bioRxiv preprint doi: https://doi.org/10.1101/2020.12.15.422887; this version posted December 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Predicting compound activity from phenotypic profiles and chemical structures

Authors

Tim Becker*, Kevin Yang*, Juan C. Caicedo*, Bridget K. Wagner, Vlado Dancik, Paul Clemons, Shantanu Singh, Anne E. Carpenter
* These authors contributed equally

Abstract

Recent advances in deep learning enable using chemical structures and phenotypic profiles to accurately predict assay results for compounds virtually, reducing the time and cost of screens in the drug discovery process. The relative strength of high-throughput data sources - chemical structures, images (Cell Painting), and gene expression profiles (L1000) - has been unknown. Here we compare their ability to predict the activity of compounds structurally different from those used in training, using a sparse dataset of 16,979 chemicals tested in 376 assays for a total of 542,648 readouts. Deep learning-based feature extraction from chemical structures provided a remarkable ability to predict assay activity for structures dissimilar to those used for training. Image-based profiling performed even better, but requires wet lab experimentation. It outperformed gene expression profiling, and at lower cost. Furthermore, the three profiling modalities are complementary, and together can predict a wide range of diverse bioactivity, including cell-based and biochemical assays. Our study shows that, for many assays, predicting compound activity from phenotypic profiles and chemical structures is an accurate and efficient way to identify potential treatments in the early stages of the drug discovery process.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.15.422887; this version posted December 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

Drug discovery is very expensive and slow. To identify a promising treatment for specific disease conditions, the theoretical landscape of possible chemical structures is prohibitively large to test in physical experiments. Pharmaceutical companies synthesize and test many millions of compounds, yet even these represent a small fraction of possible structures. Furthermore, although complex phenotypic assay systems have proven extremely valuable for identifying useful drugs for diseases where an appropriate protein target is unknown [1–3], their reliance on expensive or limited-supply biological materials, such as antibodies or human primary cells, often hinders their scalability.

What if computational models could predict the results of hundreds of expensive assays across millions of compounds at a fraction of the cost? How might this transform drug discovery? Predictive modeling shows some promise. Most attempts so far have used various representations of chemical structure alone to predict assay activity; this requires no laboratory experiments for the compounds to be predicted (neither to synthesize nor test them), so this is dramatically cheaper than physical screens and enables a huge search space. Graph convolutional architectures in particular have substantially advanced the state of the art in recent years [4–15], and were recently used to discover a novel antibiotic [16]. As impressive as these capabilities are, chemical structures do not seem to contain enough information to predict all assay readouts - their performance may depend heavily on the quantity, quality, and diversity of the given training data. Augmenting graph convolutional approaches with automatically computable descriptor sets was recently shown to improve performance in limited-data settings [15]. However, the realm of such descriptors is somewhat limited.

Considerable improvements might come from augmenting chemical structure-based features with experimental information associated with each small molecule, ideally information available in a single inexpensive, scalable assay that could be run on millions of compounds once, then used to predict assay results virtually for hundreds of individual assays. Most profiling techniques, such as those measuring a subset of the proteome or metabolome, are not scalable to millions of compounds. One exception is transcriptomic profiling by the L1000 assay [17], which has shown success for mechanism of action (MOA) prediction [18], but is untested for predicting assay outcomes.

Image-based profiling is an even less expensive high-throughput profiling technique [19] that has recently shown great success in compound activity prediction. In a landmark study, Simm et al. [20] successfully repurposed images from a compound library screen to train models to predict unrelated assays; their prospective tests yielded up to 250-fold increased hit rates while also improving structural diversity of the active compounds. More recently, Hofmarcher et al. [21] and Way et al. [22] used Cell Painting [23], an unbiased image-based profiling protocol, to predict assay outcomes using machine learning, obtaining excellent results as well. Other studies have also looked at combinations of profiling methodologies, such as the work of Trapotsi et al. [24] who aimed to combine imaging and chemical structures to complete assay readouts in a sparse matrix and Lapins and Spjuth [18] who combined L1000 and Cell Painting for MOA prediction.

Here, we test the hypothesis that computational models can powerfully predict assay outcomes at large scale when trained on advanced chemical structure representations combined with two different types of experimentally-produced phenotypic profiles, imaging (Cell Painting assay) and gene expression (L1000 assay) (Figure 1).

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.15.422887; this version posted December 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

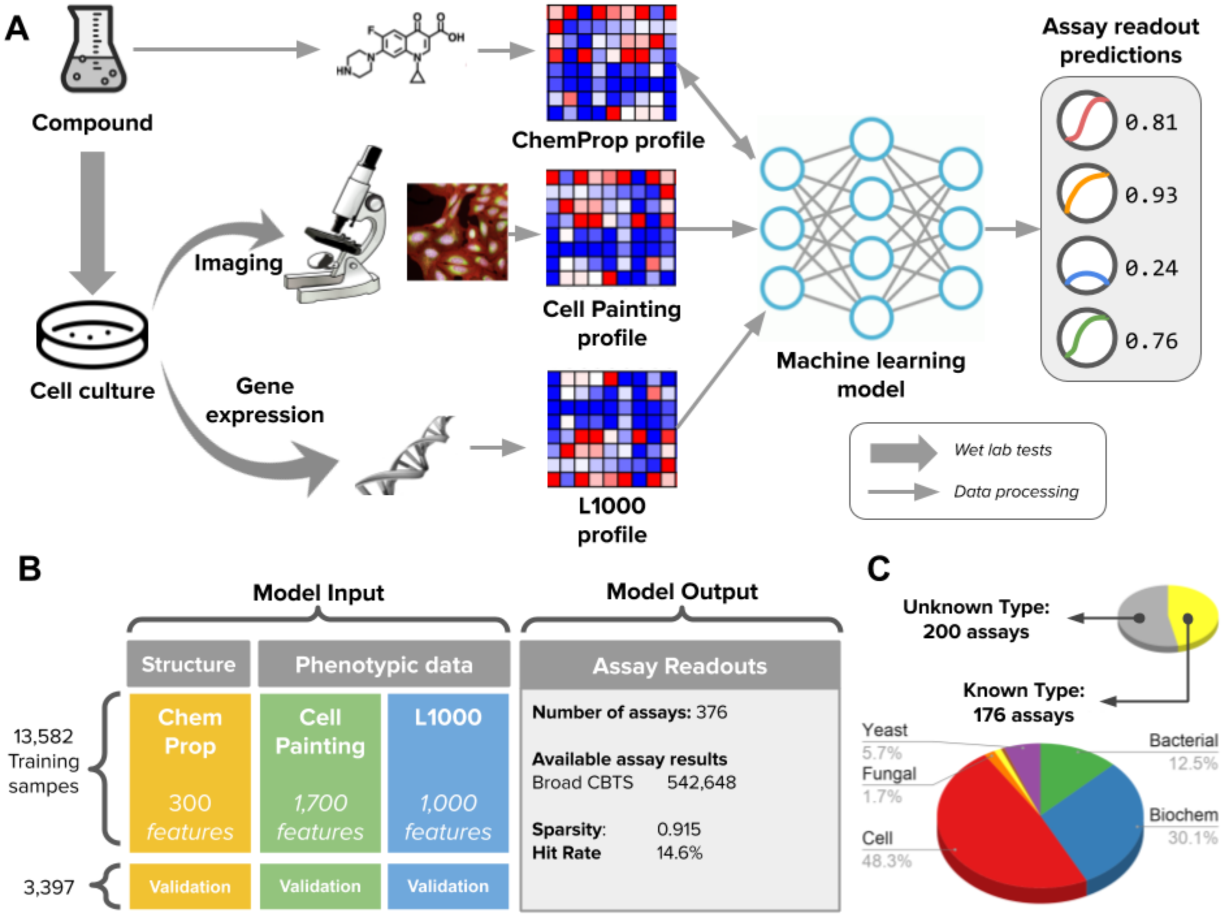


Figure 1. *Overview of the workflow and data. A) Workflow of the methodology for predicting diverse assays from perturbation experiments. B) Structure of the data used in this study. C) Types of assay readouts targeted for prediction.*

