nature methods

Resource

https://doi.org/10.1038/s41592-022-01667-0

High-dimensional gene expression and morphology profiles of cells across 28,000 genetic and chemical perturbations

Received: 7 September 2021

Accepted: 28 September 2022

Published online: 7 November 2022

Check for updates

Marzieh Haghighi ✉, Juan C. Caicedo, Beth A. Cimini, Anne E. Carpenter & Shantanu Singh ✉

Cells can be perturbed by various chemical and genetic treatments and the impact on gene expression and morphology can be measured via transcriptomic profiling and image-based assays, respectively. The patterns observed in these high-dimensional profile data can power a dozen applications in drug discovery and basic biology research, but both types of profiles are rarely available for large-scale experiments. Here, we provide a collection of four datasets with both gene expression and morphological profile data useful for developing and testing multimodal methodologies. Roughly a thousand features are measured for each of the two data types, across more than 28,000 chemical and genetic perturbations. We define biological problems that use the shared and complementary information in these two data modalities, provide baseline analysis and evaluation metrics for multi-omic applications, and make the data resource publicly available (https://broad.io/rosetta/).

Biological systems can be quantified in many different ways. For example, researchers can measure the morphology of a cell using microscopy and image analysis, or molecular details such as the levels of mRNA or protein in cells. Historically, biologists chose a single feature to measure for their cell samples, based on their previous knowledge or hypotheses. Now, 'profiling' experiments capture a high-dimensional profile of features for each sample, and hundreds to thousands of samples can be quantified. This allows the discovery of unexpected behaviors of the cell system.

Profiling experiments carried out at large scale remain expensive, even for a single profiling modality. We observed that no public dataset exists providing both genetic and chemical perturbation of cells with two different kinds of profiling readouts. Such a dataset would enable multimodal (also known as multi-omic) analyses and applications. Examples include integrating the two data sources to better predict a compound's activity in an assay[1], predicting the mechanism of action (MoA) of a drug based on its profile similarity to well-understood drugs[2], or predicting a gene's function based on its profile similarity to well-understood genes[3].

Observing a system from multiple perspectives is known to reveal patterns in data that may not be visible in individual perspectives. Machine learning methods have been explored in various fields to learn from multiple sources to make better inferences from data[4]. In biology, the advancement of technologies for measuring multi-omic data has sparked research investigating the relationship and integration of different high-dimensional readouts[5]. For example, transcriptomic, proteomic, epigenomic and metabolomic data can be combined to predict the MoAs of chemical compounds[6].

Here, we created a collection of gene expression (GE) and morphology datasets with the scale and annotations needed for machine learning research in multimodal data analysis and integration. The GE data were obtained using the L1000 assay[7] and the morphology datasets using the Cell Painting (CP) assay[8]. This Resource provides two different, rich views on the cells by providing roughly a thousand mRNA levels and a thousand morphological features when samples of cells are perturbed by hundreds to thousands of different conditions, including chemical and genetic. Furthermore, we present a framework for thinking about the utility of multimodal data by defining applications

Broad Institute of MIT and Harvard, Cambridge, MA, USA. ✉e-mail: mhaghigh@broadinstitute.org; shantanu@broadinstitute.org

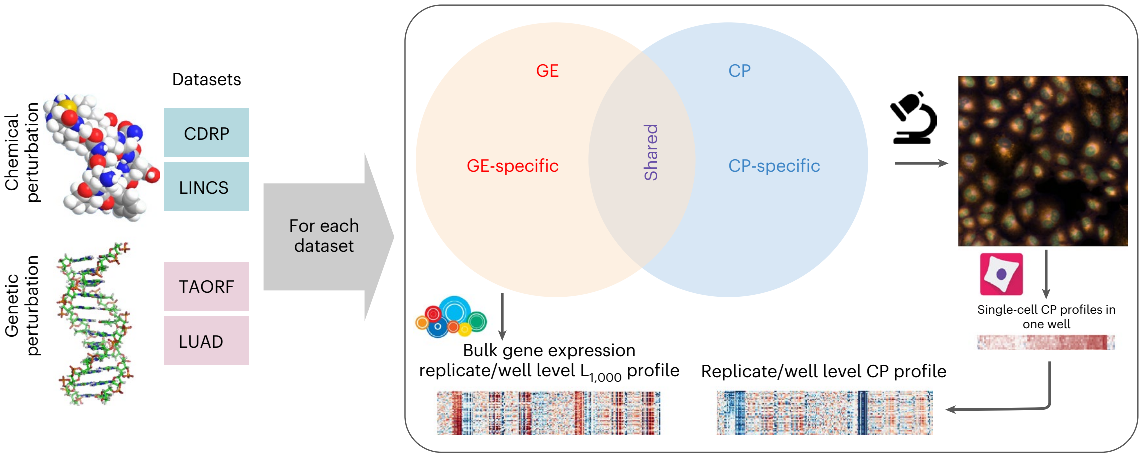

Datasets
Chemical perturbation
CDRP
LINCS
Genetic perturbation
TAORF
LUAD
For each dataset
GE
CP
GE-specific
Shared
CP-specific
Bulk gene expression replicate/well level $L_{1,000}$ profile
Replicate/well level CP profile
Single-cell CP profiles in one well

Fig. 1 | Multimodal datasets overview. Multimodal genetic and chemical perturbation datasets are valuable for many applications. For each dataset, CP and GE L1000 assays were used to collect morphological and GE representations (profiles), respectively. The datasets are described in Supplementary 1. Chemical structure by G. J. Owns, distributed under a CC0 3.0 license. DNA structure by R. Wheeler, distributed under a CC BY-SA 3.0.

where the shared information, and the complementary information, across data types can be useful, using terminology understandable to those new to the biological domain. We demonstrate example applications within each group, uncover interesting biological relationships, and provide baseline methods, code, evaluation metrics and benchmark results for each, as a foundation for future biologically oriented machine learning research.

Results

Gene expression and morphological profiles

All datasets were created at our institution (Supplementary 1) and involved one of two types of 'inputs': chemical perturbations and genetic perturbations (Fig. 1). There are also two types of high-dimensional outputs measured: GE profiles and morphological profiles, each with roughly 1,000 features measured. For each of the datasets, in a single laboratory, cells are plated into two sets of identical plates, each plate is treated with chemical (or genetic) perturbations identically, and then one set is used to measure GE and the other set to measure morphology.

We captured GE (mRNA) profiles using the L1000 assay[7]. The levels of mRNA in the cell are often biologically meaningful; collectively, mRNA levels for a cell are known as its transcriptional state. The L1000 assay reports a sample's mRNA levels for ~978 genes at high throughput, from the bulk population of cells treated with a given perturbation. These 'landmark' genes capture approximately 82% of the transcriptional variance for the entire genome[7]; the specific genes' mRNAs that are measured can be different across datasets, although largely overlapping (Methods). We note that 'genes' are an input (individual genes are overexpressed as the perturbation in some datasets) and an output (GE profiles are the measured mRNA levels for each landmark gene in the L1000 assay); this can cause confusion for researchers new to the domain.

We captured morphological profiles using the CP assay[8]. This microscopy assay captures fluorescence images of cells colored by six well-characterized fluorescent dyes to stain the actin cytoskeleton, Golgi apparatus, plasma membrane, nucleus, endoplasmic reticulum, mitochondria, nucleoli and cytoplasmic RNA in five channels of high-resolution microscopy images. Images are processed using CellProfiler software[9] to extract thousands of features of each cell's morphology such as shape, intensity and texture statistics, thus forming a high-dimensional profile for each single cell. The aggregated (population-averaged) profiles were then created for all imaged single cells in each sample well.

For both data types, aggregation of all the replicate-level (equivalent to well-level) profiles of a perturbation is called a treatment-level profile. In our study, we used treatment-level profiles for each perturbation in all experiments but have provided replicate-level profiles for researchers interested in further data exploration. Therefore, in our experiments, each perturbation in each dataset has two corresponding vectors of measurements for each modality; one treatment-level profile for GE and one treatment-level profile for morphological measurements. GE and morphological measurements are taken from two different sets of plates; therefore, no direct, one-to-one correspondence exists between the two readouts at the replicate level. Of the eight datasets provided (four datasets × two modalities), four have been used previously by researchers at our Institute[3,10,11]; here, we complete the matrix by providing the missing data type for each pair, organizing them and providing benchmarks.

Shared versus complementary information content

Cell morphology and GE are two very different kinds of measurements about a cell's state, and their relationship is known to be complex. For example, a change in morphology can induce GE changes[12] and GE changes can induce a change in cell morphology[13,14]. However, a strict relationship is not always the case; many drugs impact cells' mRNA or morphology profile, but not both[10,15,16]. Changes in protein stability or posttranslational modifications can induce changes in morphology without changes in GE; for example, in the Rho family of small GTPases, morphology changes on a timescale that is much too short to be explained by changes in mRNA[17]. Furthermore, the two data types are collected at different time points, determined as optimal for each individually. Therefore, even if technical artifacts were nonexistent, we do not expect a one-to-one map between these two modalities. We therefore hypothesized that the information in each data type consists of a shared subspace, a modality-specific complementary subspace, and noise (Fig. 1). Both subspaces can be exploited for biological applications.

Shared subspace across two modalities

The shared subspace between GE and cell morphology is beginning to be explored. For example, cross-modal autoencoders learned the shared latent space for single-cell RNA-sequencing (RNA-seq) and chromatin images to integrate and translate across modalities[18]. In another study, probabilistic canonical correlation analysis learned a shared structure in paired samples of histology images and bulk GE RNA-seq data, suggesting that shared latent variables form a composite phenotype between morphology and GE that can be useful[19]. Many uncovered relationships will not be transferable from one experimental batch to another, particularly if great differences exist: for example, histology images differ in many ways from fluorescence microscopy images, yet some features, such as nuclear shape, might be consistent across different experimental techniques.

The existence of a shared subspace enables multiple applications. Most prominently, if sufficient shared information is present, one modality can be computationally predicted (that is, inferred or estimated) using another, saving substantial experimental resources. For example, one could predict the expression level of genes of interest given their morphological profiles from already available images, even from patients whose samples are no longer available for mRNA testing. Or, one could generate images or morphological profiles from large libraries of mRNA profiles.

