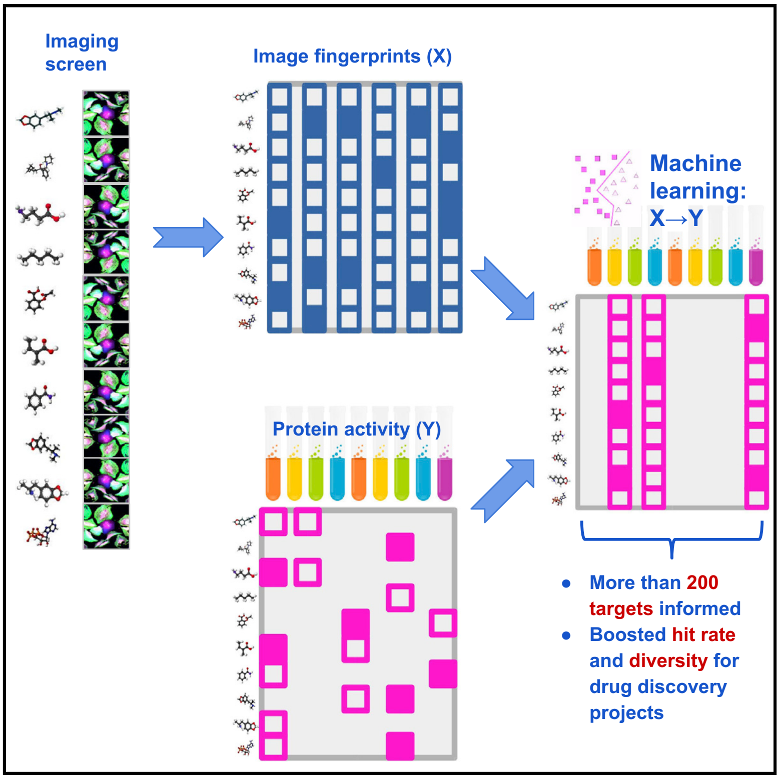
Cell Chemical Biology

Repurposing High-Throughput Image Assays Enables Biological Activity Prediction for Drug Discovery

Graphical Abstract


Imaging screen
Image fingerprints (X)
Machine learning: X→Y
Protein activity (Y)
More than 200 targets informed
Boosted hit rate and diversity for drug discovery projects

Highlights

- Scalable machine-learning-based method predicting compound activity from images
- Hundreds of assays predicted by one image screen annotating half a million compounds
- Image-based models boosted hit rate and diversity in two drug discovery projects
- Proof of concept justifying further work on image-based learning for drug discovery

Authors

Jaak Simm, Günter Klambauer,
Adam Arany, ..., Sepp Hochreiter,
Yves Moreau, Hugo Ceulemans

Correspondence

hceulema@its.jnj.com

In Brief

Simm et al. demonstrate a computational method to predict the activities of compounds in hundreds of biological assays from a single image-based screen of half a million compounds. The resulting models boosted the identification and diversity of hit compounds for two projects, encouraging further research in this field.

Simm et al., 2018, Cell Chemical Biology *25*, 1–8
May 17, 2018 © 2018 Elsevier Ltd.
https://doi.org/10.1016/j.chembiol.2018.01.015

Please cite this article in press as: Simm et al., Repurposing High-Throughput Image Assays Enables Biological Activity Prediction for Drug Discovery, Cell Chemical Biology (2018), https://doi.org/10.1016/j.chembiol.2018.01.015

Cell Chemical Biology
Resource

Repurposing High-Throughput Image Assays Enables Biological Activity Prediction for Drug Discovery

Jaak Simm,[1,8] Günter Klambauer,[2,8] Adam Arany,[1,8] Marvin Steijaert,[3] Jörg Kurt Wegner,[4] Emmanuel Gustin,[4] Vladimir Chupakhin,[4] Yolanda T. Chong,[4] Jorge Vialard,[4] Peter Buijnsters,[4] Ingrid Velter,[4] Alexander Vapirev,[5] Shantanu Singh,[6] Anne E. Carpenter,[6] Roel Wuyts,[7] Sepp Hochreiter,[2,9] Yves Moreau,[1,9] and Hugo Ceulemans[4,9,10,*]
[1]ESAT-STADIUS, KU Leuven, Kasteelpark Arenberg 10, 3001 Leuven, Belgium
[2]Institute of Bioinformatics, Johannes Kepler University Linz, Altenbergerstrasse 69, 4040 Linz, Austria
[3]Open Analytics NV, Jupiterstraat 20, 2600 Antwerp, Belgium
[4]Janssen Pharmaceutica NV, Turnhoutseweg 30, 2340 Beerse, Belgium
[5]Facilities for Research, KU Leuven, Willem de Croylaan 52c, Box 5580, 3001 Leuven, Belgium
[6]Imaging Platform, Broad Institute of Harvard and MIT, 415 Main Street, Cambridge, MA 02142, USA
[7]ExaScience Life Lab, IMEC, Kapeldreef 75, 3001 Leuven, Belgium
[8]These authors contributed equally
[9]Senior author
[10]Lead Contact
*Correspondence: hceulema@its.jnj.com
https://doi.org/10.1016/j.chembiol.2018.01.015

SUMMARY

In both academia and the pharmaceutical industry, large-scale assays for drug discovery are expensive and often impractical, particularly for the increasingly important physiologically relevant model systems that require primary cells, organoids, whole organisms, or expensive or rare reagents. We hypothesized that data from a single high-throughput imaging assay can be repurposed to predict the biological activity of compounds in other assays, even those targeting alternate pathways or biological processes. Indeed, quantitative information extracted from a three-channel microscopy-based screen for glucocorticoid receptor translocation was able to predict assay-specific biological activity in two ongoing drug discovery projects. In these projects, repurposing increased hit rates by 50- to 250-fold over that of the initial project assays while increasing the chemical structure diversity of the hits. Our results suggest that data from high-content screens are a rich source of information that can be used to predict and replace customized biological assays.

INTRODUCTION

High-throughput imaging (HTI), also known as high-content screening (HCS), captures the morphology of the cell and its organelles by microscopy and has yielded diverse biological discoveries (Pepperkok and Ellenberg, 2006; Starkuviene and Pepperkok, 2007; Walter et al., 2010). HTI is often applied to screen chemical compounds based on morphological changes they induce (Held et al., 2010; Yarrow et al., 2003). Currently, most HTI screens are designed to evaluate one specific biological process and exploit only a handful of morphological features from the image, chosen to best measure that process (Singh et al., 2014) (Figure 1).

However, any cellular system hosts many more biochemical processes and thousands of potential drug targets, all of which are exposed to the screened chemical compounds. Many of these targets and processes have an impact on cell morphology, and that morphology can to a large extent be extracted from the images (Carpenter et al., 2006). The resulting set of features, which include not just shape and spatial metrics but also the intensity and patterning of fluorescently labeled markers, can be used to describe chemical compounds and can be considered as an image-based compound fingerprint. Such fingerprints are powerful enough to accomplish a variety of important biological aims, including optimizing the diversity of compound libraries, grouping compounds by pharmacological mechanism, and grouping genes based on functional similarity (Caicedo et al., 2016).

