Dear Editorial Team,

We'd like to thank the editors for assessing our manuscript and accepting our initial revision plan. As a reminder, we had initially transferred this manuscript with three reviews from Review Commons. We had originally responded to several reviews, and we proposed a plan on how we would address the remaining comments.

In this document, we address all comments and present new experiments as we had originally planned. To facilitate understanding of exactly what we've changed, we divide our changes into two sections:

> Section 1: Response to review comments that were under the "proposed changes" section of our initial revision plan. Most changes are consistent with the original plan. In cases where our changes differ, we describe and justify our alternative approach.
>
> Section 2: All other review comments which we already addressed upon original submission. This section is exactly the same as what we had in the revision plan; we include it to maintain a record of all changes in response to reviewer comments.

Software updates and documentation for our revision plan changes can be found here: https://github.com/broadinstitute/cell-painting-vae/releases/tag/v1.1

Section 1: Analyses added as part of revision plan

Address UMAP interpretability to provide a deeper description of modeling performance

Reviewer 1: Instead of using UMAP embedding, it would be better to compare reconstruction error or show a reconstructed image with the original image to claim that models reliably approximate the underlying morphology data.
Reviewer 1: Rather than just stating that the VAE's did not span the original data distribution and saying beta-VAE performed best by eye, some simple metrics can be drawn to analyze the overlap in data for a more direct and quantified comparison. Researchers should also explain what part of the data is not being captured here. Some analysis of what the original uncaptured UMAP represents is important in understanding the limitations of the VAEs' capacity.
Reviewer 2: The authors compare generation performance based on UMAP. In the UMAP space, data tend to cluster together even though they might be far from each other in the feature space. I would like to see more quantitive metrics on how well these methods capture morphology distributions. You can compute metrics like MMD distance, kullback leibler (KL), earthmoving distance, or a simple classifier trained on actual MoA classes tested on generated data.

We agree with the reviewers that more closely evaluating reconstruction loss in addition to UMAP would improve understanding of VAE limitations and enable a better comparison of VAE

performance. Therefore, we analyzed reconstruction loss across models and added quantitative results to the manuscript. This will enable direct comparisons across models. From Reviewer 2's recommendation, we also evaluated performance using earthmoving distance. We add the following text and a new table (Table 1) to the results section:

> Based on reconstruction loss (MSE) and earth moving distance, MMD VAEs performed the best in Cell Painting data, and all architectures performed substantially better in most cases compared to randomly shuffled baselines (Table 1).

Dataset	VAE	MSE	MSE (Shuffled)	Earthmoving	Earthmoving (Shuffled)
Cell Painting level 5	Vanilla	0.00387	0.0088	0.016 (0.006, 0.040)	0.024 (0.0096, 0.060)
Cell Painting level 5	Beta	0.00272	0.0088	0.012 (0.005, 0.028)	0.024 (0.01, 0.06)
Cell Painting level 5	MMD	0.00435	0.0088	0.016 (0.005, 0.04)	0.024 (0.01, 0.06)
Cell Painting level 4	Vanilla	0.00145	0.0014	0.006 (0.004, 0.009)	0.0053 (0.004, 0.008)
Cell Painting level 4	Beta	0.00091	0.0014	0.0045 (0.002, 0.013)	0.0051 (0.004, 0.008)
Cell Painting level 4	MMD	0.00075	0.0014	0.0039 (0.0022, 0.01)	0.0051 (0.004, 0.008)
L1000	Vanilla	0.85	1.85	0.249 (0.14, 0.36)	0.61 (0.28, 1.94)
L1000	Beta	1.23	2.10	0.445 (0.23, 0.73)	0.67 (0.29, 2.25)
L1000	MMD	1.27	2.05	0.475 (0.25, 0.79)	0.64 (0.29, 2.027)

Table 1. Mean squared error (MSE) and earthmoving distance for VAE's ability to reconstruct Cell Painting and L1000 profiles. We compare these values with results derived from shuffled models. Earthmoving distance is calculated by taking the mean of the earthmoving distance of each sample. We add the 95% percentile range of earthmoving distance in parenthesis (0.05 lowest, 0.95 highest). Note that since our models required that we normalize Cell Painting and L1000 input data differently (see Methods), the metrics cannot be compared across data modalities.

The reviewer comments also bring up an important detail. We trained our VAEs using CellProfiler readouts from Cell Painting images and not the raw Cell Painting images themselves. As this was one of our primary innovations, this detail is extremely important. Therefore, we have improved clarity and added emphasis to this point in the manuscript introduction and discussion.

> Because of the success of VAEs on these various datasets, we sought to determine if VAEs could also be trained using cell morphology readouts (rather than directly on images), and further, to carry out arithmetic to predict novel treatment outcomes. We derive the cell morphology readouts using CellProfiler (McQuin et al, 2018), which measures the size, structure, texture, and intensity of cells, and use these readouts to train all models.

We determined that VAEs can be trained on cell morphology readouts rather than directly using the cell images from which they were derived. This decision comes with various trade-offs. Compared to cell images, cell morphology readouts as extracted by image analysis tools (e.g. CellProfiler) are a more manageable data type; the data are smaller, easier to distribute, substantially less expensive to analyze and store, and faster to train (McQuin et al, 2018).

In our revision plan, we planned to compare UMAP with another dimensionality reduction algorithm called PaCMAP. However, we learned that this was not feasible because PaCMAP is incompatible with our approach. The method does not have a “transform” function, and thus cannot be used to directly compare simulated and real datasets (https://github.com/YingfanWang/PaCMAP/issues/8). While this is unfortunate, we believe comparing reconstruction using MSE and earthmoving distance is sufficient to quantitatively understand the modeling performance of the different VAE variants.

More specific comparisons of MOA predictions to shuffled data and improved description of MOA label accuracy

Reviewer 1: It is difficult to know the clear threshold for successful performance is on figures like Figure 7 and SFigure 9, but by and large, it appears that the majority of predicted combination MOAs were not successful. Without the ability to either A) adequately predict most all combinations from individual profiles that were used in training or B) an explanation prior to analysis of which combination will be able to predict, it is difficult to see this method being used since the combinatorial predictions are more likely not informative.
Reviewer 1: The researchers justify the poor performance compared to shuffled data, by saying that A) MOA annotations are noisy and unreliable and B) they MOAs may only manifest in other modalities like what was seen in the L1000 vs morphology predictability. While these might be true, knowing this the researchers should make an effort to clean and de-noise their data and select MOAs that are well-known and reliable, as well as, selecting MOAs for which we have a known morphological or genetic reaction.
Reviewer 3: Figure 6 is missing error bars (standard deviation of the L2 distance) and, as such, is hard to draw conclusions from.

We thank the reviewers for raising this concern. We agree that it is critical, and we appreciate the opportunity to address it. All three of these comments relate to being unable to draw conclusions from our LSA results when most A∩B predictions appear to have no difference from shuffled controls. It is important to note that because of these comments, we realized a methodological fallacy in our results interpretation. To assess individual MOA performance, we should be comparing specific MOA combinations to their corresponding shuffled results instead of comparing “all to all”, which might artificially decrease performance when there are polypharmacology predictions that fail to measure the ground truth cell states.