Another use of shared subspace is to identify relationships between specific features of the two types. For example, a morphological feature and a specific gene's mRNA level may be tightly linked, which can yield clues as to the biological mechanisms underlying their relationship. As well, inspecting which genes can be well predicted may shine light on general relationships between mRNA levels and morphology for different classes of genes[20]; enrichment analysis of these groups of genes could also lead to biological pattern discoveries. Researchers have used linear regression and enrichment analysis to explore the association between variations in cell morphology and transcriptomic data[16].

Modality-specific, complementary subspaces

Each modality will likely have a modality-specific subspace containing information unique to that modality and unpredictable by the other. Although this property confounds applications requiring a shared subspace, it enables other applications because the fusion of two modalities should increase the overall information content, and therefore predictive power, of a profiling dataset.

Data modality fusion and integration techniques are an active area of research in machine learning[4] and could potentially yield a superior representation of samples for many different biological profiling tasks on datasets where multiple profiling modalities are available. For example, predicting assay activity might be more successful using information about the impact of that compound on cells' mRNA levels and morphology, rather than either data source alone[1]. Likewise, predicting the function of a gene based on similarities to other genes' profiles might be more successful using both data types.

Application 1: cross-modality predictions

As a baseline for finding the correspondence between GE and morphology and predicting one from the other, we modeled the relationship using a regression model in which the mRNA level of each landmark gene in the GE profile can be estimated as a function of all the morphological features in the CP profile, $\mathbf{y_l} = f(X_{cp}) + \mathbf{e_l}$; in which $\mathbf{y_l}$ is a P-dimensional (P,1) vector of expression levels for the landmark gene l across all the P perturbations in a dataset and X_{cp} is the $P \times F$-dimensional (P,F) whole morphological data matrix representing all F morphological changes/features across all the P perturbations. We used Lasso as a baseline linear model and multilayer perceptron (MLP) as a baseline nonlinear model for the regression problem.

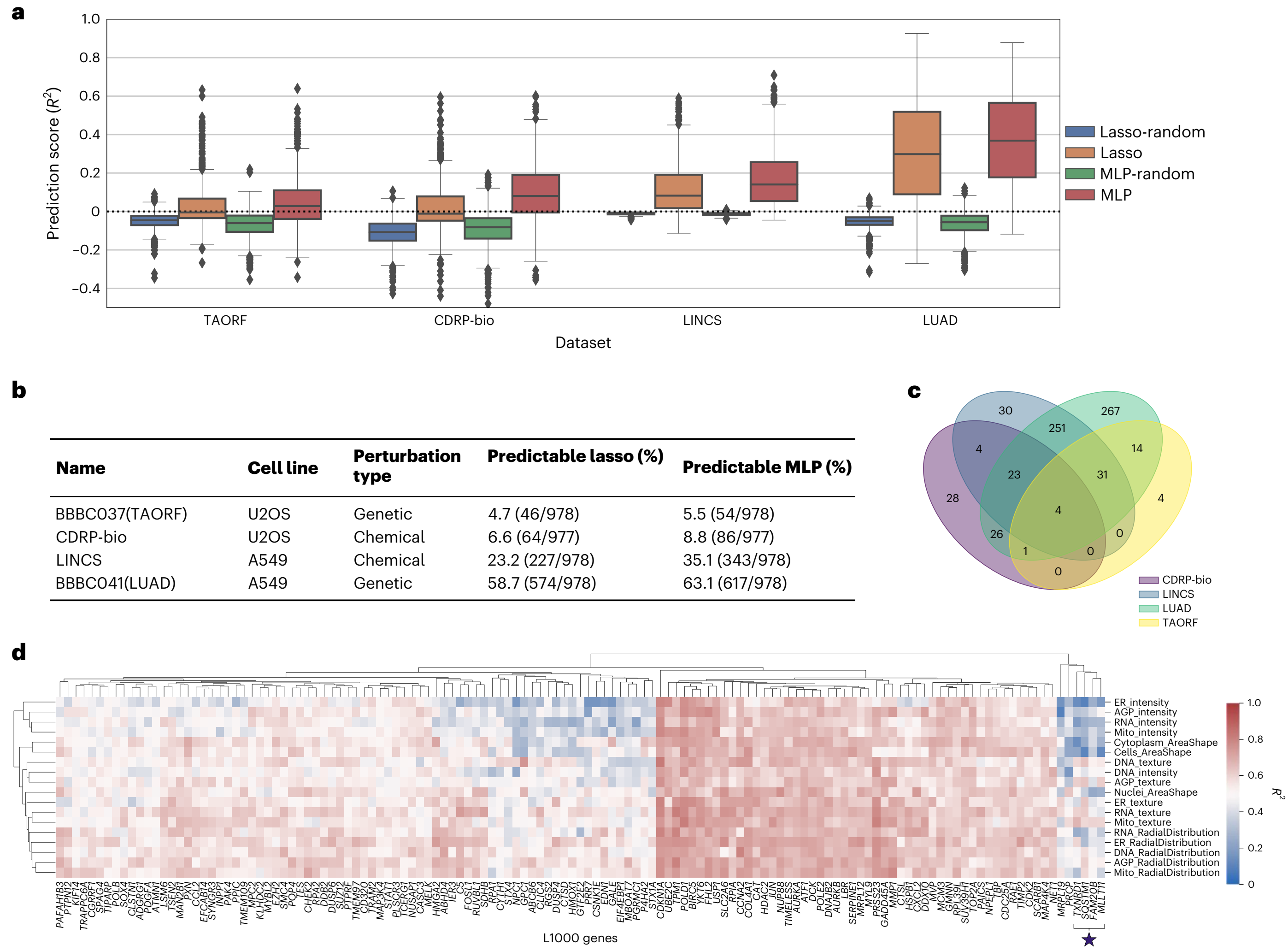
Some datasets showed excellent accuracy in predicting some mRNA levels from morphology data (and vice versa), with MLP yielding superior results to Lasso (Fig. 2a,b). Machine learning methods that can improve upon these benchmarks would be very useful to the biomedical community. Two of the datasets (LUAD and LINCS) have a markedly higher performance than the other two (TAORF and CDRP-bio), which suggests a likely poorer data quality or poorer alignment of the modalities in the latter two. Given LUAD and LINCS both use A549 cells, it is also possible that the transcription–morphology link is cell-line dependent, and that it is stronger in A549 for some reason; however, it seems even more likely that the differences in performance relate to differences in technical quality of the data. Likewise, further preprocessing and denoising techniques such as batch-effect corrections to improve alignment are another target for future machine learning research. In addition to alignment across modalities, alignment across different datasets is also necessary to translate the prediction model across different datasets. Application of a model trained on each of the highest performing datasets and tested on the other one (LUAD and LINCS) indicates poor translatability of the models across datasets (Extended Data Fig. 1). Improving model generalizability across datasets requires methods specifically designed to correct for technical variations and batch effects in the bulk-level information of the data types presented herein.

The shared information in the two modalities can be used in other ways. We can identify the overlap in landmark genes that are highly predictable according to one or more datasets (Fig. 2c); 59 landmark genes were well predicted in at least three of the four datasets. For the LUAD dataset (which has the highest cross-modal predictability), we identified the gene families for highly predictable genes (Extended Data Fig. 2). Overrepresentation analysis of LUAD's highly predictable gene set (relative to the L1000 background gene set) revealed that many overrepresented categories related to components stained in the CP assay, such as DNA and actin (Extended Data Fig. 3).

Finally, we examined prediction scores for each category of image-based feature in the experiment, to aid in understanding which features underlie prediction of which genes' mRNA levels. To do this, we first sorted CP features into four categories (intensity, texture, radial distribution and shape) and five fluorescence channels (DNA, RNA, ER, AGP and Mito), and then we calculated and displayed feature-group-specific prediction scores as a hierarchically clustered heat map of median (over k-folds) prediction scores (Fig. 2d). In this view, genes with strong red columns were predicted using many of the morphological categories of features, indicating that the genes are associated with widespread morphological changes; several of these were cell cycle related, which is known to impact morphology dramatically. Others were more selective, such as the cluster of genes including *TXNRD1*, *SQSTM1*, *FAM20B* and *MLLT11*, whose mRNA levels were strongly predictable by mitochondrial texture features (Fig. 2d). Several of these genes have functional annotations relating to mitochondria, and cells that are predicted to have (and actually do have) high levels of these four genes' mRNA were all associated with visible changes in the staining for mitochondria (Extended Data Fig. 4).

To more generally inspect if the GE–CP relationships observed (Fig. 2d) are consistent with the known biological functions of the L1000 landmark genes, we performed a Gene Ontology (GO) terms search analysis (Methods). We wondered whether landmark genes that are highly predictable by morphological features in each specific CP channel are more likely to have GO annotations related to that channel compared to the rest of CP channels; this was generally not the case (Extended Data Fig. 5 and Methods), consistent with most predictable genes showing signal across all categories of features rather than being strongly channel specific (Fig. 2d). We also wondered whether landmark genes that are more predictable than other genes are more likely to have functions associated with the particular stains in the CP assay. Indeed, among the set of 59 highly predictable genes (in at least three of the four datasets), we observed an increased chance of annotations relating to the cellular components and organelles stained in the assay (Extended Data Fig. 6). That said, many highly predictable genes were associated with no such terms, indicating that the assay probes biological impact beyond the particular labeled components, or that the genes have unannotated functions.

