nature research

Peer Review Information

Journal: Nature Methods
Manuscript Title: High-Dimensional Gene Expression and Morphology Profiles of Cells across 28,000 Genetic and Chemical Perturbations
Corresponding author name(s): Shantanu Singh, Marzieh Haghighi

Editorial Notes:

Reviewer Comments & Decisions:

Decision Letter, initial version:

Dear Anne,

Thank you for your letter detailing how you would respond to the reviewer concerns regarding your Resource, "High-Dimensional Gene Expression and Morphology Profiles of Cells across 28,000 Genetic and Chemical Perturbations". We have decided to invite you to revise your manuscript as you have outlined, before we reach a final decision on publication.

We think analyzing your data with Seurat/MOFA could be interesting for readers, so if it is not to onerous, please do add this. We do not think you need to make a Bioconductor package. Regarding the leave-one-out experiments, we leave this at your discretion. If you explore these experiments and they turn out to be informative, it would be interesting to include them.

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further.

When revising your paper:

* include a point-by-point response to the reviewers and to any editorial suggestions

* please underline/highlight any additions to the text or areas with other significant changes to facilitate review of the revised manuscript

* address the points listed described below to conform to our open science requirements

* ensure it complies with our general format requirements as set out in our guide to authors at www.nature.com/naturemethods

* resubmit all the necessary files electronically by using the link below to access your home page

[Redacted] This URL links to your confidential home page and associated information about manuscripts you may have submitted, or that you are reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage.

We hope to receive your revised paper within XX weeks [**ED TO CUSTOMIZE AS NEEDED**]. If you cannot send it within this time, please let us know. In this event, we will still be happy to reconsider your paper at a later date so long as nothing similar has been accepted for publication at Nature Methods or published elsewhere.

OPEN SCIENCE REQUIREMENTS

REPORTING SUMMARY AND EDITORIAL POLICY CHECKLISTS
When revising your manuscript, please update your reporting summary and editorial policy checklists.

Reporting summary: https://www.nature.com/documents/nr-reporting-summary.zip
Editorial policy checklist: https://www.nature.com/documents/nr-editorial-policy-checklist.zip

If your paper includes custom software, we also ask you to complete a supplemental reporting summary.

Software supplement: https://www.nature.com/documents/nr-software-policy.pdf

Please submit these with your revised manuscript. They will be available to reviewers to aid in their evaluation if the paper is re-reviewed. If you have any questions about the checklist, please see http://www.nature.com/authors/policies/availability.html or contact me.

Please note that these forms are dynamic 'smart pdfs' and must therefore be downloaded and completed in Adobe Reader. We will then flatten them for ease of use by the reviewers. If you would like to reference the guidance text as you complete the template, please access these flattened versions at http://www.nature.com/authors/policies/availability.html.

DATA AVAILABILITY

Please include a "Data availability" subsection in the Online Methods. This section should inform readers about the availability of the data used to support the conclusions of your study, including accession codes to public repositories, references to source data that may be published alongside the paper, unique identifiers such as URLs to data repository entries, or data set DOIs, and any other statement about data availability. At a minimum, you should include the following statement: "The data that support the findings of this study are available from the corresponding author upon request", describing which data is available upon request and mentioning any restrictions on availability. If DOIs are provided, please include these in the Reference list (authors, title, publisher (repository name), identifier, year). For more guidance on how to write this section please see: http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf

CODE AVAILABILITY

Please include a "Code Availability" subsection in the Online Methods which details how your custom code is made available. Only in rare cases (where code is not central to the main conclusions of the paper) is the statement "available upon request" allowed (and reasons should be specified).

We request that you deposit code in a DOI-minting repository such as Zenodo, Gigantum or Code Ocean and cite the DOI in the Reference list. We also request that you use code versioning and provide a license.

For more information on our code sharing policy and requirements, please see: https://www.nature.com/nature-research/editorial-policies/reporting-standards#availability-of-computer-code

Please do not hesitate to contact me if you have any questions or would like to discuss these revisions further. We look forward to seeing the revised manuscript and thank you for the opportunity to consider your work.

Sincerely,
Rita

Rita Strack, Ph.D.
Senior Editor
Nature Methods

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
The manuscript reported four datasets of gene expression (GE) and morphology (CP) profiles from two cell lines (A549 and U2OS) across 28,000 genetic and chemical perturbations. The authors demonstrated the values of the four datasets with two example applications: 1) using them to predict profiles from one type to the other type, and 2) integrating the two types of profiles to predict the mechanisms of actions for compounds (compound functional categories/labels).

The datasets are valuable and appear to interest a broad research community, such as biologists and bioinformaticians, for further scientific discovery. Since the raw data are complex and domain-specific (e.g., cell painting), the preprocessed data will facilitate many applications for potential users and readers.

However, the unique contribution or significance of the work is not clearly stated. Some authors were among the original paper's authors producing the cell painting assay data, but other raw data appear to be generated by others. On the other hand, although the two types of example applications are impressive, they are still conventional, and some are even oversimplified (details below).

The authors stated both shared subspace and complementary subspace among GE and CP data have their values, which provided an interesting insight into multi-omic data analysis. However, they used straightforward integration methods, either early or late fusion, which may undermine the power of the shared subspace. Seurat (for unmatched samples) or MOFA (for matched samples) will address the argument.

Besides, it is unclear whether matched samples existed for the gene expression and morphology profiles, i.e., they grew from the same cell lines in the same lab and under the same treatment (e.g., the same compound and dosage). Samples from different labs may complicate the integration and predictions. Please clarify.

Other specific concerns:

1) Can the preprocessing/integration pipeline be generalized to similar data from other cell lines if data are available?
2) Please explicitly define the concept of a sample and clarify whether it corresponds to a replicate or a treatment of multiple replicates. The same concern exists for pairs of profiles.
3) Several figures reported negative R² scores for accuracy (e.g., Fig 3a). Please double-check whether it is a correlation value as squared values are nonnegative.
4) Please describe the rows and columns of a matrix, such as X_{cp} and X_{ge}, in the section of cross-modality predictions.
5) What did the size of circles in Fig. 3b stand for?

Reviewer #2:
Remarks to the Author:
A. Description of the method or tool and any key results

The authors have organized, processed, and curated a collection of data sets with both gene expression (L1000) and cell morphology (CellPainter) profiles. The data and code have been made available and will likely be a valuable resource to the research community, especially those working to develop methods to integrate these two data modalities.

B. Originality and significance of the method and result: if not novel, please include reference

As the authors note, their own group and several others have previously jointly analyzed gene expression and cell morphology. The primary novelty here is the increased number of data sets and consistent processing and organization of these data. The results presented in the the two example applications are not examined in sufficient depth to provide novel methodological or biological insights in their own regard. While this may be beyond the scope of a Resource paper, a deeper dive into one or both of these applications would greatly increase the significance of this work by clearly demonstrating the importance of making this resource available.

C. Data & methodology: validity of approach, comparison to available techniques, ease of adoption, quality of data, quality of presentation

Obviously, it would be great to include even more cell lines and perturbations, but these data represent the largest collection of this type to my knowledge.

In Figure 5, the correlation between replicates for the gene expression data appears quite weak. I would appreciate your thoughts on why this is expected / not concerning.

I believe a key aspect of these data is that the use of cell lines increases the likelihood that roughly the same population of cells are being measured by both assays. This assumption might not hold for adjacent tissue slices used in GTEx or TCGA.

D. Appropriate use of statistics and treatment of uncertainties

The comparisons to randomized data in the assessments were interesting and potentially point to some sources of technical variation that exist in these data. If we assume these sources of technical variation may be data set specific, it would be interesting to see how the predictions performed if you performed leave-one-dataset-out CV (rather than k-fold CV) in Application 1: Predicting gene expression and morphology from each other.

It was unclear to me why leave-one-compound-out CV was used for Application 2 while k-fold CV was used for Application 1. Was this due to the computational complexity or some other aspect of these applications?

For both applications, was the QC, filtering, and preprocessing performed within the CV? If not, would that not risk over-stating the prediction performance?

In the prediction results, some care should be taken when discussing true and false positives. Both the gene expression and cell morphology data do not represent the truth — these are still estimates based on measurements and subject to their own errors and uncertainties.

E. Conclusions: robustness, validity, reliability

The main conclusion, that this resource will be useful to the research community, is reasonably well supported by the manuscript.

F. Methodological details and algorithms: everything necessary to reproduce the technique?

The authors provide a website: http://broad.io/rosetta. This website redirects to a GitHub repo with python scripts to reproduce the results, some python notebooks as well as csv and xlsx files. The data itself is available at s3 bucket s3://cellpainting-datasets/Rosetta-GE-CP. Even with only limited knowledge of python, I feel confident I could reproduce the results in the paper.

Is the data so large that it is necessary to store it on AWS? Could it be made available as say a Bioconductor experimental data package? This would provide the benefit of both ease of use and built-in version control.

Are there plans to provide web tools at http://broad.io/rosetta to interactively explore these data? The current set up likely limits the user base of this resource to people comfortable with GitHub, python notebooks / scripts, etc.

