OXFORD

*Briefings in Bioinformatics*, 2025, **26(3)**, bbaf226
https://doi.org/10.1093/bib/bbaf226
**Problem Solving Protocol**

# Large-scale information retrieval and correction of noisy pharmacogenomic datasets through residual thresholded deep matrix factorization

Zhiyue Tom Hu[1], Yaodong Yu[2], Ruoqiao Chen[3], Shan-Ju Yeh[4], Bin Chen[3,5,*], Haiyan Huang[6,*]

[1]Division of Biostatistics, University of California Berkeley, Berkeley, CA 94720, United States
[2]Department of Electrical Engineer and Computer Science, University of California Berkeley, Berkeley, CA 94720, United States
[3]Department of Pharmacology and Toxicology, Michigan State University, MI 48824, United States
[4]School of Medicine, National Tsing Hua University, Hsinchu 300044, Taiwan R.O.C.
[5]Department of Pediatrics and Human Development, Michigan State University, MI 48824, United States
[6]Department of Statistics, University of California Berkeley, Berkeley, CA 94720, United States
*Corresponding authors. Bin Chen, Department of Pharmacology and Toxicology, Michigan State University, MI 48824, United States. E-mail: chenbi12@msu.edu; Haiyan Huang, Department of Statistics, University of California Berkeley, Berkeley, CA 94720, United States. E-mail: hhuang@stat.berkeley.edu

## Abstract

Pharmacogenomics studies are attracting an increasing amount of interest from researchers in precision medicine. The advances in high-throughput experiments and multiplexed approaches allow the large-scale quantification of drug sensitivities in molecularly characterized cancer cell lines (CCLs), resulting in a number of open drug sensitivity datasets for drug biomarker discovery. However, a significant inconsistency in drug sensitivity values among these datasets has been noted. Such inconsistency indicates the presence of substantial noise, subsequently hindering downstream analyses. To address the noise in drug sensitivity data, we introduce a robust and scalable deep learning framework, Residual Thresholded Deep Matrix Factorization (RT-DMF). This method takes a single drug sensitivity data matrix as its sole input and outputs a corrected and imputed matrix. Deep matrix factorization (DMF) excels at uncovering subtle patterns, due to its minimal reliance on data structure assumptions. This attribute significantly boosts DMF's ability to identify complex hidden patterns among nuisance effects in the data, thereby facilitating the detection of signals that are therapeutically relevant. Furthermore, RT-DMF incorporates an iterative residual thresholding procedure, which plays a crucial role in retaining signals more likely to hold therapeutic importance. Validation using simulated datasets and real pharmacogenomics datasets demonstrates the effectiveness of our approach in correcting noise and imputing missing data in drug sensitivity datasets (open-source package available at https://github.com/tomwhoooo/rtdmf).

**Keywords:** pharmacogenomics datasets; drug sensitivity data; deep matrix factorization; noisy data; open-sourced

## Introduction

One goal of precision medicine is to choose the best therapy for individual cancer patients on the basis of individual molecular markers identified in clinical studies (as in [1–3]). At present, only some cancer drugs have approved biomarkers, and the process of identifying and validating a biomarker for a single drug in clinical trials takes many years [4, 5]. Emerging pharmacogenomics studies, in which drugs are tested against panels of molecularly characterized cancer cell lines (CCLs), have enabled the identification of several types of molecular biomarkers on a large scale by correlating drug sensitivity with the molecular profiles of pretreatment CCLs [6–10]. These biomarkers have the potential of identifying cell lines or even patients that will respond to a particular drug.

The datasets such as CCLE, GDSC, and CTRPv2 [7, 8, 11, 12] have made important contributions to pharmacogenomics research. Each drug sensitivity dataset can be represented as a single data matrix, where each row denotes a drug, each column symbolizes a CCL, and the values—often interpreted as drug responses—are summarized from dose–response curves. Commonly employed summarization metrics of the dose–response curves include IC50 (the concentration at which the drug inhibits 50% of maximum cell growth) and AUC (area under the dose–response curve). However, some of these matrices may contain a lot of missing data or incorrect zeros, making them difficult to use in subsequence analysis. Furthermore, re-evaluations of previously published pharmacogenomics data have exposed inconsistencies in drug response data across different studies. A comparative study by [13] between two datasets, CGP and CCLE, focusing on 15 shared drugs tested on 471 common cell lines, found that the vast majority of drugs displayed poor concordance. This inconsistency raises concerns regarding their applicability in biomarker discovery and brings into question the reliability of such data for drug discovery, as noted in [13] and [14].

Numerous attempts have been made to address this inconsistency problem. Most of the proposed ideas focused on forming better summarization metrics and/or standardizing experiments

**Received:** June 06, 2024. **Revised:** January 03, 2025. **Accepted:** April 28, 2025

and data processing pipelines (e.g. [15–17]). Although these methods have improved the cross-study consistency of drug response measures to some extent, there is still a vast inconsistency of drug response signals across the studies. The study by [18] proposed a novel semi-supervised computational method named AICM, designed to correct sensitivity data matrices. This method was inspired by an intriguing observation: when comparing the same CCLs across different datasets, the results are usually consistent, with high correlations. However, this consistency is not observed when examining the effects of the same drug across different datasets, where correlations are low. These findings suggest that while there is some level of agreement between various studies, there is also considerable noise or error in the data. AICM leverages the consistent correlations in CCLs to correct and impute missing data across different datasets. A limitation of this approach is that it requires datasets with a substantial overlap in both drugs and CCLs to be effective. As newer studies increasingly focus on specific types of cancer or particular drug libraries, this limitation becomes more significant. Therefore, rather than relying on shared information from overlapping sections of datasets, it is essential to develop a method that individually corrects each dataset by thoroughly examining its underlying structure. This approach should surpass traditional methods by capturing subtle and intricate patterns and signals. Additionally, it must accommodate large amounts of missing data and be scalable to larger datasets. In this paper, we present Residual Thresholded Deep Matrix Factorization (RT-DMF), a novel computational framework that operates in a fully unsupervised manner and aims to reveal hidden signals within individual datasets while effectively filtering out noise. Specifically, RT-DMF seeks to correct the noise in existing drug sensitivity data rather than predict sensitivity values for new chemicals and cell lines using additional data beyond the drug-sensitivity data, as done in [19, 20] Concurrently, an unsupervised approach combining linear matrix factorization (MF) with a nonlinear neural network has been proposed in [21]. While the inherent flexibility of nonlinear neural networks enables the identification of diverse hidden patterns, it also complicates the distinction between noise and meaningful signals during model training. This often results in overfitting, particularly when there is limited guidance for differentiating between noise and therapeutically significant signals.

RT-DMF is developed with the understanding that therapeutically significant interactions between drug and CCLs are sparse in drug sensitivity data matrices, leading to the assumption that most patterns identified by MF methods are likely to represent nuisance drug–CCL relationships. This framework takes a single drug sensitivity data matrix as its sole input and outputs a corrected and imputed matrix. RT-DMF has two key features: (i) It advances traditional MF methods by employing a wide linear neural network architecture that facilitates thorough exploration of the underlying structure during optimization. This eliminates the need to assume that the data matrix for nuisance effects is the inner product of two low-rank matrices, thereby allowing for a more effective representation of complex nuisance structures or patterns in the data. This, in turn, aids in the detection of signals that are more therapeutically relevant. (ii) It incorporates a residual filtering process to recover valuable data that might otherwise be categorized as outliers. Through extensive experiments, we have demonstrated that RT-DMF not only denoises individual datasets but also enhances consistency across multiple datasets. Furthermore, we have shown that RT-DMF provides significant benefits for various downstream analyses. Specifically, the data corrected and imputed by RT-DMF enhance two downstream tasks: drug sensitivity prediction and clustering of drugs that share similar mechanisms of action (MOAs). Interestingly, we also observed that the threshold parameters selected based on simple validation errors during RT-DMF training align closely with optimal settings in downstream tasks such as transfer learning, offering a straightforward yet effective approach for parameter selection. Based on these results, we believe that RT-DMF has the potential to become a standard processing pipeline for future data sets.

# Method

## Method overview

Modeling the collaborative relationships between drugs and CCLs is structurally and numerically complex, regardless of whether these relationships are nuisance or therapeutically interesting. In ideal situations with discernible patterns in datasets, we may observe randomly distributed "blocks." These blocks suggest scenarios where a group of drugs similarly affects a set of CCLs. The size of these groups can vary; for instance, nonselective drugs active across many cell lines form larger blocks, possibly indicating nuisance drug–cancer relationships. On the other hand, specialized drugs might be effective on a specific small subset of CCLs, resulting in smaller blocks, potentially signifying therapeutically relevant relationships.

Moreover, some drugs may have multiple modes of action, targeting different CCL groups, while certain CCLs could be influenced by diverse drug groups acting on distinct pathways. This implies that some drugs or CCLs might belong to multiple groups or blocks. "Stripes" may also appear, e.g. from cytotoxic drugs that show efficacy across most CCLs. Additionally, hidden nonblock patterns and various types of noise likely exist. For instance, a group of drugs targeting a common mutation in CCLs may show inconsistent efficacy due to each drug's unique and intricate MOA.

Explicit models face challenges in effectively capturing the above patterns and differentiating between nuisance and therapeutically relevant drug–CCL relationships, particularly in the presence of data noise. Therefore, a model with fewer structural assumptions but greater capability to uncover hidden nuisance patterns, while remaining robust against noise, is considered to have more promise for this task.

Considering that therapeutically relevant drug–CCL relationships are sparse in a drug sensitivity data matrix, it is reasonable to assume that patterns identified through MF methods would mostly represent nuisance drug–CCL relationships. In this study, we introduce a new method based on deep matrix factorization (DMF) to isolate these nuisance patterns within drug sensitivity data. We compared DMF with classical low-rank MF methods and found that the DMF-based approach has advantages in extracting these nuisance patterns. This extraction can, in turn, facilitate the identification of signals that are more therapeutically relevant.

Classical low-rank MF aims to approximate a data matrix $\mathcal{D} \in \mathbb{R}^{n \times p}$ as $\mathcal{D} \approx M_1 M_2$, under the assumption that $M_1$ and $M_2$ are two low-rank matrices. While this approach can be useful and provide interpretable patterns for straightforward interactions, it falls short when faced with more complex data structures, as described earlier regarding possible groups and blocks.

The primary motivation for DMF lies in its ability to combine (i) the interpretability and robustness found in classical MFs, with (ii) the capacity to extract patterns of varied signal levels and intricate structures as enabled by multilayer architectures. Single-layer matrix approximations struggle to capture patterns with

widely varying signal levels in complex datasets. DMF decomposes a data matrix $\mathcal{D} \in \mathbb{R}^{n\times p}$ as $\mathcal{D} \approx M_1 M_2 \dots M_L$, where $L$ represents the number of layers. The approximation achieved by DMF can be understood as a series of successive factorizations of $\mathcal{D}$:

$$\mathcal{D} \approx M_1 D_1, \quad D_1 \approx M_2 D_2, \quad D_2 \approx M_3 D_4, \quad \dots, \quad D_{L-1} \approx M_{L-1} D_L.$$

Here, each matrix $M_l (l = 1, \dots, L)$ can be interpreted as the feature matrix of layer $l$. Consequently, the matrix $D_l (1 \le l \le L)$ is successively factored so that key linear combinations of the principal features of the $l$th layer appear in the subsequent $M_{l+1}$. This enables a wealth of interpretations concerning the semantics concealed within the data.

