

Additional File 1: Supplementary Information

Integrating pharmacogenomics and cheminformatics with diverse disease phenotypes for cell type-guided drug discovery

Arda Halu^{1,2*}, Sarvesh Chelvanambi², Julius L. Decano², Joan T. Matamalas², Mary Whelan², Takaharu Asano², Namitra Kalicharran², Sasha A. Singh², Joseph Loscalzo^{1,3}, Masanori Aikawa^{1,2*}

¹Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School

²Center for Interdisciplinary Cardiovascular Sciences, Division of Cardiovascular Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School

³Department of Medicine, Brigham and Women's Hospital, Harvard Medical School

Supplementary Note 1: Overview of LINCS-CMap data

The building block of LINCS-CMap data is what is called an “instance,” which embodies experimental attributes such as the name and ID of the perturbagen, the concentration (dose) of the perturbagen, the cell line that was treated with the perturbagen, and the time point at which the perturbagen was administered. The same perturbagen may be administered on the same cell line, with the same dose, and at the same time point, several times, in which case they are recorded as separate instances (i.e., technical replicates). Throughout our study, we use this definition of “instance” as the basis of our analyses.

Summary of data Levels 1-5. The LINCS-CMap data consists of five data levels reflecting various data processing and normalization steps. Level 1 corresponds to the raw fluorescent intensity values; Level 2 corresponds to the gene expression values for the landmark genes obtained from the intensity values; Level 3 involves the normalization of Level 2 gene expression values for the landmark genes, as well as the inference of the expression of the >11k genes; not measured directly by the L1000 assay; Level 4 corresponds to the z-scores (denoted by ZS) calculated from Level 3 expression values with respect to the entire (384-well) plate population; Level 5 corresponds to the replicate-consensus of Level 4 expression data, denoted by MODZ. See https://clue.io/connectopedia/data_levels for more details.

Comparison of QUIZC, MODZ, and CD from the LINCS-CMap data perspective. Like MODZ, QUIZ-C also aggregates experimental replicates into a representative, consensus signature. For a brief comparison with respect to the LINCS data, the MODZ-based gene-perturbagen networks are constructed from the MODZ-aggregated Level 5 LINCS data (L1000 Level 5 (perturbagen level) → gene-perturbagen network), whereas our gene-perturbagen networks are based on Level 4 LINCS data, which is fed into QUIZ-C framework to attain a consensus network (L1000 Level 4 (replicate level) → QUIZ-C (aggregated to perturbagen level) → gene-perturbagen network).

Supplementary Note 2: Assessment of QUIZ-C networks against the statistical background

To rule out the possibility to arrive at spurious gene-perturbagen edges in QUIZ-C networks in the event that all perturbagens have non-significant effects on a given gene, which would correspond to a Level 4 ZS distribution that is very narrowly distributed around 0, we provide a

careful assessment of the ZS distributions. We sought to address this point at the macro and micro levels.

At the macro level, we performed a bird's eye survey of ZS values. In particular, we calculated the kernel density estimations of ZS distributions for all cell lines to first qualitatively determine the prevalence of genes with no or few perturbations that significantly affect them (**Figure S4A**). We observed that the vast majority of genes have ZS values between -2.5 and 2.5 with a peak density < 0.5 around zero, while a much smaller yet nontrivial proportion of genes have ZS values that are narrowly distributed around 0 with a peak density well above 0.5, suggesting that for those genes, most, if not all, perturbagens do not have a significant effect (i.e., ZS close to zero).

At the micro level, we looked at the summary statistics of the latter group of genes for which there are few or no significant perturbagens affecting them. In particular, we filtered the results for (a) cases for which the kernel density estimation could not be performed due to the low number of nonzero values, (b) cases for which all values were zeros, and (c) cases for which we have narrow peaks with a density > 0.5 . While (a) and (b) are stricter criteria identifying only those genes that have very few or no nonzero ZS values, we also included (c) to be more exhaustive on our search. Together, we call these genes “potentially spurious set of genes” in the sense that they might be incorporated in the QUIZ-C networks as a result of being deemed significant because the PZ score is the z-score normalized against all other insignificant PZ scores. We first determined the percentage of the genes identified by criteria (a)-(c) in the QUIZ-C networks with the hypothesis that their contribution should be minimal in our networks (**Table S2**). Even though (c) was the most conservative criterion, less than 1% of genes detected this way were present in these networks. For (a), this percentage was even less,

around 0.1%. Next, we compared the in-degrees of the potentially spurious genes with the other genes in our QUIZ-C networks with the hypothesis that the mean in-degree of spurious genes should be less than that of other genes. We observed that for half of the cell lines this was the case (Mann-Whitney U test p-value < 0.05) for both (a) and (c) (**Table S3**). We note that the contribution of (b) was negligible in both approaches. Overall, these results help us conclude that the contribution of these potentially spurious interactions is minimal in our networks.

Finally, we confirmed that even in the cases where the distribution is narrow (i.e., the majority of perturbations are insignificant, criterion (c), the gene-perturbation interactions that are captured in the QUIZ-C networks arise from the few highly significant perturbations (we also elaborate on this point in the context of time points and concentrations). Here, we provide an example in which the narrow peak genes that are present in our networks are indeed there as a result of several highly significant instances (**Figure S4B**). DNAJB6 is one of the genes identified as a result of criterion (c) in the HS27A cell line. We observe that, while the vast majority of perturbation instances have ZS values close to zero, one perturbation (parthenolide) results in significant down-regulation of DNAJB6, resulting in the DNAJB6-parthenolide edge in the HS27A network. In fact, parthenolide is DNAJB6's only connection in the HS27A QUIZ-C network. While it is challenging to demonstrate this exhaustively, our investigation shows that most "potentially spurious" genes identified by criterion (c) have a small number of connections that fit this particular scenario.

Supplementary Note 3: Topological properties of LINCS-CMap gene-perturbation networks

Comparison between gene-perturbation networks in terms of their edge overlap as it relates to their density. While MODZ and CD networks also have little edge overlap between cell types (**Figure S9B-C**), their edge overlap is generally higher compared to QUIZ-C (**Figure S9D**). The lower edge overlap of QUIZ-C networks cannot simply be attributed to a lower network density (i.e., the number of existing edges out of all possible edges), as MODZ networks are the densest of the three sets of networks (**Figure S9E**), whereas CD networks have the highest edge overlap (**Figure S9D**).

Correlation of out-degree with number of replicates. In addition, the out-degree was largely uncorrelated with the number of replicates in QUIZ-C networks (**Table S5**), suggesting that the potential bias caused by the varying number of experimental replicates on the topology of QUIZ-C networks was minimal.

In-degree comparison of landmark vs. inferred genes in LINCS-CMap. The mean in-degrees of landmark genes were higher than those of the inferred genes for 94%, 92%, and 56% of cell lines for QUIZ-C, MODZ, and CD, respectively (**Figure S8**).

The impact of technical variation on gene-perturbation network topology. To demonstrate that any potential correlation across technical replicates does not cause any important biases in the results, especially by affecting the gene-perturbation networks on which downstream analyses and inferences are based, we performed a sensitivity analysis. In particular, one important way that we can envision that technical replicates can impart any bias on the results is through affecting the network topology, i.e., if the network structure is affected by the wide range of replicate numbers across different cell types. To test if the number of replicates of a given perturbation biases the resulting network by inflating the number of targets of that perturbation

(i.e., out-degree), we calculated the correlation between replicate number and out-degree. We performed this analysis for the most data-heavy “core” cell lines that have the largest variation in terms of the number of replicates. Our results show that QUIZ-C generally shows the least bias in this regard (**Table S5**). In MODZ, the out-degree is generally moderately positively correlated with the number of replicates, whereas for QUIZ-C and CD, this correlation appears to be weak, with QUIZ-C being the less correlated of the two in most cases.

Supplementary Note 4: Sensitivity analysis of PACOS against intermediary disease signatures

Here, our goal was to assess (1) whether there are any correlations between the Enrichr disease signatures used in Pathopticon as intermediary phenotypes and (2) whether any such correlations affect the resulting PACOS values. To this end we carried out the following tasks:

- (1) We calculated the variance-covariance matrix of disease signatures as a visual proxy of similarities.
- (2) We compared the distributions of PACOS scores for all cell types combined at various clustering resolutions.
- (3) We compared the distributions of PACOS scores for each cell type separately at various clustering resolutions.

As can be seen in the variance-covariance matrix in (**Figure S17**), there were small groups of diseases that are similar to each other, although this was limited to a few small clusters. To see if this variance-covariance structure affects the distribution of PACOS values at the extreme scenarios, we devised an analysis in which we calculated PACOS values using the disease signatures at various hierarchical clustering cutoffs. This procedure translates to collapsing similar disease signatures to a representative signature based on the given clustering cutoff, with the hypothesis that if our method is not robust to correlations between intermediary

disease signatures, we should observe stark changes in PACOS distributions as we traverse the dendrogram of the variance-covariance matrix. Therefore, for tasks (2) and (3), we plotted PACOS distributions for a range of cophenetic distance values. PACOS values only started to get inflated somewhat after a cophenetic distance of 1.5 (**Figures S18A-C**), at which point the clusters were too much artificially combined (i.e., diseases that are not similar were forced to be included in the same cluster). This was the case for both the overall distributions (**Figure S18A**), as well as the maximal and mean PACOS values for each cell line (**Figures S18B and 18C**). Overall, results suggest that PACOS distributions are not impacted for a large range of clustering resolutions.

Supplementary Note 5: Discussion of drug candidates identified by Pathopticon for vein graft disease

Rosiglitazone, a PPAR-gamma agonist, (ranked 5th in the early vein graft (VG) disease signature) is an FDA-approved anti-diabetic. The VICTORY trial for anti-atherosclerotic effects found no evidence for a statistically significant effect of rosiglitazone on atherosclerosis (1). RS-100329 (ranked 4th in the late vein graft disease signature) is an α 1A-adrenoceptor antagonist. A1-AR antagonists (norepinephrine) were previously investigated for vein grafts (2), however, not much is known on RS-100329 itself. GW-7647, a PPAR-alpha agonist, is the only drug that appears in the top-10 predictions for both early and late VG time-course data (7th in early, 3rd in late) was previously investigated for atherosclerosis and non-alcoholic steatohepatitis (NASH) (3,4). Estriol (1st in late), is a type of estrogen. Its atheroprotective effects were investigated in the early 2000s. The clinical trial, Estrogen and Graft Atherosclerosis Research (EAGAR), showed no CVD benefit with increased risk of CAD in the remaining healthy coronary vessels (5,6). Betamethasone (1st in early) is a corticosteroid. Several studies on the

protective effects of corticosteroids in coronary artery bypass graft surgery were conducted, but mostly on dexamethasone (7,8). Sunitinib (5th in late) is a multitarget RTK inhibitor, and its effect on neointimal hyperplasia in arteriovenous grafts was investigated recently (9).

Nitrendipine (8th in early) is a calcium channel blocker (CCB), Improved radial artery outcomes were reported for CCBs (10,11), with nothing in particular for nitrendipine. AG-490 (3rd in early) and Gefitinib (8th in late) are EGFR tyrosine kinase inhibitors. There are many studies investigating anti-inflammatory properties in nephropathy and liver ischemia/reperfusion.

Improved outcomes were seen in myoblast transplantation. In general, reducing EGFR activity resulted in reduced neointimal hyperplasia (12–15). Pravastatin (9th in late) is an HMG CoA reductase inhibitor, which, like other statins, is prescribed to post-operative CABG patients (16). Recent studies showed protective effects of dephostatin (2nd in late), a protein tyrosine phosphatase 1B inhibitor. against atherosclerotic plaque formation (17), with a possible role in peripheral artery disease (PAD) as well (18). SB-203580 (10th in late), a p38 kinase inhibitor, was Investigated as a potential treatment for VG failure in the early 2000s (19).

While the majority of the top-10 predictions appear to be relevant and with potential for further investigation in the context of vein graft disease, there a few drugs seem to have the opposite effect to what would be expected in the context of vein graft and atherosclerosis, e.g., inhibitors and antagonists when agonists are expected. Splitomycin (2nd in early) is a sirtuin (SIRT1) inhibitor. Sirtuins are believed to be protective against atherosclerosis (20). Calcitriol (9th in early) is the active vitamin D metabolite. Vitamin D is believed to be a promoter of vascular calcification (21). SCH-58261 (10th in early) is an adenosine receptor A2A antagonist. A2AAR is protective against inflammation and appears to have anti-atherosclerotic effects (22–24). The remaining four drugs returned no results relevant to vein graft diseases or atherosclerosis.

Most of these drugs are approved for other uses: Loperamide (6th in early), an opioid-receptor agonist, is used as an anti-diarrheal; Formoterol (4th in early), a long-acting beta

agonist, is used as a bronchodilator; Sulpiride (6th in late), dopamine/serotonin receptor antagonist, is used as an anti-psychotic; and Paroxetine (7th in late), a selective serotonin reuptake inhibitor, is used as an anti-depressant.

Supplementary Note 6: Adipose stromal cells and macrophages

Adipose stromal cells (ASCs) have been reported to work in close tandem with macrophages in times of pro-inflammatory activation (25). ASCs can mediate proinflammatory M1-like activation of macrophages (25), while macrophages can inhibit adipogenic differentiation of these cells by producing pro-inflammatory cytokines (26). As a result, adipose stromal cells can differentiate into fibro-inflammatory phenotype when adipogenic differentiation is suppressed (27). Crosstalk between macrophages and ASCs is found in many tissue microenvironments but is particularly important in obesity, a major CAD risk factor. Shared metaflammation pathways between ASCs and macrophages, due to many layers of evolutionarily conserved interactions occurring between immune response and metabolism, may provide key targets for therapeutic intervention in vascular inflammation (28). Hence, a systems approach targeting this relationship may reveal drugs that can access these shared pathways.

Supplementary Methods

Processing the LINCS-CMap Level 4 data

The LINCS-CMap Level 4 expression data were downloaded from the Touchstone Reference Dataset Supporting Files on the CLUE platform (clue.io) (version date 5/11/2018). This dataset, named “GSE92742_Broad_LINCS_Level4_ZSPCINF_mlr12k_n1319138x12328.gctx.gz”

encompasses 1,319,138 perturbation instances (including experimental replicates), 473,647 unique perturbation signatures, 51,383 perturbations, 12,328 genes, and 76 cell lines. The .gctx data matrix file was parsed using the “parse” function (“cmapPy.pandasGEXpress.parse”) within the pandasGEXpress module of the CMapPy tool (29) in Python. The small molecule perturbations (chemical compounds) denoted by “trt_cp” were retained in the data, and the gene (“rid”)-by-perturbation instance (“inst_id”) data tables were written individually to file in the .feather format for each cell type. Out of the 76 cell types in LINCS-CMap Level 4 data, five had no small molecule compounds tested on them, but rather had control vectors or shRNAs used as perturbations, and one (MCH58) yielded no significant edges. Overall, this procedure resulted in 70 cell lines with which to carry out the downstream analyses.

Sensitivity analysis of the QUIZ-C approach

To assess the response of cell type-specific gene-perturbation network topologies to changes in QUIZ-C parameters, we performed a sensitivity analysis that considers multiple models and consensus thresholds. We considered three models: (i) The core QUIZ-C approach that relies on the instance distribution described in detail in the Main methods. For this model, we considered the consensus thresholds of 70th, 75th, 80th and 85th quantiles; (ii) Model with an additional “fixed” threshold that divides drugs into two regimes: If the number of instances is over threshold, it calculates quantiles, else if the number of instances is below threshold, a fixed number of significant replicates (e.g., 2) is needed for an edge. For this model, we considered the consensus thresholds of 75th and 80th quantiles, instance thresholds of 3 and 4, and for two significant replicates; (iii) Model with an additional “percentage” threshold that divides it into two regimes: If the number of instances is over threshold, it calculates quantiles, if the number of instances below threshold, a percentage of significant replicates (e.g. 66%) is

needed for an edge. For this model, we considered the consensus thresholds of 75th and 80th quantiles, instance thresholds of between 3 and 10, and for 66% significant replicates (**Figure S2**). We found that the second and third models corresponding to fixed and percent thresholds proved limiting for cell lines with fewer instances per drug, and that the quantile cutoff did not have a high impact on the network topologies across different regimes. More than two-thirds of the cell type-specific networks built by QUIZ-C with a quantile cutoff between 70% and 85% had over 75% edge overlap with each other. Based on these findings, we chose the simple model (model i) with a quantile threshold of 75%.

Mappings between perturbation identifiers

We used PubChem Compound IDs (CIDs) as the unifying identifier in all analyses to interface various pharmacological and cheminformatics databases. To map the LINCS-CMap perturbation names (“pert_inames”) to their corresponding CIDs, we first extracted the InChI and SMILES keys of all compounds (“trt_cp”) in the LINCS-CMap perturbation dataset (“GSE92742_Broad_LINCS_pert_info.txt”). We then used the InChI and SMILES keys as inputs in the PubChem Identifier Exchange service (<https://pubchem.ncbi.nlm.nih.gov/idexchange/idexchange.cgi>) to retrieve the corresponding CIDs. We merged the results from InChI keys and SMILES to maximize mapping coverage in cases where information about either key was missing. We then subset this global mapping to only include the drugs in our QUIZ-C networks for downstream analyses.

Construction of MODZ- and CD-based networks for benchmarking

To benchmark QUIZ-C as a method for gene-perturbation network building, we used two state-of-the-art methods which can determine significant gene-perturbation interactions from the LINCS-CMap data in a similar manner, and can therefore be used to build cell type-specific networks such as ours. The first one of these methods, namely the moderated z-score (MODZ), is the method used in the original LINCS-CMap study (30) and forms the basis of the LINCS-CMap Level 5 data. To build MODZ-based drug-gene networks, we used the differentially expressed genes identified through the MODZ procedure in the LINCS Canvas Browser (31), which we obtained through the Harmonizome database (32) (<https://maayanlab.cloud/Harmonizome/dataset/LINCS+L1000+CMap+Signatures+of+Differentially+Expressed+Genes+for+Small+Molecules>, retrieved in October 2019). The second method, namely, the characteristic direction (CD) (33), has recently been used to identify differentially expressed genes under perturbations using the LINCS-CMap Level 3 data (34). We downloaded the database of differentially expressed gene signatures calculated using the CD method from the L1000CDS² website (https://maayanlab.cloud/public/L1000CDS_download/, retrieved in October, 2019) and extracted the differentially expressed gene signatures for each cell type from the L1000CDS² database using MongoDB. After obtaining the list of up- and down-regulated genes for each perturbation in each cell line for MODZ and CD, we created edge lists for both methods to be used in the construction of the cell type-specific drug-gene networks. Since both MODZ and CD measure differential expression and yield lists of up- and down-genes, they help us build unweighted networks with positive or negative edges similar to our QUIZ-C networks, thereby offering a direct means to benchmark our method and compare the resulting gene-perturbation networks.

Structural reducibility analysis

To determine the redundancy in the information content of the LINCS-CMap-based multilayer networks, where each network layer represents a cell type, we used the recently proposed structural reducibility concept (35). An information theoretic measure, structural reducibility χ quantifies how distinguishable, or non-redundant, each layer of the multilayer network is from a purely topological perspective. We used the Python implementation of the structural reducibility algorithm provided in the public GitHub repository by the authors of the study (<https://github.com/KatolaZ/multired>). To be compatible with the input format requirements of this Python script, we converted all the gene symbols and pert_inames to integer values. We restricted our attention to the 61 layers (61 cell types) that are common to QUIZ-C, MODZ and CD. We followed the simplified analysis workflow via the python script “run_multired_simple.py” and used the resulting Jensen-Shannon divergence matrices and relative entropy quality function $q(\cdot)$ vectors in our in-house scripts to generate hierarchical clusterings of cell types for each network-building method. We extracted the branching point corresponding to the maximum value of $q(\cdot)$ as the optimal configuration of the aggregation of layers (cell types). Using this clustering configuration (denoted by the vertical dashed lines), we calculated the structural reducibility χ as the ratio of the number of layer aggregation steps (i.e. branching points) to the number of total possible aggregation steps (i.e. the total number of layers minus one)

$$\chi = \frac{M - M_{opt}}{M - 1},$$

where M_{opt} is the number of branching points at the optimal clustering configuration and $M = 51$ is the total number of layers. A maximum at a low value of $q(\cdot)$ represents a highly distinguishable (i.e., not redundant or reducible), whereas a maximum at a high value of $q(\cdot)$

represents a highly redundant (i.e., reducible) multilayer network structure. In our context, this means that the cell type-specific networks are distinct in terms of their information content and cannot be merged without loss of substantial information.

Single Tandem Repeat (STR) profile similarity

We first used the cell line identifier and corresponding synonyms information in the Cellosaurus database (36) (Version 34.0, downloaded in March 2020 from <https://www.cellosaurus.org/>) to map the LINCS-CMap cell line names to Cellosaurus identifiers. Second, we queried Cellosaurus for the following 17 STR markers (excluding Amelogenin): CSF1PO, D2S1338, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S539, D18S51, D19S433, D21S11, FGA, Penta D, Penta E, TH01, TPOX and vWA. Of the 60 LINCS-CMap cell lines we used as the layers in our QUIZ-C gene-perturbation networks, 54 had STR profile information on Cellosaurus. We then used the Cell Line Authentication using STR (CLASTR) tool (37) (<https://www.cellosaurus.org/str-search/>) to calculate the STR similarity of these 54 cell lines using the default settings (“Algorithm” set to Tanabe and “Mode” set to non-empty markers). Since CLASTR calculates the STR profile similarity of a given cell line to all the cell lines in Cellosaurus with STR profile information, rather than calculating the pairwise STR similarity of two input cell lines, we obtained results for each cell line without a score filter and for the maximum number of results (typically around 1,000 top-ranked cell). This resulted in a list of cell lines, ranked from the most similar to the least similar, for each cell line. Since Cellosaurus has STR profile information on >6,000 cell lines and our similarity results return at most around 1,000 cell lines, some of the pairwise similarities between the 54 cell lines remained below “detection threshold,” meaning the lowest similarity value in the available CLASTR output. We therefore quantified (i) the similarity scores for the cell lines above the detection threshold (i.e.

within the top ~1,000 cell lines), (ii) the similarity score corresponding to the detection threshold and the percentage of cell lines below this detection threshold (i.e. with lower similarity values than the threshold).

Cell Line Ontology (CLO) semantic similarity

To measure the ontological similarity of cell lines via the Cell Line Ontology (CLO) (38) (downloaded from <https://bioportal.bioontology.org/ontologies/CLO> in January 2020), we used a graph-based semantic similarity approach for biological ontology terms (39). We converted LINCS-CMap cell line identifiers to CLO identifiers via manual curation. We constructed the directed acyclic graph (DAG) from all the terms (cell lines) in CLO using their parent and child terms. We set the semantic contribution factor of parent-child relationship to 0.5 and calculated the S-value $S(A)$, semantic value $SV(A)$, and pairwise semantic similarity $S_{CLO}(A, B)$, as described in (39), for all cell lines in L1000.

Parsing the ChEMBL database for bioactivity records

In order to assess the selectivity and potency of the compounds in our gene-perturbation networks, we leveraged the large-scale bioactivity database ChEMBL. We downloaded all available bioactivity information from ChEMBL (Release 27, downloaded in June, 2020) and focused on the bioactivity information for compounds that target single proteins in humans by choosing “Homo Sapiens” as the Target Organism and “Single Protein” as the Target Type. These filtering steps resulted in 5,190,464 records out of a total of 16,066,124 bioactivity records in ChEMBL. After removing entries with missing Compound Keys, we converted the molecule and target ChEMBL IDs to PubChem CIDs and UniProt IDs, respectively. Molecule

ChEMBL IDs were used as input on the PubChem Identifier Exchange service to retrieve the corresponding PubChem CIDs. Target ChEMBL IDs were converted to UniProt IDs by using the mappings from ChEMBL (“chembl_uniprot_mapping.txt,” downloaded from <ftp://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/latest/>). The UniProt IDs were then converted to Entrez IDs and approved Gene Symbols using the HUGO Gene Nomenclature Committee website (<https://www.genenames.org/>, downloaded in May 2020). Overall, these ID conversions resulted in 5,107,962 activities between 1,035,402 compounds (in PubChem CIDs) and 3,448 targets (in Gene Symbols/Entrez IDs). Restricting these records to the compounds in the QUIZ-C networks resulted in 326,134 activities between 4,311 QUIZ-C compounds and 1,817 targets.

Calculating selectivity and tool scores

To calculate selectivity and tool scores, we used bioactivity data from the ChEMBL database (<https://www.ebi.ac.uk/chembl/>). We first filtered the ChEMBL bioactivity data to include only the records in which bioactivity was measured in exact nanomolar (nM) units (by setting the “Standard Unit” parameter to “nM” and the “Standard Relation” parameter to “=”). We then employed a data-cleaning step wherein we removed erroneous records that displayed negative nM values. These steps enabled us to identify, for each drug (encoded by a CID number), its targets with known binding affinities. To assess the binding selectivity of a drug to a given target, we defined two sets of targets: the “on-target” as the given target, and “off-targets” as all other targets of the drug. For example, if drug A targets genes W, X, Y and Z (i.e., has bioactivity data available in ChEMBL for these genes), when we want to calculate the binding selectivity between drug A and gene X, X would be the “on-target” and W, Y and Z would be the off-targets. Using the on-target and off-target binding affinities, we followed the approaches

taken in two recent studies (40,41) and defined selectivity as the combination of three factors:

(i) magnitude of selectivity; (ii) significance of selectivity, and (iii) data bias.

First, as a measure of the magnitude of selectivity, we calculated the “potency difference” as the difference between the lower (1st) quartiles of the $\log_{10}(\text{nM})$ activity distributions of the off- and on-targets, as originally described in (41)

$$Q1_{diff} = Q1(\log_{10} nM_{off-target}) - Q1(\log_{10} nM_{on-target}).$$

Second, we calculated the Kolmogorov-Smirnov (KS) test p-value between the $\log_{10}(\text{nM})$ distributions between off- and on-targets. We introduced a lower threshold of KS p-value at $1e-16$ and then calculated the negative logarithm of the KS p-values to serve as a measure of the significance of selectivity

$$selectivity\ significance = \frac{-\log_{10} KS\ pvalue}{16}$$

Third, we calculated the data bias as the ratio of the on-target measurements to all measurements for a given drug-target pair

$$data\ bias = \frac{N_{on-target\ activities}}{N_{on-target\ activities} + N_{off-target\ activities}}.$$

Given this definition, if there are no off-target measurements, the measurement is severely biased, therefore the data bias is 100%. The higher the number of off-target measurements, the less biased the measurement becomes (i.e. data bias tends to zero). In the Selectivity and Tool Score calculations, we use $(1 - \text{data bias})$ to favor drugs with low bias.

Bringing the three components together, we define the selectivity score as defined in (40)

$$selectivity\ score = \frac{Q1_{diff}/3 + selectivity\ significance + (1 - data\ bias)}{3}.$$

We calculated overall tool scores for each drug-target pair as the product of the selectivity score and strength

$$\text{tool score} = \text{selectivity score} \times \text{strength}$$

where, if both on- and off-targets exist of a given drug exist, we divided binding affinities into three regimes (low affinity: >1000nM; medium affinity: >100 nM and <=1000nM; high affinity: <=100nM) and assigned strength values 7, 4 and 1 if there are more than one high-affinity measurements, if there are more than four medium-affinity measurements, or if neither of these two conditions are met, respectively, following the convention in (40).

We then removed outliers from selectivity score distribution by removing points outside of the range $[Q1-4*IQR, Q3+4*IQR]$. Since tool scores were skewed by the assigned strength values, we removed the outliers in tool Scores manually by excluding tool Scores above 7.5. We finally scaled the selectivity and tool scores to the range $[-1, 1]$ to be compatible with the pathophenotypic congruity scores. The tool and selectivity scores are provided in **Additional File 2**.

Calculating chemical structure similarity

We used the cheminformatics toolkit ChemmineR (42) and its annotation package ChemmineDrugs to calculate the chemical structure similarity between drugs. We first downloaded the compound information in the form of structure-data files (SDFs) for QUIZ-C drugs using their PubChem CIDs as inputs. Using the SDFs, we computed the atom pair (AP) descriptors for each compound, which were subsequently used to generate their respective binary atom pair fingerprints (APFPs). The chemical structure similarity between compound pairs were then calculated by comparing the binary APFPs, using the Tanimoto coefficient.

These calculations were performed on the O2 High Performance Compute Cluster, supported by the Research Computing Group, at Harvard Medical School. See <https://it.hms.harvard.edu/our-services/research-computing> for more information.

Benchmarking datasets

We used the curated gene sets from the Molecular Signatures Database (MSigDB) (43) as input gene sets for benchmarking. In particular, we used the Chemical and Genetic Perturbations (CGP) data set under the Curated Gene Sets (C2) collection (MSigDB ver 7.1., downloaded June 2020), which mainly consists of gene sets separated into those induced (“up”) and repressed (“down”) by the perturbation. Out of the 3,297 gene sets in CGP, 2,116 were tagged as with either “UP” or “DN,” resulting in 1,058 unique gene sets. Thus, each of these gene sets curated from publications included both “up” and “down” genes for the same study.

We next determined the set of drugs targeting each input gene set from MSigDB using the Comparative Toxicogenomics Database (CTD) (<https://ctdbase.org/downloads/>).

