Open access • Posted Content • DOI:10.1101/2021.09.08.459417

High-Dimensional Gene Expression and Morphology Profiles of Cells across 28,000 Genetic and Chemical Perturbations — Source link

Marzieh Haghighi, Shantanu Singh, Juan C. Caicedo, Anne E. Carpenter

Institutions: Broad Institute

Published on: 08 Jun 2021 - bioRxiv (Cold Spring Harbor Laboratory)

High-Dimensional Gene Expression and Morphology Profiles of Cells across 28,000 Genetic and Chemical Perturbations

Marzieh Haghighi, Shantanu Singh, Juan Caicedo, Anne Carpenter

Broad Institute of MIT and Harvard

Abstract

Populations of cells can be perturbed by various chemical and genetic treatments and the impact on the cells' gene expression (transcription, i.e. mRNA levels) and morphology (in an image-based assay) can be measured in high dimensions. The patterns observed in this profile data can be used for more than a dozen applications in drug discovery and basic biology research, but both types of profiles are rarely available for large-scale experiments. We provide a collection of four datasets with both gene expression and morphological profile data useful for developing and testing multi-modal methodologies. Roughly a thousand features are measured for each of the two data types, across more than 28,000 thousand chemical and genetic perturbations. We define biological problems that can be investigated using 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 (http://broad.io/rosetta).

Introduction

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 prior 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 multi-modal (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 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-omics data has sparked research investigating the relationship and integration of different high-dimensional readouts [5]. For example, transcriptomics, proteomics, epigenomics and metabolomics data can be combined to predict the mechanisms of action (MoAs) of chemical compounds [6].

Here, we created a collection of gene expression and morphology datasets with the scale and annotations needed for machine learning research in multi-modal data analysis and integration. 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 multi-modal data by defining applications 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 and provide baseline methods, code, evaluation metrics, and benchmark results for each, as a foundation for future biologically-oriented machine learning research.

Results

Data generation for gene expression and morphological profiles

All datasets were created at our institution (see Methods) and involved one of two types of "inputs": chemical perturbations and genetic perturbations (Figure 1). There are also two types of high-dimensional outputs measured: gene expression profiles and morphological profiles, each with roughly 1000 features measured. We note that "genes" are an input (individual genes are overexpressed as the perturbation in some datasets)

and an output (gene expression profiles are comprised of the measured mRNA levels for each gene); this can cause confusion for researchers new to the domain.

We captured gene expression (GE) profiles using the L1000 assay [7]. Each cell's DNA is transcribed into various mRNA molecules which can be translated into proteins that carry out functions in the cell. The levels of mRNA in the cell are often biologically meaningful - collectively, mRNA levels for a cell are known as its transcriptional state; each individual mRNA level is referred to as the corresponding gene's "expression". 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 978 "landmark" genes capture approximately 82% of the transcriptional variance for the entire genome [7].

We captured morphological profiles using Cell Painting (CP) [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 profiles are then aggregated for all the single cells in the sample.

For both data types, aggregation of all the replicate-level profiles of a perturbation is called a treatment-level profile. In our study, we used treatment-level profiles in all experiments but have provided replicate-level profiles for researchers interested in further data exploration. We note that of the eight datasets provided (four datasets x two modalities), four have been the subject of previous research published by researchers at our Institute [3,10,11]; here we complete the matrix by providing the missing data type for each pair, organize them, and provide benchmarks.

Information content of data modalities: Shared versus Complementary

Cell morphology and gene expression 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 gene expression changes and gene expression changes can induce a change in cell morphology, but neither is always the case. Even if technical artifacts were non-existent, we do not expect a one-to-one map between these two modalities. We therefore hypothesize that the information in each data type consists of a shared subspace, a modality-specific complementary subspace, and noise (Figure 1). Both subspaces can be exploited for biological applications.

Shared subspace

The shared subspace between gene expression and cell morphology is beginning to be explored. For example, probabilistic canonical correlation analysis learned a shared structure in paired samples of histology images and bulk gene expression RNA-seq data, suggesting that shared latent variables form a composite phenotype between morphology and gene expression that can be useful [12]. In another study, cross-modal autoencoders learned the shared latent space for single-cell RNA-seq and chromatin images in order to integrate and translate across modalities [13].

The existence of a shared subspace enables multiple applications. Most prominently, if sufficient shared information is present, one modality can be computationally predicted (i.e. inferred, estimated) using another, saving significant 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 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[14]; 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[15].

Modality-specific, complementary subspaces

Each modality will 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 compounds 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.

Figure 1. Multi-modal datasets overview

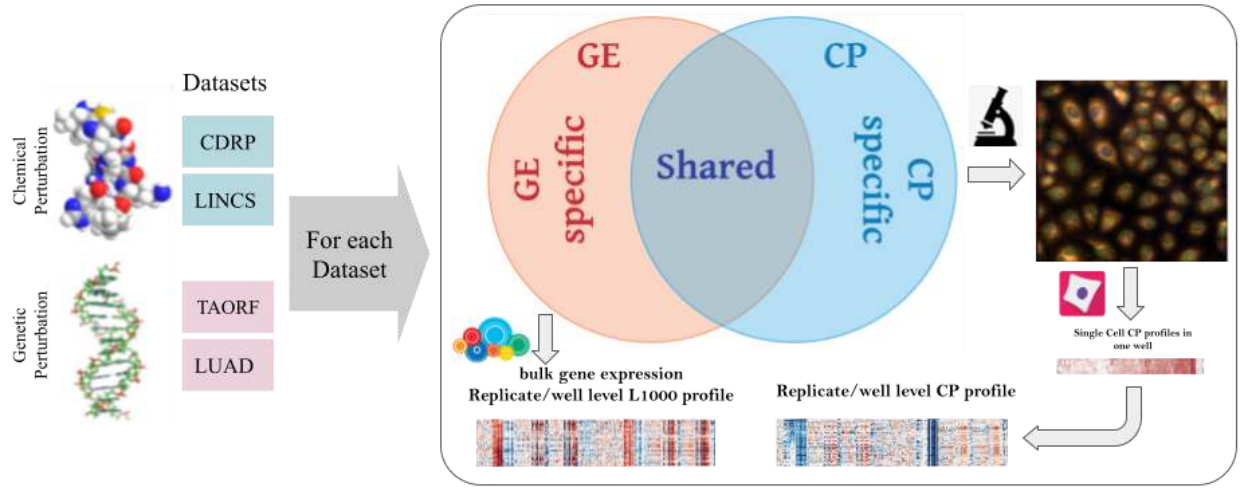
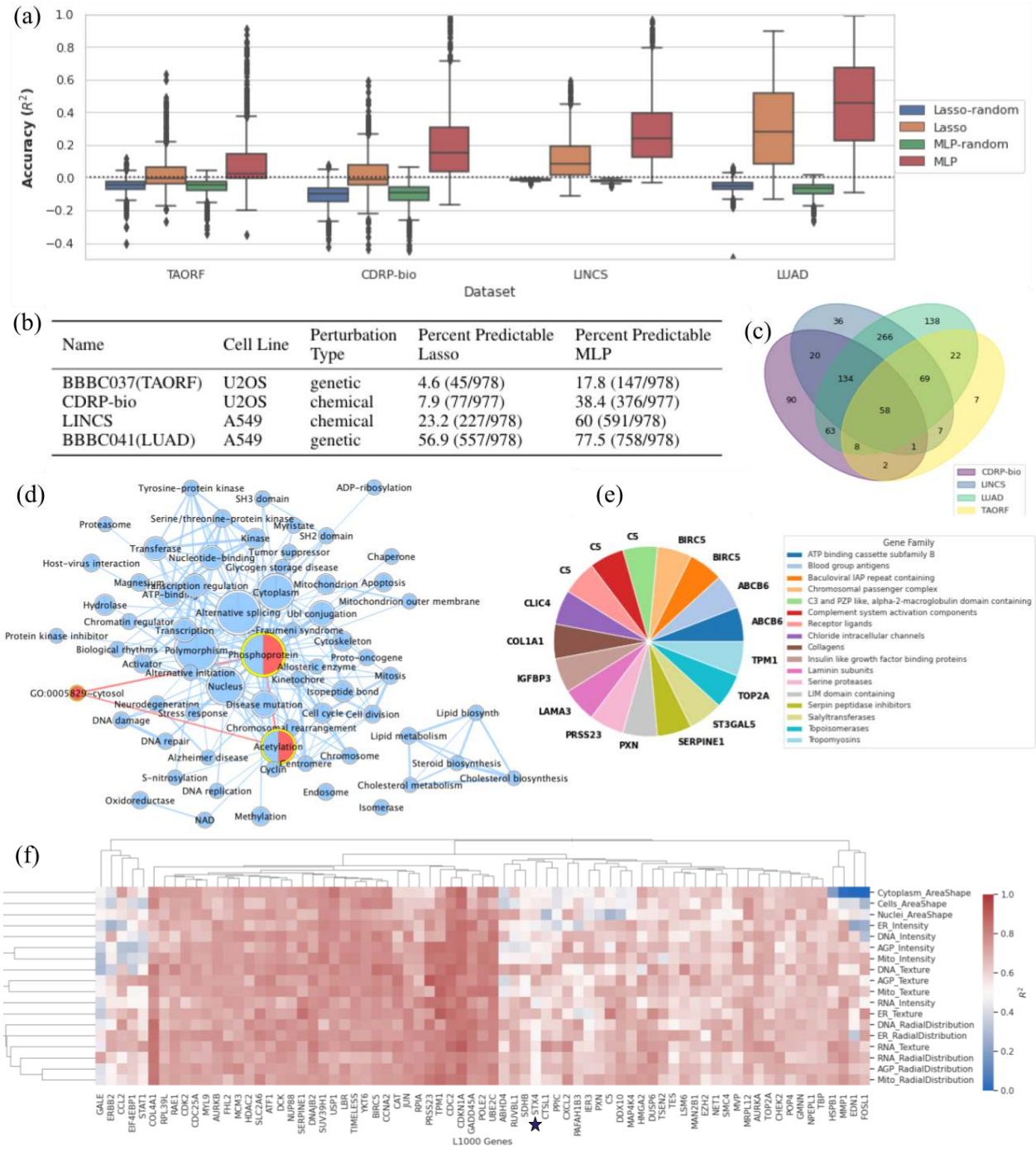