Another key feature of the DMF approximation is its optimization approach. By initializing the matrices $M_l$ (where $l = 1, \dots, L$) as high-dimensional matrices, DMF minimizes the assumptions it makes about the underlying structure. This, however, introduces the challenge of navigating an unidentifiable search space. Despite this obstacle, the use of gradient descent methods effectively regularizes the solution path, enabling DMF to autonomously arrive at a low-rank version of $M_l$'s. This phenomenon of implicit regularization—where the model optimizes itself toward a simpler, more generalizable solution—is well-studied, both theoretically and experimentally, in various research papers (e.g. [22, 23]).

In contrast to classical MF methods, which operate under stronger structural assumptions and have fewer parameters, thereby increasing the likelihood of achieving global optimization, DMF offers potentially greater fidelity to the underlying data. This is because it operates in a more relaxed, larger search space. Although global optimization is unlikely achievable, the method's greater flexibility increases the odds of finding solutions that are closer to the true underlying structure.

After applying DMF to a drug sensitivity data matrix, as we argued earlier, we assume that the DMF approximation captures structured nuisance signals. The residuals are likely composed of remaining random or unstructured nuisance noise, as well as "true" or therapeutically relevant signals. In different applications, significant deviations from DMF approximations might be dismissed as outliers or noise. However, in the context of drug sensitivity data, these substantial deviations could indicate important signals. For instance, a small number of tested CCLs might have a unique mutation that is specifically targeted by a drug, leading to "spikes" in the matrix. To preserve this crucial information, we incorporate a residual thresholding (RT) step. This process starts with identifying high residuals, marked by deviations exceeding a predetermined threshold. We then examine if the CCLs associated with these high residuals under the same drug treatment are overrepresented in known CCL groups, such as those categorized by tissue type, cancer type, or cancer subtype. We base this approach on the hypothesis that high residuals enriched within specific tissue or disease groups are more likely to hold therapeutic significance.

In summary, RT-DMF decomposes signals in a drug sensitivity data matrix by (i) using DMF to extract and remove structured nuisance signals; (ii) applying thresholding to the residuals to separate out potentially therapeutically significant signals; and, (iii) within the thresholded residuals that show substantial deviations from DMF approximations, pinpointing those that are prevalent in known tissue or disease categories. Signals identified in Step (iii) are regarded as having therapeutic interest. The final output is then the DMF approximation matrix with the high residuals identified in Step (iii) added back to the corresponding entries.

For Step (iii), in situations where reliable CCL groupings are unavailable, we suggest a more straightforward method to retain these notably deviating values: preserving values that show abrupt changes in a single training iteration. The threshold value and number of iterations can be treated as hyperparameters and tuned using a hold-out validation set during training. We provide more details on this methodology in the subsequent algorithm section.

## Algorithm

Building on the ideas outlined in the previous section, the RT-DMF algorithm is presented below. More details regarding the implementation of RT-DMF can be found in the Github repository.

**Algorithm 1** Deep Matrix Factorization with Residual Filtering

**Hyperparameter:** number of hidden layers $l$, width of each hidden layer $(w_1, w_2, \dots, w_l)$, initialization parameter $(\sigma)$, learning rate $(\lambda)$, and residual thresholding parameter $(\eta)$.
**Input:** a data matrix with $n$ rows and $p$ columns, we denote as $\mathcal{D} \in \mathbb{R}^{n\times p}$.
**Initialization:** we randomly initialize a sequence of $l+1$ matrices, $M_1, M_2, \dots M_{l+1}$, whose each entry draws from a $\mathcal{N}(0, \sigma)$ distribution. These matrices are of the size $n \times w_1$, $w_1 \times w_2$, …, $w_l \times p$, respectively. We denote the final output of these matrices multiplied together as $\hat{M} = M_1 \cdot M_2 \cdot \dots \cdot M_{l+1}$, which is of the dimension $n \times p$.
**for** Until Convergence **do**
**Forward step:** calculate $\hat{M}$ by matrix multiplying the small matrices together.
**Backward step:** calculate the partial gradient of the matrix norm of the difference between our observed matrix and fitted matrix $\nabla_{M_1,\dots,M_{l+1}} \|\|\mathcal{D} - \hat{M}\|\|$. Update each matrix by gradient descent, i.e. $M_i = M_i + \lambda \nabla_{M_i} \|\|\mathcal{D} - \hat{M}\|\|, \forall i$.
**Threshold:** for $\{i, j\} \in \mathcal{D}$, we set $M_{ij}$ to $D_{ij}$ if $|\mathcal{D}_{ij} - M_{ij}| \ge \eta$
**end for**

# Simulation results

## Overview

Unless otherwise specified, we denote a particular drug sensitivity matrix as $\mathcal{D} \in \mathbb{R}^{n\times p}$, meaning that this dataset has $n$ drugs tested on $p$ CCLs. We denote the row index of $\mathcal{D}$ as $r_1, r_2, \dots, r_n$ and column indices as $c_1, c_2, \dots c_p$. $\mathcal{D}[i, j]$ denotes the single entry on the $i$th row and $j$th column of the matrix $\mathcal{D}$, which is a summarized value of dose–response of drug $i$ on CCL $j$.

Given that our ultimate goal is to uncover signals in an unsupervised manner, we attempt to generate synthetic datasets that incorporate the complexity of drug–CCL interactions as we previously discussed to verify the effectiveness of the methods. We evaluate the performance of RT-DMF on these synthetic datasets by comparing it with the most widely used MF methods: convex matrix completion in [24] and robust principal component analysis (RPCA) in [25]. Note that the same RT techniques are used in these benchmark methods for fair comparison.

We generate our synthetic data using the following scheme:

$$\mathcal{D} = \mathcal{D}_g^1 + \mathcal{D}_g^2 + \Sigma, \tag{3.1}$$

where $\mathcal{D}_g^1$ and $\mathcal{D}_g^2$ denote different kinds of "structural information." $\mathcal{D}_g^1$ denotes more general interaction information, such as the effect of general type/group of drug(s) on a relatively large group of CCLs, thereby representing nuisance drug–CCL relationships. More specifically, we model $\mathcal{D}_g^1$ as a low-rank matrix that can be recovered as the product of multiple matrices. As we discussed earlier, modeling the nuisance interactions between drugs and CCL as the simple inner product of two matrices would be an oversimplification because such interactions are complex. Therefore, we parameterize it as a multiplication of multiple wide matrices to mimic a dataset with patterns of varied signal levels and complicated structures. Nevertheless, we also provide a case where $\mathcal{D}_g^1$ is a simple multiplication of two low-rank matrices, which is consistent with the assumptions of other benchmark methods. We model $\mathcal{D}_g^2$ as comprising combinations of "extreme values." These effects may result from a drug's strong toxicity on the majority of CCLs, which can create a row stripe in the matrix. Additionally, they can arise from signals corresponding to more specific drug–CCL interactions, which are distinct from the nuisance effects in $\mathcal{D}_g^1$ and may appear as tiny individual blocks or overlapping small blocks in the matrix. In other words, $\mathcal{D}_g^2$ can consist of high residuals that are enriched in specific tissue or disease groups of CCLs (extreme values), indicative of signals with greater therapeutic interest. We will introduce various models of $\mathcal{D}_g^2$ in the next subsection.

In summary, as previously discussed, we anticipate that $\mathcal{D}_g^1$ will represent nuisance interactions between drugs and CCLs. In contrast, $\mathcal{D}_g^2$ contains information about more specific or extreme drug–CCL interactions. Specific drug–CCL interactions are likely to be therapeutically relevant. It is important to note that accurately recovering $\mathcal{D}_g^2$ depends largely on the successful recovery of $\mathcal{D}_g^1$.

$\Sigma$ represents the noise introduced either by experimental protocols or other lab environmental factors. In the RT-DMF framework, DMF is utilized to effectively recover $\mathcal{D}_g^1$. The RT procedure is then employed to recover $\mathcal{D}_g^2$ from the residuals. The remaining residuals account for the noise as described in $\Sigma$. Generally, separating $\mathcal{D}_g^1, \mathcal{D}_g^2$ and $\Sigma$ is a challenging task. As we will demonstrate later, careful tuning of parameters can provide a viable approach to achieve this separation.

## Generation of the synthetic datasets

We outline the generation models for $\mathcal{D}_g^1, \mathcal{D}_g^2$ and $\Sigma$ below, which have dimensions of 800 × 400 (i.e. 800 drugs and 400 CCLs). Detailed information about the generation process can be found in the appendix.

- **Generation of** $\Sigma$. Each component in $\Sigma$ is sampled from a normal distribution $\mathcal{N}(0, \sigma^2)$.
- **Generation of** $\mathcal{D}_g^1$. We generate $D_g^1$ using two models: the first is a product of two nonnegative low-rank matrices, while the second is a product of multiple, such as four, low-rank matrices.
- **Generation of** $\mathcal{D}_g^2$. We consider three different block structures to generate $\mathcal{D}_g^2$: (i) **Simple two-block**. This models an ideal situation where a small group of drugs appears effective for a small group of CCLs. We generated two blocks, each with dimensions of 20 × 20, with nonoverlapping drugs or CCLs. A zoomed-in visualization of this block structure within a submatrix of $\mathcal{D}_g^2$ is shown in Fig. 1(d). Prior to applying RT-DMF, we shuffled the indices to ensure that the blocks are not immediately apparent in the data. (ii) **Mixed overlapping blocks.** This models more complex scenarios in which CCLs may respond to different drug mechanisms, resulting in blocks with overlapping CCLs as depicted in Fig. 2(d). In addition, we consider the case where there are possible big categories of cancer: that certain CCLs might respond to almost all drugs, resulting in vertical stripes in the data representation (see column 15 in the figure). (iii) **Real data residual.** This model generates a synthetic data matrix by utilizing clustered residuals from the FIMM data [15], combined with the naive low-rank approximation of the FIMM data matrix. This provides a simplified representation of real data, but with the structures of both nuisance and potentially therapeutic signals known. A visualization of a sub-matrix of $\mathcal{D}_g^2$ can be seen in Fig. 3(d).

## Comparison methods and results

We compare RT-DMF with benchmark methods including the simple convex MF proposed by [24], the RPCA proposed by [25], and the Dual Branch Deep Neural Matrix Factorization (DBDNMF) proposed by [21]. A more comprehensive description of these three methods can be found in the appendix.

For simple MF, the crucial tuning parameter is $\epsilon$, which controls the similarity between the fitted and original matrices and determines the rank of the fitted matrix. Since this basic benchmark yields inferior results. We have placed the results of the simple MF in the appendix. In RPCA, the key tuning parameter is $\tau$; as $\tau$ increases, the fitted matrix becomes sparser. In DBDNMF, the key tuning parameter is $\alpha$; as $\alpha$ trades off between linear and nonlinear signals. For RT-DMF, we adjusted the number of epochs (with a fixed learning rate) and threshold parameter $\eta$, balancing the fit to the original matrix against retaining structural low-rank information. As the number of epochs increases, the matrix recovered through RT-DMF has a higher rank to better fit the original matrix. All these parameters represent a trade-off between training error and the assumed structural sparsity.

