npj | Systems Biology and Applications

www.nature.com/npjsba

ARTICLE OPEN

Comparing structural and transcriptional drug networks reveals signatures of drug activity and toxicity in transcriptional responses

Francesco Sirci[1], Francesco Napolitano[1], Sandra Pisonero-Vaquero[1], Diego Carrella[1], Diego L. Medina[1] and Diego di Bernardo[1,2]

We performed an integrated analysis of drug chemical structures and drug-induced transcriptional responses. We demonstrated that a network representing three-dimensional structural similarities among 5452 compounds can be used to automatically group together drugs with similar scaffolds, physicochemical parameters and mode-of-action. We compared the structural network to a network representing transcriptional similarities among a subset of 1309 drugs for which transcriptional response were available in the Connectivity Map data set. Analysis of structurally similar, but transcriptionally different drugs sharing the same MOA enabled us to detect and remove weak and noisy transcriptional responses, greatly enhancing the reliability of transcription-based approaches to drug discovery and drug repositioning. Cardiac glycosides exhibited the strongest transcriptional responses with a significant induction of pathways related to epigenetic regulation, which suggests an epigenetic mechanism of action for these drugs. Drug classes with the weakest transcriptional responses tended to induce expression of cytochrome P450 enzymes, hinting at drug-induced drug resistance. Analysis of transcriptionally similar, but structurally different drugs with unrelated MOA, led us to the identification of a 'toxic' transcriptional signature indicative of lysosomal stress (lysosomotropism) and lipid accumulation (phospholipidosis) partially masking the target-specific transcriptional effects of these drugs. We found that this transcriptional signature is shared by 258 compounds and it is associated to the activation of the transcription factor TFEB, a master regulator of lysosomal biogenesis and autophagy. Finally, we built a predictive Random Forest model of these 258 compounds based on 128 physicochemical parameters, which should help in the early identification of potentially toxic drug candidates.

npj Systems Biology and Applications (2017)3:23 ; doi:10.1038/s41540-017-0022-3

INTRODUCTION

Chemoinformatics approaches to rational drug design have traditionally assumed that chemically similar molecules have similar activities. More recently, transcriptional responses of cells treated with small molecules have been used in the lead optimization phase of drug discovery projects[1] and to reveal similarities among drugs, and quickly transfer indications for drug repositioning.[2–6]

The Connectivity Map (CMAP), the largest peer-reviewed public database of gene expression profiles following treatment of five human cancer cell lines with 1309 different bioactive small molecules,[2, 7] has been extensively used by both the academic and industrial communities.[3, 8]

Whereas computational medicinal chemistry's 'pros' and 'cons' have been extensively addressed over the recent years,[9–17] in contrast, the advantages and limits of methods based on transcriptional responses have not been thoroughly addressed.[1, 3] So far, comparison of the chemical vs. transcriptional 'landscape' of small molecules has been performed to elucidate the molecular mechanisms mediating the therapeutic activity of existing drugs (MOA) and to find new off-label applications.[18–21] In this work, on the contrary, we addressed two still unanswered questions: (1) do transcriptional responses and chemical structures provide similar information on the drug mechanism of action and adverse effects? (2) If not, why does the information provided by transcriptional responses and chemical structures differ?

Answering these questions may help in addressing clinically relevant problems such as drug resistance and drug-toxicity that lie at the interface of chemoinformatics and transcriptomics.[22–24] In this work, we compared chemical structures to transcriptional responses in the CMAP dataset by first generating a 'structural' drug network by connecting pairs of structurally similar drugs, as measured by three-dimensional (3D) pharmacophore descriptors based on molecular interaction fields.[25, 26] We then compared the structural drug network to a transcriptional drug network where drugs are connected if they induce a similar transcriptional profile.[4, 27, 28]

Through the integrated analysis of chemical structures and transcriptional responses of small molecules, we revealed limitations and pitfalls of both transcriptional and structural approaches, and proposed ways to overcome them. Moreover, we found an unexpected link between drug-induced lysosomotropism and lipid accumulation, common adverse effects, and a specific transcriptional signature mediated by the transcription factor TFEB.

[1]Telethon Institute of Genetics and Medicine (TIGEM), System Biology and Bioinformatics lab. and High Content Screening facility, Via Campi Flegrei 34, 80078 Pozzuoli (NA), Italy and [2]Department of Chemical, Materials and Industrial Production Engineering, University of Naples Federico II, Piazzale Tecchio 80, 80125 Naples, Italy
Correspondence: Diego L. Bernardo (dibernardo@tigem.it)
Francesco Napolitano and Sandra Pisonero-Vaquero contributed equally to this work.

Received: 3 April 2017 Revised: 27 June 2017 Accepted: 7 July 2017
Published online: 25 August 2017

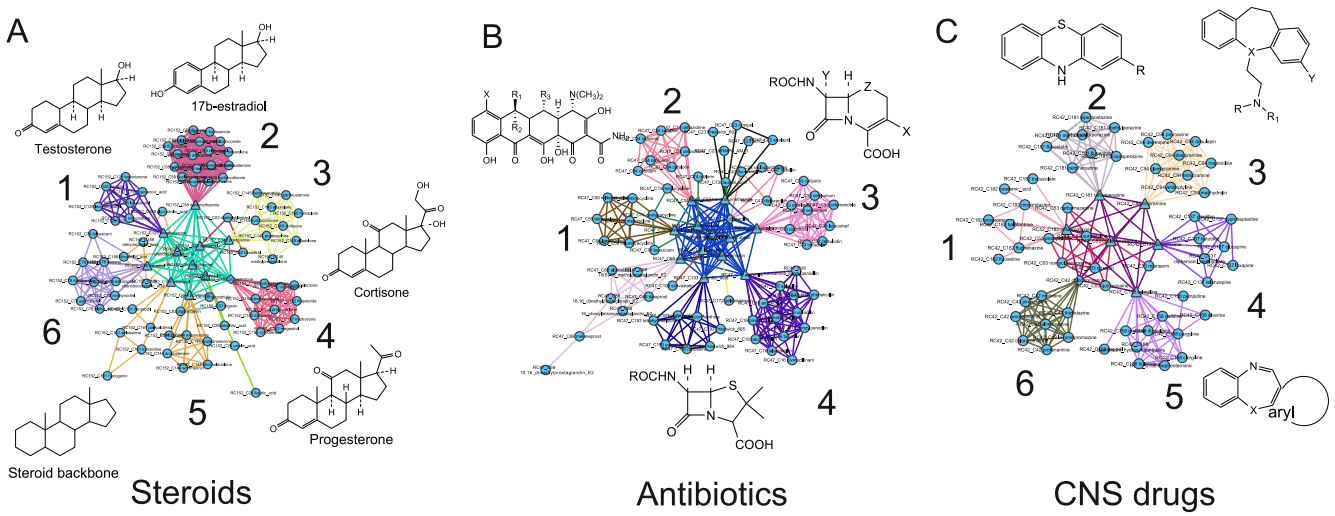

Fig. 1 The structural network among 5452 compounds. The network is partitioned into *communities* (groups of highly interconnected nodes) and *rich-clubs* (groups of communities) sharing common chemical structures and enriched for drugs with similar Mode of Action. Examples of three Rich Clubs are shown. **a** The steroids rich-club (1: testosterone scaffold, 2: estradiol scaffold, 3: cortisone scaffold, 4: progesterone scaffold, 5 and 6: mixed steroids); **b** The antibiotics rich-club (1 and 2: tetracycline scaffold, 3: cephalosporin scaffold, 4: penicillin scaffold); and **c** The CNS-acting drug rich-club (1 and 2: phenothiazine scaffold, 3–6: various tricyclic antidepressant scaffolds)

RESULTS

The CMAP data set is a collection of transcriptional responses of human cell lines to small molecules. It includes transcriptional profiles following treatment of 1309 small molecules across five different cell lines, selected to represent a broad range of activities, including both FDA-approved drugs (670 out of 1309 (51%)) and non-drug bioactive 'tool' compounds.[2] An extension of this data set to more than 5000 small molecules is being completed but it includes only 1000 genes and it has not been peer-reviewed yet (LINCS http://www.lincscloud.org).[2, 7] We selected the small molecules present in the CMAP and in the upcoming LINCS resource for a total of 5452 compounds (Supplementary Fig. 1). We then performed a physicochemical characterization of these 5452 small molecules by computing 128 physicochemical descriptors using 3D molecular interaction fields (MIFs) derived from their chemical structures.[29, 30]

Principal component analysis (PCA) of the 128 descriptors for all the 5452 compounds in Supplementary Fig. 2a reveals that the first two principal components (PC1 and PC2) explain most of the descriptors' variance (53%). PC1 (36%) is related to descriptors of hydrophobic and aromatic properties (Supplementary Fig. 2b), whereas PC2 (17%) to molecular size and shape. Most of these small molecules follow the 'Rule of Fives (RoFs)', that is the set of physicochemical features shared by biologically active drugs: MW $\leq$ 500 Da (89%); N.HBA $\leq$ 10 (93%); N.HBD $\leq$ 5 (97%); LogP $\leq$ 5 (85%) (Supplementary Fig. 3).[31, 32]

Chemical structure similarities induce a hierarchical network connecting drugs with similar scaffolds and mode of action

We derived a *structural drug network* where each small molecule is a node and an edge connects two small molecules if they have similar 3D structures. To this end, we computed the *structural distance* between each pair of small molecules based on the similarity between their 3D-pharmacophore quadruplet-based fingerprints (Methods and Supplementary Fig. 4).[33] A short structural distance (i.e., close to 0) between two compounds indicates that they are structurally similar.

We obtained a symmetric 5452 × 5452 structure-based drug-distance matrix containing 14,859,426 distances between all the possible pairs of drugs. We considered each compound as a node in the network and connected two nodes if their distance was below a threshold value (see Methods section). The resulting drug network consists of 5312 nodes and 742,971 edges, corresponding to 5% of a fully connected network with the same number of nodes (14,859,426 edges). A network representation has the advantage of offering an intuitive and interactive graphical representation of the structural similarities among compounds, enabling to visualize a compound of interest in the context of the overall chemical space. We made available an interactive website to explore and query the structural drug network (http://chemantra.tigem.it). In addition, well-established network analysis tools can be used to partition the network into communities consisting of groups of densely interconnected nodes by means of the Affinity Propagation (AP) clustering algorithm [34, 35] on the network matrix (see Methods section).[4] We thus identified 288 communities (containing more than three drugs) across 5302 drugs (out of 5452) that group together compounds sharing similar chemical functionalities, scaffolds and sub-structural fragments. The AP clustering assigns to each community an 'exemplar', i.e., the drug whose structure best represents the structures of the other drugs in the community. By iteratively applying the AP clustering on the exemplars, we could further group communities into 42 *Rich Clubs*, i.e., *clusters* of drug communities that are structurally related but with distinct characteristic functional groups (Fig. 1).

