BMC Bioinformatics

RESEARCH

Open Access

On the correspondence between the transcriptomic response of a compound and its effects on its targets

Chloe Engler Hart[1], Daniel Ence[1], David Healey[1] and Daniel Domingo-Fernández[1*]

*Correspondence: daniel.domingo-fernandez@envedabio.com

[1] Enveda Biosciences, Boulder, CO, USA

Abstract

Better understanding the transcriptomic response produced by a compound perturbing its targets can shed light on the underlying biological processes regulated by the compound. However, establishing the relationship between the induced transcriptomic response and the target of a compound is non-trivial, partly because targets are rarely differentially expressed. Therefore, connecting both modalities requires orthogonal information (e.g., pathway or functional information). Here, we present a comprehensive study aimed at exploring this relationship by leveraging thousands of transcriptomic experiments and target data for over 2000 compounds. Firstly, we confirm that compound-target information does not correlate as expected with the transcriptomic signatures induced by a compound. However, we reveal how the concordance between both modalities increases by connecting pathway and target information. Additionally, we investigate whether compounds that target the same proteins induce a similar transcriptomic response and conversely, whether compounds with similar transcriptomic responses share the same target proteins. While our findings suggest that this is generally not the case, we did observe that compounds with similar transcriptomic profiles are more likely to share at least one protein target and common therapeutic applications. Finally, we demonstrate how to exploit the relationship between both modalities for mechanism of action deconvolution by presenting a case scenario involving a few compound pairs with high similarity.

Keywords: Transcriptomics, Drug discovery, Compound target identification, Mechanism of action (MoA) deconvolution, Drug target

Introduction

Transcriptomic data informs change in transcriptional activity through differential mRNA abundance, providing a 'snapshot' of cellular signaling. In the last decade, numerous approaches leveraged this information in order to identify candidate drugs for a given indication [12, 21] and for drug repurposing applications [9]. Parallely, the large amount of bioactivity data produced by novel high-throughput techniques can reveal whether a chemical compound targets a particular protein. Both modalities are

complementary; the binding of the compound to its target(s) leads to transcriptomic changes in genes regulated or modulated by the target [23].

In recent years, several studies have examined whether transcriptomic signatures can be used to predict the target of a compound. The first study by Isik et al. [11], analyzed over 500 compounds from the Connectivity Map [14] and determined that 97% of them did not exhibit any expression changes on their targets in their corresponding drug perturbation experiments. However, by calculating the shortest path between each dysregulated gene and target in a protein–protein interactome, they found that dysregulated genes were significantly closer to the compound's target than by chance. A more recent study by Pabon et al. [19] analyzed the correlation between profiles from drug perturbation and gene knockdown experiments with the L1000 platform. In their work, they evaluated whether looking at the top correlated drug-gene knockdown profiles could identify pairs corresponding to a compound and its target. To do so, they employed 29 compounds for which their target was known, and identified eight true positives among the top 100 predicted drug-gene knockdown pairs. Additionally, they found that the number of true positives slightly increased to 10/100 by investigating correlations of the target with its interacting proteins. Similar to the results of Isik et al. this larger scale study found that some of the compounds exhibited a low correlation with the knockdown profile of its target.

While these studies have shown that transcriptomic data alone is insufficient to predict the target of a given chemical, they also revealed that leveraging prior knowledge represented as protein–protein interactions can facilitate understanding the relation between the transcriptomic signatures of a compound and its known targets. Furthermore, apart from target prediction, better understanding the relationship between these two modalities can help us understand the Mechanism of Action (MoA) of drugs, since compounds that induce a similar transcriptomic signature might share the same target(s) [20]. Lastly, it is currently unclear to what degree structurally similar compounds that target the same proteins also induce similar gene expression profiles.

Recently, a new database called ChemPert integrated protein target information and thousands of transcriptomic experiments from over one hundred non-cancer cell types [28]. Leveraging this resource, we systematically investigated the correspondence between transcriptomic and target data with a large-scale dataset containing over 2000 compounds. To do so, we represented transcriptomic and target data using several approaches and subsequently evaluated their similarity using various correlation metrics. Our comprehensive evaluation considers compound concentration, multifactorial genes, and cell lines and leverages data at different biological scales, from protein targets to downstream pathways. In line with previous work, we found that targets are rarely differentially expressed in transcriptomic experiments. However, by combining the target information of the compound with pathway data we can increase its correlation with the induced transcriptomic response. Furthermore, we find that compounds targeting the same protein do not necessarily induce a similar transcriptomic response in the same cell line. Inversely, we find that compounds with highly similar transcriptomic profiles are more likely to share at least one protein target and therapeutic indications.

Finally, we present a case scenario where we explore two natural products that exhibited the highest correlation between their transcriptomic and target vectors in order to demonstrate how to exploit this information for MoA deconvolution.