Motivation

We therefore hypothesized that image-based fingerprints of compounds derived from a given image-based cellular assay, might be leveraged to predict compound activity in seemingly unrelated assays. Effective predictors of biological activity already exist; virtual screening and quantitative structure–activity relationship (QSAR) analyses typically rely on features derived from the chemical structure of compounds to predict their activity in assays. Structure-based models are predictively performant (Cumming et al., 2013) but only for those parts of chemical space for which sufficient assay activity data are available. Unfortunately, compounds that are chemically very different from any known active compound are unlikely to be predicted as active. Because cell morphology can reflect compound-induced modulation of diverse targets and biochemical processes regardless of compound structure, we suspected that image-based models would avoid this limitation and may complement chemistry-based models in novel and poorly annotated chemical space.

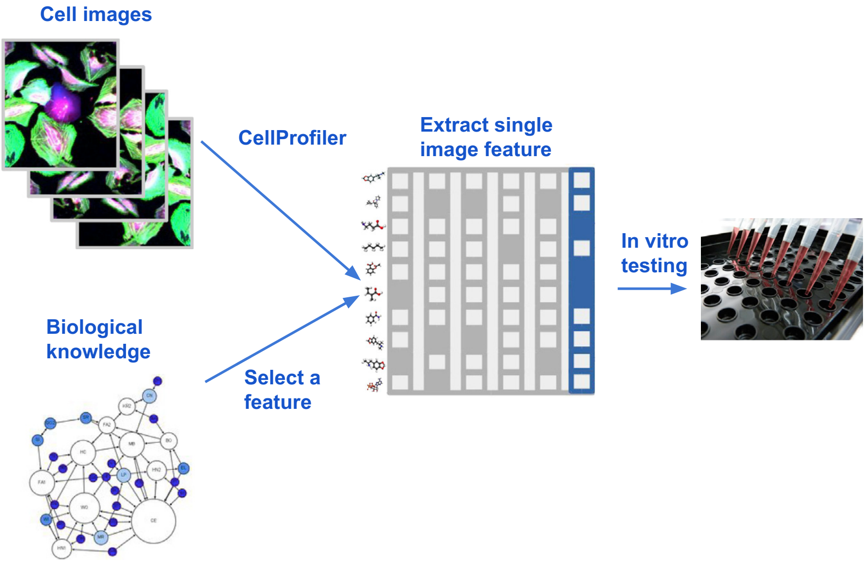


Figure 1. A Typical HTI Screen Approach
Few or single features are extracted from cellular images; the remainder of information (gray) is ignored (Ansbro et al., 2013; Evensen et al., 2010).

Decades of HCS experience indicate an ability of image-based readouts to generalize over multiple unrelated targets. Yet, most academic and commercial imaging campaigns have followed a narrowly focused classical setup depicted in Figure 1, leaving a large volume of biological information untapped. Therefore, we aimed to repurpose pre-existing imaging screens to generically predict compound activities in assays that may be unrelated to the original screening assay.

RESULTS

Overview of Proposed Repurposing Approach

We propose a pipeline (Figure 2) to leverage the rich information in existing image screens for the prediction of activity in a variety of orthogonal assays directed at seemingly unrelated proteins and processes. First, we extract an extensive image-based fingerprint of morphological features for each compound in a single, already completed large-scale imaging screen (X in Figure 2), aiming for maximal and unbiased information capture (see next section). Second, we introduce existing activity data for orthogonal assays of interest on these compounds (Y in Figure 2). Then, we train supervised machine-learning models to predict Y from X and choose models with high predictive performance. Finally, we use these high-quality models to select compounds for *in vitro* testing. Next we describe each of these steps in detail.

Extracting Image-Based Fingerprints

The goal of extracting image-based fingerprints is to capture all available information about the biological state of the cell from the image. In this work, we use previously developed software (CellProfiler) and methods (Gustafsdottir et al., 2013) to produce a feature vector for each cell, capturing general morphology, shape, and biologically important parameters (e.g., micronucleus count). For the three-channel glucocorticoid receptor (GCR) HTI assay used in the evaluation, this produced an 842-dimensional feature vector per cell. Then, for each plate we normalize each feature using the mean and the standard deviation of the corresponding feature from the negative controls (cells without treatment). Finally, for each compound, we compute a vector of feature medians across all cells in its image, producing a single image-based fingerprint.

We note that an attractive alternative procedure is to use convolutional neural networks (CNNs) to learn feature representation directly from the raw pixels of cell images. This strategy shows promise but is still exploratory for image-based profiling; considering the high computational cost and hardware requirements, we leave this direction to future research.

Machine Learning for Image-Based Fingerprints

We next use machine learning to take image-based fingerprints (X in Figure 2) and the existing bioactivity measurements on the assays of interest (Y in Figure 2) to learn a model to predict bioactivity of new compounds given their image-based fingerprints.

The simplest approach would be to model each column of the activity data separately (single-task learning). However, we can take advantage of the existence of multiple related prediction tasks by modeling them jointly (multitask learning). In the case of related tasks, multitask learning is known to improve the overall performance significantly (Caruana, 1997).

Both regression and classification methods could be used in the repurposing workflow we propose. Here, we describe two that yielded good computational and predictive performance. To document the compatibility of this generic concept with other machine-learning methods, we also carried out additional experiments with random forest (Breiman, 2001) and k-nearest neighbor classifier in our validation setup (see STAR Methods).

Bayesian Matrix Factorization

First, we explored Bayesian matrix factorization, a multitask method that does not require hyperparameterization (like regularization) and provides uncertainty estimates for predictions. Specifically, we used the Bayesian matrix factorization method Macau, which can account for side information (in this case image features). To factorize the N times M activity matrix Y, Macau represents each compound and each assay by D-dimensional latent vectors $\boldsymbol{u}_i$ and $\boldsymbol{v}_j$, respectively. The prediction for the element Y_{ij}, corresponding to the activity of compound i on assay j, is given by the scalar product $\boldsymbol{u}_i^T \boldsymbol{v}_j$. $\boldsymbol{x}_i$ is an F-dimensional features vector (F = 842) corresponding to the image-based fingerprint (see section on Extracting Image-Based Fingerprints) and is added to the prior of the latent vectors of compounds $\boldsymbol{u}_i$. Macau maps all tasks to the same