We then identified the dominant physicochemical parameters for each Rich Club by selecting those parameters whose values tend to be significantly high among compounds within the same Rich Club (see Methods section). Chemotherapeutic agents were mainly found in two Rich Clubs (RC12 and RC19) both enriched for physicochemical descriptors related to hydrophobicity (CD3, CD4, CD5, CD6, CD7) and metabolic stability (MetStab), indicative of their capacity to cross the cell membrane and exert their cytotoxic function. Antihistamines and antipsychotics were found in the same Rich Club (RC5) enriched for permeability-related descriptors (lgBBB), in agreement with their ability to cross the blood–brain barrier (BBB). Cardiac glycosides (RC23) were characterized by descriptors related to size and shape (molecular weight, volume, surface and polar surface area) in agreement with the fact that most of these compounds are large plant-derived molecules. The complete list of physicochemical parameters for each Rich Club can be found in Supplementary Table 1.

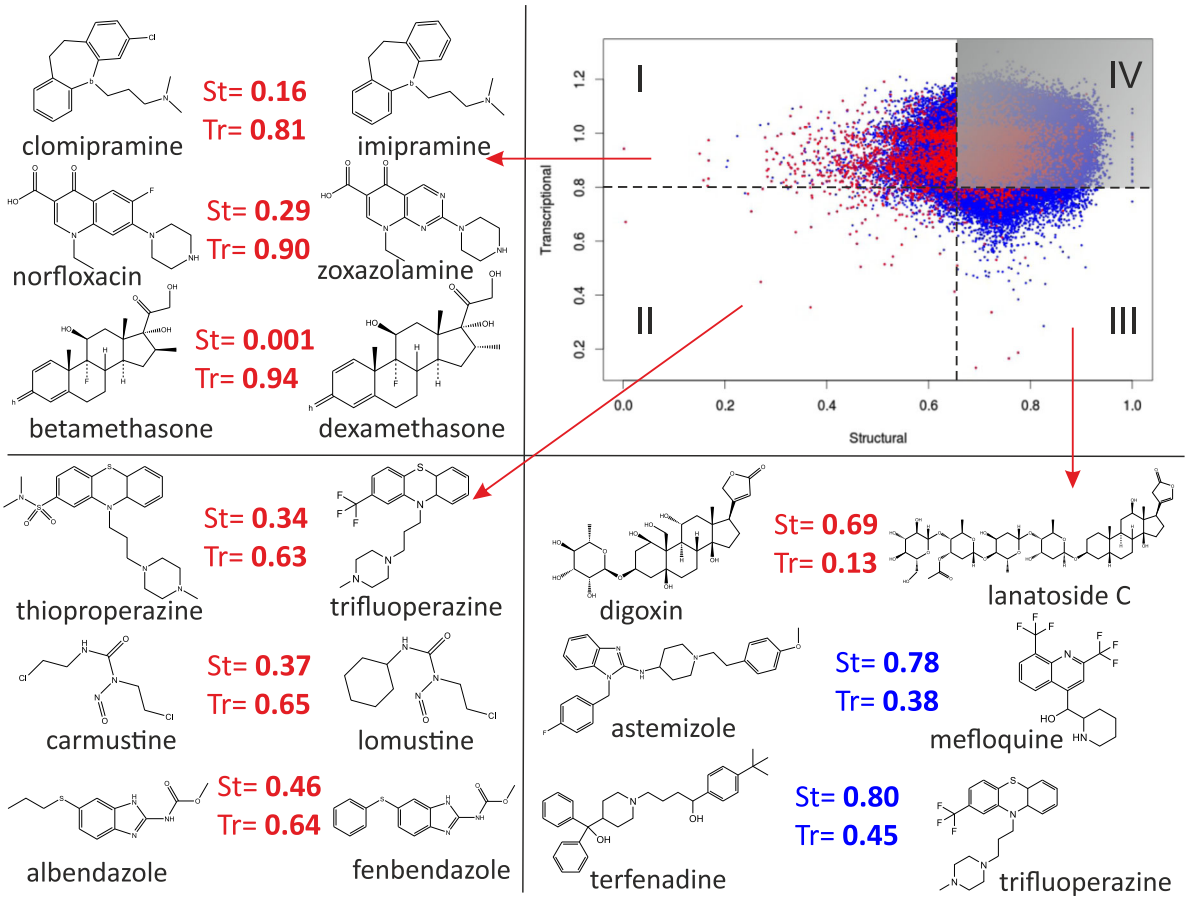

Fig. 2 Comparison of transcriptional and structural distances between 784 CMAP compounds having at least one ATC annotation. Each dot represents the structural (*x*-axis) and transcriptional (*y*-axis) distance between two compounds. A total of 306,936 drug-pairs are shown. Drug-pairs having the same clinical application as annotated by their ATC code are represented by *red dots*. *Dashed lines* represent the significance threshold for the transcriptional (*horizontal line*) and structural (*vertical line*) distance, splitting the plane into four quadrants. Representative examples of drug-pairs are shown for quadrants I, II and III: drug-pairs in quadrant I have similar structure but induce different transcriptional responses; drug-pairs in quadrant II exhibit both similar structure and similar transcriptional responses; drug-pairs in quadrant III have different structures but induce similar transcriptional responses

To assess the structural network, we collected the ATC (Anatomical Therapeutic Chemical) code, an alphanumerical hierarchical pharmacological classification, for 936 out of 5452 drugs (see Methods section). We then verified that drugs connected in the network tend to share the same ATC code (Supplementary Fig. 5). We also verified that drugs within a community share a common therapeutic application. Indeed, 230 out of 288 (80%) structural communities were significantly enriched for compounds sharing the same ATC code (false discovery rate < 0.05) (Supplementary Fig. 6).

These results demonstrate that inspection of the structural drug network can provide useful information on the drug mechanism of action and possibly help in identifying candidates for drug repositioning.

Chemical similarity between drugs is largely uncorrelated with similarity in induced transcriptional responses in CMAP

In a previous study,[4, 27] we reported on the construction of a 'transcriptional network' among 1309 small-molecules part of the CMAP dataset[2] (http://mantra.tigem.it) where two drugs are connected by an edge if they induce a similar transcriptional response. Briefly, in CMAP each transcriptional response is represented as a list of genes ranked according to their differential expression in the drug treatment vs. control. Since each drug is associated to more than one ranked list (cell, dosage, etc.), to obtain the transcriptional network, we first computed a prototype ranked list (PRL) by merging together all the ranked lists referring to the same compound to generate a single ranked list.[4] The PRL thus captures the consensus transcriptional response consistently reducing non-relevant effects due to toxicity, dosage, and cell line.[4] Transcriptional similarity among the 1309 PRLs (one for each drug) was quantified by Gene Set Enrichment Analysis and represented as a distance (i.e., 0 for identical responses, and greater than 0 if dissimilar).[4] The transcriptional network was obtained by connecting two nodes if their distance was below a significant threshold value chosen so that the total number of edges is equal to 5% of a fully connected network with the same number of nodes (856,086 edges).

Here, we compared structural and transcriptional similarities among all pairs of drugs, part of the CMAP dataset, as shown in Fig. 2 and Supplementary Fig. 7 where each point is a drug-pair and its position in the plane represents the structural (*x*-axis) and transcriptional (*y*-axis) distance between the two drugs, for a total of 856,806 drug-pairs. The structural-transcriptional plane can be subdivided into four quadrants by straight lines representing the significance thresholds for the transcriptional (*y*-axis) and structural (*x*-axis) distances: quadrant I (5.1% of drug-pairs) contains drug-pairs with similar structures but inducing different transcriptional responses; quadrant II (0.3% of drug-pairs) contains coherent drug-pairs that are both structurally and transcriptionally similar; quadrant III (4.0% of drug-pairs) consists of drug-pairs with

different structures but inducing similar transcriptional responses; finally drug-pairs different both in structure and transcription are found in quadrant IV (91% of drug-pairs). This quadrant contains most drug-pairs since two random drugs usually have no common function at all. We call drug-pairs in quadrant I and III *incoherent* because of the discrepancies between structural and transcriptional similarities, whereas drug-pair in quadrants II and IV are *coherent*.

Overall, Fig. 2 shows that the information detected by transcriptional responses and chemical structures tend to be different and independent of each other. We therefore decided to investigate the causes for this lack of correlation.

Chemically similar drugs do not induce similar transcriptional responses because of weak transcriptional effects

Drug pairs sharing highly similar chemical structures but very different transcriptional responses are found in Fig. 2 (quadrant I). The most surprising example was the betamethasone/dexamethasone drug-pair. Both drugs are glucocorticosteroids binding the glucocorticoid receptor (GR) with very high affinity and nearly identical in structural since they are enantiomers of each other. Transcriptionally, in contrast, these two drugs appear to be completely different. We then searched for the other drug-pairs composed of glucorticoids and observed that they behave similarly to the betamethasone/dexamethasone pair in that they are mostly found in quadrant I (Supplementary Fig. 8g).

One possible explanation is that these compounds cause no or weak transcriptional effects in the cell lines used in CMAP, probably because they are resistant to these compounds, and thus the measured transcriptional responses are too noisy to be informative.

To assess whether a perturbation (e.g., drug treatment) leads to a strong and informative transcriptional response, we introduce the 'transcriptional variability' score (TV). The TV score is based on the assumption that when the cellular context contains the necessary molecular *milieu* to make it responsive to a small molecule, then multiple treatments with the same compound will yield consistent and similar transcriptional responses. To obtain the TV for a small molecule, we computed the median of the transcriptional distances among its biological replicates in CMAP (see Methods section). A TV close to 0 implies very similar transcriptional responses across replicates, indicating that the small molecule induces a reliable transcriptional response. In contrast, a high TV implies a weak and unreliable transcriptional signature.

To assess whether TV is indeed able to detect informative vs. non-informative transcriptional responses to small-molecules, we exhaustively computed the TV of 1165 CMAP drugs (out of 1309) for which at least two transcriptional responses in the MCF7 cell line were available (Methods and Supplementary Table 2). Out of the 1,165,858 (73%) have a TV score greater than the significance threshold implying that most drugs in CMAP induce a weak transcriptional response (see Methods section).

