG1 gene. *Biomed Pharmacother*. 2023;164:114982.
115. Patten JJ, Keiser PT, Gysi D, et al. *Identification of Druggable Host Targets Needed for SARS-CoV-2 Infection by Combined Pharmacological Evaluation and Cellular Network Directed Prioritization Both in Vitro and in Vivo*. bioRxiv; 2022.
116. Slabicki M, Kozicka Z, Petzold G, et al. The CDK inhibitor CR8 acts as a molecular glue degrader that depletes cyclin K. *Nature*. 2020;585(7824):293–297.
117. Zeng L, Gu R, Li W, et al. Ataluren prevented bone loss induced by ovariectomy and aging in mice through the BMP-SMAD signaling pathway. *Biomed Pharmacother*. 2023;166:115332.