G. Suggested improvements: experiments, comparisons, data for possible revision

This likely exists elsewhere, but it would be helpful to provide a brief intro on how to interpret different morphological features. For example what does "Cells_Texture_InfoMeas1_RNA_3_0" mean? Maybe show a few cells spanning the range of values for this and a few other features. After quite a bit of searching, I found a wiki page in another GitHub repo with this information: https://github.com/carpenterlab/2016_bray_natprot/wiki/What-do-Cell-Painting-features-mean%3F Pointing the reader to this wiki from the article might be sufficient.

H. References: appropriate credit to previous work?

Yes.

I. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions

In general the authors did an admirable job writing for a broad audience. However, there are a few spots that may be too elementary for a Nature Methods paper. For example, "Each cell's DNA is transcribed into various mRNA molecules which can be translated into proteins that carry out functions in the cell."

There are a few typos in the manuscript, but overall it is well-written.

Reviewed by: Matthew N. McCall

Reviewer #3:
Remarks to the Author:
Haghighi et al. describe a compilation of curated and matched datasets that combine morphological and transcriptional information across genetic and chemical perturbation experiments, as assessed by the CellPainting and L1000 protocols, respectively. They showcase the utility of such multimodal datasets at two exemplary applications. The presented analyses provide a quantification of the joint and

complementary information that can be extracted from morphology and gene expression data. For the chosen task, the complementarity of the information present in the respective modalities provides little performance improvement.

About half of the included data have been published previously. Similarly, the presented analyses are based on well established tools and methods, with by and large expected results. The major significance of the manuscript thus lies in the massive scale of the compiled data and the convenient computational access via the provided tools. These resources have the potential to become a go-to reference for future methods development, as well as to fuel a broad range of research addressing numerous important biological questions. The presented analyses are technically sound and overall clearly described. The authors' efforts to clarify the technological terms in order to make them accessible to a broad readership should also be commended.

Below please find a few major issues that, in our opinion, should be addressed before publication, as well as several minor comments and corrections for the authors' consideration.

Major comments:
---

1. The existing datasets that were integrated into the resource are only poorly described, not even the acronyms (LUAD, TAORF, etc.) are spelled out. A brief description of where they come from, how they were compiled, and what exactly they contain would be important not only for interpreting the presented results, but also for using the resource.

2. Along the same lines, the experimental setups of all included data should be described with more detail, as the way the data was generated matters for handling and integrating the data correctly. This should include information on the perturbations (method employed for gene overexpression, duration and concentration of treatment, experimental QC performed, plate size and design) and deviations in the protocol from the L1000 and CellPainting publications if any. This is especially important for the new datasets provided, for which there is no publication detailing the data generation.

3. The notebooks provided give a good overview of the performed analyses. We recommend to complete and annotate them even further to make the analyses fully reproducible. Notebooks that are not meant to be reproduced could be annotated as such (e.g. if provided only for clarification of the analysis). It would be particularly useful to provide at least one thoroughly documented example of how perturbations can be matched between modalities (specifically which columns to use for joining the tables) as this is presented as a central point in the manuscript.

Minor comments:
---

4. The data conveniently includes some annotations, such as mechanisms of action for the drug screens, yet it is not clear what sources were used and what can be expected in each file, neither from the paper nor from the repository.

5. We recommend pointing out in the manuscript that the hosted files can be accessed at no cost and without the need for registration, as it may otherwise deter potential users not familiar with the AWS environment. It would also be helpful to mention how that can be done, at least in the repository (downloading the client from https://aws.amazon.com/cli/ and using the `--no-sign-request` parameter).

6. L68: "All datasets were created at our institution (see Methods)": the "Online Methods" section does not include information about the data generation (see comment #2).

7. L105: "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.": this would benefit from concrete examples and references.

8. L117: "shared latent variables form a composite phenotype between morphology and gene expression that can be useful" - The morphological descriptors from histology data (e.g. autoencoder embeddings of cropped histological images in ref #12 describe tissue organization and might differ a lot compared to the ones obtained from high-content imaging, which would describe organelle organization. Making the distinction clearer might be useful.

9. L221-224: "[...] suggests a likely poorer data quality or poorer alignment of the modalities" While displaying higher replicate correlation (which could correspond to higher data quality), the best-performing datasets are also corresponding to the experiments ran in A549 cells. Could the correlation between morphological and transcriptomic features be cell-line dependent? Figure 2c also shows a large overlap between predictable genes for LINCS and LUAD, whose expression patterns might be easier to predict in A549 cells.

10. Figure 1 and 4: please define the used acronyms in the caption.

11. Figure 2: Regressions are often used for descriptive analyses (as opposed to the predictive analysis proposed) in which R2 values are positive. Here the R2 is computed on held-out data and sometimes leads to negative R2 values. Could you comment on that and explain what it means for the model and the data?

12. For the k-fold cross validation, k is not specified (exact values can be found in the notebooks, but not the rationale behind it).

13. Figure 2b: Is there a reason why in the CRDP dataset one gene less than in L1000 is measured?

14. Figure 2d: The network lacks a legend.

15. Figure 2e: You could mention why you change the threshold compared to the previous definition of predictable genes. The source of the family name is unclear (see comment #4).

16. Figure 2d-f: Please clarify in the caption whether these panels are corresponding to the MLP model only.

17. Figure 2f: The dendrogram is truncated. Also consider using a sequential colormap instead of a diverging one.

18. Figure 3a: Why are there no shuffled controls as in figure 2a?

19. Figure 3b: Assuming the categories are of different sizes, it would be more informative to represent the percentages of predictable features or both the predictable and total number of features in each category instead.

20. Figure 3c: The "relative coefficient magnitude" does not seem to be defined.

21. Figure 3c: What do the colors of the bars represent?

22. Figure 2 and 3: As a validation of the models, it would be interesting to see if the landmark genes most likely to affect the morphology (based on the organelles stained by the CP assay) are indeed easy to predict or useful for the prediction.

23. L347: "These data, and methods derived from them, can accelerate drug discovery and therefore improve human health and reduce drug development costs". This strong claim is not backed up by the results yet (as you report that in your example L294 "Trivial early and late fusion of modalities show relatively small improvements upon the performance of the better-performing modality").

24. Appendix A: MOA completeness is reported for the chemical datasets but the number of functional annotations for LUAD and TA-ORF is missing.

25. Subsection "MoA prediction" of the "Online Methods" section: Could you please clarify what constitutes a sample (is it a single concentration of a compound or a replicate thereof)?

26. Figures have low resolution (in the document shared with the reviewers)

We also noticed a couple of typos:
---

27. L224: "of the modalities in the latter two." repeated twice.

28. L260: extra parenthesis.

29. L493: "are computing using"

30. L650,666: extra backslash

Author Rebuttal to Initial comments

Haghighi et al. Response

Thank you for the opportunity to respond! We found the reviewers' comments to be very addressable; it is great to see your "hypothesis" proven correct that a Nature Methods audience would appreciate the value of the resource. Below please find our quick informal comments about the reviews in blue.

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
The manuscript reported four datasets of gene expression (GE) and morphology (CP) profiles from two cell lines {A549 and U2OS) across 28,000 genetic and chemical perturbations.The authors demonstrated the values of the four datasets with two example applications: 1) using them to predict profiles from one type to the other type, and 2) integrating the two types of profiles to predict the mechanisms of actions for compounds (compound functional categories/labels).

The datasets are valuable and appear to interest a broad research community, such as biologists and bioinformaticians,for further scientific discovery. Since the raw data are complex and domain-specific

(e.g., cell painting), the preprocessed data will facilitate many applications for potential users and readers.

However, the unique contribution or significance of the work is not clearly stated. Some authors were among the original paper's authors producing the cell painting assay data, but other raw data appear to be generated by others. On the other hand, although the two types of example applications are impressive, they are still conventional, and some are even oversimplified (details below).

The authors stated both shared subspace and complementary subspace among GE and CP data have their values, which provided an interesting insight into multi-omic data analysis.
However, they used straightforward integration methods, either early or late fusion, which may undermine the power of the shared subspace. Seurat (for unmatched samples) or MOFA (for matched samples) will address the argument.

Indeed, our goal was to provide simple baselines upon which the community could build. Whatever algorithm we might've chosen, readers might wish we would have selected their favorite/most recent method, which might differ based on coming from a biology background versus a machine learning one. Based on our understanding of the MOFA and Seurat methods, they are built on the assumption that the multi modal

measurements are derived from the same set of samples which is not the case in our experiment (we instead have parallel pairs of plates of cells being each treated with chemical/genetic perturbations and then separately being analyzed for mRNA or morphology). However, we are open to adding any state of the art method that may be a good fit to this problem and is not already included in our work.

Besides, it is unclear whether matched samples existed for the gene expression and morphology profiles, i.e., they grew from the same cell lines in the same lab and under the same treatment (e.g., the same compound and dosage). Samples from different labs may complicate the integration and predictions. Please clarify.
For each of the datasets, in a single laboratory, cells are plated into two identical plates, each one gets treated with drugs identically and then one plate goes to L1000 and the other to Cell Painting. Sorry for not mentioning this detail in our manuscript, we will add it.

Other specific concerns:
1) Can the preprocessing/integration pipeline be generalized to similar data from other cell lines if data are available?
- The computational pipeline is cell line independent and can be used for any cell line. We demonstrated it on two cell lines in this resource.