We compared the TV of drugs belonging to different classes, which were chosen because of their expected activity, or lack thereof, in the CMAP human cancer cell lines (Fig. 3 and Supplementary Table 2). As expected, glucocorticosteroids exhibit higher values of TV when compared to the other classes of drugs. Similarly, antibiotics and NSAIDs induce very weak transcriptional responses (high TV values). Most antihistamines and antipsychotics induce weak transcriptional responses since they target-specific cell membrane receptors lowly, or not expressed, in CMAP cancer cell lines and with no direct transcriptional effects.

We observed that drugs with a high TV, hence exhibiting a weak and noisy transcriptional response, tend to have higher transcriptional distances from the other drugs in CMAP (i.e., they tend to be isolated in the network) and vice-versa (Supplementary Fig. 9).

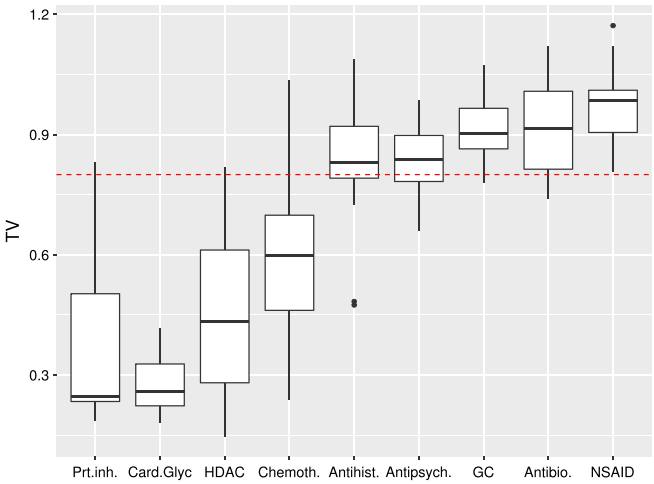

Fig. 3 The Transcriptional Variability (TV) of different drug classes. Box-plots summarizing the TV for drugs within each class. The *bold line* in each box represents the median, while the whiskers represent the 25th and the 75th percentile. Dots represent outliers. Prt.inh.: Protein synthesis inhibitors; HDAC: histone deacetylase inhibitors; Chemoth.: chemotherapeutic agents; Antibio.: antibiotics; NSAIDs: non-steroid antinflammatory agents; GC: glucocorticoids; Antipsych: antipsychotics; Antihist: antihistamines

Consistently with this observation, compounds within these drug classes tend to be found in drug-pairs belonging mostly in quadrant I (structurally similar but transcriptionally different) and quadrant IV (both structurally and transcriptionally different) as shown in Supplementary Fig. 8.

Conversely, drugs with the lowest TV (Fig. 3 and Supplementary Table 2), and thus with strong transcriptional responses, consist mostly of lipophilic molecules acting as protein synthesis inhibitors, chemotherapeutic drugs and other DNA/RNA intercalating agents, and histone deacetylase inhibitors, which all have a strong activity in most cell types. Interestingly, cardiac glycosides were also found to have a low TV. As shown in Supplementary Fig. 8, most drug-pairs within these drug classes tend to be found in quadrant III (structurally different but transcriptionally similar).

Transcriptional phenotyping of low-TV drug classes uncovers cardiac glycosides as potent modulators of epigenetic pathways

We transcriptionally phenotyped the four drug classes with the lowest TV, and hence with the most reliable transcriptional responses (Fig. 3). To this end, we applied Drug Set Enrichment Analysis (DSEA),[36] a method we recently introduced to identify, from transcriptional responses, the molecular pathways that are significantly modulated by most of the drugs in a set. DSEA highlights phenotype-specific pathways, thus helping to formulate hypotheses on the MoA shared by the drugs in the set. We chose to run DSEA using as pathway databases Gene Ontology terms: biological process (BP), cellular component (CC) and molecular function (MF). DSEA results including Enirchment Scores and *P*-value are reported in Supplementary Table 3

Protein synthesis inhibitors were enriched for pathways related to translation, such as tRNA ligase activity (MF), ribosome (CC), ER and Golgi compartments (CC), but also related to steroid biosynthesis (BP) (Supplementary Table 3). Interestingly, block of steroid synthesis is a well-known effect of protein synthesis inhibitors in vivo.[37]

Chemotherapeutic agents, as expected, strongly induced the p53-mediated DNA damage response pathway (BP), several cell cycle-related pathways (BP, CC and MF) and pathways related to the kinetochore (CC) and microtubule motor activity (MF).

In the case of *HDAC inhibitors*, DSEA found enriched pathways related to histone acetyl transferase activity (MF, BP), chromatin remodeling (BP) but also to mitochondria (CC) and RNA splicing (CC, BP), a recently discovered but still not fully dissected effect of HDAC inhibitors.[38]

The most interesting observation was made for *cardiac glycosides* that strongly modulate pathways involved in epigenetic regulation, such as histone acetylation (BP), nucleosome assembly (BP, CC) and transcription from pol II (BP) (Supplementary Table 3). This is an unexpected finding, as these drugs target Na^+/K^+ ATPase pumps. Interestingly, in a recent unbiased epigenetic drug screening using FDA-approved drug libraries, cardiac glycosides were indeed found to potently reactivate silenced gene expression via epigenetic mechanisms probably mediated by calcium signaling, and independent of their ATPase pump inhibitory effects.[39] This activity of cardiac glycosides may be the reason for the strong transcriptional responses they induce, as evidenced by the low TV of this drug class.

Lack of drug activity in high-TV drug classes is partly mediated by cytochrome P450 enzyme expression

We hypothesized that some drug classes may exhibit weak transcriptional responses (i.e., high TV), because of drug-induced drug resistance. We thus evaluated the expression of genes involved in cytochrome P450-mediated drug metabolism and in drug efflux in each of the drug classes in Fig. 3 by means of Gene Set Enrichment Analysis (Supplementary Fig. 10). We found a positive correlation (Pearson Corr. Coeff. = 0.64 in Supplementary Fig. 10) between Transcriptional Variability and the expression of Cytochrome P450-mediated drug metabolism genes across drug classes (whereas no correlation was found for drug efflux genes—data not shown). Hence, drug classes with a weak transcriptional response tend to upregulate the expression of cytochrome P450-mediated drug metabolism, which may explain, at least in part, why these drugs induce weak transcriptional responses.

Removing weak transcriptional responses from the CMAP data set improves drug classification performances

We reasoned that by removing drugs with a high TV, the performance of computational approaches based on gene expression to elucidate the MoA of a drug should improve.[4, 27] We thus partitioned the small molecules included in CMAP in two sets according to their TV score, obtaining a high-TV set and a low-TV set with the same number of drugs to facilitate the comparison. We then assessed the performance of the transcriptional distance between two drugs in correctly identifying those pairs sharing the same therapeutic application (i.e., the same ATC code), when using either drugs in the high-TV set or those in the low-TV set, as previously described.[4] As shown in Fig. 4, the low-TV set performance far exceeds the high-TV set performance, which is almost random. Moreover, the correlation between structural distance and transcriptional distance in the chemical-transcriptional landscape of small molecules in Fig. 2 increases if drugs in the low-TV set only are used (Supplementary Fig. 11).

Overall, these results show that the TV score can discriminate between informative and non-informative transcriptional responses that result from the activity, or lack thereof, of small molecules in a specific cell line.

Drugs with different chemical structures and modes of action may induce similar transcriptional responses related to lysosomal stress and phospholipidosis

Figure 2 (quadrant III) includes drug-pairs with very different molecular structures but which are transcriptionally similar. We identified at least two causes for the discrepancy between transcriptional and structural similarities: (i) drug-pairs in this quadrant tend to have at least one drug with a very large size (11% in quadrant III vs. 1% in quadrant I and 3% in quadrant II), as shown in Fig. 2 and Supplementary Fig. 12; hence, global chemical similarity metrics, such as the one used here, may fail; (ii) the direct molecular targets of two drugs in a pair may be different but act in the same pathway (e.g., purine synthesis inhibitors methotrexate/mycophenolic-acid that act on different molecular targets but both block DNA synthesis, Supplementary Fig. 12).[40–42] We estimated that this effect applies to between 32% and 61% of the drug-pairs in quadrant III, depending on the pathway database used (see Methods section).

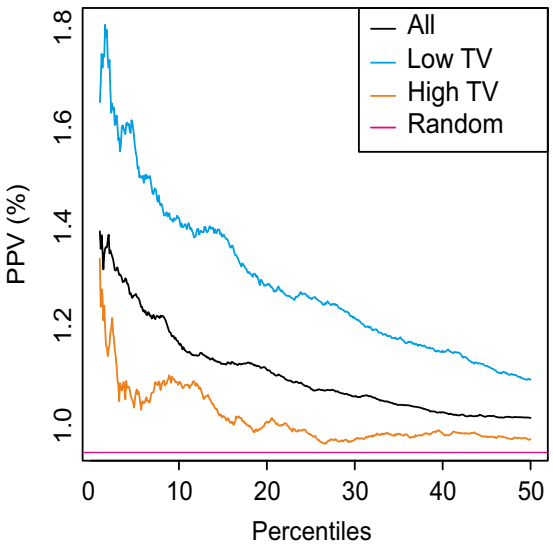

Fig. 4 Performance of the transcriptional distance in detecting drugs with the same ATC code. Compounds were divided into three sets: (all) the 1165 compounds in CMAP having at TV value; (high TV) 582 compounds with a TV higher than the median TV among all the compounds; (low TV) 582 compounds with a TV lower than the median TV. For each set, the transcriptional distance of each drug-pair was computed. Drug-pairs were then sorted according to their transcriptional distance, with drug-pairs with the smallest distance towards the origin of the *x*-axis; the positive predictive value (PPV) was computed as the percentage of true positives over false positives plus true positives and shown on the *y-axis*. The PPV obtained by randomly sorting drugs is also shown (Random)

Figure 2 (quadrant III), however, contains also a large fraction of drug-pairs that are not large molecules and do not act in the same pathway, nor share the same therapeutic application, but nevertheless have very similar transcriptional profiles. To investigate why this is the case, we ranked drug-pairs in this quadrant by their transcriptional distance in ascending order (Supplementary Table 4). We noticed that the top-ranked most transcriptionally similar drug-pairs included well-known 'lysosomotropic agents' inducing large vacuolization in cells such as astemizole, terfenadine and mefloquine (Table 1).[43–45] Among these agents, astemizole and terfenadine are no longer in use because of cardio-toxicity caused by their potassium channel blocker activity (hERG encoded by *KCNH2*), which may lead to fatal cardiac arrhythmia.[46, 47] The lysosomotropic effect of these small molecules has been attributed to their ability to cross lysosomal membrane and remain trapped within the lysosome by a mechanism known as pH partitioning.[48, 49–51] Most lysosomotropic agents belong to the class of cationic amphiphilic drugs (CADs) containing both a hydrophobic and a hydrophilic domain. CADs have increased probability to cause drug-induced phospolipidosis (PLD),[52] a lysosomal storage disorder characterized by the accumulation of phospholipids within the lysosome by unclear molecular