Figure 1. Multi-modal genetic and chemical perturbation datasets are valuable for many applications. For each dataset, Cell Painting and L1000 assays were used to collect morphological and gene expression representations (profiles), respectively.

Application 1: Predicting gene expression and morphology from each other

As a baseline for finding the correspondence between modalities 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 gene expression profile can be estimated as a function of all the morphological features in the Cell Painting profile, $y_l = f_\theta(X_{cp}) + e_l$; in which 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 representing all morphological changes across all the perturbations. We use Lasso as a baseline linear model and multilayer perceptron (MLP) as a baseline nonlinear model for the regression problem.

Figure 2. An application using the shared subspace: cross-modality predictions from CP to GE



Name	Cell Line	Perturbation Type	Percent Predictable Lasso	Percent Predictable MLP
BBBC037(TAORF)	U2OS	genetic	4.6 (45/978)	17.8 (147/978)
CDRP-bio	U2OS	chemical	7.9 (77/977)	38.4 (376/977)
LINCS	A549	chemical	23.2 (227/978)	60 (591/978)
BBBC041(LUAD)	A549	genetic	56.9 (557/978)	77.5 (758/978)

Figure 2. An application using the shared subspace: cross-modality predictions from CP to GE. (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 models are negative controls. (b) The proportion of genes that are well-predicted ($R^2 >$ (

$t_{99th} + 0.2$)) are reported as Percent Predictable for each dataset. (c) The overlap of genes predictable by the MLP model ($R^2 > (t_{99th} + 0.2)$) are shown across the four datasets; 58 are well-predicted using data from any of the four datasets. (d) Network of functional categories of all landmark genes in the L1000 assay, with the 58 highly predictable genes based on all datasets highlighted in red; they fall into the phosphoprotein, acetylation and cytosol categories. (e) Landmark genes that are well-predicted (more than 98% R^2 score) are shown along with their family names for the LUAD dataset. Multiple repetitions on the chart is due to multiple gene family labels for some of the genes in the list. (f) 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: each point on the heatmap shows the predictive power of a group of morphological features (on axis y) for predicting expression level of a landmark gene (on axis x). "Predictive power" here means the R^2 scores generated by limiting the prediction to all the features in the y axis group. For example, STX4 is marked with a star and discussed in the main text. Heatmap is limited to the genes that have R^2>0.7 for at least one of y axis groups. The complete version is provided in the github repository as a xlsx file (https://github.com/carpenterlab/2021_Haghighi_submitted/blob/main/results/SingleGene Pred_cpCategoryMap/cat_scores_maps.xlsx) that can be loaded into Morpheus [16] or Python for further exploration.

Figure 3. Experiment on shared subspace: Cross modality predictions - GE to CP

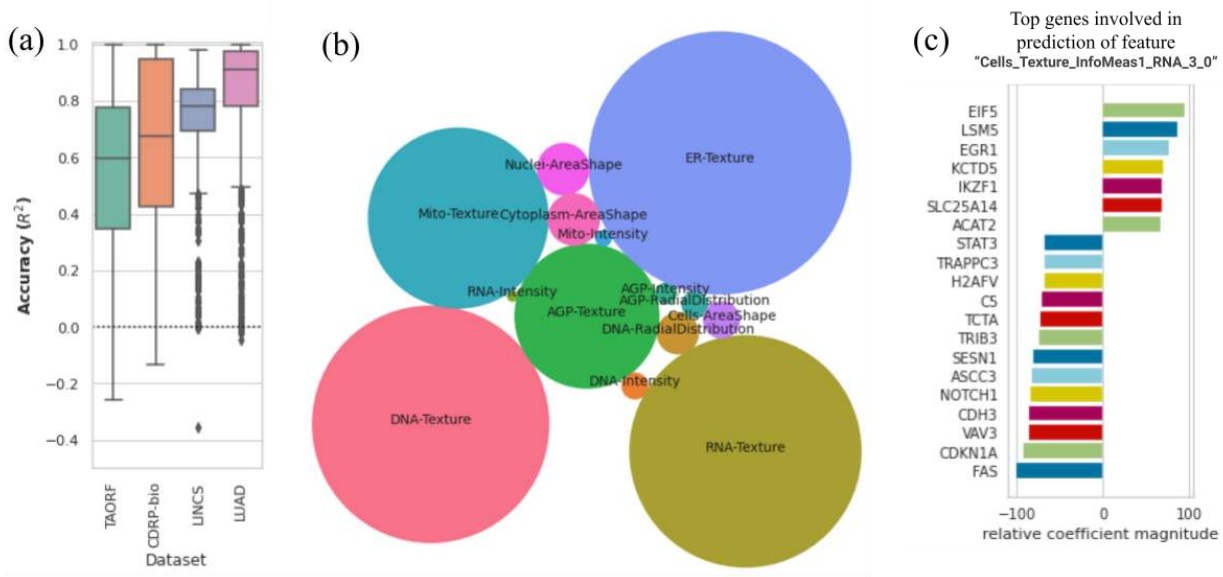


Figure 3. An application using the shared subspace: cross-modality predictions from GE to CP. (a) Distribution of R^2 prediction scores for all morphological features using the MLP model, for each dataset. (b) Categories of highly predictable CP features using GE profiles (median R^2 score across all datasets is more than 0.9). (c) Example output of explorative scripts available to researchers to see what are the most relevant genes to a given morphological feature of interest (and vice versa).

Some datasets showed excellent accuracy in predicting some mRNA levels from morphology data, with MLP yielding superior results to Lasso (Figure 2a and b, complete table in Appendices C and D). 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 of the modalities in the latter two. Likewise, further preprocessing and denoising techniques such as batch effect corrections to improve alignment are another target for future machine learning research.

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 (Figure 2c) and for the 58 well-predicted in all four, we can examine the functional categories they fall into; mainly phosphoprotein, acetylation and cytosol (Figure 2d). For the LUAD dataset (which has the highest cross-modal predictability) we also examined the gene families for highly predictable genes (Figure 2e), finding a diverse array represented, though we note the experiment contained only genes found mutated in lung cancers.

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 Cell Painting features into four categories (*intensity, texture, radial distribution, and shape*) and the five fluorescent channels (*DNA, RNA, ER, AGP, Mito*), then calculated and displayed feature-group-specific prediction scores as a hierarchically-clustered heatmap of median prediction scores (Figure 2f). In this view, genes with strong red columns are readily predicted using any of the morphological categories of features, indicating that the genes are associated with widespread morphological changes; several of these are cell cycle-related, which is known to impact morphology dramatically. Others are more selective, such as STX4, marked with a star, which is most strongly predictable by ER and AGP texture features;

this is consistent with its role in trafficking of intracellular membranes and the plasma membrane, per UniProt[17].

