# Repurposed high-throughput images enable biological activity prediction for drug discovery

Jaak Simm[1*] , Günter Klambauer[2*], Adam Arany[1*], 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 Carpenter[6], Roel Wuyts[7], Sepp Hochreiter[2#] , Yves Moreau[1#], Hugo Ceulemans[4#+]

[*] Shared first authors
[#] Shared last authors
[+] Corresponding author
[1] ESAT-STADIUS, KU Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium
[2] Institute of Bioinformatics, Johannes Kepler University Linz, Altenbergerstr 69, 4040 Linz, Austria
[3] Open Analytics NV, Jupiterstraat 20, 2600 Antwerp, Belgium
[4] Janssen Pharmaceutica NV, Turnhoutseweg 30, B-2340 Beerse, Belgium
[5] Facilities for Research, KU Leuven, Willem de Croylaan 52c, box 5580, 3001 Leuven
[6] Imaging Platform, Broad Institute of Harvard and MIT, 415 Main St, Cambridge, MA 02142, USA
[7] ExaScience Life Lab, IMEC, Kapeldreef 75, B-3001 Leuven, Belgium

**We repurpose a High-Throughput (cell) Imaging (HTI) screen of a glucocorticoid receptor assay to predict target protein activity in multiple other seemingly unrelated assays. In two ongoing drug discovery projects, our repurposing approach increased hit rates by 60- to 250-fold over that of the primary project assays while increasing the chemical structure diversity of the hits. Our results suggest that data from available HTI screens are a rich source of information that can be reused to empower drug discovery efforts.**

High-throughput (cell) Imaging (HTI) captures the morphology of the cell and its organelles by high-throughput microscopy and is successfully applied in many areas of current biological research (Walter *et al.* (2010), Pepperkok *et al.* (2006), Starkuviene and Pepperkok (2010)). In a pharmacological context, HTI is applied to screen chemical compounds based on morphological changes they induce (Yarrow *et al.* (2003), Held *et al.* (2010)). Currently, most HTI screens are designed for the single purpose of evaluating one specific biological process and exploit only a handful of morphological features from the image (Singh et al. (2014)), see **Fig. 1b**. These morphological features are understood to directly reflect that biological

process. However, any imaged 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 impact cell morphology and that morphology can to a large extent be extracted from the images (Carpenter et al. (2006)). This set of features can be used to describe chemical compounds and can be considered as an image-based fingerprint. Wawer et al. (2014) proposed the use of image-based fingerprints to optimize the diversity of medium scale compound sets. Image-based fingerprints can also be used to group compounds by pharmacological mechanism (Caicedo et al. (2016)). Thus, images provide a rich source of biological information that can be leveraged for a variety of purposes in drug discovery.

