OXFORD

Briefings in Bioinformatics, 2025, **26(1)**, bbae659

https://doi.org/10.1093/bib/bbae659

Problem Solving Protocol

CDCM: a correlation-dependent connectivity map approach to rapidly screen drugs during outbreaks of infectious diseases

Junlei Liao[1,‡], Hongyang Yi[2,‡,*], Hao Wang[3,‡], Sumei Yang[2], Duanmei Jiang[1], Xin Huang[3], Mingxia Zhang[2], Jiayin Shen[2], Hongzhou Lu[2,*], Yuanling Niu[1,*]

[1]School of Mathematics and Statistics, HNP-LAMA, Central South University, Changsha 410083, Hunan, China
[2]National Clinical Research Centre for Infectious Diseases, The Third People's Hospital of Shenzhen and The Second Affiliated Hospital of Southern University of Science and Technology, Shenzhen 518112, China
[3]Maternal-Fetal Medicine Institute, Department of Obstetrics and Gynaecology, Shenzhen Baoan Women's and Children's Hospital, Shenzhen 518133, China

*Corresponding authors. Hongyang Yi, National Clinical Research Centre for Infectious Diseases, The Third People's Hospital of Shenzhen and The Second Affiliated Hospital of Southern University of Science and Technology, Shenzhen 518112, China. E-mail: yihy2018@mail.sustech.edu.cn; Hongzhou Lu, National Clinical Research Centre for Infectious Diseases, The Third People's Hospital of Shenzhen and The Second Affiliated Hospital of Southern University of Science and Technology, Shenzhen 518112, China. E-mail: luhongzhou@fudan.edu.cn; Yuanling Niu, School of Mathematics and Statistics, HNP-LAMA, Central South University, Changsha 410083, Hunan, China. E-mail: yuanlingniu@csu.edu.cn
‡Junlei Liao, Hongyang Yi and Hao Wang contributed equally to this work.

Abstract

In the context of the global damage caused by coronavirus disease 2019 (COVID-19) and the emergence of the monkeypox virus (MPXV) outbreak as a public health emergency of international concern, research into methods that can rapidly test potential therapeutics during an outbreak of a new infectious disease is urgently needed. Computational drug discovery is an effective way to solve such problems. The existence of various large open databases has mitigated the time and resource consumption of traditional drug development and improved the speed of drug discovery. However, the diversity of cell lines used in various databases remains limited, and previous drug discovery methods are ineffective for cross-cell prediction. In this study, we propose a correlation-dependent connectivity map (CDCM) to achieve cross-cell predictions of drug similarity. The CDCM mainly identifies drug–drug or disease–drug relationships from the perspective of gene networks by exploring the correlation changes between genes and identifying similarities in the effects of drugs or diseases on gene expression. We validated the CDCM on multiple datasets and found that it performed well for drug identification across cell lines. A comparison with the Connectivity Map revealed that our method was more stable and performed better across different cell lines. In the application of the CDCM to COVID-19 and MPXV data, the predictions of potential therapeutic compounds for COVID-19 were consistent with several previous studies, and most of the predicted drugs were found to be experimentally effective against MPXV. This result confirms the practical value of the CDCM. With the ability to predict across cell lines, the CDCM outperforms the Connectivity Map, and it has wider application prospects and a reduced cost of use.

Keywords: connectivity map (CMap); correlation-dependent connectivity map (CDCM); SARS-CoV-2; monkeypox virus (MPXV); drug function prediction; breakthrough cell line boundary

Introduction

Sudden outbreaks of infectious diseases often become serious public health events. For example, the global coronavirus disease 2019 (COVID-19) pandemic has led to enormous social and economic devastation [1–4]. Monkeypox virus (MPXV) cases continue to occur in nonendemic countries and have been declared a "Public Health Emergency of International Concern" by the World Health Organization [5, 6]. However, few laboratories worldwide are qualified to study highly pathogenic infectious diseases. Thus, human society is currently unable to respond quickly to such infectious diseases, and it is necessary to develop efficient methods to address such diseases. Improvements in drug discovery methods are necessary to develop effective therapies to mitigate the damage caused by emerging infectious diseases and prepare for rapid solutions to future global public health crises caused by these diseases.

Computational methods have mitigated the high cost and failure rates that pharmaceutical research and development have long faced and have been used effectively in many drug discovery studies in recent years [7–11]. The Connectivity Map (CMap) is a useful tool for identifying novel drugs at the transcriptome level, which was made possible by the rapid development of computational methods for analyzing drug perturbation datasets [12]. The main concept of CMap is that functionally similar compounds can induce similar changes in transcriptome expression in a given cell line. To address the limitation of CMap caused by

Received: July 2, 2024. **Revised:** September 6, 2024. **Accepted:** December 3, 2024

database size and cell type [13], the Library of Integrated Network-Based Cellular Signatures (LINCS) expanded the database by using a low-cost gene expression analysis technique, which enabled the large-scale analysis of small molecule–induced transcriptome data [14–16].

As the database continues to be updated and expanded, additional studies have been conducted to improve CMap. CMap mainly uses the KS (Kolmogorov–Smirnov) method to evaluate the similarity score between the query signatures of a disease or drug and the drug signatures in the reference database. However, the KS method has limitations in accuracy and sensitivity, which has prompted researchers to develop other improved similarity scoring methods [13]. The statistically significant connectivity map calculates the weighted similarity score by assigning higher weights to significantly differentially expressed genes, combining the up- and downregulation status of genes and their ranks [17]. The eXtreme cosine uses the cosine of two signatures as the similarity measure for CMap [18]. The eXtreme sum simply sums up the compound gene expression values of the queried genes [19]. ProbCMap uses probabilistic models combined with group factor analysis for drug discovery in one or more cell lines [20]. ksRepo converts information from disease and open databases into EntrezGene identifiers, thus modifying the KS method [21]. Various improvements in the similarity measure have improved the performance of CMap [19, 22, 23]. Since CMap lacks an accurate method for creating optimal gene signatures [13], a few studies have been conducted to improve the construction of gene signatures to adapt to different research requirements [24, 25]. CMapBatch uses a meta-analysis framework to identify the best drug candidates by combining the drug list produced by multiple signatures of a disease [26]. The mode of action by NeTwoRk analysis builds a drug–drug interaction network by capturing the "consensus" of the transcriptional responses of compounds across multiple cell lines and concentrations [27]. CudaMAP uses NVIDIA graphics processing units to reduce the processing time of data analysis and ease the computing needs of improved methods [28]. However, the current version of CMap still has limitations in drug discovery across cell lineages, which hinders its wider application. Considering the cost of second-generation sequencing [29] and the limited size of existing databases, such as LINCS, the application of the current CMap version in drug discovery remains inefficient. Therefore, a method suitable for drug identification across different cell lines needs to be explored to improve the use of the existing data.

In this work, we present the correlation-dependent connectivity map (CDCM), a novel computational drug discovery method based on Pearson's correlation coefficient, for exploring the potential similarity of gene expression changes induced by drugs in different cell lines. In this study, the robustness of the CDCM's predictions of potential therapeutics for SARS-CoV-2 and MPXV was verified via query signatures and reference databases, and its performance was compared with that of the CMap method. In general, the CDCM performed well in cross-cell line drug identification and identified potential drugs against SARS-CoV-2 and MPXV, indicating its good application prospects.

Methods

CMap and its derived methods are primarily focused on the consistent effects of drugs or diseases on changes in intracellular gene expression levels. However, such methods are often limited by inherent differences in cellular environments when identifying drugs between different cell lines, resulting in a tendency for identified associations to be closely related to core cellular processes (e.g. ribosome function) [30], which may overlook the identification of drug-specific or disease-specific effects. In addition, disease-associated genes do not always show significant differential expression [31]. In contrast, the CDCM presents a new perspective that focuses on the consistency of drug or disease effects in terms of expression correlations between genes. When the focus shifted from the expression level of a single gene to the expression relationship between two genes, it was found that even if the expression level of the two genes alone did not change much, the correlation between them could change significantly, such as from a positive to a negative correlation [32, 33]. The CDCM method uses the Pearson correlation coefficient to quantify the correlation between genes, and mining the potential information in the expression profile from the perspective of gene correlation for drug discovery across cell lines.

Like CMap, CDCM includes a query signature, a reference database, and a similarity score calculation method. When calculating the similarity score, taking the drug–drug positive similarity score as an example, if the query signature and the instance signature in the reference database are highly consistent and have the same trend in the correlation change of gene pairs, it indicates that the two drugs may have similar drug effects. At this time, the similarity calculation method will assign a higher score to the instance signature. Each query will generate a ranked list of drugs based on the similarity scores, which will provide a decision-making basis for identifying potential drug candidates with target drug effects during drug screening.

