Merging Bioactivity Predictions from Cell Morphology and Chemical Fingerprint Models by Leveraging Similarity to Training Data

Srijit Seal[1], *Hongbin Yang*[1], *Maria-Anna Trapotsi*[1], *Satvik Singh*[2], *Jordi Carreras-Puigvert*[3], *Ola Spjuth*[3,]*, *Andreas Bender*[1,]*

[1] Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge, United Kingdom

[2] Department of Applied Mathematics and Theoretical Physics (DAMTP), University of Cambridge, Cambridge, United Kingdom

[3] Department of Pharmaceutical Biosciences and Science for Life Laboratory, Uppsala University, Uppsala, Sweden

* Email: ab454@cam.ac.uk, ola.spjuth@farmbio.uu.se

Machine Learning, Cell Painting, Structure, Toxicity, Bioactivity, Applicability Domain

ABSTRACT

The applicability domain of machine learning models trained on structural fingerprints for the prediction of biological endpoints is often limited by the diversity of chemical space of the training data. In this work, we developed “similarity-based merger models” which combined the output of individual models trained on cell morphology (based on Cell Painting) and chemical structure (based on chemical fingerprints). Using a combination of a decision tree and logistic regression models on the structural versus morphological feature space of the training data, which leveraged the similarity of test compounds to training compounds, the similarity-based merger models used logistic equations to weigh individual model outputs. We applied these models to predict assay hit calls of 92 assays from ChEMBL and PubChem and 89 anonymised assays released by the Broad Institute, where the required Cell Painting annotations were available. We found that for the 181 assays used in this study the similarity-based merger model improved AUC in relative terms by 16.3% compared to the models using chemical structure alone (mean AUC of 0.75 vs. 0.64), and by 21.3% compared to the models using Cell Painting data alone (mean AUC of 0.62). Our results demonstrate that similarity-based merger models combining structure and cell morphology models can more accurately predict a wide range of biological assay outcomes and expand the applicability domain by better extrapolating to new structural and morphology spaces.

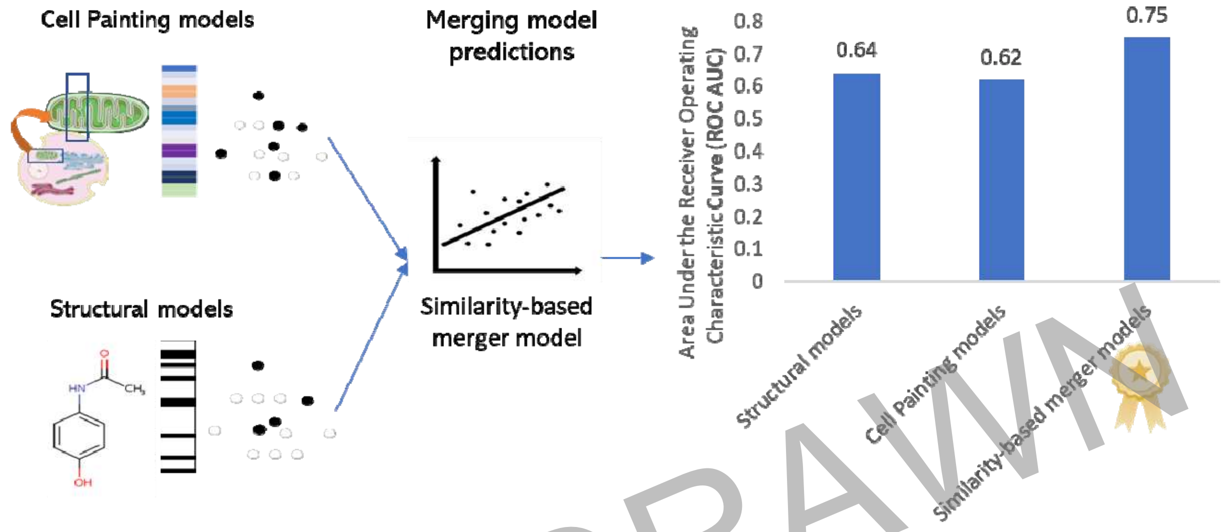

Figure: For TOC Only

MAIN

The prediction of bioactivity, mechanism of action (MOA) of compounds[1] as well as safety and toxicity[2] using only chemical structure data is challenging given that such models are limited in the diversity of the chemical space of the training data.[3] The chemical space of this data on which the model is trained is used to define the applicability domain of the model.[4] Among the various ways to calculate a model's applicability domain, Tanimoto similarity for chemical structure is commonly used as a benchmark similarity measure for compounds.[5] Tanimoto distance-based Boolean applicability has been previously used to improve the performance of classification models.[6] Expanding the applicability domain of structural models will improve the reliability of a model to predict endpoints for new compounds and one way to achieve this would be to incorporate hypothesis-free high-throughput data, such as cell morphology[7], bioactivity data[8] or predicted bioactivities[9,10] in addition to structural models which then has the potential to improve predictions for compounds structurally distant from the training data. This is because compounds having similar biological activity may not always have a similar structure; however, they may show similarities in biological response space.[11]

In recent years, relatively standardized hypothesis-free cell morphology data can now be obtained from the Cell Painting assay.[12] The Cell Painting assay uses six fluorescent dyes to capture morphological changes on eight cellular organelles imaged in five-channel microscopic images. The microscopic images are typically further processed using an image analysis software, such as Cell Profiler[13], which results in a set of circa 1700 morphological numerical features per cell. These numerical features representing morphological properties such as shape, size, area, intensity, granularity, and correlation, among many others, are considered versatile biological descriptors of a system.[7] Previous studies have shown Cell Painting data to be predictive of a wide range of bioactivity and drug safety-related endpoints such as the mechanism of action[14], cytotoxicity[24], microtubule-binding activity[15], and mitochondrial toxicity[25], and has been used to identify phenotypic signatures of PROteolysis TArgeting Chimeras (PROTACs)[16] as well as determining the impact of lung cancer variants[17]. Thus Cell Painting data can be expected to contain a signal about the biological activity of the compound perturbation,[7] and in this work, we explored how best to combine Cell Painting and chemical structural models for the prediction of a wide range of biological assay outcomes.

From the modeling perspective, several ensemble modeling techniques have been proposed to combine predictions from individual models.[18] One way to achieve this is an ensembling method shown in Figure 1a, referred to as a soft-voting ensemble in this work. This method computes the mean of predicted probabilities from individual models and thus provides equal weightage to individual model predictions. However, soft-voting ensemble models when combining two individual models give equal importance to each model.[18] This implies that if a model predicts a higher probability for a compound to be active and another model predicts the same compound to be inactive but with a lower probability, the first model prediction is considered final without considering the individual model's reliability. As shown in Figures 1b, another way to achieve this is *via* model stacking where the predictions of the individual models are used as features to build a second-level model (referred to as a hierarchical model in this work). Hierarchical models have previously been used by integrating classification and regression tasks in predicting acute oral systematic toxicity in rats.[19] The applicability domain of predictions can be estimated both by the Random Forest predicted class estimates[20] (referred to as predicted probabilities in this study) and using the similarity of the test compound to training compounds (which in turn approximates the reliability of the prediction).[21] The hypothesis of the current work is hence that using the similarity of the test compound to training compounds to weigh the predicted probabilities of an individual model can improve the model performance.

The various ways of fusing structural models with models trained on cell morphology were recently exploited by Moshkov et al.[26] who used both chemical structures and cell morphology data (from the Cell Painting assay) to predict the compound activity of 270 anonymised bioactivity assays from academic screenings in the Broad Institute. They used a late data fusion (by using a majority rule on the prediction scores similar to soft-voting ensembles) to merge predictions for individual models. The late data fusion models were able to predict 31 out of 270 assays with AUC>0.9, compared to 16 out of 270 assays for models using only structural features. This showed that fusing models built on two different feature spaces that provide complementary information was able to improve the prediction of bioactivity endpoints. Previous work has also shown that combinations of descriptors can significantly improve prediction for MOA classification[22,23,14] (using gene expression and cell morphology data), cytotoxicity[24], mitochondria toxicity[25] and anonymised assay activity[26] (using chemical, gene expression, cell morphology and predicted bioactivity data), prediction of sigma 1 (σ1) receptor antagonist[27] (using cell morphology data and thermal proteome profiling), and even organism-level toxicity[28] (using chemical, protein target and cytotoxicity qHTS data). Thus, the combination of models built from complementary feature spaces can expand a model's applicability domain by allowing predictions in new structural space.[29]

In this work, we explored merging predictions of assay hit calls from chemical structural models with predictions from another model using Cell Painting data for 92 assays from public datasets from PubChem and ChEMBL (henceforth referred to as public dataset released as Supplementary Data 1) and 89 anonymised assays from the Broad Institute[26] (henceforth referred to as Broad Institute dataset released as Supplementary Data 2). As shown in Figure 1c, we merged predictions by weighting the model that had the most-similar profile by leveraging the knowledge of a test compound's similarity to the training data of the individual models in the respective feature space, that is, by providing more weightage to the Cell Painting models when the test compound is morphologically similar to the training set but more weightage to the structural model when the structural similarity is high. Here we emphasise using similarity-based merger models to improve the applicability domain of individual models (predicting compounds that are distant to training data in respective feature spaces) and the ability to predict a wider range of assays with the combined knowledge from the chemical structure and biological descriptors from Cell Painting assay.

RESULTS AND DISCUSSIONS

We trained individual Cell Painting and structural models for both the public dataset collection comprising 92 assays, and the Broad Institute dataset comprising 89 assays. We used two baseline models for comparison, namely a soft-voting ensemble and a hierarchical model and compared the results from the individual models and baseline ensemble models to the similarity-based merger models .