Prediction can be run in the other direction as well, that is, each morphological feature can also be estimated using the 978 landmark genes as $\mathbf{y_f} = f(X_{ge}) + \mathbf{e_f}$; in which $\mathbf{y_f}$ is a P-dimensional (P,1) vector of measurements for feature f across all the P perturbations in a dataset and X_{ge} is the whole GE data matrix (P,L) representing all L landmark



Name	Cell line	Perturbation type	Predictable lasso (%)	Predictable MLP (%)
BBBC037(TAORF)	U2OS	Genetic	4.7 (46/978)	5.5 (54/978)
CDRP-bio	U2OS	Chemical	6.6 (64/977)	8.8 (86/977)
LINCS	A549	Chemical	23.2 (227/978)	35.1 (343/978)
BBBC041(LUAD)	A549	Genetic	58.7 (574/978)	63.1 (617/978)

Fig. 2 | An application using the shared subspace: cross-modality predictions from Cell Painting to gene expression. **a**, Distribution of R^2 prediction scores for all landmark genes for each Lasso and MLP model, grouped for each dataset. Many genes are well predicted, especially using MLP. The random shuffle distributions— where the outputs are shuffled in each iteration—serve as negative controls. Negative R^2 values indicate that the prediction is worse than by simply computing the mean of the output, and thus all $R^2 < 0$ values can be considered equally bad (the model does not generalize at all). The *y* axis was trimmed at −0.5 for clarity. In the box plots, the center line indicates the median, box limits represent upper and lower quartiles and whiskers denote 1.5 times the interquartile range; *n* = 978 landmark genes (977 for CDRP-bio dataset). **b**, The percentage of genes that were well predicted ($R^2 > (t_{99th} + 0.2)$; Methods) is shown for each dataset. **c**, The overlap of genes predictable by the MLP model ($R^2 > (t_{99th} + 0.2)$) is shown across the four datasets; 59 are well predicted in at least three of the four datasets. **d**, Example of interpretable maps showing the connection between the expression of each landmark gene and the activation of each category of morphological features in the LUAD dataset using the MLP model: each point on the heat map shows the predictive power of a group of morphological features (on the *y* axis) for the predicting expression level of a landmark gene (on the *x* axis). 'Predictive power' here means the R^2 scores generated by limiting the prediction to all the features in the *y* axis group. The cluster marked with an asterisk is discussed in the main text and explored in Extended Data Fig. 4. The heat map is limited to 131 genes with $R^2 > 0.6$ scores according to any of the morphological groups (on the *y* axis). The complete version is provided in the GitHub repository (cat_scores_maps.xlsx) and can be loaded into Morpheus[32] or Python for further exploration.

genes measurements across all the *P* perturbations. We found a large portion of morphological features to be highly predictable especially for the LUAD and LINCS datasets (Fig. 3a). Grouping highly predictable morphological features according to all the datasets revealed that they fell mainly in the radial distribution and texture features categories across all channels (Fig. 3b). We also provide a Jupyter notebook for exploring the list of top connections between any gene or morphological feature of interest (explore_the_link.ipynb). Users can input an L1000 landmark gene and get the list of top morphological features involved in the prediction of the input feature along with their importance score. Likewise, one can query a morphological feature to find the landmark genes whose mRNA levels are predictive. For example, the morphological feature 'Cells_Texture_InfoMeas1_RNA_3_0' relies on the levels of many genes in its prediction, including several known to be involved in mRNA processing (Fig. 3c).

Application 2: Integrating gene expression and morphology

Discerning how a compound works is a major bottleneck in drug discovery. The task is called MoA determination, and the goal is to determine the mechanism by which the drug impacts the biological system. Existing methods are often resource and time intensive, with a low success rate. As a result, few strategies have been tested systematically across a diverse set of drugs; most strategies inherently only work on a subset of drug or target types, such that multiple methods are usually pursued simultaneously to generate a hypothesis for further testing[21]. One promising method to predict MoAs is to collect a profile from cells

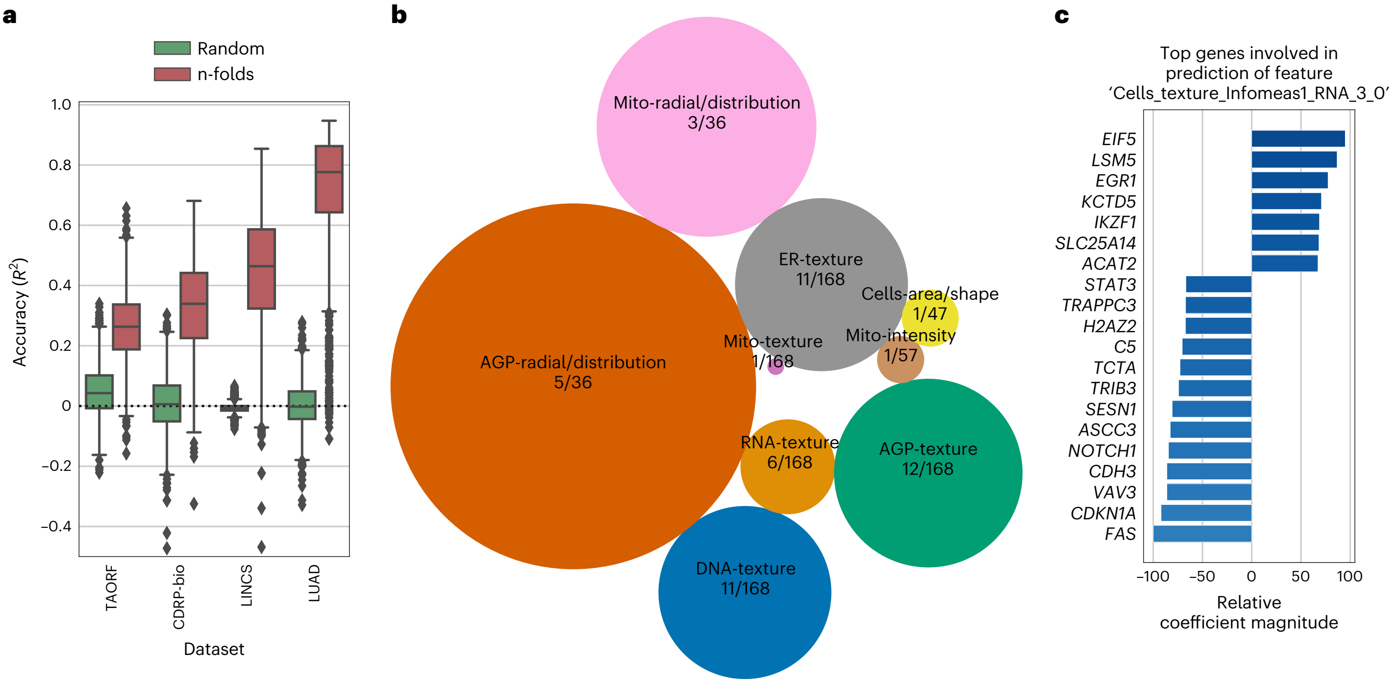


Fig. 3 | Cross-modality predictions from gene expression to Cell Painting. a, Distribution of R^2 prediction scores for all morphological features using the MLP model (red) along with the random shuffle—where the output is shuffled in each iteration—serving as negative controls (green), for each dataset. The *y* axis is trimmed at −0.5 for clarity. In the box plots, the center line indicates the median, box limits represent upper and lower quartiles and whiskers denote 1.5 times the interquartile range; the number of points or number of CP features varies among datasets; $n = 1{,}569$ (TAORF), $n = 1{,}570$ (CDRP-bio), $n = 1{,}670$ (LINCS) and $n = 1{,}569$ (LUAD). **b**, Categories of features with the highest percentage of predictable CP features using GE profiles (median R^2 score across all datasets > 0.6). The sizes of circles are proportional to the percentage of highly predictable features in each category. The number of features in each category over the total number of morphological features in that category are also shown for each circle. **c**, Example output of exploratory scripts available to researchers to see what are the most relevant genes to a given morphological feature of interest (and vice versa). The *x* axis (relative coefficient magnitude) indicates the relative importance of each feature as the percentage of the strongest feature component (here it translates to the most important landmark gene) involved in the prediction of the morphological feature under exploration. The absolute value and sign of this metric corresponds to the level of importance and direction of the linear relationship, respectively. A description of each morphological feature extracted by CellProfiler software is available on the Cell Painting wiki.

and attempt to match it to a library of profiles gathered from other chemical perturbations: a match, or close similarity, can be helpful if the compound the query matches is already well known. Likewise, a match to a genetic perturbation means that the gene, or another gene in the same pathway, is a possible target of the query compound[22].

Several studies have reported success predicting the MoA of compounds using GE or cell morphology data individually[23–26] but none of these integrated the two data types to test for improved predictive ability in a supervised or unsupervised setting. We therefore provide a benchmark for this, using the two chemical perturbation datasets in our set, CDRP-bio and LINCS. The discovery that many genes could not be well predicted based on morphology (and vice versa) in application 1 lends some support for the idea that the two modalities might carry complementary information.

In the unsupervised setting, we tested how the compounds cluster together by their MoA class, in feature spaces of each modality alone, and in the integrated space of both modalities, using several state-of-the-art modality integration methods[27]. Clustering of perturbations using each CP and GE modality alone shows CP outperformed GE in this MoA ground-truth retrieval task in both compound datasets. We observed that, although most of the integration methods increase cluster retrieval performance in the integrated space compared to the GE space, only regularized generalized canonical correlation analysis (RGCCA)[28] improved the performance over the CP space alone (Fig. 4a).

In the supervised setting, using logistic regression and MLP classifiers as the baseline models, we predicted MoA labels using each modality of data independently, with a standard *k*-fold ($k = 5$) cross-validation on a filtered subset of compounds. CP profiles resulted in higher MoA prediction performance compared to GE profiles for each of the datasets (Fig. 4b). Next, we performed the MoA prediction task on two integrated spaces: (1) trivial representation-level concatenation of profiles from the two modalities (shown as early fusion) and (2) representation-level concatenation of profiles from the two modalities in the RGCCA space (shown as RGCCA_EarlyFusion). We also performed decision-level integration of modalities for the MoA prediction task (shown as late fusion), that is, class probabilities output by classifiers trained on each modality separately were averaged before the final MoA prediction.

All three integration strategies showed relatively comparable performance in predicting MoA across the two datasets and two model types, with small average improvements upon the performance of the better-performing modality (Fig. 4b), highlighting the need for developing data fusion methods that better leverage the complementarity of the modalities.

Exploring MoA-class-specific F1-scores for the integrated modalities revealed high variation in class-specific prediction results (Fig. 4c). As already seen more generally, the integration of modalities does not always increase the performance of the MoA prediction task over the higher-performing modality alone for individual MoA categories.