We used the chemical-gene interactions file (“CTD_chem_gene_ixns.tsv”; downloaded June 2023) and filtered the interactions to include only the ones that document increase or decrease of mRNA expression in humans. We then retrieved the PubChem CIDs, Canonical SMILES, and InChI keys of the drugs in these filtered interactions using the PUG-REST API using CTD as the source depositor. To connect the CTD the chemical-gene interaction data to the LINCS-CMap compound identifier that we use (i.e., “pert_iname”), we (i) match the LINCS-CMap perturbation dataset (“GSE92742_Broad_LINCS_pert_info.txt”) with the CTD compounds using the InChI keys obtained in the previous step to obtain a CTD compound ID to pert_iname

mapping, and (ii) match the filtered interaction data with the CTD ID-to-pert_iname mapping to obtain CTD chemical-gene interaction data in terms of LINCS pert_iname identifiers.

To identify the “true positive” compounds for each gene set to be used in our benchmarks, we adopted a directional enrichment approach that quantifies the degree of overlap (and its statistical significance) between the up- and down-regulated genes in the gene set and the genes up- and down-regulated by the compound. For each benchmark gene set, we performed a Fisher’s exact test between (i) the up (down) genes of each drug in the CTD chemical-gene interaction data and the up (down) genes of the gene set in the enhance mode and (ii), the up (down) genes of each drug in the CTD chemical-gene interaction data and the down (up) genes of the gene set in the repress mode, and corrected these Fisher’s exact p-values for multiple testing using the Benjamini-Hochberg approach. Compounds with a false discovery rate (FDR) less than 0.05 were then used as “true positives” for the AUROC calculations for the cell lines. The procedure we describe here for the benchmark gene sets is also the approach Pathopticon performs for user-defined inputs.

To ensure that we have sufficient true positives in the performance assessment, we used input gene sets that had at least 100 targeting drugs in the QUIZ-C-, MODZ-, and CD-based networks. We removed the two genesets (“DIAZ_CHRONIC_MEYLOGENOUS_LEUKEMIA” and “DODD_NASOPHARYNGEAL_CARCINOMA”) that resulted in excessively high signature congruity score (SCS) values ($SCS > |0.5|$), which turned out to be also included in the pathophenotype set. Thus, the final benchmark set had 73 input signatures (**Table S1**).

As independent datasets to quantify the enrichment of known drug targets in QUIZ-C-, MODZ-, and CD-based networks, we used the Drug Repurposing Hub (44)

(<https://www.broadinstitute.org/drug-repurposing-hub>, version date 2018-09-07) and Drug-Gene Interaction Database (DGIdb) (45) (<https://www.dgldb.org/>, accessed April 2019) . We also used the Drug Repurposing Hub to annotate the mechanisms of action (MoA) and the clinical development stage of drugs in our gene-perturbagen networks. To calculate the enrichment of drug-target interactions obtained from these databases in our LINCS-CMap-based networks, we created contingency tables of edges (QUIZC/MOZ/CD +/- and DGIdb/DRH +/-) for each cell line and calculated Fisher's exact test p-values and odds ratios (OR) using these contingency tables. We considered cell lines with a Fisher's exact p-value < 0.05 and OR > 1 as significantly enriched in known drug-target interactions.

To assess the sensitivity of Pathopticon to different intermediary genesets (**Supplementary Note 4, Figure S3C**), we also used the up and down gene set libraries for (i) differentially expressed genes for kinase perturbations

(<https://maayanlab.cloud/Harmonizome/dataset/GEO+Signatures+of+Differentially+Expressed+Genes+for+Kinase+Perturbations>); (ii) differentially expressed genes for transcription factor perturbations

(<https://maayanlab.cloud/Harmonizome/dataset/GEO+Signatures+of+Differentially+Expressed+Genes+for+Transcription+Factor+Perturbations>), and (iii) GTEx tissue-specific gene expression (<https://maayanlab.cloud/Harmonizome/resource/Genotype+Tissue+Expression>) datasets from the Harmonizome website (32).

These intermediate gene sets were used to measure the predictive contribution of the scenario wherein we look for the kinases, transcription factors and tissues that mediate the effect of enhancing or repressing perturbagens on the given input signature.

Extrinsic benchmarks: Querying the L1000CDS² database

We performed automated queries to obtain the top candidates predicted by L1000CDS² for all benchmark signatures through the L1000CDS² web API (<https://maayanlab.cloud/L1000CDS2/help/#api>). We used the following configuration parameters: "aggravate": (False for Reverse; True for Forward), "searchMethod": "geneSet", "share": False, "combination": False, "db-version": "latest". We note that since the L1000CDS² server only provides the top 50 results, our comparisons in the benchmark were limited to the top 50 results of each method.

Extrinsic benchmarks: Implementation of the Multiscale Interactome and DeepDRK

As recent interactome- and deep learning-based drug repurposing approaches that are relevant to our method, we performed extrinsic benchmarks using the Multiscale Interactome (46) and DeepDRK (47). To facilitate the harmonization of our data with the disease identifiers in the Multiscale Interactome database and the publicly available gene expression datasets downloaded for use with DeepDRK (details below), we retrieved an orthogonal set of disease gene sets from Harmonizome (<https://maayanlab.cloud/Harmonizome/dataset/GEO+Signatures+of+Differentially+Expressed+Genes+for+Diseases>) to use as benchmark gene sets for these methods.

Multiscale Interactome (MI): We loaded all saved diffusion profiles from the MI database (<http://snap.stanford.edu/multiscale-interactome/data/data.tar.gz>) and, following the procedure outlined in the original study, calculated five similarity measures (L1 distance, L2 distance, Canberra distance, cosine similarity and correlation distance) between each benchmark disease and drug available in the MI database. Since MI uses Concept IDs as disease

identifiers, we performed automated PubMed searches using the rentrez R package (48). In particular, we searched the NCBI Medical Genetics (“medgen”) database given disease names via the “entrez_search” function and retrieved the Concept IDs of the benchmark gene sets from the resulting summary records. We took the intersection of these concept IDs with the concept IDs available in the MI database to form the list of benchmark gene sets to be used in the MI benchmark. For each MI benchmark gene set, we then ranked the drugs according to the above similarity measures. Based on these rankings, and using the drug indications in the MI database, we generated ROC curves for each MI benchmark dataset and calculated their respective area under the ROC curve (AUROC) values.

DeepDRK: Even though DeepDRK can incorporate multiple layers of data, we limit our use case to two layers (gene expression and chemical information) in our DeepDRK benchmark for a fair comparison with our method. For the chemical molecular descriptors to be used as features in the chemical information layer, we used ChemmineR (42) to parse the structure-data files (SDFs) we had retrieved for our chemical structure similarity analysis (see the Supplementary Methods section “Calculating chemical structure similarity”). In particular, we used the following chemical molecular descriptors as features:

"PUBCHEM_CACTVS_COMPLEXITY", "PUBCHEM_CACTVS_HBOND_ACCEPTOR",
"PUBCHEM_CACTVS_HBOND_DONOR", "PUBCHEM_CACTVS_ROTATABLE_BOND",
"PUBCHEM_XLOGP3_AA", "PUBCHEM_EXACT_MASS",
"PUBCHEM_MOLECULAR_WEIGHT", "PUBCHEM_CACTVS_TPSA",
"PUBCHEM_MONOISOTOPIC_WEIGHT", "PUBCHEM_TOTAL_CHARGE",
"PUBCHEM_HEAVY_ATOM_COUNT", "PUBCHEM_ATOM_DEF_STEREO_COUNT",
"PUBCHEM_ATOM_UDEF_STEREO_COUNT",
"PUBCHEM_BOND_DEF_STEREO_COUNT",

"PUBCHEM_BOND_UDEF_STEREO_COUNT", "PUBCHEM_ISOTOPIC_ATOM_COUNT", "PUBCHEM_COMPONENT_COUNT", "PUBCHEM_CACTVS_TAUTO_COUNT", "PUBCHEM_XLOGP3". For the gene expression layer, we log2-transformed the case expression data (i.e., we excluded controls and used only samples with disease) downloaded from the Gene Expression Omnibus (GEO) for a subset of the MI benchmark diseases. For each DeepDRK benchmark gene set, we ranked drugs according to their DeepDRK scores separately within each sample and calculated AUROCs for each sample using true positives calculated as described previously in the "Benchmarking datasets" section. Adopting a similar strategy to the maximum AUROC approach we used in our intrinsic benchmarks, we used the best AUROC value among case samples for each benchmark gene set, which allowed for a fair comparison with Pathopticon.

References

1. Bertrand OF, Poirier P, Rodés-Cabau J, Rinfret S, Title LM, Dzavik V, et al. Cardiometabolic effects of rosiglitazone in patients with type 2 diabetes and coronary artery bypass grafts: A randomized placebo-controlled clinical trial. *Atherosclerosis* [Internet]. 2010 Aug [cited 2022 Nov 17];211(2):565–73. Available from: <https://pubmed.ncbi.nlm.nih.gov/20594555/>
2. Erami C, Zhang H, Ho JG, French DM, Faber JE. α 1-adrenoceptor stimulation directly induces growth of vascular wall in vivo. *Am J Physiol - Hear Circ Physiol* [Internet]. 2002 [cited 2022 Nov 17];283(4 52-4):1577–87. Available from: <https://journals.physiology.org/doi/10.1152/ajpheart.00218.2002>
3. Nakaya K, Tohyama J, Naik SU, Tanigawa H, MacPhee C, Billheimer JT, et al. Peroxisome proliferator-activated receptor- α activation promotes macrophage reverse cholesterol transport through a liver X receptor-dependent pathway. *Arterioscler Thromb*

Vasc Biol [Internet]. 2011 Jun [cited 2022 Nov 17];31(6):1276–82. Available from:
<https://pubmed.ncbi.nlm.nih.gov/21441141/>

4. Okishio S, Yamaguchi K, Ishiba H, Tochiki N, Yano K, Takahashi A, et al. PPAR α agonist and metformin co-treatment ameliorates NASH in mice induced by a choline-deficient, amino acid-defined diet with 45% fat. Sci Reports 2020 101 [Internet]. 2020 Nov 11 [cited 2022 Nov 17];10(1):1–11. Available from: <https://www.nature.com/articles/s41598-020-75805-z>
5. Ali ES, Mangold C, Peiris AN. Estriol: emerging clinical benefits. Menopause [Internet]. 2017 Sep 1 [cited 2022 Nov 17];24(9):1081–5. Available from:
<https://pubmed.ncbi.nlm.nih.gov/28375935/>
6. Ouyang P, Tardif JC, Herrington DM, Stewart KJ, Thompson PD, Walsh MN, et al. Randomized trial of hormone therapy in women after coronary bypass surgery. Evidence of differential effect of hormone therapy on angiographic progression of disease in saphenous vein grafts and native coronary arteries. Atherosclerosis [Internet]. 2006 Dec [cited 2022 Nov 17];189(2):375–86. Available from:
<https://pubmed.ncbi.nlm.nih.gov/16442114/>
7. Schepers A, Pires NMM, Eefting D, de Vries MR, van Bockel JH, Quax PHA. Short-term dexamethasone treatment inhibits vein graft thickening in hypercholesterolemic ApoE3Leiden transgenic mice. J Vasc Surg. 2006 Apr 1;43(4):809–15.
8. Donneyong MM, Kulik A, Gagne JJ. Trends and Patterns of Corticosteroid Use During Coronary Artery Bypass Grafting Surgery in the United States. J Cardiovasc Pharmacol Ther [Internet]. 2018 May 1 [cited 2022 Nov 17];23(3):226–36. Available from:
<https://pubmed.ncbi.nlm.nih.gov/29258391/>
9. Kwon SH, Li L, He Y, Tey JCS, Li H, Zhuplatov I, et al. Prevention of Venous Neointimal Hyperplasia by a Multitarget Receptor Tyrosine Kinase Inhibitor. J Vasc Res [Internet].

2015 Mar 1 [cited 2022 Nov 17];52(4):244–56. Available from:

<https://pubmed.ncbi.nlm.nih.gov/26788996/>

10. Hall AB, Brilakis ES. Saphenous vein graft failure: seeing the bigger picture. *J Thorac Dis* [Internet]. 2019 [cited 2022 Nov 17];11(Suppl 9):S1441. Available from: </pmc/articles/PMC6560559/>
11. Gaudino M, Benedetto U, Fremes S, Hare DL, Hayward P, Moat N, et al. Effect of Calcium-Channel Blocker Therapy on Radial Artery Grafts After Coronary Bypass Surgery. *J Am Coll Cardiol* [Internet]. 2019 May 14 [cited 2022 Nov 17];73(18):2299–306. Available from: <https://pubmed.ncbi.nlm.nih.gov/31072574/>
12. Gérard C, Dufour C, Goudenege S, Skuk D, Tremblay JP. AG490 improves the survival of human myoblasts in vitro and in vivo. *Cell Transplant* [Internet]. 2012 [cited 2022 Nov 17];21(12):2665–76. Available from: <https://pubmed.ncbi.nlm.nih.gov/22963730/>
13. Freitas MCS, Uchida Y, Zhao D, Ke B, Busuttil RW, Kupiec-Weglinski JW. Blockade of Janus kinase-2 signaling ameliorates mouse liver damage due to ischemia and reperfusion. *Liver Transpl* [Internet]. 2010 May [cited 2022 Nov 17];16(5):600–10. Available from: <https://pubmed.ncbi.nlm.nih.gov/20440769/>
14. Chan AK, Kalmes A, Hawkins S, Daum G, Clowes AW. Blockade of the epidermal growth factor receptor decreases intimal hyperplasia in balloon-injured rat carotid artery. *J Vasc Surg* [Internet]. 2003 Mar 1 [cited 2022 Nov 17];37(3):644–9. Available from: <http://www.jvascsurg.org/article/S0741521402752336/fulltext>
15. Trieu VN, Narla RK, Myers DE, Uckun FM. EGF-genistein inhibits neointimal hyperplasia after vascular injury in an experimental restenosis model. *J Cardiovasc Pharmacol* [Internet]. 2000 [cited 2022 Nov 17];35(4):595–605. Available from: <https://pubmed.ncbi.nlm.nih.gov/10774791/>
16. Makuuchi H, Furuse A, Endo M, Nakamura H, Daida H, Watanabe M, et al. Effect of

pravastatin on progression of coronary atherosclerosis in patients after coronary artery bypass surgery. *Circ J* [Internet]. 2005 Jun [cited 2022 Nov 17];69(6):636–43. Available from: <https://pubmed.ncbi.nlm.nih.gov/15914938/>

17. Thompson D, Morrice N, Grant L, Le Sommer S, Lees EK, Mody N, et al. Pharmacological inhibition of protein tyrosine phosphatase 1B protects against atherosclerotic plaque formation in the LDLR^{-/-} mouse model of atherosclerosis. *Clin Sci (Lond)* [Internet]. 2017 Oct 10 [cited 2022 Nov 17];131(20):2489. Available from: </pmc/articles/PMC6365594/>
18. Mercier C, Rousseau M, Geraldès P. Growth Factor Deregulation and Emerging Role of Phosphatases in Diabetic Peripheral Artery Disease. *Front Cardiovasc Med*. 2021 Jan 7;7:380.
19. Cornelissen J, Armstrong J, Holt CM. Mechanical Stretch Induces Phosphorylation of p38-MAPK and Apoptosis in Human Saphenous Vein. *Arterioscler Thromb Vasc Biol* [Internet]. 2004 Mar 1 [cited 2022 Nov 17];24(3):451–6. Available from: <https://www.ahajournals.org/doi/abs/10.1161/01.ATV.0000116690.17017.8b>
20. Grootaert MOJ, Bennett MR. Sirtuins in atherosclerosis: guardians of healthspan and therapeutic targets. *Nat Rev Cardiol* 2022 1910 [Internet]. 2022 Mar 30 [cited 2022 Nov 28];19(10):668–83. Available from: <https://www.nature.com/articles/s41569-022-00685-x>
21. Razzaque MS. The dualistic role of vitamin D in vascular calcifications. *Kidney Int* [Internet]. 2011 [cited 2022 Nov 28];79(7):708. Available from: </pmc/articles/PMC3120050/>
22. Johnston-Cox HA, Koupenova M, Ravid K. A2 Adenosine Receptors and Vascular Pathologies. *Arterioscler Thromb Vasc Biol* [Internet]. 2012 Apr [cited 2022 Nov 28];32(4):870. Available from: </pmc/articles/PMC5755359/>
23. Wang H, Zhang W, Tang R, Zhu C, Bucher C, Blazar BR, et al. Adenosine receptor A2a

deficiency in leukocytes increases arterial neointima formation in apolipoprotein E-deficient mice. *Arterioscler Thromb Vasc Biol* [Internet]. 2010 May 1 [cited 2022 Nov 28];30(5):915–22. Available from:

<https://www.ahajournals.org/doi/abs/10.1161/ATVBAHA.109.202572>

24. Zerneck A, Bidzhekov K, Özüyan B, Fraemohs L, Liehn EA, Lüscher-Firzlaff JM, et al. CD73/Ecto-5'-Nucleotidase Protects Against Vascular Inflammation and Neointima Formation. *Circulation* [Internet]. 2006 May 2 [cited 2022 Nov 28];113(17):2120–7.

Available from:

<https://www.ahajournals.org/doi/abs/10.1161/CIRCULATIONAHA.105.595249>

25. Manferdini C, Paoletta F, Gabusi E, Gambari L, Piacentini A, Filardo G, et al. Adipose stromal cells mediated switching of the pro-inflammatory profile of M1-like macrophages is facilitated by PGE2: in vitro evaluation. *Osteoarthritis Cartilage* [Internet]. 2017 Jul 1 [cited 2022 Dec 18];25(7):1161–71. Available from: <https://pubmed.ncbi.nlm.nih.gov/28153787/>

26. Ma H, Li YN, Song L, Liu R, Li X, Shang Q, et al. Macrophages inhibit adipogenic differentiation of adipose tissue derived mesenchymal stem/stromal cells by producing pro-inflammatory cytokines. *Cell Biosci* [Internet]. 2020 Jul 20 [cited 2022 Dec 18];10(1).

Available from: <https://pubmed.ncbi.nlm.nih.gov/32699606/>

27. Joffin N, Paschoal VA, Gliniak CM, Crewe C, Elnwasany A, Szewda LI, et al. Mitochondrial metabolism is a key regulator of the fibro-inflammatory and adipogenic stromal subpopulations in white adipose tissue. *Cell Stem Cell* [Internet]. 2021 Apr 1

[cited 2022 Dec 18];28(4):702-717.e8. Available from:

<https://pubmed.ncbi.nlm.nih.gov/33539722/>

28. Hotamisligil GS. Inflammation, metaflammation and immunometabolic disorders. *Nature* [Internet]. 2017 Feb 8 [cited 2022 Dec 18];542(7640):177–85. Available from:

<https://pubmed.ncbi.nlm.nih.gov/28179656/>

29. Enache OM, Lahr DL, Natoli TE, Litichevskiy L, Wadden D, Flynn C, et al. The GCTx format and cmap{Py, R, M, J} packages: Resources for optimized storage and integrated traversal of annotated dense matrices. *Bioinformatics* [Internet]. 2019 Apr 15 [cited 2020 Oct 29];35(8):1427–9. Available from: www.hdfgroup.org/HDF5/.
30. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu X, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* [Internet]. 2017 Nov 30 [cited 2020 Oct 29];171(6):1437-1452.e17. Available from: [/pmc/articles/PMC5990023/?report=abstract](https://pubmed.ncbi.nlm.nih.gov/27900000/)
31. Duan Q, Flynn C, Niepel M, Hafner M, Muhlich JL, Fernandez NF, et al. LINCS Canvas Browser: Interactive web app to query, browse and interrogate LINCS L1000 gene expression signatures. *Nucleic Acids Res* [Internet]. 2014 Jul 1 [cited 2020 Oct 31];42(W1). Available from: <https://pubmed.ncbi.nlm.nih.gov/24906883/>
32. Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, et al. The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database (Oxford)* [Internet]. 2016 Jan 1 [cited 2020 Oct 31];2016. Available from: <https://academic.oup.com/database/article/doi/10.1093/database/baw100/2630482>
33. Clark NR, Hu KS, Feldmann AS, Kou Y, Chen EY, Duan Q, et al. The characteristic direction: a geometrical approach to identify differentially expressed genes. *BMC Bioinformatics* [Internet]. 2014 Mar 21 [cited 2020 Oct 31];15(1):79. Available from: <http://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-15-79>
34. Duan Q, Reid SP, Clark NR, Wang Z, Fernandez NF, Rouillard AD, et al. L1000CDS2: LINCS L1000 characteristic direction signatures search engine. *npj Syst Biol Appl* [Internet]. 2016 Aug 4 [cited 2020 Oct 31];2(1):16015. Available from: <http://www.nature.com/articles/npjsba201615>

35. De Domenico M, Nicosia V, Arenas A, Latora V. Structural reducibility of multilayer networks. *Nat Commun* [Internet]. 2015 Apr 23 [cited 2020 Dec 17];6(1):6864. Available from: <http://www.nature.com/articles/ncomms7864>
36. Bairoch A. The cellosaurus, a cell-line knowledge resource. *J Biomol Tech* [Internet]. 2018 Jul 1 [cited 2020 Dec 20];29(2):25–38. Available from: [/pmc/articles/PMC5945021/?report=abstract](http://pmc/articles/PMC5945021/?report=abstract)
37. Robin T, Capes-Davis A, Bairoch A. CLASTR: The Cellosaurus STR similarity search tool - A precious help for cell line authentication. *Int J Cancer* [Internet]. 2020 Mar 4 [cited 2020 Dec 20];146(5):1299–306. Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.32639>
38. Sarntivijai S, Lin Y, Xiang Z, Meehan TF, Diehl AD, Vempati UD, et al. CLO: The cell line ontology. *J Biomed Semantics* [Internet]. 2014 Aug 13 [cited 2020 Dec 20];5(1):37. Available from: [/pmc/articles/PMC4387853/?report=abstract](http://pmc/articles/PMC4387853/?report=abstract)
39. Wang JZ, Du Z, Payattakool R, Yu PS, Chen C-F. A new method to measure the semantic similarity of GO terms. *Bioinformatics* [Internet]. 2007 May 15 [cited 2020 Dec 20];23(10):1274–81. Available from: <https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btm087>
40. Moret N, Clark NA, Hafner M, Wang Y, Lounkine E, Medvedovic M, et al. Cheminformatics Tools for Analyzing and Designing Optimized Small-Molecule Collections and Libraries. *Cell Chem Biol* [Internet]. 2019 May 16 [cited 2020 Nov 1];26(5):765-777.e3. Available from: <https://doi.org/10.1016/j.chembiol.2019.02.018>
41. Wang Y, Cornett A, King FJ, Mao Y, Nigsch F, Paris CG, et al. Evidence-Based and Quantitative Prioritization of Tool Compounds in Phenotypic Drug Discovery. *Cell Chem Biol* [Internet]. 2016 Jul 21 [cited 2020 Nov 1];23(7):862–74. Available from: <http://dx.doi.org/10.1016/j.chembiol.2016.05.016><http://dx.doi.org/10.1016/j.chembiol.2016>

.05.016

42. Cao Y, Charisi A, Cheng LC, Jiang T, Girke T. ChemmineR: A compound mining framework for R. *Bioinformatics* [Internet]. 2008 Aug [cited 2020 Dec 17];24(15):1733–4. Available from: [/pmc/articles/PMC2638865/?report=abstract](https://pubmed.ncbi.nlm.nih.gov/17334/)
43. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst* [Internet]. 2015 Dec 23 [cited 2018 May 22];1(6):417–25. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/26771021>
44. 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 [Internet]. Vol. 23, *Nature Medicine*. Nature Publishing Group; 2017 [cited 2020 Dec 19]. p. 405–8. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5568558/>
45. Cotto KC, Wagner AH, Feng YY, Kiwala S, Coffman AC, Spies G, et al. DGIdb 3.0: A redesign and expansion of the drug-gene interaction database. *Nucleic Acids Res* [Internet]. 2018 Jan 1 [cited 2020 Dec 19];46(D1):D1068–73. Available from: <https://academic.oup.com/nar/article/46/D1/D1068/4634012>
46. Ruiz C, Zitnik M, Leskovec J. Identification of disease treatment mechanisms through the multiscale interactome. *Nat Commun* 2021 121 [Internet]. 2021 Mar 19 [cited 2023 Oct 31];12(1):1–15. Available from: <https://www.nature.com/articles/s41467-021-21770-8>
47. Wang Y, Yang Y, Chen S, Wang J. DeepDRK: a deep learning framework for drug repurposing through kernel-based multi-omics integration. *Brief Bioinform* [Internet]. 2021 Sep 2 [cited 2023 Oct 31];22(5):1–10. Available from: <https://dx.doi.org/10.1093/bib/bbab048>
48. Winter DJ. rentrez: An R package for the NCBI eUtils API. 2017 Aug 25 [cited 2022 Jun 7]; Available from: <https://peerj.com/preprints/3179>

Figure Legends:

Figure S1: **(A)** The distribution of experimental instances for each perturbagen for each cell line in the LINCS-CMap data; full (top panel) and zoomed-in (bottom panel) versions. **(B)** Number of perturbagens and breakdown of perturbagen type for each cell line; full (top panel) and zoomed-in (bottom panel) versions.

Figure S2: Sensitivity of QUIZ-C networks to the choice of quantiles and other network building parameters. Similarity between network topologies is quantified by the Jaccard coefficient of overlap. All comparisons are made with the network configuration used in the manuscript, i.e., a 75th quantile consensus threshold.

Figure S3: **(A)** The total number of drugs in the top 50 predictions made by PACOS and PACOS combined with tool score, as a function of the PACOS-to-tool-score-ratio r . **(B)** The AUROC for PACOS and PACOS combined with tool score, as a function of the PACOS-to-tool-score-ratio r . **(C)** Maximum AUROC values of PACOS-QUIZ-C and PACOS-QUIZ-C-Tool (in repress and enhance modes) when using GEO kinase, GEO TF and GTEx tissue datasets.

Figure S4: **(A)** Kernel density estimations of ZS distributions for all cell lines. Each line shows the Level 4 ZS distribution of a single gene across all perturbagens. **(B)** The The kernel density estimation for the Level 4 ZS distribution of *DNAJB6* in the HS27A cell line.

Figure S5: (A) Distributions of the number of doses administered for each cell line in LINCS-CMap data. (B) Distributions of the number of time points at which measurements were taken, for each cell line in LINCS-CMap data.

Figure S6: Overview of QUIZ-C cell type-specific gene-perturbation networks. Green and yellow nodes indicate perturbations and genes, respectively. Up- and down-regulating interactions are shown in red and blue edges, respectively. The positions of the drugs (green nodes) are fixed across all networks. Instructions on how to access the zoomable and interactive versions of each QUIZ-C network can be found at <https://github.com/r-duh/Pathopticon>.

Figure S7: (A) The number of cell type-specific MODZ networks a given drug appears in, as a function of the mean normalized out-degree of each drug across all MODZ networks. Every circle represents a drug, and circle size is proportional to the standard deviation of the normalized out-degree. (B) The number of cell type-specific MODZ networks a given gene appears in, as a function of the mean normalized in-degree of each gene across all MODZ networks. Every circle represents a gene, and the circle size is proportional to the standard deviation of the normalized in-degree. (C) The number of cell type-specific CD networks a given drug appears in, as a function of the mean normalized out-degree of each drug across all CD networks. Every circle represents a drug, and circle size is proportional to the standard deviation of the normalized out-degree. (D) The number of cell type-specific CD networks a given gene appears in, as a function of the mean normalized in-degree of each gene across all CD networks. Every circle represents a gene, and the circle size is proportional to the standard deviation of the normalized in-degree.

Figure S8: In-degree distributions of landmark vs. predicted genes for each cell line for QUIZ-C (A), MODZ (B), and CD (C).

Figure S9: Hierarchically clustered heatmap qualitatively showing the edge overlap between QUIZ-C (A), MODZ (B), and CD (C) networks. Each row and column correspond to a cell line. The colors correspond to the Jaccard index of edge overlap. (D) (Upper panel) Full boxplots showing the edge overlap Jaccard index of each cell line with all other cell lines. Green, red, and blue correspond to QUIZ-C, MODZ and CD networks, respectively. (Lower panel) Zoomed-in version of the boxplot without fliers. (E) Edge density of CMap-based networks. Green, red, and blue correspond to QUIZ-C, MODZ and CD networks. (F) Boxplots showing percent edge concordance (i.e., the percentage of overlapping edges with the same direction of effect) of each cell line with respect to the other cell lines. Green, red, and blue correspond to QUIZ-C, MODZ and CD networks, respectively.

Figure S10: (A) Reducibility of CD networks. (B) Reducibility of QUIZ-C networks. Left panel shows the relative entropy quality function $q(\cdot)$ of each hierarchical clustering branch, with the maximum value of $q(\cdot)$ corresponding to the optimal configuration of the aggregation of layers. Right panel shows the hierarchical clustering dendrogram and the optimal clustering threshold. Cell lines belonging to the same cluster are in the same color, and clusters consisting of only one cell line are shown in light grey.

Figure S11: (A) Hierarchically clustered heatmap showing the single Tandem Repeat (STR) profile similarity between LINCS-CMap cell lines, calculated by CLASTR. (B) (Top) Boxplots of STR similarities showing the similarity values above the detection threshold. (Bottom) The

similarity score corresponding to the detection threshold (red lines) and the percentage of cell lines that fall below this detection threshold (grey bars).

Figure S12: (A) Hierarchically clustered heatmap showing the Cell Line Ontology (CLO) similarity between LINCS-CMap cell lines. **(B)** Boxplots of CLO similarities of each cell line with respect to the other LINCS-CMap cell lines.

Figure S13: (A) Percentage of drugs in CMap-based networks that have known mechanisms of action per the Drug Repurposing Hub. Green, red, and blue circles correspond to QUIZ-C, MODZ and CD networks, respectively. **(B-C)** Overrepresentation analysis of drugs with known MoA in MODZ **(B)** and CD **(C)** networks, as measured by the p-values and odds ratios given by the two-tailed Fisher's exact test. Red text indicates significant over-representation, blue text indicates significant under-representation, and black text indicates neither over- or under-representation ($p\text{-value} > 0.05$) representation.

