IJAS

Repurposing High-Throughput Imaging Tests for Drug Discovery Allows for Biological Activity Prediction

Mani Manavalan[1], Naresh Babu Bynagari[2]

[1]*Sr. Manager, Capgemini America, 79 5th Avenue, Suite 300, New York, NY 10003, USA*
[2]*Director of Sales, Career Soft Solutions Inc, 145 Talmadge rd Edison NJ 08817, Middlesex, USA*

Abstract: A glucocorticoid receptor High-Throughout (cell) Imaging (HTI) screen to predict target protein activity in a variety of seemingly unrelated tests. The repurposing method raised hit rates by 60 to 250 times over the primary project assays in two ongoing drug discovery projects, while also boosting the chemical structure diversity of the hits using machine learning. Our findings show that data from existing HTI screens is a significant source of data that may be repurposed to aid drug discovery efforts.

Keywords: High-Throughput imaging, Drug discovery, Biological Activity prediction, Machine learning

1. INTRODUCTION

High-throughput (cell) Imaging (HTI) is a technique that uses high-throughput microscopy to record the morphology of a cell and its organelles. It has been effectively used in a variety of current biological research areas (Walter, 2010; Pepperkok and Ellenberg, 2006; Ahmed et al., 2021; Starkuviene and Pepperkok, 2007). HTI is used to screen chemical compounds based on the morphological changes they cause in a pharmacological environment (Yarrow et al., 2003; Held, 2010). Most HTI screens are currently intended to evaluate only a few morphological aspects from an image (Singh et al., 2014), as seen in Fig. 1b. These physical characteristics are thought to reflect that biological process directly. Any imaging biological system, on the other hand, has thousands of biochemical processes and possible therapeutic targets, all of which are exposed to the screened chemical compounds. Many of these targets and processes have an impact on cell morphology, which may be gleaned to a significant extent from photographs (Ganapathy et al., 2021b). This set of traits can be used to define chemical compounds and can be thought of as a fingerprint based on images. Wawer (2014) suggested that image-based fingerprints be used to improve the diversity of medium-scale compound collections. Compounds can also be grouped by pharmacological mechanism using image-based fingerprinting (Caicedo et al., 2016). As a result, pictures provide a rich source of biological data that can be used for a variety of applications in drug development.
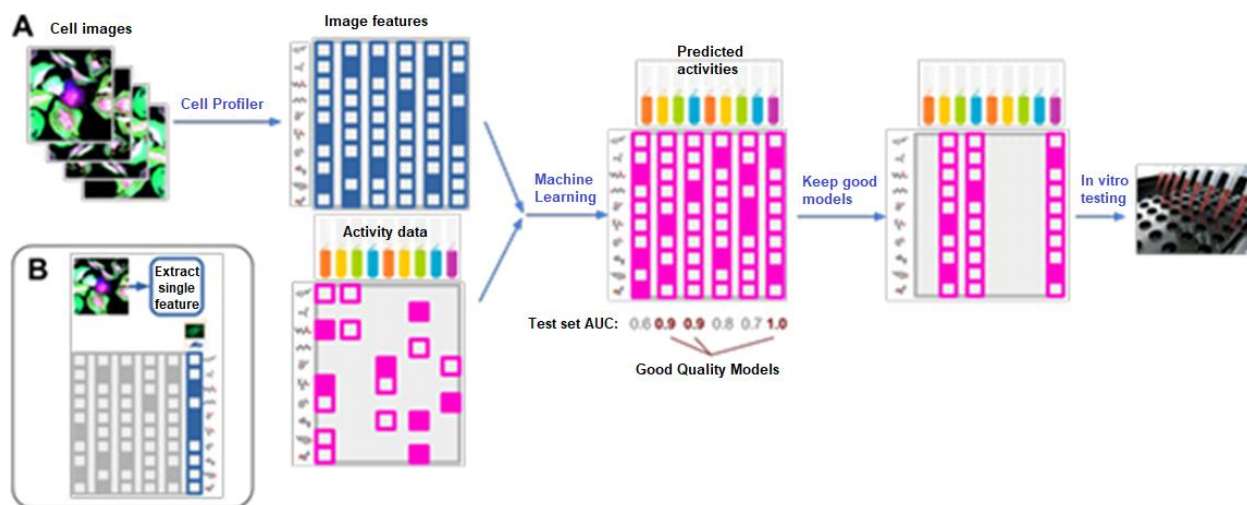
Figure 1: Repurposing imaging screens

A method for systematically evaluating image-based fingerprints from HTI screens for predictability on a large number of protein targets is presented, the majority of which were not considered during the screen's design (Figure 1). Caicedo et al. (2016) perform this by an exhaustive fingerprint of morphological traits was extracted for each compound imaged in a single screen, with the goal of capturing as much information as possible in a single screen. A machine learning approach is used to predict activity across a large number of validated target assays (hence referred to as assay) using the image-based fingerprint of the drugs, and evaluated model performance for each assay (Bynagari & Fadziso, 2018; Donepudi and Bynagari, 2021). Existing HTI screens, we argued, might be repurposed to provide information on the activity of untested drugs in assays for which a high-quality model exists (Fig. 1a).

This process is similar to virtual screening and QSAR predictive modeling approaches, but it differs in that it uses image-based fingerprints rather than chemical fingerprints that encode complex structure (Bynagari, 2018). Because chemical fingerprints do not include biological or target information, chemistry-based models are predictively performant only for those areas of the chemical space for which sufficient assay activity data is available. As a result, molecules that are chemically dissimilar to any known active ingredient are unlikely to be predicted as such. The availability of chemically comparable training samples should be less of a concern for image-based models (Bynagari, 2019; Donepudi, 2015). The model then predicts by correlating all imaged biology to biological processes. As a result, in innovative and activity-wise poorly characterized chemical space, image-based models may outperform chemistry-based models.