Correlation coefficient change vector calculation

Data sources and processing details are provided in the supplementary. In CDCM, the Pearson correlation coefficient was selected to measure the correlation between two genes. Suppose that the number of replicates under one treatment is $\boldsymbol{n}$. Each replicate contained N genes. The expression values of each gene in $\boldsymbol{n}$ replicates can construct an n-dimensional expression vector. Gene i and gene j are any two genes whose expression vectors are denoted as $\boldsymbol{G}_i$, $\boldsymbol{G}_j$. The correlation coefficient of gene i, j is defined as Equation (1):

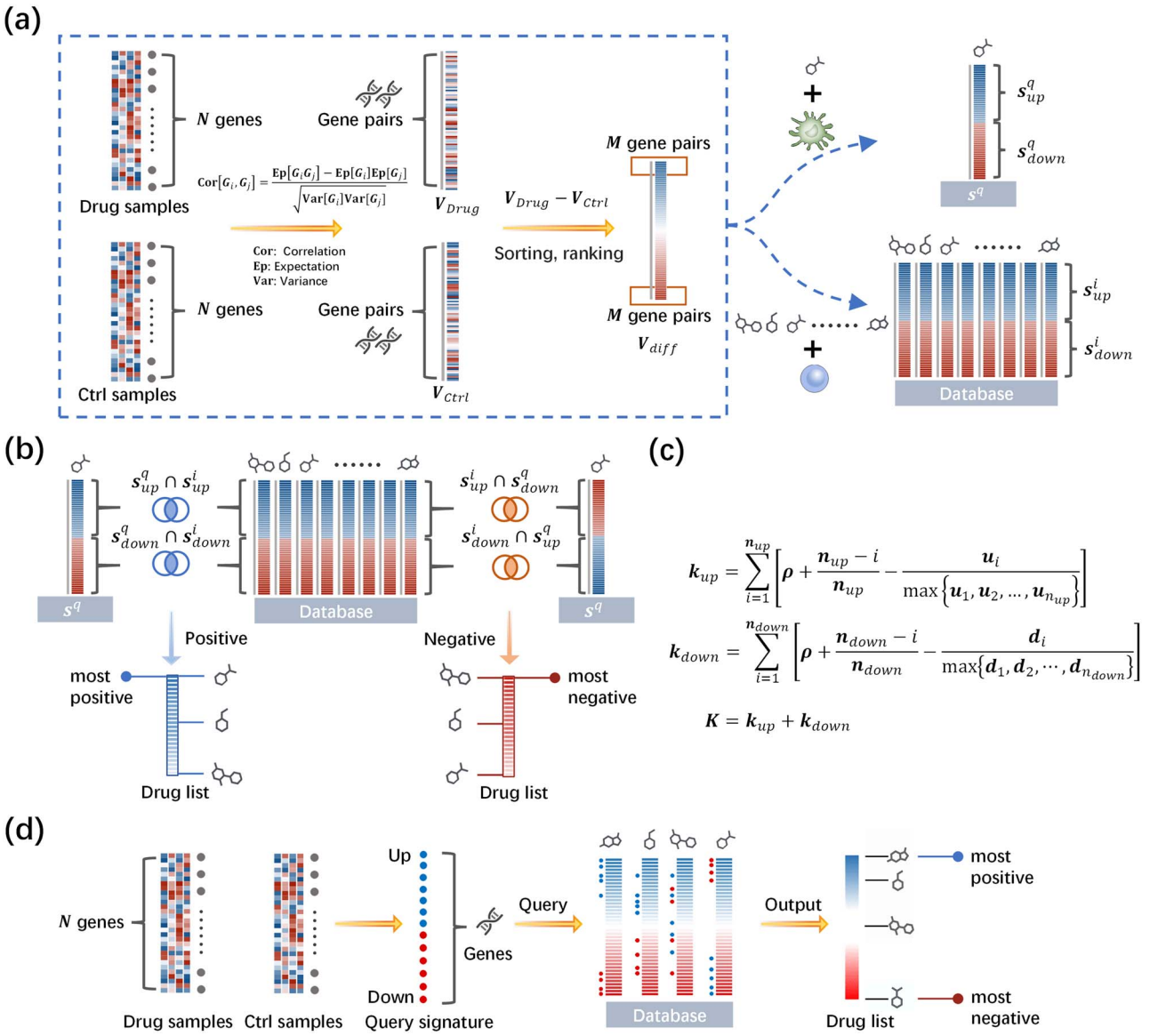
$$\mathrm{Cor}\left[\boldsymbol{G}_i, \boldsymbol{G}_j\right] = \frac{\mathrm{Ep}\left[\boldsymbol{G}_i \boldsymbol{G}_j\right] - \mathrm{Ep}\left[\boldsymbol{G}_i\right] \mathrm{Ep}\left[\boldsymbol{G}_j\right]}{\sqrt{\mathrm{Var}\left[\boldsymbol{G}_i\right] \mathrm{Var}\left[\boldsymbol{G}_j\right]}}, \tag{1}$$

where $\boldsymbol{Ep}$ is the expectation and $\boldsymbol{Var}$ is the variance. N genes can yield $\left(N^2 - N\right)/2$ different gene pair combinations, and the correlation coefficients of these pair combinations can construct a $\left(N^2 - N\right)/2$-dimensional correlation coefficient vector $\mathbf{V}$. Assume that the correlation coefficient vectors of 'drug samples' and 'control samples' are $\mathbf{V}_{Drug}$ and $\mathbf{V}_{Ctrl}$, respectively (Fig. 1a). The correlation change between them is defined as $\mathbf{V}_{diff} = \mathbf{V}_{Drug} - \mathbf{V}_{Ctrl}$. Gene pair combinations with larger absolute values in $\mathbf{V}_{diff}$ are considered to have more essential drug effects on the expression relation between genes.

Query signature and reference database

Constructing a validated drug 'query signature' (denoted as $\boldsymbol{s}^q$) requires arranging the elements of $\mathbf{V}_{diff}$ in descending order and taking the top $\mathbf{M}$ gene pairs denoted as $\boldsymbol{s}^q_{up}$ and the bottom $\mathbf{M}$ gene pairs denoted as $\boldsymbol{s}^q_{down}$ (Fig. 1a).

A reference database is a collection of drug 'instance signatures'. An instance signature (denoted as $\boldsymbol{s}^i$) is the processed data on the expression profiles under a drug treatment, which



$$k_{up} = \sum_{i=1}^{n_{up}} \left[\rho + \frac{n_{up} - i}{n_{up}} - \frac{u_i}{\max\{u_1, u_2, \ldots, u_{n_{up}}\}} \right]$$

$$k_{down} = \sum_{i=1}^{n_{down}} \left[\rho + \frac{n_{down} - i}{n_{down}} - \frac{d_i}{\max\{d_1, d_2, \cdots, d_{n_{down}}\}} \right]$$

$$K = k_{up} + k_{down}$$

Figure 1. Flowchart of the CDCM and CMap. (a) Preparation of the 'query signature' ($\mathbf{s}^q$) and reference database. The Pearson correlation coefficients of all gene pair combinations of N genes in 'drug samples' (or 'control samples') can be used to construct an $(N^2 - N)/2$-dimensional correlation coefficient vector $\mathbf{V}$. The correlation change between drug samples and control samples is defined as $\mathbf{V}_{diff} = \mathbf{V}_{Drug} - \mathbf{V}_{Ctrl}$. Constructing a validated drug query signature (denoted as $\mathbf{s}^q$) requires arranging the elements of $\mathbf{V}_{diff}$ in descending order and taking the top $\mathbf{M}$ and bottom $\mathbf{M}$ gene pairs denoted as $\mathbf{s}^q_{up}$ and $\mathbf{s}^q_{down}$, respectively. To obtain an 'instance signature' (denoted as $\mathbf{s}^i$), gene pairs in a sorted $\mathbf{V}_{diff}$ are divided into positive (≥0) and negative (<0) sets according to their corresponding values and are ranked in descending order in each set according to the absolute values. The top $\mathbf{M}$ gene pairs with the lowest ranks in the positive and negative sets, denoted as $\mathbf{s}^i_{up}$ and $\mathbf{s}^i_{down}$, respectively, are extracted to form a $2\mathbf{M}$-dimensional $\mathbf{s}^i$ in the reference database. (b) Positive and negative similarity score calculation. The positive similarity between $\mathbf{s}^q$ and $\mathbf{s}^i$ measures the similarity degree of the change in gene pair correlation between $\mathbf{s}^q_{up}$ and $\mathbf{s}^i_{up}$ and between $\mathbf{s}^q_{down}$ and $\mathbf{s}^i_{down}$ by obtaining $\mathbf{s}^q_{up} \cap \mathbf{s}^i_{up}$ and $\mathbf{s}^q_{down} \cap \mathbf{s}^i_{down}$ to calculate a positive similarity score. The negative similarity between $\mathbf{s}^q$ and $\mathbf{s}^i$ measures the similarity degree of the change in gene pair correlation between $\mathbf{s}^q_{up}$ and $\mathbf{s}^i_{down}$ and between $\mathbf{s}^q_{down}$ and $\mathbf{s}^i_{up}$. (**c**) Formula of the similarity score. Suppose there are $\boldsymbol{n}_{up}$ and $\boldsymbol{n}_{down}$ gene pairs in $\mathbf{s}^q_{up} \cap \mathbf{s}^i_{up}$ (or $\mathbf{s}^q_{up} \cap \mathbf{s}^i_{down}$) and $\mathbf{s}^q_{down} \cap \mathbf{s}^i_{down}$ (or $\mathbf{s}^q_{down} \cap \mathbf{s}^i_{up}$) for a positive similarity score (or a negative similarity score). $[\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{n_{up}}]$ are ranks of the $\boldsymbol{n}_{up}$ gene pairs in $\mathbf{s}^i$, and $[\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_{n_{down}}]$ are ranks of the $\boldsymbol{n}_{down}$ gene pairs in $\mathbf{s}^i$. (**d**) Flowchart of the CMap. Query signature is composed of genes with significantly upregulated and downregulated expression. Each instance signature in the reference database contains all the genes and their differential expression ranks. The list of drugs output shows the KS scores of all drugs in the database with the query drug.