Table 1. Drug-pairs with different chemical structures but inducing very similar transcriptional responses

Drug A	Drug B	Tr. Dist.	Str. Dist.
digoxin	lanatoside_C	0.131	0.693
digoxin	proscillaridin	0.166	0.758
lanatoside_C	proscillaridin	0.187	0.776
rifabutin	vorinostat	0.286	0.826
astemizole	*terfenadine*	0.337	0.724
astemizole	*mefloquine*	0.385	0.776
doxorubicin	mitoxantrone	0.414	0.651
mefloquine	*terfenadine*	0.421	0.767
chlorzoxazone	clindamycin	0.442	0.829
chlorzoxazone	glibenclamide	0.445	0.791
terfenadine	**trifluoperazine**	0.453	0.758
irinotecan	*phenoxybenzamine*	0.455	0.800
suloctidil	*terfenadine*	0.466	0.696
astemizole	**trifluoperazine**	0.469	0.718
protriptyline	**trifluoperazine**	0.472	0.713
niclosamide	**trifluoperazine**	0.472	0.688
mefloquine	**trifluoperazine**	0.478	0.674
doxazosin	sulconazole	0.481	0.776
lomustine	*phenoxybenzamine*	0.484	0.719

Drug-pairs in Fig. 2 (quadrant III) were ranked by transcriptional distance (Tr. Dist.). Only the top 20 ranked drugs pairs are shown together with their structural distance (Str. Dist.). Lysosomotropic drugs are shown in italic and phospholipidosis inducing drugs in bold. Shaded rows highlight when one of the member of a pair is CAD or PLD drug

mechanisms, leading to cellular stress.[53–57] Indeed among the lysosomotropic drugs involved in the most transcriptionally similar drug-pairs (Table 1), there were also three known PLD-inducing drugs (astemizole, suloctidil and trifluoperazine).

We hypothesised that 'lysosomotropic' stress induced by these compounds could explain their similarity in transcriptional responses. We therefore selected 187 CAD compounds present in CMAP according to their physicochemical properties (LogP > 3; pKa > 7.4).[50] Within these CAD compounds, we searched the literature for lysosomotropic drugs known to induce PLD,[52] which, according to our hypothesis, should elicit a strong transcriptional response. We thus identified a total of 36 compounds (PLD/CAD) (Supplementary Table 5).

We verified that PLD/CAD compounds tend to induce a stronger transcriptional response (i.e., a lower TV) (Supplementary Fig. 13) and they tend to be transcriptionally similar among them (but not structurally) despite having different MOA and therapeutic applications (Supplementary Fig. 14).

We next asked which genes were transcriptionally modulated by the majority of PLD/CAD compounds. We applied DSEA[36] to the 36 PLD/CAD compounds and found that the most significant gene set, out of about 5000 gene-sets within the Gene Ontology (GO) database, was the GO-Cell Component term 'lysosome' consisting mainly of genes coding for lysosomal enzymes and ion channels ($p = 5.03 \times 10^{-8}$—Supplementary Table 6), thus in agreement with the 'lysosomotropic' effect of these drugs.

Recently, the transcription factor E-box (TFEB) has been found to be a major player in the transcriptional control of lysosomal genes in response to a variety of cellular and environmental stresses.[58] In normal nutrient conditions TFEB is phosphorylated by the mTORC1 complex on the lysosomal surface. This phosphorylation favors TFEB binding to 14-3-3 proteins and its retention in the cytoplasm.[59–61] Upon stress signal, such as nutrient deprivation, mTOR is inhibited, the calcium-dependent phosphatase Calcineurin is activated, and TFEB is de-phosphorylated shuttling to the nucleus where it transcriptionally controls lysosomal biogenesis, exocytosis and autophagy.[59–65] Moreover, TFEB was shown to translocate to the nucleus upon amiodarone treatment, a well-known lysosomotropic agent.[60] We thus decided to investigate whether TFEB activation was responsible for the characteristic transcriptional response induced by PLD/CAD compounds.

The transcriptional response of PLD-inducing compounds is associated to TFEB translocation

We performed a panel of high content screening (HCS) assays including the TFEB nuclear translocation assay (TFEB-NT)[65] at 3 h and 24 h following drug administration at different concentrations (0.1, 1 and 10 μM) for 34 out of 36 PLD drugs (two drug was not available to us at the time). HCS assays at 24 h included LAMP-1 immunostaining and Lysotracker dye to quantify lysosomal compartment (see Methods section), GM130 and PDI immunostaining to detect morphological changes in the Golgi and ER (Endoplasmic Reticulum) compartments, both of which have been recently suggested to be involved in PLD etiology (see Methods section). We also performed the LipidTox assay at 48 h to check for the accumulation of phospholipids to confirm PLD at least in vitro (see Methods section).

Quantification of the HCS assays for the 34 PLD drugs is reported in Supplementary Table 7. Nuclear translocation of TFEB at 3 h was observed for 18 out of 34 drugs (53%) increasing to 29 drugs at 24 h (85%). Out of these 29 drugs, 27 induced an increase in lysosome size and number as evidenced by LAMP1 and Lysotracker staining, and all 29 drugs induced accumulation of phospholipids according to the Lipidtox assay (100%). Only five drugs did not induce TFEB translocation at 24 h, and just one out of these five drugs was positive in the Lysotracker assay, while four of them were positive in the Lipidtox assay. None of the drugs tested were positive for the Golgi marker and only six were positive for the ER marker, albeit marginally.

Overall, HCS confirmed a concentration dependent nuclear translocation of TFEB for 29 out of 34 drugs (85% at 24 h) with a concomitant perturbation of the lysosomal compartment for 28 out of 34 drugs (82%) occurring mostly at the highest dosage tested (10 μM). Furthermore, HCS revealed an accumulation of lipid in vitro at 48 h following treatment with the 34 drugs (100%) at the highest dosage tested (10 μM), as previously reported in the literature.[52]

These results support the role of TFEB in shaping the transcriptional response of cells treated with PLD-inducing drugs in a way completely unrelated to their MoA. We next asked whether the activation of TFEB (or TFE3, another member of the MiT family of transcription factors with similar functions) is a consequence of lysosomal stress upon compound treatment or if it is directly related to the induction of the PLD phenotype. Thus, we set up a HCS Lipidtox assay using TFEB wt vs. TFEB/TFE3 KO in HeLa cell type, administering high dosage of chloroquine (50 μM) known to induce lipids accumulation in cells at 48 h. Supplementary Fig. 15a, b show no major differences in terms of spot intensity in the Lipidtox assay, thus confirming that TFEB activation is a consequence of lysosomal stress and not an inducer of PLD. As this manuscript was under review, Lu et al. reported an increase in TFEB, TFE3 and MITF translocation to the nucleus in ARPE-19 cells together with lysosomal activation and lipid accumulation following treatment with eight lysosomotropic compounds, well in agreement with our results.[50]

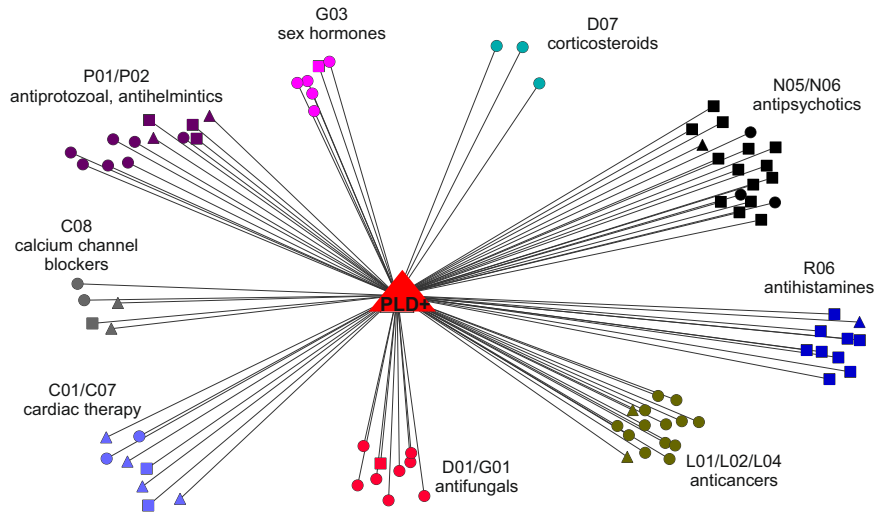

Fig. 5 Drugs inducing a lysosomotropic gene expression signature. The transcriptional responses elicited by eight lysosomotropic compounds were combined into a single node in the transcriptional drug network (*red triangle*). Transcriptional distances to this lysosomotropic gene expression signature were computed for all the 1309 drugs in CMAP. Only drugs with a transcriptional distance below the significance threshold are shown (0.8) and color-coded according to their ATC classification. *Triangles* (PLD + drugs); *squares* (CAD + drugs); *circles* (CAD- and PLD)

A PLD-specific transcriptional signature can predict compounds inducing lipid accumulation

We combined the transcriptional responses elicited by the 36 PLD/CAD compounds into a consensus transcriptional response ('PLD' signature) and computed its transcriptional distance from all the other 1273 (i.e., 1309-36) CMAP compounds (see Methods section). We reasoned that drugs inducing a transcriptional profile similar to the PLD signature should have a higher probability of inducing lipid accumulation than the other drugs. Surprisingly, 258 compounds out of 1274 (20%) cMAP compounds were found to be similar to the PLD signature (Supplementary Table 8). About a third of these drugs are CADs (77 out of 258 (30%)).