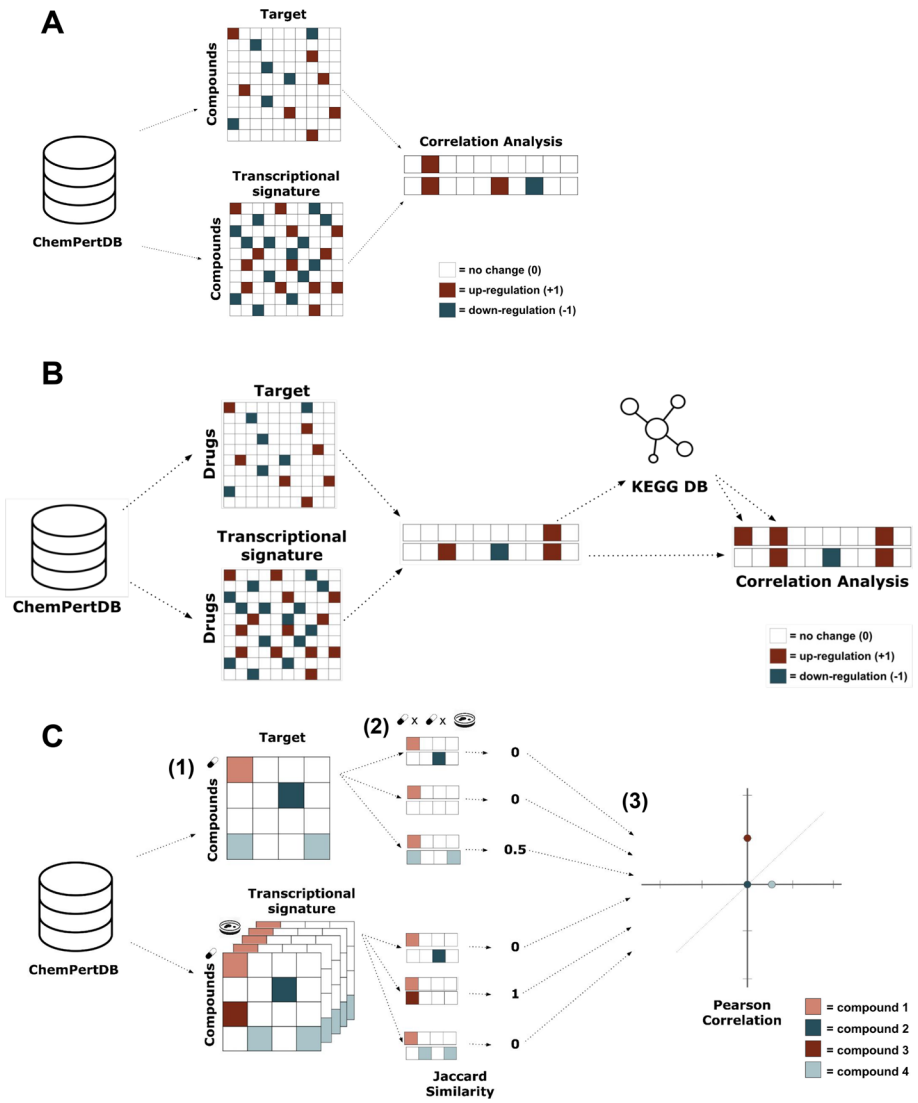

Fig. 1 Schematic illustration of our work. **A** Correlating compound-induced transcriptomic signatures with the known target(s) of the compound. In the figure, + 1 corresponds to up-regulation of a transcript or the activation of a protein by a compound, − 1 corresponds to down-regulation of a transcript or the inhibition of a protein by a compound, and 0 corresponds to no change in the transcript or no binding. **B** Correlating compound-induced transcriptomic signatures with downstream pathway information of the target(s) of the compound. **C** Assessing whether target vector similarity is correlated with transcriptomic response similarity. After the transcriptomic and target data is collected from ChemPert, Jaccard similarity is calculated for every pair of target vectors and every pair of transcriptomic vectors. To discard the inherent variability across cell lines, we filtered out compound pairs that were not tested in the same cell line. (**1**) Next, for each pair of compounds, the Jaccard similarity of their target vectors and the Jaccard similarity for their transcriptomic vectors becomes an x, y coordinate pair (**2**). Once all of the x,y pairs have been found, the Pearson correlation for all of these pairs is calculated (**3**)

Methodology

Figure 1 illustrates the different analyses conducted. All analyses are based on transcriptomic and target data for over two thousand compounds retrieved from ChemPert [28] (see "Collecting transcriptomic and target information" Section). These two modalities are represented as vectors as we intend to evaluate their correspondence for each compound. In "Representing transcriptomic and target information" Section, we describe the three approaches to represent each compound. The correlation metrics used to evaluate the agreement between both modalities are outlined in "Correlation and Similarity Metrics" Section. The following "Correlation Analyses" Section describes the correlation analyses outlined in Fig. 1. Finally, "Implementations details and code availability" Section outlines the implementation details.

Collecting transcriptomic and target information

We leveraged data from ChemPert [28], a database containing transcriptional data from thousands of experiments where cell lines and tissues were perturbed by 2508 unique perturbagens (details in Additional file 1: Text S1). Furthermore, this database contains

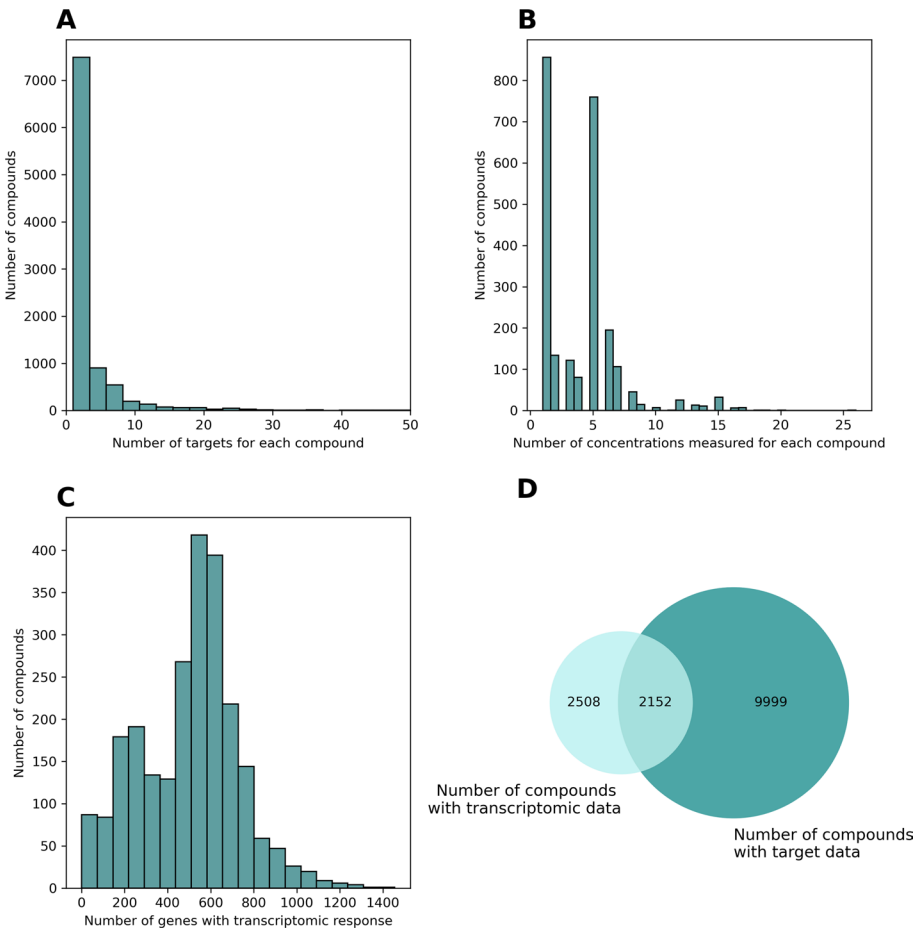

Fig. 2 **a** Distribution of targets for each compound. **b** Distribution of measured concentrations for each compound. **c** Distribution of DEG (up/down-regulated genes) per compound. **d** Number of compounds with target and transcriptomic data in ChemPert and their intersection

target information from Drug Repurposing Hub [1], DrugBank [27], and STITCH [22]. The 82,270 transcriptional signatures at different concentrations of the 2508 compounds in ChemPert were filtered to a subset of 2152 chemical compounds for which both transcriptomic and target data were available on 2022-05-04 (Fig. 2D). This subset is majorly composed of small molecules and a few peptides. Distributions of the different concentrations measured for each gene, number of Differentially Expressed Genes (DEG), and number of targets for each chemical are shown in Fig. 2.