represents the drug effect on the gene relation. As described in the previous section, every drug's $\mathbf{V}_{diff}$ can be calculated based on 'drug samples' and 'control samples'. To obtain an $\mathbf{s}^i$, gene pairs in a sorted $\mathbf{V}_{diff}$ are divided into positive (≥0) and negative (<0) sets according to their corresponding values and are ranked in descending order in each set according to the absolute values. The top $\mathbf{M}$ gene pairs with the lowest rank in positive and negative sets, denoted as $\mathbf{s}^i_{up}$ and $\mathbf{s}^i_{down}$, are extracted to form a $\mathbf{2M}$-dimensional $\mathbf{s}^i$ in the reference database.

Pattern-matching strategy

CDCM needs to calculate the positive similarity score and negative similarity score (Fig. 1b). For the positive similarity score, suppose there are $\boldsymbol{n}_{up}$ and $\boldsymbol{n}_{down}$ gene pairs in two intersections

$\boldsymbol{s}^q_{up} \cap \boldsymbol{s}^i_{up}$ and $\boldsymbol{s}^q_{down} \cap \boldsymbol{s}^i_{down}$. The corresponding ranks of these gene pairs in $\boldsymbol{s}^i$ are $[\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{n_{up}}]$ and $[\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_{n_{down}}]$. The score $\boldsymbol{K}$ is defined as follows:

$$\boldsymbol{k}_{up} = \sum_{i=1}^{\boldsymbol{n}_{up}} \left[\boldsymbol{\rho} + \frac{\boldsymbol{n}_{up} - i}{\boldsymbol{n}_{up}} - \frac{\boldsymbol{u}_i}{\max\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{n_{up}}\}} \right], \quad (2)$$

$$\boldsymbol{k}_{down} = \sum_{i=1}^{\boldsymbol{n}_{down}} \left[\boldsymbol{\rho} + \frac{\boldsymbol{n}_{down} - i}{\boldsymbol{n}_{down}} - \frac{\boldsymbol{d}_i}{\max\{\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_{n_{down}}\}} \right], \quad (3)$$

$$\boldsymbol{K} = \boldsymbol{k}_{up} + \boldsymbol{k}_{down} \quad (4)$$

Score $\boldsymbol{K}$ measures the similarity degree of gene-pair correlation change between $\boldsymbol{s}^q_{up}$ and $\boldsymbol{s}^i_{up}$, $\boldsymbol{s}^q_{down}$, and $\boldsymbol{s}^i_{down}$. $\boldsymbol{\rho}$ is a constant that adjusts the effect of the two intersection sizes on score $\boldsymbol{K}$.

It can be seen from the formula: (i) the larger the intersections are, the more gene pairs there are that change synchronously between $\boldsymbol{s}^q$ and $\boldsymbol{s}^i$, and the greater the similarity is; (ii) the smaller the ranks of gene pairs in two intersections are, the more significant the correlation change between gene pairs in $\boldsymbol{s}^q$ and $\boldsymbol{s}^i$ is, and the higher the score is; (iii) $\boldsymbol{\rho}$ can balance the contributions of ranks and intersection sizes to the score by balancing their weights; and (iv) the larger the $\boldsymbol{M}$ is, the higher the requirement is for the degree of similarity between $\boldsymbol{s}^q$ and $\boldsymbol{s}^i$.

Due to the symmetry of the method, the negative similarity score is calculated in essentially the same way as the positive similarity score. Only the objects of the intersection need to be exchanged, that is, $\boldsymbol{s}^q_{up} \cap \boldsymbol{s}^i_{down}$ and $\boldsymbol{s}^q_{down} \cap \boldsymbol{s}^i_{up}$, where the corresponding ranks of the gene pair are denoted as $[\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{n_{up}}]$ and $[\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_{n_{down}}]$, respectively. The score $\boldsymbol{K}$ is still calculated according to Equations (2–4). The essence of the negative similarity score is that gene pairs with increased (or decreased) correlation in $\boldsymbol{s}^i$ are decreased (or increased) in $\boldsymbol{s}^q$, indicating that the two drugs have opposite effects on gene expression correlation and may have certain reverse effects.

Both types of score calculations output a ranked list of drugs, showing how similar the drugs in the reference database are to the validated drug. Through small-scale validation, we find that the optimal value of $\boldsymbol{M}$ is between 300 000 and 2 000 000, and in this research, $\boldsymbol{M}$ is set to 300 000, 500 000, 1 000 000, and 2 000 000. By comparing the proportion of top-ranked results between different cell lines, $\boldsymbol{\rho} = 1$ shows a stable performance in 4, 2, 1, 0.1, 0.05, 0.02, and 0.01 and is a compromise choice that will be applied in the following large-scale validations (Fig. S1).

In fact, the core of CDCM is to exploit the similarity of gene networks. A gene pair represents two nodes in the network. The edge between two nodes is represented by the correlation or correlation change of the gene pair. All possible gene pair combinations and their edges form a complete gene network. Therefore, for each $\boldsymbol{s}^i$ in the drug rank list output by CDCM in a query with a $\boldsymbol{s}^q$, the gene pair intersection ($(\boldsymbol{s}^q_{up} \cap \boldsymbol{s}^i_{up}) \cup (\boldsymbol{s}^q_{down} \cap \boldsymbol{s}^i_{down})$ for positive score) generated by the score calculation with $\boldsymbol{s}^q$ can also be a gene network. To show the huge difference in similarity between the first $\boldsymbol{s}^i$ (denoted as $\boldsymbol{s}^{i_1}$) and the last $\boldsymbol{s}^i$ (denoted as $\boldsymbol{s}^{i_2}$) in a drug rank list from a network perspective, we extract two subsets of the two gene pair intersections from $\boldsymbol{s}^{i_1}$ and $\boldsymbol{s}^{i_2}$ to construct gene networks. Assume that the vectors of correlation change for $\boldsymbol{s}^{i_1}$, $\boldsymbol{s}^{i_2}$ and $\boldsymbol{s}^q$ are $\boldsymbol{V}^{s^{i_1}}_{diff}$, $\boldsymbol{V}^{s^{i_2}}_{diff}$, and $\boldsymbol{V}^{s^q}_{diff}$. Subset extraction refers to obtaining the top 1000 gene pairs from the gene pair intersection of an $\boldsymbol{s}^i$ based on the sum of absolute values in $\boldsymbol{V}^{s^i}_{diff}$ and $\boldsymbol{V}^{s^q}_{diff}$ for each gene pair, which could roughly reflect whether the correlation changes of the drug are significant and similar between its $\boldsymbol{s}^i$ and $\boldsymbol{s}^q$. Finally, two subsets of gene pair intersections of $\boldsymbol{s}^{i_1}$ and $\boldsymbol{s}^{i_2}$ are denoted as $\boldsymbol{S1}$ and $\boldsymbol{S2}$.