As shown in Figure 2, we found that similarity-based merger models performed with significantly improved AUC-ROC (mean AUC 0.71 using similarity-based merger models) compared to individual models (mean AUC 0.57 using Cell Painting models; mean AUC 0.58 using structural models) in the public dataset as well as the Broad Institute dataset (mean AUC 0.67 using Cell Painting models; mean AUC 0.71 using structural models; mean AUC 0.79 using similarity-based merger models). Figure S1 shows similarity-based merger models significantly improved Balanced Accuracy and F1 scores compared to individual models. Results are presented separately from the PubChem dataset comprising 92 assays (as shown in Supplementary Data 3) and the Broad Institute dataset comprising 89 assays (as shown in Supplementary Data 4) as the Broad Institute dataset is not annotated with complete biological metadata, which renders some of the more detailed analysis downstream not viable.

(A) Results from the Public dataset collection

We next analysed results from the public dataset comprising 92 assays (with at least 100 compounds for which Cell Painting annotations were available) collected from Hofmarcher et al[33] and Vollmers et al[35] (see Supplementary Data 1 for assay descriptions). As shown in Figure 3, 51 out of 92 assays achieved AUC>0.70 with the similarity-based merger model, followed by hierarchical models for 24 out of 92 assays. Structural models achieved AUC>0.70 in only 17 out of 92 assays while for the Cell Painting models this was the case in only 13 out of 92 assays. Only 3 assays out of 92 were predicted with AUC>0.70 with all methods while only 6 out of 92 assays did not achieve AUC>0.70 with similarity-based merger models but did with the other models. When considering balanced accuracy, 30 out of 92 assays achieved a balanced accuracy > 0.70 with similarity-based merger models compared to 9 out of

92 assays for hierarchical models as shown in Figure S2. Comparing performance for the Cell Painting and structural models by AUC individually (Figure S3) we observed that structural models and Cell Painting models were complementary in their predictive performance; while 48 out of 92 assays achieve a higher AUC with structural information alone, 42 out of 92 assays achieve a higher AUC using morphology alone as shown in Figure S3a. Hierarchical models outperform soft-voting ensembles for 59 out of 92 assays as shown in Figure S3b. Finally, the similarity-based merger model achieved a higher AUC score for 83 out of 92 assays compared to 8 out of 92 with soft-voting ensembles and 78 out of 92 assays compared to 13 out of 92 with hierarchical models as shown in Figures S3c and S3d. This shows that the similarity-based merger model was able to leverage information from both Cell Painting and structural models to achieve better predictions in assays where no individual models were found to be predictive thus indicating synergistic effect.

We next looked at the performance at the individual assay level (as shown in Supplementary Data 3) as indicated by the AUC scores. We looked at 90 out of 92 assays where either the similarity-based merger model or the soft-voting ensemble performed better than a random classifier ($AUC = 0.50$) We observed that for 83 out of 90 assays (individual changes in a performance recorded in Figure S4), the similarity-based merger models improved performance compared to the soft-voting ensemble (with the largest improvement recoded at 212.6%) and a decrease in performance was recorded in 6 out of 90 assays (largest decrease recorded at -11.9% in performance). Further comparisons of AUC performance in Figure S5 show that similarity-based merger models improved AUC compared to both structural models and Cell Painting models. This improvement in AUC was independent of the total number of compounds in the resampled assay as shown in Figure S6(a). Thus, we conclude that the similarity-based merger model outperformed individual models by combining the rich information contained in cell morphology and structure-based models more efficiently than baseline models.

Comparison of Performance at Gene Ontology Enrichment level

We next analysed the assays (and associated biological processes) where the Cell Painting model, structural model, and the similarity-based merger model were most predictive and therefore if there was complementary information present in both feature spaces. Figure 4a shows a protein-protein network (annotated by genes) from the STRING database labelled by the model performance where the respective individual model was better predictive (or otherwise equally predictive, which includes cases where different models are better predictive over multiple assays related to the same protein target). We found meaningful models (AUC > 0.50) were achieved when using either the Cell Painting or the structural model for 31 out of 35 gene annotations. Of these, the Cell Painting models were better predictive for 9 out of 31 gene annotations (average AUC= 0.69) compared to the structural models which were better predictive for 20 out of 31 gene annotations (average AUC=0.68). We next compared the soft-voting ensemble model to the similarity-based merger model for 35 gene annotations where either model achieved AUC>0.50. The soft-voting ensemble model performed with higher AUC (average AUC= 0.60) for only 2 out of 35 gene annotations compared to the similarity-based merger model which was better predictive for 32 out of 35 gene annotations (average AUC=0.75). Thus, we observe that similarity-based merger models performed better over a range of assays (over 33 out of 35 gene annotations) capturing a wide range of biological pathways.

In particular, Cell Painting models performed better than structural models for assays associated with 6 gene annotations, POLK, FEN1, GMNN, VDR, KAT2A and NFE2L2 (with an average AUC = 0.67 for Cell Painting models compared to AUC = 0.57 for structural models). These gene annotations were associated with molecular functions of ‘GO:0003684 Damaged DNA binding’, ‘GO:0140297 DNA-binding transcription factor binding’, and ‘GO:0008134 Transcription factor binding’, which are processes resulting in morphological changes induced by compounds that damage or bind the DNA, which were captured by Cell Painting. Among gene annotations associated with the assays better predicted by structural models are TSHR, GLP1R, DRD1, HCRTR1 and CHRM1 (with average AUC = 0.67 for structural models compared to AUC = 0.58 for Cell Painting models). These gene annotations are associated with the KEGG pathway of ‘neuroactive ligand-receptor interaction’ and the Reactome pathway of ‘amine ligand-binding receptors’ which were captured better by chemical structure. We can hence see that Cell Painting models perform better on assays capturing morphological changes in cell or cellular compartments such as the nucleus, while structural models work better for assays associated with a ligand-receptor activity. In addition, the KEGG term ‘amine ligand-binding receptors’ is defined on the chemical ligand level explicitly, making the classification of compounds falling into this category from the structural side easier. The similarity-based merger models hence combined the power of both spaces and were better predictive for both assays affecting morphological changes (average AUC= 0.71 for the similarity-based merger model) as well as related to the ligand-receptor binding activity (average AUC= 0.78 for similarity-based merger model). This is further illustrated in Figure 4b which shows enriched molecular and functional pathway terms from ClueGO[39] for the 35 gene annotations available. We observe that around 25-33% of gene annotations associated with damaged or DNA binding, transcription coregulator binding, and positive regulation of blood vessel endothelial cell migration pathways are better predicted with AUC>0.70 with the Cell Painting models. At the same time around 33-50% of gene annotations associated with G protein-coupled receptor signalling pathways and transcription coregulator binding were predicted with AUC>0.70 with structure-based models. Similarity-based merger models predicted 67-100% gene annotations associated with all four pathways with an AUC>0.70, hence underlining the utility of such merger models across a range of biological endpoints.

Applicability domain analysis

We next determined how individual and similarity-based merger model predictions differ with compounds that were structurally or morphologically similar/dissimilar to the training set. We looked at predictions for each compound from the Cell Painting and structural models over the 92 assays and grouped them based on their morphological and structural similarity to the training set respectively. We observed, as shown in Figure S7, that as the morphological similarity of test compounds with respect to the training set increased, the Cell Painting models correctly classified a higher proportion of compounds, while the structural model improved performance as the test compound becomes more structurally similar to the training set. Further, as the structural similarity of test compounds with respect to the training set increased, the models using chemical structure correctly classified a higher proportion of compounds, while the structural model improved performance as the test compound becomes more structurally similar to the training set. For example, out of 1,888 compounds with a low structural similarity between 0.25 to 0.35, models using chemical structure correctly classified 46.8% compounds while similarity-based merger models correctly classified a much greater 57.0% compounds. However, out of 63 compounds with higher structural similarity between 0.65 to 0.85, models using chemical structure correctly classified 85.7% compounds comparable to the similarity-based merger models that correctly classified 82.5% compounds. This shows that the similarity-based merger model correctly predicted a larger proportion of compounds over a wide range of structural and morphological similarities to the training set, hence demonstrating an increase in the applicability domain. For clarity of the reader, this is further illustrated in Figure S8 as in the case of a particular assay, namely PubChem assay 2686 which is a qHTS assay for lipid storage modulators in drosophila S3 cells. Here, the structural model correctly predicted compound activity when they were structurally similar to the training set. The Cell Painting model performed better over a wide range of structural similarities but was often limited when morphological similarity was low. This shows that similarity-based merger models learned and adapted weights across individual models from local regions in this structural versus morphological similarity space in a manner best suited to compounds in that region to correctly classify a wider range of compounds with lowered structural and morphological similarities to the training set.

(B) Results from the Broad Institute dataset

We next analysed the performance of 89 assays from the Broad Institute dataset in more detail (as shown in Supplementary Data 4). As shown in Figure 5, the similarity-based merger models outperformed all other models and achieved an AUC>0.70 for 74 out of 89 assays, followed by hierarchical models at 52 out of 89 assays. Structural models were able to achieve an AUC>0.70 for 48 out of 89 assays compared to Cell Painting models in 35 out of 89 assays. For 22 out of 92 assays, all methods achieved an AUC>0.70 while only 4 out of 89 assays did not achieve AUC>0.70 with the similarity-based merger models but did with the other models. Figure S10 compares the balanced accuracy of all 5 models which shows that similarity-based merger models achieved a balanced accuracy > 0.70 for 51 out of 89 assays compared to 35 out of 89 assays for hierarchical models. Hence, the similarity-based merger models predicted an additional 18 out of 89 assays with AUC>0.7 that no other method did. Similarly, we observed similarity-based merger models predicted an additional 14 out of 89 assays balanced accuracy > 0.70 as shown in Figure S9, once again demonstrating its applicability in predicting a wide range of assays. Further, as shown in Figure S10, Cell Painting models and structural models were again complementary in the assays they predicted better individually (54 out of 89 assays perform better with structural information, 35 out of 89 assays better perform using morphology). Hierarchical models outperformed soft-voting ensembles for 81 out of 89. The similarity-based merger model outperformed both baseline models (in 85 out of 89 assays compared to 4 out of 89 assays in soft-voting ensembles and 69 out of 89 assays compared to 17 out of 89 assays for hierarchical models).