2) Please explicitly define the concept of a sample and clarify whether it corresponds to a replicate or a treatment of multiple replicates. The same concern exists for pairs of profiles.

- We have defined a sample in our manuscript as a treatment level profile which is aggregation of all the replicate-level profiles of a perturbation. We have described how a replicate level profile is defined as well. We will add more clear referencing to the current version to ensure clarity of these definitions.

Based on our manuscript: "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."

3) Several figures reported negative R2 scores for accuracy (e.g., Fig 3a). Please double-check whether it is a correlation value as squared values are nonnegative.

- We understand this is confusing: RA2 or Coefficient of Determination can have negative values when the fit is worse than a constant average value of data (https://web.maths.unsw.edu.au/-adelle/Garvan/Assays/GoodnessOfFit.html). We will clarify this in the paper.

4) Please describe the rows and columns of a matrix, such as x..and Xg., in the section of cross-modality predictions.

- Will add this detail to the corresponding section.

5) What did the size of circles in Fig. 3b stand for?

Size of each circle is proportional to the number of highly predictable features in each category. Sorry that this detail is missing in our manuscript. We will add it.

Reviewer #2:

Remarks to the Author:

A. Description of the method or tool and any key results

The authors have organized, processed, and curated a collection of data sets with both gene expression (L1000) and cell morphology (CellPainter) profiles. The data and code have been made available and will likely be a valuable resource to the research community, especially those working to develop methods to integrate these two data modalities.

8. Originality and significance of the method and result: if not novel, please include reference

As the authors note, their own group and several others have previously jointly analyzed gene expression and cell morphology. The primary novelty here is the increased number of data sets and consistent processing and organization of these data. The results presented in the the two example applications are not examined in sufficient depth to provide novel methodological or biological insights in their own regard. While this may be beyond the scope of a Resource paper, a deeper dive into one or both of these applications would greatly increase the significance of this work by clearly demonstrating the importance of making this resource available.

- We agree that presenting novel insights is beyond the scope of a Resource paper :) As the reviewer appreciates, the effort of this paper was in providing four datasets of structured multi-modal data in a unified format, discussing and categorizing the connection of modalities' information content to provide an insightful framework for researchers, and exemplifying two potential applications of these multi-modal data and benchmarking them using baseline ML approaches and metrics.
- We do not think there is a question that datasets with parallel mRNA and morphology readouts are valuable, as multi-modal data analysis is increasing in popularity and no structured resources like this exist.

Nevertheless, even though the main contribution is providing a foundation for methodological advances, there are some interesting insights one can gain from this dataset.

- We would like to point out that although we did not emphasize it, we already present

many biological insights in the paper (some confirmed by literature, but others novel results); for example:

- Most importantly, that many mRNAs are predictable by cell morphology and vice versa, under the conditions of these high-throughput screening assays. This was not known and in our view is quite exciting.
- Similarly, we find that having both the mRNA and morphology data types available improves MOA prediction {though not by much, Fig 4, and integration does not typically help much beyond predictive ability using the best-performing modality alone). This has never been tested and is quite interesting as well. Many might have predicted that morphology would add no information to an mRNA profile.
- Fig 2d: The 58 genes' mRNA levels that are predictable by morphology fall into a very small number of categories: phosphoprotein, acetylation and cytosol
- Fig 2f: STX4 is a gene whose mRNA levels are strongly predictable by ER and AGP texture features, consistent with its known functions.
- Fig 3b: texture features are the most predictable morphological features of cells (given an mRNA profile).

- Fig 3c: 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 4b: particular MOAs are better-predicted using morphology, others by mRNA, and very few by integrating both. The identities of these have never been reported.

We can certainly choose a few more examples to explore for some of these (e.g. mRNAs well-predicted by morphology and vice versa), showing images of cells from populations with high and low levels of each mRNA, for example. It is unfortunately beyond the ability of our dry lab to choose a relationship to explore in mechanistic detail.

C. Data & methodology: validity of approach, comparison to available techniques, ease of adoption, quality of data, quality of presentation

Obviously, it would be great to include even more cell lines and perturbations, but these data represent the largest collection of this type to my knowledge.

In Figure 5, the correlation between replicates for the gene expression data appears quite weak. I would appreciate your thoughts on why this is expected I not concerning.

- That is an important observation, we also noticed that Cell Painting data has a much stronger "profile replicate reproducibility". As many confounders could contribute to such an observation, we didn't make any conclusion about how much of this difference might go to the existence of biological signature versus experimental noise. We instead processed both data modalities similarly to each other and removed samples with minimum replicate reproducibility. Further domain specific denoising techniques such as batch effect corrections could potentially improve the quality of profiles and that can be a target for future computational research on this dataset.

For one of the datasets (CORP), this correlation is specifically very poor and the reviewer is likely referring to the plots for that dataset. We have stated in the •Filtering samples• paragraph referring to Figure 5, that: "One of the chemical datasets
(CORP-BBBC047-Bray) has a subset of compounds that are known to be bioactive. We refer to this subset as CORP-bio-BBBC036-Bray and report the details independently for this dataset in Table 1 and Appendix A and B. We only use CORP-bio and not the full CORP set for the analysis in this paper. We did so because we believe that the quality of CORP 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)."

I believe a key aspect of these data is that the use of cell lines increases the likelihood that roughly the same population of cells are being measured by both assays. This assumption might not hold for adjacent tissue slices used in GTEx or TCGA.
Indeed! We can add this insight.

0. Appropriate use of statistics and treatment of uncertainties

The comparisons to randomized data in the assessments were interesting and potentially point to some sources of technical variation that exist in these data. If we assume these sources of technical variation may be data set specific, it would be interesting to see how the predictions performed if you performed leave-one-dataset-out CV (rather than k-fold CV) in Application 1: Predicting gene expression and morphology from each other.
It is feasible to add this experiment to the paper.

It was unclear to me why leave-one-compound-out CV was used for Application 2 while k-fold CV was used for Application 1. Was this due to the computational complexity or some other aspect of these applications?
- That is indeed because of computational complexity; in Application 2, leaving each compound together with all its doses out will result in less than a thousand folds. By contrast, in Application 1, for each landmark gene, we apply k-fold cross validation over all the genetic or chemical perturbation samples. Therefore, there are k•1000 models to be learned in each dataset and this number will be much higher if we have •k=number of samples• as in leave one perturbation out cross validation. We can clarify this in the text.

For both applications, was the QC, filtering, and preprocessing performed within the CV? If not, would that not risk over-stating the prediction performance?
- The preprocessing involves profile standardization and filtering based on replication correlation: we do not think cross validation is necessary as they are all unsupervised and there are no parameters for these steps to be tuned.

In the prediction results, some care should be taken when discussing true and false positives. Both the gene expression and cell morphology data do not represent the truth - these are still estimates based on measurements and subject to their own errors and uncertainties.

Definitely. We will emphasize this in the text.

E. Conclusions: robustness, validity, reliability

The main conclusion, that this resource will be useful to the research community, is reasonably well supported by the manuscript.

F. Methodological details and algorithms: everything necessary to reproduce the technique?

The authors provide a website: http://broad.io/rosetta. This website redirects to a GitHub repo with python scripts to reproduce the results, some python notebooks as well as csv and xlsx files. The data itself is available at s3 bucket s3://cellpainting-datasets/Rosetta-GE-CP. Even with only limited knowledge of python, I feel confident I could reproduce the results in the paper.

Is the data so large that it is necessary to store it on AWS? Could it be made available as say a Bioconductor experimental data package? This would provide the benefit of both ease of use and built-in version control.

- We are in the process of making the data available through an S3 bucket supported by AWS Open Data and it is going to be version controlled. We are not familiar with Bioconductor data packages, and are happy to hear how this might be useful to the community.

Are there plans to provide web tools at http://broad.io/rosetta to interactively explore these data? The current set up likely limits the user base of this resource to people comfortable with GitHub, python notebooks/ scripts, etc.

No, although our lab is passionate about making software available to non-experts, we don't believe it is suitable to create an interactive exploratory web user interface for this work. Most of the relevant analyses are intense and not suited for casual browsing or analyses by non-computational biologists.

G. Suggested improvements: experiments, comparisons, data for possible revision

This likely exists elsewhere, but it would be helpful to provide a brief intro on how to interpret different morphological features. For example what does "Cells_Texture_InfoMeas1_RNA_3_0" mean? Maybe show a few cells spanning the range of values for this and a few other features. After quite a bit of searching, I found a wiki page in another GitHub repo with this information: https://github.com/carpenterlab/2016 bray natproVwiki/What-do-Cell-Painting-features-mean¾ 3F Pointing the reader to this wiki from the article might be sufficient.

- Thanks for your suggestion. We will add it to the manuscript.

H. References: appropriate credit to previous work?

Yes.

I. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions

In general the authors did an admirable job writing for a broad audience. However, there are a few spots that may be too elementary for a Nature Methods paper. For example, "Each cell's DNA is transcribed into various mRNA molecules which can be translated into proteins that carry out functions in the cell."

- We revised heavily when converting the paper from a machine learning audience to a computational biology one, but apparently missed a spot :D

There are a few typos in the manuscript, but overall it is well-written.