Results

Assay predictors trained with phenotypic profiles can improve hit rates

We extracted experiment-derived profiles from two high-dimensional assays for each of 16,979 compounds, including gene expression data (GE) from the L1000 assay [25,26] and image-based morphological profiles (MO) from the Cell Painting assay [26,27] (see Figure 1B and Methods). We also computed a chemical structure profile (CS) using Chemprop [15]. We trained predictors using a subset of the data (13,582 of the compounds), for each kind of profile individually and in combinations (using late data fusion, see Methods).

We evaluated each predictor for its ability to predict quantitative outcomes for 3,397 held-out compounds in 376 assays performed at the Broad Institute for more than a decade; the assays

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.15.422887; this version posted December 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

were not selected based on any metadata and thus are representative of the activity of an academic screening center. We report the mean Area Under the Receiver Operating Characteristic Curve (AUC) and the number of assays with AUC > 0.9 (termed "well predicted"). We chose the latter as our primary metric, as in past studies of assay prediction [20], because it best matches the real-world use case where a subset of predicted compounds will be chosen and tested in an experiment. Predictors meeting AUC > 0.9 in our experiments produce on average a 25-fold enrichment of hits (compounds with the desired activity) for assays with a baseline hit rate below 1%, although we note that assays with AUC between 0.7 and 0.9 can in some cases show sufficient predictive ability to be useful (Supplementary Figure 1), and thus we report the number of assays with AUC > 0.7 as well.

It is important to note that we aimed here to test the ability of each data type to predict assays for chemical structures that are *distinct* relative to training data. This is because there is little practical value to screen for additional, similar structures (scaffolds) to compounds already known to have activity; in drug discovery, any compounds with positive activity undergo medicinal chemistry where small variations in structure are synthesized and tested to optimize the molecule. In other words, biological assay outcomes are the primary goal and chemical dissimilarity is a secondary goal. We therefore took care that similar classes of structures were not included in both the training and held-out set to test whether methods (see Methods).

Morphological profiles are the best individual modality for assay prediction

Taking each data type individually, morphological profiling was able to accurately predict the highest number of different assays, 67 out of 376 attempted (Figure 2A). For comparison, in a prior study, morphology information alone was found to accurately predict 66 of 209 assays [21], though the experiments are not directly comparable because the sets of assays likely differ in their characteristics, and the data partitions for training / validation may not be based on scaffold diversity as in our study.

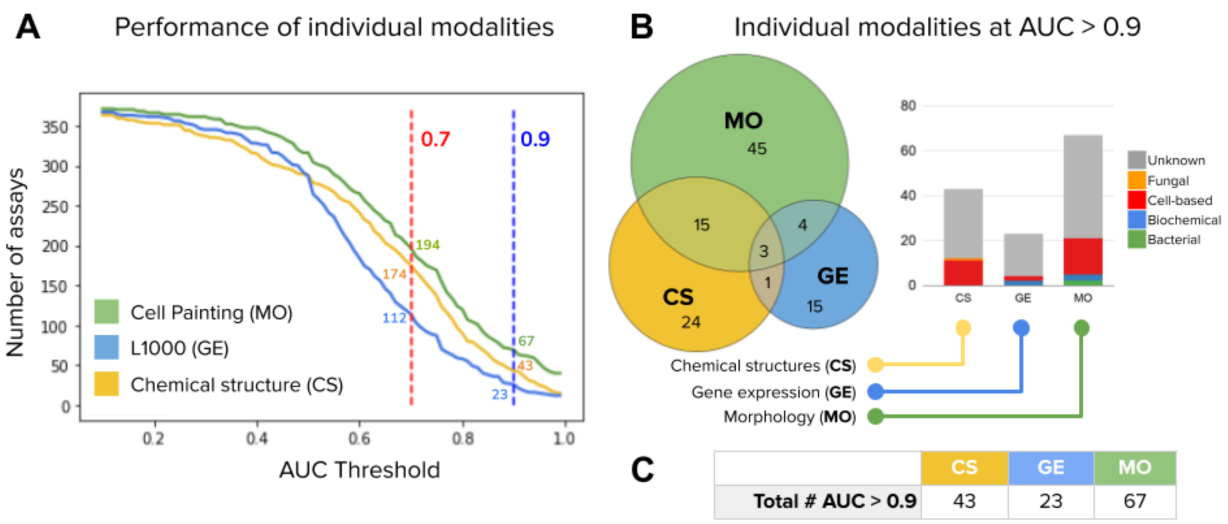
By contrast, chemical structures could predict 43 assays and gene expression profiles 23 assays. We strongly note that chemical structures would perform better without our imposed limitation to find structures different from those already known to be hits in each assay, as described in the prior section; it is remarkable that chemical structure information performs so well to find new structures.

Chemical structure, morphology, and gene expression profiles provide complementary information for prediction

We found a lack of major overlap among assays predicted by each profiling modality alone (Figure 2B). This indicates significant complementarity, that is, each profiling modality captures different biologically relevant information. In fact, only three of the 376 assays (<1%) "overlapped" - that is, were accurately predicted using any of the three profiling modalities alone. The two best-performing modalities, chemical structure and morphology, only have 18 well-predicted assays in common. Taken independently, the three profiling modalities can identify a total of 107 unique assays, far higher than the best individual modality (MO, at 67).

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.15.422887; this version posted December 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Ideally, a combined strategy should recover all of those and more, by productively integrating the data, as we explore next.



	CS	GE	MO
Total # AUC > 0.9	43	23	67

Figure 2. *Number of assays that can be accurately predicted using single profiling modalities. A) Performance of individual modalities measured as the number of assays (y axis) predicted with AUC above a certain threshold (x axis). When a higher AUC threshold is needed, the number of assays that can be predicted decreases for all profiling modalities. We define accurate assays as those with AUC greater than 0.9 (dashed vertical line in blue). B) The Venn diagrams on the right show the number of accurate assays (AUC > 0.9) that are in common or unique to each profiling technique. The bar plot shows the distribution of assay types correctly predicted by single profiling modalities. All metrics and counts are measured in the holdout set using cross validation. C) Number of assays well predicted (AUC > 0.9) by each individual modality.*

Adding morphological profiles to chemical structure information improves assay prediction ability

To understand how many assays can be predicted across data modalities, we counted the number of unique assays predicted by any of the individual profiling modalities following a retrospective assessment (Figure 3C, row “Single”). Given that all three modalities showed some amount of non-overlap with the others, having more modalities would always produce an increased number of well-predicted assays, as compared to having fewer.