We next looked at performance at individual assay level (as shown in Supplementary Data 4). All 89 assays achieved AUC>0.50 (better than a random classifier) using either the similarity-based merger model or the soft-voting ensemble. We observed that for 85 out of 92 assays (individual changes in a performance recorded in Figure S11), similarity-based merger models improved performance compared to the soft-voting ensemble (with the largest improvement recorded at 70.7%) and a decrease in performance was recorded in only 6 out of 90 assays (with the largest decrease in performance recorded at -37.1%). Further comparisons of AUC achieved by the similarity-based merger model compared to structural models and Cell Painting models show a similar improvement in performance as shown in Figure S12. Further, the improvement in AUC from using the similarity-based merger model was independent of the total number of compounds in an assay as shown in Figure S6(b). Thus, similarity-based merger models were able to outperform both baseline ensemble methods and hence can be used to capture a wide range of assay endpoints.

Comparison of Performance by Readout and Assay type

As the Broad Institute dataset was released with only information about assay type and readout type (for details see Supplementary Data 2 and Figure S13), we analysed the Cell Painting, structural and similarity-based merger model as a function of those. As shown in Figure 6, Cell Painting models perform significantly better with a relative 12.8% increase in AUC with assays measuring luminescence (mean AUC = 0.72) compared to assays measuring fluorescence (mean AUC = 0.64) while structural and similarity-based merger model show no significant differences in performances. The better predictions in the case of luminescence-based assays, which are readouts specifically designed to answer a biological question, and can be related to the use of a reporter cell line and a reagent that based on the ATP content of the cell, is converted to a luciferase substrate which leads to a cleaner datapoint.[30] On the other hand, Cell Painting is an unbiased high-content imaging assay that takes into consideration the inherent heterogeneity in cell cultures where we visualise cells (often even measuring at a single cell level), contrary to a luminescence assay where one measures the average signal of a cell population. Further Cell Painting models performed significantly better with a relative 16.2% increase in AUC for cell-based assays (mean AUC = 0.72) compared to biochemical assays (mean AUC = 0.62). This might be due to also the Cell Painting assay being a cellular assay, hence also implicitly including factors such as membrane permeability in measurements. Further, similarity-based merger models outperform individual and baseline models over most assay and readout types as shown in Figures S14 and S15. Overall, Cell Painting models can hence be considered to provide complementary information to chemical structure regarding cell-based assays, which was particularly beneficial for the significant improvement in the performance of similarity-based merger models.

Comparison of performance based on the similarity of test compounds to training data

We next compared the performance of individual and similarity-based merger models across all compounds in all 89 assays from the Broad dataset based on their morphological or structural similarity to the respective training data. Here, the results were consistent with the results obtained above for the public dataset. As shown in Figure S16, out of 1,032 compounds with a low structural similarity between 0.25 to 0.35, models using chemical structure correctly classified 50.8% compounds while similarity-based merger models correctly classified a much greater 59.7% compounds. However, out of 2,640 compounds with a higher structural similarity between 0.80 to 0.90, models using chemical structure correctly classified 80.7% of compounds compared to the similarity-based merger models that correctly classified 84.5% of compounds. Thus, once again, similarity-based merger models were able to classify a high percentage of compounds over a wider range of structural and morphological similarities to the respective training data and use the complementary information present in both feature spaces to extrapolate to novel chemical space where individual models failed.

Limitations of this work

One limitation of the study is having to balance unequal data classes by undersampling. Here, the data was therefore initially undersampled to a 1:3 ratio of majority to minority class in order to build a similarity-based merger model, which leads to some loss of experimental data. Further, after the splitting of the dataset into training and test datasets, the training data need to contain enough samples spread across the structural versus morphological similarity map for the logistic regression models to work. This was ensured by a random split; other splitting strategies such as scaffold-based splitting may not allow the use of the second level logistic regression models as the chemical space of the test data will vary significantly from the training data. Further, from the side of feature spaces, Cell Painting data is derived from U2OS cell-based assays which are usually different from the cell lines used in measuring the activity endpoint. However previous work has shown that Cell Painting data is similar across different cell lines and the versatile information present was universal, that is, the genetic background of the reporter cell line does not affect the AUC values for MOA prediction.[31] Thus Cell Painting data can be used to model different assays with different cell lines. Future studies will also benefit from larger datasets, such as the JUMP-CP consortium.[32]

CONCLUSIONS

Predictive models that use chemical structures as features are often limited in their applicability domain to compounds which are structurally similar to the training data. In this work, we developed similarity-based merger models to combine two models built on complementary feature spaces of Cell Painting and chemical structure and predicted assay hit calls from 92 assays from the public dataset and 89 assays from a dataset released by the Broad Institute for which Cell Painting data were available.

We found that Cell Painting and chemical structure contain complementary information and can predict assays associated with different biological pathways, assay types, and readout types. Cell Painting models achieved higher AUC better for cell-based assays and related to biological pathways such as damaged DNA binding. Structural models achieved a higher AUC for biochemical and ligand-receptor binding assays associated with pathways such as G protein-coupled receptor signalling pathways. The similarity-based merger models, combining the two feature spaces, achieved a higher AUC for both cell-based (mean AUC=0.83) and biochemical assays (mean AUC=0.77) as well as assays related to both biological pathways (mean AUC=0.72) and ligand-receptor based pathways (mean AUC=0.76). Further, the similarity-based merger models outperformed all other models with an additional 20-30% assays over both datasets with AUC>0.70. We also showed that the similarity-based merger models correctly predicted a larger proportion of compounds which are comparatively less structurally and morphologically similar to the training data compared to the individual models, thus being able to improve the applicability domain of the models.

In conclusion, the similarity-based merger models improved the prediction of assay outcomes by combining high predictivity of fingerprints in areas of structural space close to the training set with better generalizability of cell morphology descriptors at greater distances to the training set. Such models can hence contribute to overcoming the limitation of chemical space in drug discovery projects.

METHODS

Bioactivity Datasets

We retrieved drug bioactivity data as binary assay hit calls for 202 assays and 10,570 compounds from Hofmarcher et al[33] who searched ChEMBL[34] for assays for which cell morphology annotations from the Cell Painting assay were available as shown in Figure S17. We further added binary assay hit calls from another 30 assays not included in the source above from Vollmers et al[35] who searched PubChem[36] assays for overlap with Cell Painting annotations (see Supplementary Data 1 for assay descriptions). Additionally, we used 270 anonymised assays (with binary endpoints) from the Broad Institute[26] as shown in Figure S17 (see Supplementary Data 2 for assay descriptions). This dataset, although not annotated in with biological metadata, comprises assay screenings performed over 10 years at the Broad Institute and is representative of their academic screenings.

Gene Ontology Enrichment of Bioactivity Assays

From the public dataset of 92 assays used in this study where detailed assay data was available, 38 out of 92 assays where experiments used human-derived cell lines were annotated to 35 protein targets. Next, we determined using the STRING database[37], we annotated all 35 protein targets with the associated gene set and further obtained a set of Gene Ontology terms associated with the protein target. We used Cytoscape[38] v3.9.1 plugin ClueGO[39] to condense the protein target set by grouping them into functional groups to obtain the associated significant (using the baseline ClueGO p-value ≤0.05) molecular and functional pathway terms. In this manner, we associated individual assays to molecular and functional pathways for further evaluation of model performances.

Cell Painting Data

The Cell Painting assay used in this study, from the Broad Institute, contains 30,616 morphological profiles of small molecule perturbations.[40,41] Following the procedure from Lapins et al, we subtracted the average feature value of the neutral DMSO control from the particular compound perturbation average feature value on a plate-by-plate basis.[14] For each compound and drug combination, we calculated a median feature value. Where the same compound was replicated for different doses, we used the median feature value across all doses that were within one standard deviation of the mean dose. Finally, after SMILES standardisation and removing duplicate compounds using standard InChI calculated using RDKit[43], we obtained 1783 median Cell Painting features for 30,404 unique compounds (available on Zenodo at https://doi.org/10.5281/zenodo.6613741).

Overlap of Datasets

For both the public and Broad dataset, as shown in Figure S17 (step 1) we used MolVS[42] standardizer based on RDKit[43] to standardize and canonicalize SMILES for each compound which encompassed sanitization, normalisation, greatest fragment chooser, charge neutralisation, and tautomer enumeration described in the MolVS documentation[42]. We further removed duplicate compounds using standardised InChI calculated using RDKit[43].

Next, for the public dataset, we obtained the overlap with the Cell Painting dataset using standardised InChI Figure S17 (step 2). From this, we removed 148 assays which contained less than 100 compounds with Cell Painting datasets (which were difficult to model due to limited data) as shown in Figure S17 (step 3). Finally, we obtained the public assay data for a sparse matrix of 92 assays and 10,402 unique compounds. Similarly, for the Broad dataset, out of 270 assays provided, as shown in Figure S17, we removed 181 assays that contained less than 100 compounds resulting in a Broad Institute dataset as a sparse matrix of 16170 unique compounds over 89 assays. (Both datasets are publicly available on Zenodo at https://doi.org/10.5281/zenodo.6613741).

Structural Data

We generated Morgan Fingerprints of radius 2 and 2048 bits using RDKit[43] used as chemical fingerprint features in this work.

Feature Selection

Firstly, we performed feature selection to obtain morphological features for each compound. From 1,783 Cell Painting features, we removed 55 blocklist features that were known to be noise from Way et al.[44] For the compounds in the public assays, we further removed 1,011 features which had a very low variance below a 0.005 threshold using the scikit-learn[45] variance threshold module. Next, similar to the feature section implemented in pycytominer[46], we obtained the list of features such that no two features correlate greater than a 0.9 Pearson correlation threshold. For this, we calculated all pairwise correlations between features and removed the 488 features with the highest pairwise correlations. Finally, we removed another 45 features if their minimum or the maximum absolute value was greater than 15 (using the default threshold in pycytominer[46]). Hence, we obtained 184 Cell Painting features for 10,402 unique compounds for the dataset comprising public assays. Analogously, for the Broad Institute dataset, we obtained 192 Cell Painting features for 16,170 unique compounds (both datasets are available on Zenodo at https://doi.org/10.5281/zenodo.6613741).