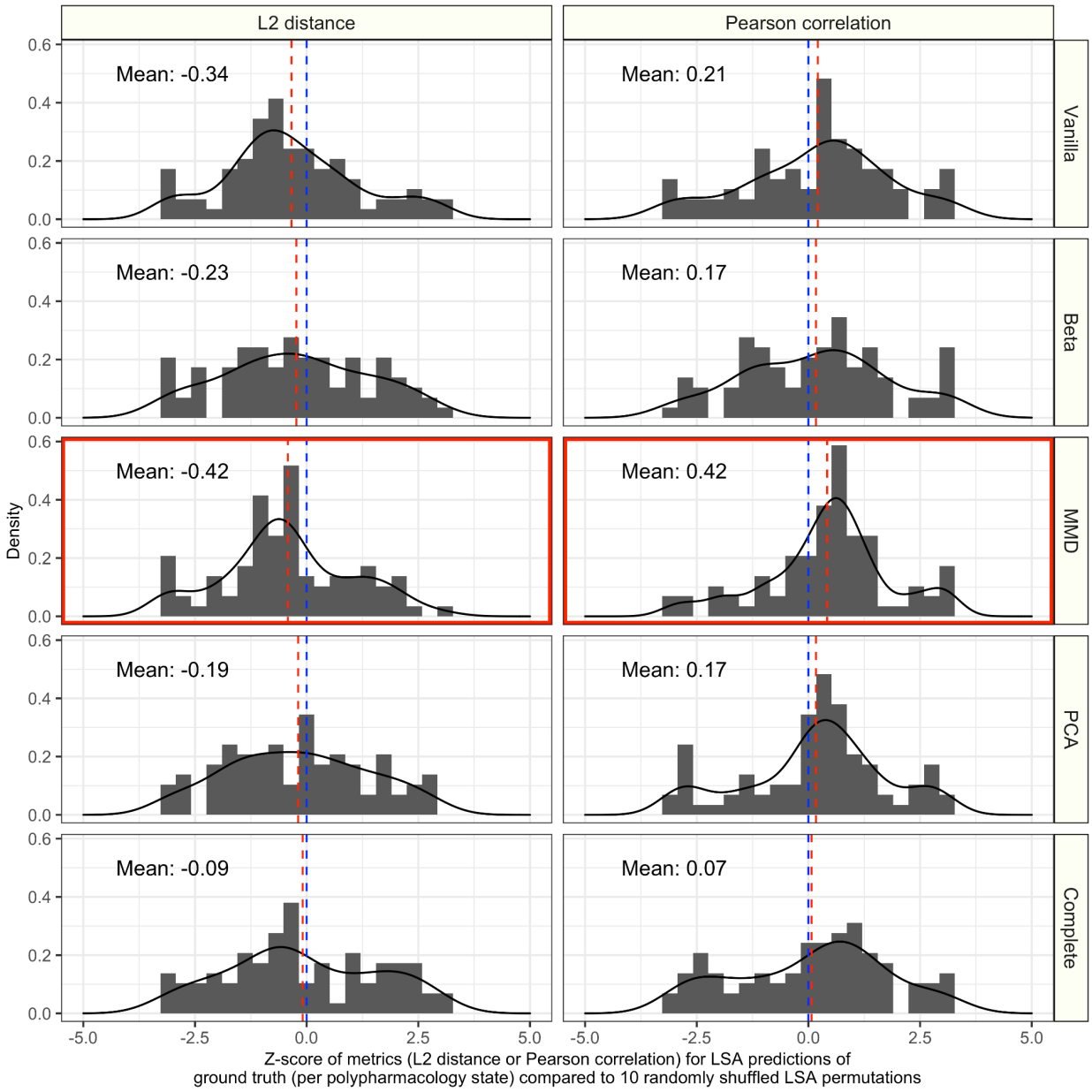
Therefore, we performed an additional analysis comparing LSA for each individual polypharmacology MOA A∩B. In our methods section, we describe this analysis as follows:

> To determine the MOAs our VAE could predict the best, we calculated a z score metric for each MOA combination. Specifically, for a given polypharmacology MOA "A ∩ B", we compared two values: 1) the L2 distance between the predicted cell state and the ground-truth cell state and 2) the distribution of L2 distances between ten cell state predictions from randomly shuffled negative controls and the ground-truth cell state. We calculated the z-score of the ground truth value compared to the randomly shuffled distribution. As well, we repeated this procedure using Pearson correlation instead of L2 distance. This evaluation measures performance for each polypharmacology MOA "A ∩ B" independently, and describes how much better LSA with real data could predict cell states compared to random.

We also add a brief description in the results section:

> Using Cell Painting level 5 VAE models, we compared LSA performance of specific polypharmacology MOAs. For each polypharmacology MOA "A ∩ B", we calculated a z score comparing 1) the L2 distance between the real and the predicted "A ∩ B" cell state against 2) a distribution of L2 distances between the real and ten predictions of "A ∩ B" cell states from randomly permuted input data (see Methods). We repeated this procedure using Pearson correlations as well. These metrics indicated that, for the majority of MOAs, we predicted polypharmacology states better than random (Supplementary Figure 11). High test statistics for L2 distance and a low test statistics for Pearson correlation indicates that the specific MOA "A ∩ B" could not be predicted, either because of incorrect annotations, non-additive or synergistic treatment effects, or a low penetrant phenotype unable to be captured in Cell Painting data.

We created a supplementary figure to present these new results, which increases confidence in the LSA experiment's ability to predict MOAs.



Supplementary Figure 11. Distribution of LSA z-scores from comparing ground-truth MOA cell state to 10 randomly shuffled LSA permutations for Cell Painting level 5 data. The blue line is centered at 0. All MOAs to the left of the blue line for the L2 distance graph are predicted better than random, and all MOAs to the right of the blue line in the Pearson correlation graph are predicted better than random. The red line indicates the mean of all the z-scores, so a lower mean for the L2 distance is better, and higher mean for Pearson correlation is better.

Additionally, we modified the manuscript to better describe potential reasons for poor LSA performance. In correspondence with Paul Clemons, the senior director Director of Computational Chemical Biology Research at the Broad Institute of MIT and Harvard, we learned that the Drug Repurposing Hub annotations are among the most well documented. Therefore, while we know that biological annotations are often incomplete, our original text

overemphasized the amount of noise contributed by inaccurate labels. We therefore added the following sentence to the discussion to clarify this important point:

> However, the Drug Repurposing Hub MOA annotations are among the most well-documented resources, so other factors like different dose concentration and non-additive effects may also contribute to weak LSA performance for some compound combinations (Corsello et al. 2017).

We also recognize that the LSA performance compared to random controls suggests that it would work poorly in many use-cases. Therefore, in the Discussion section, we emphasize the novelty of our findings, and add the following statement about the need to develop more improved tools for predicting polypharmacology cell states:

> Therefore, while VAEs model interpretable latent spaces in Cell Painting data and LSA can serve as a baseline for predicting polypharmacology, other datasets, which directly collect data on an expanded set of polypharmacology cell states, will enable training of more accurate predictive models (Caldera et al, 2019).