Morphology and gene expression profiles require wet lab experimentation, whereas chemical structures are always available, even for theoretical compounds, with the only cost being computing their profiles. Therefore we took CS as a baseline and explored the value of adding other kinds of profiles to it. In retrospection, adding MO to CS indicates potential to predict more than double the number of unique assays compared to CS alone (92 vs 43; Figure 3C, row “Single” vs Figure 2C)).

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.15.422887; this version posted December 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

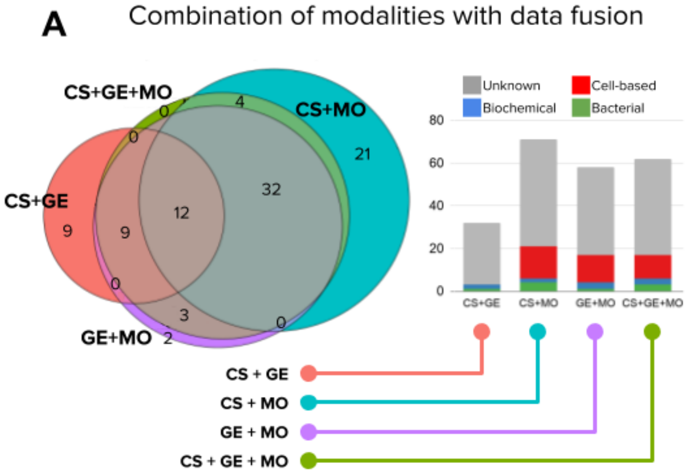
Training assay predictors from each profiling modality separately then choosing the most-predictive modality for each assay is simplistic and does not leverage their strengths and complementarity. We therefore sought an improved strategy to integrate data from different profiling methods. In practice, data fusion strategies have to balance trade-offs between preserving accuracy and introducing noise from the various data sources (Supplementary Figure 2). We therefore carried out an analysis to combine results from modalities in alternate ways. Compared to early data fusion, we found that late data fusion (see Methods) provided better performance, yielding more predictors with AUC > 0.9 for all combinations of data types (Supplementary Table 1).

Using late data fusion (Figure 3B, row “# AUC > 0.9”), we found that adding morphological profiles to chemical structures yields 71 well-predicted assays (CS+MO) as compared to 43 assays for CS alone. In retrospection, there are 24 unique assays that are well predicted using CS or MO alone that could not be captured by the data fusion model. Adding them to the list would yield 95 well-predicted assays total (Figure 3C, row “Plus fusion”), indicating potential to predict more than twice the assays compared to CS alone (43). Improvements when adding morphology profiles to chemical structures were consistently found across other evaluation metrics (mean AUC in Figure 3B and AUC > 0.7 in Supplementary Figure 4 and Supplementary Table 1) and when adding morphological profiles to all other data types and combinations (Figure 3C).

Obtaining morphological profiles requires physical experiments, whereas chemical structure-based predictions can be carried out completely virtually. Nevertheless, we conclude the biological information encoded in the inexpensive Cell Painting assay makes it worthwhile to profile the large compound libraries in both pharmaceutical companies and academia. Running this single assay costs about the same as a typical screening assay but would reduce the number of subsequent physical screens needed (at a cost of millions of dollars each in a pharmaceutical setting).

At an AUC > 0.9, the 95 unique assays that are well predicted with CS+MO in retrospection represent 25% of the total. It is currently debated whether an AUC closer to 0.7 would be acceptable; we found that for assays with a low baseline hit rate, this accuracy level may be sufficient to dramatically increase the ability to identify useful compounds in the screen (Supplementary Figure 1). If a cutoff of AUC > 0.7 was found to be acceptable, 67% of assays would be well predicted with CS+MO (253 out of 376, Supplementary Figure 4). This is remarkable given that the imaging assay uses only six dyes for cell structure components in a single cell type (U2OS), captured at a single time point and concentration of drug exposure.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.15.422887; this version posted December 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



B Performance of data fusion by cross-validation (predictive)

	CS+MO	CS+GE	GE+MO	CS+GE+MO
# AUC > 0.9	71	32	58	62

C Retrospective counts of predictors with AUC > 0.9 (no cross-validation)

	CS+MO	CS+GE	GE+MO	CS+GE+MO
Single	92	62	83	107
Plus fusion	95	64	91	115
Increase	3%	3%	10%	7%

Figure 3. *Number of assays that can be accurately predicted using combinations of profiling modalities. Accurate predictors are defined as models with accuracy greater than 0.9 AUC. We considered all four modality combinations using late data fusion in this analysis: CS+MO (chemical structures and morphology), CS+GE (chemical structures and gene expression), GE+MO (gene expression and morphology), and CS+GE+MO (all three modalities). A) The Venn diagram shows the number of accurate assays that are in common or unique to fused data modalities. The bar plots in the center show the distribution of assay types correctly predicted by the fused models. All counts are measured in the holdout set. B) The number of accurate assay predictors (AUC > 0.9) obtained for combinations of modalities (columns) using late data fusion following predictive cross-validation experiments. C) Retrospective performance of predictors. These counts indicate how many assays can be predicted with high accuracy (AUC > 0.9), whether by single or fused modalities. "Single" is the number of assays reaching AUC > 0.9 with any one of the specified modalities, i.e., take the best single-modality predictor for an assay. This count corresponds to the simple union of circles in the Venn diagram in Figure 2B, i.e., no data fusion is involved. "Plus fusion" is the same, except that it displays the number of assays that reach AUC > 0.9 with any individual or data-fused combination. This count corresponds to the union of circles in the Venn diagram in Figure 2B plus the number of additional assays that reach AUC > 0.9 when the modalities are fused. For example, the last column counts an assay if its AUC > 0.9 for any of the following: CS alone, GE alone, MO alone, data-fused CS+GE, data-fused GE+MO, data-fused CS+MO, and data-fused CS+GE+MO. "Increase" indicates the relative increase in the number of assays that can be accurately predicted when using data fusion (calculated from the two rows above it).*

Adding gene expression profiles to chemical structure information improves assay prediction ability in some cases

Gene expression profiles were the weakest profile type when tested individually: chemical structures alone were twice as powerful and morphological profiles alone were three times as powerful (23 well-predicted assays for GE compared to 43 for CS and 67 for MO; Figure 2C). In retrospection, the unique assays predicted by either CS or GE amounts to 62 in total (Figure 3C): GE would add 19 unique assays to CS's 43 well-predicted assays, a 50% improvement. Surprisingly, adding gene expression profiles to chemical structures by late data fusion actually worsened performance, yielding only 32 well-predicted assays as compared to 43 for CS alone (Figure 3B). This worsening was quite consistent across other evaluation metrics when adding gene expression profiles to all other profiling modalities and combinations by data fusion (Figure 3 and Supplementary Table 1 / Supplementary Figure 4).

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.15.422887; this version posted December 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Nevertheless, data fusion did yield some additional well-predicted assays in retrospection; adding the unique assays well predicted by CS alone and GE alone to those well predicted by the data fusion of CS+GE would add two assays, bringing the total to 64. Therefore, comparing the profiling methods, MO would yield 2.5 times as many additional unique assays when added to CS (95 assays total) as compared to GE being added to CS (64 assays total), that is, MO would add 52 whereas GE would add 21. MO is also more cost-effective, making it the better choice when planning a single profiling experiment.