Prediction can be run in the other direction as well, i.e. each morphological feature can also be estimated using the 978 landmark genes as $y_f = f_\theta(X_{GE}) + e_l$. We find a large portion of morphological features to be highly predictable especially for the LUAD and LINCS datasets (Figure 3a). Grouping highly predictable morphological features according to all the datasets reveals that they fall mainly in the texture features category across all the channels (Figure 3b). We also provide a jupyter notebook for exploring the list of top connections between any gene or morphological features of interest (https://github.com/carpenterlab/2021_Haghighi_submitted/blob/main/3-exploreTheLink.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 (Figure 3c).

Application 2: Integrating gene expression and morphology to predict the mechanism of action of compounds

Discerning how a compound works is a major bottleneck in drug discovery [18]. The task is called mechanism of action (MoA) determination, and the goal is to determine the mechanism by which the drug impacts the biological system. One promising method to predict mechanisms of action is to collect a profile from cells 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 to 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.

Several studies have reported success predicting the mechanism of action of compounds using gene expression or cell morphology data individually [19–22] but none of these integrated the two data types to test for improved predictive ability. We therefore provide here the first benchmark for this, using the two chemical perturbation datasets in our set, CDRP-bio and LINCS. Using logistic regression and multilayer perceptron (MLP) classifiers as the baseline models, we applied each for predicting MoA labels using each modality of data independently, using leave-one-compound-out cross-validation on a filtered subset of compounds.

The two profile types, GE and CP, gave relatively comparable performance in predicting MoA across the two datasets and two model types, with the exception that CP was

much stronger on the CDRP-bio dataset using the MLP model (Figure 4a). We also tested two trivial data fusion strategies (one early, one late) to combine both modalities (Figure 4a). Our early fusion baseline is a representation-level concatenation of profiles from the two modalities. Late fusion is at the decision level and averages predicted class probabilities (based on single modality learned classifiers) for making the MoA class decision for test compounds. Trivial early and late fusion of modalities show relatively small improvements upon the performance of the better-performing modality, highlighting the need for applying data fusion methods that better leverage the complementarity of the modalities.

We also test how removing the complementary information of each modality impacts MoA prediction; the yellow bars show the effect of performing Canonical Correlation Analysis (CCA) across both modalities, projecting each modality into the common (ie. shared) subspace, and then late fusion of the classifier decision trained on each subspace-projected modality. As hypothesized, limiting the model to the shared information and removing the modality-specific information reduces the performance of the MoA prediction task.

Exploring MoA-class-specific F1-scores for the late fusion of modalities reveals high variation in class specific prediction results (Figure 4b). As already seen more generally, the simple fusion of modalities does not typically increase the performance of MoA prediction task over the higher performing modality alone for individual MoA categories.

Figure 4. Utilizing Complementary Information: Data Integration for MoA prediction in compound datasets

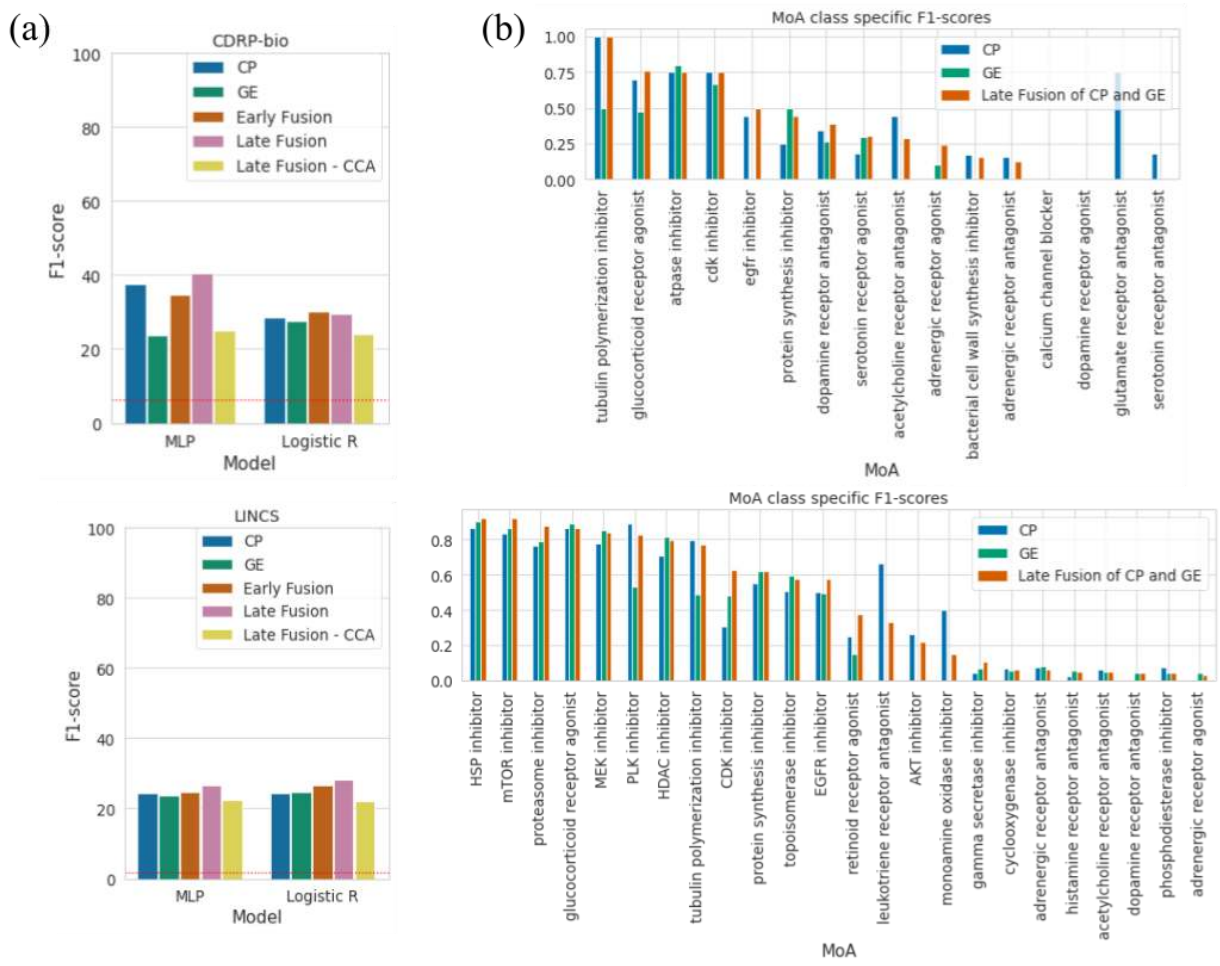


Figure 4. An application using the complementary subspaces: integrating multimodal data for mechanism of action (MoA) prediction. MoA classification of the two compound datasets (CDRP-bio and LINCS) using gene expression, morphology and their integration to predict the mechanism of action of compounds. (a) Classification performance (weighted F1-score) for the MLP and Logistic Regression classifiers using each data modality alone, as well as the two trivial fusion strategies explained in the main text and late fusion of modalities after application of CCA on the feature space of both modalities. Chance-level predictions for each dataset are shown as a horizontal red line on each dataset plot. (b) Class-specific F1-scores are shown based on the MLP model for 16 MoA categories of CDRP-bio (top) and for LINCS (bottom, where the 33 out of 57 MoA categories that resulted in zero F1-scores are excluded).

Discussion and Limitations

We provide the research community a collection of multi-modal profiling datasets with gene expression and morphology readouts, representing two cell types and two perturbation types (genetic and chemical). We define 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 gene expression and morphology profiles contain useful overlapping and distinct information about cell state. 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 (e.g. to correct batch effects) 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. In the case of MoA prediction, the ground truth is based on imperfect human knowledge.

These data, and methods derived from them, can accelerate drug discovery and therefore improve human health and reduce drug development costs. Nevertheless, we note an ethical concern: the cell types are commonly used historical lines derived from two white patients, one male (A549) and one female (U2OS). Therefore, conclusions from this data may only hold true for the demographics or genomics of those persons and not broader groups. They were chosen because the lines are both well-suited for microscopy and they offer the advantage of connecting to extensive prior studies and datasets using them.