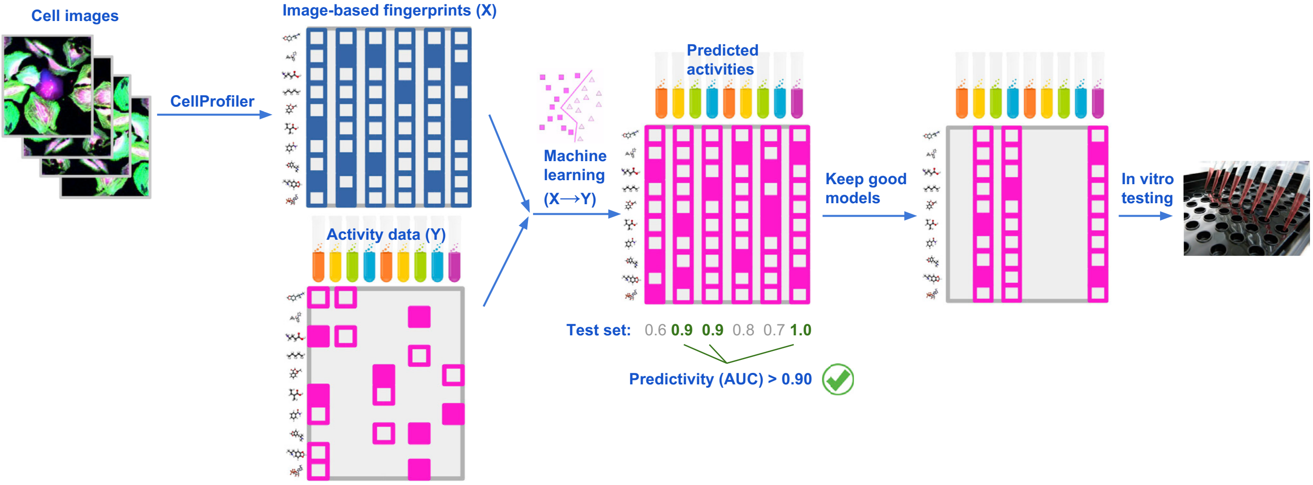


Figure 2. Strategy to Repurpose Imaging Screens to Efficiently Predict Biological Activity

Features extracted from images of cells are used by machine-learning methods to model all available activity data from previously performed assays. Assays with good predictivity on the test data are then selected for testing a relatively small number of predicted-active compounds, chosen from a large set of compounds profiled in the imaging assay.

D-dimensional latent space, therefore enabling sharing of parts of the model.

This results in a probabilistic model of

$$Y_{ij} \sim \mathcal{N}\left(\boldsymbol{u}_i^{\mathrm{T}}\boldsymbol{v}_j,\ \alpha^{-1}\right),$$

$$\boldsymbol{u}_i \sim \mathcal{N}\left(\mu_u + \beta\boldsymbol{x}_i,\ \Lambda_u^{-1}\right),$$

$$\boldsymbol{v}_j \sim \mathcal{N}\left(\mu_v,\ \Lambda_v^{-1}\right),$$

where α is the precision of the observations, parameters μ_u and Λ_u model the mean and precision of the compound latent vectors, similarly μ_v and Λ_v model the latent vectors for assays. The parameter β is a D times F dimensional matrix that maps the image features to the compound latent space. To learn β we apply a Gaussian prior on it:

$$\beta \sim \mathcal{N}\left(\mathbf{0},\ \Lambda_u \otimes \lambda_\beta I_F\right)^{-1},$$

where $\otimes$ is the Kronecker product, λ_β is a precision parameter, and I_F is the identity matrix of size F. Figure 3 depicts the plate diagram for the probabilistic model.

By deriving conditional distributions for all model variables, we obtain a Gibbs sampler that iterates over all model variables (Simm et al., 2017). For each variable, it samples a value from the conditional distribution by fixing all the others. Finally, to compute the predictions for Y_{ij}, we use each sample to compute the scalar products $\boldsymbol{u}_i^T\boldsymbol{v}_j$ and then average over the samples. We observed that the performance of the method does not degrade with choosing a high latent dimensionality D. In practice, this implies the choice of a large enough latent space; in our case D = 150.

The Macau model described here is for the regression setting, i.e., Y_{ij} are real-valued. The model can be easily modified to handle the classification setting by replacing the normal prior on Y_{ij} with a probit one. We have made the implementations for both settings available open source. The C++/Python package is available at https://github.com/jaak-s/macau.

Deep Neural Networks

The matrix factorization model described above is linear and may lack the flexibility to capture all important biological effects. Therefore, we additionally tested a multitask deep learning architecture. We implemented deep neural networks (DNNs), concretely feedforward artificial neural networks, with many layers comprising a large number of neurons and rectified linear units (Mayr et al., 2016). DNNs (Figure 4) consists of interconnected neurons that are arranged hierarchically in layers. In the first layer of the network (the input layer), the neurons obtain an input vector that is the image-based fingerprint. The intermediate layers (the hidden layers) comprise the hidden neurons that have weighted connections to the neurons of the previous level layer and can be considered as abstract features, built from features below. The last layer (the output layer) supplies the predictions of the model. Typical DNNs comprise several layers, which consist of thousands of neurons.

We used rectified linear units (ReLUs) as activation functions in the hidden layers. The output layer has sigmoid activation functions. To avoid overfitting, we employed multiple regularization techniques, concretely dropout (Srivastava et al., 2014) and early stopping. Both the dropout rate and the early-stopping parameter, i.e., the number of epochs after which learning is stopped, were determined on a validation dataset.

Deep learning naturally enables multitask learning (Caruana, 1997). In our setting, each assay is a task. Commonalities across the assays translate to shared representation in the hidden layers and can yield performance improvements (Mayr et al., 2016). We modeled each assay by a separate output unit.

We used cross-entropy as a loss function for our DNNs:

$$\sum_{i,j} m_{ij}\left(Y_{ij}\log\tilde{Y}_{ij} + (1-Y_{ij})\log\left(1-\tilde{Y}_{ij}\right)\right),$$

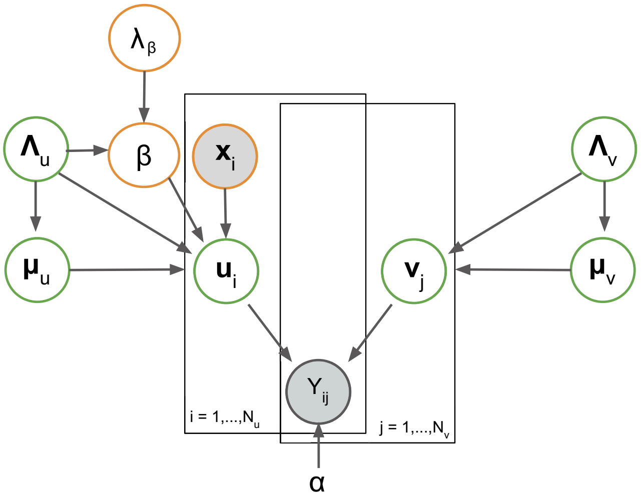


Figure 3. Diagram for the Probabilistic Model for the Bayesian Matrix Factorization Approach Macau

The shaded circles denote observed variables and the transparent circles are inferred from the data.