Reviewed by: Matthew N. McCall

Reviewer #3:
Remarks to the Author:
Haghighi et al. describe a compilation of curated and matched datasets that combine morphological and transcriptional information across genetic and chemical perturbation experiments, as assessed by the CellPainting and L1000 protocols, respectively. They showcase the utility of such multimodal datasets at two exemplary applications. The presented analyses provide a quantification of the joint and complementary information that can be extracted from morphology and gene expression data. For the chosen task, the complementarity of the information present in the respective modalities provides little performance improvement.

About half of the included data have been published previously. Similarly, the presented analyses are based on well established tools and methods, with by and large expected results. The major significance of the manuscript thus lies in the massive scale of the compiled data and the convenient computational access via the provided tools. These resources have the potential to become a go-to reference for future methods development, as well as to fuel a broad range of research addressing numerous important biological questions. The presented analyses are technically sound and overall clearly described. The authors' efforts to clarify the technological terms in order to make them accessible to a broad readership should also be commended.

Below please find a few major issues that, in our opinion, should be addressed before publication, as well as several minor comments and corrections for the authors' consideration.

Major comments:

1. The existing datasets that were integrated into the resource are only poorly described, not even the acronyms (LUAD, TAORF, etc.) are spelled out. A brief description of where they come from, how they were compiled, and what exactly they contain would be important not only for interpreting the presented results, but also for using the resource.
- We will be happy to add a brief description for each dataset to the appendix.

2. Along the same lines, the experimental setups of all included data should be described with more detail, as the way the data was generated matters for handling and integrating the data correctly. This should include information on the perturbations (method employed for gene overexpression, duration and concentration of treatment, experimental QC performed, plate size and design) and deviations in the protocol from the L1000 and CellPainting publications if any. This is especially important for the new datasets provided, for which there is no publication detailing the data generation.
Can do.

3. The notebooks provided give a good overview of the performed analyses. We recommend to complete and annotate them even further to make the analyses fully reproducible. Notebooks that are not meant to be reproduced could be annotated as such (e.g. if provided only for clarification of the analysis). It would be particularly useful to provide at least one thoroughly documented example of how perturbations can be matched between modalities (specifically which columns to use for joining the tables) as this is presented as a central point in the manuscript.
- We cancertainly add more documentation as described.

Minor comments:

- We thank the reviewer for taking the time to notice and remark on these issues below -
we can address them accordingly.

4. The data conveniently includes some annotations, such as mechanisms of action for the drug screens, yet it is not clear what sources were used and what can be expected in each file, neither from the paper nor from the repository.

5. We recommend pointing out in the manuscript that the hosted files can be accessed at no cost and without the need for registration, as it may otherwise deter potential users not familiar with the AWS environment. It would also be helpful to mention how that can be done, at least in the repository (downloading the client from https://aws.amazon.com/cli/ and using the · -no-sign-request' parameter).

6. L68: "All datasets were created at our institution (see Methods)": the "Online Methods" section does not include information about the data generation (see comment #2).

7. L105: "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.": this would benefit from concrete examples and references.

8. L117: "shared latent variables form a composite phenotype between morphology and gene expression that can be useful" - The morphological descriptors from histology data (e.g. autoencoder embeddings of cropped histological images in ref #12 describe tissue organization and might differ a lot compared to the ones obtained from high-content imaging, which would describe organelle organization. Making the distinction clearer might be useful.

9. L221-224: "[...] suggests a likely poorer data quality or poorer alignment of the modalities" While displaying higher replicate correlation (which could correspond to higher data quality), the best-performing datasets are also corresponding to the experiments ran in A549 cells. Could the correlation between morphological and transcriptomic features be cell-line dependent? Figure 2c also shows a large overlap between predictable genes for LINC$ and LUAD, whose expression patterns might be easier to predict in A549 cells.

10. Figure 1 and 4: please define the used acronyms in the caption.

11. Figure 2: Regressions are often used for descriptive analyses (as opposed to the predictive analysis proposed) in which R2 values are positive. Here the R2 is computed on held-out data and sometimes leads to negative R2 values. Could you comment on that and explain what it means for the model and the data?

12. For the k-fold cross validation, k is not specified (exact values can be found in the notebooks, but not the rationale behind it).

13. Figure 2b: Is there a reason why in the CROP dataset one gene less than in L1000 is measured?

14. Figure 2d: The network lacks a legend.

15. Figure 2e: You could mention why you change the threshold compared to the previous definition of predictable genes. The source of the family name is unclear (see comment #4).

16. Figure 2d-f: Please clarify in the caption whether these panels are corresponding to the MLP model only.

17. Figure 2f: The dendrogram is truncated. Also consider using a sequential colormap instead of a diverging one.

18. Figure 3a: Why are there no shuffled controls as in figure 2a?

19. Figure 3b: Assuming the categories are of different sizes, it would be more informative to represent the percentages of predictable features or both the predictable and total number of features in each category instead.

20. Figure 3c: The "relative coefficient magnitude" does not seem to be defined.

21. Figure 3c: What do the colors of the bars represent?

22. Figure 2 and 3: As a validation of the models, it would be interesting to see if the landmark genes most likely to affect the morphology (based on the organelles stained by the CP assay) are indeed easy to predict or useful for the prediction.

23. L347: "These data, and methods derived from them, can accelerate drug discovery and therefore improve human health and reduce drug development costs". This strong claim is not backed up by the results yet (as you report that in your example L294 ''Trivial early and late fusion of modalities show relatively small improvements upon the performance of the better-performing modality").

24. Appendix A: MOA completeness is reported for the chemical datasets but the number of functional annotations for LUAD and TA-ORF is missing.

25. Subsection "MoA prediction" of the "Online Methods" section: Could you please clarify what constitutes a sample (is it a single concentration of a compound or a replicate thereof)?

26. Figures have low resolution (in the document shared with the reviewers)

We also noticed a couple of typos:

27. L224: "of the modalities in the latter two." repeated twice.

28. L260: extra parenthesis.

29. L493: "are computing using"

30. L650,666: extra backslash

Decision Letter, first revision:

Dear Anne,

Thank you for submitting your revised manuscript "High-Dimensional Gene Expression and Morphology Profiles of Cells across 28,000 Genetic and Chemical Perturbations" (NMETH-RS46726B). It has now been seen by the original referees and their comments are below. The reviewers find that the paper has improved in revision, and therefore we'll be happy in principle to publish it in Nature Methods, pending (very) minor revisions to satisfy the referees' final requests and to comply with our editorial and formatting guidelines.

We are now performing detailed checks on your paper and will send you a checklist detailing our editorial and formatting requirements in about a week. Please do not upload the final materials and make any revisions until you receive this additional information from us.

TRANSPARENT PEER REVIEW
Nature Methods offers a transparent peer review option for new original research manuscripts submitted from 17th February 2021. We encourage increased transparency in peer review by publishing the reviewer comments, author rebuttal letters and editorial decision letters if the authors agree. Such peer review material is made available as a supplementary peer review file. Please state in the cover letter 'I wish to participate in transparent peer review' if you want to opt in, or 'I do not wish to

participate in transparent peer review' if you don't. Failure to state your preference will result in delays in accepting your manuscript for publication.
Please note: we allow redactions to authors' rebuttal and reviewer comments in the interest of confidentiality. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons. Reviewer names will be published in the peer review files if the reviewer signed the comments to authors, or if reviewers explicitly agree to release their name. For more information, please refer to our FAQ page.

Thank you again for your interest in Nature Methods Please do not hesitate to contact me if you have any questions.

Sincerely,
Rita

Rita Strack, Ph.D.
Senior Editor
Nature Methods

ORCID
IMPORTANT: Non-corresponding authors do not have to link their ORCIDs but are encouraged to do so. Please note that it will not be possible to add/modify ORCIDs at proof. Thus, please let your co-authors know that if they wish to have their ORCID added to the paper they must follow the procedure described in the following link prior to acceptance:
https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research

Reviewer #1 (Remarks to the Author):

The authors made substantial additions to this revision. My concerns have been satisfactorily addressed.

A typo on page 57: The word 'morphological' in the sentence 'X_ge is the whole morphological data matrix representing all L
landmark genes measurements across all the P perturbations' appears to be 'gene expression'.

Reviewer #2 (Remarks to the Author):

The authors have addressed all of my comments, and in several places, have gone well beyond what I expected in the thoroughness of their responses.

Reviewer #3 (Remarks to the Author):

The authors have considerably revised their manuscript and addressed, by and large, the comments raised by the reviewers. The updated content now makes the resource more easily accessible and reusable. New additions, including a prediction of gene expression across datasets and an unsupervised analysis of joint expression and imaging data, demonstrate the usefulness of the resource.

Here are a few minor improvements that could still be made:

+ Caption Fig.2: "The y-axis is trimmed at -1 for clarity". This seems to be the case in Figure 3 but not in Figure 2.

+ Figure 3b lacks a legend for the size of the circles.

+ Supplementary Figure 5: "stress fiber" is truncated.

+ Typo: "for the entire genome 7the specific genes' mRNAs"

+ The "Editorial Policy Checklist" mentions that "Box-plot elements are defined (e.g. center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers)." This is not the case everywhere.