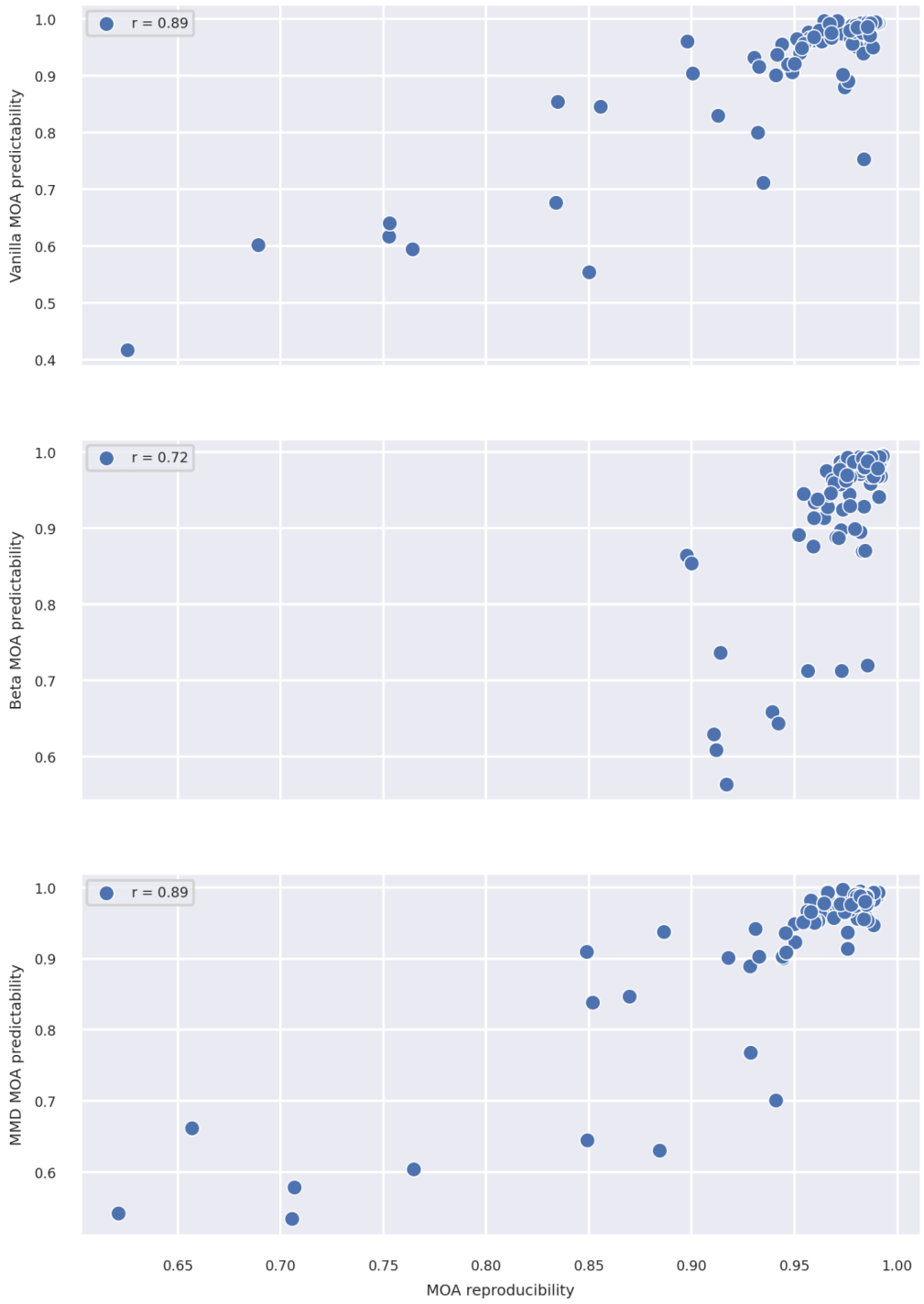
More detailed evaluation of MOA performance across drug variance and drug classes

Reviewer 1: With the small number of combinations that are successfully predicted, to build confidence in the performance, it would be necessary to explain the reason for the differences in performance. Further experimentations should be done looking into any relationship between the type of MOAs (and their features) and the resulting A|B predictability. Looking at Figure 7, the top-performing combinations are comprised entirely of inhibitor MOAs. If the noisiness of the data is a factor, there should be some measurable correlation between feature noisiness and variation and the resulting A|B predictability from LSA.

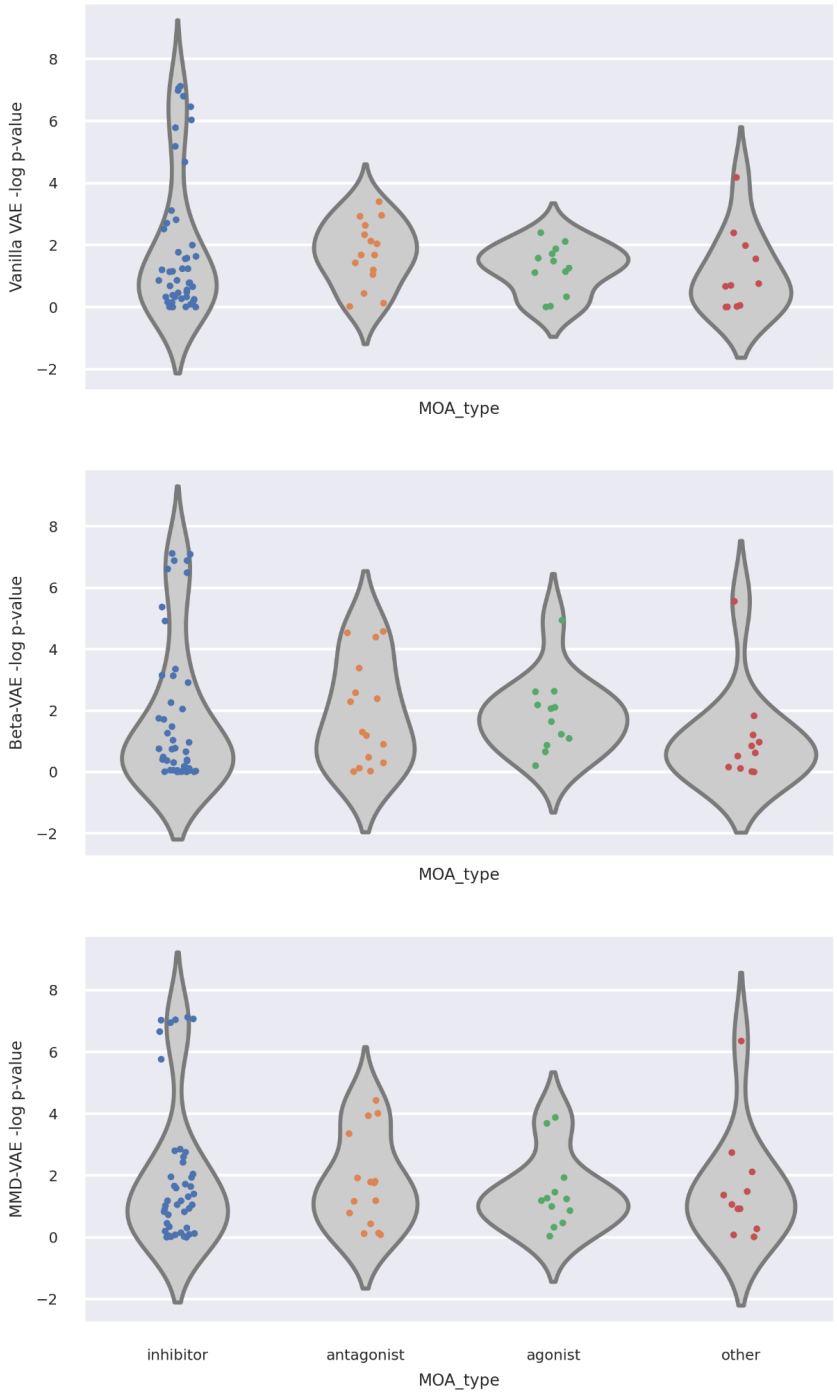
We agree with the reviewer that further experimentation would gain confidence and increase understanding of LSA performance. We have performed two different analyses to address this question and add two new supplementary figures. We introduce both in the results section:

> Finally, we performed two analyses to more specifically understand LSA performance. First, we assessed correlation between MOA reproducibility and MOA predictability. We calculated MOA reproducibility by comparing median pairwise correlation between original and reconstructed “A∩B” MOAs, and we calculated MOA predictability as the correlation between real and predicted “A∩B” MOAs from the LSA experiment. We observed a strong correlation (Pearson r = 0.8) between MOA reproducibility and MOA predictability (Supplementary Figure 13). This was expected; MOAs that we more reproducibly captured are more reliably modeled and predicted. Second, we sought to see if there were any LSA performance differences depending on the type of MOA (inhibitor, antagonist, agonist, and other). While the top performing MOAs were all

inhibitors, we did not observe significant differences across MOA types (Supplementary Figure 14).



Supplementary Figure 13. Strong correlation between MOA reproducibility (median pairwise correlation among real and reconstructed MOAs) and MOA predictability (correlation between real and predicted MOA from LSA experiment).



Supplementary Figure 14. *No significant association between MOA compound class and the predictability of MOAs.*

Higher confidence in LSA overfitting assessment

Reviewer 1: To show that the methodology works well on unseen data, researchers withheld the top 5 performing A|B MOAs (SFig 9) and showed they were still well predicted. This is not the most compelling demonstration since the data to be held out was selected with bias as the top-performing samples. It would be much more interesting to withhold an MOA that was near or only somewhat above the margin of acceptability and see how many holdouts affected the predictability of those more susceptible data points. From my best interpretation, the hold-out experiment also only held out the combination MOA groups from training. It would be better if single MOAs (for example A) which were a part of a combination of MOA (A|B) were also held out to see if predictability suffered as a result and if generalizability did extend to cells with unseen MOAs (not just cells which had already highly performing combinations of seen MOAs).

We believe our original analysis was compelling. We did remove all single MOAs from training, in addition removing compounds belonging to the polypharmacology combination. We now make this point more clear in the results and methods section:

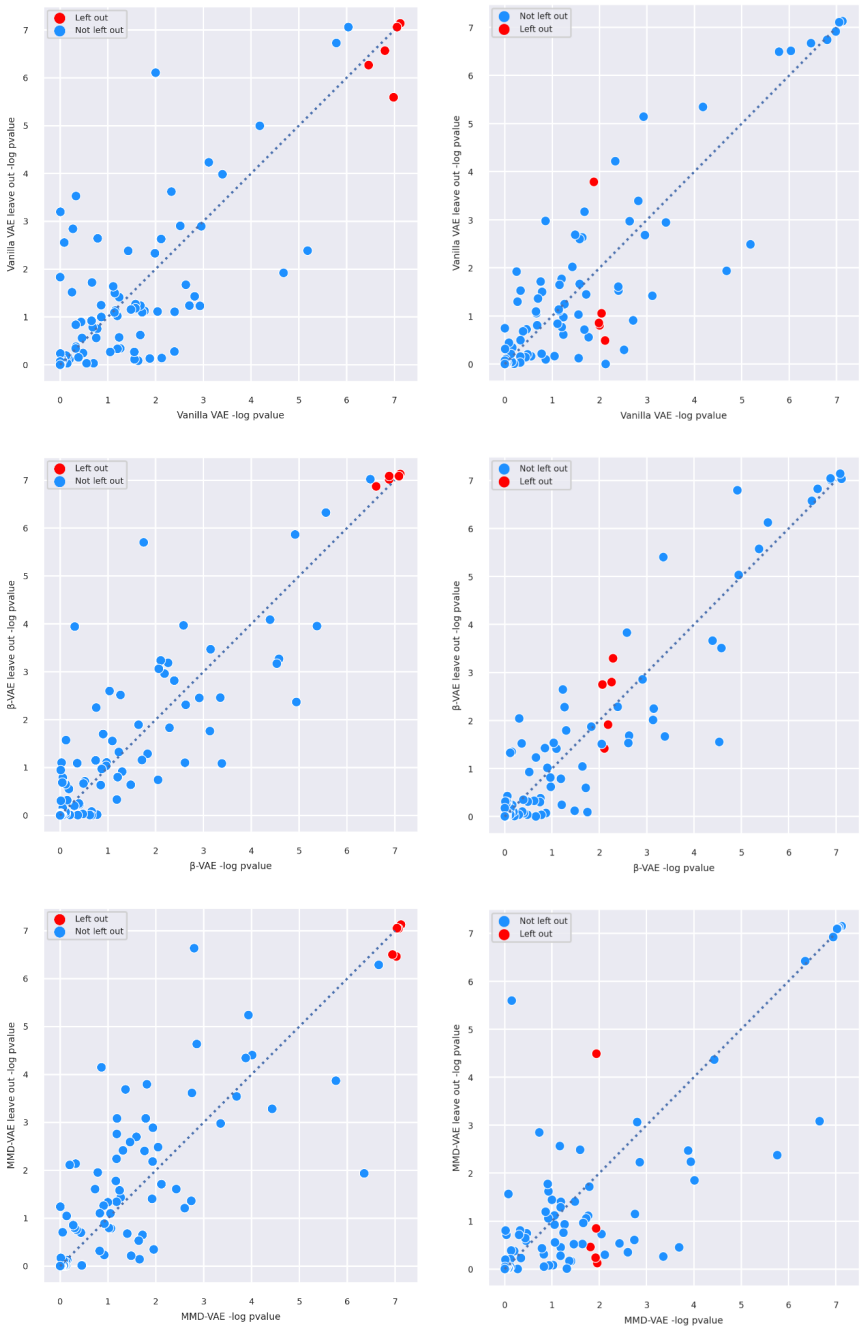
> For our work with predicting MOAs to have practical application, we needed to know that we can predict MOAs that have never been seen by the training set. So, we removed compounds annotated to the top five best-predicted MOAs, as well as the single MOAs that were part of the MOA combinations (i.e. also removing compounds annotated with MOA “A” or MOA “B” if MOA “A∩B” was in the top 5 best predicted) for each LSA evaluation.

This analysis tells us that even if we removed compounds belonging to the top MOAs *from training*, we were still able to capture their combination polypharmacology cell states through LSA. Our model is learning some fundamental data generating function that our top performing MOAs tap into regardless of if they are present or not in training. We find this similar to removing all pictures of sunglasses in an image corpus of human faces, but still being able to reliably simulate pictures of people wearing sunglasses.

