

# Large Scale Cell Painting Guided Compound Selection Reveals Activity Cliffs and Functional Relationships

Corresponding Author: Dr Auguste Genovesio

**This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.**

Version 0:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

The authors published an analysis of the public JUMP-CP data set, which contains cell painting images for more than 112K compound perturbations. They acknowledged that this data set, which was produced by multiple laboratories, suffers from batch effects and investigate how to best normalize the data on a plate level to mitigate these effects. They first chose DINOv2-Giant as a performant encoder for the cell painting images. Then they tested more than 5000 different workflows to determine the best combination of normalization steps (whitening matrix fitted to DMSO controls of each plate and then applied to all sample wells of that plate, followed by an inverse normal transformation of each feature). Their subsequent analyses show convincingly that their procedure was able to remove batch effects substantially.

I think this part of the manuscript is very interesting to a broad audience, as the JUMP-CP dataset is a one-of-a-kind resource for compound imaging data and many scientists can benefit from instructions how to best normalize the data before running any kind of analyses. Hence, I recommend the publication of the manuscript mostly based on these results.

The second part of the manuscript uses the normalized data to analyse the occurrence of activity cliffs in the data set, highlighting that structurally similar compounds can elicit (phenotypically) different biological responses. However, I am not a huge fan of how this analysis was conducted. Compounds are grouped by their Bemis Murcko scaffolds and activity cliffs are searched for within these groups. However, Bemis Murcko scaffolds can cover only a small proportion of the overall atoms in a molecule and molecules that share the same BM scaffolds can be chemically rather different. Hence, I would suggest the adaption of indices such as the SALI (<https://pmc.ncbi.nlm.nih.gov/articles/PMC3448951/>) or SARI (<https://pubmed.ncbi.nlm.nih.gov/17902636/>) to identify activity cliffs. Also, the example presented in the main text shows that some molecules sharing the same BM scaffold are cytotoxic whereas others with the same scaffold aren't. I think the story would be much more convincing if a more specific phenotype was discussed. E.g., one could take a series of molecules with a known target, look for activity cliffs against this target and then show that this differential target activity also results into different cell painting images.

Finally, the authors investigate whether compounds that are acting in the same pathway are displaying similar (or opposed) phenotypes. I think it would be nice if some conclusions about the investigated pathways could be drawn, i.e. create a catalog of pathways (phenotypes) that can be well studied via U2OS cell painting versus other pathways where compounds show only weak or no similarities in cell painting.

As mentioned previously, I see a lot of merit in the data standardization procedure that is described in the manuscript, and I would also like to highlight that the authors made the extra effort to implement a web-based tool (phenoseeker) for the community that enables the comparison of a query to all molecules in the JUMP-CP dataset using their normalized cell painting profiles as descriptors. I think the paper could be even better if the activity cliff part would be reworked and clearer messages could be phrased for the analyses of various pathways and their cell painting phenotypes.

Reviewer #3

(Remarks to the Author)

see attachment

Reviewer #4

(Remarks to the Author)

The authors present a method for pre-screening compound datasets based on a similarity metric derived from cell-painting data.

While traditional similarity-based pre-selection screens are pre-dominantly based on assessing features derived from chemical structural information, the authors base their method on phenotypic profiling of human cells. To this end, the authors utilize JUMP-CP, a database of cell painting data for 112,480 compounds on the U2OS osteosarcoma cell line, comprising a number of different assays.

The principal method proposed by the authors for extracting features from the multi-channel image data is sound. For raw feature extraction the authors evaluated a number of approaches, i.e., CellProfiler and various pre-trained feature extraction deep learning models for general an microscopic image data. A key challenge of the method is feature normalization to address the batch effects.

The authors perform a comprehensive investigation of a combination of raw feature extraction and different normalization strategies to identify an optimal combination that reduces batch effects while preserving biological signals from positive controls.