Representing transcriptomic and target information

After obtaining the subset of chemicals from ChemPert containing both transcriptomic and target data (i.e., effects of chemicals on protein targets, either activation or inhibition), we represented their transcriptomic signature and their known targets as two vectors of equal length with the ultimate goal of evaluating the correlation between both. In the following subsections, we outline the two different vector representations proposed in this work.

Original ChemPert data

Since the ChemPert data is already preprocessed and normalized, the most straightforward representation consists of directly leveraging the original vectors available in ChemPert. Therefore, for each of the 2152 compounds, we represent both its transcriptomic and target information as a vector of length X, where X corresponds to the number of genes measured (4938) (Fig. 1). In both vectors, there are three possible values for each gene provided by ChemPert's data:

- $+1$. Representing an up-regulation of the protein transcript in the case of the transcriptomic vector or activation of the protein after the chemical binds to it in the case of the chemical-target vector.
- -1. Representing a down-regulation of the protein transcript in the case of the transcriptomic vector or inhibition of the protein after the chemical binds to it in the case of the chemical-target vector.
- *0.* Corresponding to no change in the protein transcript or no binding

Additionally, since for a minority of the chemicals there are multiple transcriptomic experiments using different doses or concentrations (Fig. 2B), we represented the transcriptomic signature vector for these chemicals as the union of all differentially expressed genes. In other words, if a chemical increases the abundance of gene transcript X with a particular concentration and the same chemical decreases the abundance of gene transcript Z with a different concentration, the vector for chemical A will contain $+1$ for X and -1 for Z while the rest of the gene transcripts will have 0 as their value. For a small number of cases, genes exhibited discrepancies, i.e., upregulated in some transcriptomic experiments and downregulated in others for a given chemical. To resolve these discrepancies, we first counted the number of experiments where it was upregulated and the number of experiments where it was downregulated. If it was upregulated more than two times the number of times that it was downregulated, we set the value to $+1$. In a similar way, if the gene was downregulated more than two times the

number of times that it was upregulated, we set the value to −1. Otherwise, the vector was set to 0. While using the union allowed us to incorporate all known DEGs for a particular compound, we have additionally analyzed transcriptomic data for individual dose concentrations (e.g., dose concentration inducing the highest number of DEGs), without seeing any significant differences compared to the previously described approach (see "Using transcriptomic responses with highest number of DEGs yielded similar results" Section).

Enriching ChemPert target data with pathway information

Since compounds in the ChemPert database have only a small number of target proteins (Fig. 2A), the target vectors are considerably more sparse than the transcriptomic ones, meaning that it is unlikely that the two vectors will have a high degree of similarity. In order to make the target vectors less sparse, we enriched them using pathway information from three databases: (i) KEGG, a database that contains both topological information and gene sets for over 300 pathways [13], (ii) Reactome [6], another database containing over 2000 pathways, and (iii) WikiPathways [15]. We denote these enriched vectors as pathway vectors through the paper. Transcriptional changes can typically be seen downstream of the target, so it is reasonable to use information about the neighbors of the target protein in the target vectors [11].

The process for generating the pathway vectors went as follows. For a chemical A, we obtained a target gene, which we will call gene B, by finding a gene that corresponded to a non-zero value in the target vector for chemical A. We then found all the gene sets (pathways) that contained gene B. If any of the other 4938 genes were in any of those pathways, we changed the corresponding value of those genes in the resulting pathway vector for chemical A to match the value of gene B in the original target vector (Fig. 1B). We repeated this process until we had done this for every target gene of every chemical (pseudocode in Additional file 2: Fig. S1). Notably, we filtered out pathways with more than 300 genes (e.g., metabolic pathway) and pathways smaller than 15 genes, following pathway enrichment guidelines [17].

Enriching ChemPert target data with topological information

In addition to the three pathway databases (gene sets) used above, we modified the target vectors to account for the topology of a protein–protein interaction network (PPI) (pseudocode in Additional file 2: Fig. S2). To do so, we first used the KEGG database as a PPI to be able to infer if a protein is activated or inhibited by its neighbor using the polarity of each interaction. Leveraging this PPI, we were able to generate PPI vectors by modifying the values of the neighbors of each original target as follows:

- $+1$ if the value of the target protein in the target vector was $+1$ and the target protein activated this protein or if the value of the target protein in the target vector was -1 and the target protein inhibited this protein.
- -1 if the value of the target protein in the target vector was $+1$ and the target protein inhibited this protein or if the value of the target protein in the target vector is -1 and the target protein activated this protein.

If we apply these changes, we generate the values of the immediate neighbors of the target proteins in the PPI vector. However, it is possible that these neighbor proteins also activate or inhibit other proteins farther along in the network. Thus, we also enriched the values of the neighbors in the PPI vector by treating all of the neighbors as target proteins and repeating the above process (see example in the Additional file 2: Fig. S3). We repeated this process 2–5 times to obtain more densely populated vectors and test whether including downstream information would improve the correlation between the new PPI vectors and transcriptomic vectors.

Correlation and similarity metrics

In order to calculate the agreement between a pair of vectors (x and y) for a certain compound, we used the Pearson correlation coefficient by obtaining the mean values of each vector and applying the following equation:

$$r = \frac{\sum (x_i - \underline{x})\left(y_i - \underline{y}\right)}{\sqrt{\sum (x_i - \underline{x})^2 (y_i - \underline{y})^2}} \tag{1}$$

Equation 1: The Pearson correlation coefficient will be between − 1 and 1. It is a measure of the linear correlation between the target vector and transcriptomic vector.

We also used Jaccard similarity in order to determine the similarity between pairs of vectors.

$$J(x, y) = \frac{|x \cap y|}{|x \cup y|} \tag{2}$$

Equation 2: The Jaccard similarity scores are between 0 and 1 and measure the similarity between two vectors by finding the number of elements that they share.

We chose to include both metrics in our analysis in order to be more transparent about the results. Since the data in our vectors is discrete, Jaccard similarity allowed us to measure how many of the values matched up across the two vectors. Complementary, Pearson correlation allowed us to see if the two vectors were positively or negatively correlated.

Correlation analyses

Correlating the transcriptomic and target vectors for a given compound

We applied the aforementioned correlation metrics ("Correlation and Similarity Metrics" Section) in order to calculate the agreement between a target vector, x, and a transcriptomic vector, y, for a certain compound (Fig. 1A). After calculating the correlation coefficients for each compound, we used the permutation test to determine whether the results were statistically significant. We permuted the pairs of target vectors and transcriptomic vectors and applied the same correlation metrics to these random pairs. We repeated this process 100,000 times and obtained the mean correlation score for each iteration. We then found the p-value by using the distribution of correlation means from the permutation tests We used a significance level of 0.05 to determine whether our results were statistically significant after applying Bonferroni correction.