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 are in the few hundreds whereas there are roughly 21,000 genes in the genome and numerous variations within each. Likewise, a few thousand compounds are tested here but pharmaceutical companies often have collections of compounds numbering in the millions. The only limitation for expanding these datasets are the financial resources to carry out the experiments. In terms of the assays themselves, the gene expression profiles are captured by the L1000 assay, which is thought to capture 82% of full

transcriptome variation [7], and the Cell Painting assay includes only six stains, which is insufficient to capture the localization and morphological variation of all cellular components.

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[23]. 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, although such samples are more difficult to procure. 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 and point estimates in the other space. Although still very rare, small, and labor-intensive to create, data sets with both gene expression and morphology at single-cell resolution are beginning to become available via in situ RNA-seq methods and could accelerate the field of multi-modal biological data analysis.

Code and Data Availability

Preprocessed profiles that are augmented with gene and compound annotation are available on a public AWS S3 bucket.
Documentation on the folder structure, dataset details, instructions for accessing the data, and the source code to reproduce and build upon these results are available at http://broad.io/rosetta. We license the source code as BSD 3-Clause, and license the data, results, and figures as CC0 1.0.

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 grant 2018-183451 the Chan Zuckerberg Initiative DAF, an advised fund of the Silicon Valley Community Foundation and the National Institutes of Health NIGMS (R35 GM122547 to AEC).

Author information

M.H., S.S., and A.E.C. all contributed to drafting the manuscript. J.C. initiated the project and performed early explorations of a subset of the data. M.H. analyzed and explored the data with inputs from the other coauthors.

Competing interests

The authors declare no competing interests.

References

1. Becker, T. *et al.* Predicting compound activity from phenotypic profiles and chemical structures. *Cold Spring Harbor Laboratory* 2020.12.15.422887 (2020) doi:10.1101/2020.12.15.422887.
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**, (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.e17 (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. U. S. A.* **111**, 10911–10916 (2014).

11. Berger, A. H. *et al.* High-throughput Phenotyping of Lung Cancer Somatic Mutations. *Cancer Cell* **32**, 884 (2017).

12. Gundersen, G., Dumitrascu, B. & Ash, J. T. End-to-end training of deep probabilistic cca on paired biomedical observations. *Uncertain. Artif. Intell.* (2019).

13. Dai Yang, K. *et al.* Multi-domain translation between single-cell imaging and sequencing data using autoencoders. *Nat. Commun.* **12**, 1–10 (2021).

14. He, B. *et al.* Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat Biomed Eng* **4**, 827–834 (2020).

15. Nassiri, I. & McCall, M. N. Systematic exploration of cell morphological phenotypes associated with a transcriptomic query. *Nucleic Acids Res.* **46**, e116 (2018).

16. Tandon, G., Chan, P. & Mitra, D. MORPHEUS. in *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security - VizSEC/DMSEC '04* (ACM Press, 2004). doi:10.1145/1029208.1029212.

17. Syntaxin-4. https://www.uniprot.org/uniprot/Q12846.

18. Schenone, M., Dančík, V., Wagner, B. K. & Clemons, P. A. Target identification and mechanism of action in chemical biology and drug discovery. *Nat. Chem. Biol.* **9**,

232–240 (2013).

19. 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).

20. 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).

21. 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).

22. Lapins, M. & Spjuth, O. Evaluation of gene expression and phenotypic profiling data as quantitative descriptors for predicting drug targets and mechanisms of action. *bioRxiv* 580654 (2019) doi:10.1101/580654.

23. 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).

24. Bray, M.-A. *et al.* A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay. *Gigascience* **6**, 1–5 (2017).

25. Cell Painting Image Collection. https://registry.opendata.aws/cell-painting-image-collection/.

26. Natoli, T. *et al. broadinstitute/lincs-cell-painting: Initial Processed Data Release*. (2020). doi:10.5281/zenodo.3928744.

27. [No title]. https://clue.io/.

Online Methods

Dataset preprocessing

We gathered four available data sets that had both Cell Painting morphological (CP) and L1000 gene expression (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-datasets/Rosetta-GE-CP`
Each csv file contains a replicate level of profiles and is augmented with the metadata available for that dataset.

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://github.com/cytomining/cytominer).

We made the following three types of profiles available:

- Aggregated profiles which are the average of single-cell profiles in each sample.
- Normalized profiles which are the z-scored aggregated profiles, where the scores are computing using the distribution of negative controls as the reference.
- Normalized variable-selected which are normalized profiles with features selection applied.

For L1000, we use the previously processed 978 "landmark" genes as our input features The complete processing details are provided in [7].

Data processing for Analysis

We have used treatment-level profiles for both the gene expression (GE, using L1000) and morphology (CP, using Cell Painting) modalities for the analysis presented, 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, and replicate-level profiles are the average of single-cell measurements (in the case of CP; for GE the finest granularity available is the bulk population replicate-level profile).
We standardized replicate-level profiles per plate to have zero mean and unit variance before averaging them to form treatment-level profiles.

Measuring quality of profiles

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) differences in the datasets. Consistency of profiles of a single treatment across different batches of experiment is considered a measure of data quality. We check this consistency as follows.
After standardization of the profiles per plate, we calculate the Pearson correlation coefficient between each pair of profiles for the same perturbation. The distribution of these coefficients for each dataset and modality are illustrated in Appendix B shown as red curves. 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 non-zero dotted vertical line to the right shows the 90th percentile of the null distribution. We consider the perturbations that have an average replicate correlation more than the 90th percentile of the null distribution as high quality samples.

Filtering samples

To remove noisy samples 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 samples or higher quality samples according to both modalities. For the analysis for data integration, we used samples that are higher quality (i.e. > 90th percentile of the null distribution, as defined above) in at least one of the modalities. Definition of higher quality samples is given in the previous section. A comprehensive description of the number of samples in each modality, number of overlapping perturbations across both modalities, size of intersection and union sets of higher quality samples across both modalities are given in Appendix A and highlights are summarized in Table 1.

One of the chemical datasets (CDRP-BBBC047-Bray) has a subset of compounds that are known to be bioactive. We refer to this subset as CDRP-bio-BBBC036-Bray and report the details independently for this dataset in Table 1 and Appendix A and B. We only use CDRP-bio and not the full CDRP set for the analysis in this paper. We did so because we believe that the quality of CDRP is insufficient for either of these analyses presented given that very few samples remain after filtering for replicate reproducibility across both modalities (see Appendix B).

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: $y_l = f_\theta(X_{cp}) + e_l$; in which 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 representing all morphological changes across all the perturbations.
GE to CP: $y_f = f_\theta(X_{ge}) + e_l$; in which y_f is a vector of morphological feature f across all the perturbations in a dataset and X_{ge} is the whole L1000 data matrix representing all gene expression changes across all perturbations.

For each prediction direction (CP to GE, GE to CP) and each baseline linear (Lasso) and nonlinear (MLP) model for this regression problem, we use the coefficient of determination (R^2) and k-fold cross-validation over the perturbation samples to form a distribution of kR^2 values for each landmark gene (for CP to GE) or each morphological feature (for GE to CP). We also shuffle the vector y_l for each gene l across all the samples and apply 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 (regularisation parameter for Lasso model and number and size of hidden layers, activation function and regularization parameter for the MLP model) are selected using grid-search and cross-validation inside each of outer k folds.

In the Appendix D, the median prediction scores of each model for each landmark gene for each dataset and according to each model is presented. Distribution of MLP model prediction scores for the 50 landmark genes with the highest median scores in each dataset is shown at Appendix C.

Percent Predictable

Percent predictable is defined as the percentage of landmark genes which 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 ($t_{99th} + 0.2$) as the threshold for calling a gene "predictable". We reported the *Percent Predictable* values for each dataset in the table in Figure 2b.

MoA Prediction

For the analysis for MoA prediction, we used the samples that had high quality (i.e. > 90th percentile of the null distribution; see above) in either modality.

The LINCS dataset has MoA annotations for 1401 overlapping compounds across two

modalities. Every compound is tested at six 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 present in the dataset. We start with the filtered union set, and filter it again to keep MoA classes which have at least 4 samples in their class. It results in a set of 1655 samples across 521 compounds in 57 MoA categories. One multi-label MoA category was removed to keep the problem multi-class single label.