Overall the manuscript is well written and the analyses performed are sound. However, the author seem to frame their method in some machine learning context as training-free and transfer learning-based, which I think is not only unnecessary but also misleading

- The authors refer to their method as "training-free, transfer learning-based". The method is definitely training-free, however transfer-learning is a highly misleading designation. The "only" transfer refers from moving away from traditional structure-based similarity by developing a novel similarity metric based on a phenotypic type similarity. To this end, existing feature extraction methods (also existing deep-learning based approaches) have been utilized. No step in the author's pipeline can be described as learning. This in itself does not disqualify the method or the manuscript but the authors should rewrite the relevant sections mainly in the abstract and introduction to avoid this misleading impression.

Section "Large-scale Cell Painting-based selection boosts active compound yield"

- A key limitation of the usage of the proposed method is that it is currently limited to the 112,480 unique compounds in the JUMP-CP data set and all reported results for validation were limited to taking compounds from (cell-based) assays from other sources, identifying the overlap with JUMP-CP compounds and performing enrichment analysis using cosine-similarity rankings of the normalized features derived from the JUMP-CP dataset. While this process can be inferred from the manuscript it is not stated in sufficient clarity in the Results section.

- In the reported enrichment factors with respect to positive controls from ChEMBL, Lit-PCBA, and Institute Curie it is unclear how many trials per assay were performed. Instead of just reporting mean and maximum enrichment, standard deviations or inter-quartile ranges of enrichment factors should be reported.

Section "Clustering analogs per phenotype reveals activity cliffs" / "Clustering analogs per phenotype reveals key chemical functions"

- The authors misspell Bemis as Demis.

- While the general analysis is sound, did the author also look for activity cliffs in the validation from the external screens. These would nicely complement this section.

- In the review copies, I found figures to be either too small, or not supplied in sufficient resolution for zooming. I would be appreciated if enlarged high resolution figures of key figures, 1b, 1d would also be included in the supplementary information to allow better visual inspection of the results.

Version 1:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

The authors have addressed my concerns. I recommend the publication of the revised article in Communications Biology.

Reviewer #3

(Remarks to the Author)

This manuscript is now good to be published, the authors have addressed all my points

Reviewer #4

(Remarks to the Author)

I am satisfied with the way the authors addressed the points by the reviewers and I recommend publication of the manuscript.

**Open Access** This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

# Authors responses to the reviewers comments

We thank the reviewers for their review, we have highlighted in red the changes in the manuscript and the 6 additional supplementary figures. Our answers are in blue here below each point.

## Reviewer #2 (Remarks to the Author):

The authors published an analysis of the public JUMP-CP data set, which contains cell painting images for more than 112K compound perturbations. They acknowledged that this data set, which was produced by multiple laboratories, suffers from batch effects and investigate how to best normalize the data on a plate level to mitigate these effects. They first chose DINOv2-Giant as a performant encoder for the cell painting images. Then they tested more than 5000 different workflows to determine the best combination of normalization steps (whitening matrix fitted to DMSO controls of each plate and then applied to all sample wells of that plate, followed by an inverse normal transformation of each feature). Their subsequent analyses show convincingly that their procedure was able to remove batch effects substantially.

I think this part of the manuscript is very interesting to a broad audience, as the JUMP-CP dataset is a one-of-a-kind resource for compound imaging data and many scientists can benefit from instructions how to best normalize the data before running any kind of analyses. Hence, I recommend the publication of the manuscript mostly based on these results.

The second part of the manuscript uses the normalized data to analyse the occurrence of activity cliffs in the data set, highlighting that structurally similar compounds can elicit (phenotypically) different biological responses. However, I am not a huge fan of how this analysis was conducted. Compounds are grouped by their Bemis Murcko scaffolds and activity cliffs are searched for within these groups. However, Bemis Murcko scaffolds can cover only a small proportion of the overall atoms in a molecule and molecules that share the same BM scaffolds can be chemically rather different. Hence, I would suggest the adaption of indices such as the SALI ([Exploring Uncharted Territories - Predicting Activity Cliffs in Structure-Activity Landscapes - PMC](#)) or SARI (<https://pubmed.ncbi.nlm.nih.gov/17902636/>) to identify activity cliffs.