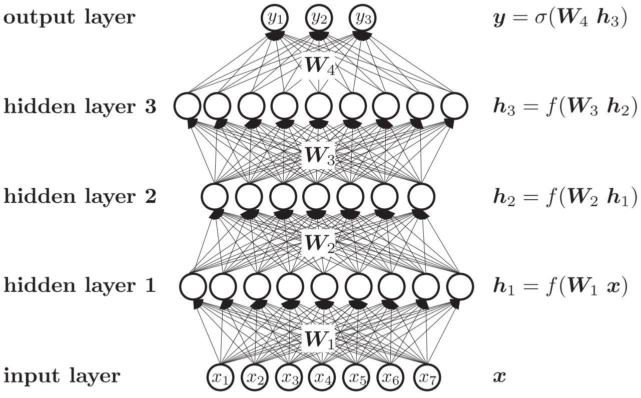


Figure 4. General Architecture of Deep Neural Networks

Variable x denotes the image-based fingerprint, y corresponds to biological activity. The tested hyperparameters of DNN are shown in Table S1.

where $\tilde{Y}_{ij}$ is the prediction for compound *i* and assay *j* and the actual label is Y_{ij}, which indicates whether the compound was active $(Y_{ij} = 1)$ or inactive $(Y_{ij} = 0)$ in the given assay. The binary variable m_{ij} indicates whether a measurement is present $(m_{ij} = 1)$ or missing $(m_{ij} = 1)$. The implementation details, optimization of architecture, and hyperparameters are given in Supplemental Information.

Selection of High-Quality Models

Next, we select only assays yielding a highly reliable model. To this end, we employ cross-validation, i.e., we split the compounds into *k* folds (here, *k* = 3). In cross-validation, the activity data for each fold are predicted using a model built on the data from the other folds. The resulting predictions enable the computation of an AUC-ROC (area under the curve-receiver operating characteristic) score, or some other performance metric of choice. We used the average of the *k* folds as the evaluation metric for each model, and focused on models with an AUC-ROC > 0.9. If a machine-learning method required an optimization of hyperparameters (e.g., choices of model architecture, kernel, dropout), we applied nested cross-validation (Mayr et al., 2016).

The simplest splitting scenario would be the random assignment of compounds to folds. However, in the case of chemistry-based modeling of pharmaceutical datasets, where compounds tend to be concentrated around attractive chemical backbones, this approach results in overoptimistic performance estimates (as close structural analogs get spread over test and validation, and performance metrics are boosted but do not hold up when applied to new chemistry). One popular mitigation approach is the use of temporal or roll-back splitting, where a timestamp is used to separate test and validation folds. In a multitask setting, however, temporal splitting is impractical because the order of measurement of the same compounds in different assays is not guaranteed to be aligned. Instead, we clustered the compounds based on chemical similarity and randomly assigned the clusters into folds (see Figure S1). Here, a Tanimoto similarity cutoff on ECFP6 features was used to ensure close analogs ended up in the same test or validation fold. A high choice of cutoff may fall short of addressing the overoptimistic performance estimation, while a low cutoff may restrict the learning potential (as machine learning relies on recognizing similarities). In our experience, a similarity cutoff of 0.7 offers an optimal trade-off.

Image-based fingerprints are insulated from the underlying chemistry. Thus, performance estimates for the resulting models are not expected to be skewed by the above-mentioned pharmaceutical chemistry bias. However, for consistency reasons, we still followed the clustered cross-validation approach.

Compound Selection for *In Vitro* Testing

Finally, we select compounds highly ranked by good-quality models. There are two main selection strategies. The first is to select all the highest ranked compounds for *in vitro* testing. Although simple, the strategy may select sets that are too homogeneous or too chemically similar to the original training set. The second strategy is to apply diversity maximization (e.g., sphere exclusion clustering) on the highly ranked compounds, and only test a diverse set. This strategy can result in novel hits, but only if the model can generalize across scaffolds. As indicated before, we hypothesized that this is the case for models that use image-based fingerprints.

Experimental Evaluation

In the following, we evaluate our HTI assay repurposing approach in a large-scale industrial context. To begin, we chose a HTI screen of 524,371 proprietary compounds originally used for the detection of GCR nuclear translocation. In this assay, each compound was applied at a concentration of 10 μM to H4 brain neuroglioma cells, incubated for 1 hr, then exposed to 1 μM hydrocortisone for 1 hr to stimulate translocation of the GCR. Cells were then fixed and imaged in three-channel fluorescence, with Hoechst to label the nucleus, CellMask deep red to delineate cell boundaries, and indirect immunofluorescence to detect GCR. From these images, our pipeline extracted 842 dimensional vectors for each compound representing the feature matrix X (see the section on Extracting Image-Based Fingerprints).

Table 1. The Number of Protein Assays above the AUC-ROC Threshold for Machine Learning Methods Macau and Deep Neural Networks (DNN)

AUC-ROC Threshold	Macau (%)	DNN (%)	Common (%)
0.9	31 (5.8)	43 (8.0)	26 (4.9)
0.7	218 (40.7)	245 (45.8)	209 (39.1)

The percentage is calculated relative to the total number of 535 assays. The Common column depicts the number of assays well predicted by both of the methods. Venn diagrams of the predicted targets are shown in Figures S2 and S3. The tested hyperparameters are described in Tables S1 and S2. The mean AUC-ROC values for Macau, DNN, random forest, and k-nearest neighbor are given in Table S3.

The bioactivity matrix Y documents the available experimental activities of 524,371 imaged compounds in about 1,200 biochemical or cellular assays that can all be interpreted as an activity on a protein target. This also means that a single compound can be measured on multiple targets. The activity is expressed as the pXC_{50} of the given compound in the given assay. The pXC_{50} is defined as $-\log_{10}$ of the molarity concentration of the compound yielding a half-maximal effect in the experimentally measured dose-response curve. Typically, a given compound is measured in a handful of assays, such that Y is sparsely populated, i.e., has many missing values. In total, more than 10 million pXC_{50} values were available for the roughly 1,200 prediction tasks, corresponding to a fill rate of Y of about 1.6%.

We evaluated all protein assays at four different thresholds of pXC_{50} (here activity is defined as a pXC_{50} value exceeding the threshold), namely 5.5, 6.5, 7.5, and 8.5. We only used assay-threshold pairs with at least 25 actives and 25 inactives. For 535 assays, at least one pXC_{50} threshold resulted in a data subset meeting this criterion. In the step of selecting high-quality models (see section on Selection of High-Quality Models), we used AUC-ROC higher than 0.9 as the cutoff. We additionally report results for a cutoff of 0.7.

Results of *In Silico* Experiments

Of the 535 assays, the described pipeline yielded 31 assays with high-quality models using Macau (run for 2,000 iterations, discarding the first 400 as burn-in) and 43 using DNN (for details, including hyperparameter tuning, see Table S1). An AUC-ROC threshold of 0.7 yielded 6–7 times as many assays (Table 1).

Both methods can successfully repurpose the original GCR HTI assay for predicting activity toward more than 30 unrelated protein targets (AUC-ROC > 0.9), and provide models of sufficient quality to enrich compound sets for (or deplete them of) activity toward a further 200 targets (AUC-ROC > 0.7). Therefore, the image-based fingerprinting of HTI assays prove a rich and hitherto untapped source of information on biological activity that is compatible with multiple machine-learning methods. If computational resources are limiting, less expensive methods such as Macau or random forest yield a predictive performance comparable with that of deep learning (see Table S3 for mean AUC-ROC values for Macau and random forest).