Figure 5 reports a breakdown by ATC classes of drugs for which an ATC code was available and that were found to induce a transcriptional response similar to the PLD signature. Some drug classes (ATC classes N05, N06 and R06 including antihistamines and antipsychotics) are enriched for known PLDs.[52, 54] Other classes cause global cellular stress responses not mediated by their physicochemical properties, but rather because of their direct molecular targets, such as anti-cancer compounds that block cell cycle (e.g., ATC class L01 composed of CDK2 and Topoisomerase I, II inhibitors). Antihelmintics (ATC P02) and antifungals (ATC D01), despite being neither CADs nor PLDs, were also found among the PLD node's neighbors. Several recent reports in the literature have found antihelmintics to induce an anti-proliferative effect in cancer cell lines by indirectly inhibiting the mTOR pathway thus inducing TFEB activity, which may explain their PLD-like transcriptional response.[61, 66–69] Calcium channel blockers were also found to induce a transcriptional response similar to PLDs, which may be expected since calcium signaling has been involved in autophagy regulation and lysosomal function.[65] Interestingly, some cardenolides (ATC C01 and C07) were also found to contain the PLD signature, despite not being CADs (median distance equal to 0.71).[70, 71]

To experimentally validate the usefulness of the PLD transcriptional signature in identifying novel PLD drugs, we selected the top quartile of the 258 drugs (i.e., 25% of 258 = 64 drugs) with the shortest transcriptional distance to the PLD node and performed HCS for lipid accumulation following drug treatment at three different concentrations (Lipidtox assay) (Supplementary Table 9). Twenty-two out of the top 64 small molecules were present in our HCS small-molecule library. Overall 11 out of 22 (50%) compounds were positive to the Lipidtox assay (Supplementary Table 9), including Terfenadine, a cardiotoxic lysosomotropic CAD, not reported to be a PLD inducer in the literature, which caused a strong accumulation of lipids, as shown in Fig. 6 (LipiTox Intensity Spot: 450.93 at a concentration of 10 μM).

Overall, our data demonstrate the value of the PLD transcriptional signature in identifying compounds potentially inducing lysosomal stress and phospholipidosis.

Improved physicochemical models of compounds causing lysosomal stress and potentially PLD could aid in rational drug design. To this end, we selected the 258 compounds most similar to the PLD transcriptional signature, and identified the physicochemical parameters that best distinguish these drugs from the rest. We applied a random-forest model to classify the 258 compounds (out of the 1309 compounds) using as features the 128 physicochemical descriptors (Supplementary Methods). The overall classification error rate (OOB, out of bag) was of 22%. Interestingly, the most important features used by the classification model were the Log-P (First ranked; Supplementary Fig. 16a, b) and the pH-specific log-D (lgD8 and lgD10 ranked Second and third; Supplementary Fig. 16a, b), which basically recapitulate the physicochemical properties of CAD drugs (LogP > 3; pKa > 7.4).[50] The other descriptors ranked from position 4 to 10 (Supplementary Fig. 16a, c-f) include solubility at various pH (lgS9, lgS8, LgS7.5 and LgS4), the volume of hydrophobic interactions at −1kcal/mol (D5) and the concentration of hydrophobic interactions on the molecular surface at two energy levels (capacity descriptors CD2 at −0.4 kcal/mol and CD5 at −1 kcal/mol). This random-forest model should be of value in the early identification of drug candidates which may potentially cause lysosomal stress and PLD.

The PLD transcriptional signature affects transcriptional responses to drug treatment in a concentration dependent manner

We next investigated whether the PLD expression signature was linked to the elevated drug concentration used in the CMAP experiments, in agreement with the HCS results indicating a dose dependent TFEB nuclear translocation (Supplementary Fig. 17 and Supplementary Table 7). Indeed 5747 out 6100 CMAP gene expression profiles (94%) were measured at high drug

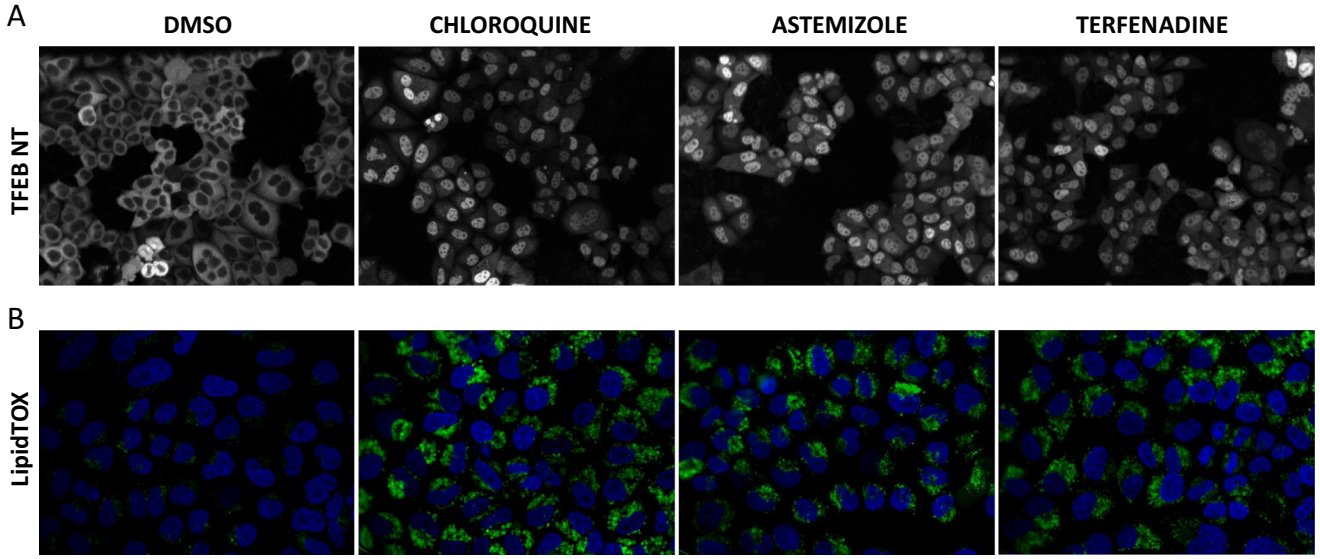

Fig. 6 Effects of drugs on TFEB nuclear translocation and LipidTOX assay. **a** TFEB localization in stably HeLa cells overexpressing TFEB-GFP and treated with DMSO or the indicated drugs. **b** Lipid accumulation in HeLa cells was detected by staining with LipidTOX reagent upon drug treatment

concentrations ranging from 1 μM to 10 mM, while the remaining 353 (6%) at lower concentrations ranging from 10 nM to 0.5 μM. We thus searched CMAP for PLD-inducing drugs for which both high and low concentration instances were present. We selected five drugs (out of 36 PLD) drugs: raloxifene (ER antagonist at 0.1 and 7.8 μM), tamoxifen (ER antagonist at 1 and 7.0 μM), amitriptyline (antidepressant 1 and 12.8 μM), thioridazine (antipsychotic at 1 and 10 μM) and chlorpromazine (antipsychotic at 1 and 11.2 μM). We then generated two additional transcriptional responses (LOW and HIGH) for each of these five drugs by analyzing separately the low and high concentration experiments (Methods, Supplementary Fig. 18 and Supplementary Table 10).

The HIGH transcriptional responses for the five drugs were more similar to the PLD signature than the corresponding LOW transcriptional responses (Supplementary Table 10), confirming an increased alteration of the transcriptional response caused by high drug dosages. Moreover, the HIGH transcriptional responses of four out of five drugs were connected to a much larger number of drugs in the transcriptional network when compared to their LOW transcriptional response counterparts (Supplementary Fig. 18). Raloxifen, a selective estrogen receptor modulator (SERM), is the only drug tested also at sub-micromolar concentrations (0.1 μM). When using the HIGH transcriptional response, raloxifene is predicted to be transcriptionally similar to 154 compounds (Supplementary Fig. 18 and Supplementary Table 10), none of which behaving as a SERM, with the most similar being trifluoperazine, an antipsychotic drug with known PLD-inducing properties. On the contrary, when the LOW transcriptional response is used, raloxifene is predicted to be transcriptionally similar only to four compounds, the most similar one being tamoxifen, a well-known SERM.

DISCUSSION

By analyzing a large set of chemical structures, we generated a network representing structural similarities among compounds that can be used to automatically group together drugs with similar scaffolds and MOA. Other methods to cluster drugs based on structural similarity have been proposed in the literature[16] but no hierarchical classification of drugs in communities and rich-clubs based on the network structure has been previously performed. We also computed 128 physicochemical parameters for each compound and identified the dominant parameters for each Rich-Club. These data are a useful resource for compound characterization based on physicochemical properties. By comparing the structural drug network with the transcriptional drug network, we observed broad differences between the two: drugs can be very similar in terms of the transcriptional response they induce, but with unrelated chemical structures, or vice-versa have very similar structures but induce diverse transcriptional responses.

Here, we identified a set of confounding factors that can hinder the usefulness of transcriptional based methods. We introduced a simple but powerful measure, 'Transcriptional Variability', to assess the strength and robustness of the transcriptional response of a cell to a drug treatment. Cardiac glycosides were among the drug classes with the lowest TV, hence inducing the strongest transcriptional responses. By analyzing their transcriptional profiles, we found a strong induction of genes involved epigenetic regulation, supporting the repositioning of cardiac glycosides as epigenetic drugs.[39]

In the original CMAP study,[2] the authors indeed recognized that although gene expression signatures can be highly sensitive, they may be uninformative if measured in cells that lack the appropriate physiological or molecular context, but offered no solution to identify such cases. We propose the use of the TV score to identify these uninformative profiles. We observed that glucocorticoids tend to have a high TV, hence uninformative transcriptional profiles. Indeed, MCF7,[72, 73] HL60 and PC3 (refs. 74, 75) cell lines used in CMAP may exhibit resistance to glucocorticosteroids.[2] Hence, if not filtered out, computational analysis of their transcriptional responses may be misleading and lead to wrong conclusions, e.g., such that betamethasone and dexamethansone have a different MOA (Fig. 2). Interestingly, the TV score could be used also to uncover cell lines that a resistant to a specific drug treatment, and hence have no or very weak transcriptional response to that drug. Interestingly, we verified that drug classes exhibiting a high TV tend to induce expression of cytochrome P450 enzymes involved in drug metabolism, suggesting that some drug classes may induce drug resistance, at least in the CMAP cell lines. In contrast, drug classes with a low TV tend to reduce expression of these enzymes (Supplementary Fig. 10).