Investigating pairs of compounds based on their transcriptomic/target vector similarity

Apart from assessing whether the transcriptomic and target vectors of a given compound have any correlation, we sought to evaluate whether there was any correlation between the target vector similarity scores and the transcriptomic vector similarity scores. In order to do this, we first calculated the Jaccard similarity for every pair of transcriptomic vectors and every pair of target vectors. To discard the inherent variability across cell lines, we filtered out compound pairs that were not tested in the same cell line. We then created a vector, which we call X, of all of the target vector correlation scores and another vector, which we call Y, of all the transcriptomic vector correlation scores. We ensured that for each compound pair the target similarity score and transcriptomic similarity score corresponding to that pair were in the same position in their respective vectors. In order to determine whether target vector similarity was correlated with transcriptomic vector similarity, we calculated the Pearson correlation coefficient for X and Y (Fig. 1C (3)). To determine whether the Pearson correlation coefficient was significant, we used the permutation test by permuting the transcriptomic similarity vector, Y, 100,000 times.

In order to better understand whether pairs of compounds with similar transcriptomic profiles also shared targets, we first calculated the Jaccard similarity (Eq. 2) for every pair of transcriptomic vectors (Fig. 1C (2)) and filtered out any pairs with a correlation coefficient less than 0.6. The rationale behind choosing 0.6 as a cutoff was we wanted to include a large sample of compound pairs in our study and there were a limited number of compound pairs (32) with a transcriptomic signature correlation score above 0.7. After selecting these pairs, we calculated the Jaccard similarity of the target vectors for each pair and used those correlation coefficients to determine whether compounds with similar transcriptomic profiles shared targets.

Following, we decided to evaluate whether compounds with the same targets had similar transcriptomic profiles measured in the same cell line. To this end, we calculated the Jaccard similarity scores for all the target vector pairs (Fig. 1C (2)) and filtered out any compound pairs whose correlation score was less than one (i.e., all the target(s) of a compound perfectly match the targets of the other one). We employed such a strict cutoff because we wanted to examine pairs that shared all of their targets and affected those targets in the same way (up-regulation or down-regulation). For the remaining pairs, we calculated the Jaccard similarity scores (Eq. 2) for their transcriptomic vectors and utilized those correlation coefficients to determine whether the compound pairs with shared targets had similar transcriptomic profiles.

Permutation analysis at pathway level

We also used a similar method to determine the statistical significance of the correlation scores between the transcriptomic vectors and the target vectors that were modified using pathway information. To do this, we first removed any pathways from the original pathways that contained more than 300 genes or less than 15 genes. We found the lengths of the remaining pathways in the dataset and filled the pathways up with random genes from the pathways. We did this process 1000 different times. We then calculated the mean Pearson correlation score and Jaccard similarity score

for each set of random pathways and compared these scores to the mean correlation scores for the original pathways.

Permutation analysis at network level

In order to further determine the statistical significance of the correlation scores between the target vectors that were modified with topological information from KEGG and the transcriptomic vectors, we generated permuted networks by using the XSWAP algorithm [7]. We then created a new set of target vectors for each random network using the method from "Enriching ChemPert target data with topological information" Section. After creating these vectors, we calculated the mean Jaccard similarity score and the mean absolute value of the Pearson correlation score between the target vectors and transcriptomic vectors for each network and compared these scores to the scores obtained with the original network. In this case, we used the mean absolute value of the Pearson correlation scores because we were more interested in knowing whether the vectors were correlated than knowing if they were positively or negatively correlated.

Implementations details and code availability

We preprocessed the datasets and generated the vectors using the Pandas Python package [16]. In the enrichment analysis, we modified these vectors by leveraging the KEGG pathways available at PathMe [2] (released date 01-03-2021). To calculate the correlations, we employed the NumPy [8] and SciPy [24] Python packages. Additionally, we plotted the visualizations using Seaborn [26] and Matplotlib [10]. Lastly, both data and source code are available at https://github.com/enveda/transcriptomic-target-correlation.

Results

Targets are generally not differentially expressed in drug perturbation transcriptomic experiments

We first investigated the correlation between the target vectors and transcriptomic vectors retrieved from ChemPert without any modification. The resulting correlation scores were very low with the majority of the Pearson correlation scores between -0.02 and 0.02 (Fig. 3A). This was not a surprising result for two reasons. First, the target vectors are very sparse (Fig. 1A) compared to the transcriptomic vectors (Additional file 2: Fig. S4). Second, most of the known targets are not differentially expressed genes (Additional file 2: Fig. S5), in line with previous work [11]. In an effort to determine whether the correlation scores were statistically significant, we used permutation tests on both the target and transcriptomic vectors and compared the results against the null distribution. This analysis confirmed that the permuted vectors yielded comparable Pearson correlations to the original ones (q-value $= 1.0$) (Fig. 3B). Similarly, the Jaccard similarity scores obtained on the permuted vectors were not significantly different (q-value $= 1.0$).