We agree that Bemis–Murcko scaffolds only capture the core structure of a molecule, and that two compounds sharing the same scaffold can still be chemically very different. Our choice of scaffold-based grouping was motivated by the practical perspective of medicinal chemistry, where optimization campaigns are typically carried out around a common scaffold or chemical series.

To further validate the presence and identification of phenotypic activity cliffs, as suggested we have additionally adapted the Structure–Activity Landscape Index (SALI) proposed by Guha [[Exploring Uncharted Territories - Predicting Activity Cliffs in Structure-Activity Landscapes - PMC](#)] to phenotypic differences. This adaptation allows us to quantify the extent of phenotypic divergence between structurally similar molecules beyond the constraints of a common scaffold.

The distribution of SALI scores across the dataset, as well as illustrative examples of phenotypic activity cliffs identified with this index, are now provided in *Supplementary Figure 20*. This point is also explicitly discussed in the revised corresponding result section.

Also, the example presented in the main text shows that some molecules sharing the same BM scaffold are cytotoxic whereas others with the same scaffold aren't. I think the story would be much more convincing if a more specific phenotype was discussed. E.g., one could take a series of molecules with a known target, look for activity cliffs against this target and then show that this differential target activity also results into different cell painting images.

Note that due to the strong batch effect of the JUMP-CP dataset, many examples of activity cliff are in practice hard to see visually. The aim of this main figure was to illustrate that obvious toxicity but also more fine grain phenotype shift could arise from tiny molecular change. That said, other examples of compounds sharing the same scaffold, yet inducing different -non toxic- phenotypes are displayed in supplementary figures 8, 9 and 10. To find more specific examples, we additionally compared the list of pairs of compounds displaying activity cliffs extracted from ChEMBL by Zhang et al [<https://arxiv.org/pdf/2302.07541>] to the compounds we have in JUMP-CP dataset to find one example (out of 6 pairs of compounds present in both dataset) of specific activity change (inhibition of Carbonic anhydrase VII) clearly reflected in the phenotypic profiles. This example has been added as a remark in the corresponding part, pointing to the new supplementary figure 21.

Finally, the authors investigate whether compounds that are acting in the same pathway are displaying similar (or opposed) phenotypes. I think it would be nice if some conclusions about the investigated pathways could be drawn, i.e. create a catalog of pathways (phenotypes) that can be well studied via U2OS cell painting versus other pathways where compounds show only weak or no similarities in cell painting.

To obtain a broader view of which pathways, gene families, or genetic functions can be effectively studied using our method and the JUMP-CP dataset (U2OS cells), we refer to the following recent work by their author "*Morphological map of under- and overexpression of genes in human cells*" [<https://www.nature.com/articles/s41592-025-02753-9#additional-information>] that we have added as reference in the Discussion.

As mentioned previously, I see a lot of merit in the data standardization procedure that is described in the manuscript, and I would also like to highlight that the authors made the extra effort to implement a web-based tool (phenoseeker) for the community that enables the comparison of a query to all molecules in the JUMP-CP dataset using their normalized cell painting profiles as descriptors. I think the paper could be even better if the activity cliff part would be reworked and clearer messages could be phrased for the analyses of various pathways and their cell painting phenotypes.

### **Reviewer #3 (Remarks to the Author):**

This is a well written article demonstrating the usefulness of cell-painting as a multi-purpose tool for compound profiling and identification of bioactive compounds. The authors demonstrated nicely the importance of careful batch normalization which is a precondition for being able to

assemble a dataset of this scale. If one would use a cell-painting protocol as an upfront characterization of screening compound on screening deck entry then three questions are important:

1. How many phenotypes/target can be discovered with this type of assay: In the discussion the authors pointed also out rightly the limitation of a dataset generated on one single immortalized cell line and analyzed the phenotypic profile distribution of compounds interacting with target on cancer signaling pathways. This kind of information would be interesting to include for other targets and pathways itself.

