Response to reviewers' comments from Round 1

- Societal implications and limitations of the dataset - We have included a section in the paper to explicitly address the societal implications and limitations of this dataset.
- Details about the dataset and data analysis - We have made several changes to the manuscript to address the reviewers' concerns about lack of details about the dataset and reproduction of the benchmarks
    - We have updated the README of our data repository https://github.com/jump-cellpainting/neurips-cpjump1/ with additional information about the content of our data repository, instructions for generating the profiles and for reproducing the benchmark analyses.
    - We have also included additional text to elaborate on how the compound and gene selection was made.
    - We have expanded on why the dataset was generated under several experimental conditions.
    - We have included an example image from our dataset as Figure 3
    - We have included a new section, "Data Splits", to provide guidelines for creating training, test and validation splits that will be useful for researchers in the ML community
- Contextualize the importance of this dataset relative to existing data – we have explained why this dataset is so exciting and useful: primarily, it is the first experiment to simultaneously assess three completely different ways to perturb cell pathways: chemical compounds (drugs), gene overexpression (ORFs), and gene knockdown (CRISPR). It is the first dataset systematically built to study the relationship between matched pairs of drugs and gene perturbations; prior datasets may have annotations for a subset of genes and drugs, but these rare samples are spread across many experimental batches and plates; here we have concentrated all relevant pairs into a single well-controlled experiment.
- Clarify data access – we had set up a "requester pays" access at the time of submission but had failed to clarify that the images is now available on a AWS Open Data bucket (s3://cellpainting-gallery) and is free to download Instructions for downloading the images is available in the README of the data repository.
- Benchmarks - The reviewers had expressed concern that we had not provided a baseline for the third application of our dataset "Benchmarking style transfer methods". As we weren't able to include a baseline, we have removed the section from the paper.
- Use of alternate methods for benchmarking -
    - For our first application, "Benchmarking perturbation-detection methods", the reviewers had suggested using anomaly detection methods. In a typical experiment, we have ~30-80% of the perturbations have a detectable phenotype, so this doesn't fit into the category of anomaly detection problems.
    - For our second application, "Benchmarking gene-compound matching methods", the reviewers had suggested using methods from re-identification and metric learning domains. We believe the real challenge here is learning more effective

representations (or metrics) that can better capture the biological state (or similarity between states). We didn't find any additional advantage in framing this as a re-identification problem – where the task is to match the same entity across different views. Additionally, we note that in such images, we are not looking for the same cells but in a different view (e.g. across two different compounds). These are cells sampled from the same population, but treated with different perturbations, and the goal is to find perturbations that induce similar phenotypes in the cells.

- Use of a single microscope - The reviewers had expressed their concerns regarding the use of a single instrument for creating this dataset. We no longer have the sample plates, so it is not possible to image using another microscope without reproducing the entire experiment. Also, the samples fade after imaging with one microscope such that a second image acquisition may be feasible but will show degraded quality compared to the first and artifacts introduced by the laser of the first microscope may carry over to imaging the second. In general, systematic artifacts are not problematic to most applications in profiling. By contrast, systematic artifacts are problematic when attempting to match across experimental datasets acquired on different microscopes, hence the entire set presented here was collected on a single microscope. Batch correction across experimental setups (including microscopes) is an important problem to be solved, but not one this benchmark dataset was designed for.
- Alternative metrics – Our choice of metrics was based on decisions made downstream; typically a fraction of the perturbations or connections among perturbations are followed up on, and these metrics give a direct readout of what this fraction is. Other information retrieval – such as Precision@K – are not suitable to report for this purpose (at the individual perturbation level) given the relatively small number of samples per class in this dataset (~n=4 for replicates of a perturbations, ~n=2 for matching genes and compounds).