The essence of CDCM is to identify and quantify the consistency of drug effects on gene expression relation networks in cells (Fig. 1). Signature is a collection of gene pairs and ranks that stores the gene relations most significantly affected by a drug, which can be divided into $\boldsymbol{s}_{up}$ and $\boldsymbol{s}_{down}$, including gene pairs with significant increases and decreases in the Pearson correlation coefficient. Ranks in $\boldsymbol{s}_{up}$ (or $\boldsymbol{s}_{down}$) show the positions of positive (or negative) correlation change in a gene pair among all gene pairs of the same correlation change type, with a larger change leading to a higher rank. This means that the gene pairs with the most positive and negative changes are ranked at the top one in $\boldsymbol{s}_{up}$ and $\boldsymbol{s}_{down}$, respectively. By comparing the number of gene pairs common in $\boldsymbol{s}^i$ and $\boldsymbol{s}^q$ with the degree of correlation change, CDCM assigns each $\boldsymbol{s}^i$ in the reference database a score and ranks these $\boldsymbol{s}^i$s according to the score.

Connectivity map method

The CMap method was used to compare it with CDCM (Fig. 1d). There has long been a lack of clear guidance on $\boldsymbol{s}^q$ length when using the CMap method [34]. However, the change in individual genes in $\boldsymbol{s}^q$ may lead to completely different results [26]. Therefore, three lengths of 200, 400, and 600 were set, which are evenly divided into the upregulated part and downregulated part of $\boldsymbol{s}^q$, and three scores were calculated for $\boldsymbol{s}^q$ to weaken the randomness of the result brought by the $\boldsymbol{s}^q$ length variation.

Experimental validation of predicted monkeypox virus drugs

The A549 cells were infected with MPVX (MOI = 0.1), and 48 h later, the intracellular RNA was extracted and the mRNA-seq was detected. After MPXV infected A549, the cells were treated with 1 μM concentration of drugs (ponatinib, dabrafenib, sunitinib, lapatinib), the nucleic acid (Daan gene, #DA0620) was extracted from the samples 48 h later, and the MPVX content in the samples was detected by qPCR (F3L-F: cttccgtcaatgtctacacaggc; F3L-R: cgttggtctacgacaatggatgc).

Results

Validation across different cell lines

We denoted the four cell lines in dataset 1 as *A*, *B*, *D*, and *E*. The reference database and $\boldsymbol{s}^q$ used for validation were constructed according to the steps shown in Fig. 1. Drugs with more than three replicate profiles in each cell line were retained to construct four reference databases. Each cell line needed three $\boldsymbol{s}^q$ sets of common drugs, with the other three cell lines used for validation. The expression profiles of the cell lines treated with the same drug were expected to be the most similar. The ideal result for a given drug list would be that the drug to be validated is at the top of the list. The minimum rank of the validated drug in the output lists with different $\boldsymbol{M}$ was regarded as the result for this method.

As shown in Fig. 2a, ~12.2% of the validated drugs ranked first in the drug lists, indicating that the method could accurately identify the most similar drugs. Nearly 21.2% of the validated drugs ranked in the top two, and 28.8% ranked in the top three. Figure 2b shows that ~20% of the validated drugs were in the top 20%, and approximately half were in the top 40%. When used for validation, many drugs ranked in the top half of the list. The resulting lists of two $\boldsymbol{s}^q$'s are provided in Tables S1 and S2, in which the validated drugs were identified and placed in the top rank.

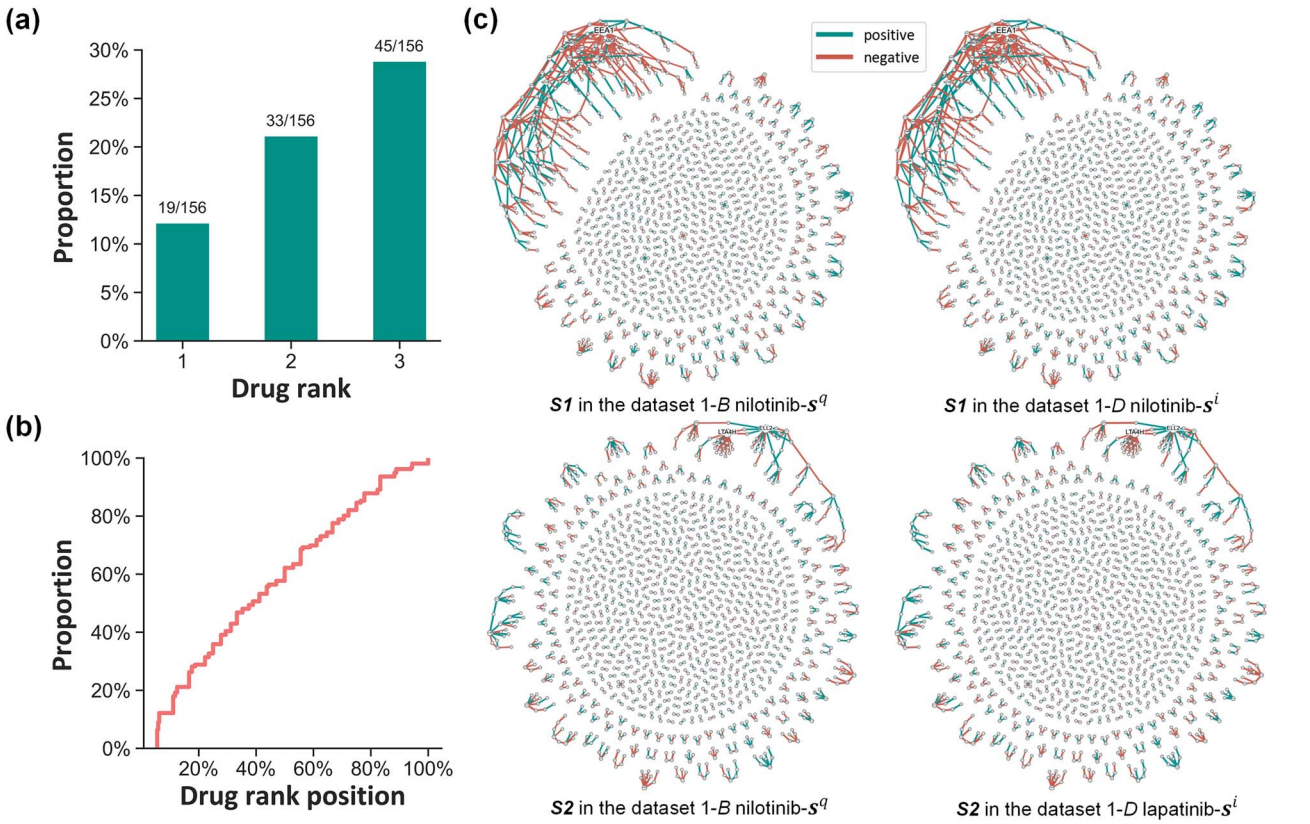


Figure 2. Validation results for Dataset 1. (a) The proportions of validated drugs ranked in the top one, two, and three. (b) Cumulative distribution of the validated drug rank positions in the list. (c) The gene expression networks of nilotinib and lapatinib are shown in Table S1. ***S1*** and ***S2*** were obtained from nilotinib-$\mathbf{s}^q$, which was validated with nilotinib-$\mathbf{s}^i$ and lapatinib-$\mathbf{s}^i$. Nodes represent genes. Lines between nodes represent correlations between genes. The line color represents the positive or negative value of the change. The node size and text label size represent the degree of the node. The line width represents the change in the gene-pair correlation coefficient.

It can be concluded that the CDCM identified validated drugs accurately with Dataset 1.

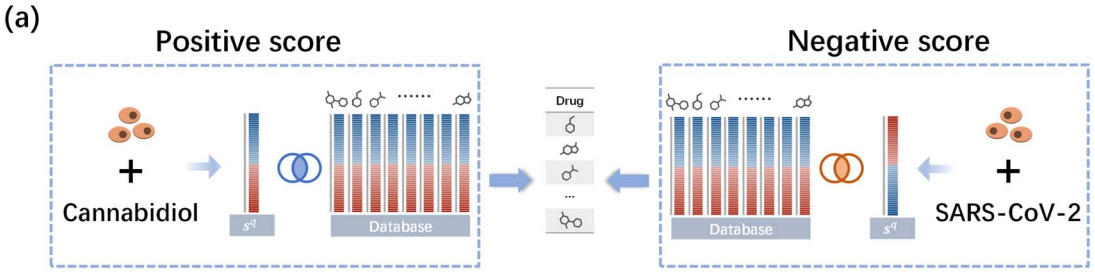
In addition, the CDCM could identify other drugs that act similarly to the validated drugs and assign them higher scores. As shown in Table S1, nilotinib ranked first in Dataset 1. Nilotinib is mainly used to treat imatinib-resistant chronic myelocytic leukemia [35], and its main targets are platelet-derived growth factor receptor (PDGFR) and c-Kit [36]. Sunitinib, which ranks second, is mainly used to treat gastrointestinal stromal tumors and metastatic renal cell carcinoma [37], and its targets include PDGFR and c-Kit [38]. Dasatinib, which ranks third, is a multikinase inhibitor with targets including c-Kit and PDGFR [39]. These drugs share similar targets. Another case is the list generated by sorafenib-$\mathbf{s}^q$ (Table S2). Sorafenib is a novel multitarget kinase inhibitor that inhibits B-Raf proto-oncogene (BRAF) and vascular endothelial growth factor receptor (VEGFR) [40]. Dabrafenib, which ranks second, has high inhibitory activity against BRAF [41]. Vandetanib, which ranks third, also acts on VEGFR targets in tumor cells [42]. These results show that the CDCM can be used to identify functionally similar drugs in different cell lines.