Complementarity across all three profiling types

We had hypothesized that data fusion of all three modalities would provide the best assay prediction ability than any individual or subset. However, data-fused CS+GE+MO yielded 62 well-predicted assays, fewer than could be obtained by data-fused CS+MO (71 assays), which itself was not far from MO alone (67 assays). All of these fall short of the 92 unique assays that, in retrospection, could be identified by taking the single best of just two of the data types, CS alone + MO alone. This highlights the need for designing improved strategies for data fusion.

In combining all three profiling modalities, therefore, we trained predictors using all possible individual modalities as well as pairwise data-fused combinations, plus a data fusion of all three modalities. In retrospection, taking the best (single or fused) of all three modalities would predict 115 unique assays, as compared to CS+MO and GE+MO, which would predict 95 and 91 assays respectively. We therefore conclude that if all modalities are available, they are all useful to increase predictive ability. We also conclude that morphological profiles are more valuable than gene expression profiles in this context, keeping in mind the caveat that the particular assays here were Cell Painting for morphology and L1000 for gene expression.

Models can predict a diversity of assay types

The morphological and gene expression profiles used for model training derive from cell-based profiling assays. We find that they can correctly predict compound activity for mammalian cell-based assays, which were the most frequent in this study (Figure 1C), but also a variety of other assay types, such as bacterial and biochemical (Figure 2B, 3A). We obtained assay type annotations for 176 assays in 7 categories, and obtained diverse results: from 85 cell-based assays, 11, 2 and 16 are accurately predicted by CS, GE and MO respectively (12%, 2%, 18%). From 53 biochemical assays, 0, 2 and 3 were predicted by CS, GE and MO respectively (0%, 3%, 5%). These results suggest that the subset of well predicted assays include diverse assay types, i.e., phenotypic profiling strategies are not constrained to predict cell-based assays only. The results also indicate, together with those shown in Figure 2B, that MO captures the widest range of assay types among all predictors.

Interestingly, most assays can be predicted with a single profiling modality, while some others benefit from combining experimental evidence of various profiling modalities. For example, MO accurately predicts two bacterial assays, and fusing CS+MO predicts four (out of 22 available). We examined four assays with increased fused accuracy more closely (Figure 4). The *Fibroblast*

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.15.422887; this version posted December 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

growth factor 22 (FGF-22) assay, a biochemical assay, can be predicted with an AUC of 0.83 and 0.86 using MO and CS respectively, but when both are combined using data fusion the prediction accuracy increases significantly to 0.97 (Figure 4). The other assays showed less dramatic improvement in AUC when combining data modalities by data fusion.

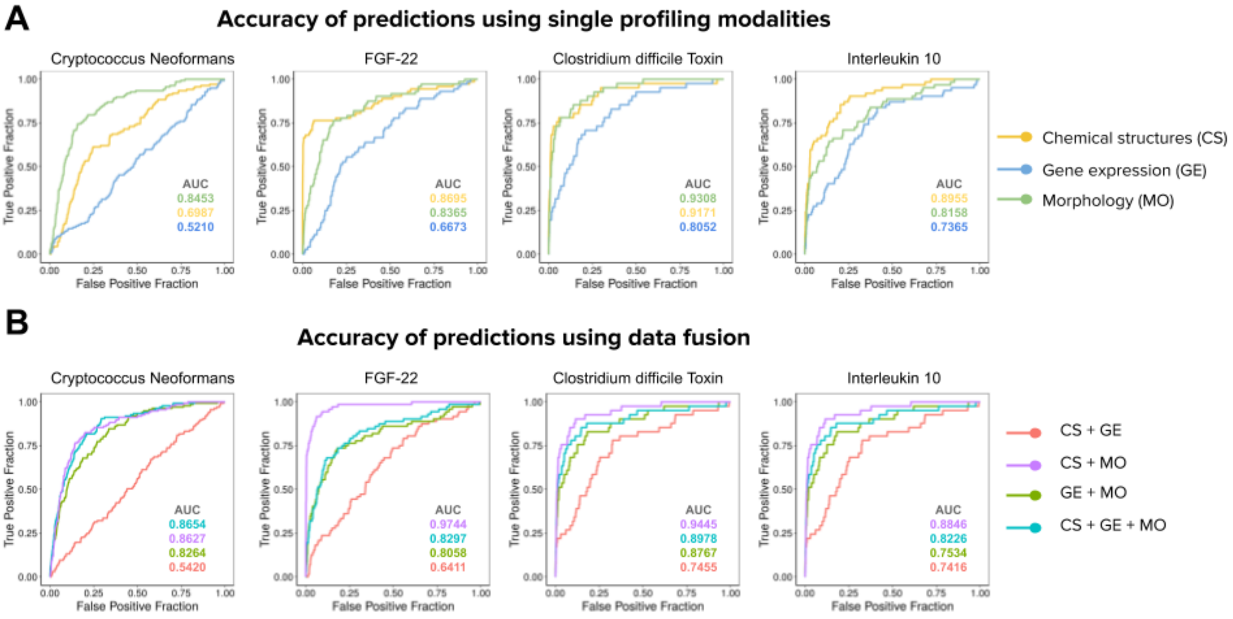


Figure 4*. Prediction accuracy of example assays of diverse types. The plots are Receiver Operating Characteristic (ROC) curves and the Area Under the Curve (AUC) is reported for each curve with the corresponding color. Four assays were selected for panels A and B, in order from left to right: Cryptococcus Neoformans (bacterial), FGF-22 (biochemical), Clostridium difficile Toxin (bacterial), Interleukin 10 (cell-based). A) Performance of predictors when using a single profiling method. B) Performance of predictors when using combinations of profiling methods.*

Discussion

Predicting bioactivity of compounds could become a powerful strategy for drug discovery in light of ever-improving computational methods (particularly, deep learning) and ever-increasing rich data sources (particularly, from profiling assays). Here, we used Chemprop, state-of-the-art software for learning predictors from chemical structures, to combine the molecule fingerprint with phenotypic profiles obtained from images (Cell Painting) and gene expression (L1000).

We discovered that all three profile modalities—chemical structure, morphology, and gene expression—offer independently useful information about perturbed cell states that enables predicting different assays. Chemical structure is always readily available for a given compound. Of the two profiling modalities that require physical experimentation, morphology was significantly more powerful in providing information about cell state than gene expression, at

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.15.422887; this version posted December 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

least using this data set and profiling assay types: Cell Painting and L1000. This finding is consistent with prior studies for other applications [18,26].

In retrospection, we found that data fusion strategies increased the number of well-predicted assays by only 3-10%, depending on the subset of modalities tested, as compared to simply using each profiling modality independently for prediction. We believe this argues for further research on how best to integrate disparate profiling modalities, capturing the strengths of each individually as well as the complementarity of their combinations. Nevertheless, using late data fusion to combine each subset of available modalities does offer some improvement versus each individually and is likely a worthwhile exercise given its ease of implementation. Our set of assays lacked sufficient metadata to assess the characteristics of the assays that were only captured by fusion; this would be interesting to explore.