Discussion

We provide the research community a collection of multimodal profiling datasets with GE and morphology readouts, representing two cell types and two perturbation types (genetic and chemical). We define

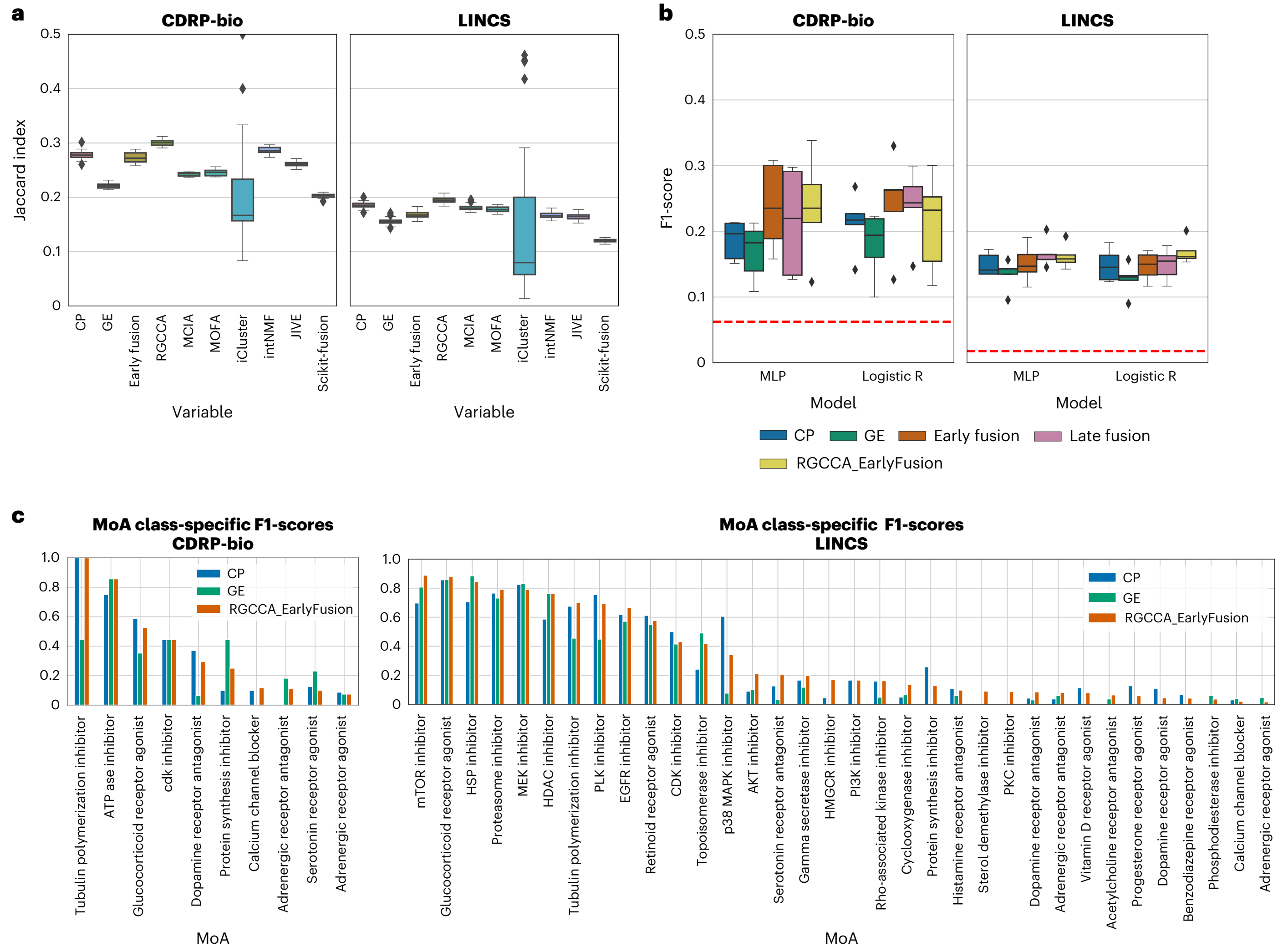


Fig. 4 | Using complementary information: data integration for MoA cluster retrieval and class prediction in compound datasets. a,**b**, An application using the complementary subspaces: integrating multimodal data for MoA unsupervised clustering retrieval (**a**) and supervised prediction (**b**). **a**, Benchmarking of data integration methods on the task of clustering compounds by their MoA categories. Distribution of the Jaccard Index values (one per MoA class; higher is better) computed between the clusters identified by the different integration methods[27] and the ground-truth MoA clusters. RGCCA improved MoA retrieval for both CDRP-bio and LINCS datasets. In the box plots, the center line indicates the median, box limits represent upper and lower quartiles and whiskers denote 1.5 times the interquartile range; $n = 16$ (CDRP-bio), $n = 57$ (LINCS). **b**, MoA classification of the two compound datasets (CDRP-bio and LINCS) using GE, morphology and their integration to predict the MoA of compounds. Classification performance (weighted F1-score) for the MLP and logistic regression classifiers using each data modality alone, the two early and late fusion strategies explained in the main text, and the early fusion of modalities after application of RGCCA on the feature space of both modalities. Chance-level predictions for each dataset are shown as a horizontal red dashed line. In the box plots, the center line indicates the median, box limits represent upper and lower quartiles and whiskers denote 1.5 times the interquartile range; $n = k = 5$. **c**, Class-specific F1-scores are shown based on the MLP model for 16 MoA categories of CDRP-bio (left; 4 of 16 MoA categories that resulted in zero F1-scores after fusion were excluded) and for LINCS (right; 23 of 57 MoA categories that resulted in zero F1-scores after fusion were excluded).

useful biological applications for this data in two categories: those using the shared information and those using modality-specific, complementary information. We provide the data, code, metrics and benchmark results for one application in each category.

The results demonstrate that GE and morphology profiles contain useful overlapping and distinct information about cell state. We were pleased to find that many mRNAs are predictable by cell morphology and vice versa, under the conditions of these high-throughput assays. Similarly, we found that morphology captures information beyond that seen in an mRNA profile; that is, the two modalities contain unique information and we identified which compounds' mechanisms are better captured by each. Although some scientists speculated that it is impossible for cells to show a morphological change without mRNA profiles changing, whether as a cause or consequence, we find this is not the case. We made a number of observations of new biology, such as which genes' mRNA levels are predicted by which particular morphological features (and vice versa). Finally, we discovered that the CP assay information can predict the mRNA levels of genes not clearly linked to the stains in the assay, pointing to its ability to capture broad biological impact.

The results also demonstrate that these applications are challenging enough to provide room for improvement. For example, the variation in the performance for prediction tasks across different datasets shows the necessity of machine learning techniques to further filter and preprocess the profiles (for example, to correct batch effects, including those resulting from the position of wells on a plate[29]) to improve performance. Such techniques might also sufficiently align the four datasets with each other, to explore generalized, dataset-independent models. Nevertheless, we note that we do not expect anywhere close to 100% accuracy for either application. For prediction across the two modalities, we do not expect the modalities to be completely overlapping in their shared information. Furthermore, we note that ground truth in this prediction task is defined only by the available experimental GE and cell morphology data, which is subject to technical variation and error and thus is not absolute truth. For MoA prediction, the application is 'notoriously challenging' and low percentage success rates are expected for any single assay; most commonly, several strategies are used to determine the MoA[30]. In addition, the ground truth is based on imperfect human knowledge.

There are multiple additional limitations for the presented datasets, aside from their data quality as already noted. The number of gene perturbations captured in these datasets was a few hundred, whereas there are roughly 21,000 genes in the genome and numerous variations within each, which could be overexpressed or knocked down. Likewise, a few thousand compounds were tested here but pharmaceutical companies often have millions of compounds. The only limitation for expanding these datasets is the financial resources to carry out the experiments. Regarding limitations of the assays, the GE profiles are captured by the L1000 assay, which is thought to capture 82% of full transcriptome variation[7], and the CP assay includes only six stains, which is insufficient to capture the localization and morphological variation of all cellular components. Finally, the cell types were commonly used historical lines derived from two white patients, one male (A549) and one female (U2OS). Therefore, conclusions from these data may only hold true for the demographics or genomics of those persons and not broader groups. The cell lines were chosen because they are both well suited for microscopy and they offer the advantage of connecting to extensive previous studies and datasets using them.

Despite these limitations, these datasets may be used to pursue many other applications of profiling in biology, as well as methods development. The complementary information used here for MoA prediction can be used for any profiling application; there are more than a dozen that can impact basic biology discovery and the development of novel therapeutics[31]. Each application can also be validated in different ways. For example, the prediction task might be extended to more complex systems, such as human tissue samples where appropriate stains have been used, although such samples are more difficult to procure, and assessing adjacent tissue slices may introduce variation not present in the cultured cell lines used in this study. In the future, multimodal profiles at the single-cell level may become widely available. In the presented datasets, single-cell information exists in one modality (images) but not in the other modality (mRNA). Therefore, the variations in one cannot be explained by the other, as we have a distribution in one space (images) and point estimates in the other space (mRNA). Although still very rare, small and labor intensive to create, datasets with both GE and morphology at single-cell resolution are beginning to become available via in situ RNA-seq methods and could accelerate the field of multimodal biological data analysis.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-022-01667-0.