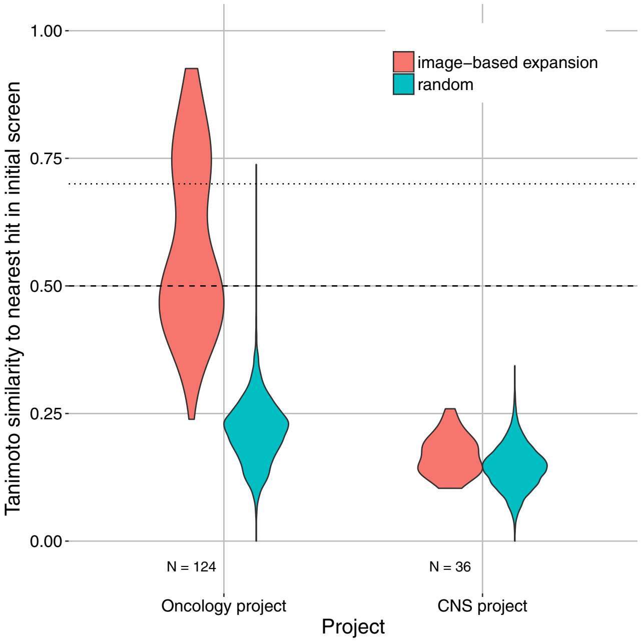


Figure 5. Image-Based Profiling Strategy Yields More Chemically Diverse Compounds than Would be Expected for Chemical Extension

In an oncology project (left) and a CNS project (right), we calculated the ECFP (radius 4)-based Tanimoto similarity of each hit to the nearest hit identified by the initial high-throughput screen (red). For reference, the blue distributions show the similarity of randomly selected compounds to the closest hit identified by the initial high-throughput screen. Note that in the CNS project, unlike the oncology project, the selection procedure involved an additional step to reduce representatives from the same chemical-structural class. The horizontal dotted lines depict the 0.5 and 0.7 similarity levels.

Results for *In Vitro* Validation

As the Macau results were readily available during the early phase of the research, we proceeded with the *in vitro* validation using these models. Among these 31 assays with high-quality predictions from Macau, two were connected to ongoing discovery projects: one oncology project and one central nervous system (CNS) project. For these two projects, we selected compounds for testing.

Results for the Oncology Project

For the oncology project, the target was a kinase with no known direct relation to the glucocorticoid receptor. Using our Macau model, we ranked about 60,000 compounds tested in the GCR assay but for which no activity measurement was available in the oncology screen. We selected the 342 highest ranking compounds for experimental follow-up (see the section on Compound Selection for *In Vitro* Testing). We found that 124 of them were submicromolar (XC_{50} < 1 μM) hits (36.3% hit rate), which corresponds to a 50-fold enrichment over the initial high-throughput screen (0.725% hit rate).

To evaluate the chemical diversity of the hits, we computed the Tanimoto similarity (based on extended-connectivity fingerprints (ECFP); Rogers and Hahn, 2010) of each hit to the nearest hit identified by the initial high-throughput screen (red distribution in Figure 5). Seventy percent of the hits are below the 0.7 similarity line, and a significant proportion is even below 0.5. Per definition, a chemical similarity search based on the initial hits would rarely yield analogs below the 0.7 line, and extremely

Table 2. Number of Murcko Scaffolds in the Two Follow-Ups

Project Name	No. of Murcko of Initial Screen	No. of Murcko of New Hits (Novel/All)	No. of New Hits
Oncology	2,660	108/117	124
CNS	57	34/34	36

rarely below the 0.5 line. We also found 108 novel Murcko scaffolds among the new hits (Table 2). Together these two facts imply that our repurposing pipeline can result in a hit set with high chemical diversity. For reference, the figure also shows the distribution for randomly selected compounds (blue distribution in Figure 5). We furthermore compared the top ranked compounds with those retrieved by chemical fingerprint-based approaches. Specifically, we used the exact same activity data with chemical structure-based features (ECFP) to train a Macau model and then ranked the untested 60,000 compounds. From the above-mentioned top 342 compounds ranked by the image features, 113 (33%) compounds were retrieved by ECFP model in its top 342. From the 124 actives, 44 (35%) compounds were found in the top 342 of the ECFP model. Moreover, to identify all 124 active compounds using the ECFP ranking, one would need to test more than 21% of the 60,000 candidate compounds, i.e., at least 13,000 compounds. This shows that the image fingerprints clearly provided an additional source of information that is not encoded in the chemical fingerprints.

Results for the CNS Project

For the CNS project, the target was a non-kinase enzyme, again without obvious relation to the glucocorticoid receptor. Using our Macau model, activity was predicted for all 500,000 image-annotated compounds, and we selected all compounds with submicromolar prediction, resulting in 1,715 compounds. Next, we kept only compounds without unfavorable properties, such as PAINS filter (Baell and Holloway, 2010) and low predicted CNS availability (see STAR Methods). For this project, to maximize compound diversity, we employed the selection strategy of grouping the remaining compounds into clusters based on structure, using sphere exclusion clustering with similarity cutoff 0.7 (see section on Compound Selection for *In Vitro* Testing). We then selected a handful of representatives from each cluster resulting in 141 compounds. We experimentally tested them and found that 36 of them were submicromolar hits (25.5% hit rate), which corresponds to a 289-fold enrichment over the hit rate of the initial high-throughput screen (0.088% rate). These compounds were highly diverse (Tanimoto similarity <0.3; Figure 5) while maintaining a relatively high hit rate. The 36 hits resulted in 34 novel Murcko scaffolds (Table 2).

DISCUSSION

In this work, we have demonstrated that HTI data enable the identification of diverse hits without the need to test the entire library in the target assay. By accessing rich morphological features of the cell, imaging screens capture diverse cellular processes, resulting in a fingerprint of biological action. Our results indicate that images from HTI screening projects that are conducted in many institutions can be repurposed to dramatically reduce the scale of screens required for other projects, even those that seem unrelated to the primary purpose of the HTI screen.

We emphasize that our approach relies on a supervised machine-learning method, and hence activity measurements and imaging data must be acquired for a reasonably sized library of compounds to train the model. Subsequently, however, it seems possible to replace many particular assays with the potentially more cost-efficient imaging technology together with machine-learning models. Specifically, one would execute one or a few image screens on the library instead of dozens of target-focused assays. This raises an interesting question of the breadth of drug targets that could be accessed by imaging screens if the screen were optimized for that purpose, or if a combination of screens was used that explored multiple cell lines or sources, culturing conditions, staining of organelles, and/or incubation times.

We leave for future work the head-to-head comparison of chemistry-based and image-based fingerprints, but can speculate based on our results. In the case of a well-covered chemical space, we would not expect image-based fingerprints to outperform a well-designed chemical fingerprint like ECFP. For example, if the compound in question has several close enough neighbors, we expect chemical fingerprints to prove predictively performant. In contrast, we expect the performance of image-based fingerprints that do not depend on chemical closeness to be superior for scaffold hopping, i.e., identifying active compounds with novel backbones, given it does not depend on the chemical closeness. Evidence for this idea includes the high chemical diversity of active compounds and the ability to identify actives that were not flagged by chemistry-based machine learning (see section on Results for the Oncology Project). Moreover, image-based fingerprinting is a feasible approach to predict the activity of not just small-molecule compounds but any agent, such as antibodies, RNA interference agents or other biologics.