Figure S14: (A) Clinical development phase breakdown of the drugs in each MODZ network. **(B)** Clinical development phase breakdown of the drugs in each CD network.

Figure S15: (A) The percentage of drugs in QUIZ-C networks with known targets in the Drug Repurposing Hub (DRH) and Drug-Gene Interaction Database (DGIdb). Light green indicates percentage in DRH only and dark green indicates the additions from DGIdb. **(B)** The odds ratio of enrichment of known drug-target interactions in each MODZ network. Error bars indicate 95% confidence intervals and cell lines in red indicate a significant enrichment. **(C)** The odds ratio of enrichment of known drug-target interactions in each CD network. Error bars indicate 95% confidence intervals and cell lines in red indicate a significant enrichment. **(D)** The

difference of log odds-ratios between QUIZ-C and MODZ networks. Green and red denote cell lines for which QUIZ-C and MODZ have a higher log odds-ratio, respectively. **(E)** The difference of log odds-ratios between QUIZ-C and CD networks. Green and blue denote cell lines for which QUIZ-C and CD have a higher log odds-ratio, respectively.

Figure S16: Scatter plots showing (left) chi-squared statistic values vs SCS values; (right) odds ratios vs SCS values.

Figure S17: Variance-covariance matrix of the Enrichr disease signatures.

Figure S18: (A) PACOS distributions across all cell lines as a function of the cophenetic distance on the hierarchical clustering dendrogram of Figure S15. **(B)** Maximum PACOS values for each cell line as a function of the cophenetic distance on the hierarchical clustering dendrogram of Figure S15. Dashed line indicates the average of maximum PACOS values over all cell lines and the numbers indicate the number of clusters (i.e., the number of disease signatures used in the PACOS calculation) at each cophenetic distance value. **(C)** Mean PACOS values for each cell line as a function of the cophenetic distance on the hierarchical clustering dendrogram of Figure S15. Dashed line indicates the average of mean PACOS values over all cell lines at each cophenetic distance value.

Figure S19: Circle heatmaps showing the AUROC values of benchmark signatures (rows) for each cell line (columns). Circle size reflects AUROC value and cell lines marked in darker shade are the significant cases for which empirical p-value < 0.05. Bars at the top show the total number of significant benchmark signatures per cell line and bars on the right show the

total number of significant cell lines per benchmark signature. **(A)** PACOS in repress mode. **(B)** PACOS in enhance mode.

Figure S20: Circle heatmaps showing the AUROC values of benchmark signatures (rows) for each cell line (columns). Circle size reflects AUROC value and cell lines marked in darker shade are the significant cases for which empirical p-value < 0.05. Bars at the top show the total number of significant benchmark signatures per cell line and bars on the right show the total number of significant cell lines per benchmark signature. **(A)** PACOS combined with tool score, in repress mode. **(B)** PACOS combined with tool score, in enhance mode.

Figure S21: Boxplots of the chemical similarity of the top 50 predicted drugs for each benchmark signature. Each boxplot contains data on the similarity of the given cell line with all other cell lines. Chemical similarity is defined as the Tanimoto coefficient of similarity between the atom pair fingerprints (APFPs) of the top predicted drugs. **(A)** PACOS combined with tool score in repress mode, run on QUIZ-C networks. Font colors indicate comparisons of panel (A) with panels (B) and (C): Green, red, and blue font colors indicate the benchmark signatures for which QUIZ-C top candidates had lower chemical similarity (i.e., higher chemical diversity) than the top candidates of both MODZ and CD networks, and only CD networks, respectively. **(B)** PACOS combined with tool score in repress mode, run on MODZ networks. **(C)** PACOS combined with tool score in repress mode, run on CD networks.

Figure S22: Boxplots of the chemical similarity of the top 50 predicted drugs for each benchmark signature. Each boxplot contains data on the similarity of the given cell line with all other cell lines. Chemical similarity is defined as the Tanimoto coefficient of similarity between the atom pair fingerprints (APFPs) of the top predicted drugs. **(A)** PACOS combined with tool

score in enhance mode, run on QUIZ-C networks. Font colors indicate comparisons of panel (A) with panels (B) and (C): Green, red, and blue font colors indicate the benchmark signatures for which QUIZ-C top candidates had lower chemical similarity (i.e., higher chemical diversity) than the top candidates of both MODZ and CD networks, only MODZ networks, and only CD networks, respectively. **(B)** PACOS combined with tool score in enhance mode, run on MODZ networks. **(C)** PACOS combined with tool score in enhance mode, run on CD networks.

Figure S23: Boxplots of the chemical similarity of the top 50 predicted drugs for each benchmark signature. Each boxplot contains data on the similarity of the given cell line with all other cell lines. Chemical similarity is defined as the Tanimoto coefficient of similarity between the atom pair fingerprints (APFPs) of the top predicted drugs. **(A)** PACOS combined with tool score in repress mode, run on QUIZ-C networks. Font colors indicate comparisons between panels (A) and (B): Green font color indicates the benchmark signatures for which QUIZ-C top candidates had lower chemical similarity (i.e., higher chemical diversity) than the top candidates of L1000CDS2. **(B)** L1000CDS2 predictions. Font colors indicate comparisons between panels (A) and (B): Magenta font color indicates the benchmark signatures for which L1000CDS2 top candidates had lower chemical similarity (i.e., higher chemical diversity) than the top candidates of PACOS combined with tool score in repress mode, run on QUIZ-C networks.

Figure S24: Boxplots of the chemical similarity of the top 50 predicted drugs for each benchmark signature. Each boxplot contains data on the similarity of the given cell line with all other cell lines. Chemical similarity is defined as the Tanimoto coefficient of similarity between the atom pair fingerprints (APFPs) of the top predicted drugs. **(A)** PACOS combined with tool score in enhance mode, run on QUIZ-C networks. Font colors indicate comparisons between

panels (A) and (B): Green font color indicates the benchmark signatures for which QUIZ-C top candidates had lower chemical similarity (i.e., higher chemical diversity) than the top candidates of L1000CDS2. **(B)** L1000CDS2 predictions. Font colors indicate comparisons between panels (A) and (B): Magenta font color indicates the benchmark signatures for which L1000CDS2 top candidates had lower chemical similarity (i.e., higher chemical diversity) than the top candidates of PACOS combined with tool score in enhance mode, run on QUIZ-C networks.

Figure S25: Heatmaps showing the results of pathway enrichment analysis for the six compounds chosen for *in vitro* experiments. Darker colors indicate higher enrichment (i.e., lower two-tailed Fisher's exact test p-values) and blank cells indicate not significant (p-value > 0.05).

Figure S26: High resolution version of the combined network in Figure 7D with labels.

Figure S27: Boxplots showing the eigenvector centrality distributions of the three identified pathways. ** indicates two-tailed Mann-Whitney U test p-value < 0.01 and *** indicates two-tailed Mann-Whitney U test p-value < 0.001.

Figure S28: q(PCR) results of the *in vitro* validation candidate drugs for the selected panel of pro-inflammatory/athero-genic genes.

Figure S29: q(PCR) results of the *in vitro* validation candidate drugs for the selected panel of anti-inflammatory/athero-protective genes.

Table S1: Alphabetical list of benchmark signatures.

Table S2: The percentages of potentially spurious genes in QUIZ-C networks by cell type.

Table S3: The statistics of in-degrees of spurious vs other genes.

Table S4: Properties of the QUIZ-C cell type-specific gene-perturbation networks.

Table S5: Pearson's r values between the number of replicates and out-degree for QUIZ-C, MODZ and CD networks.

Table S6: Enrichment of landmark genes in each QUIZ-C network. Results are shown in terms of two-tailed Fisher's exact test p -values and odds-ratios.

Table S7: Summary of extrinsic benchmarks showing the two-tailed Mann-Whitney U test p -values between pairs of methods. Green text indicates instances where our method significantly surpasses the other methods, red text indicates instances where other methods significantly surpass our method, and black font indicates where the difference between the two methods is not significant.

Table S8: Comparison of methods in terms of their chemical structure similarity, measured by the Tanimoto coefficient.

Table S9: Top therapeutic candidates to repress the early and late stage signatures of vein graft disease (VGD), predicted by pathopticon. Predictions that are significant for a given cell

line (empirical p-value ≤ 0.05) are marked in red. Lists are truncated after the top 10 significant hits.

Supplementary Figures and Tables

Integrating pharmacogenomics and cheminformatics with diverse disease phenotypes for cell type-guided drug discovery

Arda Halu^{1,2*}, Sarvesh Chelvanambi², Julius L. Decano², Joan T. Matamalas², Mary Whelan², Takaharu Asano², Namitra Kalicharran², Sasha A. Singh², Joseph Loscalzo^{1,3}, Masanori Aikawa^{1,2*}

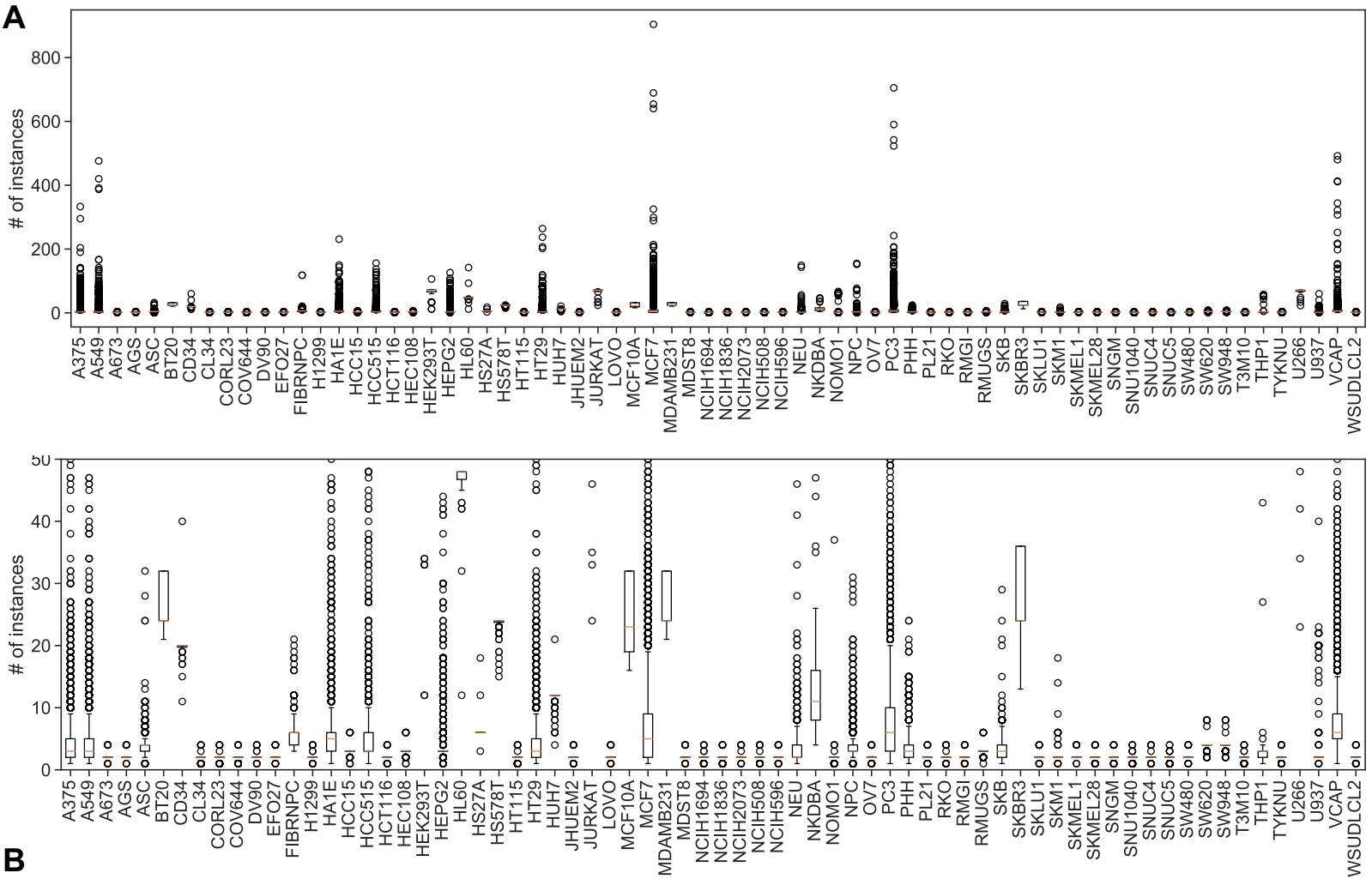
¹Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School

²Center for Interdisciplinary Cardiovascular Sciences, Division of Cardiovascular Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School

³Department of Medicine, Brigham and Women's Hospital, Harvard Medical School

Figure S1

A



B

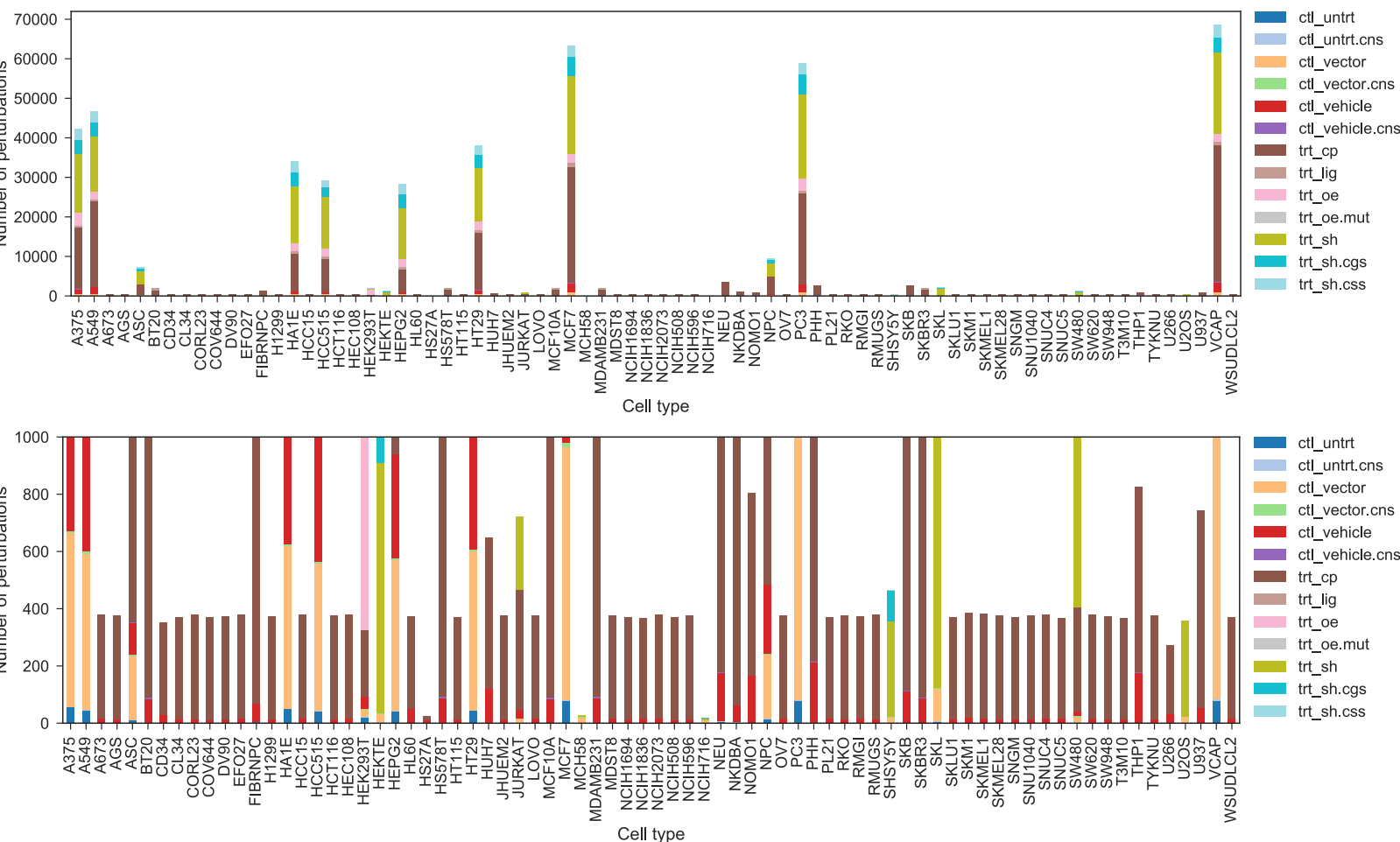


Figure S2

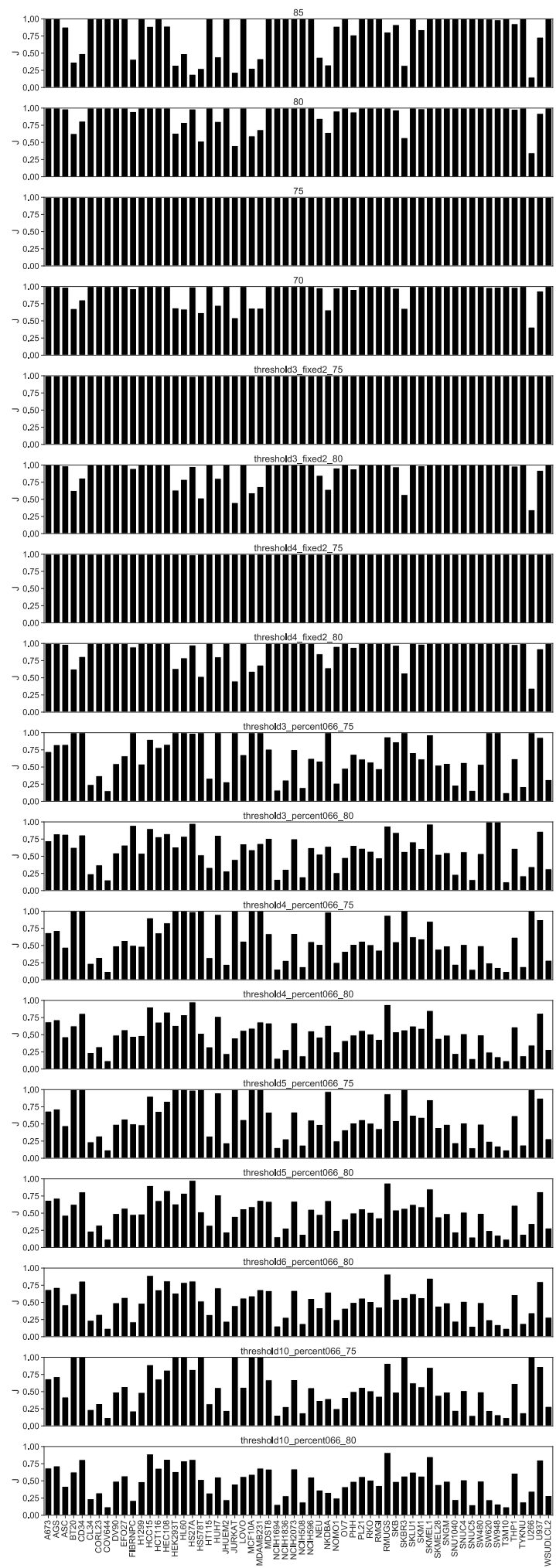
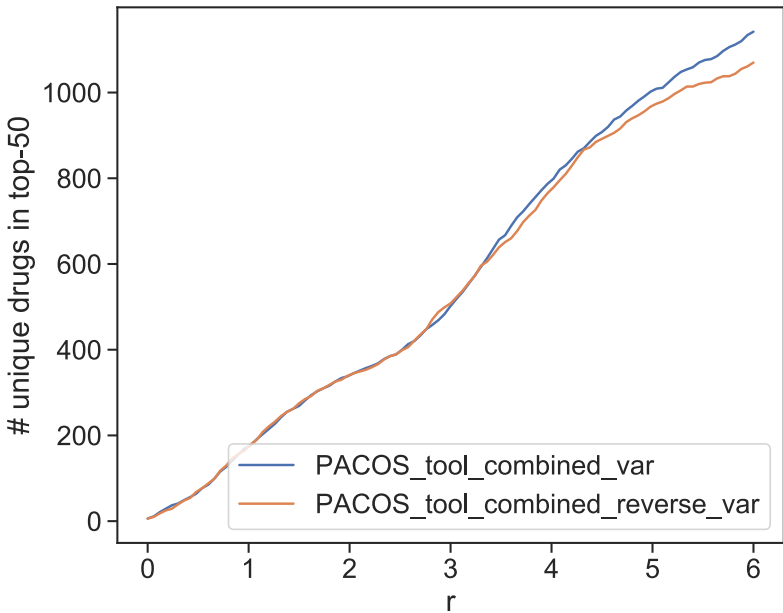
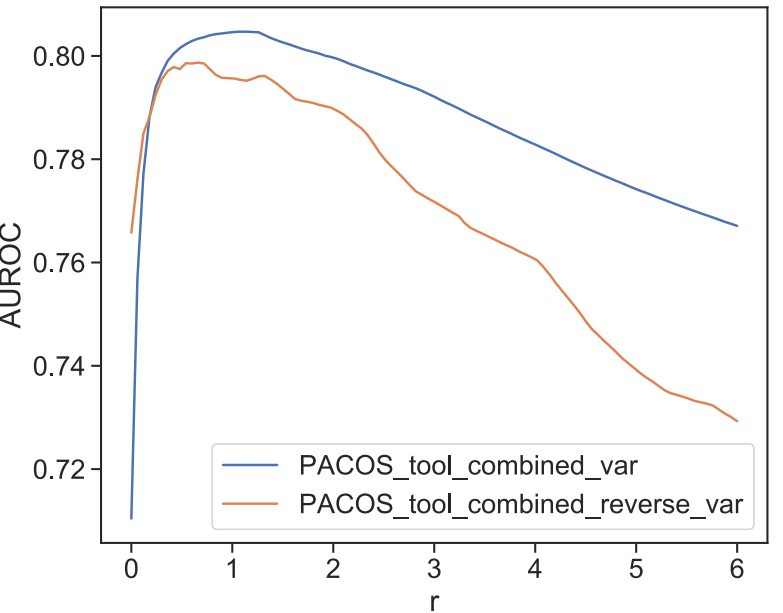


Figure S3

A



B



C

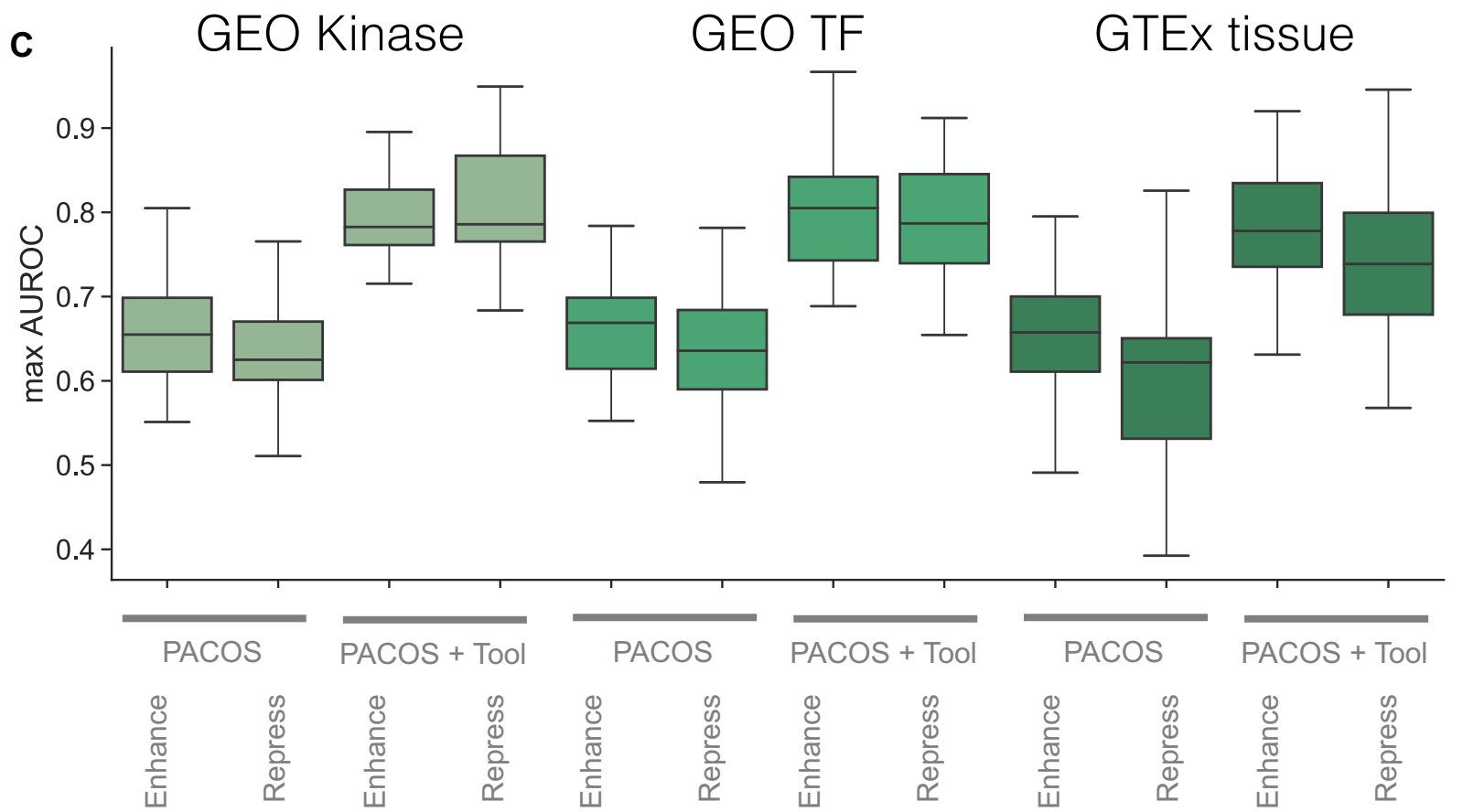
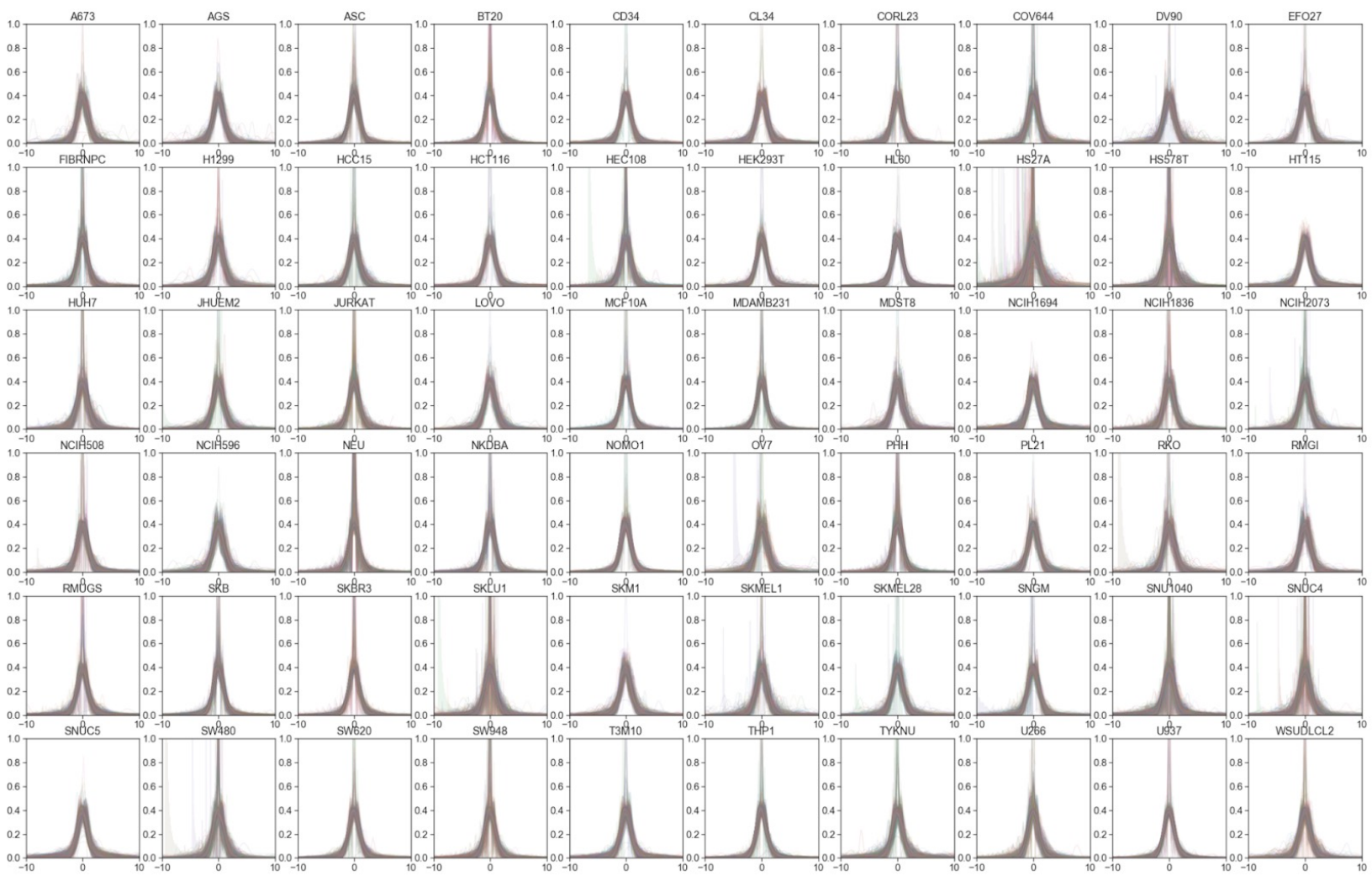