Compound selection criteria

We filtered the Repurposing Hub compounds using several criteria, of which three are most important:

1. The compounds should target genes that belong to diverse gene families (Supplementary Table 1). This is because the ideal methods would work well for many different biological pathways, not just a few that are well-characterized and/or easy to predict.
2. Each gene should be targeted by at least two compounds, so that gene-compound matching and compound-compound matching can both be performed using the dataset.
3. We additionally considered applying the constraint that each compound should target only a single gene. However, this criterion is difficult to achieve due to polypharmacology (Supplementary Table 2), which is the property for compounds to bind and impact many different gene products in the cell; this is especially common for protein kinase inhibitors in the dataset. Instead, we only filtered out the so-called “historical compounds” listed in the Chemical Probes Portal (Arrowsmith et al. 2015), which are compounds that are known to be quite non-selective (or not sufficiently potent) compared with other available chemical probes.

Our list of compounds and genes also includes both negative and positive controls. The negative controls for each perturbation modality are:

- Compounds: DMSO (Dimethyl sulfoxide), which is the solvent for all the compounds studied. In other words, all samples will have DMSO added at the same concentration but the negative controls have no *additional* compound added.
- ORFs: 15 ORFs with the weakest signature in previous image-based profiling experiments (Rohban et al. 2017).
- CRISPRs: 30 CRISPR guides that target an intergenic site (cutting controls, n = 3) or don't have a target sequence that exists in human cells (non-cutting controls, n = 27).

There are three types of compound positive controls in our list. First, we included chemical probes that are very well-studied and (unlike most compounds) are known to very selectively modulate the genes that they target (Arrowsmith et al. 2015). Second, we included compounds that strongly correlate with the correct genetic perturbation in previous image-based profiling experiments with ORFs (Rohban et al. 2017) and compounds (Bray et al. 2017). Finally, we included a set of very diverse pairs of compounds with strong intra-pair and weak inter-pair correlations, based on prior experiments (poscon_diverse).

Additionally, compounds were filtered based on availability through at least one of four compound vendors (Sigma, SelleckChem, Tocris, and MedChemEx) and genetic reagents through Broad’s Genetic Perturbation Platform portal. Lastly, we also excluded compounds on the U.S. Drug Enforcement Agency (DEA) list of controlled substances or the Organisation for the Prohibition of Chemical Weapons (OPCW) list of chemical weapons precursors. All of these

steps will enable a commercial vendor to offer the compound set so others can test the same perturbations in other contexts for comparison.

Number of gene targets (N)	Number of gene families with N gene targets in the final list
1	92
2	16
3	2

Supplementary Table 1: **Number of gene families with a given number of gene targets.** Closely related genes are called *families.* To maximize the diversity of genes, the genes were chosen such that most gene families (n=92) have only a single gene in the final list.

Number of gene targets (N)	Number of compounds in the final list targeting N gene targets
1	218
2	49
3	23
4	7
5	4
6	3
7	1
8	1

Table 2: **Number of compounds with a given number of gene targets.** The compounds were chosen such that most compounds (n=218) in the final list are annotated as having only a single target.

Target loci selection

We picked the target loci for the CRISPR experiments by selecting the top-two-ranking sgRNA sequences that maximize their on-target activity, calculated using the Azimuth 2.0 model (Doench et al. 2016) and minimize the off-target activity, calculated using the Cutting Frequency

Determination score (additional details can be found at https://portals.broadinstitute.org/gpp/public/software/sgrna-scoring-help).

Plate layout design

The first plate (Supplementary Figure 1a) is entirely **compounds**, with two compounds per gene target. Each compound is in singlicate on the plate except for a dozen or so compounds in duplicate and the negative control DMSO described above, in n=64 replicates. The second plate (Supplementary Figure 1b) is entirely **CRISPR** reagents, with two guides per gene, each arrayed in its own well and kept separate, with no within-plate replicates; it does have two replicates of the 30 CRISPR negative controls described above. The third plate (Supplementary Figure 1c) is entirely **ORFs**; because there was only one perturbation reagent per gene, there are two replicates of each per plate, plus n=4 replicates of the 15 ORF negative controls.