We also anticipate that improvements in the computational pipeline may increase the power of the method. For example, CNNs could predict activity from raw images directly rather than from features extracted from each cell using classical image processing. This would allow the model to learn the best image features for the specific task at hand and may improve results. Another future direction is to maintain the native single-cell resolution of image-based profiles instead of aggregating values. Finally, our current results are based on a single HTI screen, and we envision that data fusion across a collection of multiple HTI screens could even be more powerful for assay activity prediction, which we aim to explore in follow-up work.

Our results also encourage the creation of sufficiently large public datasets of compounds annotated with chemical structures, activity measurements in validated assays, and images. While a few efforts have publicly documented up to about 30,000 compounds with cellular images (Wawer et al., 2014), only a 10th of the compounds have been annotated with some assay activities, yielding a very sparse annotation matrix.

SIGNIFICANCE

High-throughput imaging is an affordable screening technology most often used to read out a handful of morphological features that document a single biological process of

interest. Leveraging access to a large private set of activity and image-annotated compounds, we here establish proof of concept that images from one given cellular assay support activity prediction across a spectrum of seemingly unrelated biological assays. Hence, images can inform on biological activity far beyond the intended focus of the original screen. Once a chemical library is documented with image-based fingerprints, a medium-scale screening in an expensive or tedious assay may suffice to train an image-based model that can predict the outcome for the rest of the library and enable cost-effective targeted experimental validation. Effective predictive approaches that rely on the chemical structure of compounds are well established in the context of the gradual virtualization of screening and drug discovery. Our study suggests image-based approaches can complement these structure-based ones, particularly in those cases where the latter suffer from chemical biases in training data. Moreover, they can extend predictive modeling options to agents with (bio) chemistry that elude standard structure-based approaches, such as antibodies, RNA interference agents, and other biologics. Importantly, given that the field of structure-based prediction already exploits decades of optimization and research, the pace of predictive performance gain has slowed down. In contrast, advancements like convolutional neural networks have recently boosted the performance of generic image-based machine learning. The proof of concept described in this paper justifies further research in optimizing the specific application of image-based machine learning in drug discovery. Future lines of research may aim to maximize the generic informativity by screen design or by data fusion over pre-existing screens that cover a broader range of biological contexts, and to improve feature extraction and additional learning from microscopy images.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Cell Lines
- METHOD DETAILS
  - ○ Experimental Setup of GR Assay
  - ○ Autofluorescence Filtering and CNS Availability
  - ○ Software Implementation, Training, and Tuning DNN
  - ○ Random Forests and k-nearest Neighbor
  - ○ Method Performance
- QUANTIFICATION AND STATISTICAL ANALYSIS
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes three figures, three tables, and one data file and can be found with this article online at https://doi.org/10.1016/j.chembiol.2018.01.015.

ACKNOWLEDGMENTS

This work was supported by research grants IWT135122 ChemBioBridge, IWT130405 ExaScience Life Pharma, and IWT150865 Exaptation from the Flanders Innovation and Entrepreneurship agency. The NVIDIA Corporation generously donated a GPU. J.S., A.A., and Y.M. were additionally supported by Research Council, KU Leuven: CoE PFV/10/016 SymBioSys, PhD grants and imec strategic funding 2017, and A.E.C. by the NIH (NIH R35 GM122547).

AUTHOR CONTRIBUTIONS

J.S., G.K., A.A., S.H., Y.M., and H.C. conceived this study and designed the experiments. J.S., G.K., A.A., M.S., J.K.W., E.G., V.C., Y.T.C., J.V., P.B., I.V., A.V., S.S., A.E.C., R.W., and H.C. conducted the experiments and interpreted data. S.H., Y.M., and H.C. supervised this project. J.S., G.K., A.A., A.E.C., and H.C. wrote the manuscript with input from all authors.

DECLARATION OF INTERESTS

P.B., H.C., Y.T.C., E.G., A.V., J.V., I.V., and J.K.W. are or were at the time of writing employees of Janssen Pharmaceutica. Parts of the work by A.A., G.K., J.S., M.S., S.S., and R.W. was supported by funding by or involving Janssen Pharmaceutica. R.W. is also affiliated with the KULeuven.

Received: June 23, 2017
Revised: October 31, 2017
Accepted: January 29, 2018
Published: March 1, 2018

REFERENCES

Ansbro, M.R., Shukla, S., Ambudkar, S.V., Yuspa, S.H., and Li, L. (2013). Screening compounds with a novel high-throughput ABCB1-mediated efflux assay identifies drugs with known therapeutic targets at risk for multidrug resistance interference. PLoS One *8*, e60334.

Baell, J.B., and Holloway, G.A. (2010). New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. J. Med. Chem. *53*, 2719–2740.

Baumann, D., and Baumann, K. (2014). Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. J. Cheminformatics *6*, 47.

Breiman, L. (2001). Random forests. Mach. Learn. *45*, 5–32.

Caicedo, J.C., Singh, S., and Carpenter, A.E. (2016). Applications in image-based profiling of perturbations. Curr. Opin. Biotechnol. *39*, 134–142.

Carpenter, A.E., Jones, T.R., Lamprecht, M.R., Clarke, C., Kang, I.H., Friman, O., Guertin, D.A., Chang, J.H., Lindquist, R.A., and Moffat, J. (2006). CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol. *7*, R100.

Caruana, R. (1997). Multitask learning. Mach. Learn. *28*, 41–75.

Cumming, J.G., Davis, A.M., Muresan, S., Haeberlein, M., and Chen, H. (2013). Chemical predictive modelling to improve compound quality. Nat. Rev. Drug Discov. *12*, 948.

Egan, W.J., Merz, K.M., and Baldwin, J.J. (2000). Prediction of drug absorption using multivariate statistics. J. Med. Chem. *43*, 3867–3877.

Evensen, L., Link, W., and Lorens, J.B. (2010). Imaged-based high-throughput screening for anti-angiogenic drug discovery. Curr. Pharm. Des. *16*, 3958–3963.

Garg, P., and Verma, J. (2006). In silico prediction of blood brain barrier permeability: an artificial neural network model. J. Chem. Inf. Model. *46*, 289–297.

Gustafsdottir, S.M., Ljosa, V., Sokolnicki, K.L., Wilson, J.A., Walpita, D., Kemp, M.M., Seiler, K.P., Carrel, H.A., Golub, T.R., and Schreiber, S.L. (2013). Multiplex cytological profiling assay to measure diverse cellular states. PLoS One *8*, e80999.

Held, M., Schmitz, M.H., Fischer, B., Walter, T., Neumann, B., Olma, M.H., Peter, M., Ellenberg, J., and Gerlich, D.W. (2010). CellCognition: time-resolved phenotype annotation in high-throughput live cell imaging. Nat. Methods *7*, 747–754.

Hochreiter, S., and Obermayer, K. (2004). Gene selection for microarray data. In Methods in Computational Biology, B. Schölkopf, K. Tsuda, and J.P. Vert Kernel, eds. (MIT Press), pp. 319–355.

Louppe, G. (2014). Understanding random forests: from theory to practice. arXiv, arXiv:14077502.