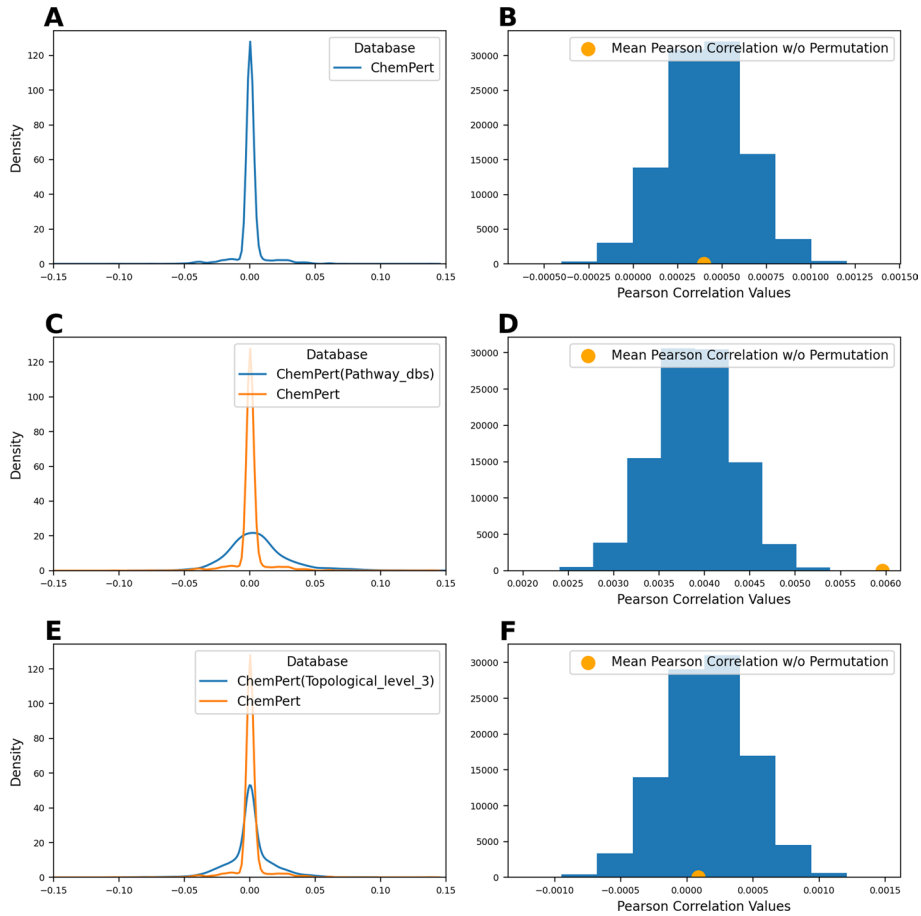

Fig. 3 **a** Distribution of Pearson correlation scores for target and transcriptomic vectors for each of the 2512 compounds. **b** Distribution of the mean Pearson correlation obtained from permutation tests (null distribution) compared against the mean correlation of the original ChemPert data. **c** Pearson correlation scores for pathway and transcriptomic vectors from ChemPert for each of the 2512 compounds. **d** Distribution of the Pearson correlation means for the permutation tests from ChemPert using pathway vectors. **e** Pearson correlation scores for target and transcriptomic vectors from ChemPert using PPI vectors going three levels downstream of the target. **f** Distribution of the Pearson correlation means for the permutation tests from ChemPert using PPI vectors going three levels downstream of the target

Using transcriptomic responses with highest number of DEGs yielded similar results

ChemPert reported several transcriptomic responses for a minority of the compounds in our dataset. In addition to representing the transcriptomic vectors as a union of the differentially expressed genes (see "Original ChemPert data" Section), we also tried using the transcriptomic response with the highest number of DEGs for each compound to create the transcriptional vectors. When we calculated the correlation between the target vectors and these transcriptomic vectors, the majority of the Pearson correlation scores were between − 0.02 and 0.02 (similar to the results in "Targets are generally not differentially expressed in drug perturbation transcriptomic experiments" Section). The Jaccard similarity scores for these vectors were also similar. Since the results for both sets of transcriptomic vectors were not significantly different, we decided to use the transcriptomic vectors that combined the different responses for the rest of the analysis.

Drug perturbation transcriptomic signatures tend to correlate with the downstream pathways targeted

After observing the low correlation coefficients for the ChemPert target vectors and transcriptional vectors, we hypothesized that enriching the target vectors with downstream pathway information might increase our correlation scores due to two main reasons: (i) transcriptomics signatures capture downstream changes at the pathway level [5, 11], and (ii) the pathway vectors are less sparse compared with the target vectors (Additional file 2: Fig. S4B). Below, we discuss the results after conducting two disparate enrichment approaches.

Pathway level analysis

The first and simplest enrichment approach consisted of using pathway vectors corresponding to proteins that participate in the same pathway as the original target(s) (Fig. 1B). Conducting this enrichment increased the average Pearson correlation scores between target and transcriptomic vectors from 0.00039 to 0.00596 compared to the original ChemPert correlation scores (Fig. 3C). While these initial results appeared to be positive, we wanted to make sure that these results were statistically significant and not a result of the pathway vectors being less sparse. Thus, we conducted 100,000 permutations experiments on the transcriptomic vectors and pathway vectors and compared their underlying distribution of Pearson correlation scores to the observed one (q-value of 0.00003 (Fig. 3D)) Similarly by comparing the Jaccard similarity scores for these permutation tests to the null distribution, we got a q-value of 0.09846.

Additionally, we wanted to ensure that the pathways from KEGG, Reactome, and WikiPathways were not only increasing the correlation scores because the enriched target vectors were less sparse than the original target vectors. Thus, we generated 1000 sets of random pathways maintaining their original size and gene occurrence and enriched the target vectors with each set of random pathways. We compared the mean similarity scores and correlation scores for these random sets of pathways to the scores for the original set of pathways. The mean Pearson correlation score and mean Jaccard similarity score for the original pathways were significant (q-value = 0.003). Therefore, we determined that the higher correlation scores resulting from enriching the target vectors with pathway information were not just a result of making the target vectors less sparse.

Network level analysis

The second enrichment approach leverages topological information from KEGG to generate protein−protein interaction (PPI) vectors. Using the PPI vectors increased the Jaccard similarity scores. Furthermore, the similarity scores continued to increase when we repeatedly applied the database up to three levels downstream of the target (Fig. 3E). When we used the permutation test on this data, the results were not statistically significant up to level two. This is not an unexpected result since the vectors were only slightly enriched at these first two levels. However, after the enrichment with topological information going three levels downstream of the target, our results were statistically significant using Jaccard similarity (q-value of 0.00015). We also calculated the Pearson correlation score between the PPI vectors and the transcriptomic vectors. However, the Pearson correlation score did not increase and our results were not statistically

significant when we used Pearson correlation (q-value of 1.0) (Fig. 3F). Thus, indicating the importance of conducting the correlation/similarity analyses with non-sparse vectors.

In order to determine whether the topological information from KEGG was significant, we generated 100 random networks of the same structure. We then enriched the target vectors with the information from each of the random networks ("Permutation analysis at pathway level" Section). We then computed the correlation scores and similarity scores between these sets of enriched target vectors and the transcriptomic vectors. When we compared these scores to the scores of the original network, the original network had a q-value of 0.15 when we used Jaccard similarity and had a q-value of 0.03 when the used the mean absolute value of the Pearson correlation scores. Thus, the topological information from KEGG added significant information to the target vectors.

In summary, our results indicate that enriching target vectors with pathway information increases their correlation with transcriptomic signatures. This is not surprising given that previous work observed that differentially expressed genes tend to be closer to the target than randomly expected [11].

Assessing the impact of pathway size

Since large pathways are more likely to have a non-zero value in the vector, we evaluated whether removing large pathways had an influence on the observed correlations. To do so, we generated different subsets of the pathway dataset where the top X largest pathways were removed (i.e., top 50, top 100, top 250) and subsets of pathways with varied numbers of genes. We used the subsets to enrich the target vectors and calculated the correlation scores between these target vectors and the transcriptomic vectors. The correlation scores were similar for all subsets of the pathways (Additional file 2: Fig. S7). We also applied permutation tests to these vectors and found that all sets had a q-value of 0.003 when we used Pearson correlation. When Jaccard correlation was used, the q-values ranged from 0.057 to 0.27 with the target vectors using the set of pathways with 100–300 genes having the highest q-value.