References

1. Moshkov, N. et al. Predicting compound activity from phenotypic profiles and chemical structures. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/2020.12.15.422887v4 (2022).
2. Breinig, M., Klein, F. A., Huber, W. & Boutros, M. A chemical–genetic interaction map of small molecules using high-throughput imaging in cancer cells. *Mol. Syst. Biol.* **11**, 846 (2015).
3. Rohban, M. H. et al. Systematic morphological profiling of human gene and allele function via Cell Painting. *Elife* **6**, e24060 (2017).
4. Meng, T., Jing, X., Yan, Z. & Pedrycz, W. A survey on machine learning for data fusion. *Inf. Fusion* **57**, 115–129 (2020).
5. Baldwin, E. et al. On fusion methods for knowledge discovery from multi-omics datasets. *Comput. Struct. Biotechnol. J.* **18**, 509–517 (2020).
6. Patel-Murray, N. L. et al. A multi-omics interpretable machine learning model reveals modes of action of small molecules. *Sci. Rep.* **10**, 954 (2020).
7. Subramanian, A. et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell* **171**, 1437–1452 (2017).
8. Bray, M. -A. et al. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat. Protoc.* **11**, 1757–1774 (2016).
9. McQuin, C. et al. CellProfiler 3.0: next-generation image processing for biology. *PLoS Biol.* **16**, e2005970 (2018).
10. Wawer, M. J. et al. Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling. *Proc. Natl Acad. Sci. USA* **111**, 10911–10916 (2014).
11. Berger, A. H. et al. High-throughput phenotyping of lung cancer somatic mutations. *Cancer Cell* **30**, 214–228 (2016).
12. Haftbaradaran Esfahani, P. & Knöll, R. Cell shape: effects on gene expression and signaling. *Biophys. Rev.* **12**, 895–901 (2020).
13. Drareni, K., Gautier, J.-F., Venteclef, N. & Alzaid, F. Transcriptional control of macrophage polarisation in type 2 diabetes. *Semin. Immunopathol.* **41**, 515–529 (2019).
14. Mota de Sá, P., Richard, A. J., Hang, H. & Stephens, J. M. Transcriptional regulation of adipogenesis. *Compr. Physiol.* **7**, 635–674 (2017).
15. Way, G. P. et al. Morphology and gene expression profiling provide complementary information for mapping cell state. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/2021.10.21.465335 (2022).
16. Nassiri, I. & McCall, M. N. Systematic exploration of cell morphological phenotypes associated with a transcriptomic query. *Nucleic Acids Res.* **46**, e116 (2018).
17. Spiering, D. & Hodgson, L. Dynamics of the Rho-family small GTPases in actin regulation and motility. *Cell Adh. Migr.* **5**, 170–180 (2011).
18. Dai Yang, K. et al. Multi-domain translation between single-cell imaging and sequencing data using autoencoders. *Nat. Commun.* **12**, 31 (2021).
19. Gundersen, G., Dumitrascu, B. & Ash, J. T. End-to-end training of deep probabilistic CCA on paired biomedical observations. In *Proceedings of PMLR* pp.945–955 (2019).
20. He, B. et al. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat. Biomed. Eng.* **4**, 827–834 (2020).
21. Pasquer, Q. T. L., Tsakoumagkos, I. A. & Hoogendoorn, S. From phenotypic hit to chemical probe: Chemical biology approaches to elucidate small molecule action in complex biological systems. *Molecules* **25**, 5702 (2020).
22. Rohban, M. H. et al. Virtual screening for small-molecule pathway regulators by image-profile matching. *Cell Syst.* **13**, 724–736 (2022).
23. Ljosa, V. et al. Comparison of methods for image-based profiling of cellular morphological responses to small-molecule treatment. *J. Biomol. Screen.* **18**, 1321–1329 (2013).
24. Warchal, S. J., Dawson, J. C. & Carragher, N. O. Evaluation of machine learning classifiers to predict compound mechanism of action when transferred across distinct cell lines. *SLAS Discov.* **24**, 224–233 (2019).

25. Aliper, A. et al. Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Mol. Pharm.* **13**, 2524–2530 (2016).
26. Lapins, M. & Spjuth, O. Evaluation of gene expression and phenotypic profiling data as quantitative descriptors for predicting drug targets and mechanisms of action. Preprint at *bioRxiv* https://doi.org/10.1101/580654 (2019).
27. Cantini, L. et al. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nat. Commun.* **12**, 124 (2021).
28. Tenenhaus, M., Tenenhaus, A. & Groenen, P. J. F. Regularized generalized canonical correlation analysis: a framework for sequential multiblock component methods. *Psychometrika* https://doi.org/10.1007/s11336-017-9573-x (2017).
29. Roselle, C., Verch, T. & Shank-Retzlaff, M. Mitigation of microtiter plate-positioning effects using a block randomization scheme. *Anal. Bioanal. Chem.* **408**, 3969–3979 (2016).
30. Lill, J. R., Mathews, W. R., Rose, C. M. & Schirle, M. Proteomics in the pharmaceutical and biotechnology industry: a look to the next decade. *Expert Rev. Proteom.* **18**, 503–526 (2021).
31. Chandrasekaran, S. N., Ceulemans, H., Boyd, J. D. & Carpenter, A. E. Image-based profiling for drug discovery: due for a machine learning upgrade? *Nat. Rev. Drug Discov.* **20**, 145–159 (2021).
32. Tandon, G., Chan, P. & Mitra, D. MORPHEUS: motif oriented representations to purge hostile events from unlabeled sequences. in *Proceedings of the 2004 ACM Workshop on Visualization and Data Mining for Computer Security* https://doi.org/10.1145/1029208.1029212 (2004).

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

Methods

Dataset preprocessing

We gathered four available datasets that had both CP morphological profiles and L1000 GE profiles, preprocessed the data from different sources and in different formats in a unified .csv format, and made the data publicly available at amazon s3 bucket: s3://cellpainting-gallery/cpg0003-rosetta/broad/workspace/preprocessed_data/.

The rows of each csv file are replicate-level (that is, well-level) profiles, augmented with metadata available for that well.

Cell Painting and L1000 profiles

Single-cell morphological (CP) profiles were created using CellProfiler software and processed to form aggregated replicate profiles using the R cytominer package (https://cran.r-project.org/package=cytominer).

We made the following three types of profiles available:

- Aggregated profiles, which are the average of single-cell profiles in each replicate well (replicate_level_cp_augmented.csv.gz).
- Normalized profiles, which are the *z*-scored aggregated profiles, where the scores are calculated using the distribution of negative controls as the reference (replicate_level_cp_normalized.csv.gz).
- Normalized variable-selected profiles, which are normalized profiles with features selection applied (replicate_level_cp_normalized_variable_selected.csv.gz).

For L1000, we used the previously processed 978 'landmark' genes as our input features. The complete processing details are provided in ref. [7]. The L1000 landmark genes in the CDRP dataset are different from the landmark genes in the other datasets, with an overlap of $n = 785$ (80%). The CDRP dataset was acquired using the so-called 'delta prime' probe set ($n = 977$). Subsequent datasets (LUAD, TAORF and LINCS) were acquired using the so-called 'epsilon' probe set ($n = 978$)[33].

The 20% of the delta prime landmark genes that are absent in epsilon can be inferred using the epsilon landmark genes[7].

To simplify our analysis, we did not perform this inference, and instead only used the landmark genes available for each dataset. When combining CDRP with other datasets, we used the intersection of the two probe sets.

Data processing for analysis

We used treatment-level profiles for both the GE (using L1000) and morphology (CP) modalities for the analysis, although replicate-level profiles are provided and could be used instead in other formulations of the problem to create more advanced models.

Treatment-level profiles are the average of replicate-level profiles. For CP, replicate-level profiles are the average of single-cell measurements of cells from that replicate well. For GE, replicate-level profiles are simply the bulk GE profile for that replicate well.

We standardized replicate-level profiles for each plate to have zero mean and unit variance before averaging them to form treatment-level profiles.

Note that, aside from some image segmentation parameters in the CellProfiler pipeline, which are adjusted for each cell type based on its baseline morphology, the computational pipelines for data processing and analysis were identical regardless of the cell type in the experiment.

Measuring quality of data points for subsequent analysis

We use treatment-level profiles for all the analysis that follows. Henceforth, 'data points' refer to treatment-level profiles unless indicated otherwise. The specific transformation of the treatment-level profiles (such as 'normalized_variable_selected') is clarified when necessary.

There are inherent differences in the biological design (type of perturbation, cell line used and time point of exposure to perturbation) and experimental parameters (different instrumentation, reagent batches and personnel running the experiments creating distinct technical artifacts such as batch effects) that lead to differences in the datasets. Consistency of profiles of a single treatment across different batches of experiment is considered a measure of data quality. We checked this consistency as follows. After standardization of the replicate-level profiles per plate, we calculate the Pearson correlation coefficient between each pair of replicate-level profiles for the same perturbation. The distribution of these coefficients for each dataset and modality is illustrated in Supplementary Fig. 1. The corresponding blue curve to each red curve is the null distribution showing the correlation coefficient between pairs of profiles that belong to different perturbations. The nonzero dotted vertical line to the right shows the 90th percentile of the null distribution. We considered the perturbations that had an average replicate correlation more than the 90th percentile of the null distribution as high-quality data points for subsequent analysis.

We note a source of systematic error present in all datasets that may affect replicability metrics: for nearly every treatment, all its replicates occurred at the same well position on the plate (because replicates in such high-throughput experiments are created by replicating the entire, and exact same, plate layout, for logistical reasons). The location of the well on the plate can impact the cells in the well. For example, wells on the edge are more likely to dry slightly, impacting cell morphology. This effect—the impact of an experiment covariate on the readout of the assay—can inflate replicability quality metrics. In our experience, well-positioned effects tend to be more pronounced in CP than L1000, and therefore the observed differences in data quality (Supplementary Data 2) can be a function of this batch effect. As noted in the discussion, correcting for batch effects could improve the prediction tasks discussed herein, and also make such comparisons of data quality more reliable.

Filtering data points

To remove noisy data points from the analysis, we used two filtering strategies for each shared subspace and data integration analysis. For cross-modality prediction experiments, we used the intersection of higher quality data points according to both modalities. For the analysis for data integration, we used data points that were higher quality (that is, >90th percentile of the null distribution, as defined above) in at least one of the modalities. A definition of higher quality data points is provided in the previous section. A comprehensive description of the data sizes in each modality, number of overlapping perturbations across both modalities, size of intersection and union sets of higher quality data points across both modalities are given in Supplementary 1 and highlights are summarized in Supplementary Table 1.

One of the chemical datasets (CDRP-BBBC047-Bray) has a subset of compounds that are known to be bioactive. We referred to this subset as CDRP-bio-BBBC036-Bray and reported the details independently for this dataset (Supplementary Data 1 and 2). We only used CDRP-bio and not the full CDRP set for the analysis, because we believe that the quality of CDRP is insufficient for either of these analyses given that very few data points remained after filtering for replicate reproducibility across both modalities (Supplementary Fig. 1).

Cross-modality predictions

For prediction of each single landmark gene using CP profiles or each single morphological feature using GE profiles, we used two regression models of: CP to GE: $\mathbf{y}_l = f(X_{cp}) + \mathbf{e}_l$; in which $\mathbf{y}_l$ is a vector of expression levels for the landmark gene l across all the perturbations in a dataset and X_{cp} is the whole morphological data matrix where each row is a treatment-level CP profile. For this prediction direction, we used the so-called 'normalized variable-selected' treatment-level CP profiles, which resulted in 601 features for CDRP-bio dataset, 291 features for LUAD dataset, 63 features for TAORF and 119 features for the LINCS dataset. The variable selection step removes features with near-zero variance and reduces redundancy in the feature set (ensuring that no pair of features has a Pearson correlation coefficient > 0.9).