Figure S4

A



B

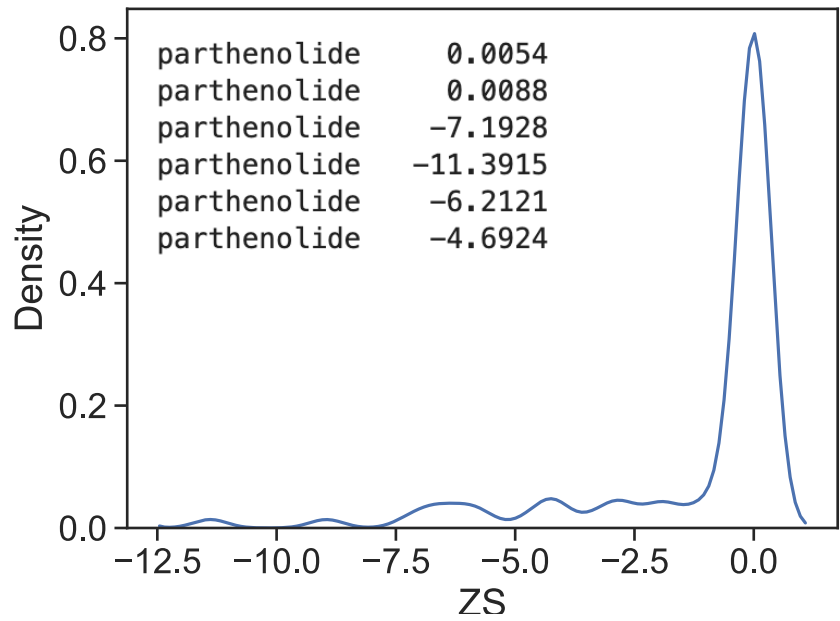
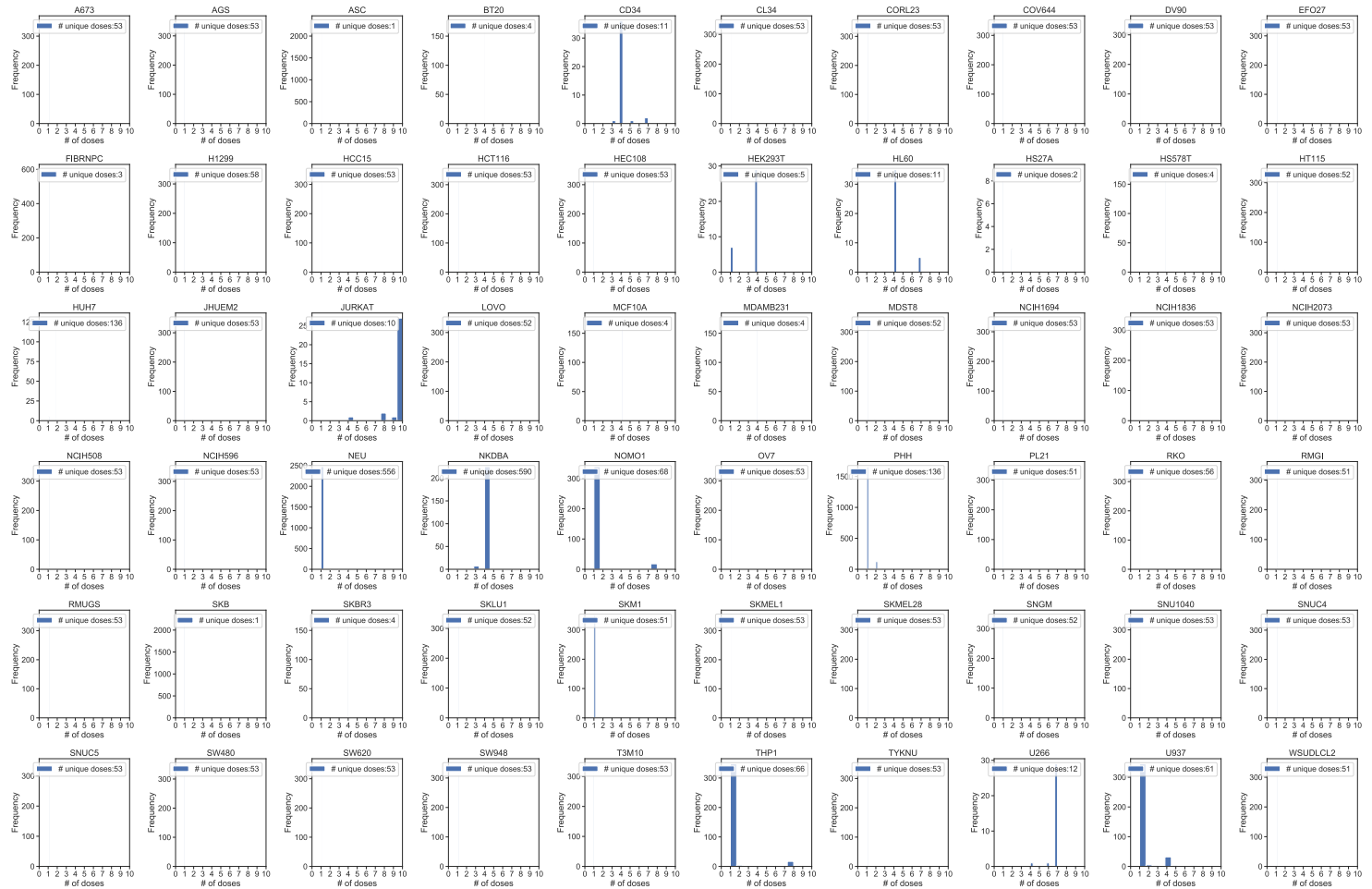


Figure S5

A



B

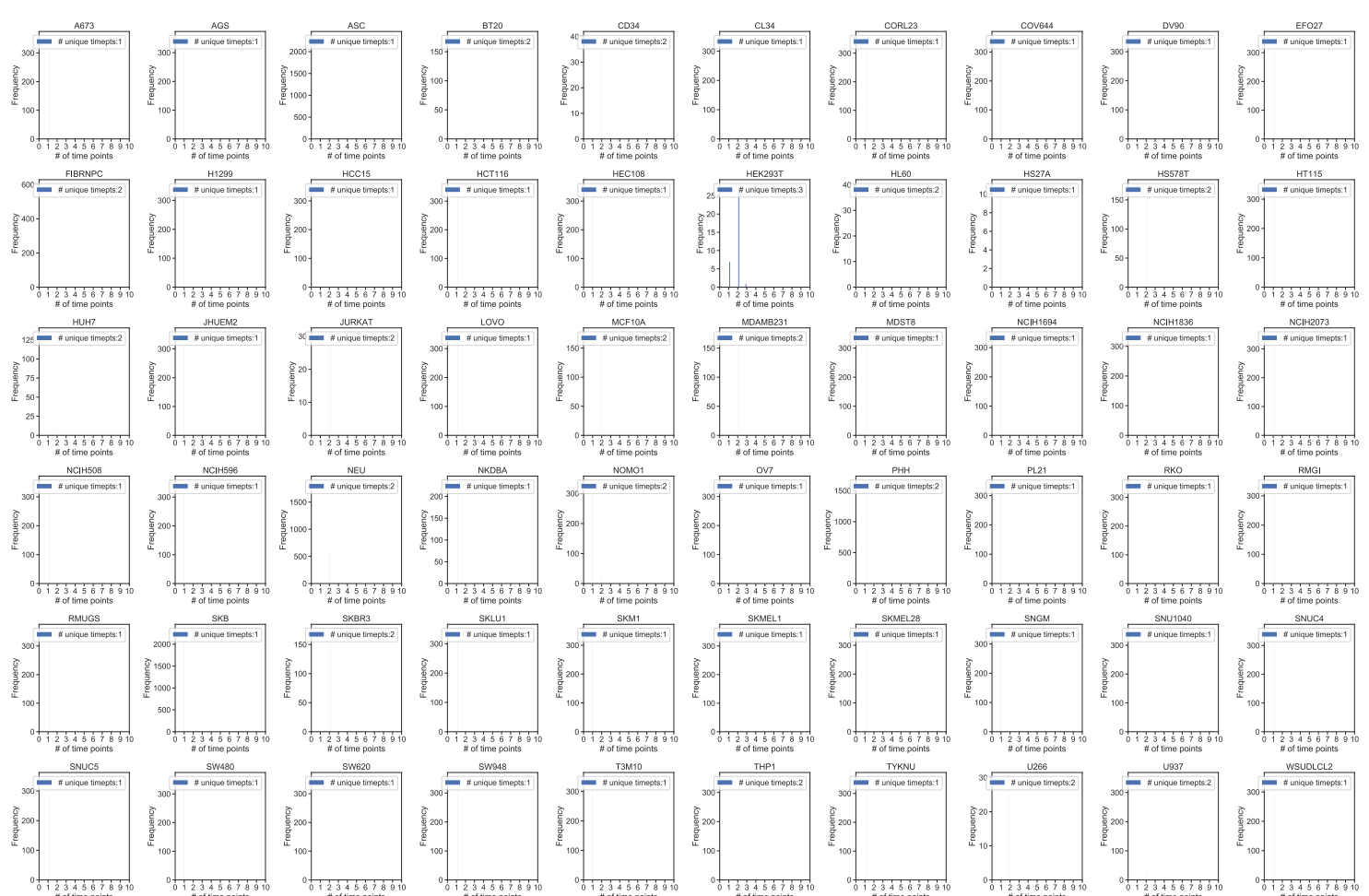


Figure S6

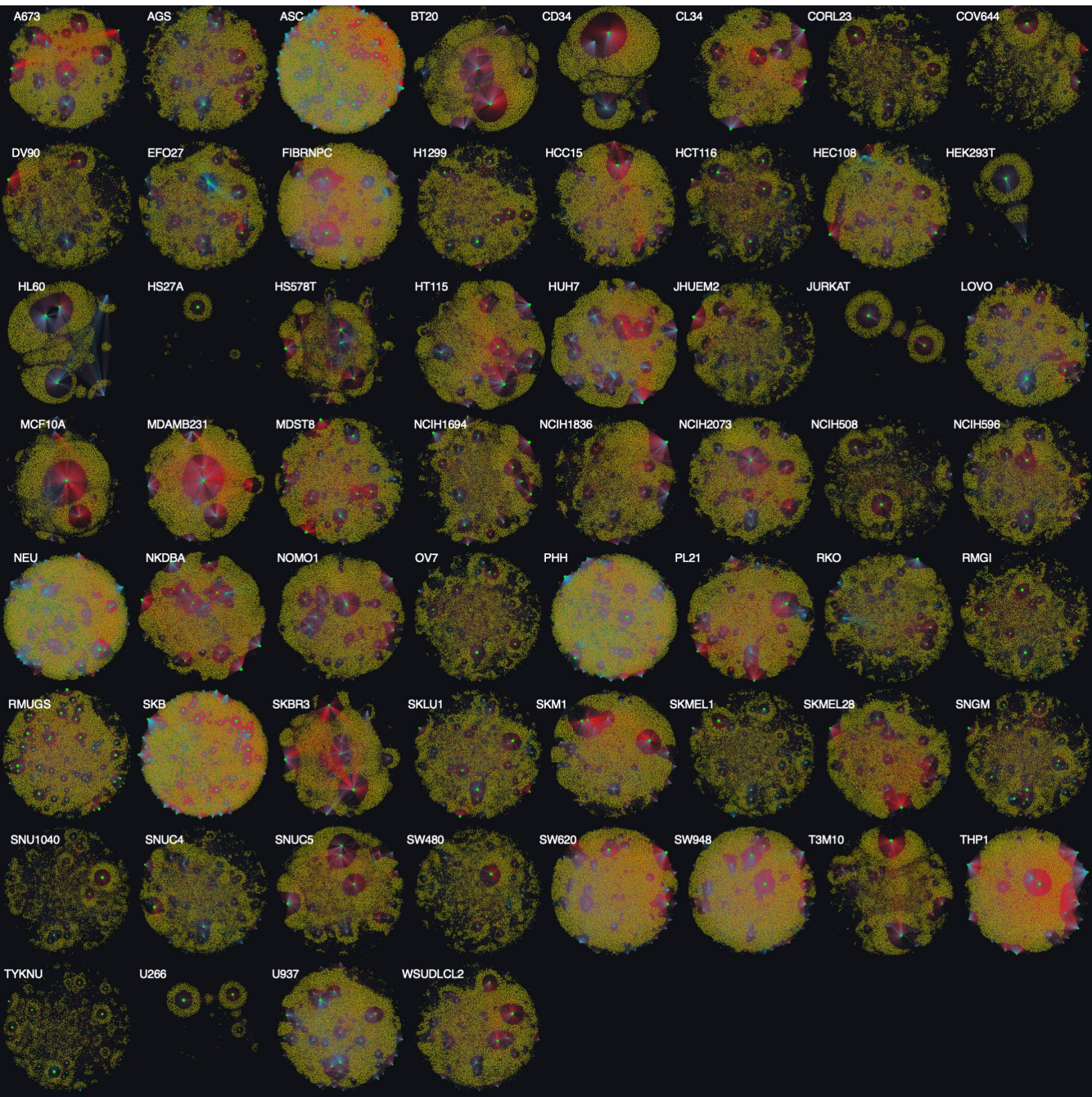
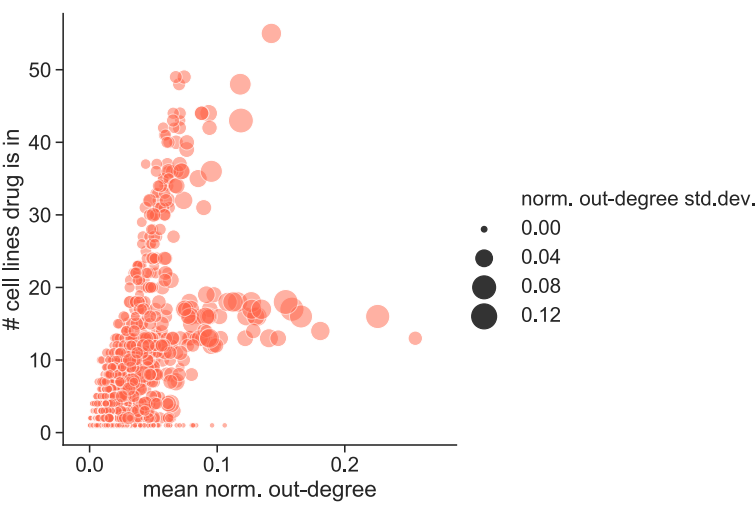
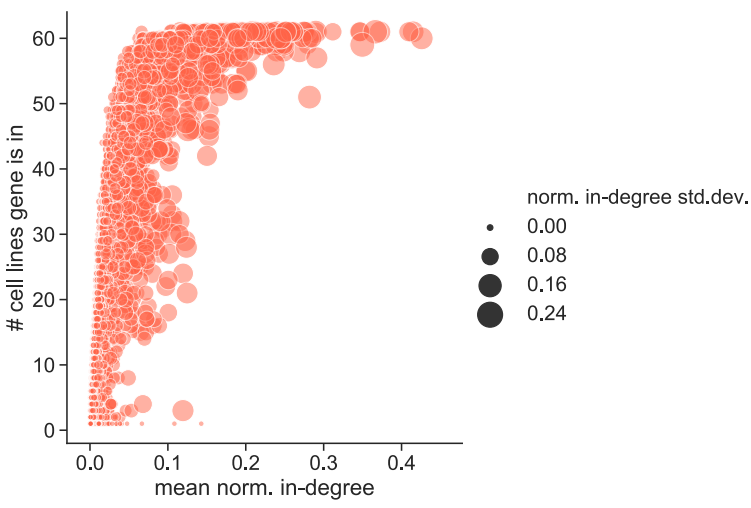


Figure S7

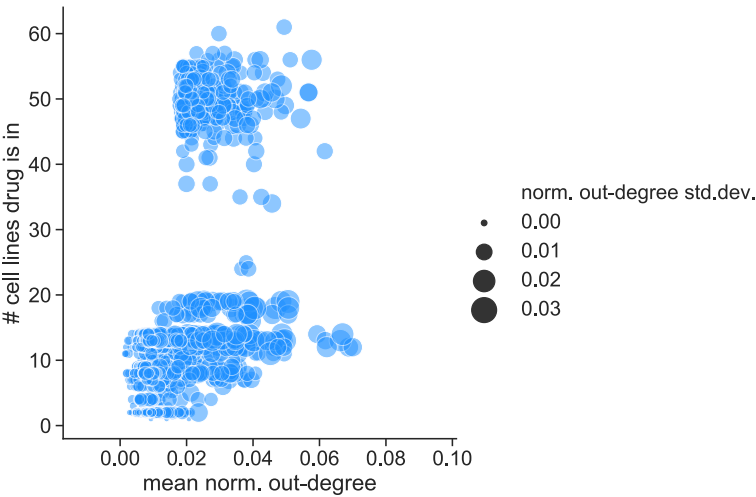
A



B



C



D

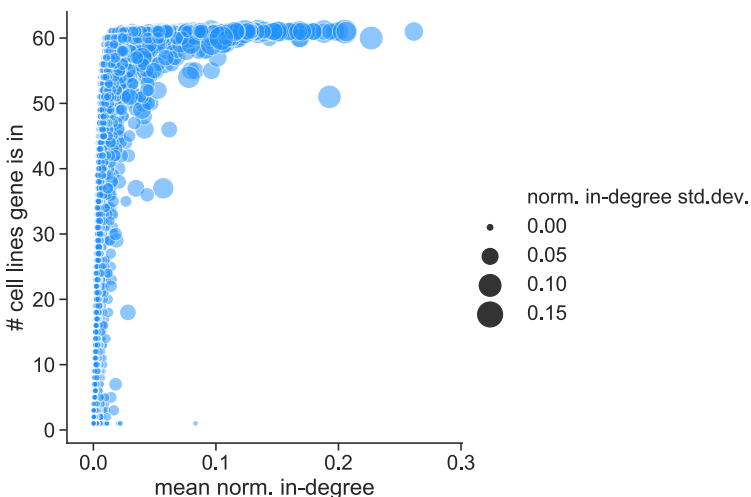
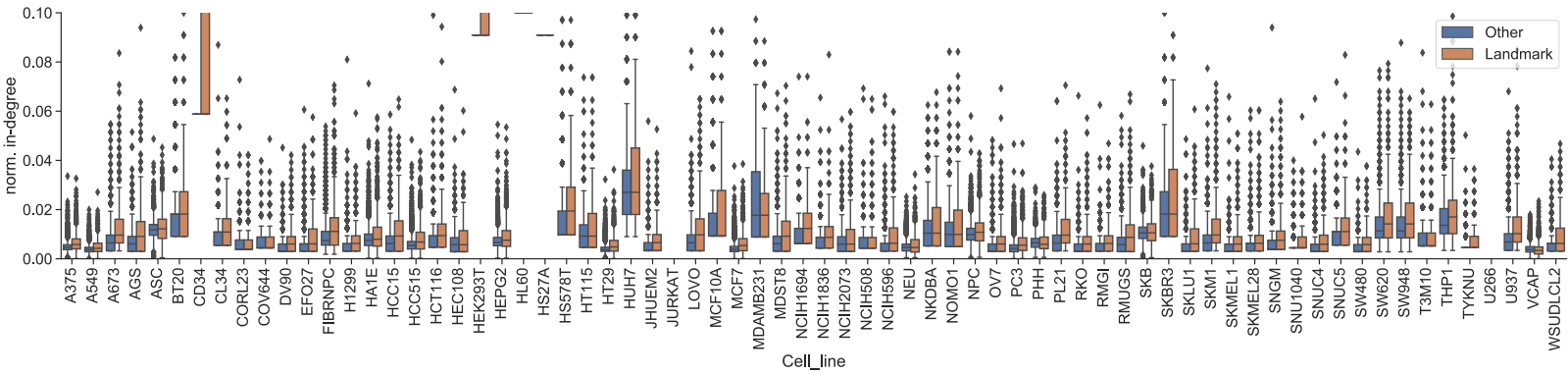
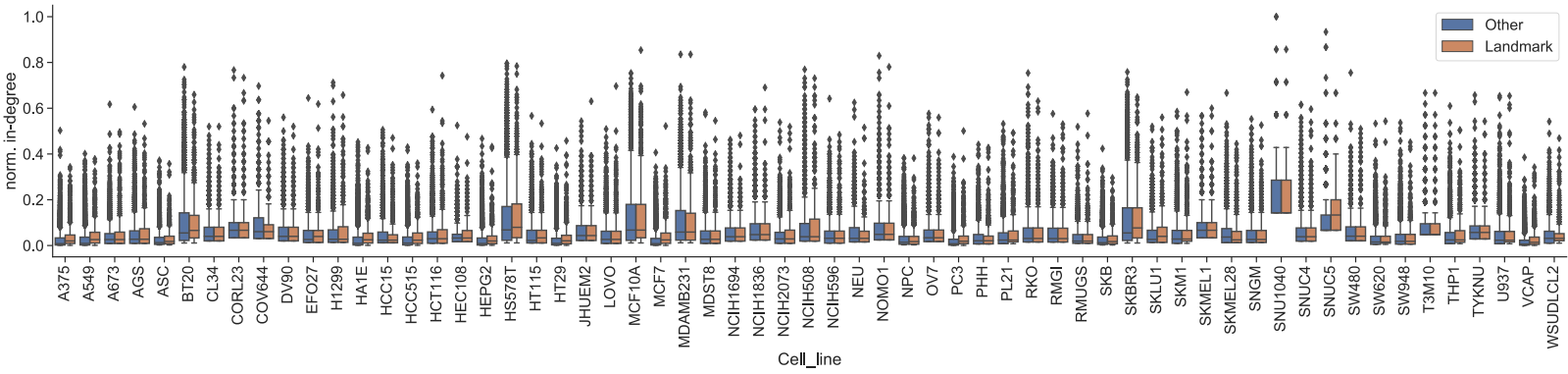