Two logistic regression and multilayer perceptron (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 baseline integration of the two. We performed leave-one-compound-out cross-validation (all doses of a compound are left out) to report F1-score as the classification performance. 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 two; to address this imbalance in the data, we used weighted logistic regression by taking into account the frequency of each class in the training set. For the MLP model, we oversampled samples in class to the number of majority samples in the training set. The leave-one-compound-out cross validation experiment results in a vector of predictions for the 1655 samples. We then calculate weighted average F1 score of MoA predictions (where we weight class-specific scores by the number of true samples in each MoA class) for each model and each data modality.

For baseline fusion, we used trivial data fusion strategies (one early, one late) to combine both modalities. Early fusion baseline is a representation-level concatenation of modalities. Late fusion is at the decision level and averages predicted class probabilities (based on single modality learned classifiers) for making the MoA class decision for test compounds.

We applied the same procedure to the CDRP-bio dataset. This dataset has MoA annotations for 1,327 out of 1,916 overlapping compounds across two modalities. After passing samples from three filters of: union higher quality across modalities, available MoA labels, being in an MoA class which have at least four compounds in the set, we will get 123 compounds in 16 MoA categories.

Supplementary information

Appendix A. Curated Datasets

Datasets List

The details of each dataset, including the type of perturbation, number of perturbations, cell line, and number of replicates, are below.

CDRP-BBBC047-Bray-CP [24]- GE [7]:

Cell line used for chemical perturbation of cells in this dataset was U2OS.
There are 30,430 and 21,782 unique compounds for CP and GE datasets, respectively.

For CP dataset, the median number of replicates for each compound in the set is 4 and there are 26,572 replicates for control wells (samples).

For GE dataset, the median number of replicates for each compound in the set is 3 and there are 3,478 replicates for control wells (samples).

20,131 compounds are present in both datasets. 6\% percent of these compounds have MoA annotations. Only 3/20,131 compounds have replicate correlation more than 90th percentile of random distribution in both modalities.

CDRP-bio-BBBC036-Bray-CP[24]-GE [7]:
This is a subset of the previous dataset, containing the bioactive subset of compounds.

There are 2,242 and 1,917 unique compounds for CP and GE datasets, respectively.

For CP dataset, The median number of replicates for each compound in the set is 8 and there are 3,528 replicates for control wells (samples).

For GE dataset, The median number of replicates for each compound in the set is 2 and there are 3,478 replicates for control wells (samples).

1,916 compounds are present in both datasets. 69\% percent of these compounds have MoA annotations. 131/1,916 compounds have replicate correlation more than 90th percentile of random distribution in both modalities.

LUAD-BBBC041-Caicedo-CP [25]- GE [11]:

Cell line used for genetic perturbation of cells in this dataset was A549.

There are 593 and 529 unique alleles for CP and GE datasets, respectively.

For CP , GE datasets, the median number of replicates for each allele in the set is 8.

525 alleles are present in both datasets. 197/525 of these alleles have replicate correlation more than 90th percentile of random distribution in both modalities.

TA-ORF-BBBC037-Rohban-CP [3] - GE:

Cell line used for genetic perturbation of cells in this dataset was U2OS.

There are 323 and 327 unique alleles for CP and GE datasets, respectively.

For CP dataset, the median number of replicates for each allele in the set is 5 and there are 268 replicates for control wells (samples).

For GE dataset, the median number of replicates for each allele in the set is 2 and there are 56 replicates for control wells (samples).

150 alleles are present in both datasets. 36/150 of these alleles have replicate correlation more than 90th percentile of random distribution in both modalities.

LINCS-Pilot1-CP [26] - GE [27]:

Cell line used for chemical perturbation of cells in this dataset was A549.

There are 1,570 unique compounds across 7 doses for CP dataset. There are 1,402 unique compounds across 7 doses for GE dataset.

There are 9,394 and 8,369 unique compounds-dose for CP and GE datasets, respectively.

For CP dataset, the median number of replicates for each compound in the set is 5 and there are 3,264 replicates for control wells (samples).

For GE dataset, the median number of replicates for each compound in the set is 3 and there are 1,485 replicates for control wells (samples).

6984 compound-dose pairs are present in both datasets.
100% of these compounds have MoA annotations.

Among 6984 unique compounds-dose overlapping compounds, 1140 compounds have replicate correlation more than 90th percentile of random distribution in both modalities.

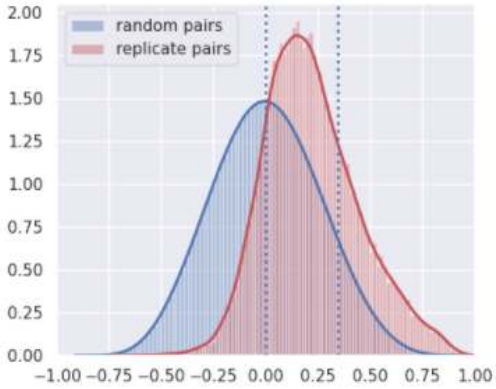
Table 1: GE and CP dataset descriptions. N_p: number of perturbations,N_r: number of replicates,N_d: number of doses

Datasets					
Name	Cell Line	Perturbation Type	$N_p/N_d/N_r$ CP	$N_p/N_d/N_r$ GE	N_p Intersection
CDRP-BBBC047-Bray	U2OS	chemical	30,430/1/4	21,782/1/3	20,131
CDRP-bio-BBBC036-Bray	U2OS	chemical	2,242/1/8	1,917/1/2	1916
LUAD-BBBC041-Caicedo	A549	genetic	593/-/8	529/8	525
TA-ORF-BBBC037-Rohban	U2OS	genetic	323/-/5	327/2	150
LINCS-Pilot1	A549	chemical	1570/ 7/ 5	1402/3	$N_{p/d}$: 6984

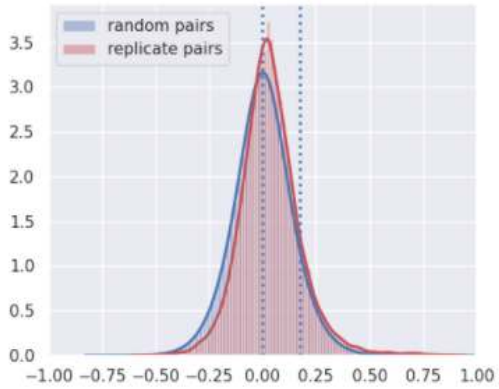
Appendix B. Data Quality: Replicate reproducibility

CDRP-BBBC047-Bray:

CP

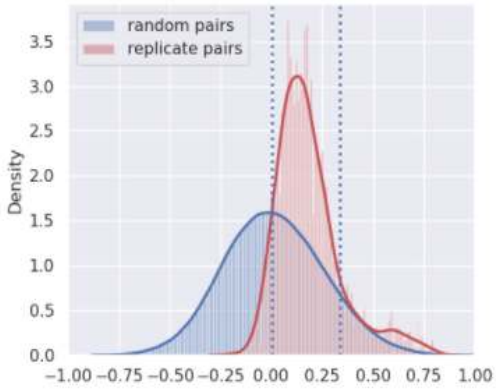


GE

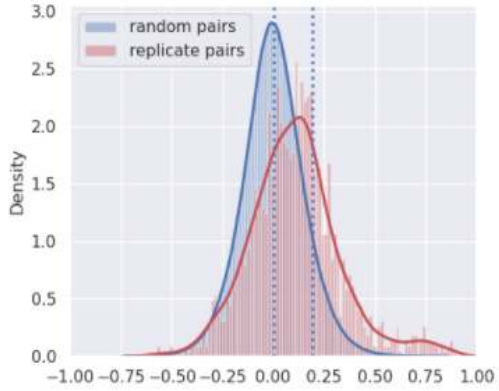


CDRPBIO-BBBC036-Bray:

CP

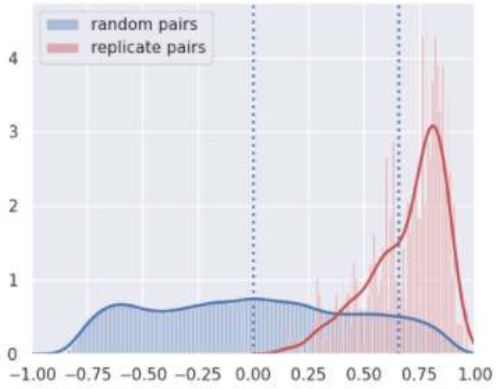


GE

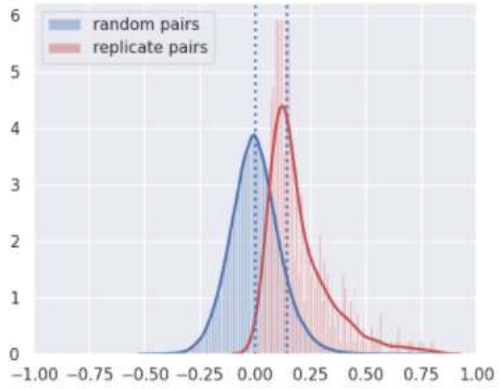


LUAD-BBBC041-Caicedo-CP-GE :