We uncovered a transcriptional signature common to a subset of transcriptionally similar but structurally distinct drugs profiled in

CMAP that is not related to their MOA, but rather to cellular toxicity caused by lysosomal stress and lipid accumulation. We also derived a predictive model based on physicochemical parameters to identify such compounds. We further demonstrated by HCS that PLD-inducing drugs have little effect on ER and Golgi morphology, but rather increase the number and size of lysosomes, as previously reported in the literature, and induce the nuclear translocation of the transcription factor TFEB, a master regulator of lysosomal biogenesis and autophagy. The transcriptional signature present in the transcriptional response of PLD-inducing drugs is likely driven by TFEB activation. These results may help in further elucidating the effect of lysosomotropic PLD-inducing drugs on autophagy.[76] Moreover, the PLD transcriptional signature may be a useful tool for identifying and repositioning drugs as inducers of TFEB activation and thus of authophagy.[63]

Our findings are directly relevant for all those studies relying on CMAP transcriptional responses to determine drug MOA and for drug repositioning. Here, we show that very high and not physiological compound concentrations, such as the ones used in the CMAP dataset, increase the chance of off-target effects including lysosomotropism and phospholipidosis. Somewhat surprisingly, despite the high concentrations used, only a minority of compounds in CMAP (~ 30%) have reproducible transcriptional responses (TV < 0.8). Notwithstanding these limitations, the CMAP still contains relevant information on drug activity if properly analyzed, allowing to correctly discriminate among different classes of drugs[3] and it can provide complementary information to that obtained by HCS.[4, 77–79]

Our results, although derived from the CMAP dataset, can be used to draw general guidelines to prevent inconsistencies and erroneous conclusion when using transcriptional responses of small molecules for drug discovery and drug repositioning: (i) the transcriptional response elicited by a drug in a specific cell line can be uninformative. Hence these responses must be detected and then excluded from further analyses. We demonstrated that this can be achieved by assessing the Transcriptional Variability (TV) of the drug-induced transcriptional response across multiple replicates; (ii) drug treatment can cause cellular stress unrelated to the drug MoA and thus affect the drug-induced transcriptional response by partially masking transcriptional changes directly related to the drug molecular targets. We generated a PLD transcriptional signature which can be used to detect these compounds. This signature is particularly strong if drug concentrations used to treat cells are above their clinically relevant concentrations. One way to avoid this is to use clinically relevant (sub-micromolar) concentrations; (iii) in the case of natural compounds, computational approaches based on transcriptional responses maybe more informative than those based on structural approaches, because of the large size and molecular complexity of these compounds.

METHODS

Compounds

We retrieved the chemical structure of 5500 small-molecules part of the Library of Integrate Network-based Cellular Signatures (LINCS—http://lincscloud.org) project in the form of SMILES string annotations (Supplementary Information). 4719 out of 5500 SMILES strings were retrieved according to their annotated ChemSpider ID (CSID) and PubChem ID (PID) in the NIH LINCS database. The remaining 779 NHS LINCS structures, for which no CSID or PID annotation was found, were retrieved by a web-API search in ChemSpider according to the molecule names. Six compounds were restricted structures. Thus, a final collection of 4927 LINCS unique structures was obtained. In addition, we retrieved chemical structures for the 1309 small-molecules part of the CMAP data set.[2, 7] 784 out of 1309 small-molecules were already present among the 4929 LINCS unique structures. Thus only 523 unique CMAP structures were retrieved as described before (Supplementary Fig. 1). The total number of chemical structure used for further analysis was thus equal to 5452.

The ChemAxon Standardizer tool (v. 14.9) was run to convert SMILES string annotations into two-dimensional multi-structure-data File (SDF) structural files.[80] The 'remove fragments' and 'neutralize' options were used to fix all the molecular structures, to remove counter-ions and other various kinds of molecular fragments, which may be present in branded drug formulation but not useful in this work (e.g., besilates, mesilates, chlorides, bromides, sulfates, etc.). Protonation state of each structure was calculated with MoKa software v. 2.0 considering physiological pH 7.4.[81]

Finally, 3D minimized conformations were generated with the MMFF4x force field in the MOE software (v. 2013)[82] and stored as 3D multi-SDF structural files. The MMFF4x is the standard force field parameterized for small organic molecules such as drugs. Partial charges are based on bond-charge increments. Conjugated nitrogens are considered as planar. Thus, a unique 3D multi-SDF file was obtained and used as input file for all the subsequent analyses.

Physicochemical and pharmacokinetic properties

Starting from the three-dimensional (3D) coordinates multi-SDF file, each structure was the imported in the Volsurf+ v.1.5 software[30] normalizing their protonation state at pH 7.4. A set of 128 physicochemical and pharmacokinetic descriptors were calculated using Volsurf+ v. 1.5, using a grid spatial resolution of 0.5 Å. A final matrix of 5452 objects (drugs and chemical substances) and 128 descriptors was thus obtained. The molecular descriptors matrix was then visualized through the PCA tool integrated in Volsurf+. Only the first five PCs were considered for the analysis. PCA score and loading plots are shown in Supplementary Fig. 2a and Fig. 2b. Analysis of the physicochemical descriptor distribution plots are shown in Supplementary Fig. 3.

3D structural similarities by pharmacophore descriptors

The software FLAP v. 2.0[33] was used to compute all-against-all pairwise 3D structural similarities among the 5452 compounds. FLAP allows 3D molecular superimposition of two molecules and computes a pairwise similarity score based on MIFs, in order to evaluate type, strength, and direction of the interactions a molecule can have. The GRID tool,[26] part of the FLAP software was used to compute the MIFs based on three interaction probes: H, DRY and OH2. The hydrogen probe H is used to compute the shape of a small molecule. The hydrophobic probe DRY finds places at which hydrophobic atoms on the surface of a target molecule will make favorable interactions with hydrophobic ligand atoms. The probe OH2 represents polar and hydrophilic interactions mainly generated by hydrogen bond donor and acceptor functional groups and charges interactions. Four-point pharmacophores derived from the MIFs were used to align molecules with specific biological activity.[33, 83, 84] The evaluation of MIF volume superimpositions between the two structures is reported as a similarity score ranging from 0 to 1 for each of the three probes. A global score (GLOB-Sum) is then obtained as the sum of the three scores of the individual probes. Higher GLOB-S values correspond to more similar structures. For this study, we transformed the GLOB-Sum similarity score matrix (**S**) of dimension 5452 × 5452 into a distance matrix defined as **D** = 1-**S**/3.

Since the distance matrix is symmetric (i.e., the distance between A and B is the same as the distance between B and A), the total number of drug-pairs to consider is 14,859,426 (5452 × 5451 /2).

Construction of the drug network

We ranked drug-pairs according to their structural distance in ascending order and considered as significant only those drug-pairs in the top 5% of the ranked list, as previously described by Iorio et al.[4] to reduce the total amount of egdes in the MANTRA network (The distance threshold is 0.51 when considering the 5452 × 5452 network or 0.65 when considering only the CMAP 1309 × 1309 sub-network). We then represented drugs as nodes connected by edges. The resulting Structural Drug Network has a giant connected component with 5312 nodes (i.e., drugs) out of 5452 and 35,527 edges, corresponding to 5% of a fully connected network with the same number of nodes (14,859,426 edges) (Supplementary Fig. 4). In order to visualize and extract useful information from the SDN, we identified communities via the AP Clustering algorithm, as implemented in the R package apcluster (v. 1.3.5).[35, 85] A community is defined as a group of nodes densely interconnected with each other and with fewer connections to nodes outside the group.[86] Each community was coded with a numerical identifier, a color, and one of its nodes was identified as the

'exemplar' of the community, i.e., the drug whose effect best represents the effects of the other drugs in the community.[4]

Enrichment of physicochemical parameters for the 42 rich-clubs

In the physicochemical Volsurf+ matrix (5452 × 128), we sorted the 5542 drugs independently in each column, thus obtaining 128 ranked lists of drugs. In each list, drugs are sorted according to their value of the Volsurf+ parameter in descending order. We then applied the Kolmogorov–Smirnov (KS) test, similar to the Gene Set Enrichment Analysis, to each of the 42 Rich-Clubs in order to identify whether drugs in the same Rich Club were significantly found in the top ranks of one, or more, of the 128 ranked lists. We thus obtained an Enrichment Score and a p-value of the 128 VolSurf+ parameters from the KS test for each Rich Club. These results are reported in Supplementary Table 1.

Validation of the structural drug network

To validate the drug structural network, we assessed whether pairs of drugs connected by an edge in the network (i.e., structurally similar according to our distance) shared a common clinical application. To this end, we collected for each drug the correspondent Anatomical Therapeutic Chemical (ATC) code (version Index 2014). This drugs classification method developed by the World Health Organization in collaboration with the Drug Statistics Methodology (WHOCC),[87] hierarchically classifies compounds according to five different levels: (first level) Organ or system on which they act; (second level) Therapeutic class; (third level) Pharmacological subgroup; (fourth level) Chemical subgroup; (fifth level) Compound identifier. ATC code collisions often occur for the same drug. For instance, Aspirin has three distinct ATC codes: A01AD05 (drug for alimentary tract and metabolism), B01AC06 (blood agent as platelet inhibitor) and N02BA01 (nervous system agent as analgesic and antipyretic). In such cases we considered multiple ATC codes for the same drug in the network. ATC codes available from the WHOCC were 936 out of 5452 drugs (17%).

We then sorted drug-pairs according their structural distance in ascending order and for each drug-pair we checked whether they shared the same ATC to assess whether it was a true positive (TP) or a false positive (FP). Supplementary Fig. 5 reports the PPV = TP/(TP+FP) vs. the drug-pair distance for different ATC code levels.

Furthermore, in order to assess whether a community in the drug network was enriched for a common ATC code, we counted the number of drugs with the same ATC code at the 4th level (pharmacological subclass) in community. We then computed a *P*-value for each community by applying the hypergeometric probability distribution test.

Transcriptional variability score

TV was computed for all the compounds having at least two profiles available in CMAP for the same cell line. The number of such small molecules for each cell line is: 1165 in MCF7, 398 in PC3, 32 in HL60, two in ssMCF7. We took advantage of the large majority of MCF7 experiments to avoid the problematic integration of TV values across different cell types and discarded all non-MCF7 data. About 15% of the CMAP small molecules have more than two profiles in MCF7 cells, producing an average of 16.08 'within-molecule' profile pairs and a maximum of 630 (for tanespimycin). We computed the TV of a small-molecule as follows: given M transcriptional responses to the same small-molecule in the same cell line (i.e., ranked list of differentially expressed genes as in CMAP), we evaluate the transcriptional distances between all the M(M-1)/2 pairs of transcriptional responses and then take their median value as a measure of TV (if M = 2 then TV is defined as the maximum distance). The pairwise transcriptional distance is based on the enrichment of the top (bottom) genes of one profile among the top (bottom genes) of the other profile and vice-versa, as detailed in Iorio et al.[4] Since the TV is based on the same transcriptional distance measure used to derive the transcriptional network in Iorio et al.,[4] we set as a significance threshold for the TV the same threshold used to derive the transcriptional network ($TV_{th} = 0.8$).