GE to CP: $\mathbf{y_f} = f(X_{\mathrm{ge}}) + \mathbf{e_f}$; in which $\mathbf{y_f}$ is a vector of morphological feature l across all the perturbations in a dataset and X_{ge} is the whole GE data matrix where each row is a treatment-level L1000 profile. For this prediction direction, we have not performed any dimensionality reduction on the GE data.

For each prediction direction (CP to GE, GE to CP) and each baseline linear (Lasso) and nonlinear (MLP) model for this regression problem, we used the coefficient of determination (R^2) and nested k-fold cross-validation over the data points for evaluating the prediction model performance. Therefore, for each landmark gene (for CP to GE) or each morphological feature (for GE to CP), we can form a distribution of k, R^2 values. We also shuffled the vector y_l for each gene l across all the data points and applied the same cross-validation procedure to form a null distribution for each gene. The same procedure on y_f will result in the null distribution for each morphological feature. Model parameters were selected using grid search and cross-validation on each training set for each of the k test folds.

In Supplementary Data 4, the median prediction scores of each model for each landmark gene for each dataset and according to each model are presented. The distribution of MLP model prediction scores for the 50 landmark genes with the highest median scores in each dataset is shown in Supplementary Data 3.

Percentage predictable

Percentage predictable is defined as the percentage of landmark genes that have a median of R^2 predictability score more than a defined threshold. The threshold is based on the null distribution of predictability scores for each dataset. The dataset-specific null is formed using medians of single gene null distributions. We take the 99th percentile of this null distribution plus a 0.2 margin ($t_{99\mathrm{th}} + 0.2$) as the threshold for calling a gene 'predictable'. We reported the 'percentage predictable' values for each dataset in Fig. 2b.

Modality integration

For the analysis for MoA prediction, we used the data points that had high quality according to their replicate-level profiles (that is, >90th percentile of the null distribution; see above) in either modality. In compound datasets, each perturbation is tested at multiple doses and therefore there are multiple data points corresponding to each compound. A data point here is a treatment-level profile corresponding to a dose of a compound.

The LINCS dataset has MoA annotations for 1,401 overlapping compounds across two modalities. Every compound is tested at seven different doses, increasing the chances of detecting the expected behavior of the compound at one of them. Each compound can have multiple mechanisms; therefore, we have multiple labels for a subset of compounds. The set of labels comprises 478 unique MoAs. There are 568 unique combinations of these labels in the dataset. We started with the filtered union set, and filtered it again to keep MoA classes that had at least five data points in their class. Because this process resulted in only one MoA category that was a multi-label one (that is, composed of multiple MoAs), we removed this category to simplify the problem as being multi-class but single label; that is, we effectively used only compounds labeled with a single MoA. The filtered set included 1,655 data points across 521 compounds in 57 MoA categories.

The CDRP-bio dataset included MoA annotations for 1,327 of 1,916 overlapping compounds across two modalities. After passing data points from three filters—union of higher quality data across modalities, available MoA labels, and being in an MoA class that has at least five compounds in the set—we were left with 123 compounds in 16 MoA categories.

Unsupervised joint dimensionality reduction of modalities and mechanism of action of cluster retrieval

k-means clustering (k = number of MoA classes) was performed on each modality space and integrated spaces using representation-level concatenation of modalities (early fusion) and seven state-of-the-art modality integration methods[27]. Jaccard Index values between k-means clustering results and the MoA annotation labels were used as a measure of ground-truth cluster retrieval for this unsupervised clustering task.

Supervised mechanism of action prediction

For the multi-class MoA classification problem, two logistic regression and MLP classifiers were used as baseline models; we apply each model for predicting MoA labels using each modality of data independently as well as the baselines for integration of the two. We performed stratified nested k-fold cross-validation ($k = 5$) to evaluate the classification performance using the F1-score metric. Note that all doses of a compound should be in the same fold in this data partitioning scheme. Model hyper-parameters were optimized using grid search and cross-validation in each training fold.

Some MoAs have several tens of compounds, whereas others have as few as five. To address this imbalance in the data, for both logistic regression and MLP models, we oversampled data points in each class to match the size of the majority class in the training set. The k-fold cross-validation experiment resulted in k vector of multi-class predictions. We then calculated F1-scores for each class independently and the averaged class-specific F1-scores within each fold formed k F1-scores (Fig. 4b).

For representation-level integration strategies, we simply concatenated CP and GE profiles in their original spaces (early fusion) or projected into RGCCA space (RGCCA_EarlyFusion) to integrate both modalities. On the other hand, late fusion is at the decision level and averages predicted class probabilities (based on the output of classifiers trained on each modality separately) for making the MoA class decision for test compounds.

Gene Ontology terms search analysis

The goal of this analysis was to see if the observed CP–GE link is consistent with the known functional characteristic of L1000 landmark genes under study in this work. We used the DAVID[34] Functional Annotation Tool (2021 update) to form a table of GO annotation terms for 978 landmark genes in the LUAD, LINCS and TAORF datasets. For 921 DAVID IDs detected, GO categories of 'GOTERM_BP_DIRECT', 'GOTERM_CC_DIRECT' and 'GOTERM_MF_DIRECT' were selected and the 'Functional Annotations Table' was used as the source file for the following search analysis.

For each landmark gene and each CP channel, we searched for all relevant keywords to that channel and formed channel-specific annotation columns. While mRNA-level prediction of landmark genes using specific categories of morphological features were created for the LUAD dataset (Fig. 2d), we formed channel-specific GE prediction scores by rearrangement of these categories to CP channels and taking the maximum prediction score within each channel-specific category of features. We then have access to channel-specific GO functional annotations and channel-specific prediction scores for each of the landmark genes. We discretized prediction scores into three categories of predictability: high ($R^2 > 0.8$), medium ($0.1 \le R^2 \le 0.8$), and low ($R^2 < 0.1$). Using Fisher's exact test, we calculated the odds ratio as a measure of association between a landmark gene being predictable and having channel-specific GO annotations. An odds ratio > 1 indicates an increased chance of having an organelle annotation (using GO terms) for a highly predictable landmark gene.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Preprocessed profiles that are augmented with gene and compound annotation are freely available through the 'Registry of Open Data on AWS' on a public S3 bucket. Documentation on the folder structure, dataset details and instructions for accessing the data are available at https://broad.io/rosetta/. Datasets are described and referenced in Supplementary Data 1. Source data are provided with this paper.

Code availability

Source code to reproduce and build upon the presented results is available at https://broad.io/rosetta/. We licensed the source code as BSD 3-Clause, and licensed the data, results and figures as CC0 1.0.

References

33. Broad Institute. Guide to LINCS data release into NCBI GEO—L1000. Connectopedia. https://clue.io/connectopedia/guide_to_geo_l1000_data
34. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).

Acknowledgements

We thank all the researchers who created and shared the data, who are mentioned in their respective publications cited in the paper. Funding was provided by grants (2018-183451 to A.E.C. and 2020-225720 to B.A.C.) from the Chan Zuckerberg Initiative DAF, an advised fund of the Silicon Valley Community Foundation and the National Institutes of Health NIGMS (R35 GM122547, to A.E.C.).

Author contributions

M.H., S.S., B.A.C. and A.E.C. contributed to drafting the manuscript and designing the research. J.C.C. initiated the project and performed early explorations of the LUAD dataset. M.H. analyzed and explored the data with inputs from the other co-authors.

Competing interests

S.S. and A.E.C. serve as scientific advisors for companies that use image-based profiling and Cell Painting (A.E.C.: Recursion, S.S.: Waypoint Bio, Dewpoint Therapeutics), and receive honoraria for occasional talks at pharmaceutical and biotechnology companies.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41592-022-01667-0.

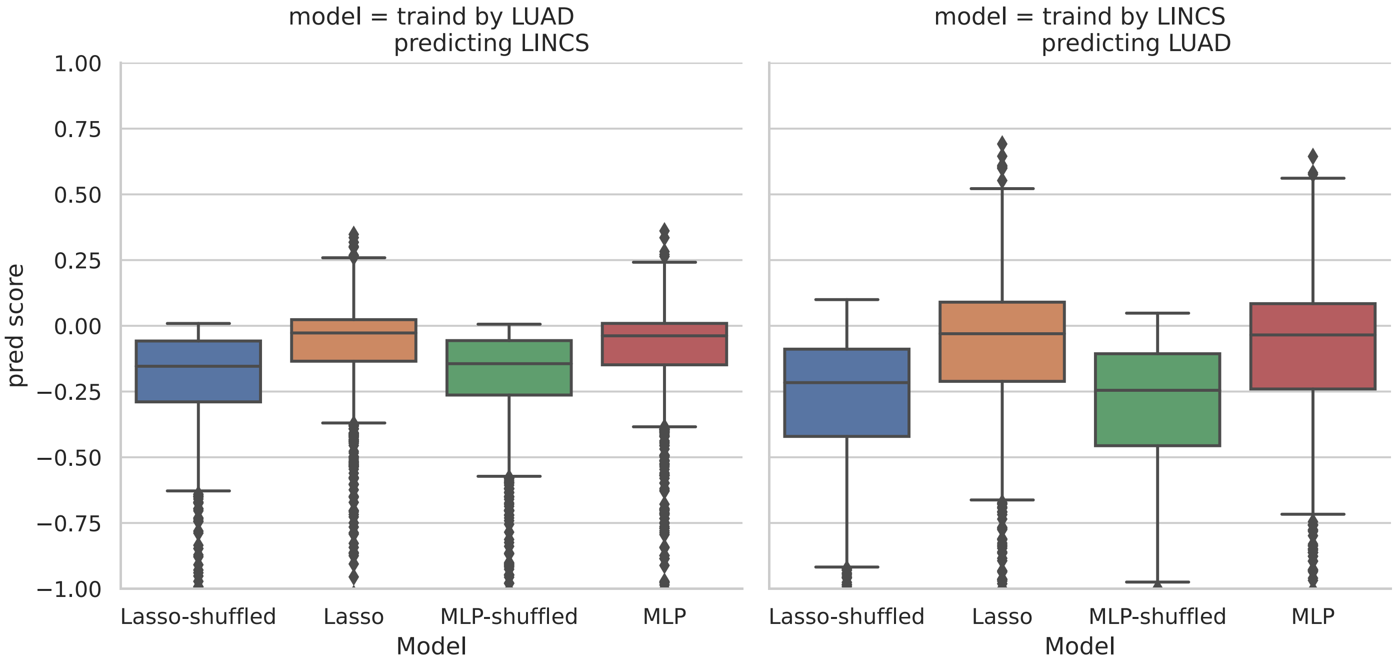
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-022-01667-0.