We believe these findings support widespread adoption of morphological profiling early in the drug discovery and chemical biology process. Given the low cost of carrying out Cell Painting, it is practical in many settings to profile an entire institution’s chemical library. Then, a modest-sized library of a few thousand compounds would be tested in each new assay of interest. Researchers would assess whether a sufficiently accurate predictor could be trained on this data, using CS alone, MO alone, or a data-fused combination of CS+MO. Taking into account the baseline hit rate for the assay, researchers could decide whether the predictor will increase the hit rate sufficiently to warrant a virtual screen against a large compound library for which morphological profiles are already available (within an institution, or publicly available profiles [28]), followed by cherry picking a small set of predicted hits and testing them for actual activity in the assay. Although we suggest a few thousand compounds for the training set based on the data shown in Supplementary Figure 3, it remains to be fully evaluated how many training points are needed to achieve strong predictivity - in fact, it is likely that the number and structural diversity of hits in the training set more strongly influences predictivity than the total number of assay data points. Nevertheless, in most academic and industry screening centers, preparing a training/test set of ~17,000 compounds, as we used here, is practical.

Based on our results, and depending on whether an AUC of 0.7 or 0.9 is the lower threshold for accuracy needed given the baseline hit rate of the assay, 25-67% of assays should be predictable using a combination of chemical structures and morphology (CS+MO), saving the enormous expense of screening these assays against a full compound library. Especially considering potential improvements in data integration techniques and deep learning for image feature extraction, it is clear that this strategy will be fruitful and accelerate the discovery of useful chemical matter.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.15.422887; this version posted December 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1. Moffat, J. G., Vincent, F., Lee, J. A., Eder, J. & Prunotto, M. Opportunities and challenges in phenotypic drug discovery: an industry perspective. *Nat. Rev. Drug Discov.* **16**, 531–543 (2017).
2. Haasen, D. *et al.* How Phenotypic Screening Influenced Drug Discovery: Lessons from Five Years of Practice. *Assay Drug Dev. Technol.* **15**, 239–246 (2017).
3. Warchal, S. J., Unciti-Broceta, A. & Carragher, N. O. Next-generation phenotypic screening. *Future Med. Chem.* **8**, 1331–1347 (2016).
4. Bruna, J., Zaremba, W., Szlam, A. & LeCun, Y. Spectral Networks and Locally Connected Networks on Graphs. *arXiv [cs.LG]* (2013).
5. Duvenaud, D. K. *et al.* Convolutional Networks on Graphs for Learning Molecular Fingerprints. in *Advances in Neural Information Processing Systems 28* (eds. Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M. & Garnett, R.) 2224–2232 (Curran Associates, Inc., 2015).
6. Li, Y., Tarlow, D., Brockschmidt, M. & Zemel, R. Gated Graph Sequence Neural Networks. *arXiv [cs.LG]* (2015).
7. Kearnes, S., McCloskey, K., Berndl, M., Pande, V. & Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput. Aided Mol. Des.* **30**, 595–608 (2016).
8. Defferrard, M., Bresson, X. & Vandergheynst, P. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. in *Advances in Neural Information Processing Systems 29* (eds. Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I. & Garnett, R.) 3844–3852 (Curran Associates, Inc., 2016).
9. Kipf, T. N. & Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv [cs.LG]* (2016).
10. Battaglia, P., Pascanu, R., Lai, M., Rezende, D. J. & Others. Interaction networks for

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.15.422887; this version posted December 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

learning about objects, relations and physics. in *Advances in neural information processing systems* 4502–4510 (papers.nips.cc, 2016).

11. Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R. & Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **8**, 13890 (2017).
12. Gilmer, J., Schoenholz, S. S., Riley, P. F. & Vinyals, O. Neural message passing for quantum chemistry. *Proceedings of the 34th* (2017).
13. Coley, C. W., Barzilay, R., Green, W. H., Jaakkola, T. S. & Jensen, K. F. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *J. Chem. Inf. Model.* **57**, 1757–1772 (2017).
14. Wu, Z. *et al.* MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).
15. Yang, K. *et al.* Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).
16. Stokes, J. M. *et al.* A Deep Learning Approach to Antibiotic Discovery. *Cell* **180**, 688–702.e13 (2020).
17. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437–1452.e17 (2017).
18. Lapins, M. & Spjuth, O. Evaluation of Gene Expression and Phenotypic Profiling Data as Quantitative Descriptors for Predicting Drug Targets and Mechanisms of Action. *bioRxiv* 580654 (2019) doi:10.1101/580654.
19. 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.* (in-press).
20. Simm, J. *et al.* Repurposing High-Throughput Image Assays Enables Biological Activity Prediction for Drug Discovery. *Cell Chem Biol* **25**, 611–618.e3 (2018).
21. Hofmarcher, M., Rumetshofer, E. & Clevert, D. A. Accurate prediction of biological assays

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.15.422887; this version posted December 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

with high-throughput microscopy images and convolutional networks. *Journal of chemical* (2019).

22. Way, G. P. *et al.* Predicting cell health phenotypes using image-based morphology profiling. 2020.07.08.193938 (2020) doi:10.1101/2020.07.08.193938.
23. Bray, M.-A. *et al.* Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *bioRxiv* 049817 (2016) doi:10.1101/049817.
24. Trapotsi, M.-A. *et al.* Multitask Bioactivity Predictions Using Structural Chemical and Cell Morphology Information. (2020) doi:10.26434/chemrxiv.12571241.v1.
25. Golub, T. L1000 gene expression profiling assay - DOS small molecule perturbagens. (2014).
26. Wawer, M. J. *et al.* Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling. *Proceedings of the National Academy of Sciences* **111**, 10911–10916 (2014).
27. Bray, M.-A. *et al.* A dataset of images and morphological profiles of 30 000 small-molecule treatments using the Cell Painting assay. *Gigascience* **6**, 1–5 (2017).
28. Mullard, A. Machine learning brings cell imaging promises into focus. *Nat. Rev. Drug Discov.* **18**, 653–655 (2019).
29. Dančík, V. *et al.* Connecting Small Molecules with Similar Assay Performance Profiles Leads to New Biological Hypotheses. *J. Biomol. Screen.* **19**, 771–781 (2014).
30. Yang, K. *et al.* Improved Conditional Flow Models for Molecule to Image Synthesis. *arXiv [q-bio.BM]* (2020).
31. McQuin, C. *et al.* CellProfiler 3.0: next generation image processing for biology. *PLoS Comput. Biol.* (2018).
32. Caicedo, J. C. *et al.* Data-analysis strategies for image-based cell profiling. *Nat. Methods* **14**, 849–863 (2017).

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.15.422887; this version posted December 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Methods

Datasets

For this study, we used a compound library of over 30,000 chemicals screened at high-throughput [26]. Of these compounds, about 10,000 came from the Molecular Libraries Small Molecules Repository, other 2,200 were drugs and small molecules, and the remaining 18,000 were novel compounds derived from diversity oriented synthesis. U2OS cells were plated in 384-well plates and treated with these chemicals in 5 replicates, using DMSO as a negative control. The Cell Painting and L1000 platforms were used to generate morphological and transcriptional profiling data, respectively, as previously described [26].

On the assays side, we collected a list of more than 500 assays from drug discovery projects conducted at the Broad Institute at different scales, and we kept those where at least a subset of the chemicals in the compound library described above was tested. In addition, we kept assays that had at least 1 hit identified in the hold out set for evaluation. That resulted in a final list of 376 assays with their corresponding readout results, and the compound-assay matrix had 8.5% of known entries (91.5% sparsity). We prepared assay performance profiles following a double sigmoid normalization procedure to ensure that all readouts are scaled in the same range [29].

The total number of chemicals in the library that had the three types of information required to conduct the experiments in our project (Cell Painting images, L1000 profiles, and assay readouts) was 16,979. From this subset of chemicals, we created a training set with 13,582 examples and a hold-out set for validations with 3,397 examples. This partition was created following a scaffold-based approach to minimize the similarity between chemicals in the training and hold out sets.