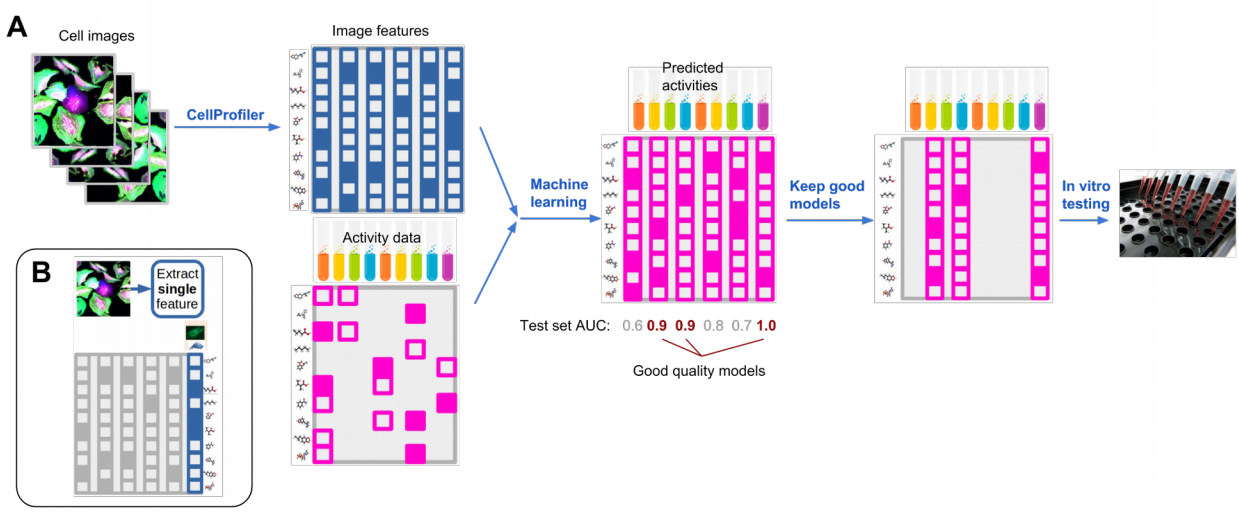


***Figure 1:*** *Repurposing imaging screens. Panel A: Our repurposing approach is depicted. A large number of features are extracted from images of cells which are then used by machine learning methods to model all available activity data from previously performed assays. Targets with good predictivity on the test data are then selected for in vitro validation. Panel B: A typical HTI screen approach is depicted (Evensen et al. (2010), Ansbro et al. (2013)). Few or single features are extracted from cellular images.*

Here, we propose the systematic evaluation of image-based fingerprints from HTI screens for predictivity on a large collection of protein targets, most of which were not considered during the design of the screen (**Fig. 1a**). To this end, an extensive fingerprint of morphological features was extracted for each compound imaged in a single screen, aiming for maximal and unbiased information capture. We then deployed a machine learning approach to predict the activity across a broad set of validated target assays (from now referred to as assay) based on the image-based fingerprint of the compounds, and evaluated model performance for each assay. In this way, we hypothesized existing HTI screens can be repurposed to inform on the

activity of untested compounds in assays for which a high-quality model is available (**Fig. 1a**).

This procedure is reminiscent of the predictive modeling approaches used for virtual screening and QSAR, but differs in that it uses image-based fingerprints rather than chemical fingerprints that encode the structure of compounds. Chemistry-based models are predictively performant, but only for those parts of chemical space for which sufficient assay activity data is available, because chemical fingerprints themselves do not incorporate biological or target information. Hence, compounds that are chemically very different from any known active compound are unlikely to be predicted as active. Image-based models are expected to be less dependent on the availability of chemically similar training examples. The model then correlates all imaged biology to the biological activities to predict. Therefore, image-based models could outperform chemistry-based models in novel and activity-wise poorly annotated chemical space.

In our study, we repurposed a high-throughput imaging screen of 524,371 compounds originally used for the detection of glucocorticoid receptor (GCR) nuclear translocation. Each compound was applied in a concentration of 10µM to H4 brain neuroglioma cells, incubated for one hour, before the addition of 1µM hydrocortisone to stimulate translocation of the GCR. After an additional 1 hour of incubation, cells were fixed and imaged in 3-channel fluorescence, with a nuclear stain (Hoechst), CellMask Deep Red (Invitrogen) to delineate cell boundaries, and indirect immunofluorescence detection of GCR. For repurposing the screen, the images were post-processed using CellProfiler software. Using a pipeline similar to that of Gustafsdottir and colleagues (2013), we extracted unbiased maximally-informative features from the images. For each cell in the image, the pipeline computed an image-based fingerprint of 842 features. In our machine learning models, each compound was then represented by the vector of feature medians across all cells in the image.

We then built a machine-learning model using Bayesian matrix factorization with the image-based fingerprint as side information (Online Methods). This model was evaluated for its predictivity across assays that map to more than 600 drug targets leveraging more than ten million activity measurements. We assessed the performance of this model using nested cluster cross-validation (Mayr, 2016). The model was predictive for 37.3% of the assays (AUC > 0.7 on 225 assays, Online Methods) and offered high predictivity for 5.6% of the assays (AUC > 0.9 on 34 assays, Online Methods). Among these 34 assays with high-quality predictions, two were running in ongoing discovery projects: one oncology project and one central nervous system (CNS) project. We used the matrix factorization model to select compounds for testing by these two projects.

For the oncology project, the target was a kinase with no known direct relation to the glucocorticoid receptor. Using our matrix factorization model, we ranked about

60,000 compounds tested in the GR assay but for which no activity measurement was available in the oncology screen. We then experimentally validated the 342 compounds ranked highest by our matrix factorization method, roughly the amount of non-control wells on a plate. This resulted in 141 submicromolar hits (41% hit rate), which corresponds to a 60-fold enrichment over the initial HTS (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 HTS. When compared to that of the initial hits (red distribution in **Fig. 2a**), the distribution of these similarities implies a substantially improved chemical structure diversity (shift to the left) of the image-based hits (green). As a reference, the figure also depicts the distribution for randomly selected compounds (blue). Thus, the HTI matrix factorization model selected candidate compounds with a high hit rate, and diversified the hit space.