Correspondence and requests for materials should be addressed to Marzieh Haghighi or Shantanu Singh.

Peer review information *Nature Methods* thanks Haiquan Li, Matthew McCall, and the other, anonymous, reviewer for their contribution to the peer review of this work. Primary Handling Editor: Rita Strack, in collaboration with the *Nature Methods* team. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

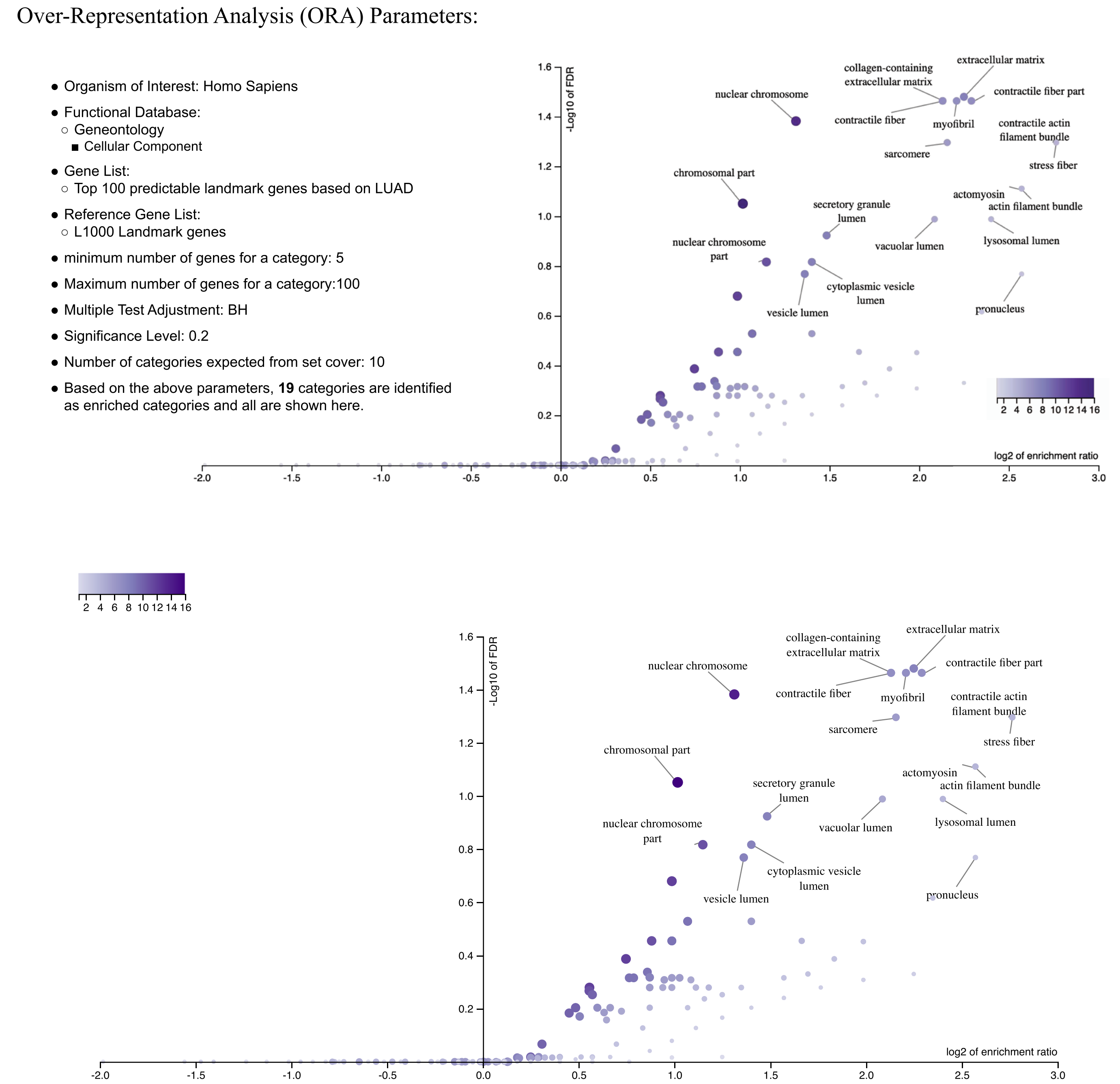
https://doi.org/10.1038/s41592-022-01667-0



Extended Data Fig. 1 | Generalizability of the prediction model across datasets. Prediction of each L1000 mRNA level by Cell Painting features in dataset A, using a model trained on dataset B. We have trained Lasso and MLP models on each of LUAD and LINCS datasets and checked the prediction results on the other dataset which was not used in model training. Distribution of R2 prediction scores for all landmark genes are shown. Comparison of the results here with Fig. 2 indicates weakness of the prediction model in generalizability across datasets. This is an indication of dataset-specific technical variations (batch effects) that need exploration of experimental alignment techniques (batch-effect correction), which is an active area of research. We also observe that the model's prediction power is stronger when the model is trained on the LINCS dataset and tested on the LUAD dataset. This is expected as the LUAD dataset is limited to a narrow set of genes associated with lung adenocarcinoma cancer; however, the LINCS dataset contains a wide variety of compounds with different mechanisms and known phenotypes. The y-axis is trimmed at −1 for clarity. Distributions are presented as boxplots, with center line being median, box limits being upper and lower quartiles and whiskers being 1.5× interquartile range; n = 978 landmark genes for each boxplot.

gene_group_name	Genes	gene_group_name	Genes
LIM domain containing	FHL2,PXN,TES	SRY-box transcription factors	SOX4
Receptor ligands	C5,MMP1,RGS2	SNAREs	YKT6
Basic leucine zipper proteins	ATF1,CEBPD,JUN	Minor histocompatibility antigens	ERBB2
MicroRNA protein coding host genes	ERBB2,GPC1,SMC4	AAA ATPases	RUVBL1
Nucleoporins	NUP88,RAE1	Mannosidases alpha class 2	MAN2B1
Protein phosphatase 1 regulatory subunits	AURKA,AURKB	MCM family	MCM3
Chromosomal passenger complex	AURKB,BIRC5	Cyclophilin peptidylprolyl isomerases	PPIC
Cyclins	CCNA2,CCND3	Condensin II subunits	SMC4
Solute carriers	SLC2A6,SLC35A1	Condensin I subunits	SMC4
Pleckstrin homology domain containing	GRB10,NET1	Complement system activation components	C5
MAP kinase phosphatases	DUSP4,DUSP6	Class III Cys-based CDC25 phosphatases	CDC25A
Ubiquitin conjugating enzymes E2	UBE2C,UBE2L6	Chemokine ligands	CCL2
Ankyrin repeat domain containing	ILK,RAI14	Cathepsins	CTSD
Myb/SANT domain containing	EZH2,MYBL2	Canonical high mobility group	HMGA2
SH2 domain containing	GRB10,STAT1	CD molecules	ERBB2
Collagens	COL1A1,COL4A1	CCAAT/enhancer binding proteins	CEBPD
Purinosome	PAICS	C3 and PZP like, alpha-2-macroglobulin domain c...	C5
RNA binding motif containing	RNPS1	Blood group antigens	ABCB6
R2TP complex	RUVBL1	Baculoviral IAP repeat containing	BIRC5
Poly(ADP-ribose) polymerases	TIPARP	BCH domain containing	ARHGAP1
Protein tyrosine phosphatases non-receptor type	PTPN12	Armadillo-like helical domain containing	RRP12
Polycomb repressive complex 2	EZH2	Abhydrolase domain containing	ABHD4
Rho GTPase activating proteins	ARHGAP1	ATP binding cassette subfamily B	ABCB6
Phosphoinositide phosphatases	INPP1	DEAD-box helicases	DDX10
NuRD complex	HDAC2	DNA helicases	RUVBL1
Myosin light chains	MYL9	DNA polymerases	POLE2
Mitochondrial ribosomal proteins	MRPL12	Glypicans	GPC1
Regulators of G-protein signaling	RGS2	ASAP complex	RNPS1
SRCAP complex	RUVBL1	M10 matrix metallopeptidases	MMP1
SET domain containing	EZH2	Lysine methyltransferases	EZH2
SIN3 histone deacetylase complex subunits	HDAC2	L ribosomal proteins	RPL39L
Zinc fingers ZZ-type	SQSTM1	Jun transcription factor family	JUN
WD repeat domain containing	RAE1	INO80 complex	RUVBL1
Ubiquitin specific peptidases	USP1	Histone deacetylases, class I	HDAC2
Tudor domain containing	LBR	Glutaredoxin domain containing	TXNRD1
Tropomyosins	TPM1	DNAJ (HSP40) heat shock proteins	DNAJB2
Topoisomerases	TOP2A	Erb-b2 receptor tyrosine kinases	ERBB2
Tissue inhibitor of metallopeptidases	TIMP2	Endothelins	EDN1
Structural maintenance of chromosomes proteins	SMC4	EMSY complex	HDAC2
Small heat shock proteins	HSPB1	EF-hand domain containing	MYL9
Serpin peptidase inhibitors	SERPINE1	Diphthamide biosynthesis pathway genes	DPH2
Serine proteases	PRSS23	Deoxyribonucleoside kinases	DCK
Selenoproteins	TXNRD1	Dbl family Rho GEFs	NET1
Scavenger receptors	SCARB1	tRNA-splicing endonuclease subunits	TSEN2