Beyond the cancer signaling pathways analyzed in our study, we now include in the Discussion a reference to the *"Morphological map of under- and overexpression of genes in human cells"* [<https://www.nature.com/articles/s41592-025-02753-9#additional-information>], which assessed which pathways or genetic functions can be robustly captured in U2OS cells by Cell Painting.

2. How many compounds are active in some way, how much profiling data do I need to collect to able to find compounds of interest (without using a supervised structure – phenotype model)?

In the newly added Supplementary Figure 22 (inspired by *Nat. Methods* 2025), we report the fraction of compounds whose replicates can be reliably retrieved from one another. We also note that, given the high dimensionality of phenotypic profiles, some compounds may still appear close to DMSO by global similarity metrics while nevertheless inducing distinct phenotypes.

Regarding the amount of profiling data required, our method only needs the phenotype induced by a single reference compound in Cell Painting on U2OS cells. As shown in the second part of the Results, this enables any researcher to identify compounds with a desired bioactivity using only one positive control. Testing the reference compound at multiple concentrations or time points could further increase the number of hits, however this could not be evaluated here as perturbations in the JUMP-CP dataset are plated at a single concentration.

One key piece of information not explicitly mentioned in the manuscript is the fraction of random compounds without any known bioactivity (such as reported in ChEMBL) that have a non-DMSO phenotype.

As shown in *Supplementary Figure 13*, approximately 80,000 compounds in JUMP-CP have no reported activity in ChEMBL. For these compounds, we performed the same analysis as in *Supplementary Figure 22* (similarity to DMSO vs. DMSO replicates), and the results are now provided in the new *Supplementary Figure 23*. This demonstrates that a substantial fraction of compounds without known bioactivity nevertheless display non-DMSO phenotypes.

In earlier conceptional studies typically datasets enriched with known reference compounds there used, hence these were also enriched with compounds having a non-DMSO phenotypic profile, and this is not the case in this study

We are not sure to understand this point clearly and if it raises a question. Maybe what is said is that in earlier conceptual studies, datasets were typically enriched with known reference

compounds. As a result, they contained many compounds with phenotypic profiles distinct from DMSO in contrast to the JUMP-CP. If so, we partly agree with this statement but emphasize that the JUMP-CP dataset was assembled from libraries that were themselves designed to be used on existing assays in several different labs, so at the end the collection forms an active library that we agree could be considered as rather unfocused.

3. Can I control the variability of the assay over multiple experimental batches, so that I can run this as a regular profiling assay. This the authors have demonstrated.

There are some minor points:

The author missed out on citing some earlier work in the phenotypic profiling area

Reisen, F., Zhang, (7), 415–427. <https://doi.org/10.1089/adt.2015.656>

Typo: Page 9: “Demis-Murcko scaffold” should be “Bemis-Murcko scaffold”

We added this reference. And corrected the “Demis” typo.

#### **Reviewer #4 (Remarks to the Author):**

The authors present a method for pre-screening compound datasets based on a similarity metric derived from cell-painting data.

While traditional similarity-based pre-selection screens are pre-dominantly based on assessing features derived from chemical structural information, the authors base their method on phenotypic profiling of human cells. To this end, the authors utilize JUMP-CP, a database of cell painting data for 112,480 compounds on the U2OS osteosarcoma cell line, comprising a number of different assays.

The principal method proposed by the authors for extracting features from the multi-channel image data is sound. For raw feature extraction the authors evaluated a number of approaches, i.e., CellProfiler and various pre-trained feature extraction deep learning models for general an microscopic image data. A key challenge of the method is feature normalization to address the batch effects.