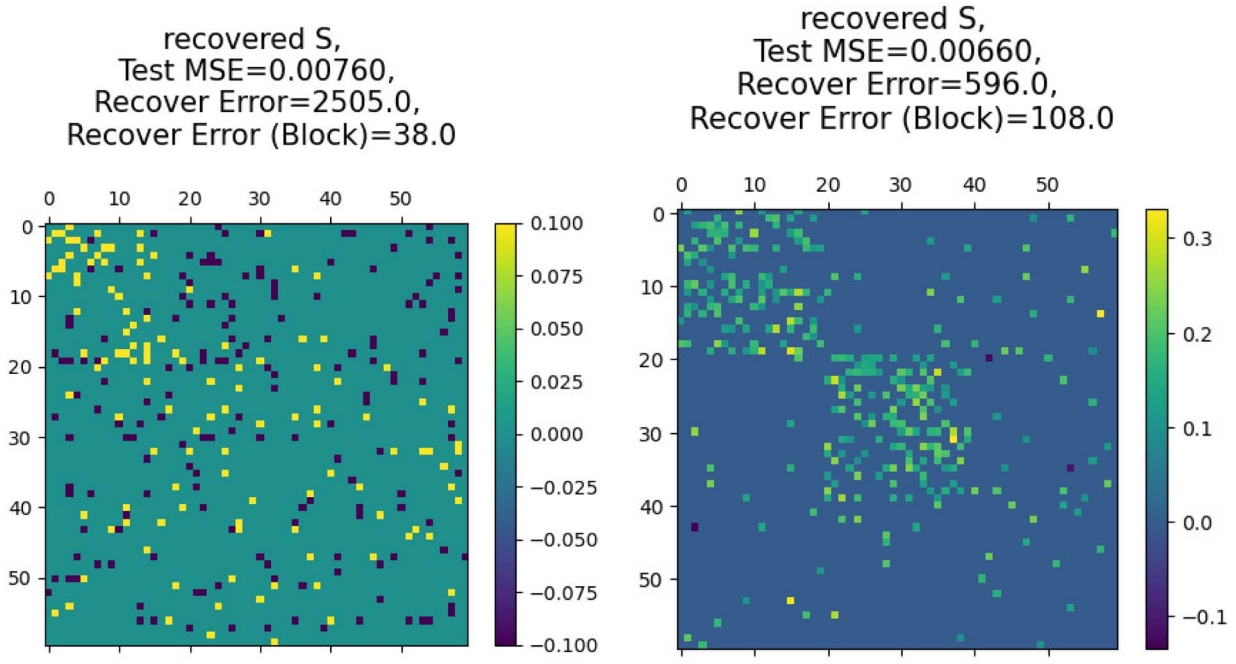
For all four methods, we tune the parameters based on the validation mean square error (v-MSE). Specifically, we use only 70% of the data for training. The remaining 30% of the data in the matrix is not utilized during the optimization process. Of this, 15% is used as validation entries for parameter selection, and the other 15% serves as test entries for evaluating the model's performance using the test MSE. We select the parameter that yields the best v-MSE for testing purposes.

We assess the performance of the methods using three different metrics: test mean square error (MSE), recovery error (RE), and in-block recovery error (IBRE). Test MSE refers to the MSE calculated on the test entries. RE quantifies the overall difference between the recovered matrix and the original matrix. IBRE measures the difference specifically for entries that are nonzero in $\mathcal{D}_g^2$. IBRE is particularly important, as our primary interest lies in determining whether these methods can accurately recover meaningful drug–CCL interactions, which we primarily characterize as "tiny blocks and stripes" in $\mathcal{D}_g^2$.

The comparative results for three synthetic datasets are presented in Figs 1, 2, and 3. These datasets were generated using the three models for $\mathcal{D}_g^2$ and a product of two low-rank matrices for $\mathcal{D}_g^1$ as described in the appendix. Additional studies on synthetic data, utilizing a different model for $\mathcal{D}_g^1$ and varying block sizes, are available on GitHub.

In the case of the synthetic data with the simple two-block model for $\mathcal{D}_g^2$ depicted in Fig. 1, RT-DMF consistently surpasses all other methods in terms of RE and IBRE, even though the other methods also demonstrate satisfactory performance.

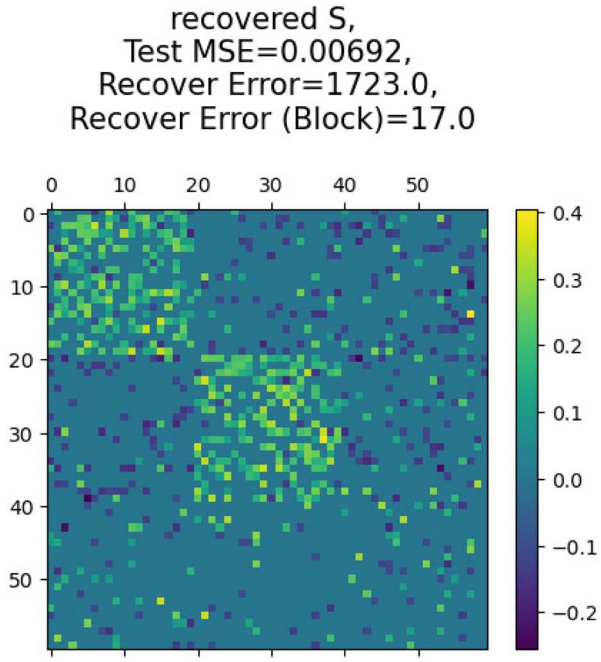
In the case of the synthetic data with mixed overlapping blocks for $\mathcal{D}_g^2$, as shown in Fig. 2, both MF and DBDNMF nearly completely fail to perform effectively, likely because the data do not align with



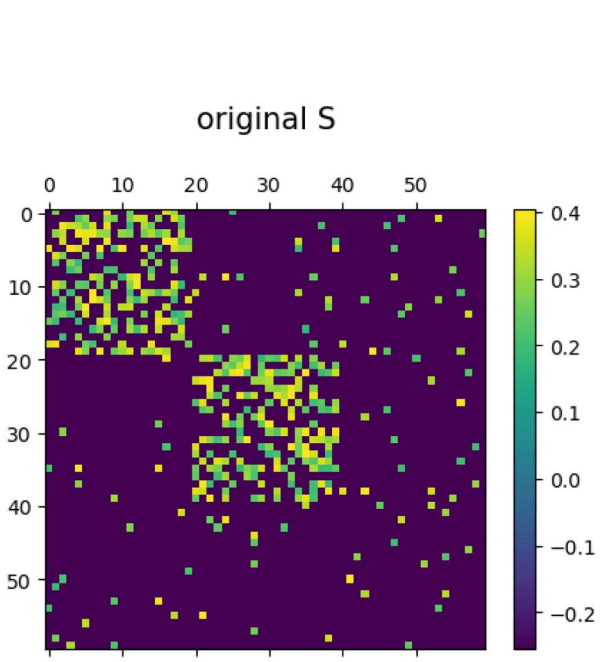

(a) Recovered $S$ with DBDNMF




(b) Recovered $S$ with RPCA




(c) Recovered $S$ with RTDMF



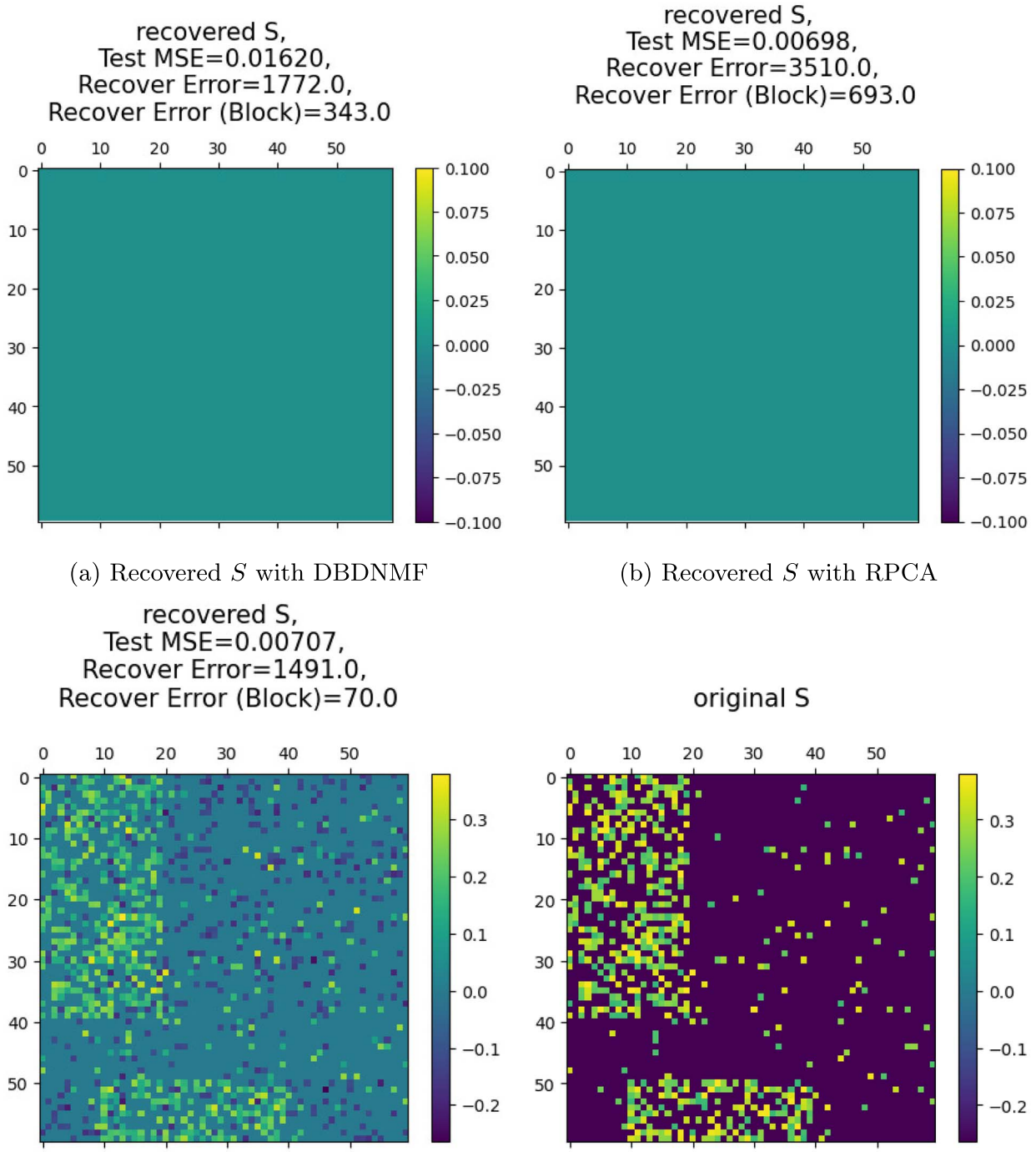

(d) Ground truth $S$ generated according to the simple two-block model

Figure 1. $\mathcal{D}_g^2$ model 1: simple two-block model.

the underlying assumptions of MF and DBDNMF. However, this synthetic dataset represents an important real-world scenario: CCLs might respond to various drug mechanisms, leading to blocks with overlapping CCLs. Similarly, drugs that share a common mode of action could have mechanisms that overlap other drugs, resulting in blocks with overlapping drugs. The results of this dataset also imply that MSE may not be a suitable metric for comparing methods in the context of our study. Despite the inability of MF and RPCA to recover any substantial information, they still manage to achieve lower test MSE compared with RT-DMF.

For the synthetic data incorporating real data residuals shown in Fig. 3, RT-DMF achieves impressively low IBRE, indicating its ability to recover true signals even in the presence of relatively complex noise.