Author Rebuttal, first revision:

Haghighi et al. Response

We thank the reviewers for their insightful comments. Below please find our response to reviewers in blue text; for many responses we also added new text to the manuscript. This is highlighted in cyan in the response, except for cases where the new added text was very long or needs to be seen in the context of the main text; all the added or updated text is highlighted in cyan in the main manuscript as well.

Reviewers' Comments:

Reviewer #1:
Remarks to the Author:
The manuscript reported four datasets of gene expression (GE) and morphology (CP) profiles from two cell lines (A549 and U2OS) across 28,000 genetic and chemical perturbations. The authors demonstrated the values of the four datasets with two example applications: 1) using them to predict profiles from one type to the other type, and 2) integrating the two types of profiles to predict the mechanisms of actions for compounds (compound functional categories/labels).

The datasets are valuable and appear to interest a broad research community, such as biologists and bioinformaticians, for further scientific discovery. Since the raw data are complex and domain-specific (e.g., cell painting), the preprocessed data will facilitate many applications for potential users and readers.

However, the unique contribution or significance of the work is not clearly stated. Some authors were among the original paper's authors producing the cell painting assay data, but other raw data appear to be generated by others.

- Thank you for the encouragement to summarize the unique contribution or significance of our work here (we have also adjusted the main text to be more clear)):

Dataset:

- As a resource paper, the unique contribution of this work is in the curation and combination of existing datasets with multiple modalities into a readily usable format, which is not trivial. Providing all four datasets in one place and format provides for the first time a collection with both genetic and chemical perturbation of cells with two different kinds of profiling readouts. All of the data was made at our Institute, though involving many others beyond the authors of this paper. Reviewer 2 emphasized this and

Reviewer 3 stated it more eloquently than we did: “The major significance of the manuscript thus lies in the massive scale of the compiled data and the convenient computational access via the provided tools. These resources have the potential to become a go-to reference for future methods development, as well as to fuel a broad range of research addressing numerous important biological questions.”

Benchmarking on two applications (including code and metrics):

- We discuss and categorize the connection of modalities’ information content to provide an insightful framework for thinking about multi-modal data, exemplifying two potential applications and benchmarking them using baseline ML approaches and metrics.

- The task of exploring the association between changes in cell morphology and transcriptomic data has been briefly studied in the literature for one dataset (related works are cited in the paper). However, the lack of defined baseline tasks and evaluation metrics and multiple independent datasets made it difficult for the ML community to build on the existing literature and datasets until now.

- Mechanism-of-action prediction is an important and challenging step in the drug discovery process. Much effort has been devoted to it, including in the machine learning community. However, our paper is the first to our knowledge to attempt MoA prediction by integrating gene expression and morphology together and thus provides the dataset and evaluation metrics to assess future methods for multi-modal data integration.

Biological insights:

- We gained interesting biological insights from this dataset (some confirmed by literature, but others novel results). We had not emphasized these for a computational audience but have now revised the paper to do so. While it is unfortunately beyond the ability of our dry lab to choose a relationship to explore in mechanistic detail, we also performed several new analyses and dug into some biological stories to add to the paper. Please see Reviewer 2 response on "B. Originality and significance of the method and result" for a comprehensive description of these.

On the other hand, although the two types of example applications are impressive, they are still conventional, and some are even oversimplified (details below).
The authors stated both shared subspace and complementary subspace among GE and CP data have their values, which provided an interesting insight into multi-omic data analysis. However, they used straightforward integration methods, either early or late fusion, which may undermine the power of the shared subspace. Seurat (for unmatched samples) or MOFA (for matched samples) will address the argument.

- We understand the reviewer's concern about data integration methods. Our rationale for choosing relatively simple early and late fusion methods was that the application where we integrate data modalities is MOA classification – a supervised task. In early or late

fusion strategies the classifier is supposed to find the important features/dimensions and focus on them, minimizing the need for the fusion/integration method itself to be particularly sophisticated.

- Nevertheless, in our revision, we also include performance on clustering – an unsupervised task for which data integration is much more critical. We evaluate the information content of a clustering based on how well the clusters align with MOA labels. Complementarity is thus quantified by the increase in the information content when integrating modalities vs. a single modality.

- In our revision, motivated by this suggestion, we now include seven state-of-the-art methods (including MOFA) for data integration of matched samples (Cantini et al. 2021 https://doi.org/10.1038/s41467-020-20430-7), with results now shown in Figure 4a. We then used the best-performing method (for this unsupervised task; it was RGCCA) to integrate the modalities in the "original" supervised task of MOA prediction. Figure 4b shows the improvement of data integration for MOA classification using this strategy of picking an "optimal" integration method for the data, addressing the source of the reviewer's concern.

Besides, it is unclear whether matched samples existed for the gene expression and morphology profiles, i.e., they grew from the same cell lines in the same lab and under the same treatment (e.g., the same compound and dosage). Samples from different labs may complicate the integration and predictions. Please clarify.

- For each of the datasets, in a single laboratory, cells are plated into two sets of identical plates, each plate gets treated with drugs identically and then one set goes to L1000 and the other to Cell Painting. We have now added this detail to our manuscript:
- For each of the datasets, in a single laboratory, cells are plated into two sets of identical
plates, each plate gets treated with chemical (or genetic) perturbations identically and then one set is used to measure gene expression and the other set to measure morphology.

Other specific concerns:

1) Can the preprocessing/integration pipeline be generalized to similar data from other cell lines if data are available?

- The computational pipeline is cell line independent and can be used for any cell line (including the two in this resource) - we have added this comment to the manuscript:
- 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 pipeline for data processing and analysis were identical regardless of the cell type in the experiment.

2) Please explicitly define the concept of a sample and clarify whether it corresponds to a replicate or a treatment of multiple replicates. The same concern exists for pairs of profiles.

- Thank you for pointing out the murky definition. We have now changed the text as following to minimize the confusion in definitions:

- In the main text, we use the term sample to refer to a sample of cells (e.g. a sample of cells in a well).

- In Online Methods, we changed the sample in the statistical sense to “data point”, and we have clarified what it means in the “Online methods” section:
In the rest of the text, unless indicated otherwise, “data points” refer to treatment-level profiles.
- We have now clarified “pair of profiles” in “Measuring quality of data points for subsequent analysis” section, as “pair of replicate-level profiles” which mean each random pair selected from the replicate-level profiles of a given perturbation.

3) Several figures reported negative R2 scores for accuracy (e.g., Fig 3a). Please double-check whether it is a correlation value as squared values are nonnegative.
- We understand this is confusing! R^2 or Coefficient of Determination can have negative values when the fit is worse than simply fitting a flat line, with the y-intercept equal to the average of y_i’s. R^2 values cannot be negative when the model is applied to the same data on which it is trained (unless the model is missing a constant term https://web.maths.unsw.edu.au/~adelle/Garvan/Assays/GoodnessOfFit.html; our models do have a constant term). But R^2 values can be negative when the model is trained on one set of data and applied to another (test) set – this is the case for us. In other words, in our case when the R^2 is negative, it is because the fit on the test set is worse than simply fitting the average value (of the response variable) of the test set, indicating that the model does not generalize at all.

4) Please describe the rows and columns of a matrix, such as Xcp and Xge, in the section of cross-modality predictions.
- We have added more details as requested to the corresponding section.

5) What did the size of circles in Fig. 3b stand for?
- The size of each circle was proportional to the number of highly-predictable features in each category. We have now changed the figure in this version so that the sizes of circles are proportional to the percentage of highly-predictable features in each category instead of numbers. We have added the details to our manuscript.

Reviewer #2:
Remarks to the Author:
A. Description of the method or tool and any key results

The authors have organized, processed, and curated a collection of data sets with both gene expression (L1000) and cell morphology (CellPainter) profiles. The data and code have been

made available and will likely be a valuable resource to the research community, especially those working to develop methods to integrate these two data modalities.

B. Originality and significance of the method and result: if not novel, please include reference

As the authors note, their own group and several others have previously jointly analyzed gene expression and cell morphology. The primary novelty here is the increased number of data sets and consistent processing and organization of these data. The results presented in the two example applications are not examined in sufficient depth to provide novel methodological or biological insights in their own regard. While this may be beyond the scope of a Resource paper, a deeper dive into one or both of these applications would greatly increase the significance of this work by clearly demonstrating the importance of making this resource available.