The expression correlation of gene pairs under drug action can be observed in the gene network. The networks were constructed from the gene expression data of the first-ranked drug (nilotinib) and the last-ranked drug (lapatinib) from Table S1 (Fig. 2c). The intersection sizes of nilotinib-$\mathbf{s}^i$ and lapatinib-$\mathbf{s}^i$ generated in the calculation with $\mathbf{s}^q$ were 23 000 and 19 000, respectively, suggesting one reason for such a large gap between the scores of the two $\mathbf{s}^i$'s. ***S1*** and ***S2*** had the same number of gene pairs (1000) but different numbers of gene nodes (1484 for ***S1*** and 1686 for ***S2***), indicating that ***S1*** had higher network complexity than ***S2***. The gene networks in ***S1*** had a larger network aggregation module, which were denser than those in ***S2***, and they could better demonstrate the direct or indirect effects of drugs on genes. Relatively, the correlation changes among the gene nodes in ***S2*** were not as strong as those in ***S1***, reflecting the lower coincidence degree between nilotinib-$\mathbf{s}^q$ and lapatinib-$\mathbf{s}^i$.

Prediction results applied to the COVID-19 dataset

COVID-19 has caused millions of infections and deaths worldwide. We used Datasets 1, 2, 3, and 4 to identify potential effective therapies for COVID-19 and explore the utility of the CDCM in discovering drug–drug and disease–drug relationships. Nguyen *et al.* studied the inhibitory effects of cannabinol (CBD) against SARS-CoV-2 (the causative pathogen of COVID-19) in A549 cells in Dataset 4 and highlighted CBD as a potential preventive agent for early SARS-CoV-2 infection. We used the expression profiles of the treatment samples and control samples of CBD and SARS-CoV-2 in this dataset to construct CBD-$\mathbf{s}^q$ and SARS-CoV-2-$\mathbf{s}^q$. The CDCM was used to obtain positive and negative similarity scores by querying CBD-$\mathbf{s}^q$ and SARS-CoV-2-$\mathbf{s}^q$ in the reference databases of Datasets 1, 2, and 3 (Fig. 3a).

For the predictions of CBD-$\mathbf{s}^q$ and SARS-CoV-2-$\mathbf{s}^q$ in the four reference databases of Dataset 1, we found that dabrafenib in G1-B ranked first in the output lists of both drug–drug similarity and disease–drug similarity calculations (Fig. 3b and c). These findings



(b)

Compound	Rank	Normalized score
Dabrafenib	**1**	**1.0000**
Nilotinib	2	0.9177
Erlotinib	3	0.8976
Tofacitinib	4	0.8973
Vandetanib	5	0.8973
Lapatinib	6	0.8895
Pazopanib	7	0.8868
Ponatinib	8	0.8838
Afatinib	9	0.8829
Sorafenib	10	0.8791
Trametinib	11	0.8776
Dasatinib	12	0.8757
…	…	…
Gefitinib	18	0.8325

Note: $\boldsymbol{M}$=1000000, $\boldsymbol{\rho}$=1.

(c)

Compound	Rank	Normalized score
Dabrafenib	**1**	**1.0000**
Nilotinib	2	0.8868
Ponatinib	3	0.8842
Pazopanib	4	0.8808
Erlotinib	5	0.8802
Cabozantinib	6	0.8726
Afatinib	7	0.8611
Imatinib	8	0.8561
Axitinib	9	0.8537
Trametinib	10	0.8482
Crizotinib	11	0.8070
Dasatinib	12	0.8068
…	…	…
Tofacitinib	18	0.7454

Note: $\boldsymbol{M}$=1000000, $\boldsymbol{\rho}$=1.

Figure 3. Drug prediction by the CDCM in the SARS-CoV-2 dataset. (a) Flowchart for SARS-CoV-2 drug prediction via the CDCM. (b) Output lists of CBD-$\boldsymbol{s}^q$ validated in reference databases of Dataset 1-*B*. (c) Output lists of SARS-CoV-2-$\boldsymbol{s}^q$ validated in reference databases of Dataset 1-*B*.

indicate that the effect of dabrafenib on gene correlation in neural cells is consistent with the effect of CBD on gene correlation in A549 cells. However, this effect was opposite to that of SARS-CoV-2 on the gene expression correlation changes in A549 cells. In other words, the CDCM predicted that dabrafenib has an inhibitory effect on SARS-CoV-2 similar to that of CBD. In addition, as with dabrafenib, lapatinib (dataset 1-*E*) (Tables S3 and S4), sorafenib (dataset 2) (Tables S5 and S6), and trichostatin A (dataset 3-MCF7) (Tables S7 and S8) ranked first in the lists for both queries. Previous studies have confirmed that these predicted drugs have certain inhibitory effects on SARS-CoV-2 [43–48]. Therefore, these compounds deserve further exploration and validation as potential therapeutic agents for COVID-19. The prediction results also indicated that the CDCM has a strong potential for identifying therapeutic drugs for viral diseases.

Prediction results applied to the monkeypox virus dataset

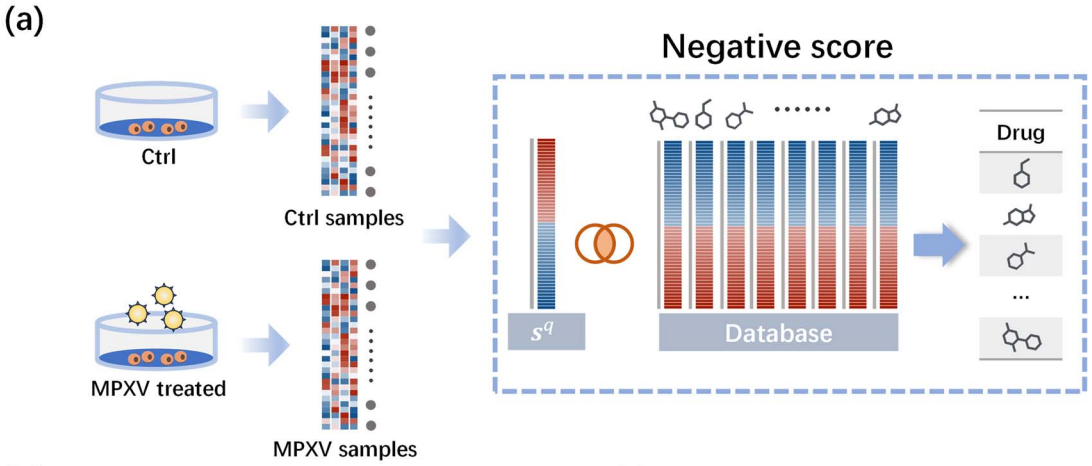
We applied CDCM to predict potential drugs for MPXV and conducted experiments to verify the inhibitory effects of the predicted drugs on MPXV. First, MPXV was used to infect A549 cells (experimental group), and RNA sequencing was performed to obtain the expression profiles of the experimental and control groups to construct MPXV-$\boldsymbol{s}^q$ (Fig. 4a). Then, MPXV-$\boldsymbol{s}^q$ was used to calculate the negative similarity score in the reference databases of Dataset 1. Four drugs (ponatinib, dabrafenib, sunitinib, and lapatinib), which were in the top 5% of the output lists, were chosen for subsequent experimental validation (Tables S9–S11 and Fig. 4b).

A549 cells infected with MPXV were treated with the selected predicted drugs, and three of them (ponatinib, sunitinib, and lapatinib) exhibited varying degrees of inhibitory effects on MPXV (Fig. 4c). Ponatinib, sunitinib, and lapatinib decreased the proliferation rate of MPXV by 173.1, 264.3, and 19.9 times, respectively. Dabrafenib had no effect on the proliferation of MPXV. This finding highlights the usefulness of the CDCM for drug repurposing. Currently, patients infected with MPXV lack effective treatment options and have to rely on natural recovery; therefore, the potential use of these drugs warrants further exploration.