However, we agree with the reviewer that withholding intermediate-performing MOAs would also be informative. Unlike the best predicted MOAs, the intermediate MOAs are likely more susceptible to changes in training data composition, so it would be interesting to determine if intermediate MOAs’ performance is a result of overfitting instead of truly learning aspects of the data generating function. We performed this new analysis and have added the results to Supplementary Figure 12 as a subpanel and added a full description of the approach and results in the results section.

> We also retrained all models after removing compounds with MOA annotations that were predicted with intermediate performance. Specifically, we removed the five MOAs that were predicted with intermediate scores of above one standard deviation better than random. In general, performance dropped more for these intermediate data points than

the high performers, but overall performance per MOA remained consistent particularly in our β-VAE (Supplementary Figure 12B). Despite the top MOAs retaining signal, lower performing MOAs experienced some signal loss, which suggests that intermediate performing MOAs are more susceptible to changes in training data composition and require more data or different approaches to model.



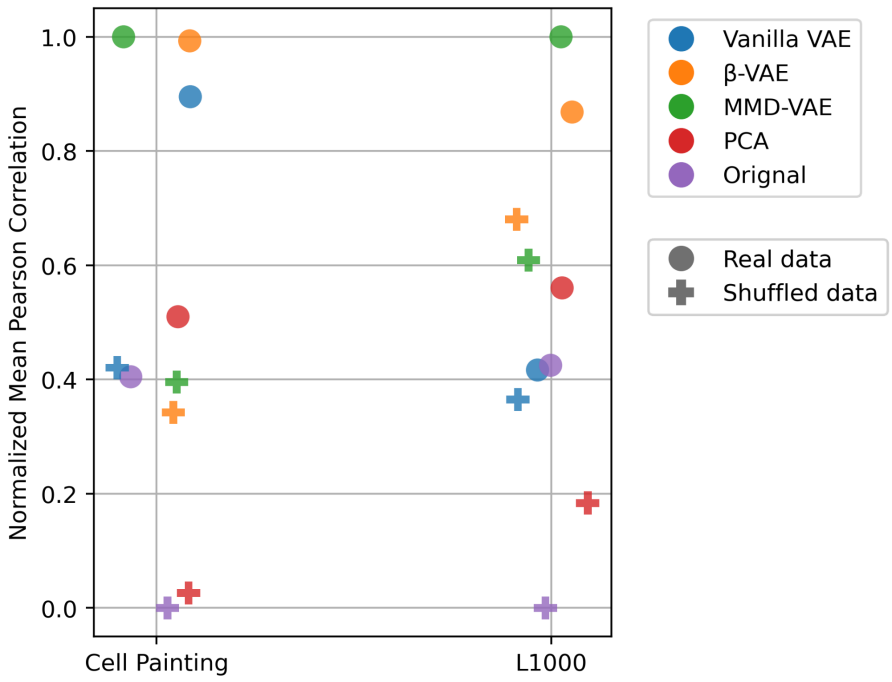
Supplementary Figure 12. Leave-five-out MOA experiments demonstrate robust learning in the absence of certain compounds.

Additional metrics to evaluate LSA predictions to provide more confident interpretation

Reviewer 2: The predictions are evaluated using L2 distances, which I find not that informative. I would like to see other metrics (correlation or L1 or distribution distances in previous comments)

We agree with the reviewer that using more than one metric would be helpful because oftentimes a single metric does not tell a complete story. In response, we added Supplementary Figure 9, which visualizes performance across models using Pearson correlation. While L2 distances will tell us how linearly far apart our predictions are to ground truth, Pearson correlations will tell us how close our predictions trend in the same direction as ground truth. We reference this finding in the results section as follows:

> We observed similar results when measuring MOA similarity using Pearson correlation instead (Supplementary Figure 9).



Supplementary Figure 9. Mean Pearson correlation (higher is better) between real and predicted profiles annotated with known polypharmacology ("A ∩ B") mechanisms of action (MOAs) for three different VAE architectures, PCA, and original input space. We used level 4 Cell Painting input data for LSA predictions.

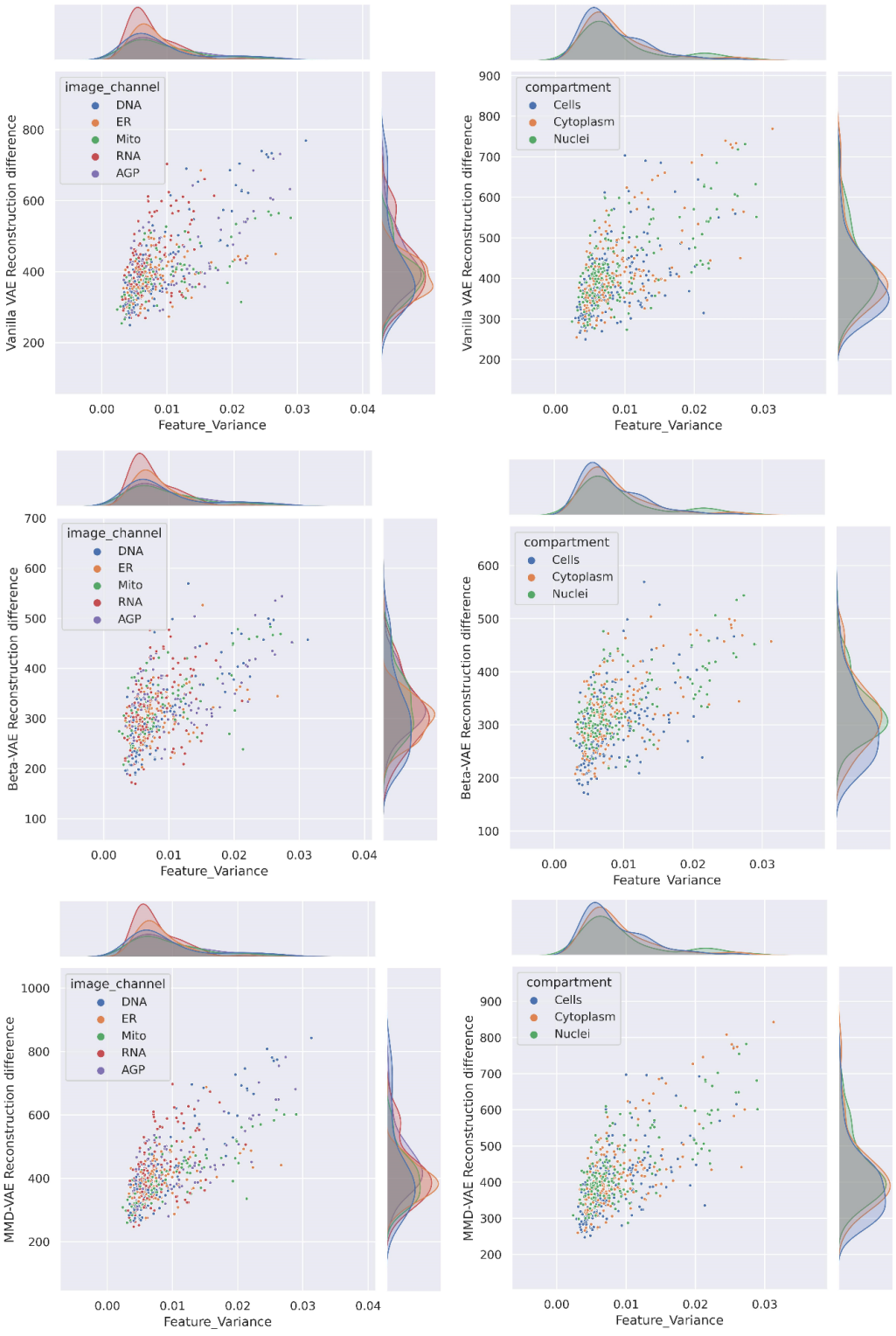
Adding a performance-driven feature level analysis to categorize per-feature modeling ability

Reviewer 2: I would like to see feature-level analysis, which features are well predicted and which ones are more challenging to predict?