Target and transcriptomic vector similarity are slightly correlated

Our next question was whether compounds with similar targets had similar transcriptomic responses or vice versa. To discard the inherent variability across cell lines, we filtered out compound pairs that were not tested in the same cell line. We then calculated the Jaccard similarity scores between each target vector pair and each transcriptomic vector pair to identify pairs of compounds targeting the same proteins. In order to determine whether the target similarity scores were correlated to the transcriptomic similarity scores for each compound pair, we created x,y coordinate pairs of the target and transcriptomic similarity scores and computed the Pearson correlation coefficient for all of these pairs (see Methods; Fig. 1C (3)). When we did this with the vectors from the ChemPert database, we obtained a Pearson correlation coefficient of 0.012. We used a permutation test with 1000 permutations and determined that this correlation coefficient was significant with a significance level of 0.05 (q-value of 0.003). We also repeated this process after we had enriched the target vectors with pathway information and we computed a Pearson correlation coefficient of 0.052 which was also statistically

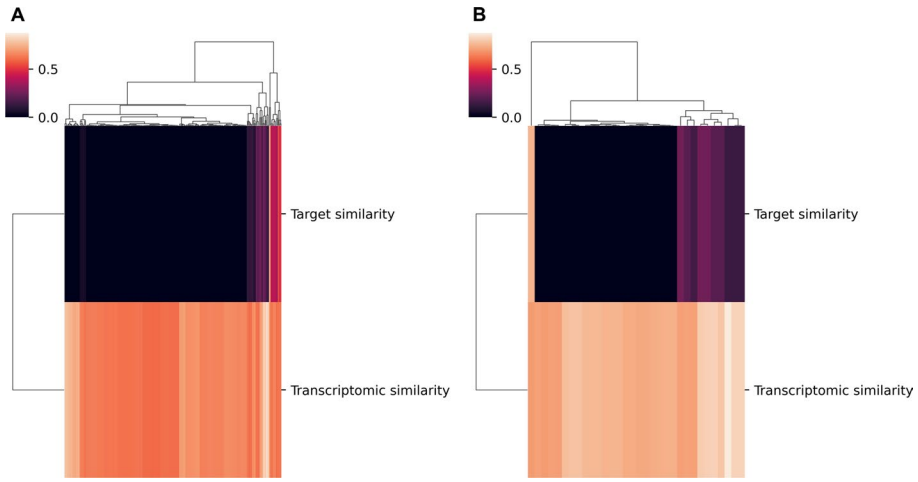

Fig. 4 Target similarity scores for compound pairs with highly similar transcriptomic signatures. **a** This clustered heatmap contains all compound pairs with a higher transcriptomic similarity than 0.6. On the right side of the heatmap, where most of the pairs of compounds with some degree of target overlap are clustered, transcriptomic similarity tends to be also higher (lighter colors). **b** This clustered heatmap containing all compound pairs with transcriptomic similarity higher than 0.7. Compounds with transcriptomic similarity scores in this range are more likely to share targets as seen by the comparative abundance of non-zero (colors other than black) target similarity scores for this graph

significant (q-value of 0.003). To summarize, these results revealed that there is a slight correlation between the target similarity for a compound pair and the transcriptomic similarity for that same compound pair.

Compounds with highly similar transcriptomic profiles are more likely to share targets

We decided to further investigate whether compounds that induced similar transcriptomic profiles target the same proteins by filtering out those pairs with a transcriptomic correlation score less than 0.6 (Fig. 4) (Additional file 1). Among these over 240 pairs, the average target correlation score was 0.044 compared to the correlation scores for all possible compound pairs which was 0.005. Thus, the average target correlation score did increase when we only looked at pairs with high transcriptomic similarity. Interestingly, the percentage of compound pairs that shared at least one target also increased from 2.88 to 18.85%. Additionally, when we increased the cutoff for transcriptomic signature similarity from 0.6 to 0.7, the percentage of pairs with at least one shared target increased to 34.3% (11/32). Furthermore, all of the compound pairs (6) with transcriptomic signature similarity scores over 0.8 share at least one target. For instance, the pair with the highest correlation is Alvocidib and Cgp-60474, both of which are Cyclin Dependent Kinase (CDK) inhibitors. Alvocidib is a flavonoid alkaloid CDK9 kinase inhibitor extracted from two plants and CGP60474 a potent CDK inhibitor. Taken together, our findings suggest that pairs of chemicals which have a high correspondence on the transcriptomic vector tend to share at least one target.

Compounds targeting the same targets typically induce disparate transcriptomic responses

Similarly, we also decided to evaluate whether compounds with the same targets induced similar transcriptomic responses. In order to find the compounds with the

same targets, we calculated the Jaccard similarity for all possible pairs of target vectors and filtered out any compound pairs whose correlation score was less than one. After identifying those compound pairs that share targets (2435) (Additional File 1), we calculated their Jaccard similarity using their transcriptomic vectors. Interestingly, the average transcriptomic correlation score was 0.139 which we compared to the average correlation score of 0.130 for all possible compound pairs. Furthermore, we found that all the transcriptomic correlation scores were lower than 0.6, the threshold previously used in "Target and transcriptomic vector similarity are slightly correlated" Section. This suggests that even when compounds share all of their target(s), they do not necessarily induce a similar transcriptomic response.