These simulation studies emphasize RT-DMF's capability to distinguish and retrieve meaningful signals from complex data across diverse scenarios. This makes RT-DMF an invaluable tool




(a) Recovered $S$ with DBDNMF

(b) Recovered $S$ with RPCA

(c) Recovered $S$ with RTDMF

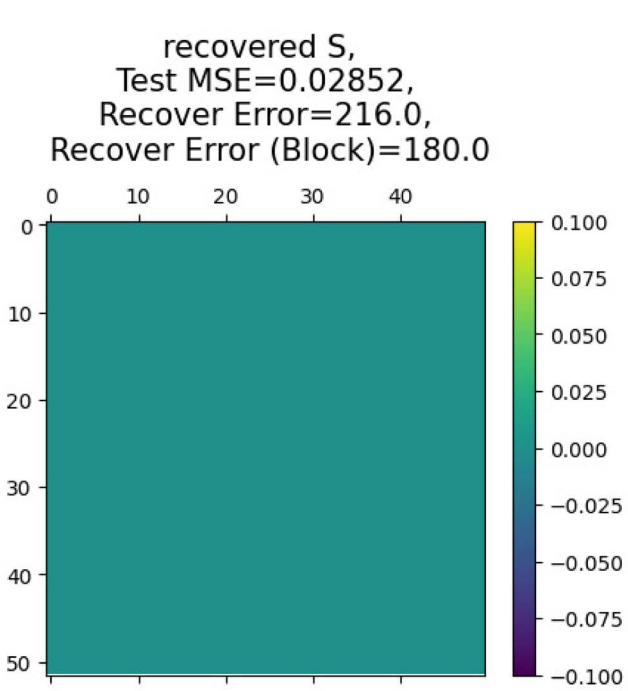
(d) Ground truth $S$ generated according to the mixed overlapping block model

Figure 2. $\mathcal{D}_g^2$ model 2: mixed overlapping block model.

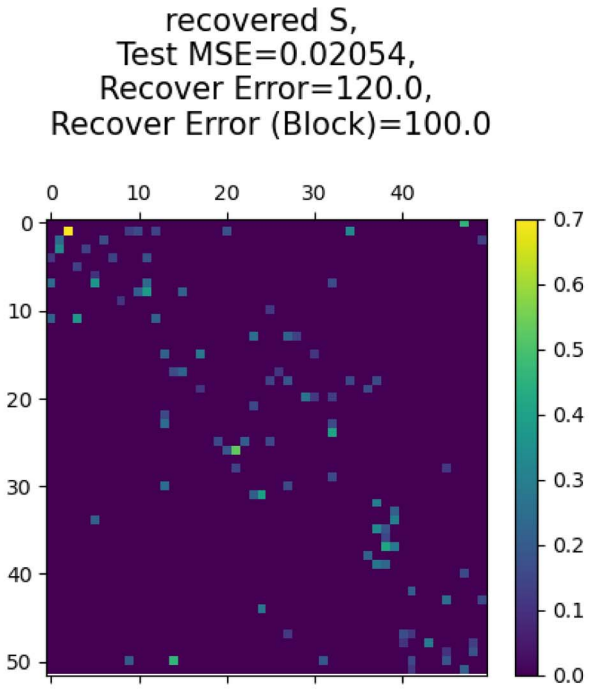
in the analysis of drug sensitivity data, where discerning subtle yet important signals is crucial for identifying potentially effective therapeutic interventions.

## Real data results

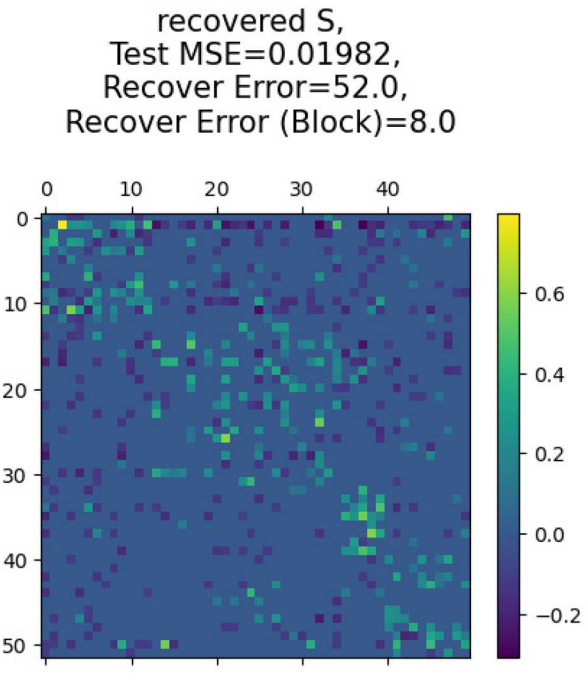
We present our results from multiple perspectives to show that RT-DMF successfully retains meaningful information in real data. We apply RT-DMF to five relatively large and well-prepared datasets individually: GDSC1000, CTRPv2, FIMM, GRAY, and PRISM. For each pair of the datasets, we use the overlapping part to calculate the correlation between the drug response profiles in the two datasets for each overlapping drug. We show that the percentage of significantly correlated drugs increases after RT-DMF is applied. We then present some case studies of individual drugs on how RT-DMF can reveal meaningful information after correcting the data. We also demonstrate the validity of RT-DMF on cell lines that have known mutations. Lastly, we demonstrate its effectiveness on a downstream analysis scenario to show that not only the correction part of RT-DMF provides meaningful



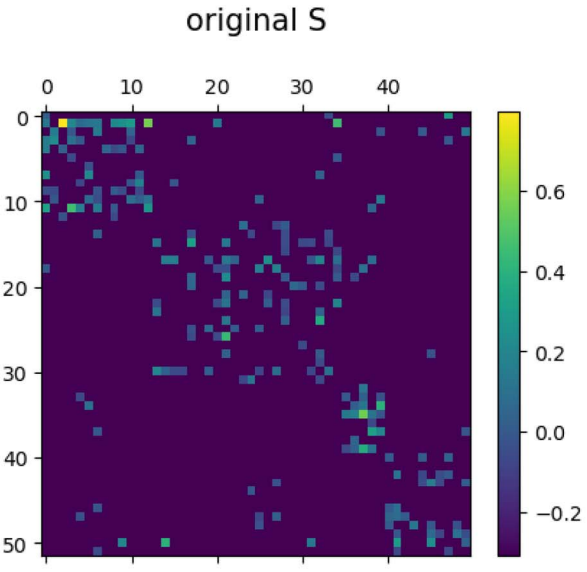

(a) Recovered $S$ with DBDNMF




(b) Recovered $S$ with RPCA




(c) Recovered $S$ with RTDMF




(d) Ground truth $S$ generated according to the real data residual model

Figure 3. $\mathcal{D}_9^2$ model 3: residual fetched from co-clustered FIMM residual.

information, but also the imputation part can help improve downstream tasks.

## Increase of significantly correlated drugs

We use a two-tailed Spearman's rank correlation test presented in [26] with $P$-value 0.01 to determine if a particular drug's Spearman's rank correlation is significant across two datasets. In Table 1, we summarize the percentage of drugs that show significant Spearman's correlation in the original and corrected datasets. For comparison, we denote the correlation before correction with a parenthesis under the correlation after correction. We observe that in general, RTDMF increases the percentage of drugs that show significant correlation across different datasets. Notably, over 20% more drugs are correlated between GDSC1000

Table 1. Percentage of drugs that show significant Spearman's correlation before (values in parentheses) and after (values above) the application of RT-DMF

| | GDSC1000 | CTRPv2 | FIMM | GRAY | PRISM |
|---|---|---|---|---|---|
| GDSC1000 | – | 83.3% (62.5%) | 43.8% (18.8%) | 42.1% (21.1%) | 47.5% (35.6%) |
| CTRPv2 | – | – | 66.7% (60%) | 50% (35.7%) | 57.6% (49.1%) |
| FIMM | – | – | – | 5.26% (5.26%) | 6.67% (6.67%) |
| GRAY | – | – | – | – | 28.8% (18.6%) |
| PRISM | – | – | – | – | – |

and CTRPV2, two highly cited pharmacogenomics datasets. However, one might notice there are a few exceptions occurring at FIMM and GRAY. This is because they are very small datasets, and hence have very few overlapping cell lines with other datasets in general. For example, FIMM and GRAY only have 11 overlapping cell lines for the computation of Spearman's rank correlation. The critical value gets more stringent when the vector size gets smaller, namely two drugs need to show extremely strong Spearman correlation to be considered significantly correlated when the vector size is small. Nevertheless, for datasets that contain abundant drugs and CCL's, such as GDSC1000, CTRPv2, and PRISM, we observe a considerable improvement trend. We also present scatter plots for some drugs individually for a better illustration of the performance of RT-DMF in the appendix.

We present a typical biomarker discovery case to demonstrate the utilization of the corrected data. It is known that BRAF mutation is a biomarker to predict the efficacy of Dabrafinib, a BRAF inhibitor; hence we expect to see a difference in Dabrafinib's sensitivity between BRAF-mutated and wild-type cell lines (as denoted by 1 and 0). We observe that in Fig. 4, after RT-DMF is applied, the difference between the two groups in GDSC becomes even more profound compared with the original. This is specifically demonstrated that the wild-type group in RT-DMF processed data is much less varied than the original data, which leads to better separation. This demonstrates that the correction of RT-DMF does provide biologically more meaningful data.

## Application of drug-sensitivity data on other tasks

### *Drug sensitivity prediction*

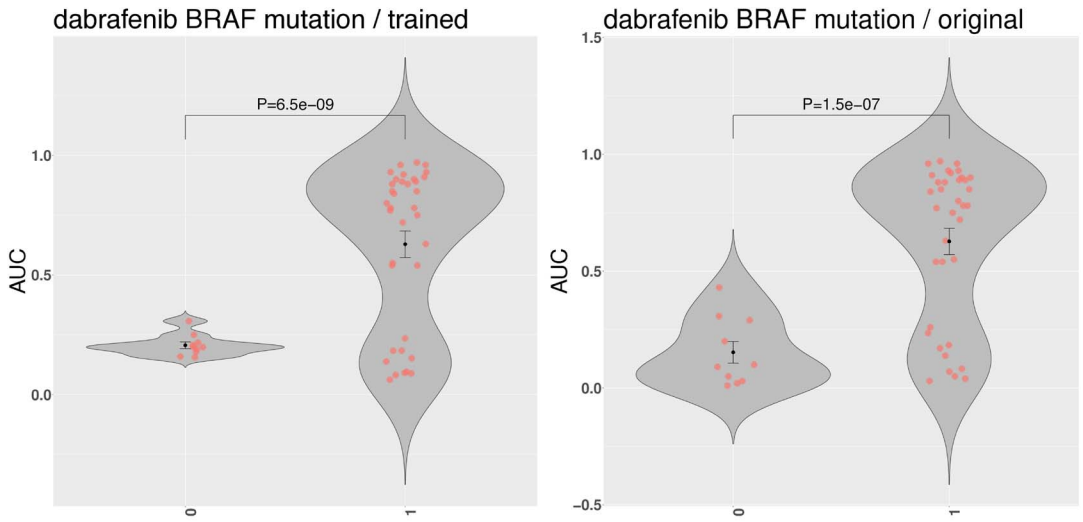
Cell line gene expression profiling is performed routinely in the laboratory, while testing the sensitivity of a large number of drugs in a cell line is costly and time-consuming. It would be helpful if a machine learning model could predict drug sensitivity in a cell line based on its gene expression profile. However, the model's performance may be limited by the poor quality of existing drug sensitivity data used to train it. We first investigated the performance of RT-DMF imputation compared with simple K-NN imputation in the elastic net algorithm as in [27]. RT-DMF demonstrated superior performance, reducing the root mean squared error (RMSE) from 0.023 to 0.009 in CTRP and from 0.047 to 0.031 in GDSC, while also decreasing variation. To further evaluate the impact of corrected datasets, we employed a state-of-the-art deep transfer learning model as in [28] on the CTRP and GDSC datasets. After preprocessing to identify common cell lines, our final dataset comprised 722 samples from CTRP and 596 samples from GDSC. We randomly selected 100 drugs from each dataset (CTRP and GDSC) and developed drug sensitivity prediction models using CCLE gene expression profiles. Using RMSE as our evaluation metric, five-fold cross-validation showed that RT-DMF achieved an average RMSE of 0.035 (95% CI: 0.031-0.040) for CTRP predictions (Fig. 5). For the GDSC dataset, a threshold setting of 0.2 achieved the best performance, yielding an average RMSE of 0.055 (95% CI: 0.048-0.063) in five-fold cross-validation (Fig. 5). It is worth noting that this optimal threshold value was independently confirmed by our hold-out validation dataset during model training. Further details regarding threshold selection are provided in the appendix. In brief, for both datasets, RT-DMF yielded corrected datasets that led to better predictive models than those using the original datasets.