Figure S8

A



B



C

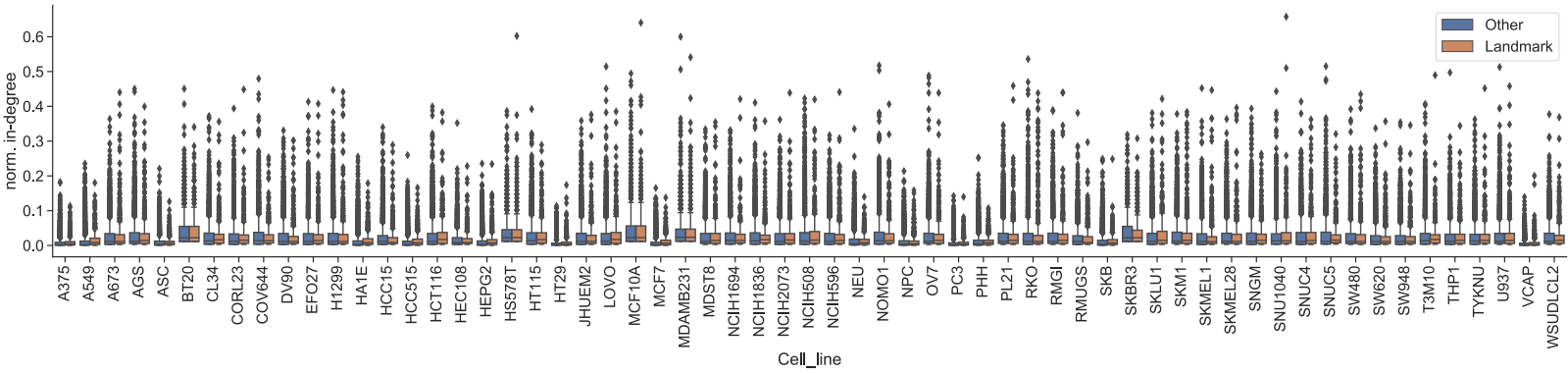


Figure S9

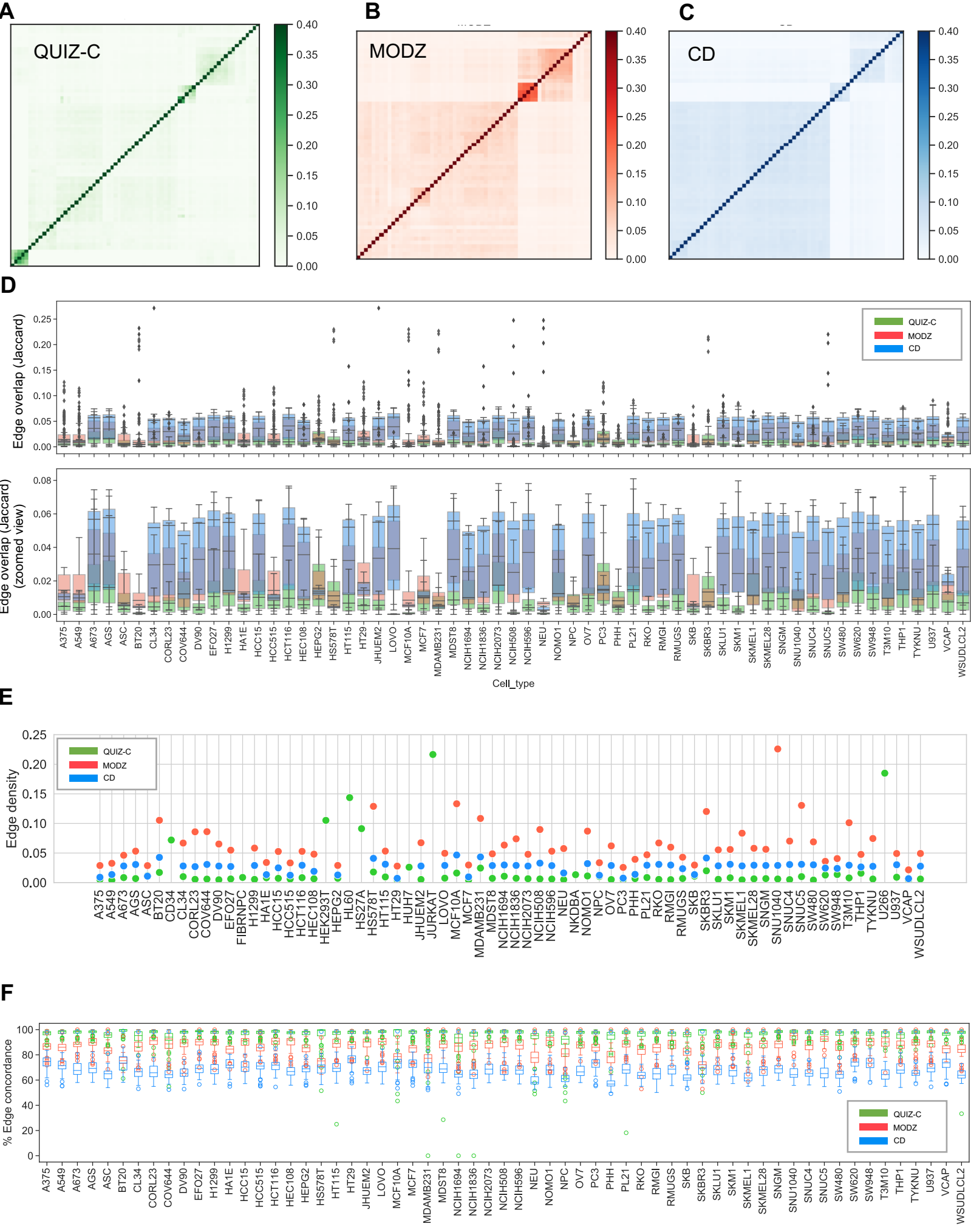


Figure S10

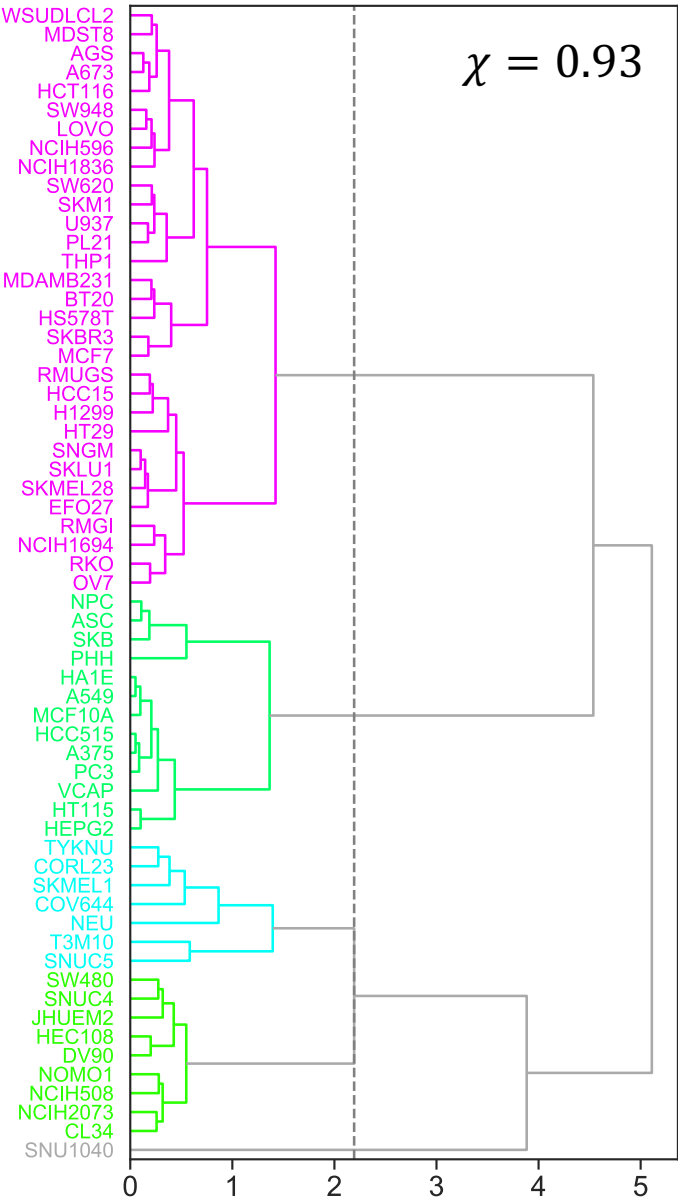
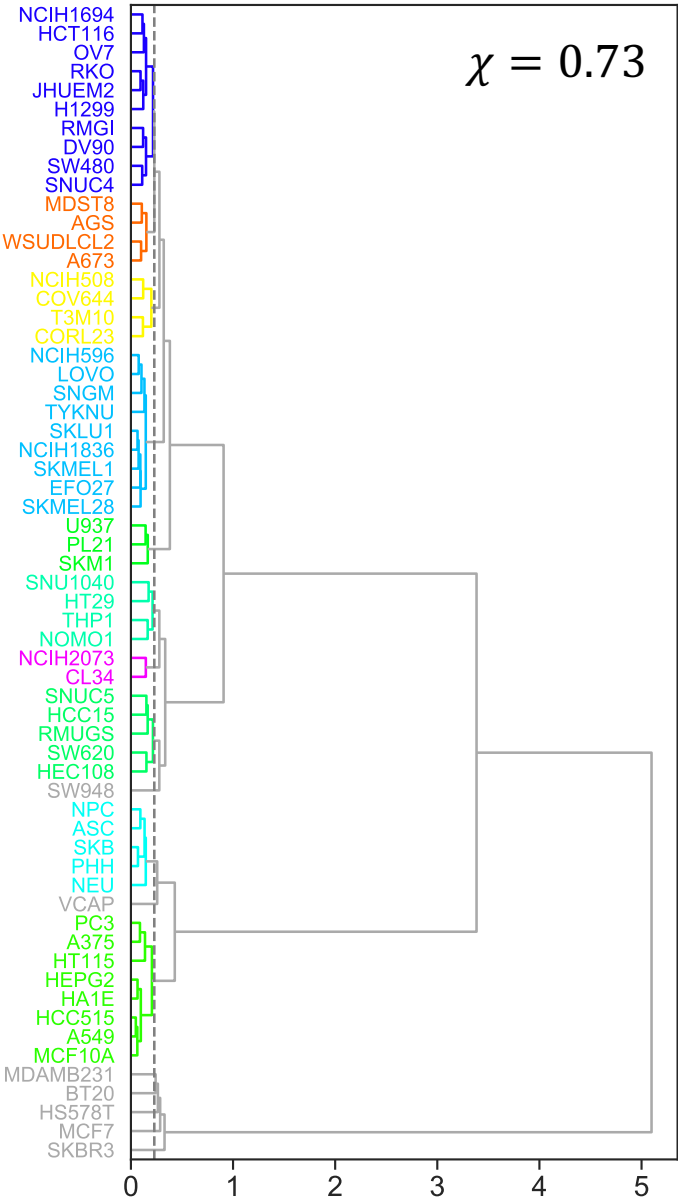
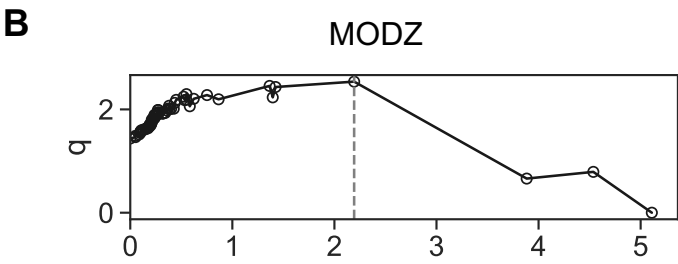
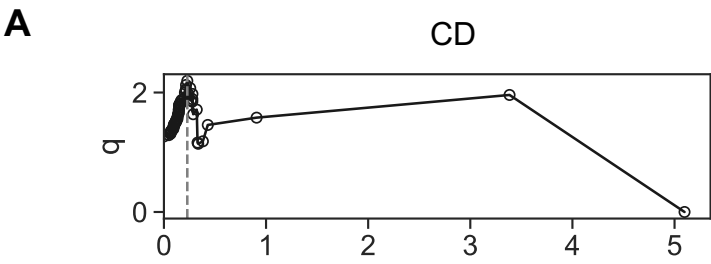


Figure S11

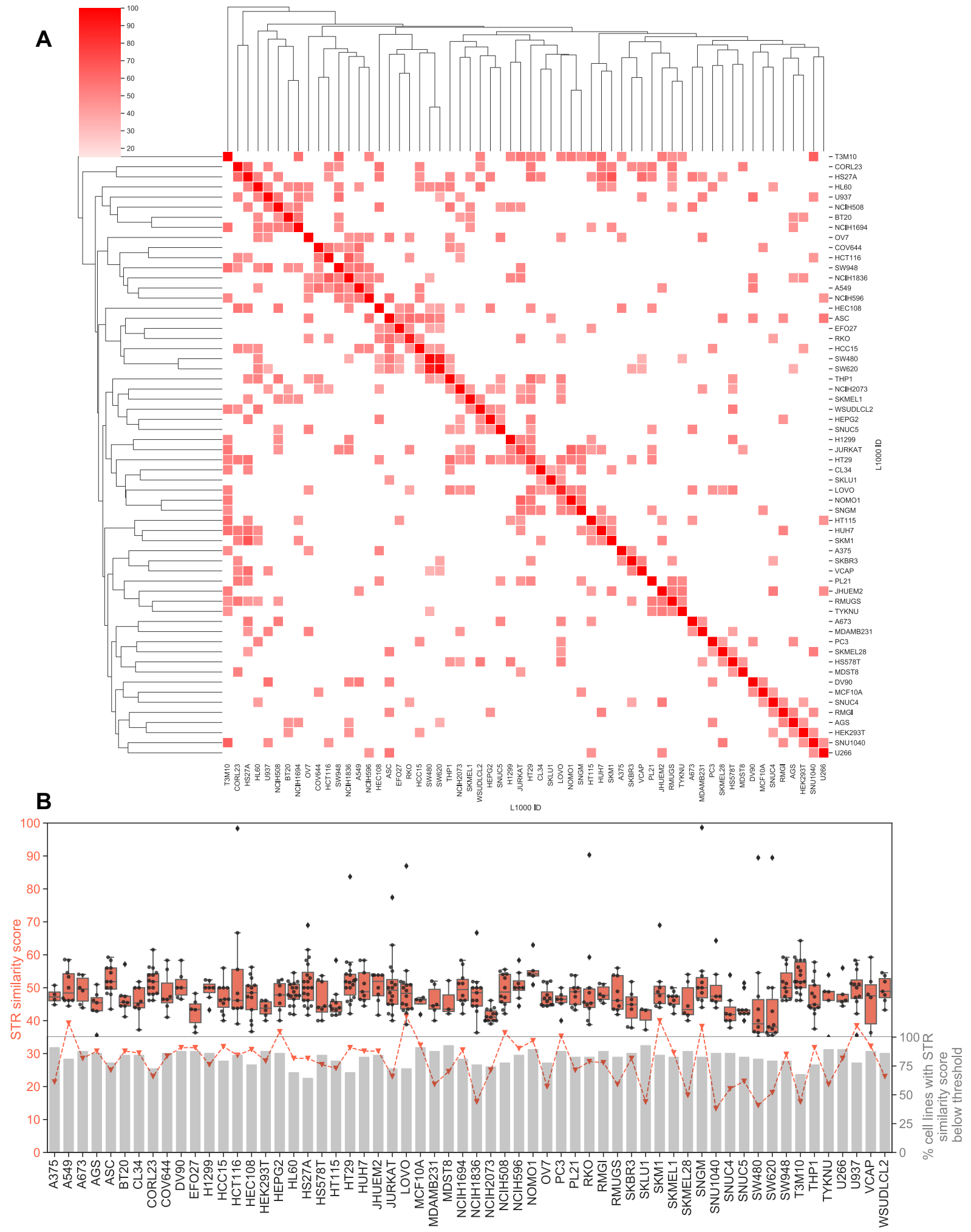
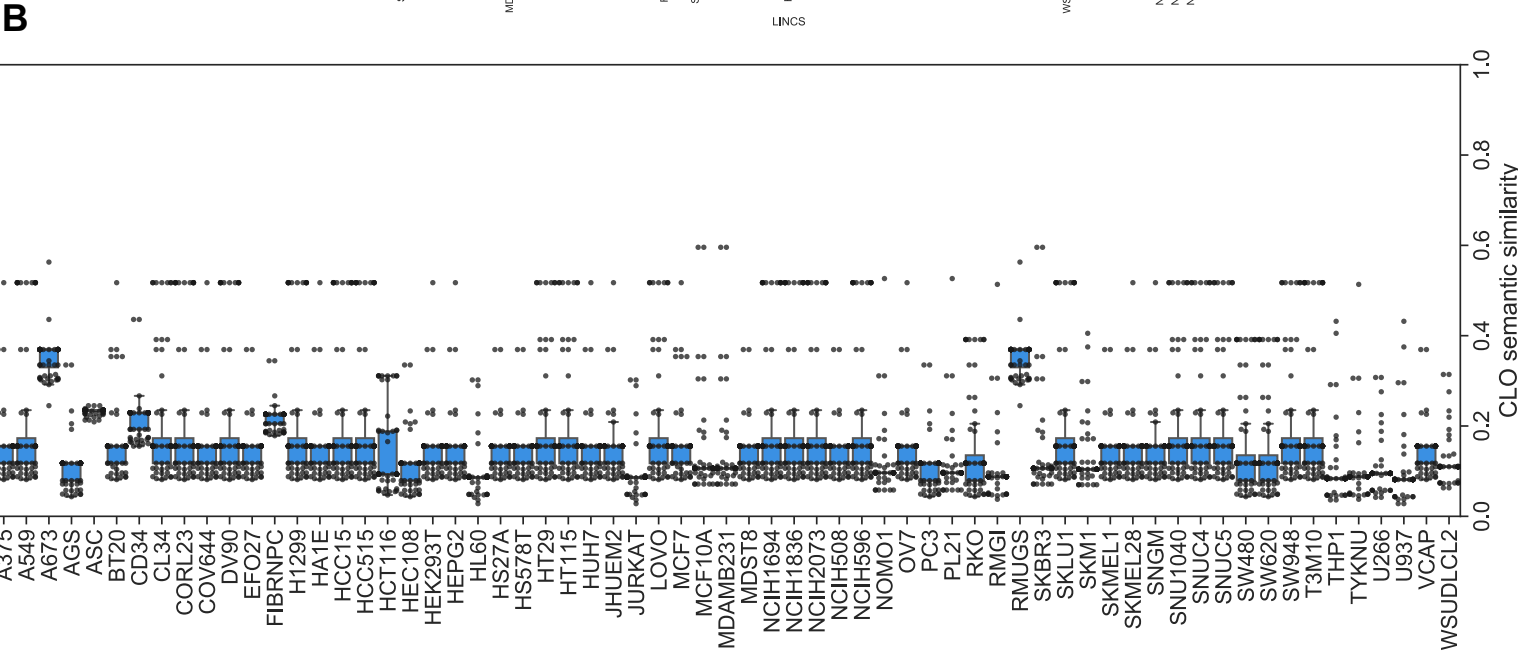
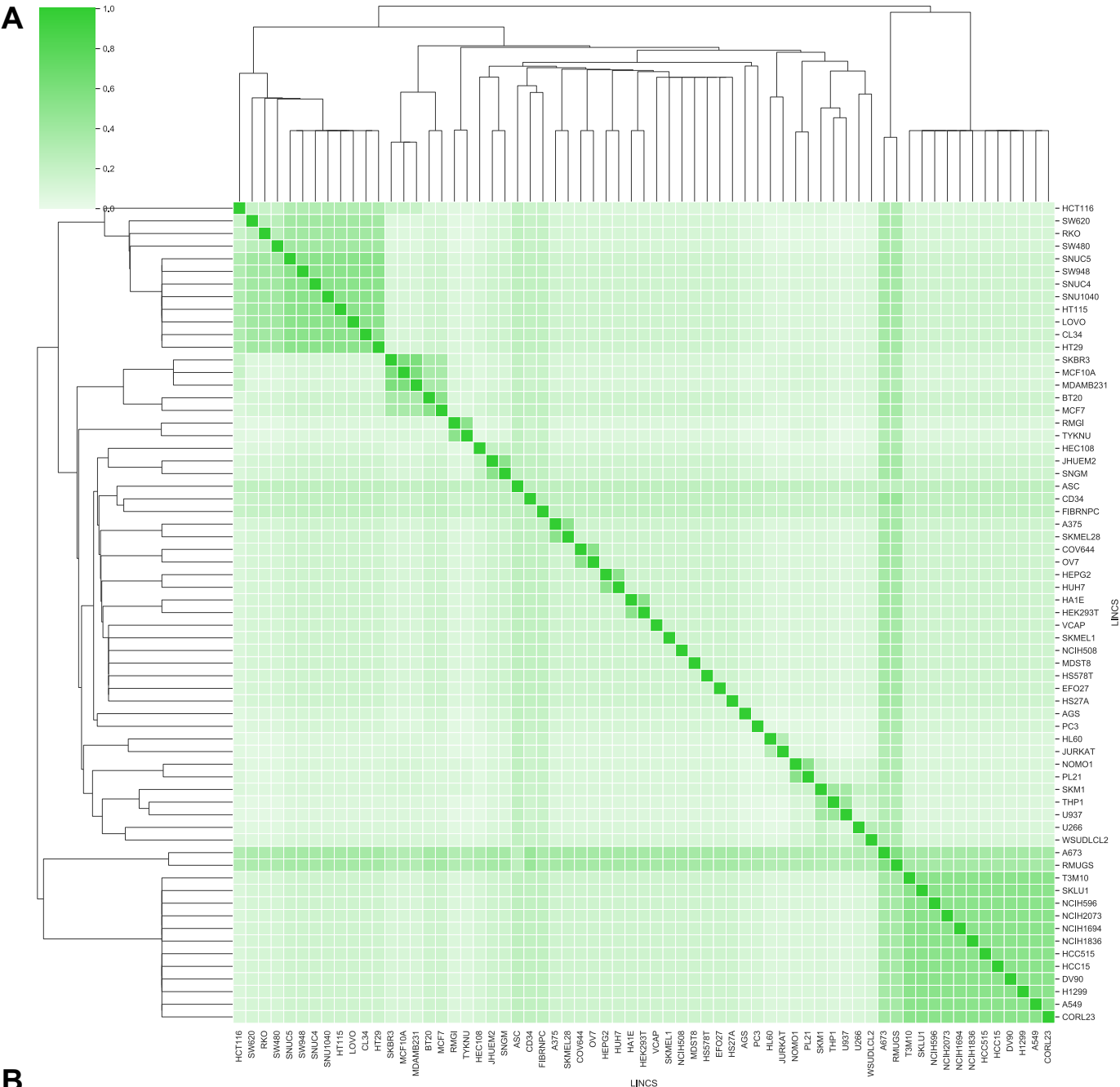


Figure S12



A

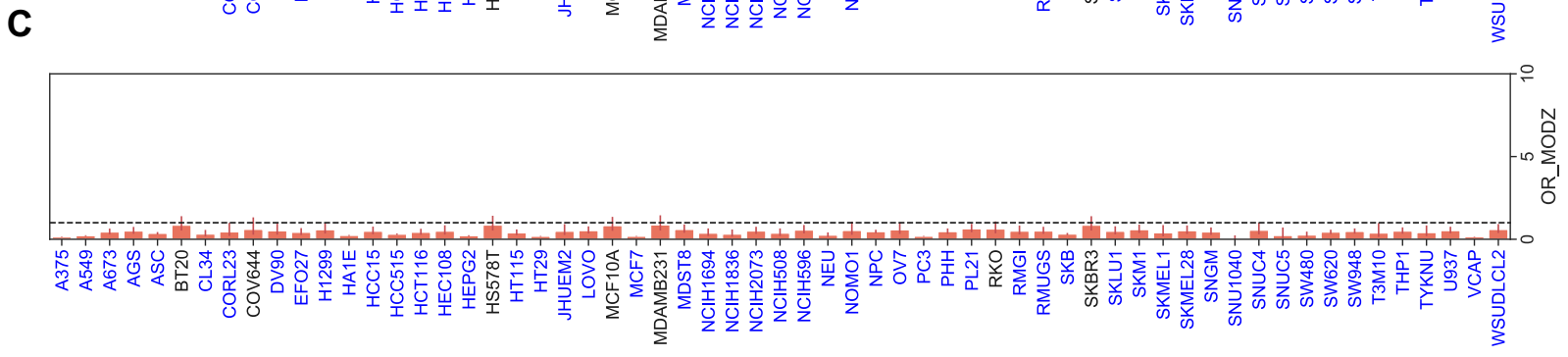


Figure S14

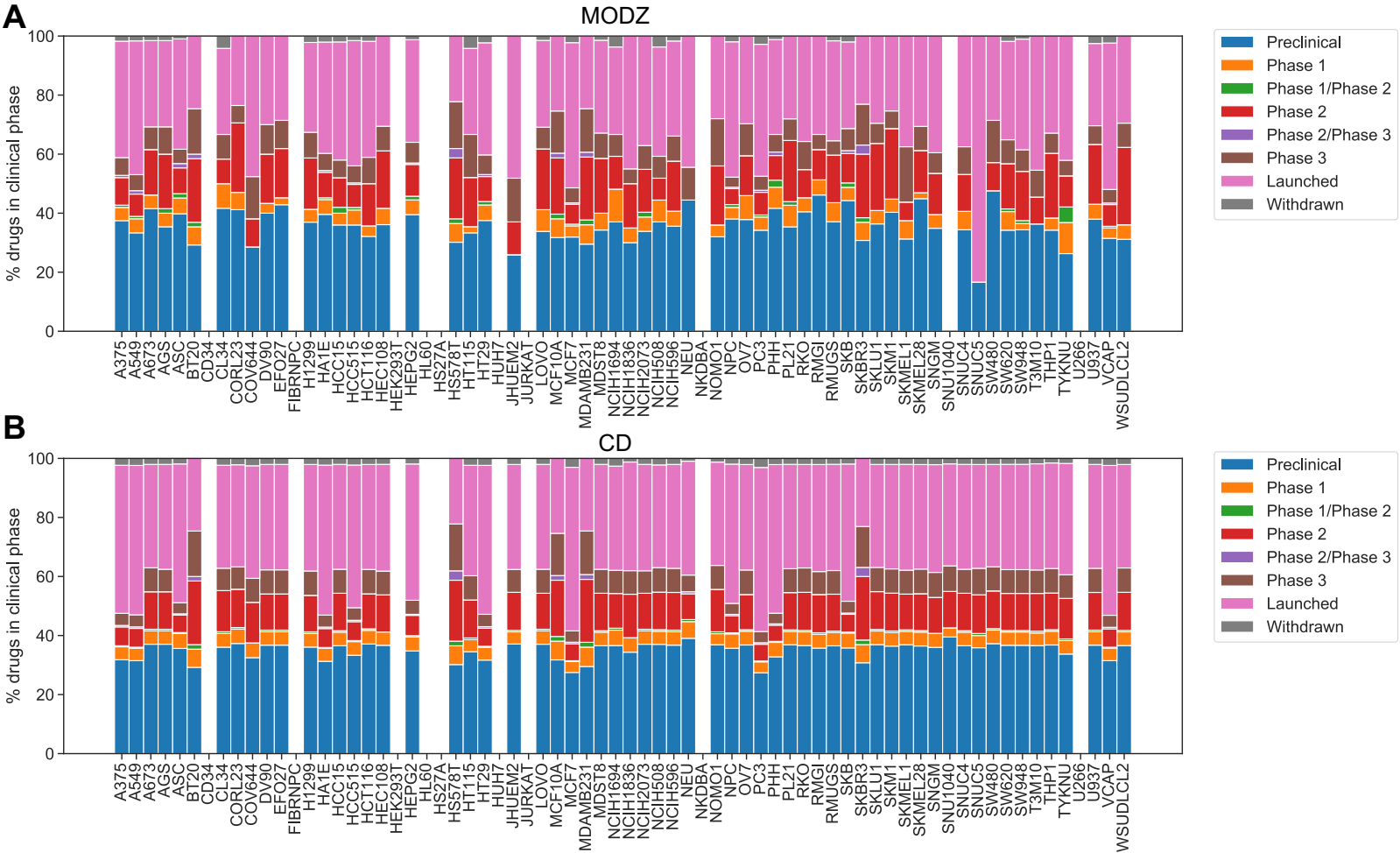
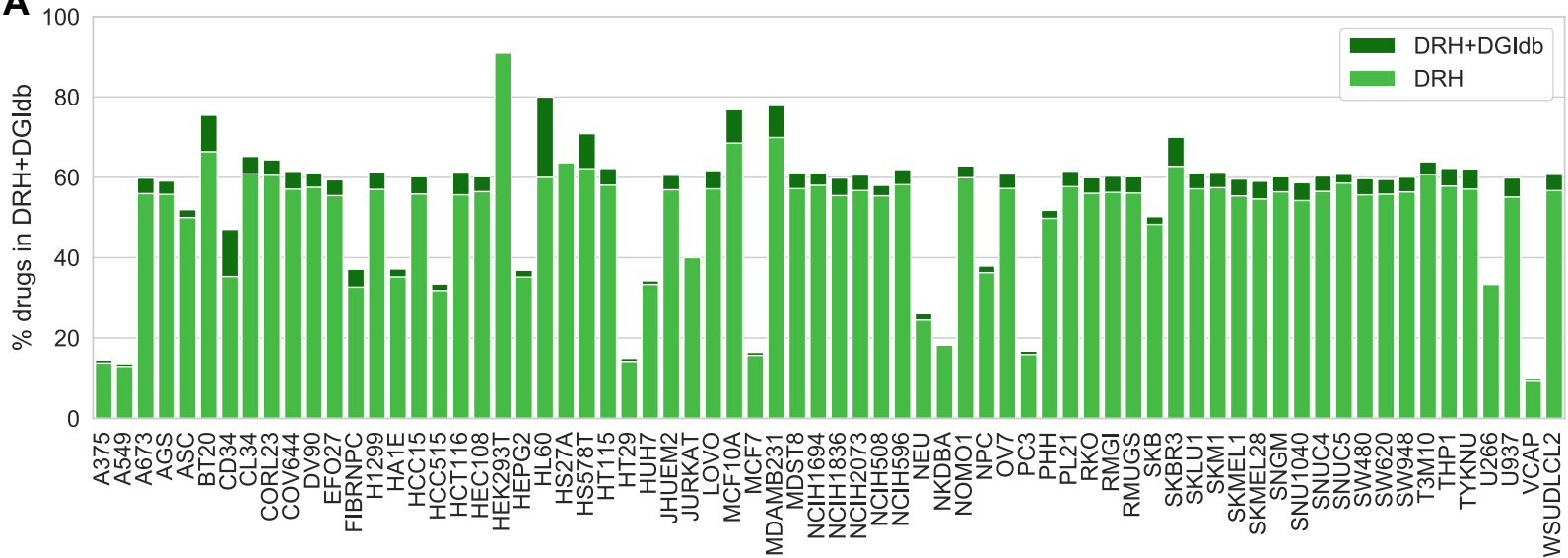
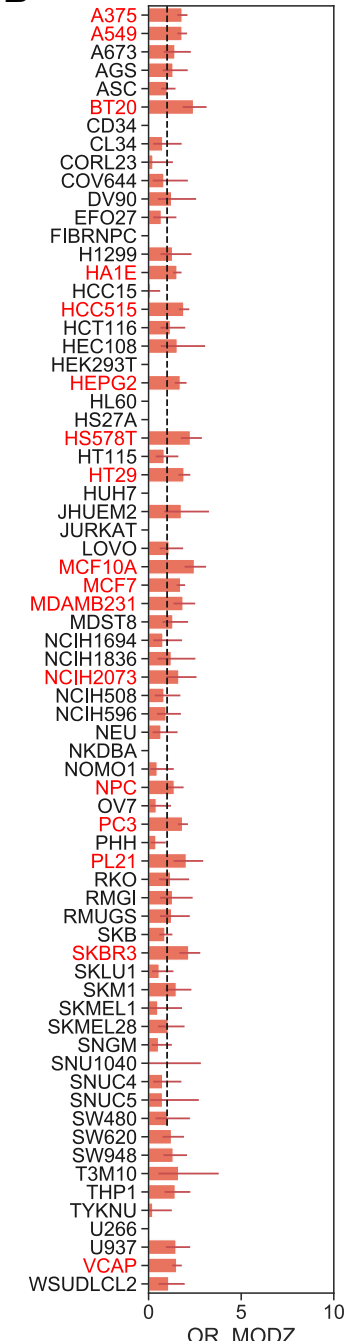