- We agree that presenting novel biological insights is beyond the scope of a Resource paper :) As the reviewer appreciates, the effort of this paper was in providing four datasets of structured multi-modal data in a unified format, discussing and categorizing the connection of modalities' information content to provide an insightful framework for researchers, and exemplifying two potential applications of these multi-modal data and benchmarking them using baseline ML approaches and metrics. We believe that the scientific community will appreciate these datasets with parallel mRNA and morphology readouts, as multi-modal data analysis is increasing in popularity and no structured resources like this exist.
- Nevertheless, beyond those main contributions, there are some interesting biological insights we gained from this dataset (some confirmed by literature, but others novel results). We had not emphasized these for a computational audience but have now revised the paper to do so. While it is unfortunately beyond the ability of our dry lab to choose a relationship to explore in mechanistic detail, we also performed several new analyses and dug into some biological stories to add to the paper. We added text throughout the paper and especially in the Discussion section. In summary, our discoveries are:
  - Most importantly, we discovered that many mRNAs are predictable by cell morphology and vice versa, under the conditions of these high-throughput assays. This was not known and in our view is quite exciting.
  - Similarly, many of our colleagues told us that they predicted that morphology profiles would capture no information that couldn't already be seen in an mRNA profile - that it was impossible for cells to show a morphological change without mRNA profiles changing, whether as a cause or consequence. Our study shows this is not the case, and that having both the mRNA and morphology data types available improves MOA prediction (though not dramatically, Fig 4).
  - We identified particular genes' mRNAs that are predictable by particular morphological features (and vice versa). Most remarkably, we discovered a cluster of genes (SPP1,

CDK4, S100A4, BNIP3L, HDAC6 and S100A13) whose mRNA levels are strongly predictable by RNA texture features. Despite these genes having little existing functional annotation in common, cells that are

predicted to have (and actually do have) high levels of these six genes' mRNA all are associated with visible changes in the staining for cytoplasmic RNA and nucleoli. We included images of cells from populations with high and low levels of each mRNA in Supplementary Figure 6.

- We found that highly predictable morphological features of cells tend to fall into the AreaShape feature category (given an mRNA profile; Fig 3b).
- We discovered that highly predictable mRNAs tend to be – in general – associated with compartments that are directly labeled in the Cell Painting assay (Supplementary Table 3 and 4, GO terms search analysis in Online Methods). However, we also discovered that many highly predictable genes were not clearly linked to Cell Painting stains, indicating that the assay probes biological impact beyond the particular labeled components, or that the genes have unannotated functions. This is a novel insight.
- We also find sensible patterns in the set of genes whose expression levels predict specific phenotypes. For example, the morphological feature "Cells_Texture_InfoMeas1_RNA_3_0" (a measure of the smoothness of RNA staining throughout the cell, via the Cell Painting dye SYTO14 which labels RNA) relies on the levels of many genes in its prediction, including several known to be involved in mRNA processing (Fig 3c).
- We found that particular compounds' mechanisms of action are better-predicted using morphology, others by mRNA, and very few by integrating both. The identities of these have never been reported (Fig 3c).

C. Data & methodology: validity of approach, comparison to available techniques, ease of adoption, quality of data, quality of presentation

Obviously, it would be great to include even more cell lines and perturbations, but these data represent the largest collection of this type to my knowledge.

In Figure 5, the correlation between replicates for the gene expression data appears quite weak. I would appreciate your thoughts on why this is expected / not concerning.

- That is an important observation, we also noticed that Cell Painting data has a much stronger correlation between replicates compared to the gene expression data. As many confounders could contribute to such an observation, we didn't make any conclusion about how much of this difference might be explained by stronger biological responses in a given modality versus experimental noise (except for one new note about well position effect; below). We instead processed both data modalities similarly to each other and proceeded with samples ("data points" in this version) that were above a certain threshold of replicate reproducibility (measured by Pearson correlation) for

the applications. Domain-specific denoising techniques such as batch-effect corrections could potentially improve the quality of profiles and that is a target for future computational research on this dataset, as covered in the Discussion.

- For one of the datasets (CDRP), this correlation is specifically very poor and the reviewer is likely referring to the plots for that dataset. We stated in the "Filtering data

points" paragraph referring to Supplementary Figure 1 (previously labeled as Figure 5), that: "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 Supplementary 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 data points remain after filtering for replicate reproducibility across both modalities (see Supplementary B)."

- We have additionally made a note in Online Methods about one likely confounder present in all datasets, that may bias replicability quality metrics in favor of Cell Painting: We note a source of systematic error present in all datasets that may affect replicability metrics: for nearly every treatment, all its replicates occur 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-position effects tend to be more pronounced in Cell Painting than L1000, and therefore the observed differences in data quality – as reported in Supplementary B – can be a function of this batch effect. As we note in the discussion, correcting for batch effects could improve the prediction tasks discussed in this paper, and also make such comparisons of data quality more reliable.

I believe a key aspect of these data is that the use of cell lines increases the likelihood that roughly the same population of cells are being measured by both assays. This assumption might not hold for adjacent tissue slices used in GTEx or TCGA.

- Indeed! We now add this insight:

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.

D. Appropriate use of statistics and treatment of uncertainties

The comparisons to randomized data in the assessments were interesting and potentially point to some sources of technical variation that exist in these data. If we assume these sources of technical variation may be data set specific, it would be interesting to see how the predictions performed if you performed leave-one-dataset-out CV (rather than k-fold CV) in Application 1: Predicting gene expression and morphology from each other.

- Two of the top-performing datasets were from the same cell type (LINCS and LUAD, both in A549 cells) which made them a good candidate to check generalizability of the model across datasets. We added this experiment to the paper (Supplementary Figure 4. Supplementary F), and find that if we train a model using one of these datasets and

test on the other dataset, we are able to predict the mRNA level of a much smaller subset of the genes compared to the intra–dataset predictions. This prediction power is stronger in one direction versus the other: 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. Please also note this text in the Discussion about potential future directions for addressing this challenge:
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, including those resulting from the position of wells on a plate30) to improve performance. Such techniques might also sufficiently align the four datasets with each other, to explore generalized, dataset-independent models.

It was unclear to me why leave-one-compound-out CV was used for Application 2 while k-fold CV was used for Application 1. Was this due to the computational complexity or some other aspect of these applications?

- That was initially because of computational complexity; we wanted to use most of the data in the training set to have a less biased estimate. However, to be consistent for the bias-variance trade off in test estimates of both applications, we used k-fold cross validation for both estimates in this revised version.
- Some notes on computational complexity of experiments for Application 1 versus Application 2: in Application 2, leaving each compound together with all its doses out will result in less than a thousand folds. By contrast, in Application 1, for each landmark gene, we apply nested k-fold cross validation over all the genetic or chemical perturbation samples. Therefore, there are k*1000 models to be learned (plus cross validation for each for hyperparameter tuning) in each dataset and this number will be much higher if we have “k=number of samples“ than in leave one perturbation out cross validation.

For both applications, was the QC, filtering, and preprocessing performed within the CV? If not, would that not risk over-stating the prediction performance?

- There are two steps of data preprocessing, standardization and filtering/QC:
- Standardization of profiles per plate: each feature in a well-level profile (i.e. replicate level) profile is z-scored with respect to the negative control wells on the same plate. The negative control wells are not included in the prediction tasks. Therefore performing this operation within each cross-validation fold is equivalent to performing it prior to the cross-validation.
- The filtering process removes treatments with low replicate correlation. The only parameter in this process is the 90th percentile of the null distribution, which is computed once across the whole dataset. Computing this parameter within each

fold does not seem useful because doing so would result in varying thresholds for quality, and confound the results.

In the prediction results, some care should be taken when discussing true and false positives. Both the gene expression and cell morphology data do not represent the truth — these are still estimates based on measurements and subject to their own errors and uncertainties.

- Definitely. We have emphasized this in the text now.

Furthermore, we note that ground truth in this prediction task is defined only by the available experimental gene expression and cell morphology data, which is subject to technical variation and error and therefore is not absolute truth. In the case of 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 mechanism of action 31. In addition, the ground truth is based on imperfect human knowledge.

E. Conclusions: robustness, validity, reliability

The main conclusion, that this resource will be useful to the research community, is reasonably well supported by the manuscript.

F. Methodological details and algorithms: everything necessary to reproduce the technique?

The authors provide a website: http://broad.io/rosetta. This website redirects to a GitHub repo with python scripts to reproduce the results, some python notebooks as well as csv and xlsx files. The data itself is available at s3 bucket s3://cellpainting-datasets/Rosetta-GE-CP. Even with only limited knowledge of python, I feel confident I could reproduce the results in the paper.

Is the data so large that it is necessary to store it on AWS? Could it be made available as say a Bioconductor experimental data package? This would provide the benefit of both ease of use and built-in version control.

- The data are available on a public AWS S3 bucket at no cost and no need for registration, within individual files directly access through an S3 URL (e.g. https://cellpainting-gallery.s3.amazonaws.com/rosetta/broad/workspace/preprocessed_d ata/LINCS-Pilot1/CellPainting/replicate_level_cp_augmented.csv.gz). In our opinion, storing the data in this manner has several benefits over Bioconductor:
- Individual data files are directly available as S3 objects that can be loaded using most CSV APIs (e.g. readr::read_csv("https://cellpainting-gallery.s3.amazonaws.com/rosetta/broad/wor kspace/preprocessed_data/LINCS-Pilot1/CellPainting/replicate_level_cp_augme nted.csv.gz") just works) independent of programming language being used for analysis.
- Storing it in a widely used format (CSVs) will make it accessible for the broader data science community.

- The editor advised us that making a Bioconductor package was not necessary, but we agree that something equivalent to BIoconductor's built-in version control is really helpful. Although there is no equivalent here, we have provided the Etag (the MD5 equivalent) for each of the objects in the bucket https://github.com/carpenterlab/2021_Haghighi_submitted/blob/b0f603b1ad04c319ca52 4c05b744d1072517755f/README.md#data-version

Are there plans to provide web tools at http://broad.io/rosetta to interactively explore these data? The current set up likely limits the user base of this resource to people comfortable with GitHub, python notebooks / scripts, etc.