### *Clustering analysis*

Another application of drug sensitivity data lies in providing insights into the relations among drugs or cell lines via clustering analysis [29]. Such an analysis facilitates the grouping of drugs sharing similar MOAs, which can foster the identification of alternative treatments, as well as facilitate the discovery of new drugs and targets. On the other hand, clustering cell lines with similar drug response patterns creates opportunities for exploring underlying biological mechanisms and discovering potential personalized therapies by predicting individual patient responses using cell line data.

To demonstrate that RT-DMF could improve clustering accuracy, we performed clustering analysis using RT-DMF-corrected and the original drug sensitivity data from GDSC and CTRP. The missing values in the original data were imputed by KNN-Imputer before clustering as in [27]. We compared the clusters of drugs/cell lines with their predefined categories, and quantitatively evaluate the clustering performance by computing cluster accuracy (CA) as in [30].

For the drug clustering, we utilized GDSC drugs that had predefined MOAs associated with epidermal growth factor receptor (EGFR) signaling and metabolism, and CTRP drugs whose MOAs were defined as inhibitors of IKK-2 and gamma-secretase. These paths and proteins are distinctly characterized by different functions, mechanisms, and associated diseases. EGFR signaling primarily regulates cell growth, proliferation, and survival through signal transduction cascades, whereas metabolism plays a role in energy production, molecule biosynthesis, and cellular homeostasis. Similarly, while IKK-2 primarily controls NF-$\kappa$B signaling and participates in inflammation and immune responses, gamma-secretase is responsible for proteolytic processing of transmembrane proteins, playing a significant role in A$\beta$ production (associated with Alzheimer's disease) and Notch signaling. Given this distinctiveness, we anticipated that these selected drugs would




(a) CCL's value distribution in GDSC after RT-DMF is applied.

(b) CCL's value distribution in GDSC before RT-DMF is applied.

Figure 4. We can see that the 0 (non-mutation group) and 1 (mutation group) have a more significant difference after RT-DMF is applied to GDSC.

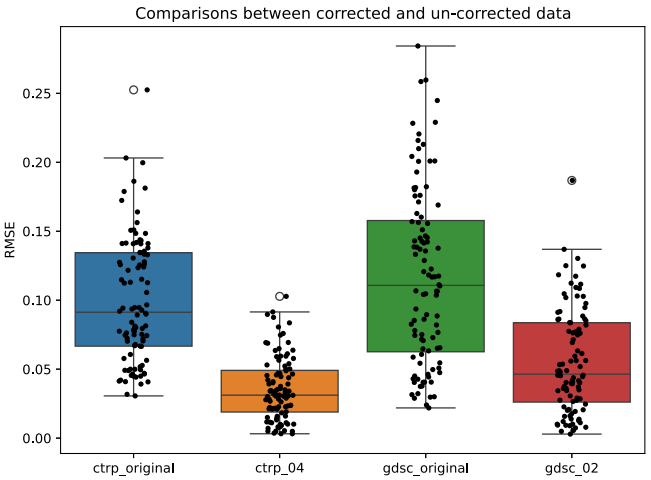



Figure 5. Comparison of RMSE between predicted and actual drug sensitivity values across GDSC and CTRP datasets; RT-DMF-corrected datasets demonstrate superior prediction accuracy compared with baseline methods.

display unique patterns between categories. Our findings confirmed this, as we observed a CA of 0.571 for the original GDSC drug data and 0.714 for the RT-DMF-corrected data, and a CA of 1 for both original and corrected CTRP drug data as demonstrated in 6a and 6b.

For cell line clustering, we used GDSC CCLs sourced from chondrosarcoma, chronic myeloid leukemia, and B cell leukemia, and CTRP CCLs sourced from acute lymphoblastic B cell leukemia and Grade IV astrocytoma. These represent distinctly different cancers with diverse cell and tissue origins. Acute lymphoblastic B cell leukemia is a blood cancer originating from B cells, whereas Grade IV astrocytoma is a brain tumor originating from astrocytes. Similarly, chondrosarcoma is a bone cancer that begins in chondrocytes, chronic myeloid leukemia arises from myeloid cells, and B cell leukemia originates from B cells. Based on these differences, we anticipated unique patterns between these CCLs. This expectation was met with a CA of 1 for both original and RT-DMF-corrected CTRP data, and a CA of 0.630 for original GDSC data and 0.704 for the RT-DMF-corrected data as demonstrated in Fig. 6c and 6d.

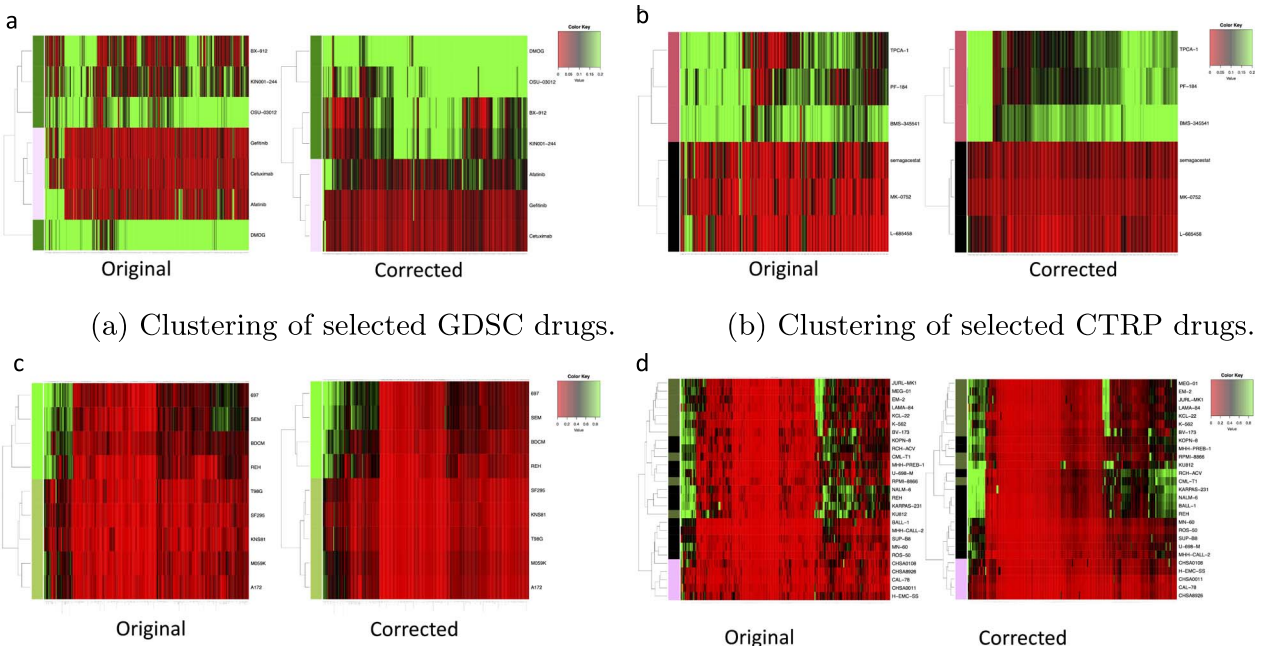
The analysis shows an enhancement in clustering accuracy when the RT-DMF-corrected drug sensitivity data are applied, compared with the original data. This implies that data correction via the RT-DMF method holds significant potential for improving downstream tasks, thereby enriching the impact and outcomes.

## Conclusions

In this work, we develop a deep learning-based framework called RTDMF. This framework is able to retain useful information and filter out noise by exploiting the underlying structure of each dataset. RTDMF is stable, scalable, and easy to apply compared with traditional MF methods. Additionally, its theory is relatively well-developed compared with other deep learning methods. We have shown that RTDMF works well under various assumptions on designated-to-recover datasets. It can improve drug-wise correlation across different large-scale pharmacogenomics studies and support biomarker discovery. Furthermore, RTDMF can serve as a valid imputation method, outcompeting many *ad hoc* imputation methods such as kNN when the downstream analysis requires a dataset without missing values.

RTDMF can be potentially applied to a variety of datasets beyond drug–dose response datasets, as long as those datasets exhibit a similar noise structure. It is particularly useful in situations where signal is of heterogeneous structure and the signal-to-noise level can vary locally. Such datasets are commonly seen in the biomedical field, such as those available in DepMap (e.g. protein expression, genetic dependency). However, we acknowledge that our method of dealing with local intricate values/structures is a first attempt with no domain-specific enrichment. It may be possible to provide more refined thresholding methods with domain-specific knowledge.

Another potential following work could be examining each individual W's hidden semantics in the DMF step. We hypothesize




(a) Clustering of selected GDSC drugs.

(b) Clustering of selected CTRP drugs.

(c) Clustering of selected CTRP cell lines.

(d) Clustering of selected GDSC cell lines.

Figure 6. Heatmaps colored by the drug sensitivity values show the clustering results. Left: using the original data. Right: using the RT-DMF-corrected data.

that each W could potentially represent a certain local pattern contributing to the overall varying signal. Therefore, by studying the connection of individual W's and certain biological facts, we may discover potentially interesting signals that can enhance our understanding of drug efficacy.

**Key Points**

- We introduced a novel computational method, Residual Thresholded Deep Matrix Factorization (RT-DMF), for denoising and imputing drug sensitivity data matrices by leveraging their underlying structure.
- RT-DMF outperforms other traditional matrix factorization methods in recovering meaningful signals and patterns in synthetic datasets designed to mimic the complexity of real drug sensitivity data.
- When applied to major pharmacogenomics datasets, RT-DMF enhances consistency across studies, enables discovery of therapeutically relevant signals, and improves the performance of downstream analyses such as drug sensitivity prediction and drug/cell line clustering.
- RT-DMF has the potential to become a standard processing pipeline for drug sensitivity data, facilitating biomarker discovery and precision medicine applications.

Conflict of interest: None declared.

## Funding

This work is supported by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number R01GM134307. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

# Appendix

## Implementation details

We discuss about some implementation details of RT-DMF and some other methods we used to benchmark, specifically, we discuss the meaning of parameters and why a good $\mathcal{D}_g^1$ will lead to a good $\mathcal{D}_g^2$. In the sections we also discuss the data synthesis details.

One might notice that the neural network in RT-DMF is a wide one. In fact, the width does not violate the low-rank assumption. This is due to an implicit regularization during the optimization of this model. What is more, such a linear DMF is scalable and surprisingly stable to initialization as long as the stopping criterion is consistent.