We also considered the impact of edge effects, or plate-layout effects, in our design. Edge effects are the technical artifact whereby different samples will yield different behavior depending on where they are located on a plate; generally this is most observed in the outer two rows and columns of the plate, and the problem persists despite efforts to mitigate it experimentally (Lundholt, Scudder, and Pagliaro 2003). While designing the plate layout, we divided the plate into outer and inner wells where the outer wells are the two rows and columns closest to the edge of the plate and the inner wells are the rest of the wells on the plate. Then we applied the following constraints in order to minimize the impact of edge effects:

- Both of the compounds that target the same gene will either be in the inner wells or in the outer wells. They will not be split such that one of the compounds is in the inner well while the other is in the outer well.
- The gene target of outer well compounds will be in the outer wells of the genetic perturbation plate.
- All the positive control compounds are in the inner wells.

If preferable, with this design, an analysis can be constrained to the inner wells only, to ensure that edge effects have minimal influence on the results. For the four U2OS 48-hour replicate plates, we find that our performance metrics, *Percent Replicating* and *Percent Matching* (described later), increase by ~7% and ~16%, respectively, for the outer wells when compared to the inner wells, indicating that indeed edge effects may inflate correlations among samples.

Full list of experimental conditions

1. Four replicate plates of compounds and CRISPRs and two replicate plates of ORFs (which, as mentioned, contain two replicates within each plate) at two time points and two cell lines each. The short and long time points were different for each perturbation type: compounds (24-hour, 48-hour), ORFs (48-hour, 96-hour) and CRISPRs (96-hour, 144-hour). The two cell lines were U2OS and A549.

2. One plate of the A549 96-hour ORF plate where the cells have been additionally treated with Blasticidin (a drug that kills cells that have not been properly infected with the genetic reagent).
3. Two replicate plates of the A549 144-hour CRISPR plate where the cells have been additionally treated with Puromycin (a drug that kills cells that have not been properly infected with the genetic reagent).
4. Two replicate plates of the A549 48-hour compound plate with 20% higher cell seeding density than the baseline.
5. Two replicate plates of the A549 48-hour compound plate with 20% lower cell seeding density than the baseline.
6. Four replicate plates of the A549 24-hour compound plate were imaged six additional times to test photobleaching from repeated imaging.
7. Two replicates of the ORF plates in U2OS and A549 at 96-hour and 144-hour were imaged four additional times, once on each of days 1, 4, 14, 28 after the first imaging, to test the stability of samples over time.
8. Four replicate plates monoclonal Cas9 cell lines treated with compounds.

References

Arrowsmith, Cheryl H., James E. Audia, Christopher Austin, Jonathan Baell, Jonathan Bennett, Julian Blagg, Chas Bountra, et al. 2015. “The Promise and Peril of Chemical Probes.” *Nature Chemical Biology* 11 (8): 536–41.

Bray, Mark-Anthony, Sigrun M. Gustafsdottir, Mohammad H. Rohban, Shantanu Singh, Vebjorn Ljosa, Katherine L. Sokolnicki, Joshua A. Bittker, et al. 2017. “A Dataset of Images and Morphological Profiles of 30 000 Small-Molecule Treatments Using the Cell Painting Assay.” *GigaScience* 6 (12): 1–5.

Doench, John G., Nicolo Fusi, Meagan Sullender, Mudra Hegde, Emma W. Vaimberg, Katherine F. Donovan, Ian Smith, et al. 2016. “Optimized sgRNA Design to Maximize Activity and Minimize off-Target Effects of CRISPR-Cas9.” *Nature Biotechnology* 34 (2): 184–91.

Lundholt, Betina Kerstin, Kurt M. Scudder, and Len Pagliaro. 2003. “A Simple Technique for Reducing Edge Effect in Cell-Based Assays.” *Journal of Biomolecular Screening* 8 (5): 566–70.

Rohban, Mohammad Hossein, Shantanu Singh, Xiaoyun Wu, Julia B. Berthet, Mark-Anthony Bray, Yashaswi Shrestha, Xaralabos Varelas, Jesse S. Boehm, and Anne E. Carpenter. 2017. “Systematic Morphological Profiling of Human Gene and Allele Function via Cell Painting.” *eLife* 6 (March). https://doi.org/10.7554/eLife.24060.