Next, we performed feature selection for structural features of Morgan fingerprints. For the public assays, we removed 1,883 bits that did not pass a near-zero variance (0.05) threshold since they were considered to have less predictive power. Finally, we obtained Morgan fingerprints of 165 bits for 10,402 unique compounds. Analogously, for the Broad Institute dataset, we obtained Morgan fingerprints of 273 bits for 16,170 unique compounds (both datasets are available on Zenodo at https://doi.org/10.5281/zenodo.6613741).

Chemical and Morphological Similarity

We next defined the structural similarity score of a compound as the median Tanimoto similarity of the 5 most similar compounds of the same class. The morphological similarity score of a compound was calculated as the median Pearson correlation to the 5 most positively correlated compounds of the same class.

Model Training

For each assay, the majority class (most often the negative class) was randomly sampled to maintain a minimum 3:1 ratio with the minority class to ensure that models are fairly balanced (which is henceforth referred to as resampled assay in this study). Figure S18 shows the distribution of the total number of compounds in the resampled assay for both the public dataset (comprising 92 assays) and the Broad Institute dataset (comprising 89 assays).

Then, the data was split into training data (80%) and held out test data (20%) using a stratified splitting based on the assay hit call. First, on the training data, we performed 5-fold nested cross-validation keeping aside one of these folds as a test-fold, on the rest of the 4 folds. We trained separate models, as shown in Figure 1c step (1) and step (2), using Morgan fingerprints (165 bits for the public dataset; 273 bits for the Broad Institute dataset) and Cell Painting data (184 features for the public dataset, 192 features for the Broad Institute dataset) respectively for each assay. In this inner fold of the nested-cross validation, we trained separately, Random Forest models on the rest of the 4 folds with Cell Painting and Morgan fingerprints. These models were hyperparameter optimised (with parameter spaces as shown in Supplementary Data 5) using cross-validation with shuffled 5-fold stratified splitting. For hyperparameter optimisation, we used a randomized search on hyperparameters as implemented in scikit-learn 1.0.1[45]. This optimisation method iteratively increases resources to select the best candidates, using the most resources on the candidates that are better at prediction.[47] The hyperparameter optimised model was used to predict the test-fold. To account for threshold balancing of Random Forest predicted probabilities (which is common in an imbalanced prediction problem), we calculated on the 4 folds, the Youden's J statistic[48] (J = True Positive Rate – False Positive Rate) to detect an optimal threshold. The threshold for the highest J statistic value was used such that the model would no longer be biased towards one class and give equal weights to sensitivity and specificity without favouring one of them. This optimal threshold was then used for the test-fold predictions, and this was repeated 5 times for both models using Morgan fingerprints and Cell Painting features until predictions were attained for the entire training data in the nested cross-validation manner. As the optimal thresholds for each fold were different, the predicted probability values were scaled using a min-max scaling such that this optimal threshold was adjusted back to 0.50 on the new scale. Further for each test-fold in the cross-validation, as shown in Figure 1c step (3) and step (4), we also calculated the chemical and morphological similarity (as described above in the “Chemical and Morphological Similarity” section) for each compound in this test-fold with respect to the compounds in the remaining of the 4 folds. This was repeated 5 times until chemical and morphological similarity scores were obtained for the entire training data.

Finally, on the entire training data, two Random Forest models were trained with Cell Painting and Morgan fingerprints with hyperparameter-optimised (in the same way as above using 5-fold cross-validation). This was used to predict the held-out data, as shown in Figure 1c step (5) (with threshold balancing performed from cross-validated predicted probabilities of the training data). We calculated the chemical and morphological similarity of each compound in the held-out data compared to all compounds in the training data and these were recorded as the chemical and morphological similarity scores respectively of the particular compound in the held-out dataset as shown in Figure 1c step (6). The predicted probability values were again adjusted using a min-max scaling such that this optimal threshold was 0.50 on the new scale.

Similarity-based merger model

The distance-based merger models presented here are a combination of multiple hierarchical models trained in different regions on the morphology versus structural similarity space of the test set with respect to the training set. In particular, for each assay, we built the similarity-based merger model on the held-out data using information from the training data to avoid any data or model leakage. First, the training data was further resampled to have a 1:1 ratio of active to inactive (further reducing the number of compounds but essential to make sure we favour the distance boundaries in a balanced manner for both majority and minority classes). As shown in Figure 1c step (7), we trained a single decision tree classifier with a maximum depth of 2 (using scikit-learn[45] DecisionTreeClassifier) on the training data using input features of the structural similarity score and morphological similarity scores and endpoints as two Boolean variables indicating if the two individual models predicted the assay hit call correctly. The decision tree maximum depth was set as 2 to ensure that there was a maximum of 4 end nodes formed. We hard-coded this decision tree and used this on the held-out data using structural similarity scores and morphological similarity scores (as defined in the *Chemical and Morphological Similarity* Section above**)** to predict which of the up to 4 end leaf node classes each compound of the held-out test set would fall be in. These node classes are effectively decision boundaries in the structural and morphological similarity space that were defined using the training data only. There is no leak of any held-out data assay hit call information but only its structural similarity and morphological similarity to the training data, which can be easily calculated for any compound with a known structure.

For each node class where held-out compounds were present, we predict the assay hit call of the compound using a soft-voting ensemble model for the following conditions: (a) if no training compounds were present but held-out compounds were, (b) only training compounds of one class were present. In these two cases, we cannot use the similarity-based merger model. For all other cases, we used baseline logistic regression models (with baseline parameters of L2 penalty and an inverse of regularization strength of 5) using the Cell Painting and Morgan fingerprints models' individual scaled predicted probabilities as features and the endpoint as the assay hit call of the compound, as shown in Figure 1c step (8). We used the training compounds of the particular node class to fit a logistic regression model in an attempt to determine which model should be given more weightage in which node class. Finally, this logistic equation was used to predict the assay hit call of the held-out compounds (which we henceforth call the similarity-based merger model prediction) and an associated predicted probability (which we henceforth call similarity-based merger model predicted probability), as shown in Figure 1c step (9).

Baseline Models

For baseline models, we used two models, namely a soft-voting ensemble[18] and a hierarchical model[19]. The soft-voting ensemble, as shown in Figure 1a, combines predictions from both the Cell Painting and Morgan fingerprints models using a majority rule on the predicted probabilities. In particular, for each compound, we averaged the re-scaled predicted probabilities of two individual models, thus in effect creating an ensemble with soft-voting. We applied a threshold of 0.50 (since predicted probabilities from individual models were also scaled to the optimal threshold of 0.50 as described above) to obtain the corresponding soft-voting ensemble prediction.

For the hierarchical model, as shown in Figure 1b, we fit a baseline logistic regression model on the (with L2 penalty and an inverse of regularization strength of 5), as implemented in scikit-learn, on the scaled predicted probabilities both individual the Cell Painting and Morgan fingerprints models for the entire training data from cross-validation. We applied this logistic equation to the held-out test set compounds to predict the assay hit call (and a corresponding model predicted probability) which we henceforth call the hierarchical model prediction (and a corresponding hierarchical model predicted probability).

Model evaluation

We evaluated all models (both individual models, soft-voting ensemble, hierarchical and similarity-based merger model) based on precision, sensitivity, F1 scores of the minority class, specificity, balanced accuracy, Matthew's Correlation Coefficient (MCC) and Area Under Curve- Receiver Operating Characteristic (AUC) scores.

Statistics and Reproducibility

A detailed description of each analysis' steps and statistics is contained in the methods section of the paper. Statistical methods were implemented using the pandas Python package.[49] Machine learning models, hyperparameter optimisation and evaluation metrics were implemented using scikit-learn[45], a Python-based package. We have released the datasets used in this study which are publicly available at Zenodo (https://doi.org/10.5281/zenodo.6613741). We released the python code for the models which are publicly available on GitHub (https://github.com/srijitseal/Merging-Predictions-from-Cell-Morphology-and-Structural-Models-by-Leveraging-Similarity).

ASSOCIATED CONTENT

SUPPORTING INFORMATION:

We have released the datasets used in this study which are publicly available at Zenodo at https://doi.org/10.5281/zenodo.6613741. We released the python code for the models which are publicly available on GitHub at https://github.com/srijitseal/Merging-Predictions-from-Cell-Morphology-and-Structural-Models-by-Leveraging-Similarity

Supplementary Data

Supplementary Data 1: Assay descriptions of the public dataset comprising 92 assays.

Supplementary Data 2: Assay types and readout types of Broad Institute dataset comprising 89 assays.

Supplementary Data 3: Performance of Cell Painting, structural models, soft-voting ensembles, hierarchical models, and the similarity-based merger models over the public dataset comprising 92 assays.

Supplementary Data 4: Performance of Cell Painting, structural models, soft-voting ensembles, hierarchical models, and the similarity-based merger models over the Broad Institute dataset comprising 89 assays.

Supplementary Data 5: Hyperparameters considered for optimising Random Forests

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGMENT

S.S. acknowledges the Cambridge Commonwealth, European and International Trust, Boak Student Support Fund (Clare Hall), Jawaharlal Nehru Memorial Fund, Allen, Meek and Read Fund, and Trinity Henry Barlow (Trinity College) for providing funding for this study. This project received funding from the Swedish Research Council (grants 2020-03731 and 2020-01865), and FORMAS (grant 2018-00924). The authors are grateful for guidance from Anne Carpenter and Juan Caicedo which helped improve the manuscript greatly. This work was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service (www.csd3.cam.ac.uk), provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/P020259/1), and DiRAC funding from the Science and Technology Facilities Council (www.dirac.ac.uk). This work was also performed using the Nest compute server (https://www.ch.cam.ac.uk/computing/nest-compute-server) provided by the Yusuf Hamied Department of Chemistry.