Representation of chemical structures (CS) using Chemprop

We used the Chemprop software (http://chemprop.csail.mit.edu/) to train directed-message passing neural networks for learning chemical structure embeddings. The software reconstructs a molecular graph of chemicals from their SMILES string representation, where atoms are nodes and bonds are edges. From this graph, a model applies a series of message passing steps to aggregate information from neighboring atoms and bonds to create a better representation of the molecule. For more details about the model and the software, we refer the reader to prior work [15,16,30].

The representation of chemical structures is learned from the set of ~13,000 training examples, unlike morphological or gene expression features, which were obtained without learning

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.15.422887; this version posted December 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

methods (hand-engineered features). The scaffold split used in our experiments may pose an apparent disadvantage to the learning of chemical structure representations because it may not learn to represent important chemical features in new scaffolds. However, previous research by Yang et al. [15] has shown that Chemprop can generalize to new scaffolds accurately. In addition, the chemicals may also generate new phenotypes in the morphological and gene expression space, which are not seen by the models during training, resulting in a fair comparison of representation power among all modalities. In any event, the goal of this work was to test the ability to identify chemical matter with structures that differ from already-known hit compounds for each assay.

Image-based morphological (MO) profiles from the Cell Painting assay

The Cell Painting assay captures fluorescence images of cells using six dyes to label eight major cell compartments. The five-channel, high-resolution images are processed using the CellProfiler software (https://cellprofiler.org/) to segment the cells and compute a set of 1,500+ morphological features at the single-cell level. These features are aggregated into well- and treatment-level profiles that capture the central statistics of the response of cells to the treatment. In our study, we used treatment level profiles in all experiments. For more details about Cell Painting [23], CellProfiler [31], and the profiling steps [32], see the corresponding references.

Gene expression (GE) profiles from the L1000 assay

The L1000 assay measures transcriptional activity of perturbed populations of cells at high-throughput. These profiles contain mRNA levels for 978 landmark genes that capture approximately 80% of the transcriptional variance [17]. The assay was used to measure gene expression in U2OS cells treated with the set of compounds in our library. Both the profiles and the tools to process this information are available at https://clue.io/ .

Predictive model and data fusion

The predictive model is a feedforward, fully connected neural network which takes as input features and produces as output the hit probabilities for each compound for each assay. The hyperparameters are optimized on a validation set for each feature grouping. The model is trained in a multi-task manner, allocating a binary classifier for each assay. During training, the model computes gradients and backpropagates errors for each classifier independently using the available assay readouts. This setup facilitates learning predictive models with sparse assay readouts.

The input to the neural network can be the features of one or all modalities used in our experiments. To combine features from multiple data modalities, we used two strategies: 1) early data fusion, where feature vectors from two or three modalities are concatenated into a

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.15.422887; this version posted December 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

single vector. 2) Late data fusion, where each modality is used to train a separate model, and then the prediction scores for a new sample are aggregated using the maximum operator. Our results show that, despite its simplicity, late data fusion works best in practice (see Supplementary Table 1), but the results also suggest that more research needs to be done to effectively combine multiple data modalities.

Performance metrics

To evaluate the performance of assay predictors we used the Area Under the Receiving Operating Characteristic (ROC) curve, also known as the AUC metric. We define a threshold of AUC > 0.9 to identify assays that can be accurately predicted. With this threshold, our second performance metric is focused on counting how many assays, from the list of 376 in our study, can be accurately predicted. In addition, we measured hit-rate improvement for individual assays as the ratio between the hit rate obtained using the computational predictors and the hit rate observed in the lab (the “baseline” hit rate):

$$Improvement = \frac{Predictor\ Hit\ Rate}{Baseline\ Hit\ Rate}$$

Predictor hit rates are calculated as the proportion of hits observed in the top 1% of the ranked list of predictions, while baseline hit rates are calculated as the number of hits identified in the complete set of compounds tested for that assay in the original experiment.

Data and code availability

The morphological and gene expression profiles were originally created and published by Wawer, M. J. et al. [26], and can be downloaded from: http://www.broadinstitute.org/mlpcn/data/Broad.PNAS2014.ProfilingData.zip

The Cell Painting images were also made available by Bray et al. [27], and can be obtained from the following link: http://gigadb.org/dataset/100351

The latest version of morphological profiles is also available in the following AWS S3 bucket: https://registry.opendata.aws/cell-painting-image-collection/

The Chemprop software and source code used for training machine learning models can be found in the following link: http://chemprop.csail.mit.edu/

Anonymized assay data and code to reproduce the analysis in the paper will be made available soon.

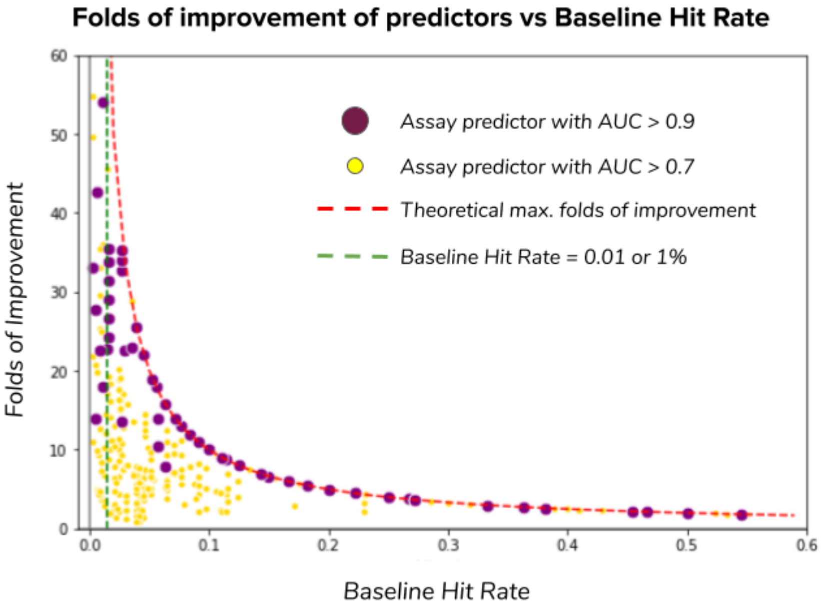
bioRxiv preprint doi: https://doi.org/10.1101/2020.12.15.422887; this version posted December 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Acknowledgements

This study was supported by a grant from the National Institutes of Health (R35 GM122547 to AEC). The authors are grateful for guidance and thoughtful discussions with Regina Barzilay and Tommi Jaakkola which improved the analysis and manuscript.