CP

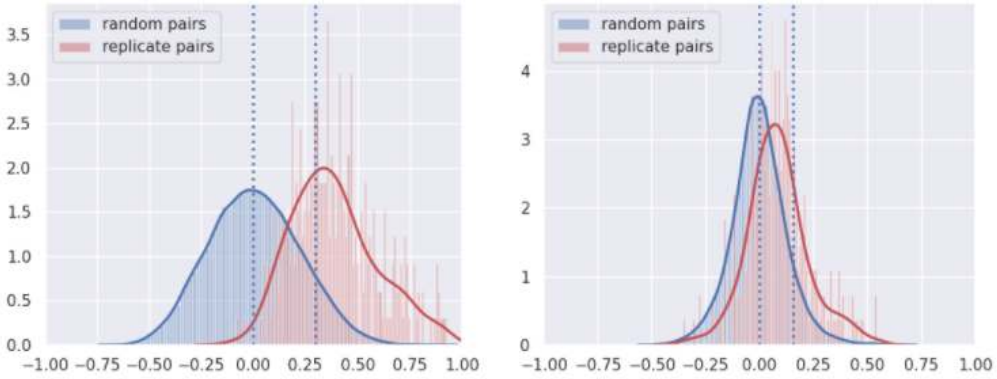


GE



TA-ORF-BBBC037-Rohban-CP-GE :

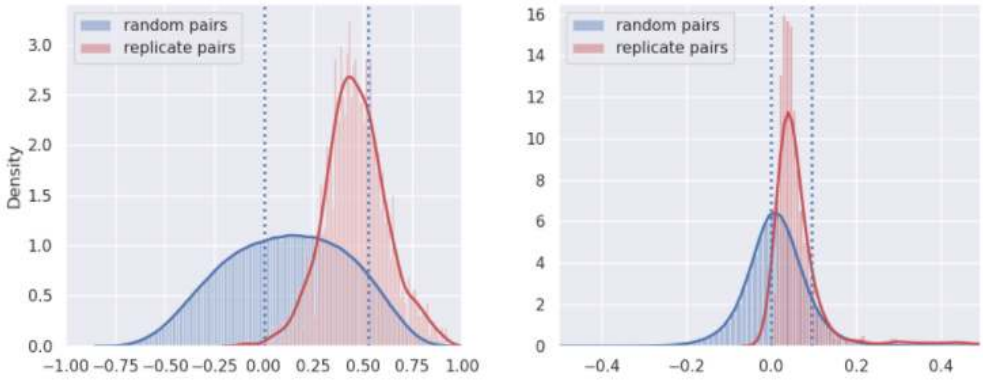
CP GE



LINCS-Pilot1-CP-GE :

- CP - All doses together (Metadata_pert_id)
- GE - All doses together (pert_id)

CP GE



- CP - each sample/dose (Metadata_pert_id_dose)
- GE - each sample/dose (pert_id_dose)

CP GE

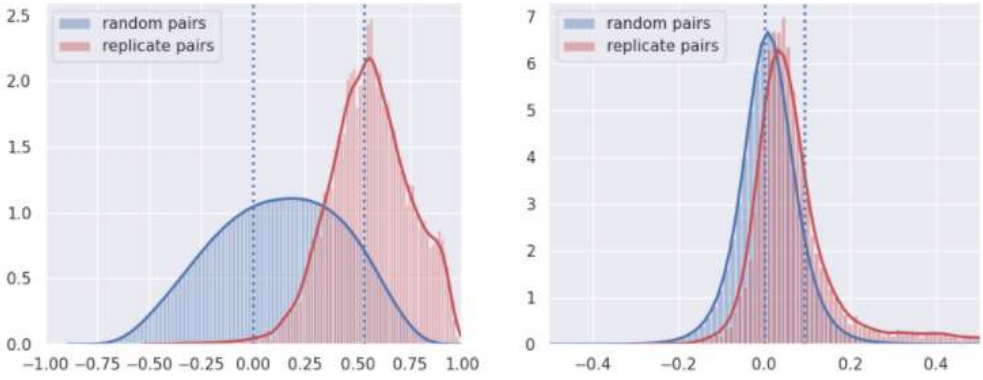


Figure 5. To inspect the quality of each dataset, we calculate the consistency of profiles across different replicates of the same perturbation as follows. We standardized the profiles per plate to have zero mean and unit variance. Next, we calculated the Pearson correlation coefficient between each pair of profiles for the same perturbation (red curve) and for different perturbations (blue curve). Dotted vertical lines are shown at zero and 90th percentile of the random pairs (blue) distribution.

Appendix C. Top 50 highly predictable L1000 genes by Cell Painting morphological features