In fact, the tuning parameters in all three methods can be seen as a trade-off between the fitting of an appropriate low-rank $\mathcal{D}_g^1$: we want $\mathcal{D}_g^1$ to be informatically similar to the original matrix in terms of intrinsic properties of drugs and CCLs and their interactions. A poor fitting of $\mathcal{D}_g^1$ will lead to the hardness in determining appropriate $\mathcal{D}_g^2$, hence losing the potential therapeutically information that biologists are seeking for. As what we have observed in the 3, other matrix completion methods are not successful in terms of recovering $\mathcal{D}_g^1$ and $\mathcal{D}_g^2$.

The early stopping step helps the determination of the $\mathcal{D}_g^1$ and $\mathcal{D}_g^2$ under the consideration/assumption that general interactions between drugs and cancer cell lines or systematic structure/-pattern, which are expected to present in $\mathcal{D}_g^1$, are less likely interesting therapeutically. In contrast, $\mathcal{D}_g^2$ consists of information reflecting more specific interactions between drugs and cancer cell lines that are more likely to be therapeutically relevant.

The thresholding step helps better separate $\mathcal{D}_g^2$ and $\mathbf{\Sigma}$ using magnitude differences and enrichment in preknown drug and CCL groups.

After recovering, we do not know which is the signal and which is the noise, i.e. it is hard to discriminate between $D_g^2$ and $\mathbf{\Sigma}$. However, though $\mathcal{D}_g^2$ would be more of interest, a poor recovery of $\mathcal{D}_g^1$ will lead to a poor discovery of $D_g^2$. As specifically, we only consider the drug–cell interaction data and there is no other information

when training the model. Hence our $\mathcal{D}_g^2$ can be only recovered by the residual information between the original dataset and $\mathcal{D}_g^1$. The low-ranked structure corresponding to larger (more general effect) will be less likely to be interesting and the smaller (more specific group design) will be more likely to be interesting.

### *Generation details of $\mathcal{D}_g^1$*

We generate $\mathcal{D}_g^1$ in two different ways:

- As a product of two nonnegative low-rank matrices: we first generate a matrix $L$ where each component $L_{ij}$ is uniformly sampled from $[0, 1]$, next we perform nonnegative matrix factorization on $L$ as in [31], i.e. $L \approx U_{\text{NMF}} V_{\text{NMF}}^{\text{T}}$, where $U_{\text{NMF}}$ and $V_{\text{NMF}}$ are low-rank matrix with rank 10. Then we let $\mathcal{D}_g^1 = U_{\text{NMF}} V_{\text{NMF}}^{\text{T}}$.
- As a product of multiple nonnegative low rank matrices: we first generate a matrix $L$ where each component $L_{ij}$ is uniformly sampled from $[0, 1]$, next we perform deep matrix factorization on $L$ with four matrix components, i.e. $L \approx W_1 W_2 W_3 W_4$, where $W_1, W_2, W_3, W_4$ are low-rank matrices with rank found by DMF. Then we let $\mathcal{D}_g^1 = W_1 W_2 W_3 W_4$.

We used the $\mathcal{D}_g^1$ generated by the first method in the simulation experiments, as we do not observe much differences between the two.

### *Generation details of $\mathcal{D}_g^2$ under simple two block assumptions*

First we generate an all-zero matrix. Then, for entries $\mathcal{D}_g^2[i, j]$ such that $i, j \in [\{r_1, \ldots, r_{\text{bsize}}\} \times \{c_1, c_2, \ldots, c_{\text{bsize}}\}] \cup [\{r_{\text{bsize}+1}, \ldots, r_{2\cdot\text{bsize}}\} \times \{c_{\text{bsize}+1}, \ldots, c_{2\cdot\text{bsize}}\}]$ (in block, and bsize denotes the block size), we have $\mathcal{D}_l[i, j]$ have a probability of $p_{\text{inside}}$ being magnitude of $s \in \text{Unif}(0.2, 0.3)$. We pick such an interval because it is slightly larger than 2 standard deviations above average for $\mathcal{D}_g^1$. Out-block entries have a probability of $p_{\text{outside}} = 0.05$ being magnitude of $s = 0.3$ to replace the 0 value. We set $p_{\text{inside}} = 0.6, p_{\text{outside}} = 0.05$ *and* $s = 0.3$ across all block sizes.

### *Generation details of $\mathcal{D}_g^2$ under mixed overlapping block assumptions*

We generate an all-zero matrix first again. Then, for entries $\mathcal{D}_g^2[i, j]$ such that $i, j \in \{[\{r_1, \ldots, r_{20}\} \times \{c_1, c_2, \ldots, c_{10}\}] \cup [\{r_5, r_6, \ldots, r_{20}\} \times \{c_{25}, \ldots, c_{35}\}] \cup [\{r_{30}, \ldots, r_{35}\} \times \{c_0, c_1, \ldots, c_{10}\}] \cup [\{r_0, \ldots, r_{40}\} \times c_{40}\}]$ (in block), we have $\mathcal{D}_g^2[i, j]$ have a probability of $p_{\text{inside}}$ being magnitude of $s$, while out-block entries have a probability of $p_{\text{outside}}$ being magnitude of $s$ to replace the original 0. We set $p_{\text{inside}} = 0.6, p_{\text{outside}} = 0.05, s = 0.3$.

### *Generation details of $\mathcal{D}_g^2$ under real data assumptions*

This model directly takes the residual from the FIMM data in [15] and their naive low-rank representation. In particular, we find a naive low-rank representation of FIMM data and fetch the residual through subtracting the real data from the low-rank data entry-wise. We then do a simple bi-clustering to the residual through spectral methods. We select the found bi-clusters that are most significant in terms of in-group size and take them as true signals. The rest of the entries have a certain probability $P$ to be kept. All other entries are set to 0 so that we assume they are irrelevant noises. This is a simplified version of real data noise as we assume the residual that matters only lies in the groups we have found. A visualization of $\mathcal{D}_g^2$ under this situation when $P = 0.1$ is demonstrated in Fig. 3(d).

### *Implementation details of baseline methods*

For basic matrix factorization, we used the formulation as in [32]. We use a slightly relaxed form of $\mathbf{min}_L \|L\|^*$ subject to $\|L - M\| = \epsilon$—this will not require us to prespecify the ranks of $L$, and instead a low-rank solution will be pursued automatically when minimizing the nuclear norm. For RPCA, we used the formulation as in [25], namely $\mathbf{min}_{L,S} \|L\|^* + \tau \|S\|_1$ subject to $L + S = M$. In this formulation, minimizing the $\ell_1$ norm will enforce sparsity, and minimizing the nuclear norm will favor low-rank solutions and $\tau$ serves as a parameter to balance these two parts. For DBDNMF, we use the formulation $\mathbf{min}_{\theta,L,R} \|M - \alpha f_\theta(L) - LR\|$, where $f_\theta$ is a nonlinear neural network parametrized by $\theta$. $\alpha$ is used to balance the linear and nonlinear effects. For more implementation details regarding these baselines, please refer to the GitHub repository's Baseline directory.

### *Simulation results for matrix factorization*

The result of matrix factorization is demonstrated in A1. For the simple two-block case, the ground truth is demonstrated in A1(b) and matrix factorization's result is demonstrated in A1(a). Similarly, the ground truth of the mixed block case is demonstrated in A1(d) and matrix factorization's result is demonstrated in A1(c). The ground truth of the mixed block case is demonstrated in A1(f) and matrix factorization's result is demonstrated in A1(e). As we can see, vanilla matrix factorization performs lackluster in all three case, yielding relatively high RE and IBRE.

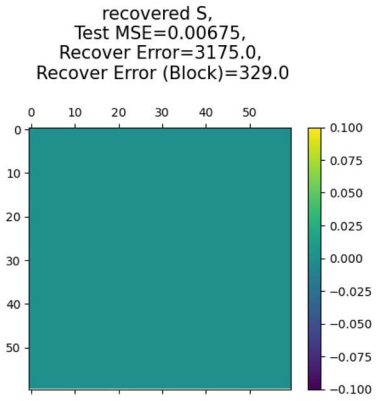
## Stability studies

Deep learning methods are often criticized for their instability [33], as they typically involve hundreds of iterations of random samples for stochastic gradient descent following a randomized initialization of weights. However, RT-DMF does not suffer from random sampling as we utilize the full gradient in each iteration. In this section, we demonstrate empirically that RT-DMF is robust with respect to initialization randomization—it can recover nearly identical matrices given different initializations with an appropriate stopping criterion.

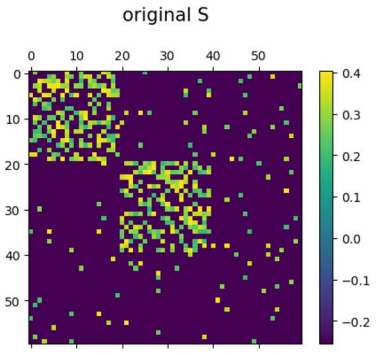
We use the same real data residual model for $\mathcal{D}_g^2$ as in the simulation section but vary the initialization by setting three different seeds. The results are illustrated in Fig. A2. Although different stopping epochs are employed (choosing the epoch just before the validation MSE begins to increase), our recovered matrices show minimal differences in terms of the evaluation metrics, as demonstrated by the figure.

## Effectiveness of RT-DMF imputation in elastic-net model

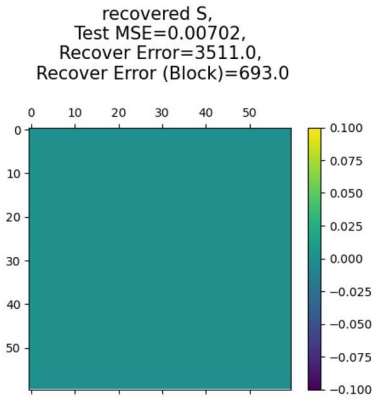
Building upon previous research that identified elastic net as an effective method for drug sensitivity prediction for smaller datasets as in [27], we evaluated RT-DMF's imputation performance against the traditional $k$-nearest-neighbor imputation. Since elastic net requires a complete matrix for predictions, the original implementation relied on KNN-Imputer for handling missing values [27]. Our analysis demonstrates that RT-DMF significantly improves prediction accuracy across both datasets: the root mean squared error decreased from 0.023 to 0.009 in CTRP and from 0.047 to 0.031 in GDSC (Fig. A3). Beyond the reduction in RMSE, RT-DMF also achieved notably lower variance in predictions, indicating more stable and reliable model performance. These improvements suggest that RT-DMF-processed data provide a more robust foundation for downstream analysis tasks.