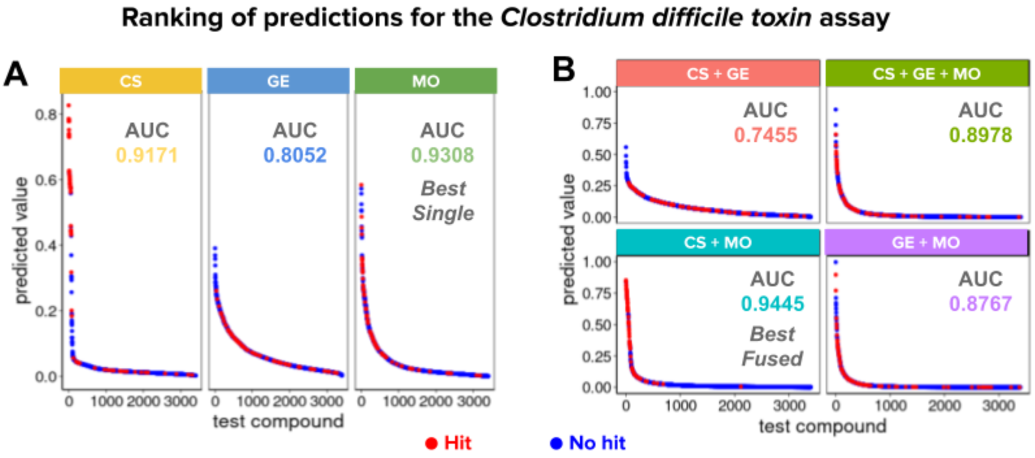
bioRxiv preprint doi: https://doi.org/10.1101/2020.12.15.422887; this version posted December 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Material

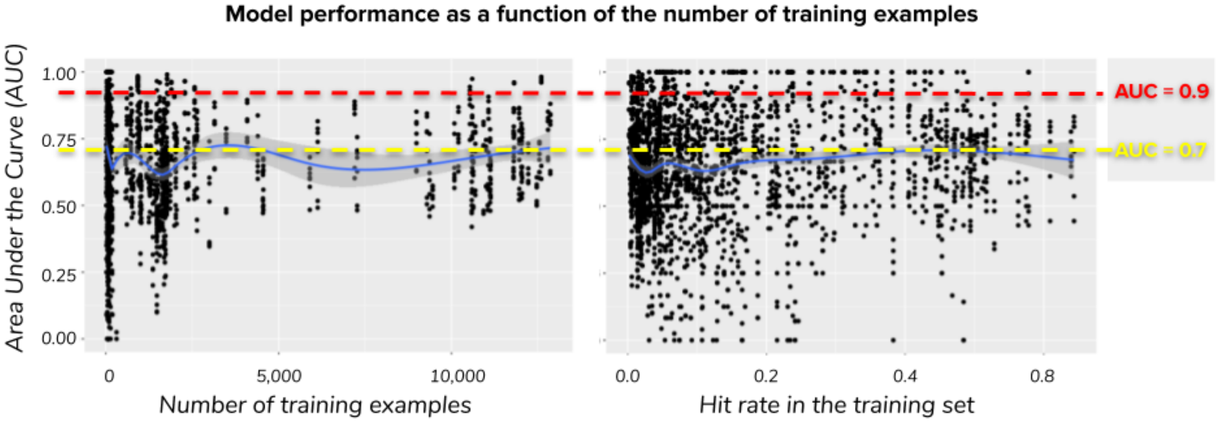


Supplementary Figure 1. Improvement of hit rates for the assays in the dataset. Each point in the plot represents one assay being predicted with one of the three descriptors (CS, GE or MO) or combinations of them. Assay predictors with AUC > 0.7 are displayed in yellow and predictors with AUC > 0.9 are displayed in purple. Assay predictors with AUC < 0.7 are not shown. The x-axis represents the baseline hit rate, i.e., the proportion of compounds found to be hits in the set of tested compounds for an assay. Primary assays in real world scenarios usually have less than 1% baseline hit rates (green dashed line at 0.01); the assays in this dataset with baseline hit rates substantially higher than 1% are likely validation/confirmatory or secondary assays. The y-axis presents the folds of improvement of assay predictions obtained with a machine learning predictor as a function of the baseline hit rate. Accurate predictors (AUC > 0.9) often offer improvements up to the theoretical maximum (100% divided by the assay's baseline hit rate), and higher-fold improvements are only possible for assays with a lower baseline hit rate.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.15.422887; this version posted December 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Supplementary Figure 2. Illustration of the impact of data fusion.
Each plot shows the hit probability for all compounds in the holdout set produced by seven different predictors for one example assay: *Clostridium difficile toxin*. The x-axis lists ranked compounds and the y-axis shows probability scores. The points in the plots are each of the compounds in the holdout set ranked according to their probability of being a hit. Actual hit compounds are in red and non-hit compounds are in blue (red points maybe hidden behind blue points). Each plot represents a predictor based on the corresponding profiling modality or their combinations (CS: chemical structures, GE: gene expression, MO: morphology). A) Predictions made using single modalities. In this example, the CS predictor alone displays great confidence to rank hit compounds in the top of the list, dropping the probabilities quickly for other compounds. MO exhibits a smoother transition of highly ranked compounds and achieves the best AUC among single modalities. B) Predictions obtained with data fusion. The combination of CS+MO appears to capture the individual modalities' strengths (high confidence and high AUC), improving the overall ranking even further (see Figure 2 in the main text).

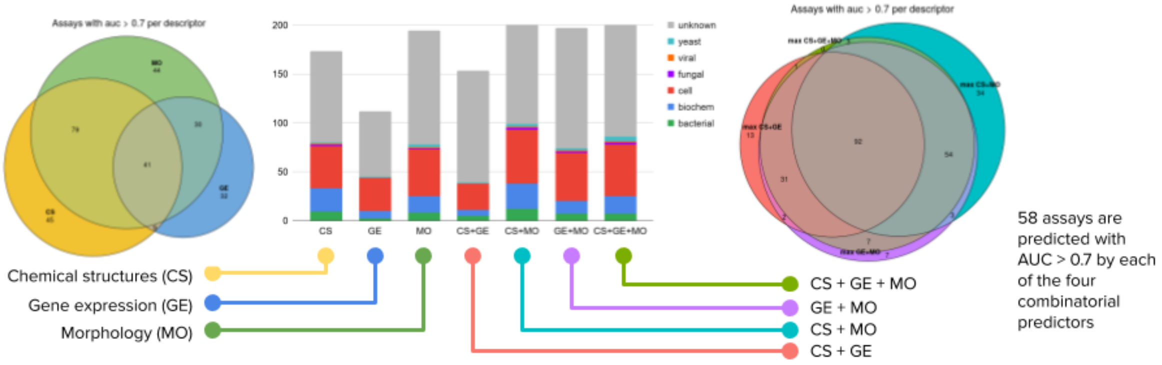


Supplementary Figure 3. The performance of predictive models is independent of the number of available training examples; some assays can be predicted with high accuracy (AUC > 0.9) using only a few example hits. The plots show on the y-axis the test set accuracy as a function of (A) the number of training examples, and (B) the hit rate in the training set. The trend lines (in blue) suggest that there is no correlation between the number of training examples (positives or

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.15.422887; this version posted December 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

total) and the performance of assay predictors.