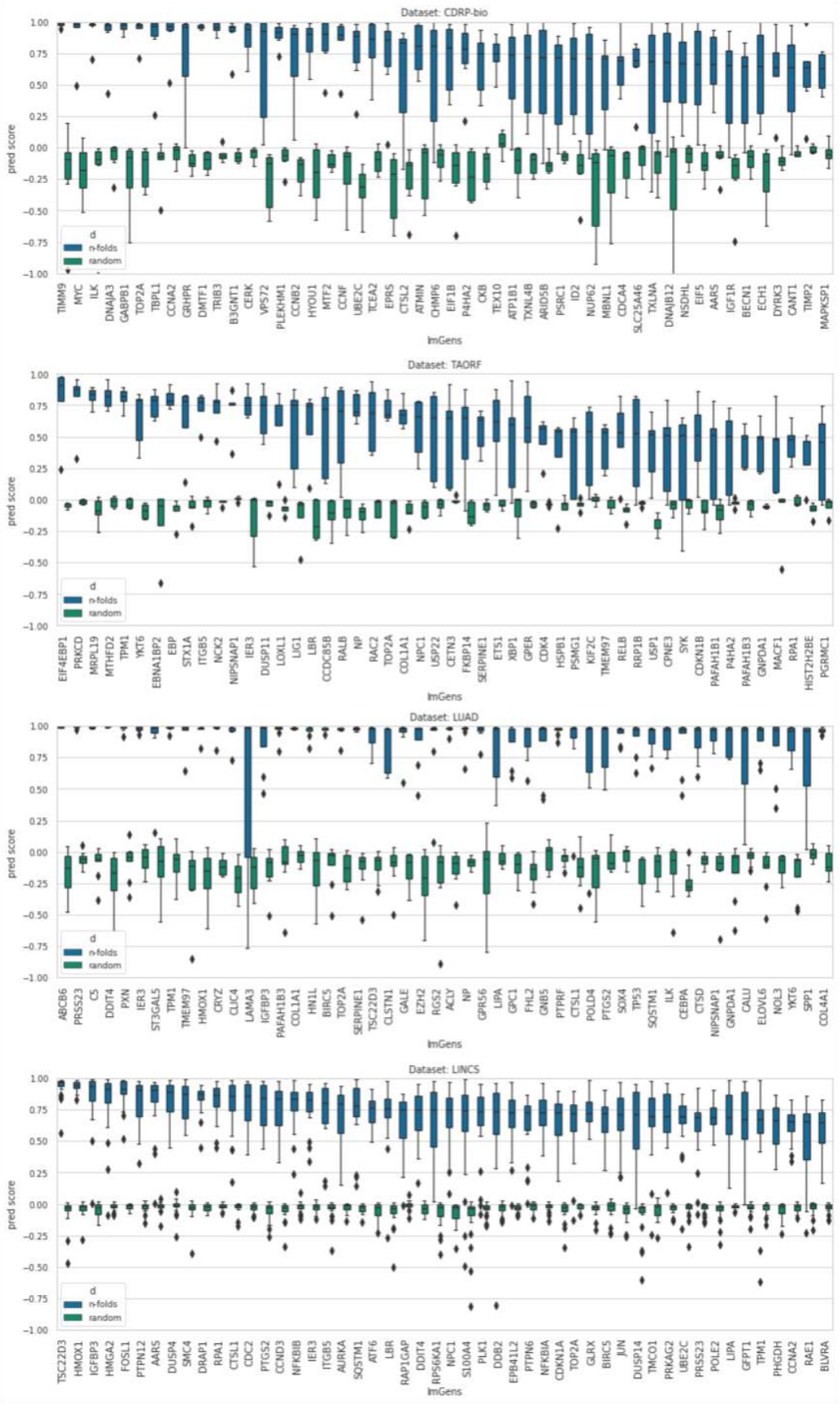
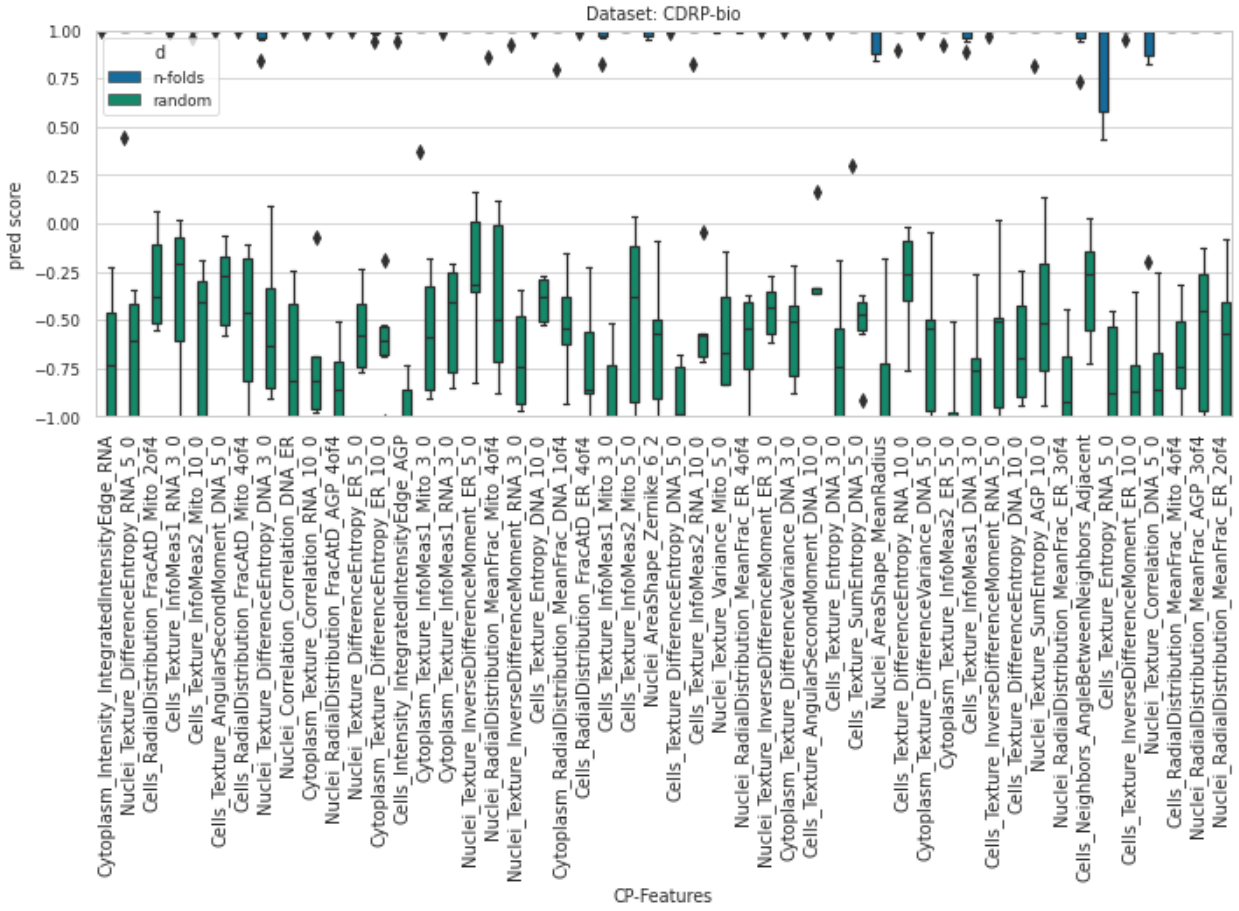


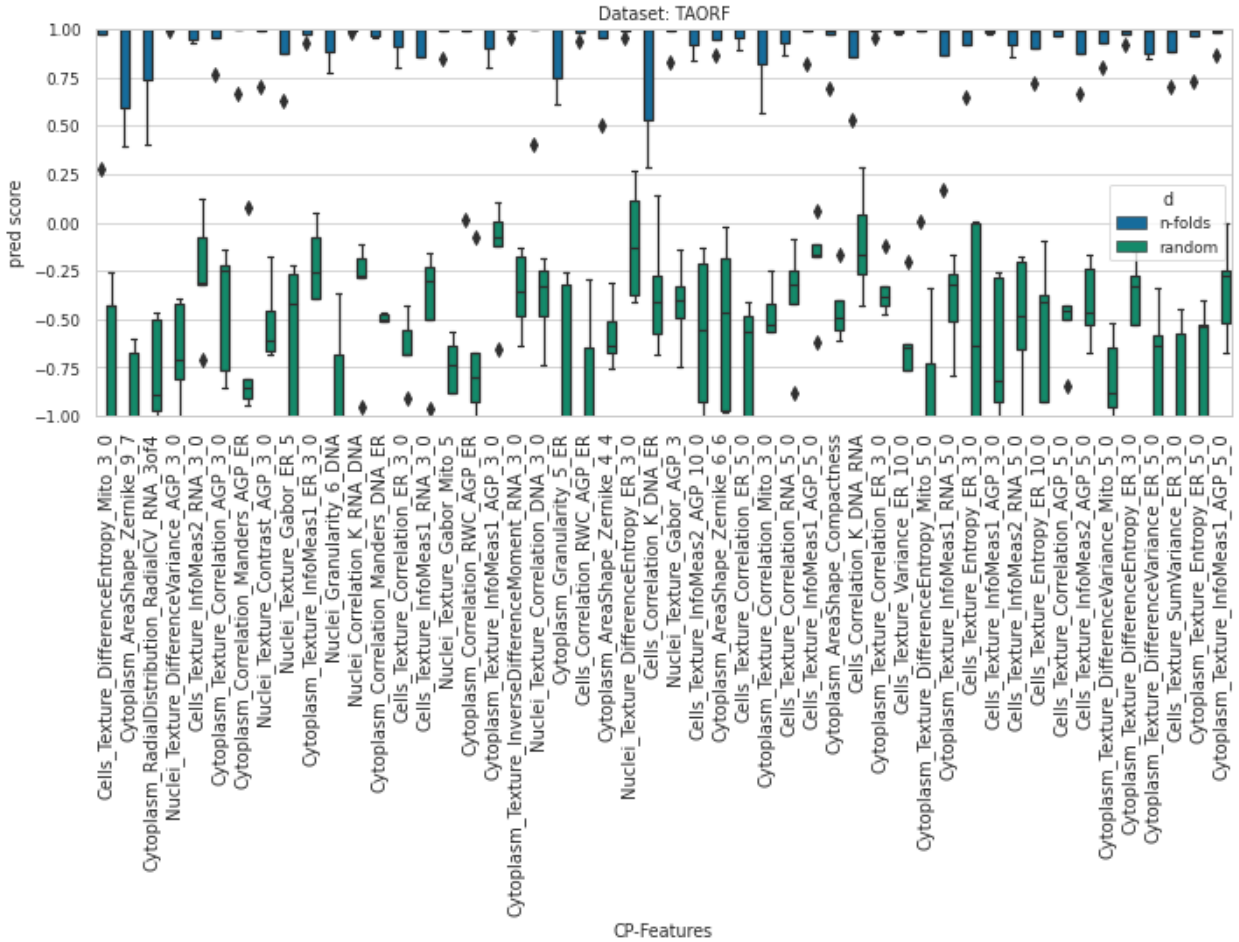
Figure 6. Prediction of L1000 mRNA levels by Cell Painting features: for each dataset, the distribution of MLP baseline prediction scores for the ordered top 50 landmark genes with the highest R2 median prediction scores are provided. Each distribution consists of k, R2 values corresponding to application of k-fold cross validation for each landmark gene in each dataset, which is shown as blue boxes. We also shuffle the landmark gene vector across all the samples and apply the same cross-validation procedure to form a null distribution for each gene which is shown as green boxes in each plot.