The authors perform a comprehensive investigation of a combination of raw feature extraction and different normalization strategies to identify an optimal combination that reduces batch effects while preserving biological signals from positive controls.

Overall the manuscript is well written and the analyses performed are sound. However, the author seem to frame their method in some machine learning context as training-free and transfer learning-based, which I think is not only unnecessary but also misleading

- The authors refer to their method as "training-free, transfer learning-based". The method is definitely training-free, however transfer-learning is a highly misleading designation. The "only" transfer refers from moving away from traditional structure-based similarity by developing a novel similarity metric based on a phenotypic type similarity.

To this end, existing feature extraction methods (also existing deep-learning based approaches) have been utilized. No step in the author's pipeline can be described as learning. This in itself

does not disqualify the method or the manuscript but the authors should rewrite the relevant sections mainly in the abstract and introduction to avoid this misleading impression.

We respectfully disagree with the reviewer's assessment regarding the use of the term *transfer learning*. The structure vs phenotype similarity is disconnected from that naming that refers to the machine learning type of method. As defined by Pan and Yang (*A Survey on Transfer Learning*, IEEE TKDE, 22(10), 1345–1359, DOI:10.1109/TKDE.2009.191) and others, transfer learning occurs when knowledge gained in one domain or task is leveraged to improve performance in a different but related domain or task. In our case, this naming reflect the fact the feature extractor (DINOv2) was pretrained on large-scale natural image datasets, and we \*transfer\* this \*learnt\* knowledge to the microscopy image domain without additional training. So we and the user do not perform training at any point (this is what was aimed at !), but there definitely was learning through the training of DINOv2. Our method is therefore both training-free (as no fine-tuning/training is required) and transfer learning-based (since it relies on features learnt from natural images and transferred to microscopy). We hope this is clearer, we carefully checked this point after this comment and don't see a more accurate way to describe the approach.

Section "Large-scale Cell Painting-based selection boosts active compound yield"

- A key limitation of the usage of the proposed method is that it is currently limited to the 112,480 unique compounds in the JUMP-CP data set and all reported results for validation were limited to taking compounds from (cell-based) assays from other sources, identifying the overlap with JUMP-CP compounds and performing enrichment analysis using cosine-similarity rankings of the normalized features derived from the JUMP-CP dataset. While this process can be inferred from the manuscript it is not stated in sufficient clarity in the Results section.

We agree with this point, we have clarified it in the corresponding Results section.

- In the reported enrichment factors with respect to positive controls from ChEMBL, Lit-PCBA, and Institute Curie it is unclear how many trials per assay were performed. Instead of just reporting mean and maximum enrichment, standard deviations or inter-quartile ranges of enrichment factors should be reported.

The number of trials corresponds to the number of potential positive controls, i.e. the bioactive hits identified in the external assays. We now report all individual enrichment factor values (not only mean and maximum) inter-quartiles values in the new supplementary figure 24.

Section "Clustering analogs per phenotype reveals activity cliffs" / "Clustering analogs per phenotype reveals key chemical functions"

- The authors misspell Bemis as Demis.

This typo has been corrected.

- While the general analysis is sound, did the author also look for activity cliffs in the validation from the external screens. These would nicely complement this section.

To explore this point, we intersected our dataset with the 21,352 activity cliff pairs reported in “*Activity Cliff Prediction: Dataset and Benchmark*” by Zhang et al. Only 6 pairs were found to be present in both datasets. As an illustrative example, we show in the new *Supplementary Figure 21* two compounds reported to induce different activities on Carbonic anhydrase VII, which also display distinct phenotypic profiles in our assay (cosine similarity =  $-0.005$ , cosine distance =  $1.005$ ).

- In the review copies, I found figures to be either too small, or not supplied in sufficient resolution for zooming. I would be appreciated if enlarged high resolution figures of key figures, 1b, 1d would also be included in the supplementary information to allow better visual inspection of the results.

High resolution figures 1b and 1d are now in supplementary figure 25