Mayr, A., Klambauer, G., Unterthiner, T., and Hochreiter, S. (2016). DeepTox: toxicity prediction using deep learning. Front. Environ. Sci. *3*, https://doi.org/10.3389/fenvs.2015.00080.

Oshiro, T.M., Perez, P.S., and Baranauskas, J.A. (2012). How many trees in a random forest? Paper presented at: MLDM (Springer).

Pepperkok, R., and Ellenberg, J. (2006). High-throughput fluorescence microscopy for systems biology. Nat. Rev. Mol. Cell Biol. *7*, 690.

Polishchuk, P.G., Muratov, E.N., Artemenko, A.G., Kolumbin, O.G., Muratov, N.N., and Kuz'min, V.E. (2009). Application of random forest approach to QSAR prediction of aquatic toxicity. J. Chem. Inf. Model. *49*, 2481–2488.

Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. J. Chem. Inf. Model. *50*, 742–754.

Simm, J., Arany, A., Zakeri, P., Haber, T., Wegner, J.K., Chupakhin, V., Ceulemans, H., and Moreau, Y. (2017). Macau: scalable Bayesian factorization with high-dimensional side information using MCMC. Proceedings of 2017 IEEE International Workshop on Machine Learning for Signal Processing. IEEE. https://doi.org/10.1109/MLSP.2017.8168143.

Singh, S., Carpenter, A.E., and Genovesio, A. (2014). Increasing the content of high-content screening: an overview. J. Biomol. Screen. *19*, 640–650.

Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. *15*, 1929–1958.

Starkuviene, V., and Pepperkok, R. (2007). The potential of high-content high-throughput microscopy in drug discovery. Br. J. Pharmacol. *152*, 62–71.

Walter, T., Shattuck, D.W., Baldock, R., Bastin, M.E., Carpenter, A.E., Duce, S., Ellenberg, J., Fraser, A., Hamilton, N., and Pieper, S. (2010). Visualization of image data from cells to organisms. Nat. Methods *7*, S26–S41.

Wang, Y., Xing, J., Xu, Y., Zhou, N., Peng, J., Xiong, Z., Liu, X., Luo, X., Luo, C., and Chen, K. (2015). In silico ADME/T modelling for rational drug design. Q. Rev. Biophys. *48*, 488–515.

Wawer, M.J., Li, K., Gustafsdottir, S.M., Ljosa, V., Bodycombe, N.E., Marton, M.A., Sokolnicki, K.L., Bray, M.-A., Kemp, M.M., and Winchester, E. (2014). Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling. Proc. Natl. Acad. Sci. USA *111*, 10911–10916.

Wright, M.N., and Ziegler, A. (2015). Ranger: a fast implementation of random forests for high dimensional data in C++ and R. arXiv, arXiv:150804409.

Yarrow, J., Feng, Y., Perlman, Z., Kirchhausen, T., and Mitchison, T. (2003). Phenotypic screening of small molecule libraries by high throughput cell imaging. Comb. Chem. High Throughput Screen. *6*, 279–286.

Please cite this article in press as: Simm et al., Repurposing High-Throughput Image Assays Enables Biological Activity Prediction for Drug Discovery, Cell Chemical Biology (2018), https://doi.org/10.1016/j.chembiol.2018.01.015

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, Peptides, and Recombinant Proteins		
RPMI with phenol red (for complete medium)	Sigma-Aldrich	Cat# D5671
RPMI medium without phenol red (for CSS medium)	Sigma-Aldrich	Cat# D1145
FCS serum, heat inactivated	Hyclone	Cat# SV30160.03
charcoal stripped serum (CSS)	Sigma-Aldrich	Cat# F6765 lot 12H047
10.000 U/ml penicillin 10.000 μg/ml streptomycin	Gibco	Cat# 15070-063
L-glutamine 200 mM stock	Sigma-Aldrich	Cat# G7513
100mM sodium pyruvate	Sigma-Aldrich	Cat# S8636
0.05% trypsin-EDTA (1x)	Gibco	Cat# 25300-054
Hydro-cortisone (H-cortisone)	Sigma-Aldrich	Cat# H4001
Formaldehyde 10%	Polysciences	Cat# 04018
Triton X-100	Sigma-Aldrich	Cat# T-9284
10x PBS wo Ca/Mg	Roche	Cat# 11666789001
goat serum	Sigma-Aldrich	Cat# G9023
GR Ab H-300 rabbit	Santa Cruz	Cat# sc-8992; RRID: AB_2155784
Alexa Fluor 568 goat anti-rabbit	Invitrogen	Cat# A11011; RRID: AB_143157
Hoechst 33528	Invitrogen	Cat# H3569; RRID: AB_2651133
HCS CellMask Deep Red stain	Invitrogen	Cat# H32721
Experimental Models: Cell Lines		
H4 Homo sapiens brain neuroglioma	ATCC	Cat# HTB-148; RRID: CVCL_1239
Software and Algorithms		
scikit-learn 0.18.2	scikit-learn project	http://scikit-learn.org/
CellProfiler	CellProfiler team	http://cellprofiler.org/
Pipeline for extracting imaging features	This paper	https://github.com/ExaScience/process-plate
Bayesian matrix factorization Macau	This paper	https://github.com/jaak-s/macau
Deep neural network (DNN) code	This paper	https://github.com/gklambauer/nnet-gmatrix

CONTACT FOR REAGENT AND RESOURCE SHARING

Requests for information and resources should be directed to the lead contact, Hugo Ceulemans (hceulema@its.jnj.com). Due to the proprietary nature of the drug development process, we are unable to disclose specific information related to the chemical compounds and specific protein targets. We gladly share all used software code, and laboratory and imaging methodologies.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Cell Lines

The H4 brain epithelial neuroglioma cell line, originating from a male patient, was used for the cellular imaging assays. Authenticated cell stocks were obtained from ATCC.

Cells were cultured in Flacon T175 flasks at 37C using DMEM medium with phenol red, with addition of 10% fetal calf serum, 100 U/ml penicillin, 100 μg/ml streptomycin, 2 mM glutamine, and 1 mM sodium pyruvate.

METHOD DETAILS

Experimental Setup of GR Assay

For the assay, the cells were trypsinized and harvested from the T175 flasks, then resuspended in DMEM without phenol red, 5% charcoal-stripped serum (CSS), 100 U/ml penicillin, 100 μg/ml streptomycin, 2 mM glutamine, and 1 mM sodium pyruvate.

They were seeded in PerkinElmer CellCarrier 384-well plates at a density of 2000 cells per well in a 30 μl volume, and incubated at 37C for 16-18 hours to promote cell attachment. Then 3 μl of compounds dissolved in DMSO and pre-diluted in PBS were added, to a final concentration in the well of 10 μM with 0.2% final DMSO content. After 1 hour of incubation at 37 C, 3 μl of 12 μM hydrocortisone in PBS was added, to a final concentration of 1 μM to stimulate translocation of the GCR, except in positive control wells where pure PBS was added. After 1 hour of incubation at 37C, the process was stopped by fixation.