Analysis of drug-pairs in quadrant III

To quantify how many drug-pairs in this quadrant have targets that despite being different are found in the same pathway, we considered the subset of the drug-pairs in quadrant III (excluding those ones including large molecules) with known molecular targets according to the EMBL-STITCH database (http://stitch.embl.de) for a total of 8065 drug-pairs (out of a total of 33994 drug-pairs). We then mapped the targets of each drug-pair to three pathway databases (Gene Ontology BP, MF, and CC). We then quantified how many drug-pairs had targets in the same pathway, which yielded 61% of drug-pairs when considering the GO:BP database, 35% of drug-pairs according to GO:MF and 32% of drug-pairs according to GO:CC.

Phospholipidosis stress signature

The PLD stress signature was built by merging together 36 PRLs, corresponding to drugs searched in the literature known to induce PLD,[52] into a single node using the Kruskal Algorithm strategy and the Borda Merging Method implemented the online tool MANTRA (http://mantra.tigem.it) and previously described [3]. Briefly, the algorithm first searches for the two ranked lists with the smallest Spearman's Footrule distance. Then it merges them using the Borda Merging Method, obtaining a new ranked list of genes. The process restarts until only one list remains.

Random forest modeling

We generated a predictive Random Forest model using as features the 128 VolSurf+ physicochemical parameters of the 1309 CMAP compounds and the subset of 258 compounds exhibiting the PLD 'transcriptional signature' for testing and training (Supplementary Table 11). To this end, we used the R (v.3.4) programming environment with the *RandomForest* function (v. 4.6-12). We optimized parameters to build a classification model with 500 number of trees and 11 variables tried at each split, with downsampling to obtain a balanced training set. For more details, please refer to the Supplementary Material.

HCS (high content screening) assays

TFEB nuclear translocation: To quantify TFEB subcellular localization, a high-content assay upon the compound treatments indicated was performed using stable HeLa cells overexpressing TFEB-GFP according to our previous protocols (Medina et al., 2015). *Lysosome, Golgi and Endoplasmic Reticulum assays:* HeLa cells were seeded in a 384-well plate, incubated for 24 h and treated with the different compounds at 0.1, 1 and 10 μM for additional 24 h. After that cells were fixed with 4% paraformaldehyde (for LAMP1 and GM130 stainings) or ice-cold methanol (for PDI staining) and permeabilized/blocked with 0.05% (w/v) saponin, 0.5% (w/v) BSA and 50 mM NH4Cl in PBS (blocking buffer). LAMP-1, GM130 and PDI detection was performed by incubating with the corresponding primary antibodies (anti-LAMP1, Santa Cruz Biotechnology; anti-GM130 and anti-PDI, Cell Signaling Technology) followed by the incubation with an AlexaFluor-conjugated secondary antibodies (Life Technologies) diluted in blocking buffer. LysoTracker Red DND-99 (Life Technologies) staining was performed by the incubating the dye for the last 30 min before fixation. DAPI and CellMask Deep Red Plasma membrane Stain (Life Technologies) were used for nuclei and plasma membrane staining, respectively. Images of of lysosomes (LAMP-1 and LysoTracker Red DND-99), Golgi (GM130) and ER (PDI) were acquired using an automated confocal microscopy (Opera High Content System, Perkin-Elmer). The fluorescent intensity and area of the different stainings were analyzed by using dedicated scripts developed in the Columbus Image Data Management and Analysis Software (Perkin-Elmer).

High Content Lipid accumulation assay: LipidTOX green phospholipidosis detection reagent (Life Technologies) was added to the cells along with the different compounds at the indicated concentrations for 48 h before fixation with 4% paraformaldehyde. DAPI and CellMask Deep Red Plasma membrane Stain (Life Technologies) were used for nuclei and plasma membrane staining, respectively. Lysosomal phospholipid accumulation was analyzed by measuring fluorescent dye intensity using an automated confocal microscopy (Opera High Content System, Perkin-Elmer) and a Columbus Image Data Management and Analysis Software (Perkin-Elmer).

Code availability and supplementary information

Supplementary information and the relative programming code are available without any restrictions on the npj Systems Biology and Application website. All the relevant data are available from the authors and from the website: http://chemantra.tigem.it. The CMAP dataset is available at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE5258. The VolSurf+ physics chemical features are available as Supplementary Table 11

ACKNOWLEDGMENTS

We are grateful to the Bioinformatics Core (TIGEM). This work was supported by a Fondazione Telethon Grant (TGM11SB1) to D.d.B.

AUTHOR CONTRIBUTIONS

D.d.B. conceived the idea and supervised the work. F.S. generated and analyzed the structural drug network. F.S. and F.N. performed computational and statistical data analysis. D.M. and S.P.V. carried out the HCS experiments. D.C. developed the website and database. F.S, D.M. and D.d.B. wrote the manuscript.

ADDITIONAL INFORMATION

Supplementary Information accompanies the paper on the *npj Systems Biology and Applications* website (doi:10.1038/s41540-017-0022-3).

Competing interests: The authors declare that they have no competing financial interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Verbist, B. et al. Using transcriptomics to guide lead optimization in drug discovery projects: Lessons learned from the QSTAR project. *Drug. Discov. Today* **20**, 505–513 (2015).
2. Lamb, J. et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
3. Cheng, J., Yang, L., Kumar, V. & Agarwal, P. Systematic evaluation of connectivity map for disease indications. *Genome Med.* **6**, 95 (2014).
4. Iorio, F. et al. Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl. Acad. Sci. USA* **107**, 14621–14626 (2010).
5. Woo, J. H. et al. Elucidating compound mechanism of action by network perturbation analysis. *Cell* **162**, 441–451 (2015).
6. Kidd, B. A. et al. Mapping the effects of drugs on the immune system. *Nat. Biotechnol.* **34**, 47–54 (2016).
7. Lamb, J. The connectivity map: a new tool for biomedical research. *Nat. Rev. Cancer* **7**, 54–60 (2007).
8. Iorio, F., Rittman, T., Ge, H., Menden, M. & Saez-Rodriguez, J. Transcriptional data: a new gateway to drug repositioning? *Drug. Discov. Today* **18**, 350–357 (2013).
9. Bajorath, J. et al. Navigating structure–activity landscapes. *Drug. Discov. Today* **14**, 698–705 (2009).
10. Geppert, H., Vogt, M. & Bajorath, J. Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inform. Model.* **50**, 205–216 (2010).
11. Heikamp, K. & Bajorath, J. The future of virtual compound screening. *Chem. Biol. Drug Des.* **81**, 33–40 (2013).
12. Shim, J. & Mackerell, A. D. Jr. Computational ligand-based rational design: Role of conformational sampling and force fields in model development. *Medchemcomm* **2**, 356–370 (2011).
13. Sirci, F. et al. Virtual fragment screening: discovery of histamine H3 receptor ligands using ligand-based and protein-based molecular fingerprints. *J. Chem. Inform. Model.* **52**, 3308–3324 (2012).
14. Stumpfe, D. & Bajorath, J. Activity cliff networks for medicinal chemistry. *Drug. Dev. Res.* **75**, 291–298 (2014).
15. Vogt, M. & Bajorath, J. Chemoinformatics: a view of the field and current trends in method development. *Bioorg. Med. Chem.* **20**, 5317–5323 (2012).
16. Backman, T. W., Cao, Y. & Girke, T. ChemMine tools: an online service for analyzing and clustering small molecules. *Nucleic Acids Res.* **39**, W486–W491 (2011).
17. Ma, X. H. et al. Comparative analysis of machine learning methods in ligand-based virtual screening of large compound libraries. *Comb. Chem. High Throughput Screen.* **12**, 344–357 (2009).
18. Ravindranath, A. C. et al. Connecting gene expression data from connectivity map and in silico target predictions for small molecule mechanism-of-action analysis. *Mol. Biosyst.* **11**, 86–96 (2015).
19. Khan, S. A. et al. Identification of structural features in chemicals associated with cancer drug response: a systematic data-driven analysis. *Bioinformatics* **30**, i497–i504 (2014).
20. Iskar, M. et al. Drug-induced regulation of target expression. *PLoS. Comput. Biol.* **6**, 1–8 (2010).
21. Hizukuri, Y., Sawada, R. & Yamanishi, Y. Predicting target proteins for drug candidate compounds based on drug-induced gene expression data in a chemical structure-independent manner. *BMC Med. Genomics* **8**, 82 (2015).
22. Sulli, G., Di Micco, R. & d'Adda di Fagagna, F. Crosstalk between chromatin state and DNA damage response in cellular senescence and cancer. *Nat. Rev. Cancer* **12**, 709–720 (2012).
23. Kirchmair, J. et al. Predicting drug metabolism: experiment and/or computation? *Nat. Rev. Drug Discov.* **14**, 387–404 (2015).
24. Szakacs, G., Paterson, J. K., Ludwig, J. A., Booth-Genthe, C. & Gottesman, M. M. Targeting multidrug resistance in cancer. *Nat. Rev. Drug Discov.* **5**, 219–234 (2006).
25. Carosati, E., Sciabola, S. & Cruciani, G. Hydrogen bonding interactions of covalently bonded fluorine atoms: from crystallographic data to a new angular function in the GRID force field. *J. Med. Chem.* **47**, 5114–5125 (2004).
26. Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **28**, 849–857 (1985).
27. Carrella, D. et al. Mantra 2.0: an online collaborative resource for drug mode of action and repurposing by network analysis. *Bioinformatics* **30**, 1787–1788 (2014).
28. Iorio, F., Isacchi, A., di Bernardo, D. & Brunetti-Pierri, N. Identification of small molecules enhancing autophagic function from drug network analysis. *Autophagy* **6**, 1204–1205 (2010).
29. Cruciani, G., Crivori, P., Carrupt, P. A. & Testa, B. Molecular fields in quantitative structure–permeation relationships: the VolSurf approach. *J. Mol. Struct. THEOCHEM* **503**, 17–30 (2000).
30. Cruciani, G., Pastor, M. & Guba, W. VolSurf: a new tool for the pharmacokinetic optimization of lead compounds. *Eur. J. Pharm. Sci.* **11**, S29–S39 (2000). Supplement 2.
31. Lipinski, C. A. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov. Today Technol.* **1**, 337–341 (2004).
32. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug. Deliv. Rev.* **23**, 3–25 (1997).
33. Baroni, M., Cruciani, G., Sciabola, S., Perruccio, F. & Mason, J. S. A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for ligands and proteins (FLAP): theory and application. *J. Chem. Inform. Model.* **47**, 279–294 (2007).
34. Bodenhofer, U., Kothmeier, A. & Hochreiter, S. APCluster: an R package for affinity propagation clustering. *Bioinformatics* **27**, 2463–2464 (2011).
35. Frey, B. J. & Dueck, D. Clustering by passing messages between data points. *Science* **315**, 972–976 (2007).
36. Napolitano, F., Sirci, F., Carrella, D. & di Bernardo, D. Drug-set enrichment analysis: a novel tool to investigate drug mode of action. *Bioinformatics*. **32**, 235–241 (2015).
37. Davis, W. W. & Garren, L. D. On the mechanism of action of adrenocorticotropic hormone. The inhibitory site of cycloheximide in the pathway of steroid biosynthesis. *J. Biol. Chem.* **243**, 5153–5157 (1968).
38. Matilainen, O., Quiros, P. M. & Auwerx, J. Mitochondria and epigenetics - crosstalk in homeostasis and stress. *Trends Cell Biol.* **27**, 453–463 (2017).
39. Raynal, N. J. et al. Targeting calcium signaling induces epigenetic reactivation of tumor suppressor genes in cancer. *Cancer Res.* **76**, 1494–1505 (2016).
40. Baliga, B. S., Pronczuk, A. W. & Munro, H. N. Mechanism of cycloheximide inhibition of protein synthesis in a cell-free system prepared from rat liver. *J. Biol. Chem.* **244**, 4480–4489 (1969).
41. Jimenez, A., Carrasco, L. & Vazquez, D. Enzymic and nonenzymic translocation by yeast polysomes. Site of action of a number of inhibitors. *Biochemistry* **16**, 4727–4730 (1977).
42. McKeehan, W. & Hardesty, B. The mechanism of cycloheximide inhibition of protein synthesis in rabbit reticulocytes. *Biochem. Biophys. Res. Commun.* **36**, 625–630 (1969).
43. Nadanaciva, S. et al. A high content screening assay for identifying lysosomotropic compounds. *Toxicol. In. Vitr.* **25**, 715–723 (2011).
44. Petersen, Nikolaj H.T. et al. Transformation-Associated Changes in Sphingolipid Metabolism Sensitize Cells to Lysosomal Cell Death Induced by Inhibitors of Acid Sphingomyelinase. *Cancer Cell.* **24**, 379–393 (2013).
45. Ellegaard, A.-M. et al. Repurposing Cationic Amphiphilic Antihistamines for Cancer Treatment. *EBioMedicine*. **9**, 130–139 (2016).
46. Roy, M., Dumaine, R. & Brown, A. M. HERG, a primary human ventricular target of the nonsedating antihistamine terfenadine. *Circulation* **94**, 817–823 (1996).
47. Zhou, Z., Vorperian, V. R., Gong, Q., Zhang, S. & January, C. T. Block of HERG potassium channels by the antihistamine astemizole and its metabolites desmethylastemizole and norastemizole. *J. Cardiovasc. Electrophysiol.* **10**, 836–843 (1999).
48. Morissette, G., Lodge, R. & Marceau, F. Intense pseudotransport of a cationic drug mediated by vacuolar ATPase: procainamide-induced autophagic cell vacuolization. *Toxicol. Appl. Pharmacol.* **228**, 364–377 (2008).
49. Ashoor, R., Yafawi, R., Jessen, B. & Lu, S. The contribution of Lysosomotropism to autophagy perturbation. *PLoS One* **8**, e82481 (2013).