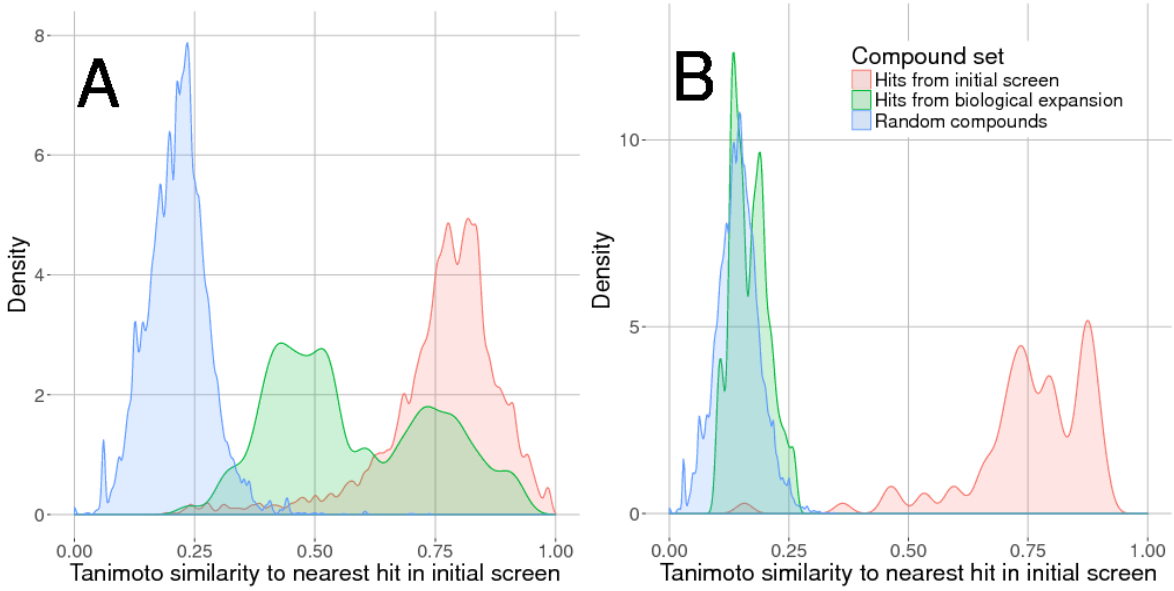


***Figure 2:*** *For each compound in a set, the ECFP (radius 4) based Tanimoto similarity to the nearest hit from the initial screen was considered. Similarity densities for the set of initial hits, the set of hits from our biological extension and a randomly selected set of compounds are depicted in red, green and blue, respectively, for the oncology project in Panel A and for the CNS project in Panel B. 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. Overall, the hits discovered by our approach are chemically highly diverse.*

For the CNS project, the target was a non-kinase enzyme again without obvious relation to the glucocorticoid receptor. Using our matrix factorization model, activity was predicted for all 500,000 image-annotated compounds. Compounds with

predicted submicromolar activity were filtered to deplete for unfavorable properties, like autofluorescence and low predicted central nervous system availability (Online Methods), and the remaining compounds were grouped in chemical clusters from which we randomly selected a handful of representatives from each cluster (Online Methods). The 141 compounds resulting from this procedure were experimentally tested, and for 37 of them, submicromolar activity was confirmed, resulting in a 22.7% hit rate or a 250-fold enrichment over the hit rate of the initial HTS (0.088%). Importantly, the 37 hits mapped to 32 Murcko scaffolds that were not represented in the set of initial hits. The distribution of Tanimoto similarities to the nearest hit in the initial screen (**Fig. 2b**) supports that conclusion. These results again suggested that an image-empowered compound selection strategy can boost hit rate and hit diversity.