Cells were fixed by addition of 5% formaldehyde for 15 minutes at room temperature (RT), washed, and permeabilized with 0.3% Triton-X for 10 minutes at RT. The plates were then washed 3 times, and 20 μl of the primary antibody for the GCR (Santa Cruz sc-8992) was applied at a dilution of 1/200 in PBS with 5% goat serum, and left on for 24 hours at 4 C. Thereafter, plates were washed again three times, and 20 μl of a reagent mix in PBS with 5% goat serum was added, which contained Hoechst 33258 (Invitrogen H3569, dilution 1/5000) to label the nucleus, CellMask Deep Red (Invitrogen H32721, dissolved in 100 μl DMSO, then diluted 1/4000) to delineate cell boundaries, and an Alexa-568 labeled goat anti-rabbit secondary antibody (Invitrogen A11011, 1/500) to detect the GCR. This was left on for 1 hour at RT. Plates were then washed two times, filled with 80 μl of PBS, sealed, and stored at 4C.

Plates were imaged at RT on a Yokogawa CellVoyager 7000, at 10x magnification, acquiring 2 fields per well. For Hoechst, a 405-nm laser was used and a 445/45 bandpass emission filter; for Alexa 568 a 561-nm excitation and a 600/37 filter, and for CellMask Deep Red a 635-nm laser and a 676/29 filter.

Autofluorescence Filtering and CNS Availability

Frequent hitters are compounds that are promiscuously active, e.g. based on certain substructure motifs they might contain. Also, some compounds might be dyes themselves, be reactive species, or interfere with particular assay technologies as Fluorescent or AlphaScreen readouts. Baell/Holloway (Baell and Holloway, 2010) suggested a Pan Assay Interference Compounds (PAINS) filter for removing such promiscuous compounds from HTS hits.

The Blood-Brain-Barrier (BBB) is a critical membrane to separate the blood from the brain in the central nervous system (CNS). Drugs for CNS disease indications should pass the BBB, while drugs for non-CNS indications should not pass the BBB for preventing unwanted side-effects. The BBB allows the passage of water and lipid-soluble molecules by passive diffusion. Two major estimations for BBB permeability are therefore based on passive diffusion models based on logP and polar surface area (PSA) of compounds (Egan et al., 2000), or active transport via a P-glycoprotein (P-gp) substrate probability of compounds (Garg and Verma, 2006; Wang et al., 2015). We filtered out all compounds that do not exhibit BBB permeability according to standard pharmaceutical practice.

Software Implementation, Training, and Tuning DNN

We used minibatch Stochastic Gradient Descent (SGD) to train the DNNs. Hence, we implemented the DNNs using the CUDA parallel computing platform and employed NVIDIA Tesla K40 GPUs to achieve speed-ups of 20-100x compared to CPU implementations.

We optimized the DNN architecture and hyperparameters, such as learning rate, early-stopping parameter and Dropout rate on a validation set in a nested cross-validation procedure (Baumann and Baumann, 2014; Hochreiter and Obermayer, 2004). This procedure produces performance estimates that are unbiased by hyperparameter selection since the hyperparameters are optimized on the inner folds and only the best performing hyperparameters are tested on the outer folds. We considered 1, 2 or 3 layer networks with 1024, 2048, or 4096 units in each layer. The tested learning rates were 0.01, 0.05, and 0.1. The Dropout rates were either set to zero, or to 10% dropout in the input layer and 50% dropout in the hidden layers. Additionally, we tested whether the dropout rate should be arithmetically increased from 0 by 0.005 after each epoch ("dropout schedule") until the given dropout rate or whether the dropout rates were constant ("no dropout schedule") during learning. Table S1 summarizes these hyperparameters and architecture design parameters that were used for the DNNs, together with their search ranges.

The hyperparameters that performed best when averaged across the three cross-validation folds were: 3 layers with 2,048 units, learning rate 0.05, Dropout: yes, Dropout-schedule: yes. The early stopping-parameter was determined to be 63 epochs.

Random Forests and k-nearest Neighbor

Random forests (RF) work well with different types of descriptors (Breiman, 2001) at a large variety of tasks and their performance is relatively robust with respect to hyperparameter settings (Polishchuk et al., 2009). We used a high number of trees to obtain a stable model with high performance (Oshiro et al., 2012). The critical parameter is the number of features considered at each split (Louppe, 2014) which we adjusted in the established nested cross-validation setting. We trained and assessed models for each assay individually in our established framework using different hyperparameters given in Table S2 and the Random Forest implementation "ranger" (Wright and Ziegler, 2015).

The k-nearest neighbor (kNN) method is a popular approach for similarity search based predictions. We applied kNN to measure how well a similarity search ranking would work on images. The number of neighbors *k* is chosen for each assay-threshold pair using the nested cross-validation, and the considered values of *k* were 7, 13, 21 and 33.

Method Performance

We compared Macau, a regression method based on Bayesian matrix factorization with side information, random forest classification (Breiman, 2001), kNN and deep neural networks (Mayr et al., 2016) for predictive performance on assay-threshold pairs with at least 25 actives and 25 inactives, see Data S1 (an external spreadsheet) for the detailed results.

To summarize the results we count how many assays had at least one threshold (5.5, 6.5, 7.5 or 8.5) with AUC-ROC above 0.9 threshold. We found that the Macau, DNN and RF performed similarly with respect to which assays could be predicted accurately. Concretely, 10 out of total 535 assays could be predicted with AUC-ROC above 0.9 by all three methods (see Venn diagram in Figure S2). Similarly, 181 assays had performance of AUC-ROC above 0.7 by all three methods (see Venn diagram in Figure S3). The numbers of assays where only a single or a pair of methods gave an AUC-ROC above 0.7 are all comparably smaller. Therefore, we conclude that the performance is mainly driven by imaging features rather than the machine learning method.

However, kNN was not able to predict a single assays with AUC-ROC above 0.9 and predicted only 93 assays with AUC-ROC above 0.7. A likely reason for the poor performance is that the image-based fingerprints are quite high-dimensional and noisy, which makes similarity-based ranking inaccurate. In contrast, the other tested methods (Macau, DNN and RF) learn to focus only on specific features of the fingerprints and thus pick up the signal.

For summary, Table S3 reports the number of protein targets that are well predicted and the mean AUC-ROC for the four methods over the 535 protein targets.

QUANTIFICATION AND STATISTICAL ANALYSIS

The details of the clustered cross-validation scheme are presented in Section “Selection of High Quality Models” and depicted in Figure S1. For the hyperparameter selection of the DNN see STAR Methods Section “Software Implementation, Training, and Tuning DNN” and Table S1. For random forest tuning see STAR Methods Section “Random Forests and k-nearest Neighbor” and Table S2.

The AUC-ROC values presented were calculated using scikit-learn.

DATA AND SOFTWARE AVAILABILITY

We offer all developed software and pipelines under open source licences.

- Macau implementation: https://github.com/jaak-s/macau
- DNN implementation: https://github.com/gklambauer/nnet-gmatrix
- CellProfiler pipeline: https://github.com/ExaScience/process-plate

Data S1 describes the AUC-ROC values for the assays-threshold pairs with at least 25 active and 25 inactive compounds.