Comparison of the correlation-dependent connectivity map with Connectivity Map

We compared the performance of the CDCM and CMap on Datasets 1, 2, and 3. There were multiple drug concentrations in these three datasets, and the number of replicates for each concentration differed. To compare the validation results of the two methods reasonably, we averaged all the expression profiles for each drug in each dataset to obtain a new expression profile, which was used to construct the $\boldsymbol{s}^q$s and $\boldsymbol{s}^i$s used in the CMap method. The gene expression values of 0 were replaced by 0.001. Since only a positive similarity comparison was performed, all nonpositive scores in CMap were regarded as invalid scores, and their corresponding ranks were modified to the maximum ranks in the lists. $\boldsymbol{M}$ of the CDCM was set to one of three values: 300 000, 500 000, and 1 000 000. $\boldsymbol{\rho}$ was set to 1. Only the minimum rank generated under the three lengths was retained for the statistical analysis.

As shown in Fig. 5, the consistent cell types in dataset 1 allowed CMap to identify certain drugs within the top 10% of the lists well, but most validated drugs $\boldsymbol{s}^i$ received invalid scores, making the CDCM far more efficient than CMap for the top 20% and beyond (Fig. 5a–d). When Dataset 2 was validated in the reference database of dataset 1, many of the CMap outputs were invalid,



(b)

Compound	Rank	Normalized score
Ponatinib	**1**	**1.0000**
Sorafenib	2	0.9936
Bosutinib	3	0.9922
Sunitinib	4	0.9916
Vemurafenib	5	0.9893
Ceritinib	6	0.9777
Vandetanib	7	0.9718
Dabrafenib	8	0.9627
Ruxolitinib	9	0.9593
Afatinib	10	0.9542
Axitinib	11	0.9496
Imatinib	12	0.9342
Nilotinib	13	0.9319
Gefitinib	14	0.9301
Tofacitinib	15	0.9217
Crizotinib	16	0.9214

Note: $\boldsymbol{M}$=1000000, $\boldsymbol{\rho}$=1.

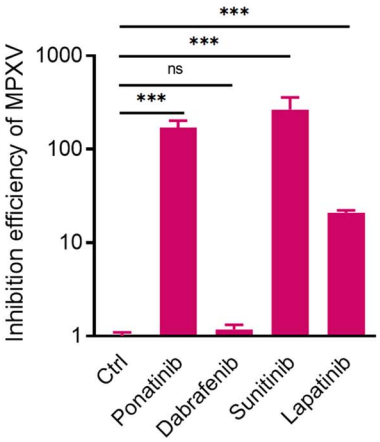


Figure 4. Drug prediction by the CDCM in the MPXV dataset. (a) Flowchart for MPXV drug prediction via the CDCM. (b) Output lists of MPXV-$\boldsymbol{s}^q$ validated in reference databases of Datasets 1-*A* and 1-*D*. (c) Inhibitory effects of four predicted drugs on MPXV in the experiment.

whereas the CDCM had the obvious advantage of being able to break through such large cell differences (Fig. 5e–h). In Dataset 3, the performance of CMap improved, and the advantage of the CDCM was that it focused mainly on identifying drugs in the top 30% of the list (Fig. 5i and j). Notably, this dataset is produced for CMap and may be more suitable for the CMap method. Overall, the CDCM performed better than the previous CMap version in cross-cell lineage applications.

Discussion

In this work, we present a novel drug discovery approach, the CDCM, designed to expedite the identification of potential treatments for viral diseases. This method holds promise for advancing the prevention and treatment of infectious diseases, including COVID-19 and MPXV, in future public health efforts. Recently, the list of pandemic pathogens released by the World Health Organization (WHO) covers >30 pathogens, such as MPXV [49] and highlights the urgency of rapidly developing methods for identifying potential therapeutic drugs that are less sensitive to the cellular context of the data. At the time of an infectious disease outbreak, time constraints often prevent us from customizing data for methods such as CMap. However, the use of available data for prediction may affect the stability and efficiency of drug discovery due to the influence of cell type or other experimental conditions. The CDCM method enables drug prediction across cell lines and has unique advantages in emergency situations with limited data.

The CDCM showed good identification performance in the prediction across different cell lines. It scores drugs by calculating the similarity of changes in expression correlations between genes, making full use of the information in expression profiles. For CMap, the different degrees of transcriptome expression caused by different cellular physiological environments may exceed the perturbations generated by drugs, resulting in poor performance in cross-cell line prediction. However, the gene–gene network interactions are substantially similar because they involve the same genome. Therefore, by calculating the correlation changes among genes, the CDCM can eliminate the dependence on the original values of the cellular gene expression, thus enabling more accurate predictions of drug functions across different cell lines.

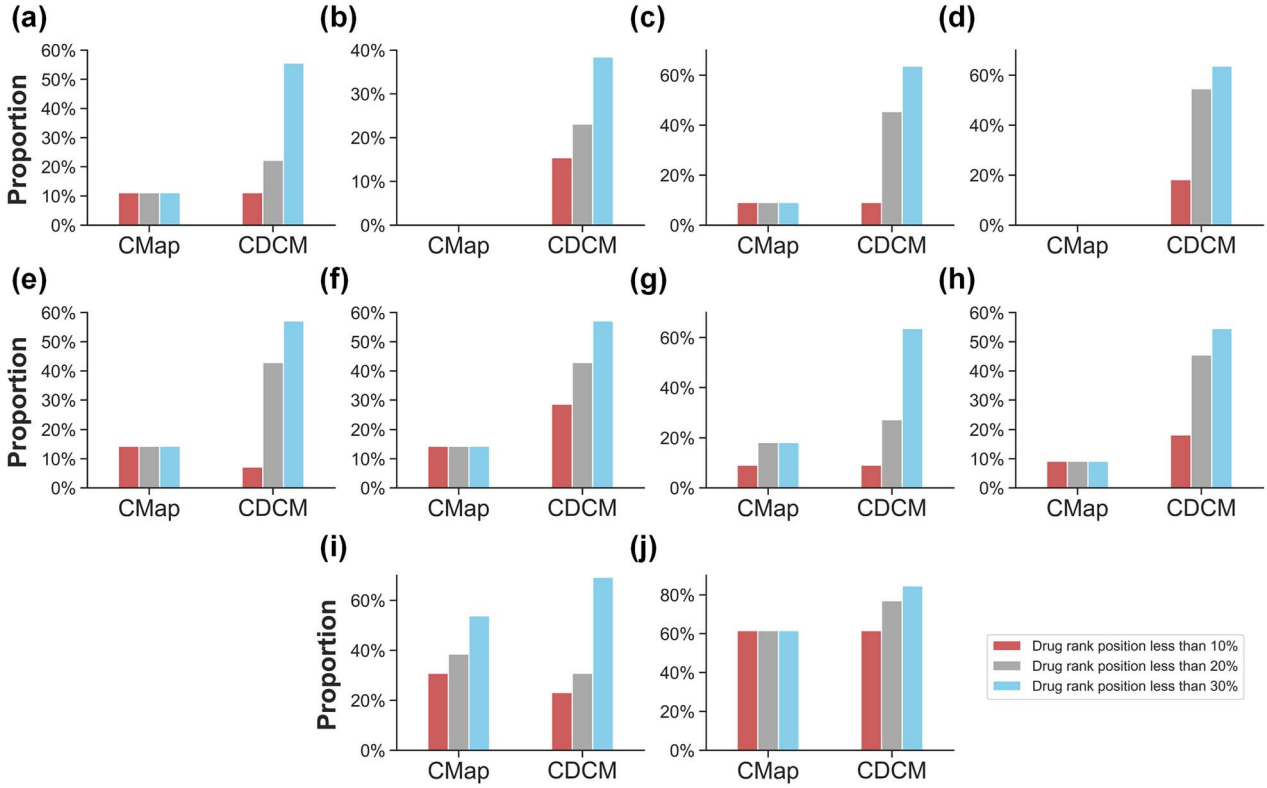


Figure 5. Comparison of CMap and the CDCM in terms of the proportion of top drug rank positions in the validation results between different cell lines. (a) Dataset 2 validated in Dataset 1-*A*. (b) Dataset 2 validated in Dataset 1-*B*. (c) Dataset 2 validated in Dataset 1-*D*. (d) Dataset 2 validated in Dataset 1-*E*. (e) Dataset 1-*A* validated in Dataset 1-*D*. (f) Dataset 1-*D* validated in Dataset 1-*A*. (g) Dataset 1-*A* validated in Dataset 1-*E*. (h) Dataset 1-*E* validated in Dataset 1-*A*. (i) Dataset 3-PC3 validated in Dataset 3-HL60. (j) Dataset 3-HL60 validated in Dataset 3-PC3.