We agree with the reviewer that feature level analysis would be interesting to study. We believe that understanding the modeling difficulty of different features could give insight into the nature of Cell Painting features and why certain MOAs are predicted better than others.

However, any direct analysis of feature performance would be impacted by feature variance, because features that have less variation are likely to be inherently easier to model. Therefore, we compared individual feature reconstruction loss vs. feature variance across profiles and observed strong positive associations. We compared different feature subcategories, including compartments (cells, cytoplasm, nuclei) or channels (DNA, ER, Mito, RNA, AGP), and identified significant differences across a few subcategories.This analysis not only demonstrates the relationship between feature variance and modeling ability, but also provides insight into the difficulty of modeling certain categories of CellProfiler features. We added the following description and supplementary figure:

> We also analyzed our VAE's ability to reconstruct specific CellProfiler features. As expected, we observed that features with lower variance were more easily reconstructed (Supplementary Figure 6). For all VAE variants, we found a high performance diversity, with many features reconstructing nearly perfectly, while others had poor reconstruction. The DNA channel was among the best reconstructed, while the AGP image channel was among the worst reconstructed, although this relationship was not significant (DNA < AGP with one-sample t-test p-value 0.37, 0.056, 0.20 for vanilla, β, and MMD-VAEs, respectively). The cells compartment was significantly better reconstructed compared to cytoplasm and nuclei for all VAE variants (Cells < non-Cells with one-sample t-test p-value 2.7e-4, 5.1e-7, 1.8e-5 for vanilla, β, and MMD-VAEs, respectively).



Supplementary Figure 6. Comparing feature variance with feature reconstruction for Cell Painting level 4 input data across three VAE variants. We stratified feature categories by image channel and compartment.

SECTION 2: REVIEW COMMENTS ALREADY ADDRESSED IN OUR REVISION PLAN

Documenting positive feedback as provided by the three reviewers

Reviewer 1: With access to the dataset, the posted GitHub, and documentation in the paper, I believe that the experiments are reproducible.
Reviewer 1: The experiments are adequately replicated statistically for conventions of deep learning.
Reviewer 1: This paper proposes a conceptually and technically unique proposal in terms of application, taking existing technologies of VAEs and LSA and, and as far as I know, uses them in a novel area of application (predicting and simulating combination MOAs for compound treatments). If this work is shown to work more broadly and effectively, is seen through to it completion, and is eventually successfully implemented, it will help to evaluate the effects of drugs used in combination on gene expression and cell morphology. An audience in the realm of biological deep learning applications as well as an audience working in the compound and drug testing would be interested in the results of this paper. Authors successfully place their work within the context of existing literature, referencing the numerous VAE applications that they build off of and fit into the field of (Lafarge et al, 2018; Ternes et al, 2021, etc...), citing the applications of LSA in the computer vision community (Radford et al, 2015, Goldsborough et al, 2017), and discussing the biological context that they are working in (Chandrasekaran et al, 2021).
Reviewer 2: The main novelty of the work is applying VAEs on cell painting data to predict drug perturbations. The final use case could be guiding experimental design by predicting unseen data. However, the authors do not show such an example and use case which is understandable due to the need for doing further experiments to validate computational results and maybe not the main focus of this paper. The authors did a good job of citing existing methods and relevant. The potential audience could be the computational biology and applied machine learning community.
Reviewer 3: The manuscript is beautifully written in a crystal clear manner. The authors have made a visible effort towards making their work understandable. The methods section is clear and comprehensive. All experiments are rigorously conducted and the validation procedures are sound. The conclusions of the paper are convincing and most of them are well supported by the data. Both the data and the code required to reproduce this work are freely available. Overall, the article is of high quality and relevance to several scientific communities.

We thank the reviewers for their encouraging remarks and overall positive sentiment. As early-career researchers, we feel empowered by these words.

Moved Figure 2 to supplement and removed Figure 5

Reviewer 1: Fig 2 is not informative so it can go to supplementary.

Reviewer 2: I liked the paper's GitHub repo, the authors did a good job making everything available and reproducible. As a suggestion, you can move the learning curves in two the sup figures cause they might not be the most exciting piece of info for the non-technical reader.
Reviewer 3: I would suggest removing Figure 5 (or moving it to the supplementary) as it revisits the content of Figure 1 and does not bring much extra information.

We agree that Figure 2 might not be informative to a non-technical reader, so we have accepted this suggestion by both reviewers 1 and 2, and we have moved Figure 2 to supplementary.

We agree with the reviewer and have removed Figure 5.

Clarified our data source as CellProfiler readouts, not raw Cell Painting images

Reviewer 1: In Fig 4, it would be useful to show a few sample representative images with respect to CellProfiler feature groups.
Reviewer 1: Figure 6, what does it means original input space? Does it mean raw pixel image? As researchers extracted CellProfiler feature groups already, it would be interesting to compare mean L2 distance based on CellProfiler features so that whether VAE improves performance or not (compared to handcrafted features) as a baseline.
Reviewer 3: While what "morphological readouts" concretely mean becomes clearer later on in the paper, it would be useful to give a couple of examples early on when introducing the considered datasets.

We thank the reviewer for these suggestions, which bring to light a common source of confusion, which we must alleviate. We are working with CellProfiler readouts (features extracted using classical algorithms) of the Cell Painting images and not the images themselves. We have made several edits throughout the manuscript to improve clarity and remove this confusion, including the introduction, in which we clearly state our model input data:

> “Because of the success of VAEs on these various datasets, we sought to determine if VAEs could also be trained using cell morphology readouts (rather than directly on images), and further, to carry out arithmetic to predict novel treatment outcomes. We derive the cell morphology readouts using CellProfiler (McQuin et al. 2018), which measures the size, structure, texture, and intensity of cells, and use these readouts to train all models.”

This decision comes with tradeoffs: The benefit of using CellProfiler readouts instead of images is that they are more manageable but we might lose some information. We more thoroughly discuss this important tradeoff in the discussion section:

> “We determined that VAEs can be trained on cell morphology readouts rather than directly using the cell images from which they were derived. This decision comes with various trade-offs. Compared to cell images, cell morphology readouts as extracted by

> image analysis tools (e.g. CellProfiler) are a more manageable data type; the data are smaller, easier to distribute, substantially less expensive to analyze and store, and faster to train (McQuin et al. 2018). However, it is likely some biological information is lost, because these tools might fail to measure all morphology signals. The so-called image-based profiling pipeline also loses information, by nature of aggregating inherently single-cell data to bulk consensus signatures (Caicedo et al. 2017)."