50. Kazmi, F. et al. Lysosomal sequestration (trapping) of lipophilic amine (cationic amphiphilic) drugs in immortalized human hepatocytes (Fa2N-4 Cells). *Drug Metab. Dispos.* **41**, 897–905 (2013).
51. Marceau, F. et al. Cation trapping by cellular acidic compartments: beyond the concept of lysosomotropic drugs. *Toxicol. Appl. Pharmacol.* **259**, 1–12 (2012).
52. Muehlbacher, M., Tripal, P., Roas, F. & Kornhuber, J. Identification of drugs inducing phospholipidosis by novel in vitro data. *ChemMedChem* **7**, 1925–1934 (2012).
53. Halliwell, W. H. Cationic amphiphilic drug-induced phospholipidosis. *Toxicol. Pathol.* **25**, 53–60 (1997).
54. Goracci, L., Ceccarelli, M., Bonelli, D. & Cruciani, G. Modeling phospholipidosis induction: reliability and warnings. *J. Chem. Inform. Model.* **53**, 1436–1446 (2013).
55. Sun, H. et al. Are hERG channel blockers also phospholipidosis inducers? *Bioorg. Med. Chem. Lett.* **23**, 4587–4590 (2013).
56. Anderson, N. & Borlak, J. Drug-induced phospholipidosis. *FEBS Lett.* **580**, 5533–5540 (2006).
57. Lu, S., Sung, T., Lin, N., Abraham, R. T. & Jessen, B. A. Lysosomal adaptation: how cells respond to lysosomotropic compounds. *PLoS One* **12**, e0173771 (2017).
58. Napolitano, G. & Ballabio, A. TFEB at a glance. *J. Cell Sci.* **129**, 2475–2481 (2016).
59. Martina, J. A., Chen, Y., Gucek, M. & Puertollano, R. MTORC1 functions as a transcriptional regulator of autophagy by preventing nuclear transport of TFEB. *Autophagy* **8**, 903–914 (2012).
60. Roczniak-Ferguson, A. et al. The transcription factor TFEB links mTORC1 signaling to transcriptional control of lysosome homeostasis. *Sci. Signal.* **5**, ra42 (2012).
61. Settembre, C. et al. A lysosome-to-nucleus signalling mechanism senses and regulates the lysosome via mTOR and TFEB. *EMBO J.* **31**, 1095–1108 (2012).
62. Sardiello, M. et al. A gene network regulating lysosomal biogenesis and function. *Science* **325**, 473–477 (2009).
63. Settembre, C. et al. TFEB links autophagy to lysosomal biogenesis. *Science* **332**, 1429–1433 (2011).
64. Medina, DiegoL. et al. Transcriptional activation of lysosomal exocytosis promotes cellular clearance. *Dev. Cell.* **21**, 421–430 (2011).
65. Medina, D. L. et al. Lysosomal calcium signalling regulates autophagy through calcineurin and TFEB. *Nat. Cell Biol.* **17**, 288–299 (2015).
66. Carrella, D. et al. Computational drugs repositioning identifies inhibitors of oncogenic PI3K/AKT/P70S6K-dependent pathways among FDA-approved compounds. *Oncotarget*. **7**, 58743–58758 (2016).
67. Jin, Y. et al. Antineoplastic mechanisms of niclosamide in acute myelogenous leukemia stem cells: inactivation of the NF-kappaB pathway and generation of reactive oxygen species. *Cancer Res.* **70**, 2516–2527 (2010).
68. Ishii, I., Harada, Y. & Kasahara, T. Reprofiling a classical anthelmintic, pyrvinium pamoate, as an anti-cancer drug targeting mitochondrial respiration. *Front. Oncol.* **2**, 137 (2012).
69. Fonseca, B. D. et al. Structure-activity analysis of niclosamide reveals potential role for cytoplasmic pH in control of mammalian target of rapamycin complex 1 (mTORC1) signaling. *J. Biol. Chem.* **287**, 17530–17545 (2012).
70. Newman, R. A., Yang, P., Pawlus, A. D. & Block, K. I. Cardiac glycosides as novel cancer therapeutic agents. *Mol. Interv.* **8**, 36–49 (2008).
71. Wang, Y. C., Chen, S. L., Deng, N. Y. & Wang, Y. Network predicting drug's anatomical therapeutic chemical code. *Bioinformatics* **29**, 1317–1324 (2013).
72. Krishnan, A. V., Swami, S. & Feldman, D. Estradiol inhibits glucocorticoid receptor expression and induces glucocorticoid resistance in MCF-7 human breast cancer cells. *J. Steroid Biochem. Mol. Biol.* **77**, 29–37 (2001).
73. Zhang, Y., Leung, D. Y. M., Nordeen, S. K. & Goleva, E. Estrogen inhibits glucocorticoid action via protein phosphatase 5 (PP5)-mediated glucocorticoid receptor dephosphorylation. *J. Biol. Chem.* **284**, 24542–24552 (2009).
74. Carollo, M., Parente, L. & D'Alessandro, N. Dexamethasone-induced cytotoxic activity and drug resistance effects in androgen-independent prostate tumor PC-3 cells are mediated by lipocortin 1. *Oncol. Res.* **10**, 245–254 (1998).
75. Zhang, C. et al. Corticosteroid-induced chemotherapy resistance in urological cancers. *Cancer Biol. Ther.* **5**, 59–64 (2006).
76. Hamid, N. & Krise, J. P. *Lysosomes: Biology, Diseases, and Therapeutics* 423-444 (Wiley, 2016).
77. Liu, J., Lee, J., Hernandez, M. A. S., Mazitschek, R. & Ozcan, U. Treatment of obesity with celastrol. *Cell* **161**, 999–1011 (2015).
78. Chen, B. & Butte, A. J. Leveraging big data to transform target selection and drug discovery. *Clin. Pharmacol. Ther.* **99**, 285–297 (2016).
79. Sirota, M. et al. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.* **3**, 96ra77 (2011).
80. JChem 14.9.15, 2014, ChemAxon (http://www.chemaxon.com)"
81. Milletti, F., Storchi, L., Sforna, G. & Cruciani, G. New and original pKa prediction method using grid molecular interaction fields. *J. Chem. Inform. Model.* **47**, 2172–2181 (2007).
82. Molecular Operating Environment (MOE), 2013.08; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2016.
83. Cross, S., Baroni, M., Carosati, E., Benedetti, P. & Clementi, S. FLAP: GRID molecular interaction fields in virtual screening. validation using the DUD data set. *J. Chem. Inform. Model.* **50**, 1442–1450 (2010).
84. Cross, S. & Cruciani, G. Grid-derived structure-based 3D pharmacophores and their performance compared to docking. *Drug Discov. Today Technol.* **7**, e213–e219 (2010).
85. De Baets, B. & Mesiar, R. Metrics and T-equalities. *J. Math. Anal. Appl.* **267**, 531–547 (2002).
86. Newman, M. E. J. Modularity and community structure in networks. *Proc. Natl Acad. Sci. USA* **103**, 8577–8582 (2006).
87. WHO Collaborating Centre for Drug Statistics Methodology, ATC classification index with DDDs, 2014. Oslo 2014.