In this study, the CDCM was validated in three datasets, and it could place validated drugs at the top of the list with a high probability. The identification efficiency was affected by experimental factors such as data quality and cell line differences, but the fluctuation range was small. The CDCM has also been used to identify SARS-CoV-2 and MPXV therapeutics. With respect to the CDCM drug predictions for SARS-CoV-2, some of the top drugs have been shown to be effective against the virus in recent studies, and most of the predicted MPXV drugs have also been shown to be effective in subsequent experiments. The efficiency of the CDCM and CMap for drug identification was also compared, and the results revealed that CMap was less stable and accurate than the CDCM.

The CDCM considers the similarity of $\mathbf{s}^q$ and $\mathbf{s}^i$ and has greater application value in drug discovery between different cell lines. However, many problems necessitate further study. The first is the number of replicates. The requirement of sufficient replicates is difficult to meet because of cost constraints. Therefore, the necessary and appropriate number of replicates should be explored to achieve the best effect while minimizing the cost of data acquisition as much as possible. The diversity of sample concentrations used also affects the quality of $\mathbf{s}^q$ and $\mathbf{s}^i$. Appropriate concentration ranges and quantities may need to be considered in relation to the specific drug to allow the range of gene expression changes at different concentrations to be reasonably extended to highlight the correlation between genes without influence from extreme concentrations and causing misinterpretation of drug effects. The sizes of $\mathbf{s}^q$ and $\mathbf{s}^i$ can also be determined by setting a threshold for Pearson's correlation coefficient on the basis of experience rather than retaining it as the same setting of $\mathbf{M}$. In addition, the reference databases constructed with multicell lines lead to global differences in the validation results, possibly due to systemic differences in the biological environments of the cell lines. The advantage of the CDCM lies in effectively calculating the similarity between $\mathbf{s}^q$ and the reference database across cell lines. For reference database construction, we recommend that the data should be obtained from the same cell line.

In the future, the CDCM may be used to integrate public data from multiple platforms to construct a large-scale comprehensive reference database for researchers to explore drugs for various diseases. In addition to helping address public health events, the CDCM could be applied in drug development projects for diseases caused by highly pathogenic microorganisms, such as HIV and *Mycobacterium tuberculosis*, which are strictly restricted to BSL-3/4 laboratories and thus are not applicable in large-scale screening projects. Thus, the CDCM can be used to perform large-scale virtual screening before studies in BSL-3/4 laboratories, which may accelerate drug development.

In terms of applications, CDCM can be extended by combining data such as phenotype, drug structure, and protein structure [50–52] to narrow the scope of drug screening, making it a valuable tool for revealing drug–drug and disease–drug relationships in pharmaceutical discovery.

Conclusion

We developed a signature-based drug identification method, CDCM, which enables drug discovery across cell lines by analyzing the similarity of correlated changes between gene expression. Especially at the critical time of the outbreak of infectious diseases, the CDCM has greater advantages in drug discovery under the condition of limited data. The validation of the CDCM

across different cell types showed good drug discovery ability, and the CDCM was also able to identify potential therapeutic drugs in the drug identification application cases of SARS-CoV-2 and MPXV. Compared with CMap, the CDCM has better accuracy and stability in cross-cell lineage drug identification, showing potential for drug discovery and personalized medicine.

Key Points

- The correlation-dependent connectivity map (CDCM) overcomes the physiological barriers inherent in different cell types for drug identification.
- The identification accuracy and stability of the CDCM are greater than those of CMap.
- The CDCM is less sensitive to drug concentration differences than CMap.
- The CDCM accurately revealed that sorafenib could be a potential therapeutic agent for COVID-19; sorafenib has previously been demonstrated by several laboratories worldwide to inhibit SARS-CoV-2.
- In the Biosafety Level III laboratory experiment, ponatinib, sunitinib, and lapatinib, which were predicted by the CDCM, were shown to inhibit monkeypox virus proliferation with an efficacy rate of 75%.

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Acknowledgements

This work was supported in part by the High Performance Computing Center of Central South University and the Biosafety Level III Laboratory of Shenzhen Third People's Hospital.

Funding

This work was supported by the National Natural Science Foundation of China [Nos. 12471394, 32300659, 12071488, 12371417, 92169119, 82070420, 92469203]; Shenzhen Science and Technology Innovation Commission Project [No. JCYJ20230807143302004]; Shanghai Science and Technology Innovation Action Plan [No. 22N31900800]; Guangdong Province Science and Technology Plan Project "Biosafety Technology" Special Project [No. 2022B111010003]; Shenzhen High-level Hospital Construction Fund [No. XKJS-CRGRK-011]; and the Grant of 2021 Guangdong Recruitment Program of Foreign Experts (Hao Wang).

Data availability

Most of the experimental datasets used in this study, except for MPXV-related data, are publicly available. Details on how to access them can be found in the text that first mentions the dataset in the supplementary material. The MPXV-related data could be used with the consent of the author; please contact the author.

Author contributions

H.Y., Y.N., H.L., J.L., and HW developed the concept of the project and wrote the paper. J.L. and D.J. performed bioinformatic analysis. S.Y., X.H., M.Z., and J.S. assisted in performing experiments. H.Y., M.Z., J.S., H.L., and Y.N. supervised the project. All authors have read and agreed to the published version of the manuscript.

References

1. Karim SSA, Karim QA. Omicron SARS-CoV-2 variant: A new chapter in the COVID-19 pandemic. *Lancet* 2021;**398**:2126–8. https://doi.org/10.1016/S0140-6736(21)02758-6.
2. Irons NJ, Raftery AE. Estimating SARS-CoV-2 infections from deaths, confirmed cases, tests, and random surveys. *Proc Natl Acad Sci USA* 2021;**118**:e2103272118. https://doi.org/10.1073/pnas.2103272118.
3. Liu H, Wang S, Yang S. *et al.* Characteristics of the severe acute respiratory syndrome coronavirus 2 omicron BA.2 subvariant in Jilin, China from March to May 2022. *J Transl Int Med* 2022;**10**:349–58.
4. Zhang M, He Y, Jie Z. Delta variant: Partially sensitive to vaccination, but still worth global attention. *J Transl Int Med* 2022;**10**:227–35. https://doi.org/10.2478/jtim-2022-0026.
5. Georgiadou SP, Giamarellos-Bourboulis EJ, Dalekos GN. Monkeypox: A real new warning or just a sign of times? *J Transl Int Med* 2023;**11**:15–8. https://doi.org/10.2478/jtim-2023-0005.
6. Lin X, Wu X. Monkeypox: Clinical issues of concern. *J Transl Int Med* 2022;**10**:297–9. https://doi.org/10.2478/jtim-2022-0038.
7. Mohs RC, Greig NH. Drug discovery and development: Role of basic biological research. *Alzheimers Dement (N Y)* 2017;**3**:651–7. https://doi.org/10.1016/j.trci.2017.10.005.
8. Zhang L, Kang W, Lu X. *et al.* 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. https://doi.org/10.1016/j.gene.2018.09.040.
9. Churchman ML, Low J, Qu C. *et al.* Efficacy of retinoids in IKZF1-mutated BCR-ABL1 acute lymphoblastic Leukemia. *Cancer Cell* 2015;**28**:343–56. https://doi.org/10.1016/j.ccell.2015.07.016.
10. Nygren P, Fryknäs M, Agerup B. *et al.* Repositioning of the anthelmintic drug mebendazole for the treatment for colon cancer. *J Cancer Res Clin Oncol* 2013;**139**:2133–40. https://doi.org/10.1007/s00432-013-1539-5.
11. Smalley JL, Breda C, Mason RP. *et al.* Connectivity mapping uncovers small molecules that modulate neurodegeneration in Huntington's disease models. *J Mol Med (Berl)* 2016;**94**:235–45. https://doi.org/10.1007/s00109-015-1344-5.
12. 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**:1929–35. https://doi.org/10.1126/science.1132939.
13. Musa A, Ghoraie LS, Zhang SD. *et al.* A review of connectivity map and computational approaches in pharmacogenomics. *Brief Bioinform* 2018;**19**:506–23. https://doi.org/10.1093/bib/bbw112.
14. Kim IW, Kim JH, Oh JM. Screening of drug repositioning candidates for castration resistant prostate cancer. *Front Oncol* 2019;**9**:661. https://doi.org/10.3389/fonc.2019.00661.
15. Huang YM, Cheng CH, Pan SL. *et al.* Gene expression signature-based approach identifies antifungal drug ciclopirox As a novel inhibitor of HMGA2 in colorectal cancer. *Biomolecules* 2019;**9**:9. https://doi.org/10.3390/biom9110688.
16. Choi B, Kang CK, Park S. *et al.* Single-cell transcriptome analyses reveal distinct gene expression signatures of severe COVID-19 in the presence of clonal hematopoiesis. *Exp Mol Med* 2022;**54**:1756–65. https://doi.org/10.1038/s12276-022-00866-1.
17. Zhang SD, Gant TW. A simple and robust method for connecting small-molecule drugs using gene-expression signatures. *BMC Bioinformatics* 2008;**9**:258. https://doi.org/10.1186/1471-2105-9-258.
18. Cheng J, Xie Q, Kumar V. *et al.* Evaluation of analytical methods for connectivity map data. *Pac Symp Biocomput* 2013;**22**:5–16.