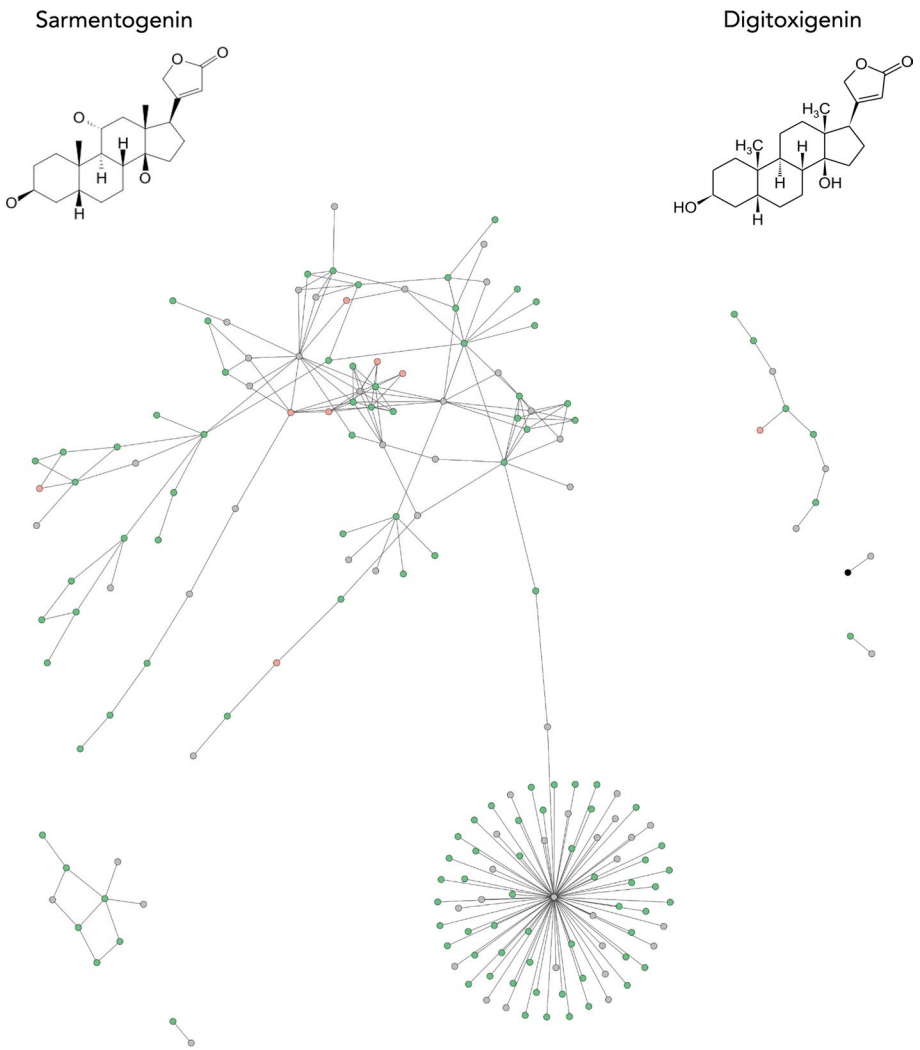

Fig. 5 Differentially expressed genes (DEGs) from the digitoxigenin-sarmentogenin pair overlaid into the PPI network derived from KEGG. Concordant DEGs (e.g., genes down-regulated by both compounds) are marked in green, while DEGs discordant ones (e.g., genes up-regulated by digitoxigenin and down-regulated by sarmentogenin) are marked in red. Finally, the target, ATP1A1, is marked in black and genes that were only measured for one of the two compounds are marked in gray

Despite the low correlation observed between transcriptomic signatures of compounds sharing targets, we investigated the eight pairs of compounds that had a correlation equal to or higher than 0.5. Among these, we found four pairs of natural products with well-known uses for treating arrhythmias among other indications (i.e., digitoxigenin, sarmentogenin, cymarin, proscillaridin). These compounds are structurally related (i.e., Tanimoto coefficient of 0.73 using Morgan fingerprints and share the same Murcko scaffold) and are cardenolides, a family of heart poison compounds from plants. The highest correlated pair is digitoxigenin and sarmentogenin (Fig. 5). Both are used due to their ability to inhibit ATP1A1 [3]. Additionally, we found a pair of DNA topoisomerase I inhibitors (i.e., Genz-644282 and SN-38), a pair of JNK inhibitors (i.e., ZG-10 and JNK-9L), and a pair of dual PYK2/FAK inhibitor (i.e., PF-431396 and PF-562271).

Discussion

In this work, we systematically evaluated correlations between target and transcriptomic data for 2152 compounds. In line with prior work, we found that target and transcriptomic signatures do not correlate, as targets are rarely differentially expressed in transcriptomic experiments. However, once target information is enriched with pathway information, the correlation between both increases. Additionally, we investigated whether compounds that target the same proteins induce a similar transcriptomic response. We found that pairs of compounds sharing the same target show a slightly increased transcriptomic correlation than average. Inversely, we found that compounds with similar transcriptomic profiles are more likely to target at least one shared protein and often treat the same indication. Lastly, we focused on two natural products (i.e., sarmentogenin and digitoxigenin, which are two members of the cardenolide family) that share all their target proteins, and also have the highest transcriptional similarity in order to demonstrate how exploring their high correlations can elucidate their MoA.

Nonetheless, the presented work is not without its limitations. First, our analysis could have been complemented by leveraging the raw transcriptomic profiles, as we have only employed the transcriptomic profiles already processed by ChemPert. This would have allowed us to lower the significance threshold and assess whether the threshold plays a role in the observed correlations between transcriptomic and target vectors. However, it would require significant effort as thousands of experiments would have to be processed. Second, since some compounds have transcriptomic profiles for multiple concentrations, one can take several approaches to correlate the profiles (e.g., use the one with the largest number of DEGs, the one with the highest concentration, etc.). While our main results were generated by taking the union of all differentially expressed genes in these profiles, we also tested alternative approaches using the profile with the largest number of DEGs and observed similar results ("Using transcriptomic responses with highest number of DEGs yielded similar results" Section). Third, it is important to note that the transcriptomic profiles were obtained from different cell lines. While it is well-known that gene expression patterns vary among cell lines [4], Pabon et al. [19] found that the cell line exhibiting the lowest correlation with respect to the control yielded the best results for target prediction. Fourthly, while we could only employ a subset of ChemPert (i.e., over 2000 compounds), this is still a larger number of profiles than previous studies that investigated

transcriptomic correlations [9, 11]. Lastly, the data we leveraged from ChemPert was conducted on the L1000 platform which infers the expression of 11,350 genes from 978 measured genes. This, combined with the fact that only 4938 of these genes were targets for any of the 2152 compounds, restricted the protein space of our exploration since not all genes/proteins were represented in the vectors. However, this set of approximately 5000 genes that was used to conduct this study highly overlap with the genes included in pathway databases; thus, indicating that they correspond to the ones for which we have more functional information about.

In the future, we ambition multiple possible extensions of our work. Firstly, a prospective study could validate our findings by leveraging an additional database. Secondly, additional *omics* modalities beyond transcriptomics such as proteomics and metabolomics could be incorporated in our analysis to explore whether higher correlations are observed. Thirdly, we could assess if the correlations between the transcriptomic and target information tend to be higher when the transcriptomic experiment has been measured in a cell line characteristic to the particular tissue where a drug acts (e.g., neuron or glial cells for drugs treating neurological disorders). If this would be the case, these higher correlations could be used as a proxy to identify candidate repurposing drugs [18, 25]. Alternatively, as demonstrated in our case study, by correlating both transcriptomic and target information we can better understand the MoA of a drug. Additionally, one can use the transcriptomic or target vectors as compound features and conduct a classification task using machine learning models to, for instance, predict the type of compound, their targets, etc.. Lastly, another possible application to the problem of repurposing compounds would be to infer novel activities of combinations of compounds by concatenating their transcriptomic profiles.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-023-05337-6.

Additional file 1. **1) target_high_correlation_cell_lines:** Transcriptomic similarity for chemicals that share the same target. The transcriptomic similarity is calculated on the same cell line. **2) transcriptomic_high_correlation_cell_lines:** Target similarity for the pairs of drugs with the highest transcriptomic similarity (>0.6) observed in the same cell line.

Additional file 2. Supplementary text and figures.

Acknowledgements
We would like to thank the entire ChemPert team for releasing the data that was used as the foundation of this work and the reviewers for the helpful comments.