Clarified future directions to infer cell health readouts from simulated polypharmacology cell states

Reviewer 1: Authors also make the claim that they can infer toxicity and simulate the mechanism of how two compounds might react. This is a claim that would not be supported even if the method were able to successfully predict morphology or gene profiles. Drug interaction and toxicity are quite complex and goes beyond just morphology and expression. VAEs predicting a small set of features would not be able to capture information beyond the readouts, especially when dealing with potentially unseen compounds for which toxicity is not yet known. For example, two compounds might produce a morphology that appears similar to other safe compounds but has other factors that contribute to toxicity. Further, here they show no evidence of toxicity or interaction analysis.

The reviewer is correct that such a claim is unsupported by our research. Our message was actually that inferring toxicity could be a potential future application of our work. Specifically, for example, we can apply orthogonal models of cell toxicity that we previously derived using other data (Way, Kost-Alimova, et al. 2021) to our inferred polypharmacology cell states. We thank this reviewer for noticing our lack of clarity, and we have made changes in the discussion to make it clear that inferring toxicity is something we may do in the future and is not something that is discussed in the manuscript:

> "In the future, by predicting cell states of inferred polypharmacology, we can also infer toxicity using orthogonal models (e.g. Way et al. 2021) and simulate the mechanisms of how two compounds might interact."

Clarified our method of splitting data, and noting how a future analysis will answer overfitting extent

Reviewer 2: Could authors outline detailed data splits? Which MoA are in train and which are held out from training? As I understood, there were samples from MoAs that were supposed to be predicted in the calculation of LSA? Generally, the predicted MoA should not be seen during training and not in LSA calculation.

We now more explicitly detail how we split our data in the methods:

> “As input into our machine learning models, we split the data into an 80% training, 10% validation, and 10% test set, stratified by plate for Cell Painting and stratified by cell line for L1000. In effect, this procedure evenly distributes compounds and MOAs across data splits.”

We also thank the reviewer for this comment, because they express an important concern about making sure that we are not overfitting to the data. We have explained in the manuscript that because of lack of data, MOAs were repeated in training and LSA. However, we believe overfitting is not playing a large role in model performance. Through our hold 5 out experiment, we are able to show that our models are able to predict the same MOAs irrespective of whether they were in the training data, indicating that we did not overfit to the distribution of certain MOAs.

Reviewer 1 also suggested that we do the hold 5 out experiment on A∩Bs that were barely predicted. After we do that, we will explicitly demonstrate the extent of overfitting.

Introduced acronyms when they first appear in the manuscript

Reviewer 3: The Kullback-Leibler divergence is properly introduced in the methods part, but not at all in the introduction (it directly appears as "the KL divergence"). To enhance readability, it would be better to fully spell it before using the acronym, and maybe give a one-sentence intuition of what it is about before pointing out to the methods part for more details.

We thank the reviewer for bringing this to our attention. We have carefully reviewed the entire manuscript and have corrected such instances of clear introductions to acronyms.

Fixed minor text changes

Reviewer 3: In Figure 1, I would recommend changing "compression algorithms" to "dimension reduction algorithm" or "embedding algorithm". In a compression setting, I would expect the focus to be on the number of bits of information each method requires (or the dimension of the resulting embedding) to encode the data while guaranteeing a certain quality threshold. This is obviously not the case here as the dimension of the embedding is fixed and the focus is on exploring how the embedding is constructed (eg how much it decorrelates the different features, etc) - which may be misleading.
Reviewer 3: I recommend using "A n B" or "A & B" or "(A, B)" to denote the combination of two independent modes of action A and B. The current notation "A | B" overloads the statistical "A given B" which appears in the VAE loss and is therefore misleading.

We agree with the reviewer, and aim to minimize all sources of potential confusion. We have made the change in the figure.

We also agree that our current notation can be confusing. We have updated all instances of "A|B" with "A ∩ B".

Added hypothesis of MMD-VAE oscillations to supplementary figure legend

Reviewer 3: Do the authors have a hypothesis of what may be causing MMD-VAE to oscillate during validation when data are shuffled? This seems to be the case on two of the three considered datasets (Figure 2 and SuppFigure 1) and is not observed for the other models. Including a few sentences on that in the text would be interesting.

We believe a big reason for this is because of the fact that the optimal MMD-VAE had a much higher regularization term, which puts a greater emphasis on forming normal latent distributions, than the optimal Beta or Vanilla VAE. Forcing the VAE to encode a shuffled distribution into a normally distributed latent distribution would be difficult to do consistently across different randomly shuffled data subsets, and therefore might cause oscillations in the training curve across epochs when the penalty for that term is high. As these observations may be interesting to a certain population of readers, we have incorporated this explanation into the supplementary figure legend (which is where this figure is shown):

> "Forcing the VAE to consistently encode a shuffled distribution into a normally distributed latent distribution would be difficult, and therefore might cause oscillations in the training curve across epochs."

Explained our selection of VAE variants

Reviewer 3: The different types of considered VAE and their differences are very clearly introduced. It may however be good to motivate a bit more the focus on beta-VAE and MMD-VAE among all the possible VAE models. This is partly done through examples in the second paragraph of page 2, but could be elaborated further.

We thank the author for their encouraging remarks. We have made edits to the manuscript's introduction, explaining why we chose these two variants out of all the possible choices:

> "We trained vanilla-VAEs, β-VAEs, and MMD-VAEs only, and not other VAE variants and other generative model architectures, such as generative adversarial networks (GANs), because these VAE variants are known to facilitate latent space interpretability."

Description of analyses that authors prefer not to carry out

We will not explore additional latent space dimensions in more detail, as this is out of scope

Reviewer 1: As both reconstructed and simulated data did not span the full original data distribution, it might be better to look at reconstruction error and increase the dimension of latent space.

We thank the reviewer for bringing up this important point. Our VAE loss function consists of the sum of reconstruction error and some form of KL divergence. Specifically, this reviewer is suggesting that if we only minimize reconstruction error (or focus more on reconstruction over KLD by lowering beta), a higher latent dimension would result in better overall reconstruction. This is true, but doing so would have negative consequences. While we would perhaps get the UMAPs to show the full data distribution, the UMAPs are not our focus; predicting polypharmacology through LSA is. We found that when we have a higher focus on the reconstruction term, we have more feature entanglement, as indicated by lower performance when simulating data and overlapping feature contribution per latent feature. The fact that simulating data would logically require less disentanglement than performing LSA shows that we require higher regularization (and hence lower focus on reconstruction) than the one we got from simulating data.

Essentially, while the reviewer's comments would improve reconstruction and allow us to improve the UMAPs, doing so would likely worsen LSA performance, which is the main focus of the project. Also, increasing the latent dimension without changing beta would likely have caused little to no change because since beta is encouraging disentanglement, it would cause the newly added dimension to have little variation and encode little new information that wasn't already encoded before.

We have also previously explored the concept of toggling the latent dimensions in a separate project (Way et al. 2020). We are very interested in this area of research in general, and any additional analyses (beyond hyperparameter optimization) deserves a much deeper dive than what we can provide in this paper.

Lastly, we intend to include a deeper description and analysis of reconstruction loss across models, datasets, and MOAs as was suggested by a previous reviewer comment (see section 2.1 above)

We will not review Gaussian distribution assumptions of the VAE as we feel it is not informative

Reviewer 1: By looking at SFigure 6, I am wondering whether latent distribution actually met gaussian distribution (assumption of VAE). It may show skew distribution as some of latent features shows low contribution.