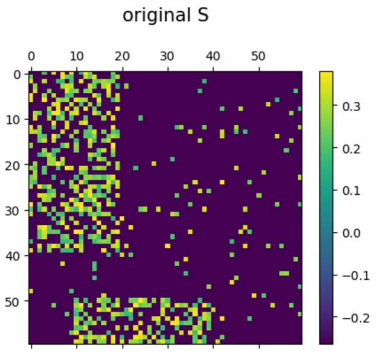

(a) Recovered $S$ with MF for simple two-block $\mathcal{D}_g^2$



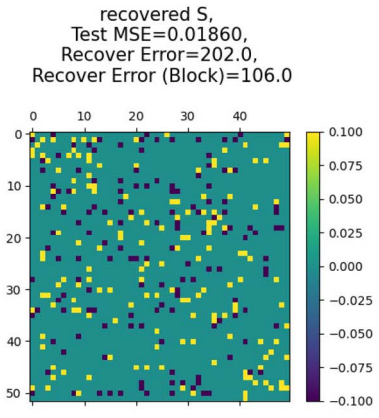

(b) Ground truth $S$ generated according to the simple two-block model



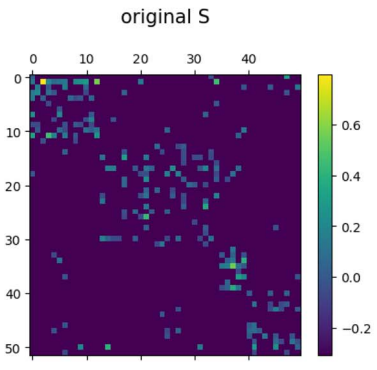

(c) Recovered $S$ with MF for mixed overlapping $\mathcal{D}_g^2$




(d) Ground truth $S$ generated according to the mixed overlapping block model

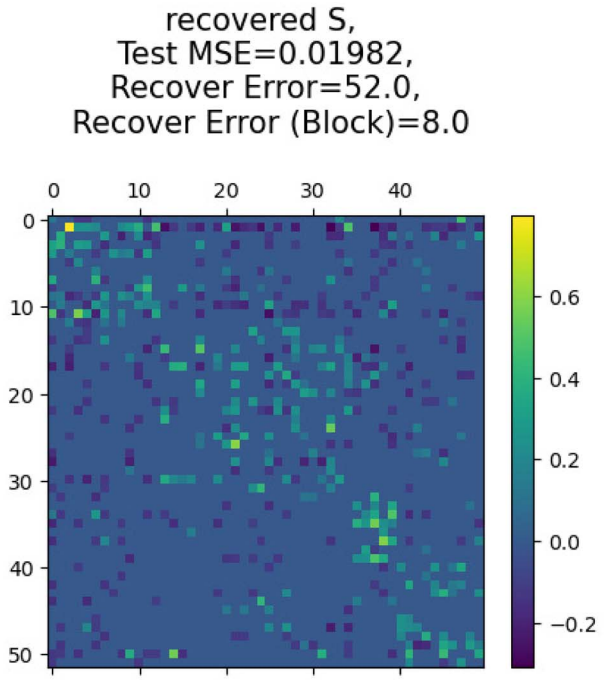



(e) Recovered $S$ for the real data residual $\mathcal{D}_g^2$

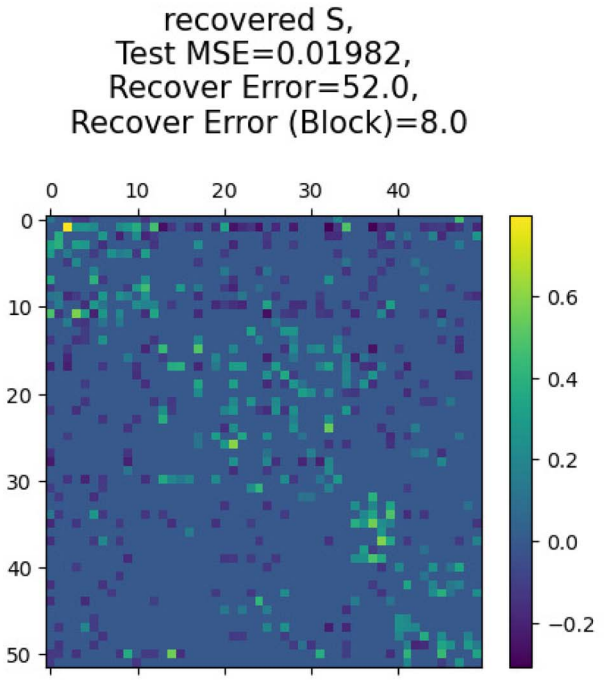



(f) Ground truth $S$ generated according to the real data residual model

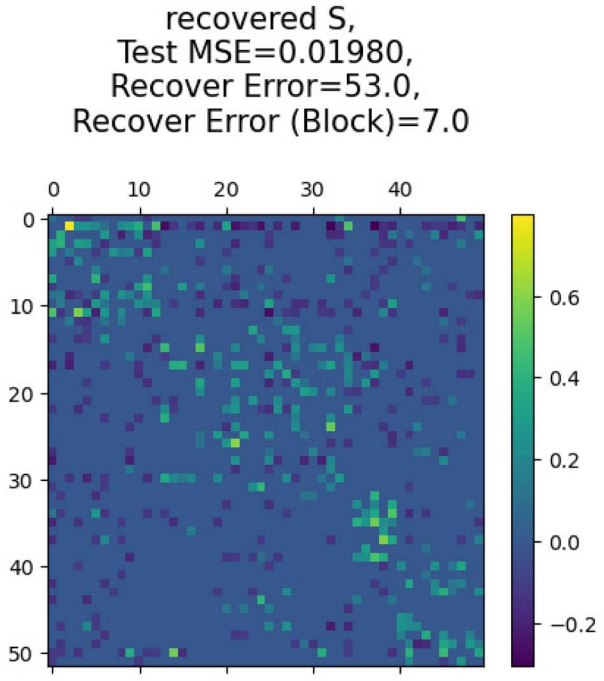
Figure A1. Matrix factorization results for simulation.



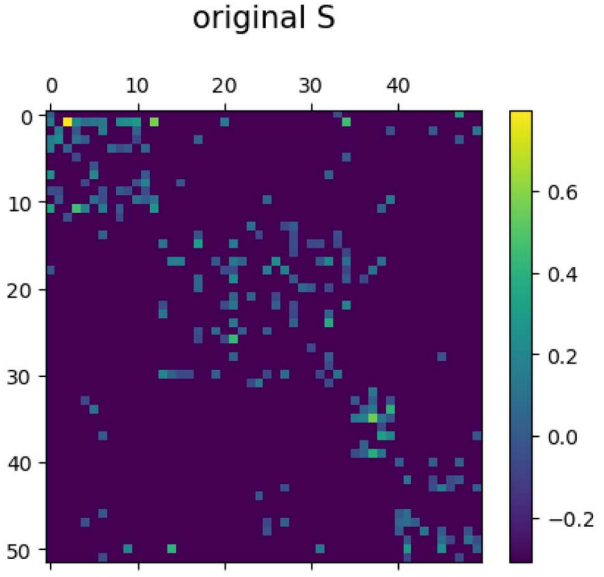

(a) Recovered $S$ with seed 1, at iteration 9000




(b) Recovered $S$ with seed 11, at iteration 11000

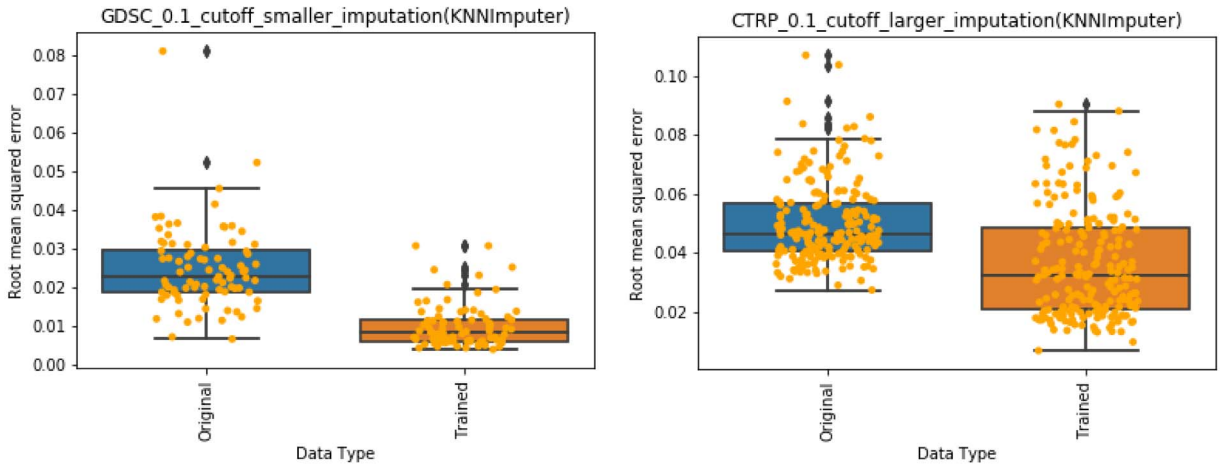



(c) Recovered $S$ with seed 111, at iteration 13000




(d) Ground truth $S$

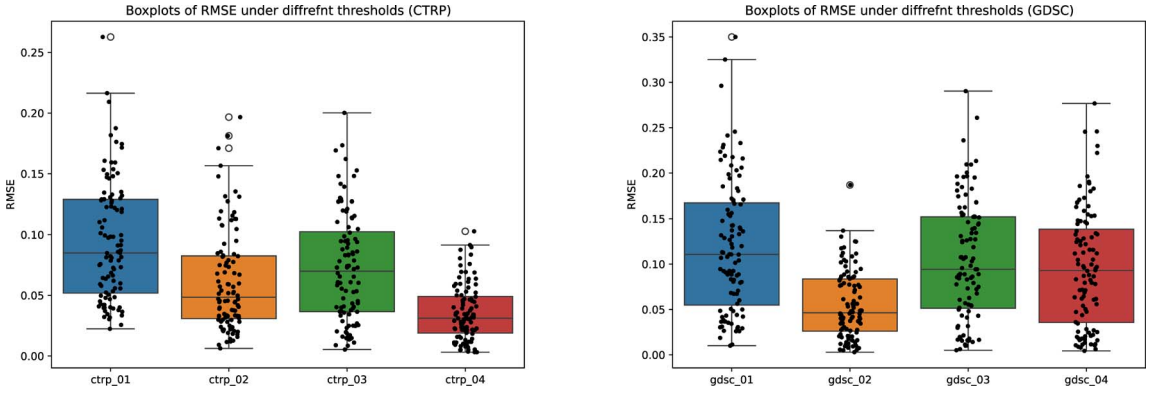
Figure A2. Replicate of recovering S using different initialization with the same stop criterion—the last epoch where validation MSE starts to rise.




(a) Comparison of RMSE between predicted and actual drug sensitivity values in GDSC dataset using RT-DMF versus baseline imputation.

(b) Comparison of RMSE between predicted and actual drug sensitivity values in CTRP dataset using RT-DMF versus baseline imputation.

Figure A3. Performance comparison of RT-DMF versus baseline imputation methods across datasets, demonstrating consistent improvement in prediction accuracy as measured by RMSE.




(a) Impact of varying threshold values on RMSE in CTRP dataset predictions.

(b) Impact of varying threshold values on RMSE in GDSC dataset predictions.

Figure A4. (a) Impact of varying threshold values on RMSE in CTRP dataset predictions. (b) Impact of varying threshold values on RMSE in GDSC dataset predictions.