A



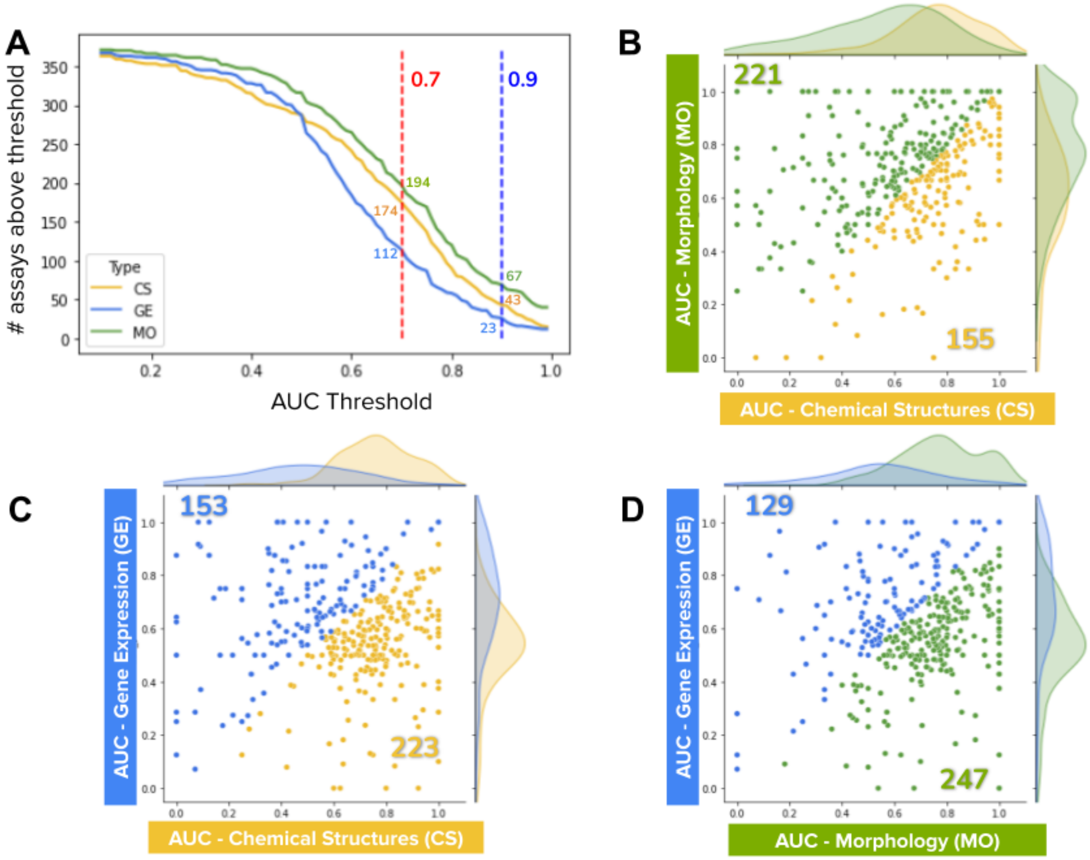
58 assays are predicted with AUC > 0.7 by each of the four combinatorial predictors

B

	CS	GE	MO	CS+GE	CS+MO	GE+MO	CS+GE+MO
Mean AUC	0.636	0.599	0.693	0.642	0.684	0.687	0.694
AUC > 0.7	174	112	194	153	200	197	200
Retrospective	-	-	-	240	253	245	292

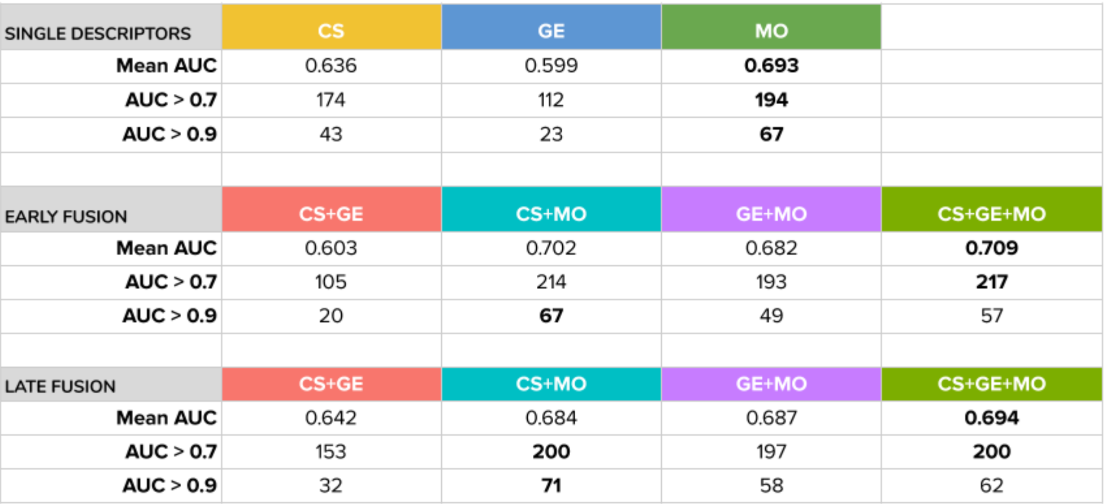
Supplementary Figure 4. Summary of the number of assays predicted with models that have AUC > 0.7, which is a lower performance threshold than the one used throughout our study. The major trends identified and presented in Figures 2 and 3 (main text) hold with this lower threshold, except that the total number of acceptable assay predictors increases. Importantly, these predictors are also capable of improving hit rates in many cases (see yellow points in Supplementary Figure 1). The row "Retrospective" in Table B presents the number of assays with AUC > 0.7 that would be predicted by any of the modalities individually or their combinations.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.15.422887; this version posted December 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Supplementary Figure 5. Area Under the Curve (AUC) performance of the three individual modalities evaluated in our study: Chemical Structures (CS), Gene Expression (GE), and Morphology (MO). A) Number of assays predicted by each modality at specific AUC thresholds. As the AUC threshold is increased, the number of assays meeting the threshold decreases for all modalities. The two thresholds discussed in this paper are highlighted in red (0.7) and blue (0.9). MO consistently outperformed CS and GE at all thresholds. B, C, D) Scatter plots of AUC for pairs of modalities. Each point in the plots represents an assay, the x coordinate indicates the AUC obtained in one modality, and the y axis represents the AUC obtained in the other modality. Colors represent the three individual modalities: CS (yellow), GE (blue) and MO (green). Points (assays) above or below the diagonal (equal performance) are colored according to the modality that has the highest AUC. The two colored numbers inside the plot indicate the total number of assays with higher AUC with respect to the other modality in the same plot. Note that there are many points far off the diagonal, indicating high AUC in one modality but low in the other. This indicates potential for complementary and fusion among the different data modalities.

bioRxiv preprint doi: https://doi.org/10.1101/2020.12.15.422887; this version posted December 16, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



SINGLE DESCRIPTORS	CS	GE	MO	
Mean AUC	0.636	0.599	**0.693**	
AUC > 0.7	174	112	**194**	
AUC > 0.9	43	23	**67**	
EARLY FUSION	CS+GE	CS+MO	GE+MO	CS+GE+MO
Mean AUC	0.603	0.702	0.682	**0.709**
AUC > 0.7	105	214	193	**217**
AUC > 0.9	20	**67**	49	57
LATE FUSION	CS+GE	CS+MO	GE+MO	CS+GE+MO
Mean AUC	0.642	0.684	0.687	**0.694**
AUC > 0.7	153	**200**	197	**200**
AUC > 0.9	32	**71**	58	62

Supplementary Table 1. Overall performance of profiling modalities and their combinations presented in the columns of the tables. Early fusion refers to concatenation of feature vectors before training predictive models, while late fusion refers to keeping the maximum prediction of individual models. The tables present three performance metrics: Mean AUC, number of assays predicted with AUC > 0.7, and number of assays predicted with AUC > 0.9. In the three cases, higher numbers indicate better performance, and the best result is in bold for each row. Late fusion yields the largest number of predictors with AUC > 0.9 overall, and also for all combinations of descriptors.