REFERENCES

(1) Trapotsi, M. A.; Hosseini-Gerami, L.; Bender, A. Computational Analyses of Mechanism of Action (MoA): Data, Methods and Integration. *RSC Chem. Biol.* **2022**, *3* (2), 170–200.

(2) Sazonovas, A.; Japertas, P.; Didziapetris, R. Estimation of Reliability of Predictions and Model Applicability Domain Evaluation in the Analysis of Acute Toxicity (LD50). *SAR QSAR Environ. Res.* **2010**, *21* (1–2), 127–148.

(3) Kar, S.; Roy, K.; Leszczynski, J. Applicability Domain: A Step toward Confident Predictions and Decidability for QSAR Modeling. Methods Mol. Biol. 2018, 1800, 141–169.

(4) Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O. A Stepwise Approach for Defining the Applicability Domain of SAR and QSAR Models. *J. Chem. Inf. Model.* **2005**, *45* (4), 839–849.

(5) Bajusz, D.; Rácz, A.; Héberger, K. Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations? *J. Cheminform.* **2015**, *7* (1), 1–13.

(6) Berenger, F.; Yamanishi, Y. A Distance-Based Boolean Applicability Domain for Classification of High Throughput Screening Data. *J. Chem. Inf. Model.* **2018**, *59* (1), 463–476.

(7) Chandrasekaran, S. N.; Ceulemans, H.; Boyd, J. D.; Carpenter, A. E. Image-Based Profiling for Drug Discovery: Due for a Machine-Learning Upgrade? *Nat. Rev. Drug Discov.* **2021**, *20* (2), 145–159..

(8) Kauvar, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engqvist-Goldstein, Å.; Bukar, R.; Bauer, K. E.; Dilley, H.; Rocke, D. M. Predicting Ligand Binding to Proteins by Affinity Fingerprinting. *Chem. Biol.* **1995**, *2* (2), 107–118.

(9) Norinder, U.; Spjuth, O.; Svensson, F. Using Predicted Bioactivity Profiles to Improve Predictive Modeling. *J. Chem. Inf. Model.* **2020**, *60* (6), 2830–2837.

(10) Bender, A.; Jenkins, J. L.; Glick, M.; Zhan, D.; Nettles, J. H.; Davies, J. W. "Bayes Affinity Fingerprints" Improve Retrieval Rates in Virtual Screening and Define Orthogonal Bioactivity Space: When Are Multitarget Drugs a Feasible Concept? *J. Chem. Inf. Model.* **2006**, *46* (6), 2445–2456. https://doi.org/10.1021/ci600197y.

(11) Petrone, P. M.; Simms, B.; Nigsch, F.; Lounkine, E.; Kutchukian, P.; Cornett, A.; Deng, Z.; Davies, J. W.; Jenkins, J. L.; Glick, M. Rethinking Molecular Similarity: Comparing Compounds on the Basis of Biological Activity. *ACS Chem. Biol.* **2012**, *7* (8), 1399–1409.

(12) Bray, M. A.; Singh, S.; Han, H.; Davis, C. T.; Borgeson, B.; Hartland, C.; Kost-Alimova, M.; Gustafsdottir, S. M.; Gibson, C. C.; Carpenter, A. E. Cell Painting, a High-Content Image-Based Assay for Morphological Profiling Using Multiplexed Fluorescent Dyes. *Nat. Protoc. 2016 119* **2016**, *11* (9), 1757–1774.

(13) McQuin, C.; Goodman, A.; Chernyshev, V.; Kamentsky, L.; Cimini, B. A.; Karhohs, K. W.; Doan, M.; Ding, L.; Rafelski, S. M.; Thirstrup, D.; Wiegraebe, W.; Singh, S.; Moshkov, T.; Caicedo, J. C.; Carpenter, A. E. CellProfiler 3.0: Next-Generation Image Processing for Biology. *PLoS Biol.* B, *16* (7), e2005970.

(14) Lapins, M.; Spjuth, O. Evaluation of Gene Expression and Phenotypic Profiling Data as Quantitative Descriptors for Predicting Drug Targets and Mechanisms of Action. bioRxiv. bioRxiv March 17, 2019, p 580654.

(15) Akbarzadeh, M.; Deipenwisch, I.; Schoelermann, B.; Pahl, A.; Sievers, S.; Ziegler, S.; Waldmann, H. Morphological Profiling by Means of the Cell Painting Assay Enables Identification of Tubulin-Targeting Compounds. *Cell Chem. Biol.* **2022**, *29* (6), 1053–1064..

(16) Trapotsi, M.-A.; Mouchet, E.; Williams, G.; Monteverde, T.; Juhani, K.; Turkki, R.; Miljković, F. M.; Martinsson, A.; Mervin, L.; Pryde, K. R.; Mu□, E.; Barrett, I.; Engkvist, O.; Bender, A.; Moreau, K. Cell Morphological Profiling Enables High-Throughput Screening for PROteolysis TArgeting Chimera (PROTAC) Phenotypic Signature. *ACS Chem. Biol.* **2022**, *17* (7), 1733–1744.

(17) Caicedo, J. C.; Arevalo, J.; Piccioni, F.; Bray, M.-A.; Hartland, C. L.; Wu, X.; Brooks, A. N.; Berger, A. H.; Boehm, J. S.; Carpenter, A. E.; Singh, S. Cell Painting Predicts Impact of Lung Cancer Variants. *Mol. Biol. Cell* **2022**, *33* (6).

(18) Dietterich, T. G. Ensemble Methods in Machine Learning. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **2000**, *1857 LNCS*, 1–15.

(19) Li, X.; Kleinstreuer, N. C.; Fourches, D. Hierarchical Quantitative Structure-Activity Relationship Modeling Approach for Integrating Binary, Multiclass, and Regression Models of Acute Oral Systemic Toxicity. *Chem. Res. Toxicol.* **2020**, *33* (2), 353–366.

(20) Klingspohn, W.; Mathea, M.; Ter Laak, A.; Heinrich, N.; Baumann, K. Efficiency of Different Measures for Defining the Applicability Domain of Classification Models. *J. Cheminform.* **2017**, *9* (1), 1–17.

(21) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR. J. Chem. Inf. Comput. Sci. **2004**, 44 (6), 1912–1928.

(22) Way, G. P.; Natoli, T.; Adeboye, A.; Litichevskiy, L.; Yang, A.; Lu, X.; Caicedo, J. C.; Cimini, B. A.; Karhohs, K.; Logan, D. J.; Rohban, M.; Kost-Alimova, M.; Hartland, K.; Bornholdt, M.; Chandrasekaran, N.; Haghighi, M.; Singh, S.; Subramanian, A.; Carpenter, A. E. Morphology and Gene Expression Profiling Provide Complementary Information for Mapping Cell State. *bioRxiv* **2021**, 2021.10.21.465335..

(23) Haghighi, M.; Singh, S.; Caicedo, J.; Carpenter, A. High-Dimensional Gene Expression and Morphology Profiles of Cells across 28,000 Genetic and Chemical Perturbations. *bioRxiv* **2021**, 2021.09.08.459417.

(24) Seal, S.; Yang, H.; Vollmers, L.; Bender, A. Comparison of Cellular Morphological Descriptors and Molecular Fingerprints for the Prediction of Cytotoxicity- And Proliferation-Related Assays. *Chem. Res. Toxicol.* **2021**, *34* (2), 422–437.

(25) Seal, S.; Carreras-Puigvert, J.; Trapotsi, M.-A.; Yang, H.; Spjuth, O.; Bender, A. Integrating Cell Morphology with Gene Expression and Chemical Structure to Aid Mitochondrial Toxicity Detection. bioRxiv 2022, 2022.01.07.475326.

(26) Moshkov, N.; Moshkov, T.; Yang, K.; Horvath, P.; Dancik, V.; Wagner, B. K.; Clemons, P. A.; Singh, S.; Carpenter, A. E.; Caicedo, J. C. Predicting Compound Activity from Phenotypic Profiles and Chemical Structures. bioRxiv 2022, 2020.12.15.422887.

(27) Wilke, J.; Kawamura, T.; Xu, H.; Brause, A.; Friese, A.; Metz, M.; Schepmann, D.; Wünsch, B.; Artacho-Cordón, A.; Nieto, F. R.; Watanabe, N.; Osada, H.; Ziegler, S.; Waldmann, H. Discovery of a Σ1 Receptor Antagonist by Combination of Unbiased Cell Painting and Thermal Proteome Profiling. Cell Chem. Biol. **2021**, *28* (6) 848-854.e5.

(28) Allen, C. H. G.; Koutsoukas, A.; Cortés-Ciriano, I.; Murrell, D. S.; Malliavin, T. E.; Glen, R. C.; Bender, A. Improving the Prediction of Organism-Level Toxicity through Integration of Chemical, Protein Target and Cytotoxicity QHTS Data. *Toxicol. Res. (Camb).* **2016**, *5* (3), 883–894.

(29) Liu, R.; Wallqvist, A. Merging Applicability Domains for in Silico Assessment of Chemical Mutagenicity. *J. Chem. Inf. Model.* **2014**, *54* (3), 793–800.

(30) Fan, F.; Wood, K. V. Bioluminescent Assays for High-Throughput Screening. *Assay Drug Dev. Technol.* **2007**, *5* (1), 127–136.

(31) Cox, M. J.; Jaensch, S.; Van de Waeter, J.; Cougnaud, L.; Seynaeve, D.; Benalla, S.; Koo, S. J.; Van Den Wyngaert, I.; Neefs, J. M.; Malkov, D.; Bittremieux, M.; Steemans, M.; Peeters, P. J.; Wegner, J. K.; Ceulemans, H.; Gustin, E.; Chong, Y. T.; Göhlmann, H. W. H. Tales of 1,008 Small Molecules: Phenomic Profiling through Live-Cell Imaging in a Panel of Reporter Cell Lines. *Sci. Rep.* **2020**, *10* (1), 1–14.