19. Cheng J, Yang L, Kumar V. *et al.* Systematic evaluation of connectivity map for disease indications. *Genome Med* 2014;**6**:540. https://doi.org/10.1186/s13073-014-0095-1.
20. Parkkinen JA, Kaski S. Probabilistic drug connectivity mapping. *BMC Bioinformatics* 2014;**15**:113. https://doi.org/10.1186/1471-2105-15-113.
21. Brown AS, Patel CJ. A standard database for drug repositioning. *Sci Data* 2017;**4**:4. https://doi.org/10.1038/sdata.2017.29.
22. Chen B, Ma L, Paik H. *et al.* Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. *Nat Commun* 2017;**8**:16022. https://doi.org/10.1038/ncomms16022.
23. Duan Q, Reid SP, Clark NR. *et al.* L1000CDS(2): LINCS L1000 characteristic direction signatures search engine. *NPJ Syst Biol Appl* 2016;**2**:1–12. https://doi.org/10.1038/npjsba.2016.15.
24. Liu C, Su J, Yang F. *et al.* Compound signature detection on LINCS L1000 big data. *Mol BioSyst* 2015;**11**:714–22. https://doi.org/10.1039/C4MB00677A.
25. Qiu Y, Lu T, Lim H. *et al.* A Bayesian approach to accurate and robust signature detection on LINCS L1000 data. *Bioinformatics* 2020;**36**:2787–95. https://doi.org/10.1093/bioinformatics/btaa064.
26. Fortney K, Griesman J, Kotlyar M. *et al.* Prioritizing therapeutics for lung cancer: An integrative meta-analysis of cancer gene signatures and chemogenomic data. *PLoS Comput Biol* 2015;**11**:e1004068. https://doi.org/10.1371/journal.pcbi.1004068.
27. 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**:14621–6. https://doi.org/10.1073/pnas.1000138107.
28. McArt DG, Bankhead P, Dunne PD. *et al.* cudaMap: A GPU accelerated program for gene expression connectivity mapping. *BMC Bioinformatics* 2013;**14**:305. https://doi.org/10.1186/1471-2105-14-305.
29. Ye C, Ho DJ, Neri M. *et al.* DRUG-seq for miniaturized high-throughput transcriptome profiling in drug discovery. *Nat Commun* 2018;**9**:4307. https://doi.org/10.1038/s41467-018-06500-x.
30. 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**:1437–1452.e1417. https://doi.org/10.1016/j.cell.2017.10.049.
31. Zhang YX, Zhao YL. Pathogenic network analysis predicts candidate genes for cervical cancer. *Comput Math Methods Med* 2016;**2016**:3186051. https://doi.org/10.1038/s41467-018-06500-x.
32. Braun R, Cope L, Parmigiani G. Identifying differential correlation in gene/pathway combinations. *BMC Bioinformatics* 2008;**9**:488. https://doi.org/10.1186/1471-2105-9-488.
33. Dettling M, Gabrielson E, Parmigiani G. Searching for differentially expressed gene combinations. *Genome Biol* 2005;**6**:R88. https://doi.org/10.1186/gb-2005-6-10-r88.
34. Chan J, Wang X, Turner JA. *et al.* Breaking the paradigm: Dr insight empowers signature-free, enhanced drug repurposing. *Bioinformatics* 2019;**35**:2818–26. https://doi.org/10.1093/bioinformatics/btz006.
35. Giles FJ, Kantarjian HM, le Coutre PD. *et al.* Nilotinib is effective in imatinib-resistant or -intolerant patients with chronic myeloid leukemia in blastic phase. *Leukemia* 2012;**26**:959–62. https://doi.org/10.1038/leu.2011.355.
36. Manley PW, Brueggen J, Fabbro D. *et al.* Extended kinase profiling of the Bcr-Abl inhibitor nilotinib. *Proceedings of the American Association for Cancer Research Annual Meeting* 2007;**48**:772–3. https://doi.org/10.1186/gb-2005-6-10-r88.
37. Adams VR, Leggas M. Sunitinib malate for the treatment of metastatic renal cell carcinoma and gastrointestinal stromal tumors. *Clin Ther* 2007;**29**:1338–53. https://doi.org/10.1016/j.clinthera.2007.07.022.
38. Chintalgattu V, Patel SS, Khakoo AY. Cardiovascular effects of tyrosine kinase inhibitors used for gastrointestinal stromal tumors, *Hematol Oncol Clin North Am* 2009;**23**:97–107. https://doi.org/10.1016/j.hoc.2008.11.004.
39. Gnoni A, Marech I, Silvestris N. *et al.* Dasatinib: An anti-tumour agent via Src inhibition. *Curr Drug Targets* 2011;**12**:563–78. https://doi.org/10.2174/138945011794751591.
40. Schneider TC, Abdulrahman RM, Corssmit EP. *et al.* Long-term analysis of the efficacy and tolerability of sorafenib in advanced radio-iodine refractory differentiated thyroid carcinoma: Final results of a phase II trial. *Eur J Endocrinol* 2012;**167**:643–50. https://doi.org/10.1530/EJE-12-0405.
41. Corcoran RB, André T, Atreya CE. *et al.* Combined BRAF, EGFR, and MEK inhibition in patients with BRAF(V600E)-mutant colorectal cancer. *Cancer Discov* 2018;**8**:428–43. https://doi.org/10.1158/2159-8290.CD-17-1226.
42. Morabito A, Piccirillo MC, Falasconi F. *et al.* Vandetanib (ZD6474), a dual inhibitor of vascular endothelial growth factor receptor (VEGFR) and epidermal growth factor receptor (EGFR) tyrosine kinases: Current status and future directions. *Oncologist* 2009;**14**:378–90. https://doi.org/10.1634/theoncologist.2008-0261.
43. Klann K, Bojkova D, Tascher G. *et al.* Growth factor receptor Signaling inhibition prevents SARS-CoV-2 replication. *Mol Cell* 2020;**80**:164–174.e164. https://doi.org/10.1016/j.molcel.2020.08.006.
44. Dwivedy A, Mariadasse R, Ahmad M. *et al.* Characterization of the NiRAN domain from RNA-dependent RNA polymerase provides insights into a potential therapeutic target against SARS-CoV-2. *PLoS Comput Biol* 2021;**17**:e1009384. https://doi.org/10.1371/journal.pcbi.1009384.
45. Raymonda MH, Ciesla JH, Monaghan M. *et al.* Pharmacologic profiling reveals lapatinib as a novel antiviral against SARS-CoV-2 in vitro. *Virology* 2022;**566**:60–8. https://doi.org/10.1016/j.virol.2021.11.008.
46. Islam MA, Kibria MK, Hossen MB. *et al.* Bioinformatics-based investigation on the genetic influence between SARS-CoV-2 infections and idiopathic pulmonary fibrosis (IPF) diseases, and drug repurposing. *Sci Rep* 2023;**13**:4685. https://doi.org/10.1038/s41598-023-31276-6.
47. Wen L, Tang K, Chik KK. *et al.* In silico structure-based discovery of a SARS-CoV-2 main protease inhibitor. *Int J Biol Sci* 2021;**17**:1555–64. https://doi.org/10.7150/ijbs.59191.
48. Wan W, Zhu S, Li S. *et al.* High-throughput screening of an FDA-approved drug library identifies inhibitors against arenaviruses and SARS-CoV-2. *ACS Infect Dis* 2021;**7**:1409–22. https://doi.org/10.1021/acsinfecdis.0c00486.
49. Mallapaty S. The pathogens that could spark the next pandemic. *Nature* 2024;**632**:488. https://doi.org/10.1038/d41586-024-03886-1.
50. Nero TL, Parker MW, Morton CJ. Protein structure and computational drug discovery. *Biochem Soc Trans* 2018;**46**:1367–79. https://doi.org/10.1042/BST20180202.
51. Swamidass SJ. Mining small-molecule screens to repurpose drugs. *Brief Bioinform* 2011;**12**:327–35. https://doi.org/10.1093/bib/bbr028.
52. Ye H, Liu Q, Wei J. Construction of drug network based on side effects and its application for drug repositioning. *PLoS One* 2014;**9**:e87864. https://doi.org/10.1371/journal.pone.0087864.

Briefings in Bioinformatics, 2025, **26(1)**, bbae659
https://doi.org/10.1093/bib/bbae659
Problem Solving Protocol