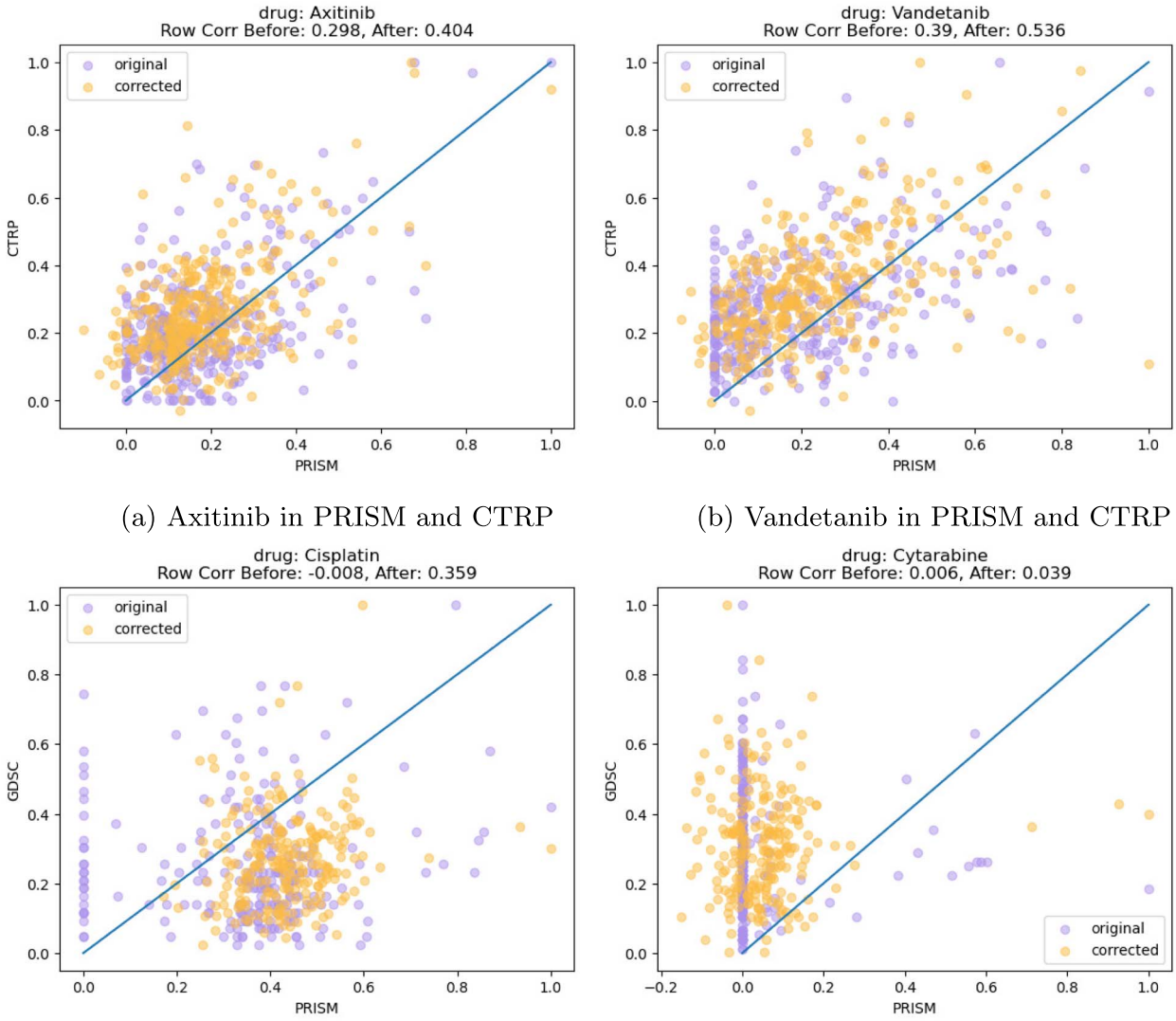
## Threshold evaluation

To systematically evaluate the impact of thresholding on downstream analysis, we extended our investigation using the transfer learning approach described in [28]. We applied this transfer learning methodology across multiple threshold values for RT-DMF applied pm both CTRP and GDSC datasets, following identical procedures for each threshold setting. Our comprehensive analysis revealed that the optimal thresholds identified through validation errors during the RT-DMF training phase corresponded closely with the best-performing thresholds in downstream tasks. For CTRP, the optimal threshold selected by validation error was indeed best-performing as demonstrated in A4(a). Similarly, in GDSC, the 0.2 threshold value identified through validation errors proved optimal for transfer learning outcomes as in A4(b).

These findings validate our threshold ($\eta$) selection approach, confirming that the use of validation errors as a criterion for threshold parameter selection yields a straightforward yet reliable method for optimizing the overall model performance. The alignment between validation-based threshold selection and transfer learning outcomes suggests that this simple approach effectively captures meaningful patterns in drug sensitivity data.

## Case studies of individual drugs

We also present scatter plots for some drugs individually for a better illustration of the performance of RT-DMF in Fig. A5. In this set of illustrations, each data point represents the value of summarized dose–response values in the dataset labeled on axes, hence the blue line $y = x$ represents the perfect situation that




(a) Axitinib in PRISM and CTRP

(b) Vandetanib in PRISM and CTRP

(c) Cisplatin in PRISM and GDSC

(d) Cytarabine in PRISM and GDSC

Figure A5. Individual drug scatter plots across different datasets. Each dot represents the value of one cell-line corresponding to such drug in the x-axis dataset and y-axis dataset.

the values of such drug cell-line responses are exactly the same between two datasets. The different colors represent the value in original and RT-DMF-corrected datasets. We may observe that the reason why correlation increases is because the rightmost points are well corrected by the method, as it is much closer to the $y = x$ line. From Fig. A5(a), A5(b), and A5(c), we observe that RT-DMF is able to alleviate the datapoints that form a "stripe" in the original scatterplot. Figure A5(d) shows an exceptional case, where most of the data points in PRISM have a number around 0 but they vary a lot in the GDSC dataset, likely due to an erroneous input in PRISM. However, we still see RTDMF attempts to adjust some values that are low in GDSC but high in PRISM.

## References

1. Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med* 2015;**372**:793–5. https://doi.org/10.1056/NEJMp1500523
2. Lowy DR, Collins FS. Aiming high—changing the trajectory for cancer. *N Engl J Med* 2016;**374**:1901–4. https://doi.org/10.1056/NEJMp1600894
3. Chen B, Butte AJ. Leveraging big data to transform target selection and drug discovery. *Clin Pharmacol Ther* 2016;**99**:285–97. https://doi.org/10.1002/cpt.318
4. Yothers G, O'Connell MJ, Lee M. *et al.* Validation of the 12-gene colon cancer recurrence score in NSABP C-07 as a predictor of recurrence in patients with stage II and III colon cancer treated with fluorouracil and leucovorin (FU/LV) and FU/LV plus oxaliplatin. *J Clin Oncol* 2013;**31**:4512–9. https://doi.org/10.1200/JCO.2012.47.3116
5. de Gramont A, Watson S, Ellis LM. *et al.* Pragmatic issues in biomarker evaluation for targeted therapies in cancer. *Nat Rev Clin Oncol* 2015;**12**:197–212. https://doi.org/10.1038/nrclinonc.2014.202
6. Garnett MJ, Edelman EJ, Heidorn SJ. *et al.* Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* 2012;**483**:570–5. https://doi.org/10.1038/nature11005

7. Barretina J, Caponigro G, Stransky N. *et al.* The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 2012;**483**:603–7. https://doi.org/10.1038/nature11003
8. Basu A, Bodycombe NE, Cheah JH. *et al.* An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell* 2013;**154**:1151–61. https://doi.org/10.1016/j.cell.2013.08.003
9. Iorio F, Knijnenburg TA, Vis DJ. *et al.* A landscape of pharmacogenomic interactions in cancer. *Cell* 2016;**166**:740–54. https://doi.org/10.1016/j.cell.2016.06.017
10. Niepel M, Hafner M, Pace EA. *et al.* Profiles of basal and stimulated receptor signaling networks predict drug response in breast cancer lines. *Sci Signal* 2013;**6**:ra84. https://doi.org/10.1126/scisignal.2004379
11. Rees MG, Seashore-Ludlow B, Cheah JH. *et al.* Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat Chem Biol* 2016;**12**:109–16. https://doi.org/10.1038/nchembio.1986
12. Seashore-Ludlow B, Rees MG, Cheah JH. *et al.* Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer Discov* 2015;**5**:1210–23. https://doi.org/10.1158/2159-8290.CD-15-0235
13. Haibe-Kains B, El-Hachem N, Birkbak NJ. *et al.* Inconsistency in large pharmacogenomic studies. *Nature* 2013;**504**:389–93. https://doi.org/10.1038/nature12831
14. Safikhani Z, Smirnov P, Freeman M. *et al.* Revisiting inconsistency in large pharmacogenomic studies. *F1000Res* 2016;**5**:2333.
15. Mpindi JP, Yadav B, Östling P. *et al.* Consistency in drug response profiling. *Nature* 2016;**540**:E5 EP.
16. Bouhaddou M, DiStefano MS, Riesel EA. *et al.* Drug response consistency in CCLE and CGP. *Nature* 2016;**540**:E9 EP.
17. Hafner M, Niepel M, Chung M. *et al.* Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. *Nat Methods* 2016;**13**:521–7. https://doi.org/10.1038/nmeth.3853
18. Hu ZT, Ye Y, Newbury Patrick A. *et al.* AICM: a genuine framework for correcting inconsistency between large pharmacogenomics. *Datasets* 2019;248–59.
19. Jia P, Hu R, Pei G. *et al.* Deep generative neural network for accurate drug response imputation. *Nat Commun* 2021;**12**:1740. https://doi.org/10.1038/s41467-021-21997-5
20. Peng W, Chen T, Dai W. Predicting drug response based on multi-omics fusion and graph convolution. *IEEE J Biomed Health Inform* 2022;**26**:1384–93. https://doi.org/10.1109/JBHI.2021.3102186
21. Liu H, Wang F, Yu J. *et al.* DBDNMF: a dual branch deep neural matrix factorization method for drug response prediction. *PLoS Comput Biol* 2024;**20**:1–22. https://doi.org/10.1371/journal.pcbi.1012012
22. Neyshabur B. Implicit regularization in deep learning. *CoRR* 2017. abs/1709.01953.
23. Arora S, Cohen N, Hu W. *et al. Implicit Regularization in Deep Matrix Factorization*. Red Hook, NY, USA: Curran Associates Inc, 2019.
24. Candès EJ, Recht B. Exact matrix completion via convex optimization. *Found Comput Math* 2009;**9**:717–72. https://doi.org/10.1007/s10208-009-9045-5
25. Candès EJ, Li X, Ma Y. *et al.* Robust principal component analysis? *J ACM* 2011;**58**:1–37. https://doi.org/10.1145/1970392.1970395
26. Zar JH. Significance testing of the spearman rank correlation coefficient. *J Am Stat Assoc* 1972;**67**:578–80. https://doi.org/10.1080/01621459.1972.10481251
27. Yeh S-J, Chen R, Xing J. *et al.* TransCell: in silico characterization of genomic landscape and cellular responses from gene expressions through a two-step transfer learning. *Cancer Res* 2022;**82**:1927–7. https://doi.org/10.1158/1538-7445.AM2022-1927
28. Yeh S-J, Paithankar S, Chen R. *et al.* TransCell: in silico characterization of genomic landscape and cellular responses by deep transfer learning. *Genomics Proteomics Bioinformatics* 2024;**22**:qzad008. https://doi.org/10.1093/gpbjnl/qzad008
29. Wang S, Huang E, Cairns J. *et al.* Identification of pathways associated with chemosensitivity through network embedding. *PLoS Comput Biol* **15**.
30. Fahad A, Alshatri N, Tari Z. *et al.* A survey of clustering algorithms for big data: taxonomy and empirical analysis. *IEEE Trans Emerg Top Comput* 2014;**2**:267–79. https://doi.org/10.1109/TETC.2014.2330519
31. Pedregosa F, Varoquaux G, Gramfort A. *et al.* Scikit-learn: machine learning in python. *J Mach Learn Res* 2011;**12**:2825–30.
32. Diamond S, Boyd S. CVXPY: a python-embedded modeling language for convex optimization. *J Mach Learn Res* 2016;**17**:1–5.
33. Yu B, Kumbier K. Veridical data science. *Proc Natl Acad Sci* 2020;**117**:3920–9. https://doi.org/10.1073/pnas.1901326117



*Briefings in Bioinformatics*, 2025, **26(3)**, bbaf226
https://doi.org/10.1093/bib/bbaf226
Problem Solving Protocol