(32) JUMP-Cell Painting Consortium. https://jump-cellpainting.broadinstitute.org/.(accessed May 2, 2022)

(33) Hofmarcher, M.; Rumetshofer, E.; Clevert, D. A.; Hochreiter, S.; Klambauer, G. Accurate Prediction of Biological Assays with High-Throughput Microscopy Images and Convolutional Networks. *J. Chem. Inf. Model*. 2019, *59* (3), 1163–1171.

(34) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945– D954.

(35) Vollmers, L. Prediction of Cytotoxicity Related PubChem Assays Using High-Content-Imaging Descriptors derived from Cell-Painting [Unpublished master's thesis]. **2021**, TU Darmstadt.

(36) PubChem https://pubchem.ncbi.nlm.nih.gov/ (accessed Jun 4, 2022).

(37) Szklarczyk, D.; Gable, A. L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N. T.; Morris, J. H.; Bork, P.; Jensen, L. J.; Von Mering, C. STRING V11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets. *Nucleic Acids Res.* **2019**, *47* (D1), D607–D613.

(38) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13* (11), 2498–2504.

(39) Bindea, G.; Mlecnik, B.; Hackl, H.; Charoentong, P.; Tosolini, M.; Kirilovsky, A.; Fridman, W. H.; Pagès, F.; Trajanoski, Z.; Galon, J. ClueGO: A Cytoscape Plug-in to Decipher Functionally Grouped Gene Ontology and Pathway Annotation Networks. *Bioinformatics* **2009**, *25* (8), 1091–1093.

(40) Bray, M. A.; Gustafsdottir, S. M.; Rohban, M. H.; Singh, S.; Ljosa, V.; Sokolnicki, K. L.; Bittker, J. A.; Bodycombe, N. E.; Dančík, V.; Hasaka, T. P.; Hon, C. S.; Kemp, M. M.; Li, K.; Walpita, D.; Wawer, M. J.; Golub, T. R.; Schreiber, S. L.; Clemons, P. A.; Shamji, A. F.; Carpenter, A. E. A Dataset of Images and Morphological Profiles of 30 000 Small-Molecule Treatments Using the Cell Painting Assay. GigaScience. **2017**, pp 1–5.

(41) GigaDB Dataset - DOI 10.5524/100351 - Supporting data for "A Dataset of Images and Morphological Profiles of 30 000 Small-Molecule Treatments Using the Cell Painting Assay" http://gigadb.org/dataset/100351 (accessed Jul 22, 2020).

(42) Swain, M. MolVS: Molecule Validation and Standardization https://molvs.readthedocs.io/en/latest/ (accessed Apr 15, 2021).

(43) RDKit: Open-source cheminformatics. https://rdkit.org/ (accessed Mar 2, 2022).

(44) Way, G. P. Blocklist Features - Cell Profiler https://figshare.com/articles/dataset/Blacklist_Features_-_Cell_Profiler/10255811 (accessed Apr 11, 2021).

(45) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. (2011) Scikit-learn: machine learning in Python. *J Mach Learning Res* ,12, 2825–2830

(46) PyCytominer: cytomining/pycytominer: cytominer python package https://github.com/cytomining/pycytominer (accessed Jun 4, 2022).

(47) Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.

(48) Fluss, R.; Faraggi, D.; Reiser, B. Estimation of the Youden Index and Its Associated Cutoff Point. Biometrical J. 2005, 47 (4), 458–472.

(49) API reference — pandas 1.3.1 documentation https://pandas.pydata.org/pandas-docs/stable/reference/index.html (accessed Jul 29, 2021).

FIGURES

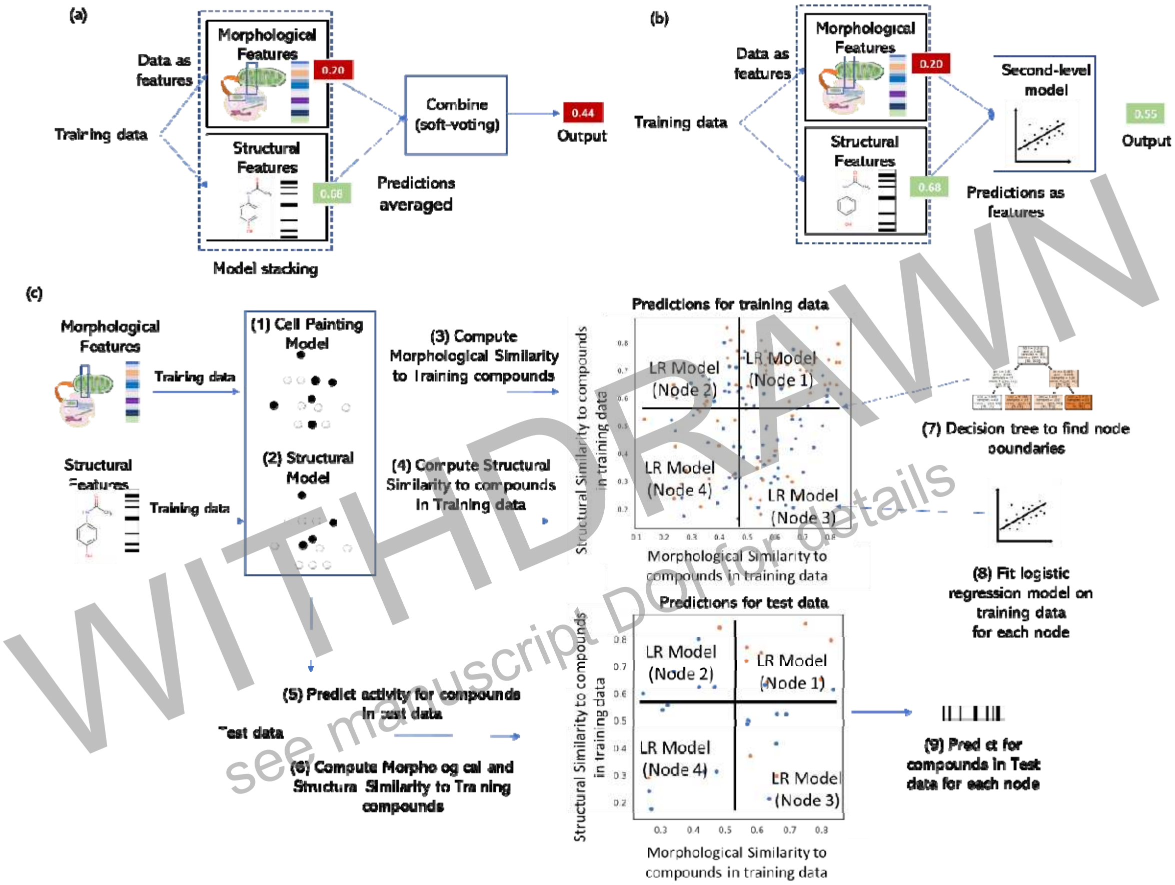
(a)
Data as features
Training data
Morphological Features
0.20
Structural Features
0.68
Model stacking
Combine (soft-voting)
Predictions averaged
0.44
Output
(b)
Data as features
Training data
Morphological Features
0.20
Structural Features
0.68
Second-level model
Predictions as features
0.55
Output
(c)
Morphological Features
Training data
Structural Features
Training data
(1) Cell Painting Model
(2) Structural Model
(3) Compute Morphological Similarity to training compounds
(4) Compute Structural Similarity to compounds in Training data
Predictions for training data
LR Model (Node 2)
LR Model (Node 1)
LR Model (Node 4)
LR Model (Node 3)
Structural Similarity to compounds in training data
Morphological Similarity to compounds in training data
(7) Decision tree to find node boundaries
(8) Fit logistic regression model on training data for each node
Predictions for test data
LR Model (Node 2)
LR Model (Node 1)
LR Model (Node 4)
LR Model (Node 3)
Structural Similarity to compounds in training data
Morphological Similarity to compounds in training data
(5) Predict activity for compounds in test data
Test data
(6) Compute Morphological and Structural Similarity to Training compounds
(9) Predict for compounds in Test data for each node

Figure 1: Schematic Representation of workflow in this study to build (a) hierarchical models where the predictions of the individual models are used as features to build a second-level logistic regression model, and (b) soft-voting ensemble models that compute the mean of predicted probabilities from individual models and (c) the similarity-based merger model. The similarity-based merger model combined predictions from individual models by weighting them in proportion to their similarity to training data in morphology and structural space. These weights are computed by different logistic regression models on the predictions from out-of-fold cross-validation compounds in the training data in different regions of the morphology and structural space.