This reviewer's comment is interesting, but we do not believe it would change the findings of our study. Suppose we find that the latent dimensions aren't normally distributed. This wouldn't change much; a gaussian distribution isn't the most critical to perform LSA. We need the latent code to be disentangled, but having normally distributed latent features doesn't necessarily mean that we have good disentanglement (see https://towardsdatascience.com/what-a-disentangled-net-we-weave-representation-learning-in-vaes-pt-1-9e5dbc205bd1)

In this paper, we will not train or compare conditional VAEs nor cycle GANs

Reviewer 2: While authors provided a comparison between vanilla VAE and MMD-VAE, B-VEA, there are other methods capable of doing similar tasks (data simulation, counterfactual predictions), I would like to see a comparison with those methods such as conditional VAE(https://papers.nips.cc/paper/2015/hash/8d55a249e6baa5c06772297520da2051-Abstract.html, CVAE + MMD : https://academic.oup.com/bioinformatics/article/36/Supplement_2/i610/6055927?login=true) or cycle GANs(https://arxiv.org/abs/1703.10593).

While such comparisons would be interesting, they are not the main focus of the manuscript, which is to benchmark the use of VAEs in cell morphology readouts and to predict polypharmacology.

We think that CVAE would not be appropriate for our study. In a CVAE, the encoder and decoder are both conditioned to some variable. In our situation where we are predicting the cell states of different MOAs, it would make most sense to condition on the MOA. However, because we're using the MOA labels in our LSA experiment, conditioning on them is likely to bias our results and not be effective for MOAs outside the conditioning.

For cycle GANs, we have found that training using these data, in a separate study in our lab, is extremely difficult. Our lab has not published this yet, but once we are able to better understand cycleGAN behavior in these data, it will require a separate paper in which we compare performance and dissect model properties in much greater detail.

Nevertheless, we have added citations to multi-modal approaches like cycle GANs (see section 4.4) as they will point a reader to useful resources for future directions.

We will not be comparing with multi-modal integration, but we clarified our focus on Cell Painting VAE novelty and added multi-modal citations

Reviewer 1: Researchers found that the optimal VAE architectures were very different between morphology and gene expression, suggesting that the lessons learned training gene expression VAEs might not necessarily translate to morphology. It would be interesting to compare the result with multimodal integration as baseline (i.e., Seurat).

Our focus in this paper was to train and benchmark different variational autoencoder (VAE) architectures using Cell Painting data and to demonstrate an important, unsolved application in predicting polypharmacology that we show is now possible for a subset of compounds. It was a natural and useful extension to compare Cell Painting VAE performance with L1000 VAE performance especially since our data set contained equivalent drug perturbations. We feel that

any extension including multi-modal data integration will distract focus away from the Cell Painting VAE novelty, and requires a much deeper dive beyond scope of our current manuscript.

Additionally, there have been other, more in-depth and very recent multi-modal data integration efforts using the same or similar datasets (Caicedo et al. 2021; Haghighi et al. 2021). In a separate paper that we just recently submitted, we also dive much deeper to answer the question of how the two modalities complement one another in various ways and for various tasks (Way, Natoli, et al. 2021). These two papers already provide a deeper and more informative exploration of Cell Painting and L1000 data integration.

Therefore, because multi-modal data integration, while certainly interesting, will distract from the Cell Painting VAE novelty and is redundant with other recent publications, we feel it is beyond scope of this current paper.

Nevertheless, multi-modal data integration is important to mention, so we add it to the discussion. Specifically, we discuss how multi-modal data integration might help with predicting polypharmacology in the future and include pertinent citations so that we, or another reader, might be able to follow-up in the future. The new section reads:

> "Because we had access to the same perturbations with L1000 readouts, we were able to compare cell morphology and gene expression results. We found that both models capture complementary information when predicting polypharmacology, which is a similar observation to recent work comparing the different technologies' information content (Way et al, 2021). We did not explore multi-modal data integration in this project; this has been explored in more detail in other recent publications (Caicedo et al, 2021; Haghighi et al, 2021). However, using multi-modal data integration with models like CycleGAN or other style transfer algorithms might provide more confidence in our ability to predict polypharmacology in the future (Zhu et al. 2017)."

1. References

Caicedo, Juan C., Sam Cooper, Florian Heigwer, Scott Warchal, Peng Qiu, Csaba Molnar, Aliaksei S. Vasilevich, et al. 2017. "Data-Analysis Strategies for Image-Based Cell Profiling." *Nature Methods* 14 (9): 849–63.

Caicedo, Juan C., Nikita Moshkov, Tim Becker, Kevin Yang, Peter Horvath, Vlado Dancik, Bridget K. Wagner, Paul A. Clemons, Shantanu Singh, and Anne E. Carpenter. 2021. "Predicting Compound Activity from Phenotypic Profiles and Chemical Structures." *bioRxiv*. https://doi.org/10.1101/2020.12.15.422887.

Caldera, Michael, Felix Müller, Isabel Kaltenbrunner, Marco P. Licciardello, Charles-Hugues Lardeau, Stefan Kubicek, and Jörg Menche. 2019. "Mapping the Perturbome Network of Cellular Perturbations." *Nature Communications* 10 (1): 5140.

Corsello, Steven M., Joshua A. Bittker, Zihan Liu, Joshua Gould, Patrick McCarren, Jodi E. Hirschman, Stephen E. Johnston, et al. 2017. "The Drug Repurposing Hub: A next-Generation Drug Library and Information Resource." *Nature Medicine* 23 (4): 405–8.

Haghighi, Marzieh, Shantanu Singh, Juan Caicedo, and Anne Carpenter. 2021. "High-Dimensional Gene Expression and Morphology Profiles of Cells across 28,000 Genetic and Chemical Perturbations." *bioRxiv*. https://doi.org/10.1101/2021.09.08.459417.

McQuin, Claire, Allen Goodman, Vasiliy Chernyshev, Lee Kamentsky, Beth A. Cimini, Kyle W. Karhohs, Minh Doan, et al. 2018. "CellProfiler 3.0: Next-Generation Image Processing for Biology." *PLoS Biology* 16 (7): e2005970.

Way, Gregory P., Maria Kost-Alimova, Tsukasa Shibue, William F. Harrington, Stanley Gill, Federica Piccioni, Tim Becker, et al. 2021. "Predicting Cell Health Phenotypes Using Image-Based Morphology Profiling." *Molecular Biology of the Cell* 32 (9): 995–1005.

Way, Gregory P., Ted Natoli, Adeniyi Adeboye, Lev Litichevskiy, Andrew Yang, Xiaodong Lu, Juan C. Caicedo, et al. 2021. "Morphology and Gene Expression Profiling Provide Complementary Information for Mapping Cell State." *bioRxiv*. https://doi.org/10.1101/2021.10.21.465335.

Way, Gregory P., Michael Zietz, Vincent Rubinetti, Daniel S. Himmelstein, and Casey S. Greene. 2020. "Compressing Gene Expression Data Using Multiple Latent Space Dimensionalities Learns Complementary Biological Representations." *Genome Biology* 21 (1): 109.

Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks." *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1703.10593.