Figure S15

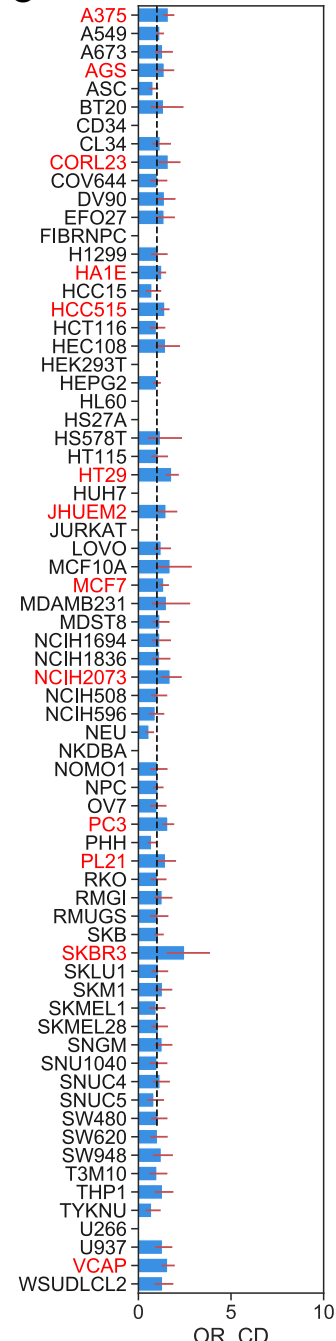
A



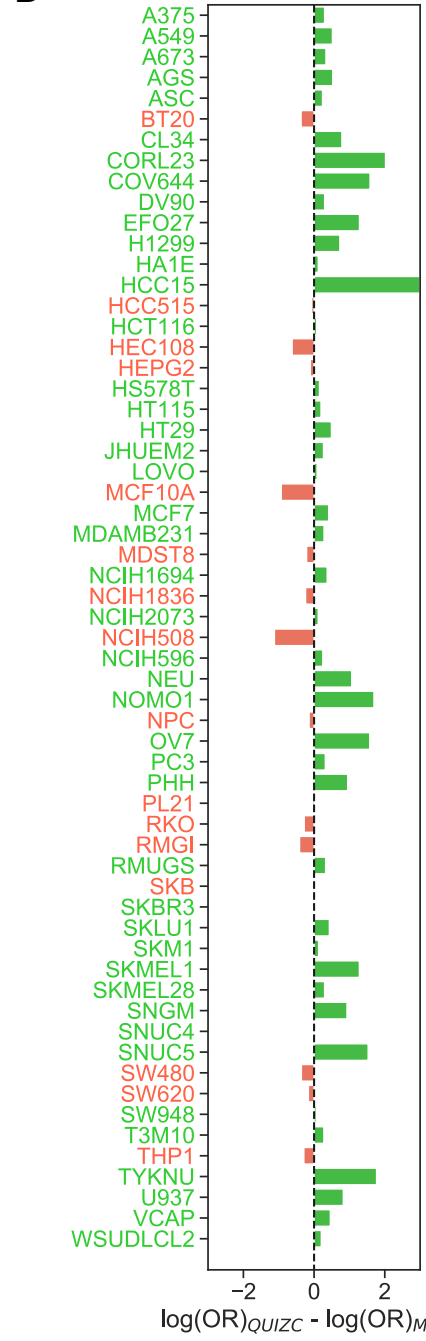
B



C



D



E

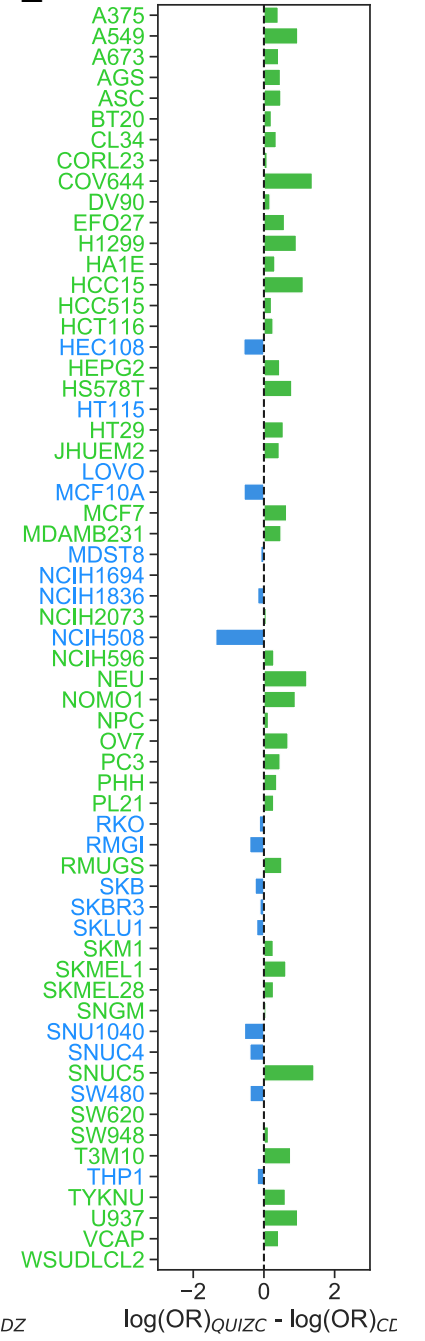


Figure S16

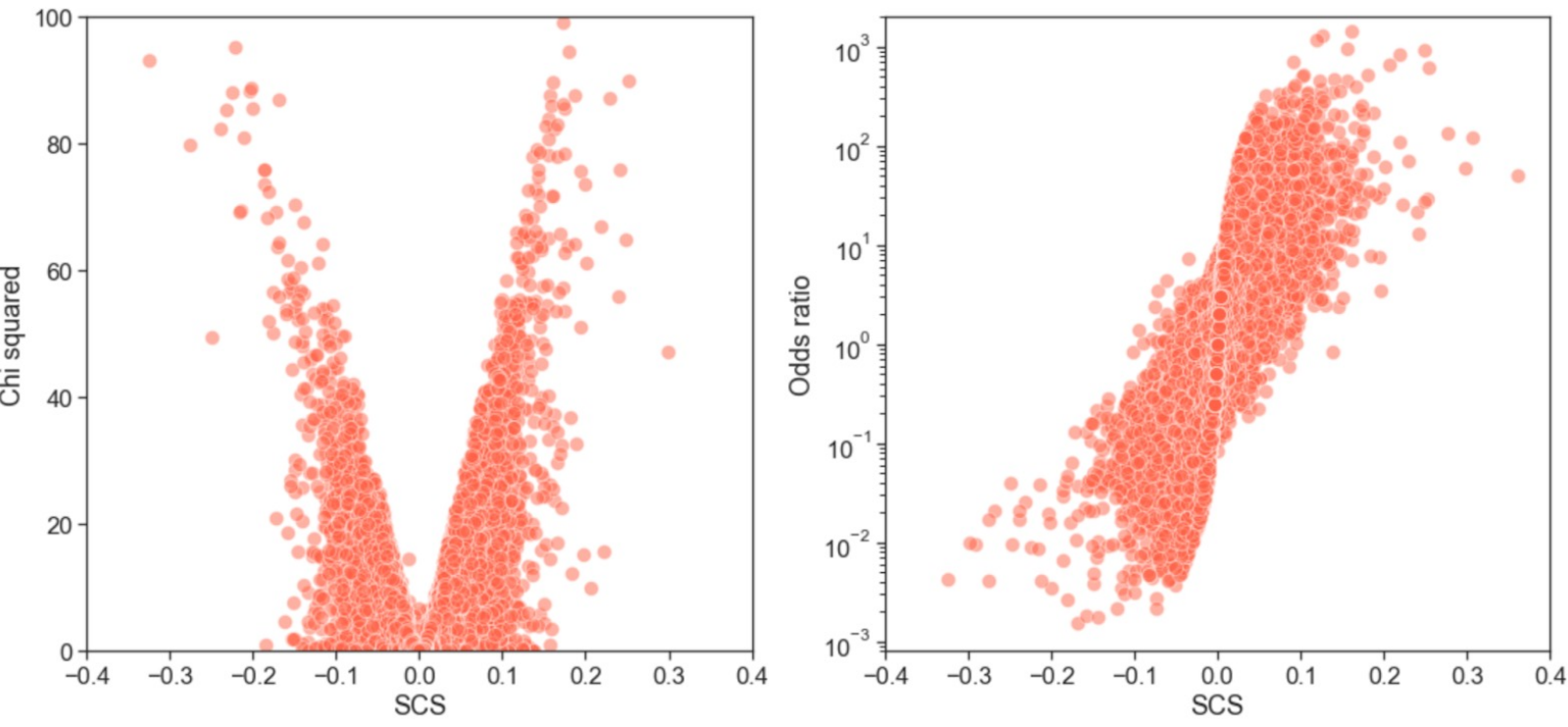


Figure S17

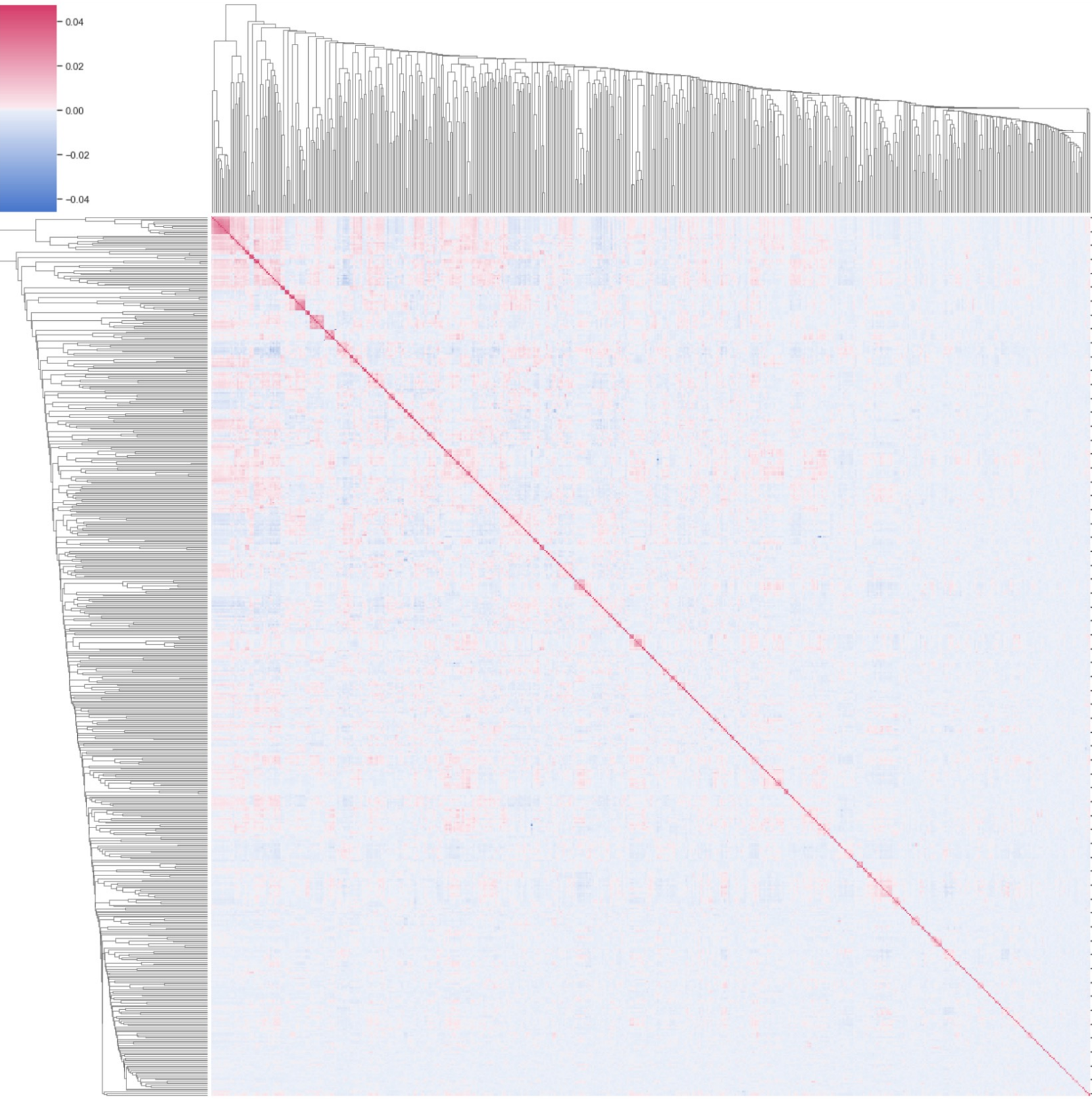
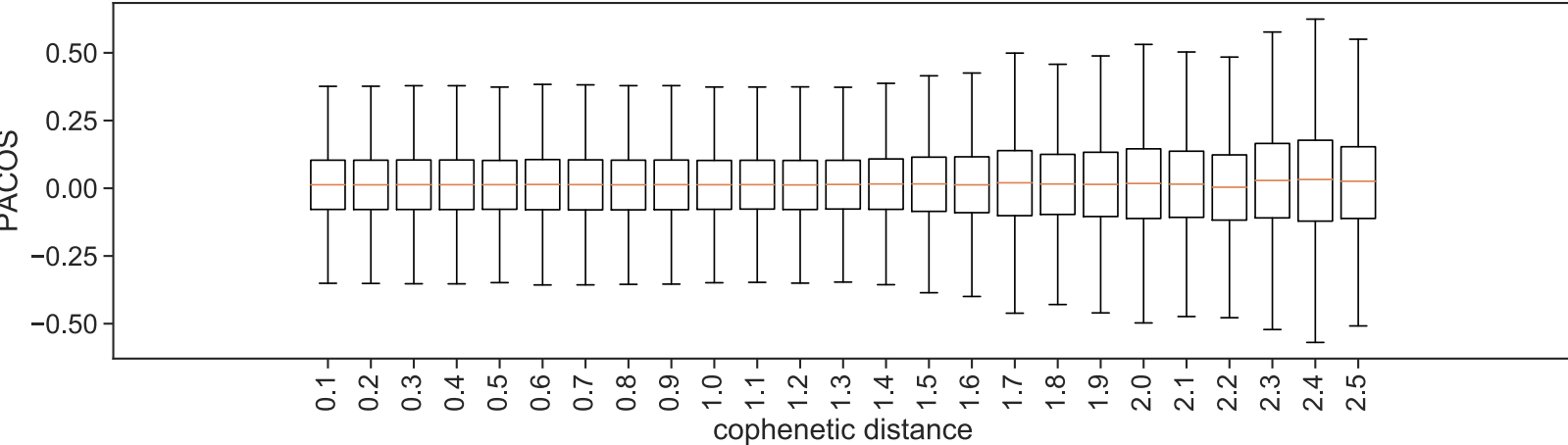
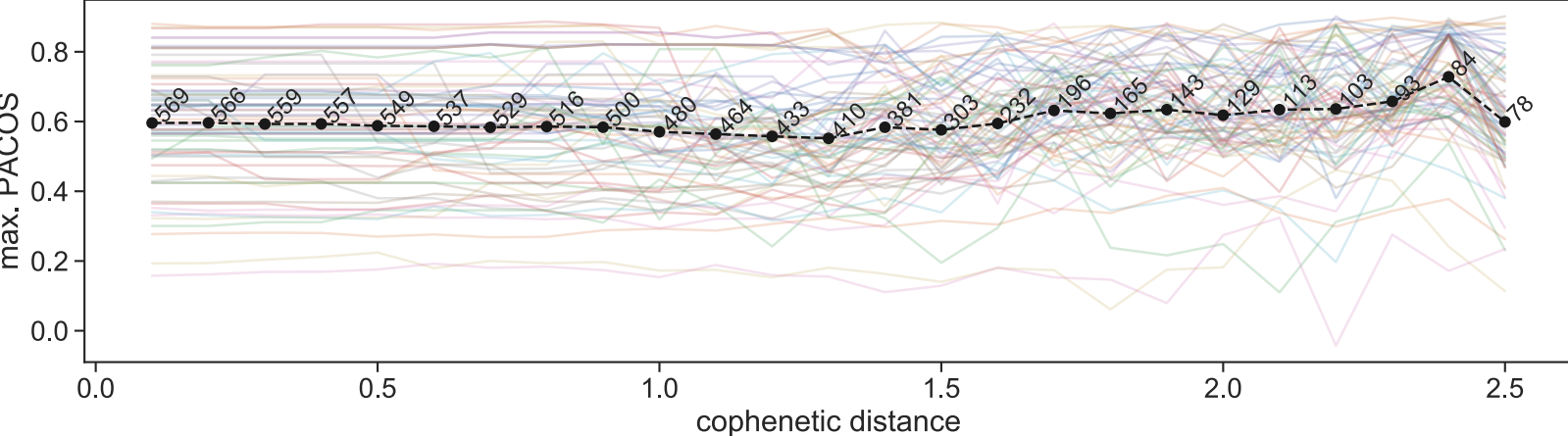


Figure S18

A



B



C

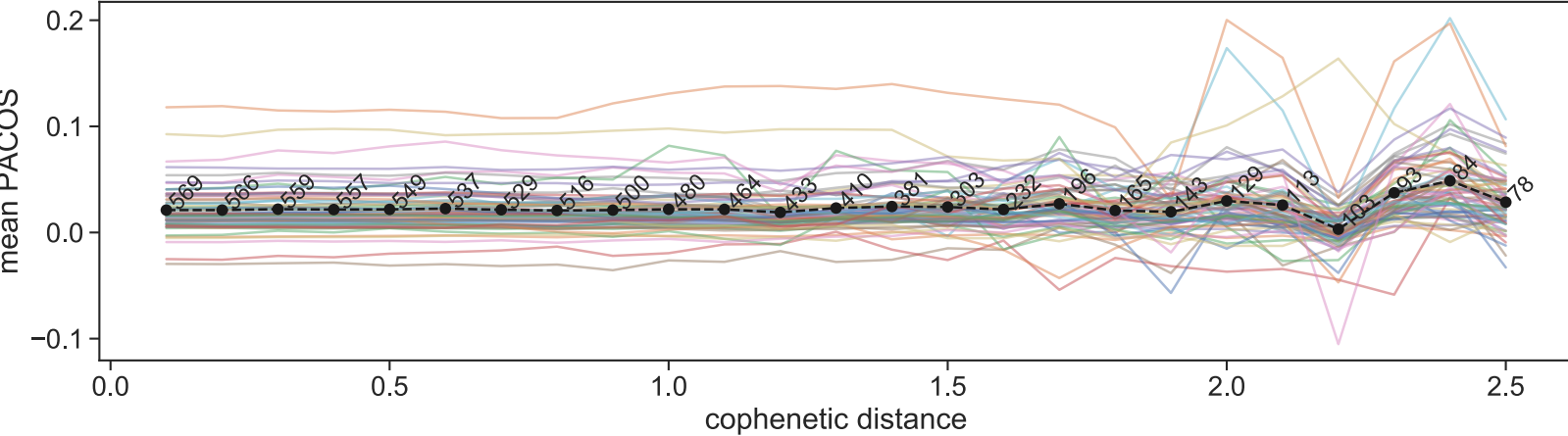
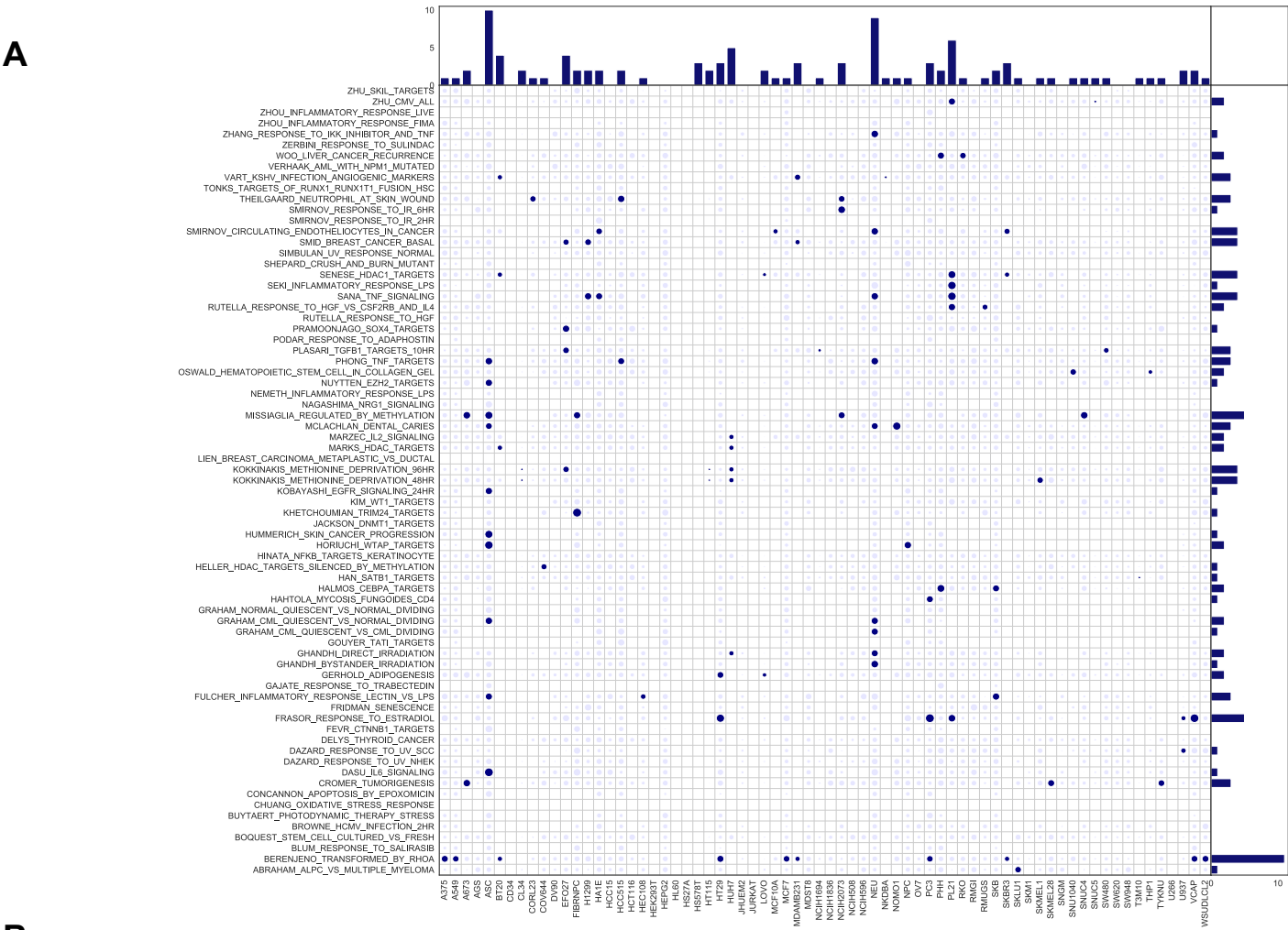


Figure S19

A



B

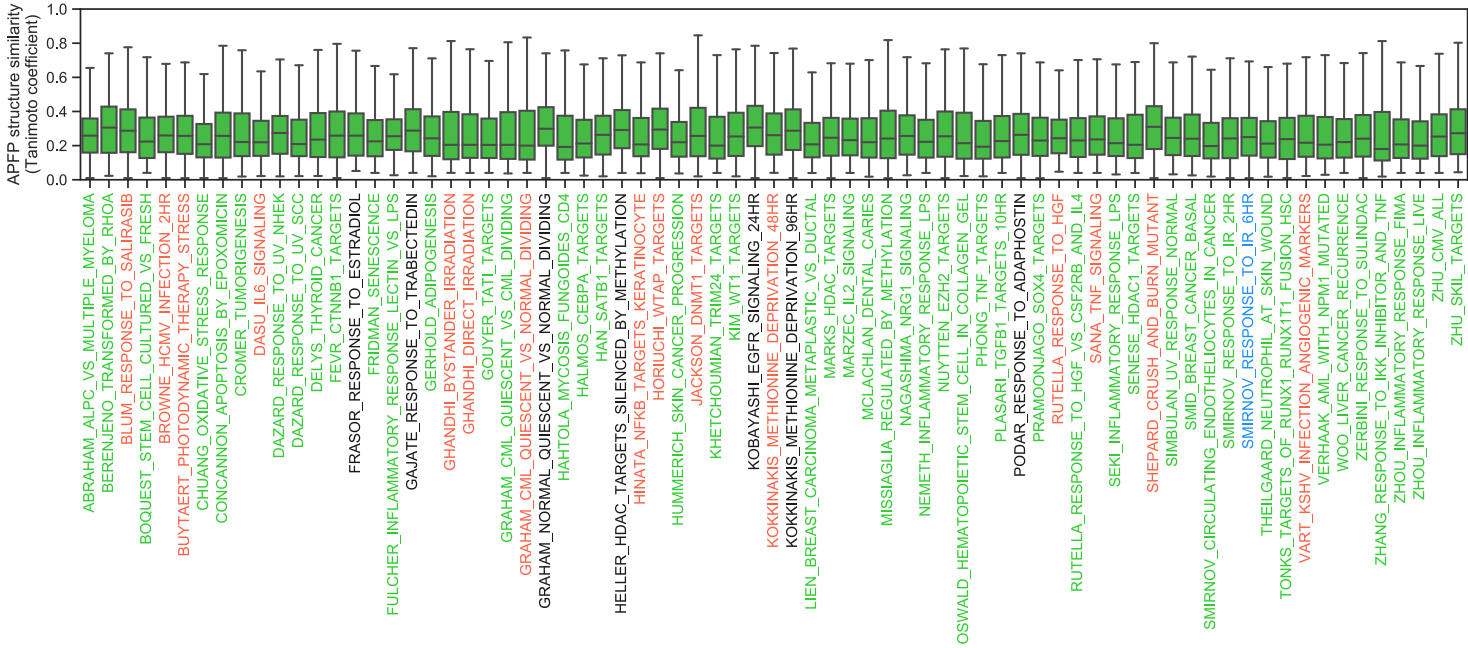


A

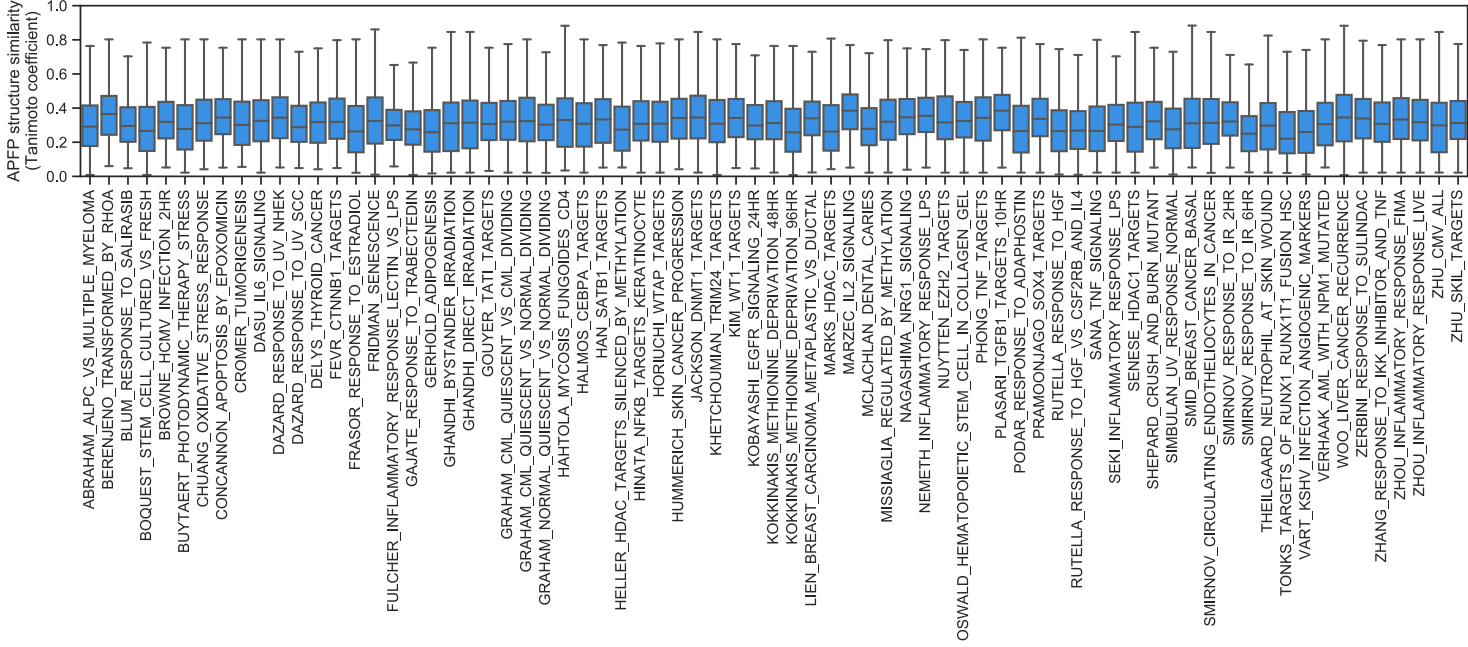


Figure S21

A



B



C

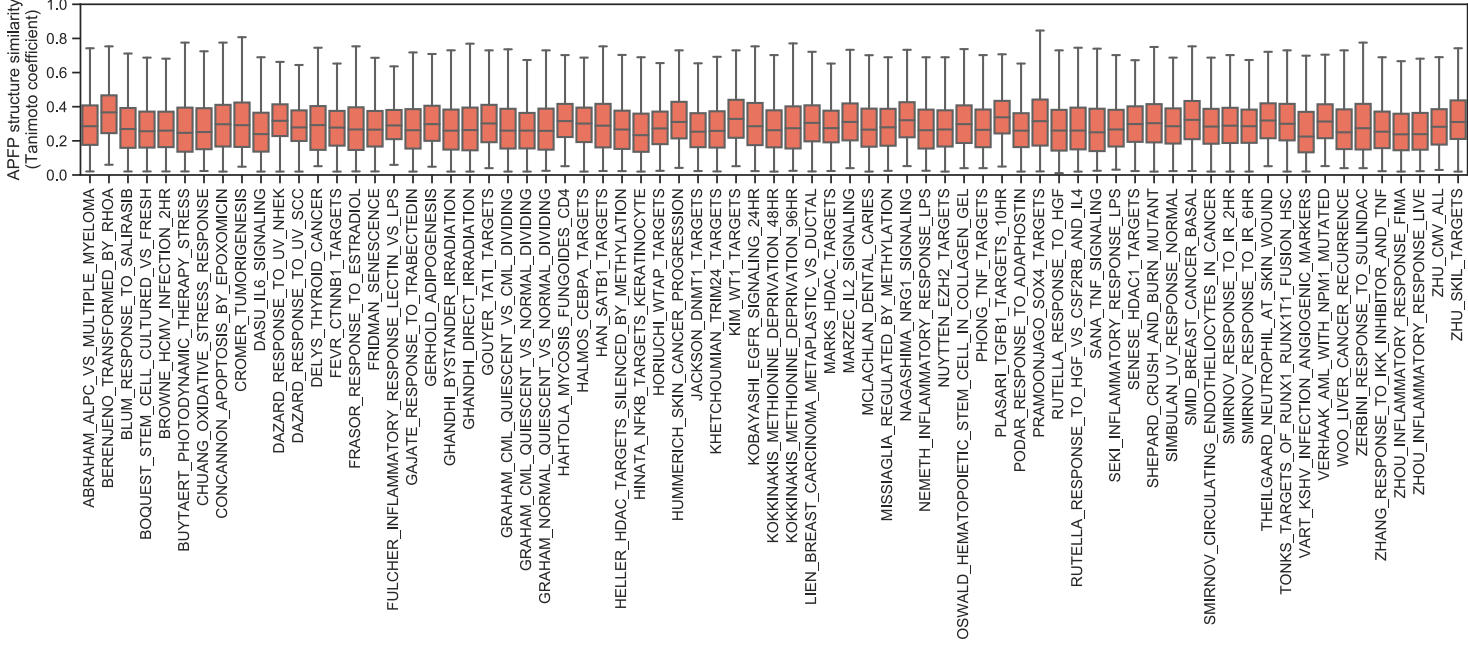
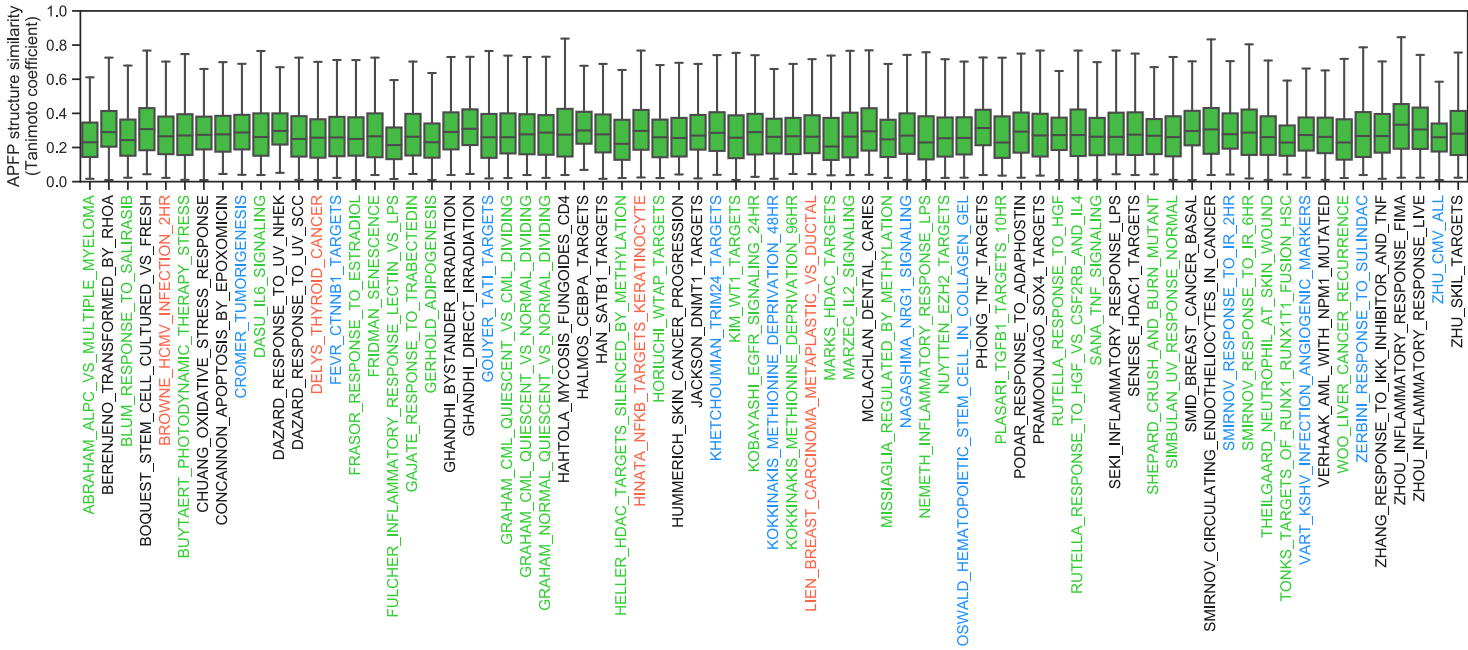
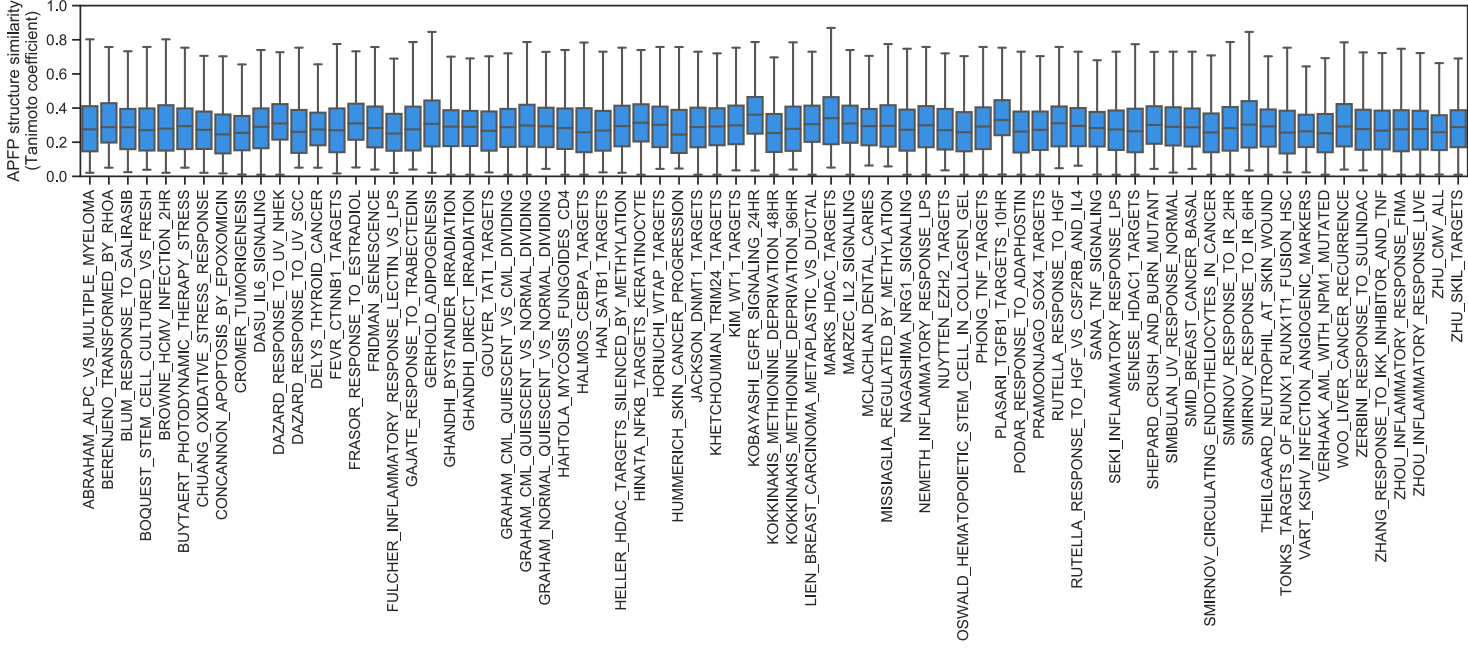