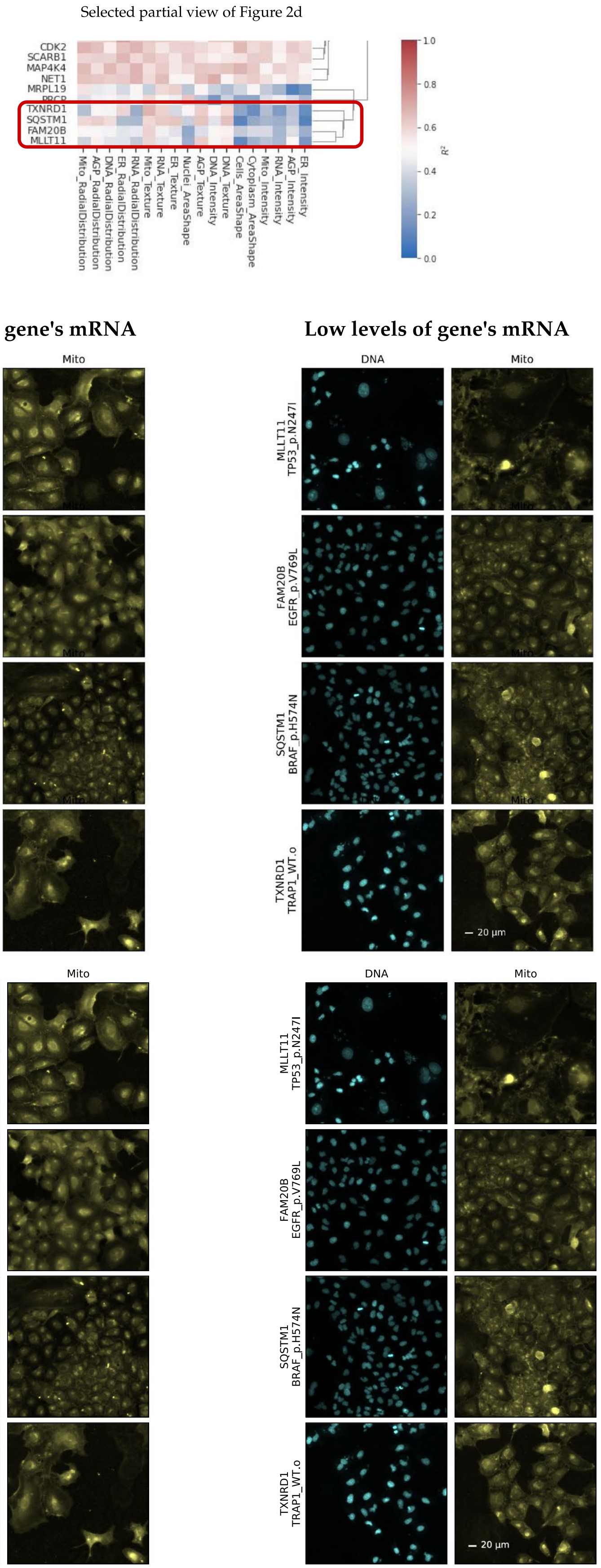
Extended Data Fig. 2 | Gene group names for top 100 predictable landmark genes in LUAD dataset. Top 100 predictable landmark genes by MLP model are shown along with their gene group names (based on HGNC Database41) for the LUAD dataset, finding a diverse array represented, though we note the perturbations in this experiment included only genes found mutated in lung cancers.



Over-Representation Analysis (ORA) Parameters:

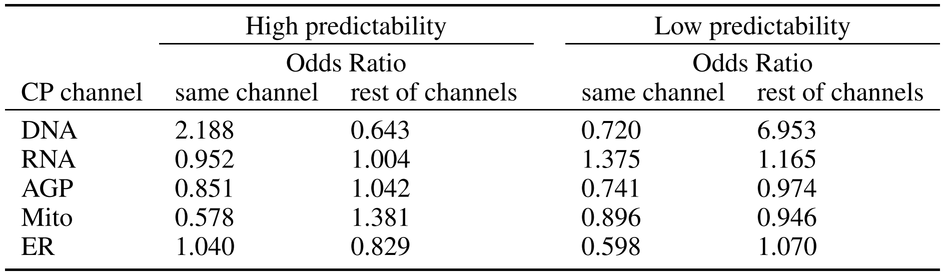
- Organism of Interest: Homo Sapiens
- Functional Database:
  - ○ Geneontology
    - ■ Cellular Component
- Gene List:
  - ○ Top 100 predictable landmark genes based on LUAD
- Reference Gene List:
  - ○ L1000 Landmark genes
- minimum number of genes for a category: 5
- Maximum number of genes for a category:100
- Multiple Test Adjustment: BH
- Significance Level: 0.2
- Number of categories expected from set cover: 10
- Based on the above parameters, **19** categories are identified as enriched categories and all are shown here.

Extended Data Fig. 3 | Over-Representation Analysis (ORA) of highly predictable (top 100) landmark genes in LUAD dataset. Over-Representation Analysis of top 100 highly predictable landmark genes according to the MLP model applied on the LUAD dataset. ORA analysis was performed by WebGestalt analysis toolkit 42. Nineteen enriched categories (FDR < 0.2) are labeled in the volcano plot.



Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Visualization of cells in a cluster of landmark genes that are tightly correlated with RNA texture category of morphological features. For the cluster of landmark genes shown in the top heatmap, which is a partial snapshot of Fig. 2d, we have shown example cell images for perturbations that have high and low predicted values for each gene in that cluster. We have filtered perturbations to those that have low prediction errors prior to that selection. We can observe that cells that are predicted to have (and actually do have) high levels of these five genes' mRNA all are associated with visible changes in the staining for mitochondria, even though only half of these genes already have functional annotations related to the mitochondria.



	High predictability		Low predictability	
	Odds Ratio		Odds Ratio	
CP channel	same channel	rest of channels	same channel	rest of channels
DNA	2.188	0.643	0.720	6.953
RNA	0.952	1.004	1.375	1.165
AGP	0.851	1.042	0.741	0.974
Mito	0.578	1.381	0.896	0.946
ER	1.040	0.829	0.598	1.070

Extended Data Fig. 5 | Validation of the observed GE-CP relationship by GO-terms search analysis. Landmark genes highly predictable according to morphological features in each specific Cell Painting channel are more likely to have GO annotation related to that channel compared to the rest of CP channels. For each channel in the rows of the table, the first column shows the Odds Ratio (OR) derived from the Fisher's exact test for associations between the landmark genes being highly predictable (R2 > 0.6) by CP features in a channel and having GO annotations for that channel. The second column shows the association between the same set of highly predictable genes and having GO annotation for any channel but not the target row channel. Higher values in the first column compared to the second column show that highly predictable genes according to features in a CP channel are more likely to have GO annotations for that channel compared to the rest of the channels. This pattern holds for DNA and ER channels but not for the rest of CP channels. The third and fourth columns show the same associations but for low-predictability genes (R2 < 0). Lower values in the third column compared to the fourth column show that non-predictable genes according to features in a CP channel are less likely to have GO annotations for that channel compared to the rest of the channels. This pattern holds for all CP channels except for RNA. The CP channel specific predictability map used for this analysis was derived from the result of the experiment and results presented partially in Fig. 2d. As we can observe from the map, usually multiple categories of morphological features contribute to the predictability of a gene, which explains the lack of a simple relationship between a given channel's predictability and GO term associations presented in this table.

Stains	Mitochondria	Golgi	Membrane	Cytoskeleton/Actin	Endoplasmic	RNA/Nucleoli	DNA	Any stain	No stain
Odds Ratio	1.58	0.69	1.22	1.59	1.02	0.64	1.95	2.5	0.4

Extended Data Fig. 6 | Association between landmark gene predictability and having gene ontology annotations related to Cell-Painting stains. Landmark genes that are predictable according to at least three of the four datasets (59 genes shown in Fig. 2c) are more likely to have GO annotations related to any of the stains in the Cell Painting assay compared to a random subset of landmark genes.

nature research

Corresponding author(s):	Marzieh haghighi Shantanu Singh
Last updated by author(s):	Aug 10, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

☐ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☒ ☐ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☐ ☒ A description of all covariates tested

☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☒ ☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☒ ☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	See "Supplementary A. Curated Datasets" section.
Data analysis	We used open source CellProfiler software (version 2.1 and 2.2, exact versions per data set are specified in Supplementary A) for extracting single cell features from images of each of the datasets and Cytominer (Cytominer v0.1.0 R 3.4.1) package for generating replicate level profiles . The code for analysis of data is public at: https://github.com/carpenterlab/2021_Haghighi_submitted

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Preprocessed profiles that are augmented with gene and compound annotation are available through the Registry of Open Data on AWS on a public S3 bucket at no cost and no need for registration. Documentation on the folder structure, dataset details and instructions for accessing the data are available at http://broad.io/rosetta. Source datasets are described and referenced in Supplementary A.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was determined by practical limitations (how many cells fit in one sample well).
Data exclusions	No data were excluded. Data quality issues were noted in the paper.
Replication	For each provided dataset, each sample has at least 2 replicates. All replicate experiments performed are described in this paper, none were omitted.
Randomization	Treatments, including negative controls, were generally randomly distributed across well positions on the 384-well plates. We note two non-random patterns LINCS: All 6 doses of a treatment are in adjacent cells in the same row. This was due to constraints imposed by compound management robotics LUAD: All alleles of a gene are generally on the same plate. This was done intentionally so that comparing the wild-type and mutant forms of the gene were not subject to plate-to-plate variation.
Blinding	Image segmentation workflows were designed by experts who were blinded at the time to the identity of each sample; a small number of blinded samples were sub-selected to choose parameters to apply to all samples in a batch. All downstream steps were then performed identically for all samples in a data set.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
□	☒	Eukaryotic cell lines
☒	□	Palaeontology and archaeology
☒	□	Animals and other organisms
☒	□	Human research participants
☒	□	Clinical data
☒	□	Dual use research of concern

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	CDRP utilized U-2OS cells from ATCC (HTB-96). LINCS utilized A549 cells from ATCC (CCL-185). TAORF utilized U-2 OS cells originally obtained from ATCC and propagated in the William Hahn lab. LUAD utilized A549 cells from ATCC and propagated in the Genetic Perturbation Platform at Broad Institute.
Authentication	The LUAD A549 cells were part of the Cancer Cell Line Encyclopedia project, which involved genetic/sequencing analysis. We are unaware of any other authentication done on other sets.
Mycoplasma contamination	Testing for mycoplasma was confirmed for the CDRP, CDRP-bio, and TAORF data sets. We lack the historical information to confirm whether the LUAD and LINCS sets were tested.
Commonly misidentified lines (See ICLAC register)	None are used (we used A549 and U2OS).