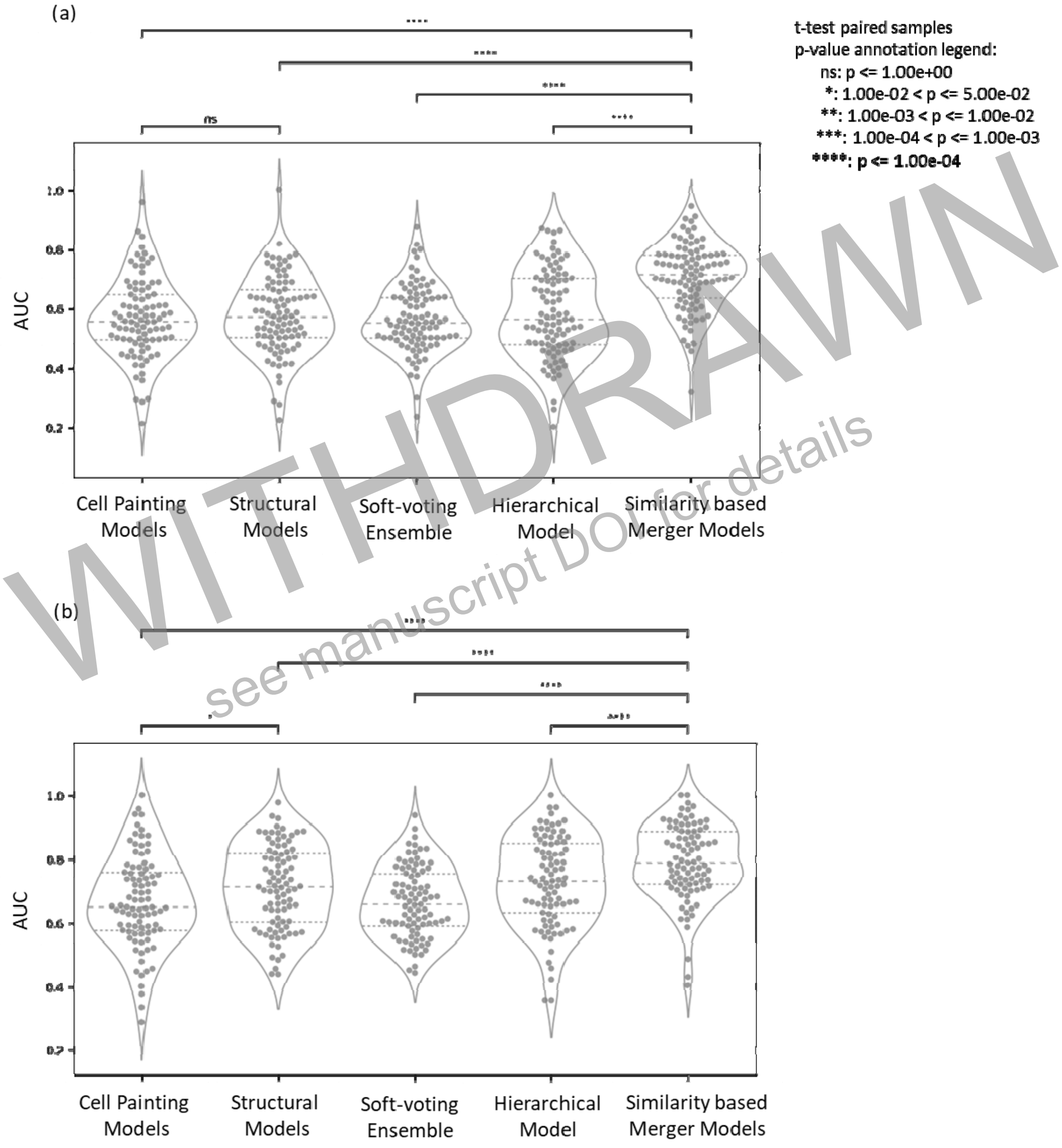

Figure 2: Distribution of AUC of all models, Cell Painting, Morgan Fingerprints, baseline models of a soft-voting ensemble, hierarchical model, and the similarity-based merger model, over (a) the public dataset comprising 92 assays and (b) Broad Institute dataset comprising 89 assays.

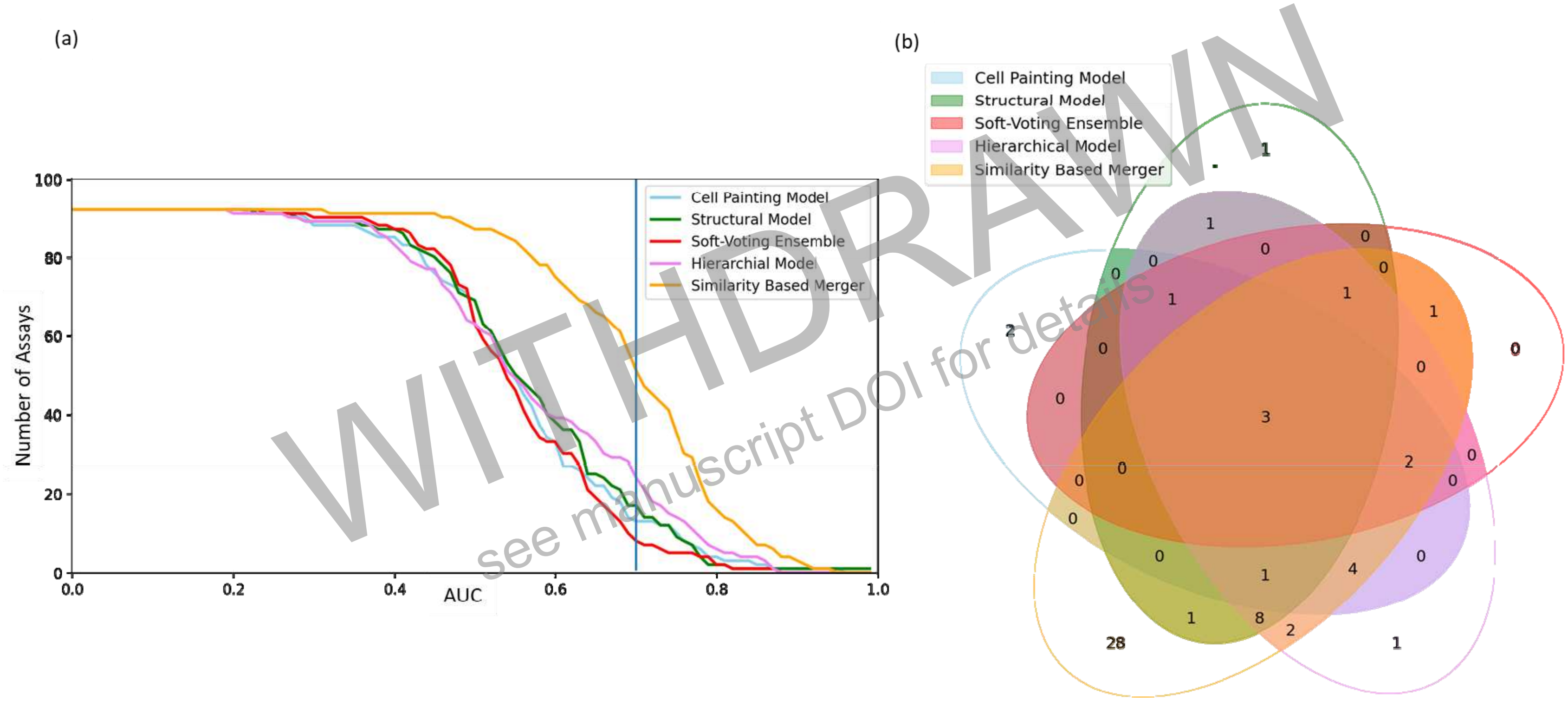

Figure 3: (a) Number of assays predicted with an AUC above a given threshold. (b) Distribution of assays with AUC > 0.70 common and unique to all models, Cell Painting, Morgan Fingerprint, baseline models of a soft-voting ensemble, hierarchical model, and the similarity-based merger model, over the public dataset comprising 92 assays.

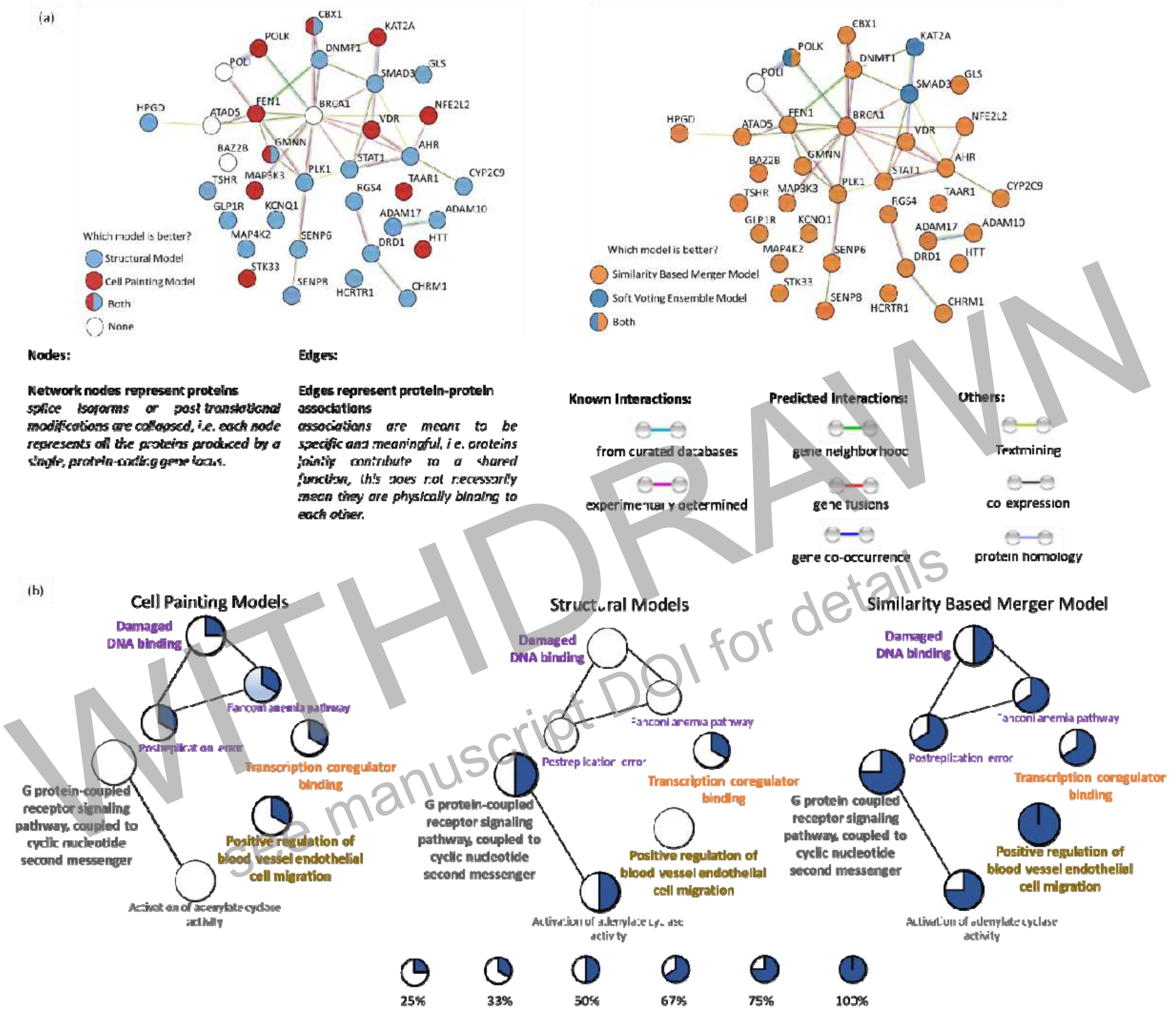
(a)
CBX1
KAT2A
POLK
DNMT1
POLI
GLS
SMAD3
HPGD
ATAD5
FEN1
BRCA1
NFE2L2
VDR
BAZ2B
GMNN
AHR
STAT1
TSHR
MAP3K3
PLK1
TAAR1
CYP2C9
GLP1R
KCNQ1
RGS4
ADAM17
ADAM10
MAP4K2
SENP6
DRD1
HTT
STK33
SENP8
HCRTR1
CHRM1
Which model is better?
Structural Model
Cell Painting Model
Both
None
Similarity Based Merger Model
Soft Voting Ensemble Model
Both
Nodes:
Network nodes represent proteins
splice isoforms or post translational modifications are collapsed, i.e. each node represents all the proteins produced by a single, protein-coding gene locus.
Edges:
Edges represent protein-protein associations
associations are meant to be specific and meaningful, i.e. proteins jointly contribute to a shared function; this does not necessarily mean they are physically binding to each other.
Known Interactions:
from curated databases
experimentally determined
Predicted Interactions:
gene neighborhood
gene fusions
gene co-occurrence
Others:
Textmining
co expression
protein homology
(b)
Cell Painting Models
Structural Models
Similarity Based Merger Model
Damaged DNA binding
Fanconi anemia pathway
Postreplication error
Transcription coregulator binding
G protein-coupled receptor signaling pathway, coupled to cyclic nucleotide second messenger
Positive regulation of blood vessel endothelial cell migration
Activation of adenylate cyclase activity
25%
33%
50%
67%
75%
100%