To check whether the success of our approach arises from the machine learning method or from the description of chemical compounds by imaging features, we applied three different machine learning methods. We used Macau (Simm, 2015), a regression method based on Bayesian matrix factorization with side information (Online Methods), random forest classification (Breiman, 2001) and deep neural networks (Mayr, 2016) (Online Methods). These machine learning models performed similarly in terms of the assays that could be predicted accurately (**Supplementary Fig. S4 and S5**). Our cross-validation setup also guarantees that the predictive performance does not come from the activity data of the same compound across other targets (Online Methods). Therefore, we conclude that the description of compounds by imaging features is the essential contribution to the success of our approach.

We emphasize that the method is a supervised machine learning method and hence output labels (in this context, activity measurements) are required to train the model. This requires that activity measurements be acquired for a reasonably sized library of compounds.

Our results indicate that images from HTI screening projects that are conducted in many institutions can be repurposed for increasing hit rates in other projects, even those that seem unrelated to the primary purpose of the HTI screen. Consequently, it might be possible to replace particular assays with the potentially more cost-efficient imaging technology together with machine learning models. By accessing rich morphological features of the cell, imaging screens provide a view over various cellular processes resulting in a fingerprint of biological action. This raises an interesting question of the breadth of targets that could be accessed by imaging screens if different cell lines, culturing conditions, staining of organelles and/or incubation times are used.

The focus of this report was to demonstrate that the use of HTI data enables the identification of diverse hits without the need to retest the entire library in the target assay. We note that our models may also support target deconvolution for

phenotypic screens, through the prioritization of targets with predicted activities that match phenotypic observations.

Moreover, in the light of recent advances of *convolutional neural networks,* raw images might be used directly in the activity prediction pipeline. This would allow the model to learn the best image features for the specific task at hand and may strengthen our approach. Furthermore, our results are based on a single HTI screen and we envision that a collection of multiple HTI screens could even be more powerful with respect to assay activity prediction. Finally, our imaging features are median values across all cells from an image. However, models based on the distribution of the feature values (*e.g.* quantiles) or even single cell analysis could prove to have higher predictive power and will be investigated in the near future.

ACKNOWLEDGEMENTS

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.

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.C., R.W., and H.C. conducted the experiments. S.H., Y.M., and H.C. supervised this project. J.S., G.K., A.A. and H.C. wrote the manuscript with input from all authors.

Ansbro, M.R., Suneet S., Suresh V.A., Stuart H.Y., and Luowei L. *PLoS One*. **8(4)**, e60334 (2013).

Baell, J.B. & Holloway, G.A. *J. Med. Chem*. **53**, 2719-2740 (2010).

Bemis, G.W. & Murcko, M.A. *J. Med. Chem*. **39**, 2887-2893 (1996).

Breiman, L. Random forests. *Machine learning*, ***45.1***, 5-32 (2001).

Caicedo, J.C., Singh S. & Carpenter A.E. *Curr. Opin. Biotechnol.* **39**, 134-142 (2016).

Carpenter, A.E. *et al*. *Genome Biol*. **7**, R100 (2006).

Evensen, L., Link W., & Lorens B.J. *Curr. Pharm. Des.* **16.35**, 3958-3963 (2010).

Gustafsdottir, S.M. *et al*. *PLoS ONE*. **8.12**, e80999 (2013).

Held, M. *et al*. *Nature methods.* ***7.9***, 747-754 (2010).

Mayr, A., Klambauer, G., Unterthiner, T., & Hochreiter, S. *Front. Environ. Sci*. **3**, 80 (2016).

Pepperkok, R. & Ellenberg, J. *Nat. Rev. Mol. Cell Biol.* **7.9**, 690-696 (2006).

Rogers, D. & Hahn, M. *J. Chem. Inform. Model.* **50.5**, 742-754 (2010).

Simm, J., Arany, A., Zakeri, P., Haber, T., Wegner, J.K., Chupakhin, V., Ceulemans H., Moreau Y. *arXiv:1509.04610v2*.

Singh, S., Carpenter, A.E. & Auguste, G. *J. Biomol. Screen*. **19.5**, 640-650 (2014).

Starkuviene, V., & R. Pepperkok. *Br. J. Pharmacol.* **152.1**, 62-71 (2007).

Yarrow, J.C., Feng, Y., Perlman, Z.E., Kirchhausen, T. & Mitchison, T.J. *Comb. Chem. High Throughput Screen*. ***6.4***, 279-286 (2003).

Walter, T. *et al*. *Nature methods*. **7**, S26-S41 (2010).

Wawer, M.J., *et al*. *PNAS*. **111.30**, 10911-10916 (2014).