Author contributions
CEH implemented and performed the analyses. CEH and DDF prepared and generated the datasets and networks. CEH, DE and DDF interpreted the results. DDF conceived the study. DE and DDF designed the study. DE, DH and DDF supervised the study. CEH, DE and DDF wrote the paper. All authors have read and approved the final manuscript.

Funding
Not applicable.

Availability of data and materials
All data supporting the conclusions of this article are available at https://zenodo.org/record/7164118 and scripts can be found at https://github.com/enveda/transcriptomic-target-correlation.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
All authors were employees of Enveda Biosciences Inc. during the course of this work and have real or potential ownership interest in the company.

Received: 30 January 2023 Accepted: 14 May 2023
Published online: 19 May 2023

References

1. Corsello SM, Bittker JA, Liu Z, Gould J, McCarren P, Hirschman JE, et al. The drug Repurposing hub: a next-generation drug library and information resource. Nat Med. 2017;23(4):405–8. https://doi.org/10.1038/nm.4306.
2. Domingo-Fernández D, Mubeen S, Marín-Llaó J, Hoyt CT, Hofmann-Apitius M. PathMe: merging and exploring mechanistic pathway knowledge. BMC Bioinform. 2019;20(1):1–12. https://doi.org/10.1186/s12859-019-2863-9.
3. El-Seedi HR, Khalifa SA, Taher EA, Farag MA, Saeed A, Gamal M, et al. Cardenolides: insights from chemical structure and pharmacological utility. Pharmacol Res. 2019;141:123–75. https://doi.org/10.1016/j.phrs.2018.12.015.
4. Figueiredo RQ, del Ser SD, Raschka T, Hofmann-Apitius M, Mubeen S, et al. Elucidating gene expression patterns across multiple biological contexts through a large-scale investigation of transcriptomic datasets. BMC Bioinform. 2022;23(1):231. https://doi.org/10.1186/s12859-022-04765-0.
5. Garrido-Rodriguez M, Zirngibl K, Ivanova O, Lobentanzer S, Saez-Rodriguez J. Integrating knowledge and omics to decipher mechanisms via large-scale models of signaling networks. Mol Syst Biol. 2022;18(7):e11036. https://doi.org/10.15252/msb.202211036.
6. Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, et al. The reactome pathway knowledgebase 2022. Nucleic Acids Res. 2022;50(D1):D687–92. https://doi.org/10.1093/nar/gkab1028.
7. Hanhijärvi S, Garriga GC, Puolamäki K. Randomization techniques for graphs. In: Proceedings of the 2009 SIAM international conference on data mining. 2009;780–791. https://doi.org/10.1137/1.9781611972795.67.
8. Harris CR, Millman KJ, Van Der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020;585(7825):357–62. https://doi.org/10.1038/s41586-020-2649-2.
9. Hosseini-Gerami L, Collier DA, Laing E, Evans D, Broughton H, Bender A. Benchmarking causal reasoning algorithms for gene expression-based compound mechanism of action analysis. BMC Bioinform. 2022. https://doi.org/10.1186/s12859-023-05277-1.
10. Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng. 2007;9(03):90–5.
11. Isik Z, Baldow C, Cannistraci CV, Schroeder M. Drug target prioritization by perturbed gene expression and network information. Sci Rep. 2015;5(1):1–13. https://doi.org/10.1038/srep17417.
12. Iorio F, Bosotti R, Scacheri E, Belcastro V, Mithbaokar P, Ferriero R, et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. Proc Natl Acad Sci. 2010;107(33):14621–6. https://doi.org/10.1073/pnas.1000138107.
13. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res. 2021;49(D1):D545–51. https://doi.org/10.1093/nar/gkaa970.
14. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science. 2006;313(5795):1929–35. https://doi.org/10.1126/science.1132939.
15. Martens M, Ammar A, Riutta A, Waagmeester A, Slenter DN, Hanspers K, et al. WikiPathways: connecting communities. Nucleic Acids Res. 2021;49(D1):D613–21. https://doi.org/10.1093/nar/gkaa1024.
16. McKinney W. Data structures for statistical computing in python. In: Proceedings of the 9th Python in science conference. 2010;445(1): 51–56.
17. Mubeen S, Tom Kodamullil A, Hofmann-Apitius M, Domingo-Fernández D. On the influence of several factors on pathway enrichment analysis. Brief Bioinform. 2022;23(3):143. https://doi.org/10.1093/bib/bbac143.
18. Namba S, Iwata M, Yamanishi Y. From drug repositioning to target repositioning: prediction of therapeutic targets using genetically perturbed transcriptomic signatures. Bioinformatics. 2022;38(1):i68–76. https://doi.org/10.1093/bioinformatics/btac240.
19. Pabon NA, Xia Y, Estabrooks SK, Ye Z, Herbrand AK, Süß E, et al. Predicting protein targets for drug-like compounds using transcriptomics. PLoS Comput Biol. 2018;14(12):e1006651. https://doi.org/10.1371/journal.pcbi.1006651.
20. Rees MG, Seashore-Ludlow B, Cheah JH, Adams DJ, Price EV, Gill S, et al. Correlating chemical sensitivity and basal gene expression reveals mechanism of action. Nat Chem Biol. 2016;12(2):109–16. https://doi.org/10.1038/nchembio.1986.
21. Sirota M, Dudley JT, Kim J, Chiang AP, Morgan AA, Sweet-Cordero A, et al. Discovery and preclinical validation of drug indications using compendia of public gene expression data. Sci Transl Med. 2011;3(96):96ra77-96ra77. https://doi.org/10.1126/scitranslmed.3001318.
22. Szklarczyk D, Santos A, Von Mering C, Jensen LJ, Bork P, Kuhn M. STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. Nucleic Acids Res. 2016;44(D1):D380–4. https://doi.org/10.1093/nar/gkv1277.
23. Trapotsi MA, Hosseini-Gerami L, Bender A. Computational analyses of mechanism of action (MoA): data, methods and integration. RSC Chem Biol. 2022;3:170–200. https://doi.org/10.1039/D1CB00069A.
24. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17(3):261–72. https://doi.org/10.1038/s41592-019-0686-2.

25. Wagner A, Cohen N, Kelder T, Amit U, Liebman E, Steinberg DM, et al. Drugs that reverse disease transcriptomic signatures are more effective in a mouse model of dyslipidemia. Mol Syst Biol. 2015;11(3):791. https://doi.org/10.15252/msb.20145486.
26. Waskom ML. Seaborn: statistical data visualization. J Open Source Softw. 2021;6(60):3021.
27. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46(D1):D1074–82. https://doi.org/10.1093/nar/gkx1037.
28. Zheng M, Okawa S, Bravo M, Chen F, Martínez-Chantar ML, del Sol A. ChemPert: mapping between chemical perturbation and transcriptional response for non-cancer cells. Nucleic Acids Res. 2022. https://doi.org/10.1093/nar/gkac862.

Publisher's Note