Figure 4: (a) STRING gene-gene interaction networks for 35 Genes annotations associated with 38 assays in the public dataset labelled by the model which was better predictive compared to the other models and a random classifier with an AUC>0.50 (b) Molecular and functional pathway terms related to the 38 assays using the Cytoscape[38] v3.9.1 plugin ClueGO[39] labelled by percentage of gene annotations where an AUC>0.70 was achieved by the Cell Painting, structural and similarity-based merger models.

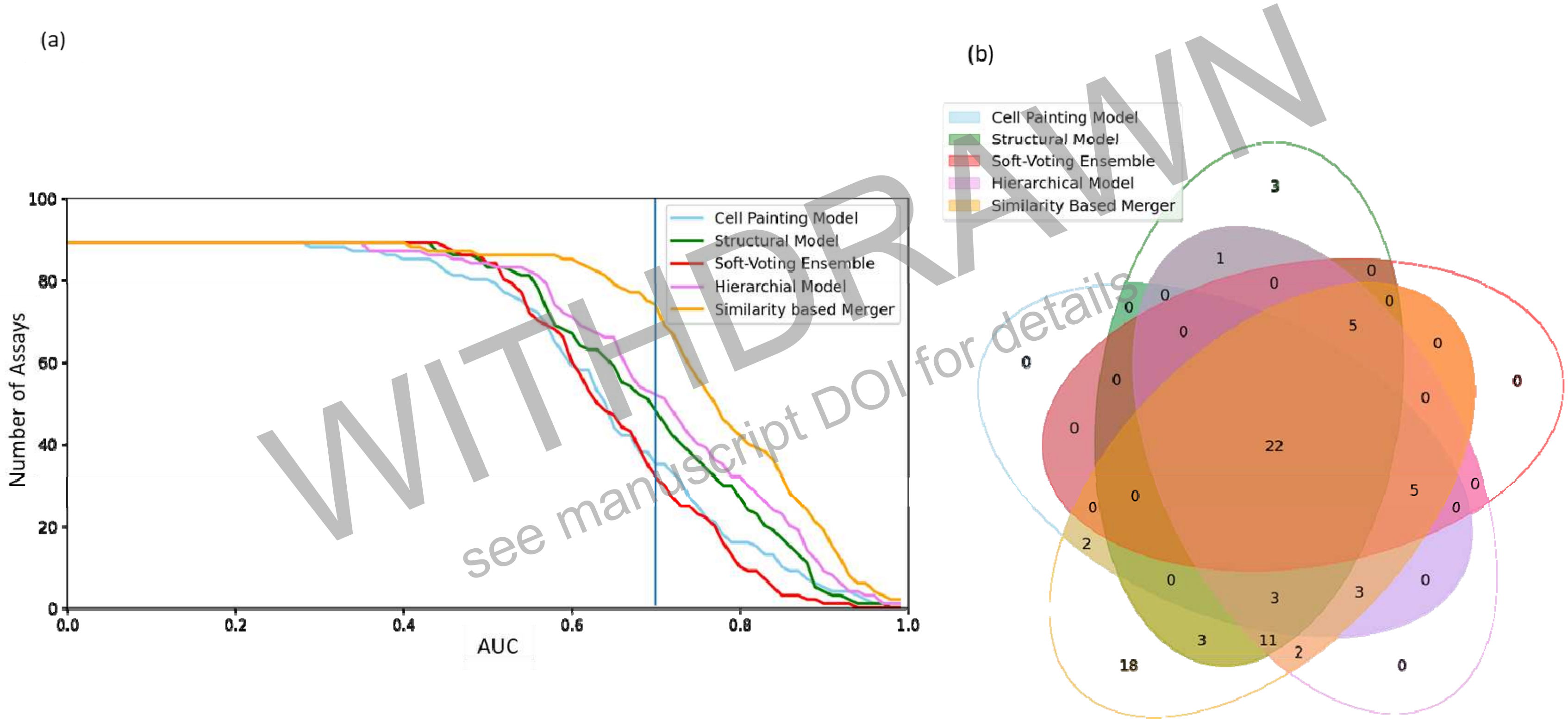

Figure 5: (a) Number of assays predicted with an AUC above a given threshold. (b) Distribution of assays with AUC > 0.70 common and unique to all models, Cell Painting, Morgan Fingerprint, baseline models of the soft-voting ensemble, hierarchical model, and the similarity-based merger model, over the Broad Institute dataset comprising 89 assays.

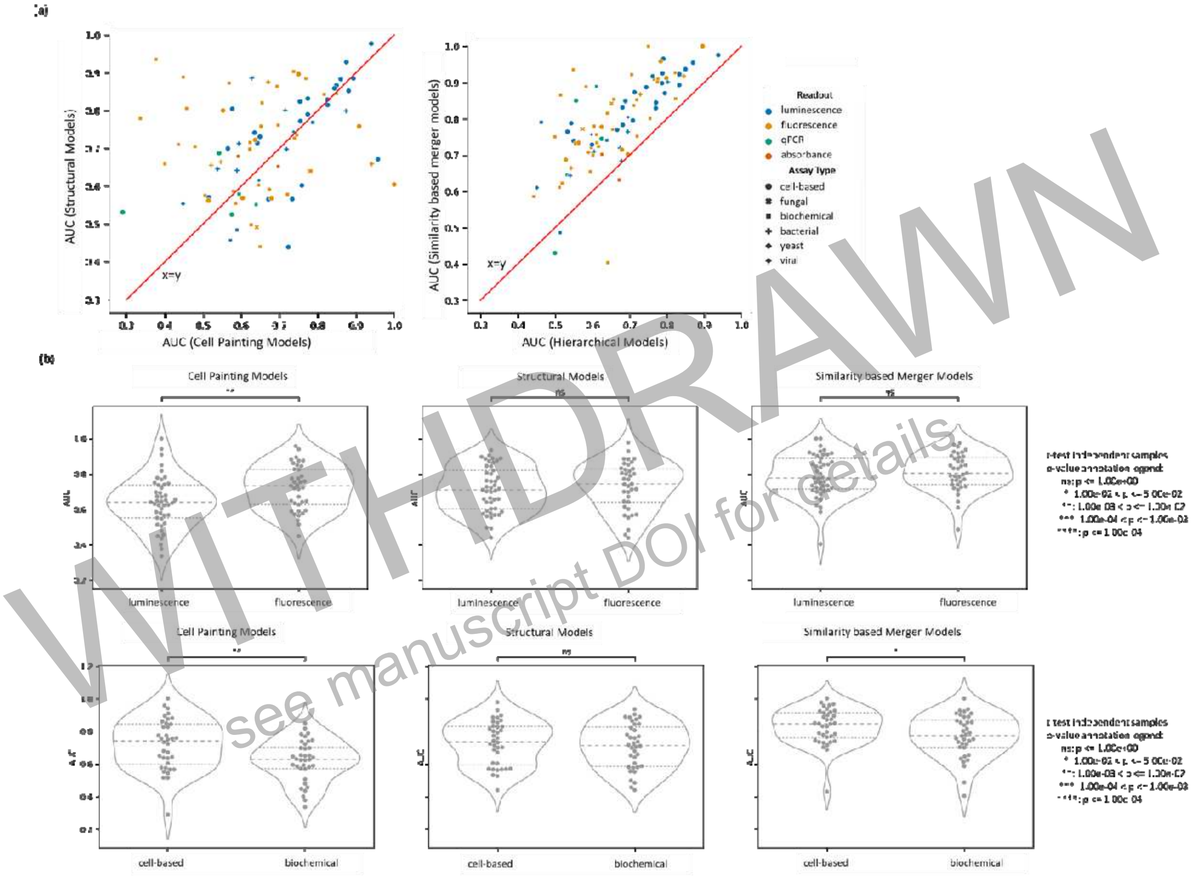

Figure 6: AUC performance of models using Cell Painting, structural models, and similarity-based merger model for 89 assays in the Broad Institute dataset based on readout type (fluorescence and luminescence) or the assay type (cell-based and biochemical). Further details are shown in Figures S14 and S15.