- No, although our lab is passionate about making software available to non-experts, we don't believe it is suitable to create an interactive exploratory web user interface for this work. Most of the relevant analyses are intense and not suited for casual browsing or analysis by non-computational biologists.

G. Suggested improvements: experiments, comparisons, data for possible revision

This likely exists elsewhere, but it would be helpful to provide a brief intro on how to interpret different morphological features. For example what does "Cells_Texture_InfoMeas1_RNA_3_0" mean? Maybe show a few cells spanning the range of values for this and a few other features. After quite a bit of searching, I found a wiki page in another GitHub repo with this information: https://github.com/carpenterlab/2016_bray_natprot/wiki/What-do-Cell-Painting-features-mean% 3F Pointing the reader to this wiki from the article might be sufficient.

- Great suggestion. We have added it to the manuscript.

H. References: appropriate credit to previous work?

Yes.

I. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions

In general the authors did an admirable job writing for a broad audience. However, there are a few spots that may be too elementary for a Nature Methods paper. For example, "Each cell's DNA is transcribed into various mRNA molecules which can be translated into proteins that carry out functions in the cell."

- We revised heavily when converting the paper from a machine learning audience to a computational biology one, but apparently missed a spot! We reviewed the manuscript again and updated the language mentioned (and a few other places).

There are a few typos in the manuscript, but overall it is well-written.

Reviewed by: Matthew N. McCall

Reviewer #3:
Remarks to the Author:
Haghighi et al. describe a compilation of curated and matched datasets that combine morphological and transcriptional information across genetic and chemical perturbation experiments, as assessed by the CellPainting and L1000 protocols, respectively. They showcase the utility of such multimodal datasets at two exemplary applications. The presented analyses provide a quantification of the joint and complementary information that can be extracted from morphology and gene expression data. For the chosen task, the complementarity of the information present in the respective modalities provides little performance improvement.

About half of the included data have been published previously. Similarly, the presented analyses are based on well established tools and methods, with by and large expected results. The major significance

of the manuscript thus lies in the massive scale of the compiled data and the convenient computational access via the provided tools. These resources have the potential to become a go-to reference for future methods development, as well as to fuel a broad range of research addressing numerous important biological questions. The presented analyses are technically sound and overall clearly described. The authors' efforts to clarify the technological terms in order to make them accessible to a broad readership should also be commended.

Below please find a few major issues that, in our opinion, should be addressed before publication, as well as several minor comments and corrections for the authors' consideration.

Major comments:

1. The existing datasets that were integrated into the resource are only poorly described, not even the acronyms (LUAD, TAORF, etc.) are spelled out. A brief description of where they come from, how they were compiled, and what exactly they contain would be important not only for interpreting the presented results, but also for using the resource.

- We agree the details were too sparse. We have now improved the description of each dataset by providing references to the assay protocols and additional dataset-specific details related to the assay protocol where necessary.
- We did not spell out the acronyms because all except one (LUAD) are entirely unrelated to the dataset itself – they originate from the project names where they were first created:

- LUAD = Lung Adenocarcinoma
- TA-ORF = Target Accelerator – Open Reading Frames
- CDRP = Center-Driven Research Program
- LINCS = Library of Integrated Network-Based Cellular Signatures

2. Along the same lines, the experimental setups of all included data should be described with more detail, as the way the data was generated matters for handling and integrating the data correctly. This should include information on the perturbations (method employed for gene overexpression, duration and concentration of treatment, experimental QC performed, plate size and design) and deviations in the protocol from the L1000 and CellPainting publications if any. This is especially important for the new datasets provided, for which there is no publication detailing the data generation.

- We agree these are crucial and valuable details.
- Method for gene overexpression: now included
- Duration and concentration of treatments: now included

- Deviations in the protocol from L1000 and Cell Painting: We cite the version of the protocol for each of the two assays, for each dataset.
- Plate size and design; we added:
All experiments were performed in multi-well plates (384-well). The plate design varies across the datasets. The design can be visualized for each plate by inspecting any of the CSV files for each dataset (for example, for Cell Painting, Metadata_Plate and Metadata_Well provide the coordinates, and the rest of the columns with a “Metadata_” prefix have details of each perturbation).

3. The notebooks provided give a good overview of the performed analyses. We recommend to complete and annotate them even further to make the analyses fully reproducible. Notebooks that are not meant to be reproduced could be annotated as such (e.g. if provided only for clarification of the analysis). It would be particularly useful to provide at least one thoroughly documented example of how perturbations can be matched between modalities (specifically which columns to use for joining the tables) as this is presented as a central point in the manuscript.
- We have added more documentation as suggested. We restructured the repository and added easy to use functions for reading and pairing profiles and a notebook to exemplify how these functions can be used (https://github.com/carpenterlab/2021_Haghighi_submitted/blob/main/read_and_match_profiles.ipynb).

Minor comments:

- We thank the reviewer for taking the time to notice and remark on these issues below - we have addressed them accordingly.

4. The data conveniently includes some annotations, such as mechanisms of action for the drug screens, yet it is not clear what sources were used and what can be expected in each file, neither from the paper nor from the repository.
- We have now included this information in the repository https://github.com/carpenterlab/2021_Haghighi_submitted#metadata-information

5. We recommend pointing out in the manuscript that the hosted files can be accessed at no cost and without the need for registration, as it may otherwise deter potential users not familiar with the AWS environment. It would also be helpful to mention how that can be done, at least in the repository (downloading the client from https://aws.amazon.com/cli/ and using the `--no-sign-request` parameter).

- We added more instructions and emphasized that the dataset is available at no cost and with no need for registration.

6. L68: "All datasets were created at our institution (see Methods)": the "Online Methods" section does not include information about the data generation (see comment #2).
- Thanks for noting that; we have added information about data generation to Supplementary A.

7. L105: "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.": this would benefit from concrete examples and references.
- Thanks for the nudge. We have added references and examples as follows: For example, a change in morphology can induce gene expression changes 12 and gene expression 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. For example, changes in protein stability or post-translational

modifications can induce changes in morphology without changes in gene expression; the Rho-family small GTPases are one example where morphology changes on a timescale much too short to be explained by changes in mRNA 17.

8. L117: "shared latent variables form a composite phenotype between morphology and gene expression that can be useful" - The morphological descriptors from histology data (e.g. autoencoder embeddings of cropped histological images in ref #12 describe tissue organization and might differ a lot compared to the ones obtained from high-content imaging, which would describe organelle organization. Making the distinction clearer might be useful.
- Good catch. We have added an appropriate caveat: "...that can be useful; 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."

9. L221-224: "[...] suggests a likely poorer data quality or poorer alignment of the modalities" While displaying higher replicate correlation (which could correspond to higher data quality), the best-performing datasets are also corresponding to the experiments ran in A549 cells. Could the correlation between morphological and transcriptomic features be cell-line dependent?

Figure 2c also shows a large overlap between predictable genes for LINCS and LUAD, whose expression patterns might be easier to predict in A549 cells.

- It is possible! Though it seems even more likely to us that technical variation across experiments is causative. Nevertheless we mention this possibility:
Given LUAD and LINCS are both using 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 to us that the differences in performance relate to differences in technical quality of the data.

10. Figure 1 and 4: please define the used acronyms in the caption.

- We have done so.

11. Figure 2: Regressions are often used for descriptive analyses (as opposed to the predictive analysis proposed) in which R2 values are positive. Here the R2 is computed on held-out data and sometimes leads to negative R2 values. Could you comment on that and explain what it means for the model and the data?

- It means the regression model fitted on the training samples ("data points" in this version) is performing worse than a constant line on the left out sample. We now explain this in Fig 2, where the negative R2 values are first encountered:
Negative R^2 values indicate that the prediction is worse than simply computing the mean of the output, and therefore all R^2 <0 can be considered equally bad (the model does not generalize at all).

12. For the k-fold cross validation, k is not specified (exact values can be found in the notebooks, but not the rationale behind it).

- For MoA analysis, we clarify this in the text:
In the supervised setting, 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 a standard K-fold (K=5) cross-validation on a filtered subset of compounds.
- For feature prediction analysis, given the computational complexity of the feature prediction task, we adjusted the K for each dataset so that the training split is large enough for learning, but not too large so as to be computationally impractical.

13. Figure 2b: Is there a reason why in the CRDP dataset one gene less than in L1000 is measured?

- The landmark gene set used in CDRP is actually different from the rest of the datasets. In “Online Methods” section, “Cell Painting and L1000 Profiles” subsection, we have described the source of difference and have linked to the relevant references:

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, 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.

14. Figure 2d: The network lacks a legend.
- We replaced the network figure with an Over-Representation Analysis of highly predictable landmark genes (Supplementary Figure 5) in the revised version. The change was made since we thought the new analysis would be more accurately interpretable.

15. Figure 2e: You could mention why you change the threshold compared to the previous definition of predictable genes. The source of the family name is unclear (see comment #4).
- As the goal of the presented map was an example map showing the relationship between the expression of each landmark gene and the activation of each category of morphological features, so we simply selected the number of genes based on what could fit into a reasonably sized figure panel for the paper, in this case 153 genes corresponding to R2>0.8.
- The source of the family name is also added to the figure caption.