Figure S22

A



B



C

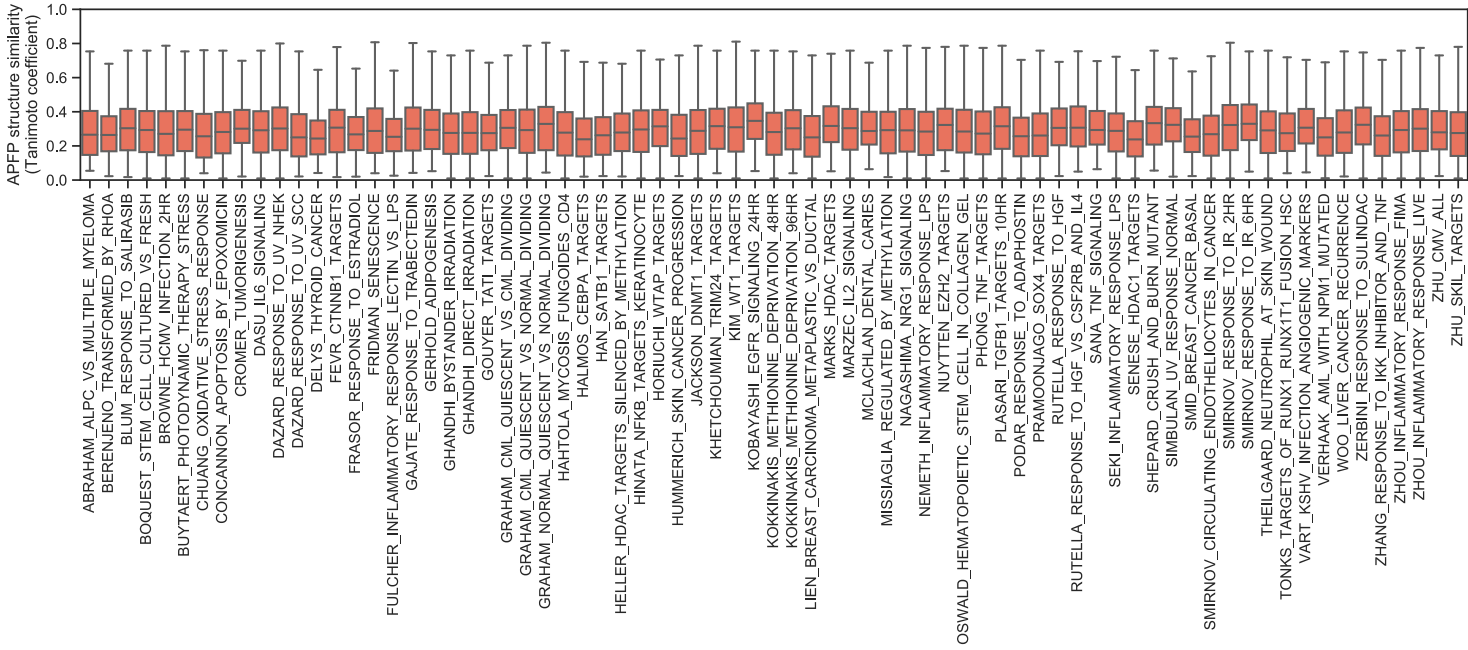


Figure S23

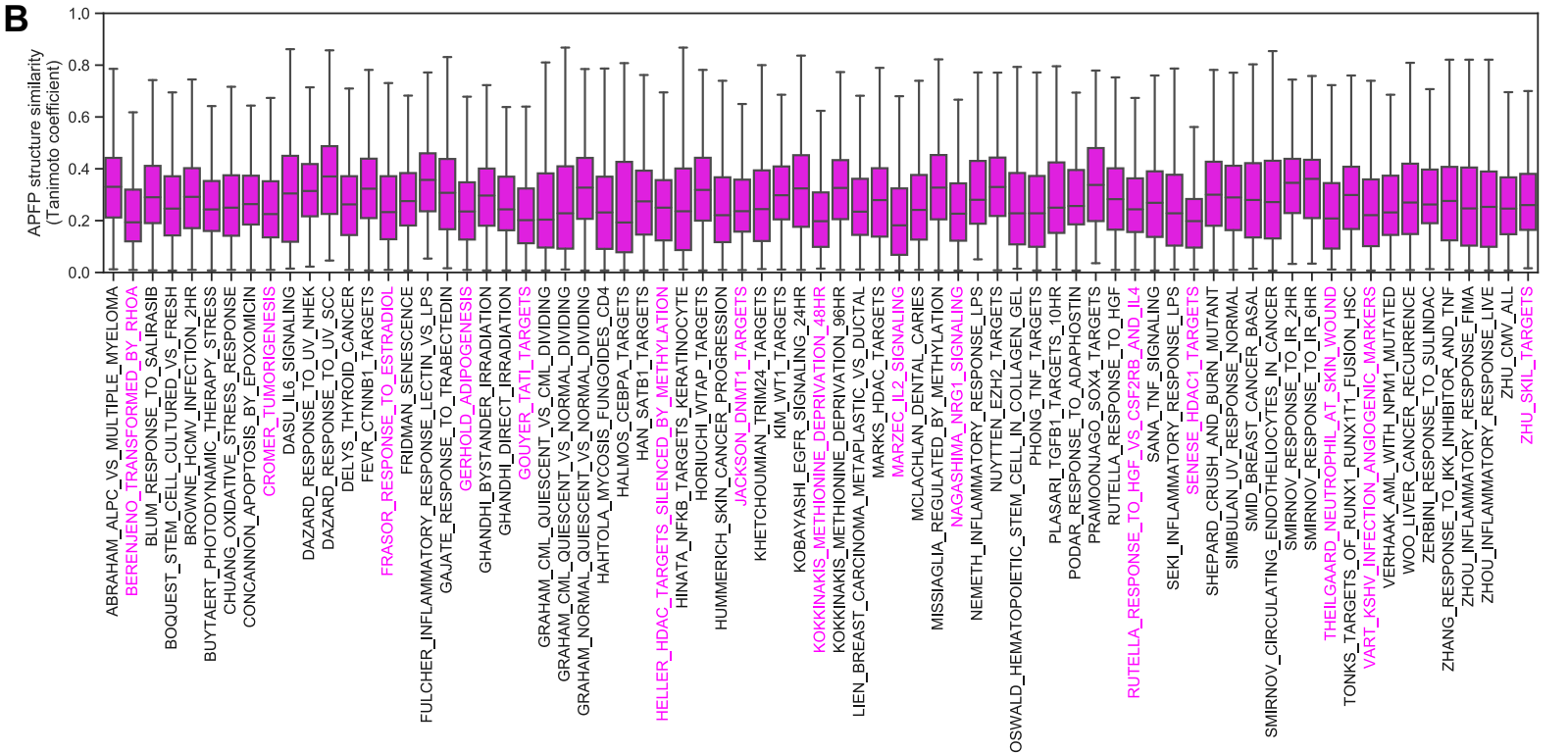
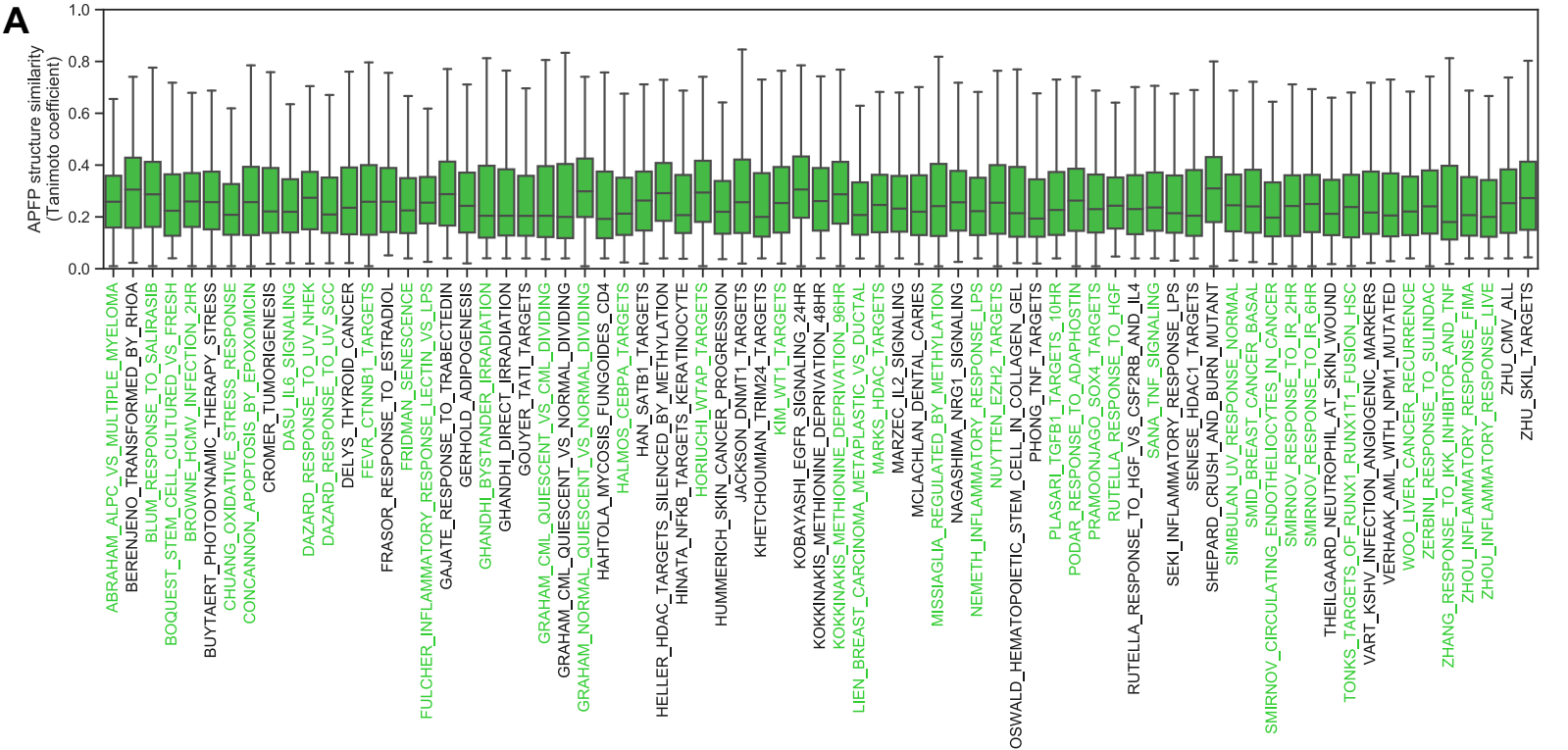


Figure S24

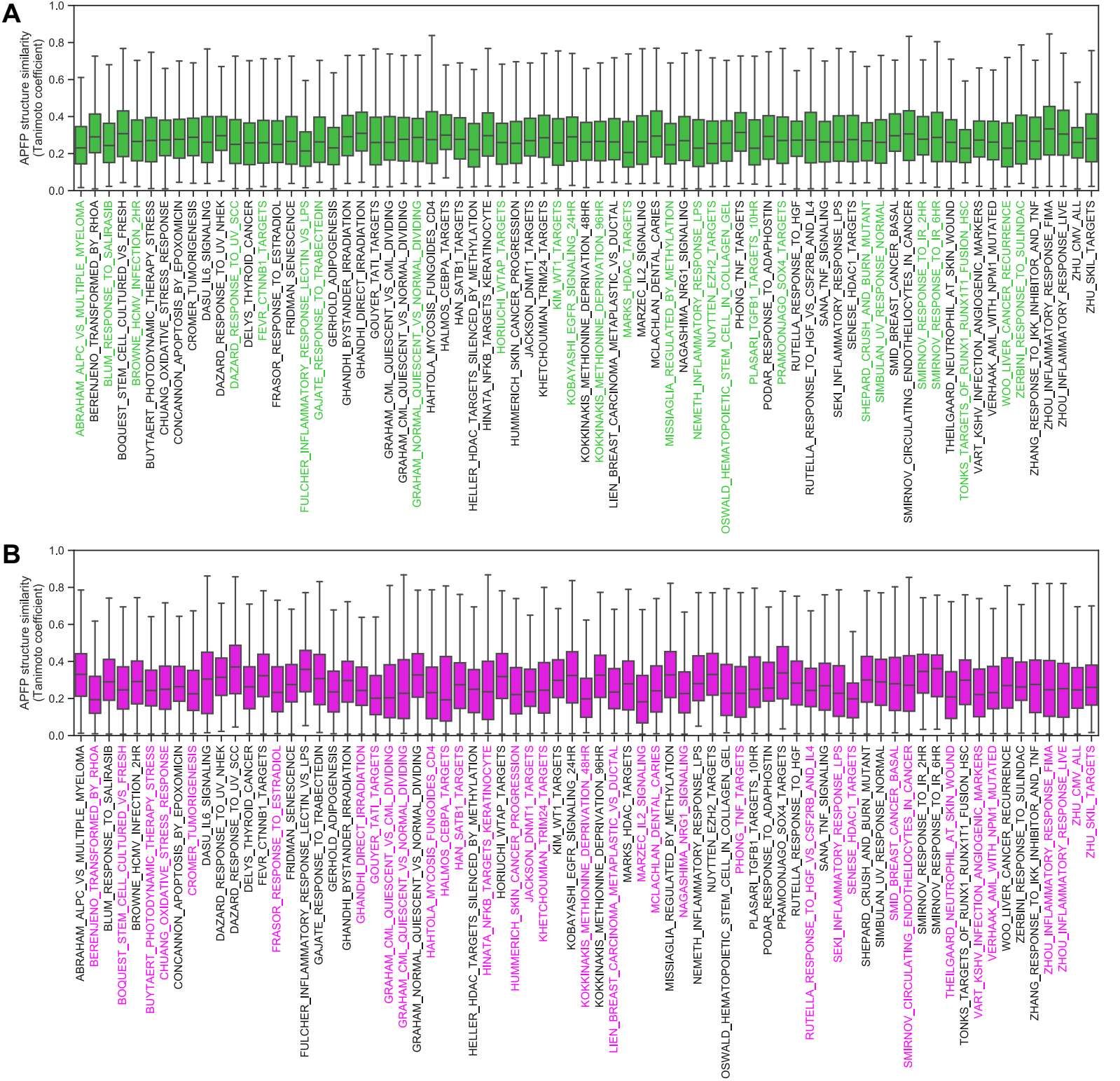


Figure S25

Tyrphostin-AG-1478



Cyclosporin-a



Figure S25 (cont'd)

simvastatin



neratinib



Figure S25 (cont'd)