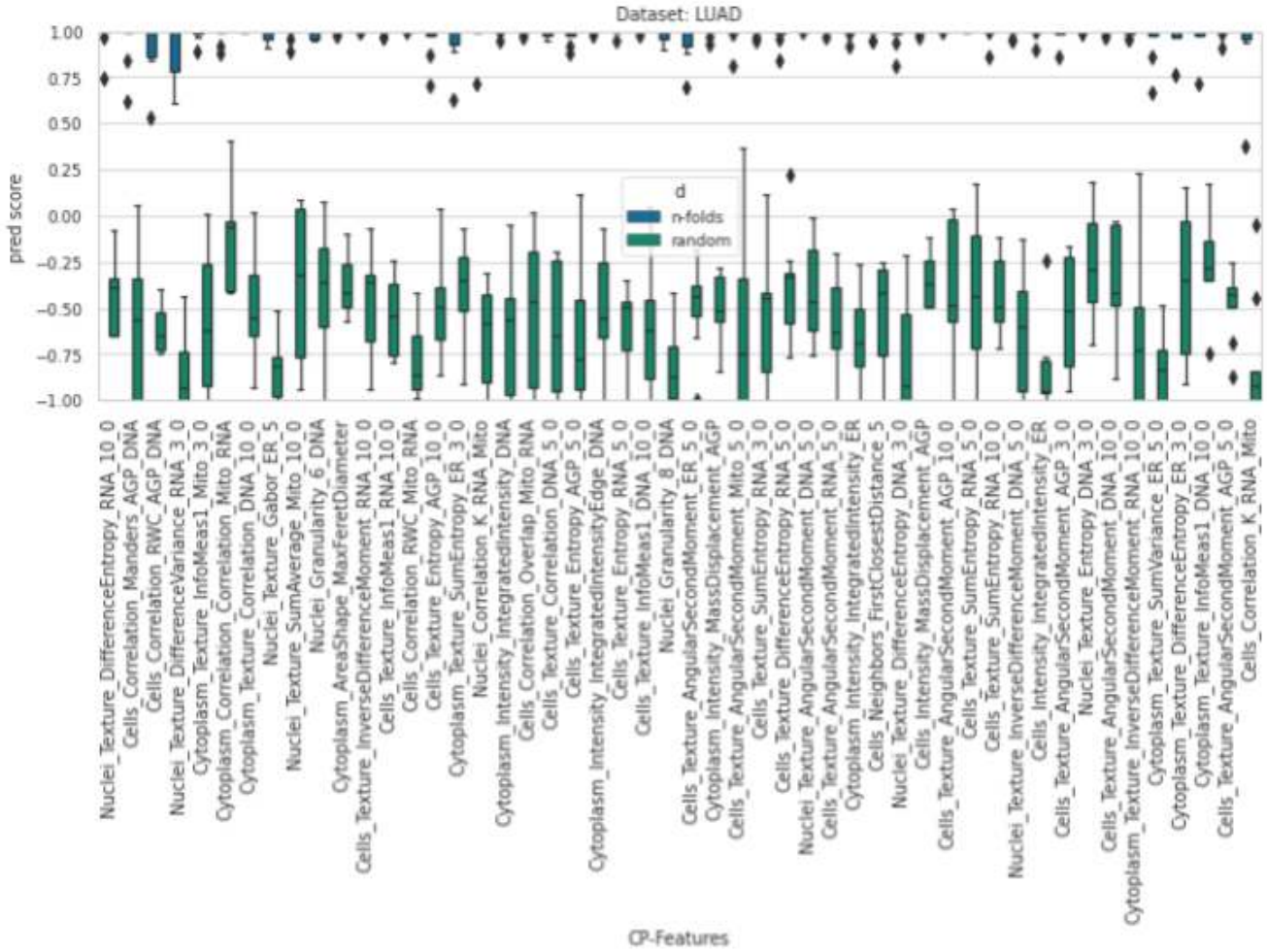
Appendix D. Median Prediction scores for each landmark gene across each datasets and models

Appendix_D.csv

Appendix E. Top 50 highly predictable Cell Painting morphological features by L1000 genes







CP-Features

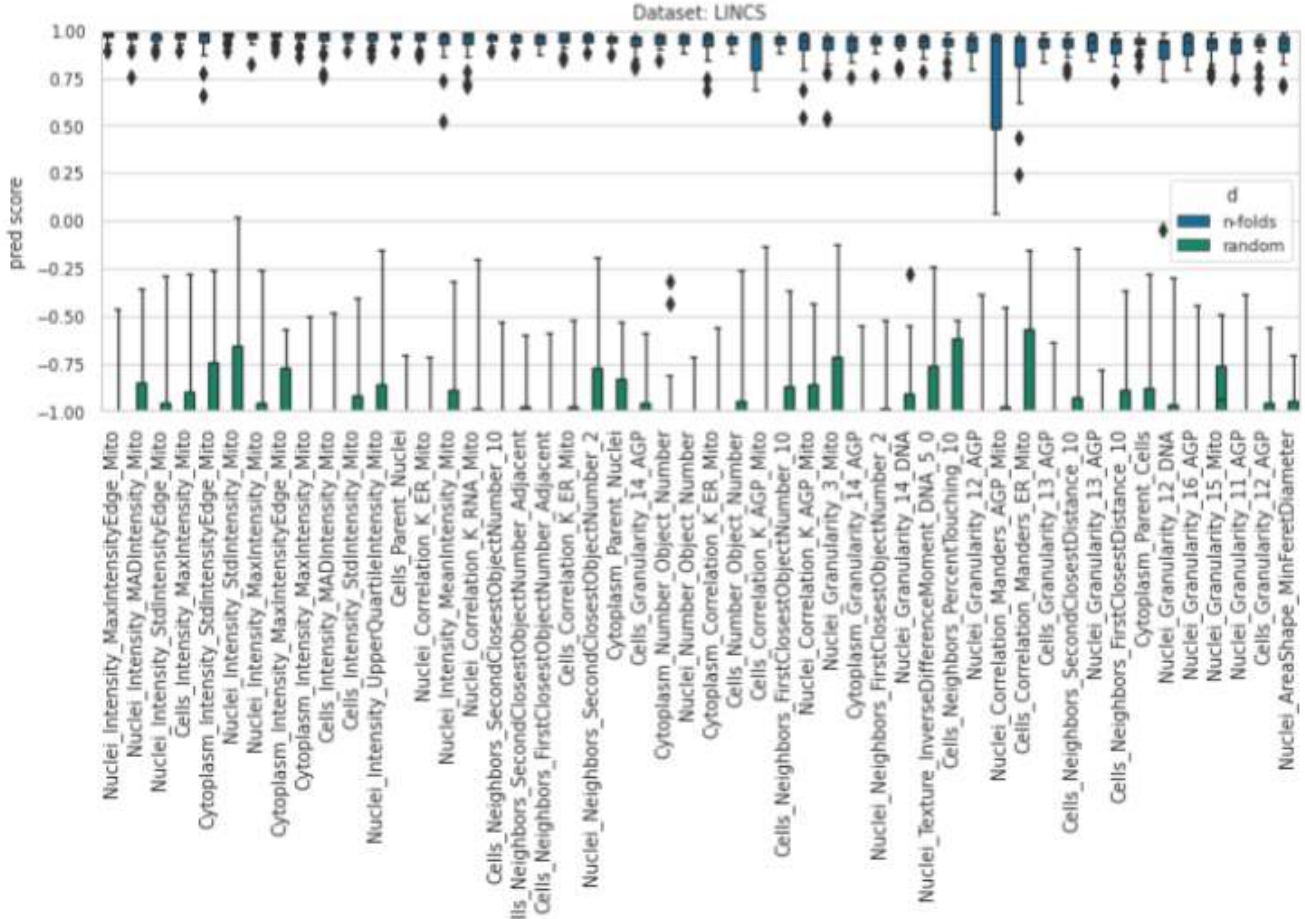


Figure 7. Prediction of each cell painting feature by L1000 mRNA levels: for each dataset, the distribution of MLP baseline prediction scores for the ordered top 50 cell painting features with the highest R2 median prediction scores are provided. Each distribution consists of k, R2 values corresponding to application of k-fold cross validation for each single CP feature in each dataset, which is shown as blue box plots. We also shuffle the CP feature vector across all the samples and apply the same cross-validation procedure to form a null distribution for each gene which is shown as green boxes in each plot.