16. Figure 2d-f: Please clarify in the caption whether these panels are corresponding to the MLP model only.
- We included this information in the updated manuscript.

17. Figure 2f: The dendrogram is truncated. Also consider using a sequential colormap instead of a diverging one.
- We have updated the figure.

18. Figure 3a: Why are there no shuffled controls as in figure 2a?

- We have updated the figure in this version, adding the shuffled control distribution.

19. Figure 3b: Assuming the categories are of different sizes, it would be more informative to represent the percentages of predictable features or both the predictable and total number of features in each category instead.
- That is indeed a great suggestion. We have modified the figure to percentages accordingly and noted this in the caption.

20. Figure 3c: The "relative coefficient magnitude" does not seem to be defined.
- We have updated the caption.

21. Figure 3c: What do the colors of the bars represent?
- They had no meaning - we changed them to a single color to avoid confusion.

22. Figure 2 and 3: As a validation of the models, it would be interesting to see if the landmark genes most likely to affect the morphology (based on the organelles stained by the CP assay) are indeed easy to predict or useful for the prediction.
- This was a neat suggestion. We designed two experiments for testing this hypothesis ("Gene Ontology (GO) terms search analysis" described in the online methods section and the results presented in Supplementary I). The analysis revealed that indeed there is an association, as we now describe in the paper:
To inspect if the observed GE-CP relationships are consistent with the known biological functions of the L1000 landmark genes, we performed a Gene Ontology (GO) terms search analysis (see Online Methods). We wondered whether landmark genes that are highly predictable by morphological features in each specific Cell Painting channel are more likely to have Gene Ontology (GO) annotations related to that channel compared to the rest of CP channels. We indeed saw this to be the case (Supplementary Table 3, Supplementary I; see Online Methods). We also wondered whether landmark genes that are more readily predictable than other genes are more likely to have functions associated with the particular stains in the Cell Painting assay. Indeed, among the set of 58 highly predictable genes (across all the datasets) we saw an increased chance of annotations relating to the cellular components and organelle stained in the assay (Supplementary Table 4, Supplementary I). 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.

23. L347: "These data, and methods derived from them, can accelerate drug discovery and therefore improve human health and reduce drug development costs". This strong claim is not backed up by the results yet (as you report that in your example L294 "Trivial early and late fusion

of modalities show relatively small improvements upon the performance of the better-performing modality").

- We have changed the sentence to avoid this strong claim. It was intended to be long-term focused but we realize it sounded overstated.

24. Appendix A: MOA completeness is reported for the chemical datasets but the number of functional annotations for LUAD and TA-ORF is missing.

- We do not provide functional annotations for LUAD and TA-ORF. We provide the Entrez Gene ID for each gene, which in turn can be used to query databases that provide this information.

25. Subsection "MoA prediction" of the "Online Methods" section: Could you please clarify what constitutes a sample (is it a single concentration of a compound or a replicate thereof)?

- In Online Methods, we changed the sample in the statistical sense to “data point”, and we have clarified what it means in the “Online methods” section:
In the rest of the text, unless indicated otherwise, “data points” refer to treatment-level profiles.

- And we clarify what each data point is in the MoA prediction section:
Note that 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.

26. Figures have low resolution (in the document shared with the reviewers)

- They look good on our side; we attempted to improve their translation to the review-ready PDFs and if this does not solve the problem, we hope the journal will work with us to ensure they are high quality in production.

We also noticed a couple of typos:

27. L224: "of the modalities in the latter two." repeated twice.
28. L260: extra parenthesis.
29. L493: "are computing using"
30. L650,666: extra backslash

- Thanks for noting these. We have corrected the typos.

Author Rebuttal, second revision:

Haghighi et al. Response

We apologize for the delay in preparing the final submission of the manuscript. In the final code

review, we noticed a bug related to the MLP model for cross-modality predictions, which caused a subset of the plots to change. We took some time to regenerate the figures and carefully ensure the reproducibility of the data. The tracked changes version of the manuscript shows the changes, which do not impact the paper's main message. We have also checked the paper against the Nature Methods checklist for formatting.

Reviewers' Comments:

Reviewer #1:

Remarks to the Author:

The authors made substantial additions to this revision. My concerns have been satisfactorily addressed.

A typo on page 57: The word 'morphological' in the sentence 'X_ge is the whole morphological data matrix representing all L
landmark genes measurements across all the P perturbations' appears to be 'gene expression'.

- Thanks for noting this. We have corrected it in this version.

Reviewer #2:

Remarks to the Author:

The authors have addressed all of my comments, and in several places, have gone well beyond what I expected in the thoroughness of their responses.

Reviewer #3:

Remarks to the Author:

The authors have considerably revised their manuscript and addressed, by and large, the comments raised by the reviewers. The updated content now makes the resource more easily accessible and reusable. New additions, including a prediction of gene expression across datasets and an unsupervised analysis of joint expression and imaging data, demonstrate the usefulness of the resource.

Here are a few minor improvements that could still be made:

+ Caption Fig.2: "The y-axis is trimmed at -1 for clarity". This seems to be the case in Figure 3 but not in Figure 2.

- We changed the note in the caption to: "The y-axis is trimmed at -0.5 for clarity"

+ Figure 3b lacks a legend for the size of the circles.

- In the figure caption we have: " The sizes of circles are proportional to the percentage of highly-predictable features in each category." We have also added "the actual number of features in each category/the total number of morphological features in that category" to each circle in that figure.

+ Supplementary Figure 5: "stress fiber" is truncated.

- Thanks for noting this truncation. We have updated the figure in this version.

+ Typo: "for the entire genome 7the specific genes' mRNAs"

- We added a semicolon that was missing: "for the entire genome [7]; the specific genes…'"

+ The "Editorial Policy Checklist" mentions that "Box-plot elements are defined (e.g. center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points, outliers)." This is not the case everywhere.
- We have added this information for all the boxplots of the paper.

Final Decision Letter:

Dear Shantanu,

I am pleased to inform you that your Resource, "High-Dimensional Gene Expression and Morphology Profiles of Cells across 28,000 Genetic and Chemical Perturbations", has now been accepted for publication in Nature Methods. Your paper is tentatively scheduled for publication in our December print issue, and will be published online prior to that. The received and accepted dates will be September 7, 2021 and September 28, 2022. This note is intended to let you know what to expect from us over the next month or so, and to let you know where to address any further questions.

Acceptance is conditional on the data in the manuscript not being published elsewhere, or announced in the print or electronic media, until the embargo/publication date. These restrictions are not intended to deter you from presenting your data at academic meetings and conferences, but any enquiries from the media about papers not yet scheduled for publication should be referred to us.

Once your paper is typeset, you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

Please note that Nature Methods is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including self-archiving policies. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

You will not receive your proofs until the publishing agreement has been received through our system.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

Your paper will now be copyedited to ensure that it conforms to Nature Methods style. Once proofs are generated, they will be sent to you electronically and you will be asked to send a corrected version within 24 hours. It is extremely important that you let us know now whether you will be difficult to contact over the next month. If this is the case, we ask that you send us the contact information (email, phone and fax) of someone who will be able to check the proofs and deal with any last-minute problems.

If, when you receive your proof, you cannot meet the deadline, please inform us at rjsproduction@springernature.com immediately.

Once your manuscript is typeset and you have completed the appropriate grant of rights, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately.

Once your paper has been scheduled for online publication, the Nature press office will be in touch to confirm the details.

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

Once your paper has been scheduled for online publication, the Nature press office will be in touch to confirm the details.

Content is published online weekly on Mondays and Thursdays, and the embargo is set at 16:00 London time (GMT)/11:00 am US Eastern time (EST) on the day of publication. If you need to know the exact publication date or when the news embargo will be lifted, please contact our press office after you have submitted your proof corrections. Now is the time to inform your Public Relations or Press Office about your paper, as they might be interested in promoting its publication. This will allow them time to prepare an accurate and satisfactory press release. Include your manuscript tracking number NMETH-RS46726C and the name of the journal, which they will need when they contact our office.

About one week before your paper is published online, we shall be distributing a press release to news organizations worldwide, which may include details of your work. We are happy for your institution or funding agency to prepare its own press release, but it must mention the embargo date and Nature Methods. Our Press Office will contact you closer to the time of publication, but if you or your Press Office have any inquiries in the meantime, please contact press@nature.com.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.
As soon as your article is published, you will receive an automated email with your shareable link.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

Nature Portfolio journals encourage authors to share their step-by-step experimental protocols on a protocol sharing platform of their choice. Nature Portfolio 's Protocol Exchange is a free-to-use and open resource for protocols; protocols deposited in Protocol Exchange are citable and can be linked from the published article. More details can found at www.nature.com/protocolexchange/about.

Please note that you and any of your coauthors will be able to order reprints and single copies of the issue containing your article through Nature Portfolio 's reprint website, which is located at http://www.nature.com/reprints/author-reprints.html. If there are any questions about reprints please send an email to author-reprints@nature.com and someone will assist you.

Please feel free to contact me if you have questions about any of these points.

Best regards,
Rita

Rita Strack, Ph.D.
Senior Editor
Nature Methods