GW-405833

TG-101348



Figure S26

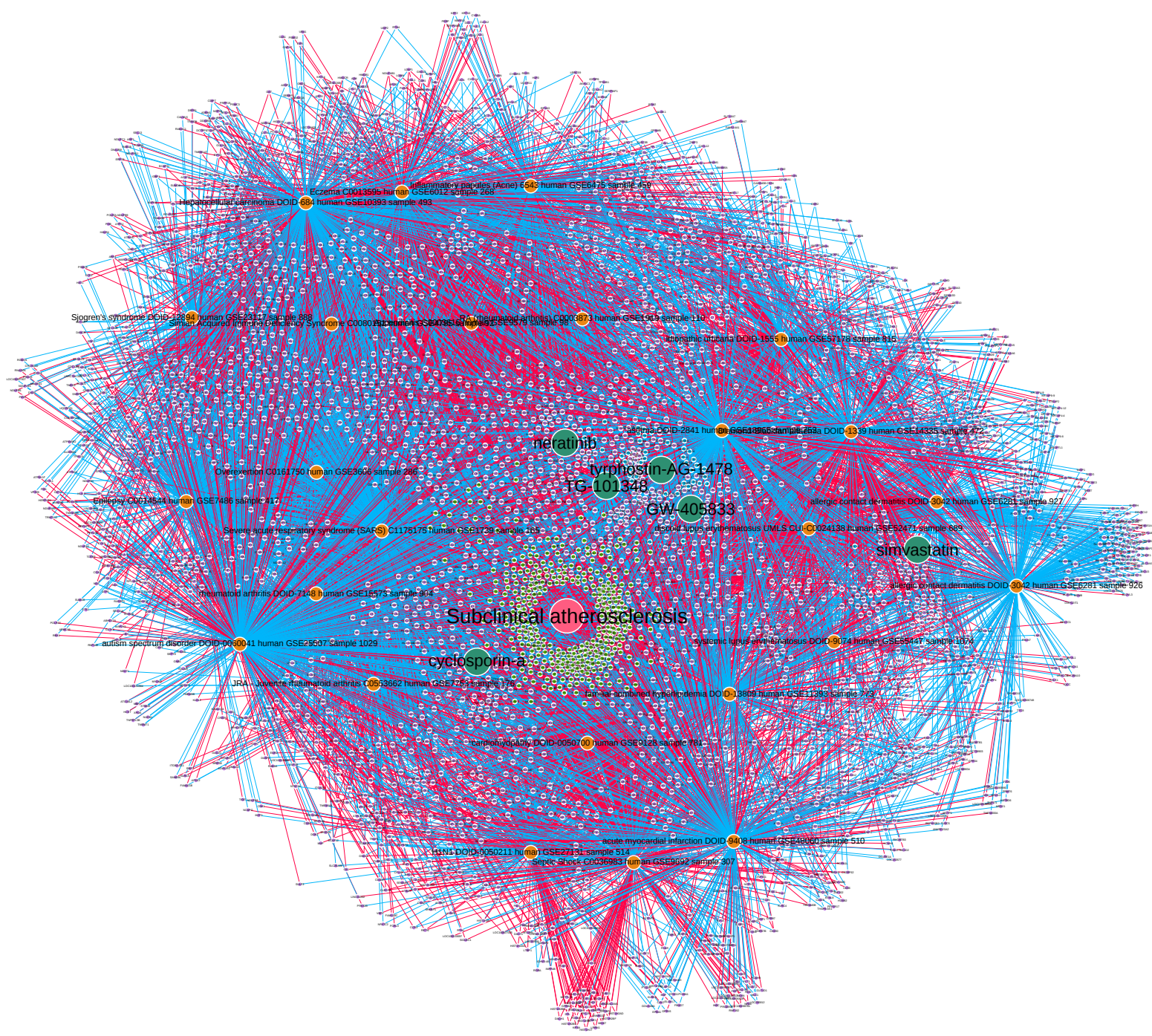


Figure S27

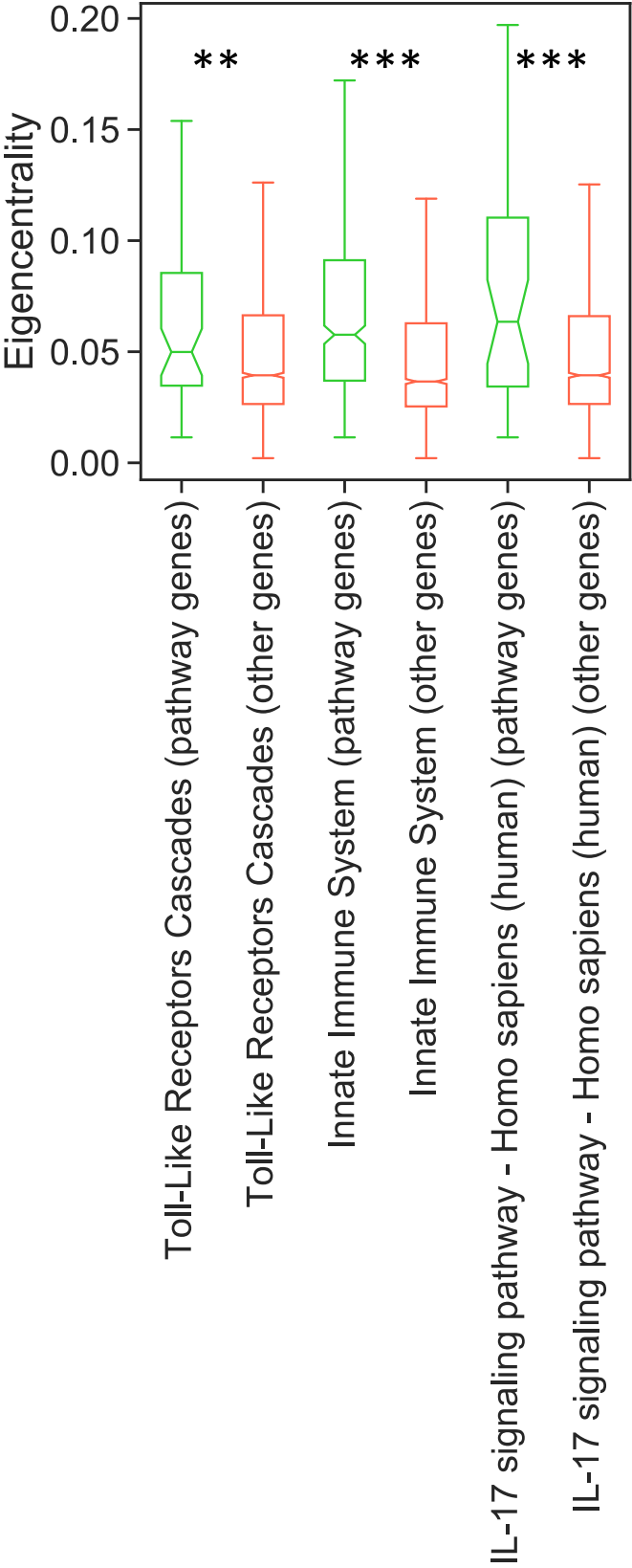


Figure S28

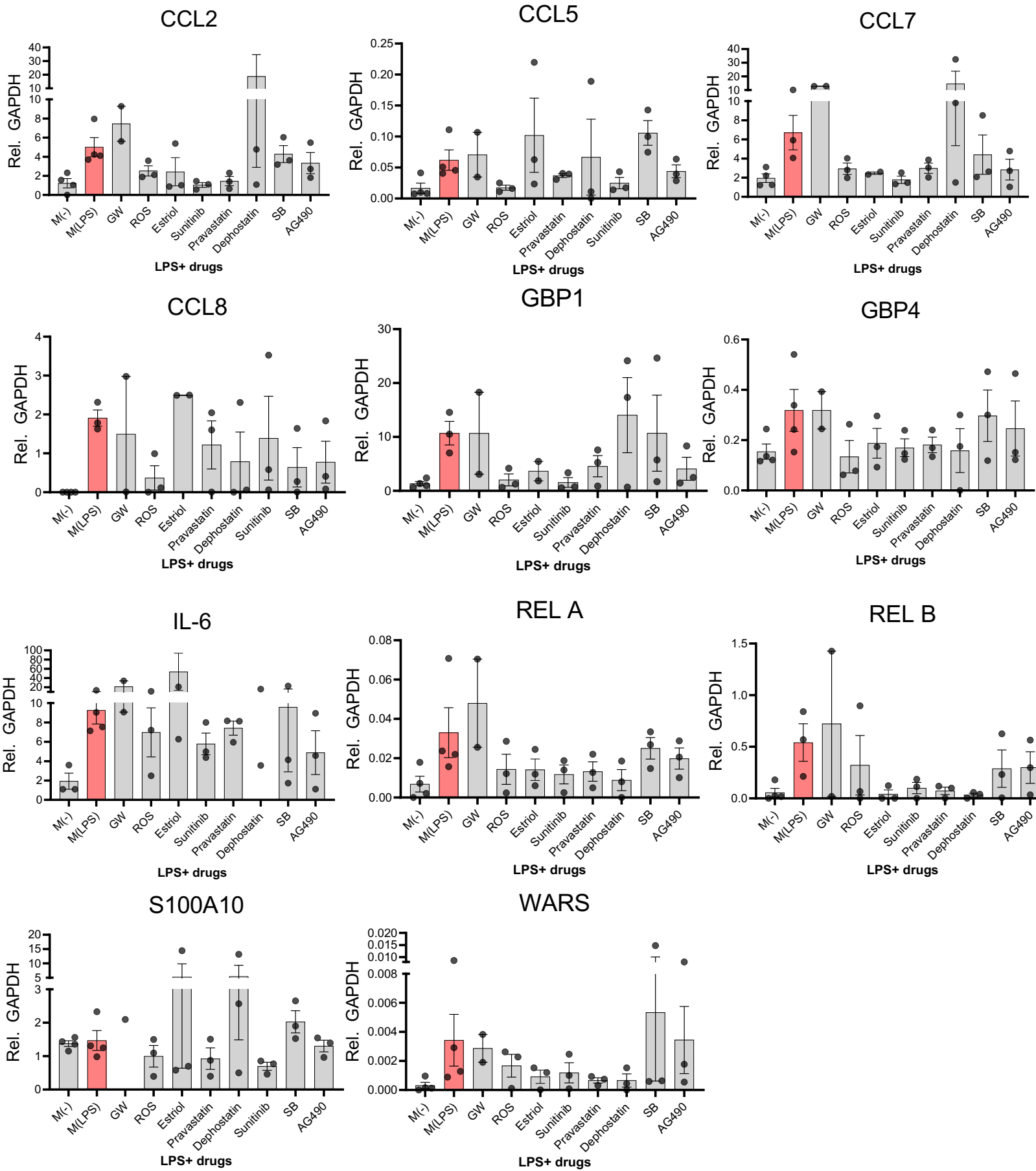


Figure S29

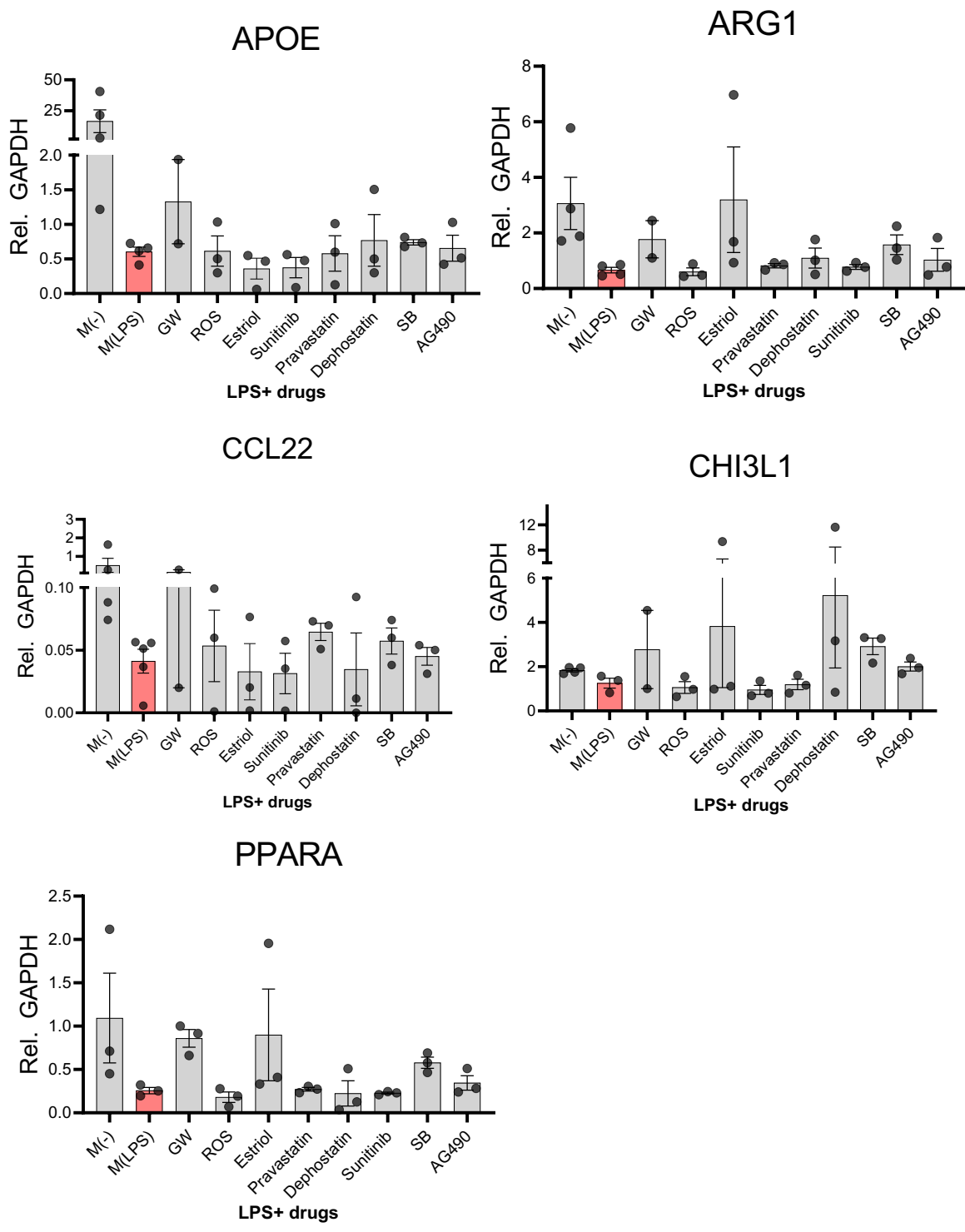


Table S1

Geneset name (page 1/2)
ABRAHAM_ALPC_VS_MULTIPLE_MYELOMA
BERENJENO_TRANSFORMED_BY_RHOA
BLUM_RESPONSE_TO_SALIRASIB
BOQUEST_STEM_CELL_CULTURED_VS_FRESH
BROWNE_HCMV_INFECTION_2HR
BUYTAERT_PHOTODYNAMIC_THERAPY_STRESS
CHUANG_OXIDATIVE_STRESS_RESPONSE
CONCANNON_APOPTOSIS_BY_EPOXOMICIN
CROMER_TUMORIGENESIS
DASU_IL6_SIGNALING
DAZARD_RESPONSE_TO_UV_NHEK
DAZARD_RESPONSE_TO_UV_SCC
DELYS_THYROID_CANCER
FEVR_CTNNB1_TARGETS
FRASOR_RESPONSE_TO ESTRADIOL
FRIDMAN_SENESCENCE
FULCHER_INFLAMMATORY_RESPONSE_LECTIN_VS_LPS
GAJATE_RESPONSE_TO TRABECTEDIN
GERHOLD_ADIPOGENESIS
GHANDHI_BYSTANDER_IRRADIATION
GHANDHI_DIRECT_IRRADIATION
GOUYER_TATI_TARGETS
GRAHAM_CML_QUIESCENT_VS_CML_DIVIDING
GRAHAM_CML_QUIESCENT_VS_NORMAL_DIVIDING
GRAHAM_NORMAL_QUIESCENT_VS_NORMAL_DIVIDING
HAHTOLA_MYCOSIS_FUNGOIDES_CD4
HALMOS_CEBPA_TARGETS
HAN_SATB1_TARGETS
HELLER_HDAC_TARGETS_SILENCED_BY_METHYLATION
HINATA_NFKB_TARGETS_KERATINOCYTE
HORIUCHI_WTAP_TARGETS
HUMMERICH_SKIN_CANCER_PROGRESSION
JACKSON_DNMT1_TARGETS
KHETCHOUMIAN_TRIM24_TARGETS
KIM_WT1_TARGETS
KOBAYASHI_EGFR_SIGNALING_24HR
KOKKINAKIS_METHIONINE_DEPRIVATION_48HR
KOKKINAKIS_METHIONINE_DEPRIVATION_96HR
LIEN_BREAST_CARCINOMA_METAPLASTIC_VS_DUCTAL
MARKS_HDAC_TARGETS
MARZEC_IL2_SIGNALING
MCLACHLAN_DENTAL_CARIES
MISSIAGLIA_REGULATED_BY_METHYLATION
NAGASHIMA_NRG1_SIGNALING
NEMETH_INFLAMMATORY_RESPONSE_LPS
NUYTEN_EZH2_TARGETS
OSWALD_HEMATOPOIETIC_STEM_CELL_IN_COLLAGEN_GEL
PHONG_TNF_TARGETS
PLASARI_TGFB1_TARGETS_10HR
PODAR_RESPONSE_TO_ADAPHOSTIN
PRAMOONJAGO_SOX4_TARGETS
RUTELLA_RESPONSE_TO_HGF
RUTELLA_RESPONSE_TO_HGF_VS_CSF2RB_AND_IL4
SANA_TNF_SIGNALING
SEKI_INFLAMMATORY_RESPONSE_LPS
SENESE_HDAC1_TARGETS
SHEPARD_CRUSH_AND_BURN_MUTANT
SIMBULAN_UV_RESPONSE_NORMAL
SMID_BREAST_CANCER_BASAL
SMIRNOV_CIRCULATING_ENDOTHELIOCYTES_IN_CANCER
SMIRNOV_RESPONSE_TO_IR_2HR
SMIRNOV_RESPONSE_TO_IR_6HR

Table S1 (cont'd)

Geneset name (page 2/2)
THEILGAARD_NEUTROPHIL_AT_SKIN_WOUND
TONKS_TARGETS_OF_RUNX1_RUNX1T1_FUSION_HSC
VART_KSHV_INFECTION_ANGIOGENIC_MARKERS
VERHAAK_AML_WITH_NPM1_MUTATED
WOO_LIVER_CANCER_RECURRENCE
ZERBINI_RESPONSE_TO_SULINDAC
ZHANG_RESPONSE_TO_IKK_INHIBITOR_AND_TNF
ZHOU_INFLAMMATORY_RESPONSE_FIMA
ZHOU_INFLAMMATORY_RESPONSE_LIVE
ZHU_CMV_ALL
ZHU_SKIL_TARGETS

Table S2

	high_peak_genes (%)	all_zeros_genes (%)	KDE_zero_bw_genes (%)
A673	0.18877298	0	0.01987084
AGS	0.10447137	0	0.05223569
ASC	0.4625497	0	0.08114907
BT20	0.63761955	0	0.04723108
CD34	0.11893749	0	0.03964583
CL34	0.11953923	0	0.0108672
CORL23	0.1193114	0	0.01704449
COV644	0.42452029	0	0.06792325
DV90	0.09045096	0	0.01292157
EFO27	0.16137708	0	0.02151694
FIBRNPC	0.70621469	0	0.15785975
H1299	0.20338121	0	0.10169061
HCC15	0.16778523	0	0.01973944
HCT116	0.27782923	0	0.0370439
HEC108	0.66318328	0	0.06028939
HEK293T	0.11357183	0	0
HL60	0.10884777	0	0.01554968
HS27A	1.3986014	0	0.52447552
HS578T	0.88395712	0	0.15046079
HT115	0.07971303	0	0.03985652
HUH7	0.37071362	0	0.07414272
JHUEM2	0.17114271	0	0.02632965
JURKAT	0.44228218	0	0.08845644
LOVO	0.09310024	0	0
MCF10A	0.2608162	0	0
MDAMB231	0.3109729	0	0.03331853
MDST8	0.19379845	0	0.09179927
NCIH1694	0.07465472	0	0.04976981
NCIH1836	0.30105973	0	0.08429672
NCIH2073	0.28537689	0	0.10824641
NCIH508	0.19900498	0	0
NCIH596	0.1484781	0	0.03711952
NEU	1.42191714	0	0.28601781
NKDBA	0.52483235	0	0.07775294
NOMO1	0.25502975	0	0.05667328
OV7	0.36671368	0	0.09873061
PHH	1.13867426	0	0.1301342
PL21	0.11418784	0	0.01903131
RKO	0.34526413	0	0.10357924
RMGI	0.18824211	0	0.02896032
RMUGS	0.2295253	0	0
SKB	0.64987815	0	0.12997563
SKBR3	0.43887957	0	0.01290822
SKLU1	0.46645138	0	0.09568233
SKM1	0.14272122	0	0
SKMEL1	0.270027	0	0.0150015
SKMEL28	0.10877842	0	0.04351137
SNGM	0.18518519	0	0.12345679
SNU1040	0.99330598	0	0.12956165
SNUC4	0.41728362	0	0.04038229
SNUC5	0.04546488	0	0
SW480	0.23020258	0	0.06138735
SW620	0.23090879	0	0.04123371
SW948	0.57798695	0	0.08256956
T3M10	0.16027617	0	0.03698681
THP1	0.20783107	0	0.00831324
TYKNU	0.2920871	0	0
U266	0.68965517	0	0.22988506
U937	0.15273394	0	0.02036452
WSUDLCL2	0.08097166	0	0.03470214

Table S3

	high_peak_genes_	allexcept_high_peak_genes	high_peak_genes_MW	all_zeros_genes_	allexcept_all_zeros_genes	all_zeros_genes_MW	KDE_zero_bw_genes	allexcept_KDE_zero_bw_gene	KDE_zero_bw_genes_MW
	mean	mean	U_pval	mean	_mean	_pval	mean	s_mean	U_pval
A673	0.0081232	0.00852029	0.45100953	0.00851954	0	0.00643087	0.00851996	0.22876689	
AGS	0.00515152	0.00683292	0.21757643	0.00683116	0	0.00848485	0.00683303	0.08813223	
ASC	0.01084085	0.0120211	0.00828131	0.01201564	0	0.00896432	0.01201812	0.02215591	
BT20	0.01683502	0.01715335	0.35295491	0.01715132	0	0.01363636	0.01715298	0.04104807	
CD34	0.06535948	0.07196893	0.29600239	0.07196107	0	0.05882353	0.07196628	0.160979	
CL34	0.01185771	0.01010677	0.08555112	0.01010886	0	0.01086957	0.01010878	0.12610388	
CORL23	0.00492611	0.00552548	0.04873894	0.00552476	0	0.00383142	0.00552505	0.44604985	
COV644	0.00672566	0.00592184	0.19693183	0.00592525	0	0.00553097	0.00592552	0.38620338	
DV90	0.00344234	0.00481336	0.02418376	0.00481212	0	0.00903614	0.00481157	0.0138139	
EFO27	0.00484848	0.00618959	0.30625909	0.00618743	0	0.00909091	0.0061868	0.03940261	
FIBRNPC	0.01084798	0.00873449	3.47E-05	0.00874942	0	0.00869056	0.00874951	0.18876279	
H1299	0.00623053	0.00541709	0.0272025	0.00541875	0	0.00584112	0.00541832	0.38416544	
HCC15	0.00526507	0.00722829	0.03970148	0.007225	0	0.01080247	0.00722429	0.00256158	
HCT116	0.00943396	0.00834058	0.00717357	0.00834361	0	0.00707547	0.00834408	0.40344182	
HEC108	0.005557	0.00643467	3.11E-23	0.00642885	0	0.01002865	0.00642668	0.0079508	
HEK293T	0.09090909	0.10512171	0.03390329	0.10510557	0	0.10510557	0.10510557	0.064547	
HL60	0.15714286	0.14360212	0.23271297	0.14361686	0	0.3	0.14359253	0.06062379	
HS27A	0.09090909	0.09123146	3.10E-11	0.09122695	0.0016283	0.12121212	0.09106886	2.39E-08	
HS578T	0.02354885	0.01784602	0.44064014	0.01789643	0	0.02063107	0.01789231	0.18766	
HT115	0.01382488	0.00959893	0.41944946	0.0096023	0	0.00806452	0.00960291	0.14528538	
HUH7	0.0222973	0.02603562	0.08465668	0.02602176	0	0.02477477	0.02602268	0.38499923	
JHUEM2	0.00379555	0.00525934	0.43595881	0.00525684	0	0.00328947	0.00525735	0.22145711	
JURKAT	0.22	0.21652599	8.84E-15	0.21654135	0	0.2	0.216556	0.20731738	
LOVO	0.00505051	0.00752219	0.08805311	0.00751989	0	0.005989	0.00751989	0.00921609	
MCF10A	0.0288671	0.01612857	0.31297171	0.0161618	0	0.0161618	0.0161618	0.00839739	
MDAMB231	0.0199115	0.02458572	0.19489009	0.02457118	0	0.01474926	0.02457446	0.07928082	
MDST8	0.00820859	0.007161	0.00468777	0.00716303	0	0.00716555	0.00716555	0.23302548	
NCIH1694	0.01234568	0.01076538	0.42531903	0.01076656	0	0.01080247	0.01076654	0.3836963	
NCIH1836	0.00733624	0.00733006	0.17791959	0.00733008	0	0.00686213	0.00733048	0.16354376	
NCIH2073	0.00627895	0.00652722	0.01836307	0.00652651	0	0.00379919	0.00652946	0.00434808	
NCIH508	0.0047619	0.00613916	0.01842969	0.00613642	0	0.00613642	0.00613642	0.01738742	
NCIH596	0.00786164	0.00596549	0.06303371	0.00596831	0	0.00628931	0.00596819	0.28188119	
NEU	0.00435707	0.00506282	0.01755304	0.00505279	0	0.00561482	0.00505118	0.35805876	
NKDBA	0.01716821	0.0133427	0.01870362	0.01367277	0	0.01367187	0.01336253	0.3672164	
NOMO1	0.0113678	0.011223	0.03629421	0.01122337	0	0.01815182	0.01121944	0.0470331	
OV7	0.00567655	0.00514786	0.33930839	0.0051498	0	0.0055938	0.00514936	0.47638207	
PHH	0.005807	0.00668751	0.00010129	0.00667749	0	0.00593658	0.00667845	0.12094995	
PL21	0.00614316	0.00859037	0.14140745	0.00858757	0	0.00641026	0.0085799	0.23347409	
RKO	0.00552209	0.00539518	0.09506579	0.00539562	0	0.00502008	0.00539601	0.4174155	
RMGI	0.00576923	0.00502457	0.43850647	0.00502597	0	0.0078125	0.00502517	0.11292575	
RMUGS	0.00634249	0.0069089	0.01189513	0.0069076	0	0.0069076	0.0069076	0.01244176	
SKB	0.00717401	0.0109255	6.12E-14	0.01090112	0	0.00677227	0.0109065	0.00012077	
SKBR3	0.02406417	0.01990854	0.35109168	0.01992678	0	0.02727273	0.01992583	0.03679895	
SKLU1	0.00459824	0.00557354	0.01910598	0.00556899	0.03782825	0.00455927	0.00556996	0.33550538	
SKM1	0.00709677	0.00810493	0.29070016	0.0081035	0	0.0081035	0.0081035	0.0386978	
SKMEL1	0.00598802	0.0046608	0.22224889	0.00466439	0	0.00299401	0.00466464	0.09565804	
SKMEL28	0.00349206	0.00635508	0.3479443	0.00635197	0	0.00634921	0.00635197	0.42520343	
SNGM	0.00595238	0.0067818	0.37675028	0.00678026	0	0.00798872	0.00677877	0.38787173	
SNU1040	0.00570048	0.00611947	2.13E-10	0.00611531	0	0.00518519	0.00611652	0.03880334	
SNUC4	0.00563085	0.00506254	0.23566846	0.00506491	0	0.00493097	0.00506496	0.49717886	
SNUC5	0.01104972	0.01010734	0.03088707	0.01010777	0	0.01010777	0.01010777	0.03772155	
SW480	0.00425121	0.00429009	0.00106259	0.00429	0	0.00724638	0.00428818	0.15929313	
SW620	0.01092675	0.01276848	0.24555581	0.01276423	0	0.01076487	0.01276505	0.19427792	
SW948	0.01339539	0.01369306	0.03262539	0.01369134	0	0.00849858	0.01369563	0.06104383	
T3M10	0.00805477	0.00791806	0.07868461	0.00791828	0	0.0052356	0.00791927	0.12156519	
THP1	0.0122449	0.01584676	0.06496666	0.01583927	0	0.00340136	0.01584031	0.46231685	
TYKNU	0.00622665	0.0060595	0.00801674	0.00605999	0	0.00605999	0.00605999	0.13780825	
U266	0.2962963	0.18415638	8.50E-08	0.18492976	0	0.22222222	0.18484383	0.11209766	
U937	0.00997732	0.00909134	0.28795543	0.00909269	0	0.0085034	0.00909281	0.08391389	
WSDULCL2	0.00756564	0.0060145	0.00064519	0.00601576	0	0.00623053	0.00601568	0.26840975	

Table S4

Cell_type	N_drugs	N_targets	M	Density
A375	9646	12328	597410	0.0050238
A549	10524	12325	511185	0.00394104
A673	311	10065	26669	0.00851986
AGS	330	9572	21579	0.00683148
ASC	2298	12323	340291	0.01201667
BT20	110	8469	15978	0.01715132
CD34	17	7567	9257	0.07196107
CL34	184	9202	17116	0.01010886
CORL23	261	5867	8460	0.00552476
COV644	226	5889	7886	0.00592525
DV90	332	7739	12364	0.00481212
EFO27	330	9295	18979	0.00618743
FIBRNPC	539	12036	56767	0.00875034
H1299	321	7867	13685	0.00541914
HA1E	3889	12323	392873	0.00819781
HCC15	324	10132	23718	0.007225
HCC515	4258	12317	317050	0.00604529
HCT116	212	5399	9550	0.00834361
HEC108	349	9952	22329	0.00642885
HEK293T	11	1761	2037	0.10515719
HEPG2	3572	12318	322618	0.00733224
HL60	10	6431	9238	0.14364796
HS27A	11	572	574	0.09122695
HS578T	103	5317	9801	0.01789643
HT115	217	10036	20912	0.0096023
HT29	8931	12324	460953	0.00418798
HUH7	111	10790	31166	0.02602176
JHUEM2	304	7596	12139	0.00525684
JURKAT	5	2261	2448	0.21654135
LOVO	308	9667	22390	0.00751989
MCF10A	108	6518	11377	0.0161618
MCF7	8535	12325	459618	0.00436925
MDAMB231	113	9004	25005	0.0245761
MDST8	327	9804	22964	0.00716303
NCIH1694	162	8037	14018	0.01076656

Cell_type	N_drugs	N_targets	M	Density
NCIH1836	229	8304	13939	0.00733008
NCIH2073	335	10162	22218	0.00652651
NCIH508	231	5025	7123	0.00613642
NCIH596	318	8082	15339	0.00596831
NEU	2183	12237	134986	0.00505312
NKDBA	192	10289	26398	0.01336277
NOMO1	202	10587	24002	0.01122337
NPC	3152	12326	402682	0.01036463
OV7	332	7090	12122	0.0051498
PC3	7956	12324	456036	0.00465107
PHH	1695	12295	139171	0.00667806
PL21	312	10509	28157	0.00858757
RKO	332	8689	15565	0.00539562
RMGI	320	6906	11108	0.00502643
RMUGS	344	9585	22776	0.0069076
SKB	2178	12310	292285	0.01090161
SKBR3	110	7747	16984	0.0199303
SKLU1	329	8361	15322	0.00557008
SKM1	310	10510	26402	0.0081035
SKMEL1	334	6666	10385	0.00466439
SKMEL28	315	9193	18395	0.00635231
SNGM	266	6480	11687	0.00678026
SNU1040	225	4631	6372	0.00611531
SNUC4	338	7429	12718	0.00506491
SNUC5	181	8798	16096	0.01010777
SW480	345	6516	9644	0.00429
SW620	353	12126	54640	0.01276493
SW948	353	12111	58536	0.01369205
T3M10	191	8111	12267	0.00791828
THP1	294	12029	56016	0.01583927
TYKNU	219	3766	4998	0.00605999
U266	6	1305	1448	0.18492976
U937	294	9821	26254	0.00909269
VCAP	10757	12321	531718	0.00401185
WSUDLCL2	321	8645	16694	0.00601576

Table S5

	QUIZ-C	MODZ	CD
A375	0.17318026	0.43394167	0.14063392
A549	0.08284464	0.53185456	0.2326157
HA1E	0.04472389	0.38324468	0.21245392
HCC515	0.06362027	0.34470904	0.19416589
HEPG2	0.06651719	0.27735884	0.21753127
HT29	0.1338308	0.40817537	0.21546634
NPC	0.12116929	0.42538201	0.15460199

Table S6

	p-value	OR
A375	0.07932517	0
A549	0.28153292	0.2582834
A673	0.12117565	0.87655228
AGS	1	1.00448929
ASC	0.39103534	0.43054821
BT20	5.36E-09	0.66695125
CD34	0.03383775	0.86537369
CL34	0.00042009	0.7696678
CORL23	0.04185701	1.14572574
COV644	0.01018405	1.18853768
DV90	3.24E-07	0.70785529
EFO27	2.86E-05	0.73243934
FIBRNPC	0.00039288	0.51888656
H1299	0.57915996	0.96218392
HA1E	0.00829268	0.08591822
HCC15	0.16353393	1.13639947
HCC515	0.06376676	0.25789102
HCT116	2.66E-25	2.01202039
HEC108	1.92E-13	0.56489317
HEK293T	2.28E-54	3.38105731
HEPG2	0.05073009	0.22921626
HL60	6.50E-06	1.35735691
HS27A	0.0001319	1.69528638
HS578T	0.31286263	1.07111413
HT115	0.93174395	0.99192657
HT29	0.00018441	0.02145374
HUH7	1.38E-07	0.61704146
JHUEM2	0.09327996	0.89216087
JURKAT	0.06411877	1.16672977
LOVO	0.09674849	0.87587792
MCF10A	7.12E-09	0.67833874
MCF7	0.28153292	0.2582834
MDAMB231	1.29E-05	0.72914685
MDST8	0.02317086	0.83453955
NCIH1694	0.32739102	1.07406906

	p-value	OR
NCIH1836	2.46E-07	0.70040319
NCIH2073	0.00251124	0.77737087
NCIH508	6.54E-11	1.55004402
NCIH596	0.00044902	1.28999532
NEU	6.14E-09	0.21414403
NKDBA	1	0.99858076
NOMO1	1.58E-05	0.68144369
NPC	0.21961357	0.17217376
OV7	0.00214936	1.23433375
PC3	0.33854394	0.34440821
PHH	1.10E-07	0.12153763
PL21	5.13E-11	0.57287456
RKO	0.28957846	1.08308187
RMGI	0.54581612	1.04234493
RMUGS	0.03720425	0.84967708
SKB	0.0004166	0.14680055
SKBR3	0.44812202	1.05640647
SKLU1	7.42E-05	0.75897841
SKM1	0.88788785	0.98510108
SKMEL1	0.28472802	1.07544428
SKMEL28	0.00086463	0.78121816
SNGM	6.16E-14	1.66995469
SNU1040	0.00725086	1.20265525
SNUC4	0.65814912	1.03161513
SNUC5	0.48372924	1.05699217
SW480	0.24268615	0.92350241
SW620	0.19086973	0.7425911
SW948	0.00031839	0.47445166
T3M10	0.30868889	1.0754411
THP1	0.27806124	0.8003828
TYKNU	9.80E-08	1.45290536
U266	3.99E-14	2.03437579
U937	0.11581196	0.88108225
VCAP	0.02058444	0.14322228
WSUDLCL2	0.00221236	0.80330925

Table S7

	PACOS-QUIZ-C (enhance)	PACOS-QUIZ-C-Tool (enhance)	PACOS-QUIZ-C (repress)	PACOS-QUIZ-C-Tool (repress)
L1000CDS2 (forward)	2.56E-10	1.36E-09		
L1000CDS2 (reverse)			1.95E-05	0.027
DeepDRK (forward)	0.043	0.043		
DeepDRK (reverse)			0.22	0.043
MI (correlation)	0.021	0.0044	0.0076	0.0076

Table S8

The number of benchmark signatures (out of 73) in which the first method has significantly higher (two-sided Mann-Whitney U p <= 0.05) chemical diversity compared to the second method.

	PACOS (enhance)	PACOS (repress)	PACOS + Tool (enhance)	PACOS + Tool (repress)
QUIZ-C vs MODZ CD	43	67	47	66
QUIZ-C vs L1000CDS2	32	49	26	34
L1000CDS2 vs QUIZ-C	25	13	35	19

Table S9

Perturbation	Drug description/MoA	Cell type	PACOS + Tool combined score (Repress)	AUROC	AUROC emp. pval	VGD stage	References
betamethasone	Corticosteroid	SKB	0.655199891	0.8125	0.0025	Early	7, 8
tamoxifen		CORL23	0.643488296	0.679338843	0.0875	Early	
estriol		ASC	0.607079937	0.640532757	0.1325	Early	
thalidomide		A673	0.601328083	0.618666667	0.175	Early	
7-nitroindazole		RMGI	0.590180388	0.549677419	0.36	Early	
splitomycin	SIRT1 inhibitor	NEU	0.573809314	0.697330791	0.015	Early	20
AG-490	EGFR tyrosine kinase inhibitor	SKB	0.491281184	0.8125	0.0025	Early	12-15
GW-7647		ASC	0.485314527	0.640532757	0.1325	Early	
formoterol	Long-acting beta agonist	NEU	0.474404473	0.697330791	0.015	Early	
rosiglitazone	PPAR- γ agonist	NCIH2073	0.472365851	0.698780488	0.05	Early	1
loperamide	Opioid-receptor agonist	HEC108	0.471989585	0.763841808	0.0125	Early	
GW-7647	PPAR- α agonist	SKB	0.47185258	0.8125	0.0025	Early	3, 4
nitrendipine	Calcium channel blocker	HEC108	0.466840953	0.763841808	0.0125	Early	10, 11
GW-843682X		RMGI	0.466694863	0.549677419	0.36	Early	
capsaicin		NCIH596	0.463770687	0.588996764	0.2725	Early	
cyclosporin-a		SW620	0.462206053	0.662395543	0.1075	Early	
DG-041		HCC15	0.461541243	0.624605678	0.1775	Early	
calcitriol	Active vitamin D metabolite	SNUC4	0.452233834	0.746107784	0.025	Early	21
dephostatin		ASC	0.449709245	0.640532757	0.1325	Early	
sildenafil		SKMEL1	0.448945213	0.576219512	0.2625	Early	
SCH-58261	Adenosine receptor A2A antagonist	SKB	0.431723728	0.8125	0.0025	Early	22-24
Perturbation	Drug description/MoA	Cell type	PACOS + Tool combined score (Repress)	AUROC	AUROC emp. pval	VGD stage	References
estriol	Steroid (Estrogen)	ASC	0.58611692	0.73981002	0.0325	Late	5, 6
splitomycin		NEU	0.5148704	0.62552381	0.1625	Late	
dephostatin	Protein tyrosine phosphatase 1B inhibitor	ASC	0.47709501	0.73981002	0.0325	Late	17, 18
GW-7647	PPAR- α agonist	ASC	0.47022795	0.73981002	0.0325	Late	3,4
RS-100329	α 1-AR antagonist	ASC	0.43838991	0.73981002	0.0325	Late	2
sunitinib	Multitarget RTK inhibitor	ASC	0.42459856	0.73981002	0.0325	Late	9
formoterol		NEU	0.41836029	0.62552381	0.1625	Late	
sulpiride	Dopamine/serotonin receptor antagonist	ASC	0.41523462	0.73981002	0.0325	Late	
paroxetine	Selective serotonin reuptake inhibitor	ASC	0.40442384	0.73981002	0.0325	Late	
gefitinib	EGFR tyrosine kinase inhibitor	ASC	0.40350811	0.73981002	0.0325	Late	12-15
pravastatin	HMG CoA reductase inhibitor	ASC	0.40282203	0.73981002	0.0325	Late	16
gefitinib		NEU	0.39579329	0.62552381	0.1625	Late	
tozasertib		NEU	0.39447828	0.62552381	0.1625	Late	
SB-203580	p38 kinase inhibitor	ASC	0.39028988	0.73981002	0.0325	Late	19