2. OBJECTIVES OF THE STUDY

This study is aimed at repurposing high-throughput imaging tests for drug discovery allows for biological activity prediction. To achieve this, a comprehensive pattern of morphological characters will be extracted for each composite imaged in a single screen, with the objective of capturing as much information as possible in a single screen. A machine learning approach is used to predict activity across a large number of validated target assays.

3. LITERATURE REVIEW

Macau: Bayesian matrix factorization

The Macau Bayesian matrix factorization method, which uses image data from Cell Profiler as side information. Macau represents each chemical and each assay by D-dimensional latent vectors u_i and v_j, respectively, to factorize the 524, 370 X 1, 199 activity matrix Y. The scalar product $u_i^T v_j$ gives the prediction for the element Y_{ij}, which corresponds to the activity of compound í on assay j. The picture characteristics x_i are 841 dimensional vectors that are added to the prior of the latent vectors of compounds u_i (Donepudi and Bynagari, 2021).

This outcomes in the probabilistic prototypical of

$Y_{ij} \sim N(u_i^T v_j, \alpha^{-1}$

$u_i \sim N(\mu_u + \beta x_i, \Lambda_u^{-1})$

$v_j \sim N(\mu_u, \Lambda_u^{-1})$

where α is the precision of the observations, μ_u and Λ_u model the mean and precision of compound latent vectors, μ_u and Λ_v model the latent vectors for assays, respectively. The parameter β is a D x 841 dimensional matrix that maps image characteristics to compound latent space (Bynagari and Asadullah, 2021). We use a Gaussian prior on it to learn:

$\beta \sim N(0, (\Lambda_u \otimes \lambda_\beta I_{841})^{-1}$

where $\otimes$ is the Kronecker product, λ_β is the accuracy parameter, and I_{841} is the 841-dimensional identity matrix. The plate diagram for the probabilistic model is shown in Figure 2.
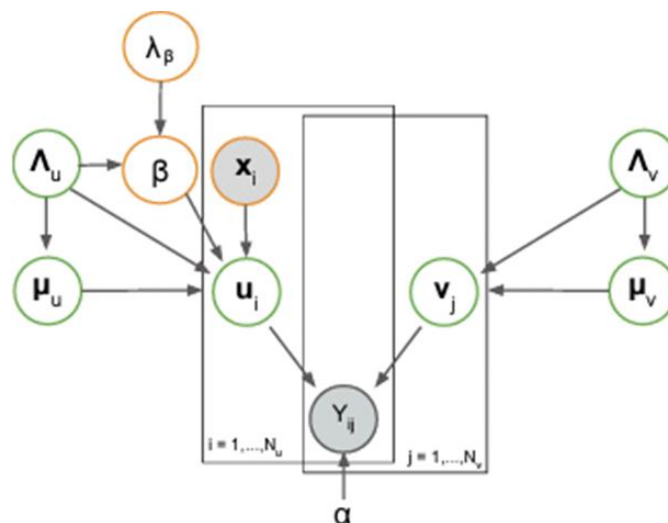

Figure 2: Diagram for the probabilistic model for the used matrix factorization approach Macau

Simm (2015) expresses Gibbs sampler for the model by deriving conditional probabilities. Except for and D, which are fixed at 10 and 150, the Gibbs sampler was utilized to learn the model's parameters. We ran Gibbs sampler for 2000 iterations in the experiments and discarded the first 400 as burn-in. To arrive at the final answer, we construct a prediction for Y_{ij} using each Gibbs sample and then average the results. A 3-fold cross-validation was utilized to assess the model's performance, as explained in the next section.

Cross-validation and measuring predictively

A 3-fold cross-validation was employed in our in silico tests, where all of the compounds in a

chemical series are in a single fold (see Figure 3). The 0.7 Tanimoto similarity cutoff on ECFP6 characteristics was used to define each chemical series (Bynagari, 2021). As a result, no compound in the test set is chemically similar to a compound in the training set. It can assess the model's ability to generalize across series using this arrangement.
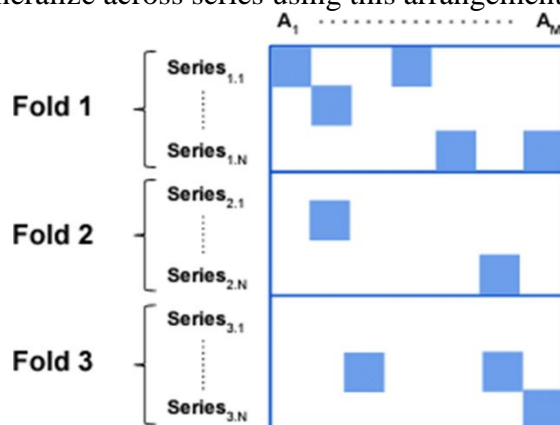

Figure 3: Clustered cross-validation for compound-protein activity prediction

The test AUC-ROC scores was constructed for each cross-validation fold for each target assay at pIC50 levels of 5.5, 6.5, 7.5 and 8.5. For each assay-threshold pair, we utilize the average of the three folds as the assessment metric. Only assay-threshold pairings with at least 25 actives and 25 inactives are kept, and the model is considered to have high (moderate) predictivity on an assay if one of the thresholds has an AUC-ROC of greater than 0.9. (0.7).

Auto fluorescence filtering and CNS availability

Compounds that are promiscuously active, for example, due to the presence of particular substructure motifs, are known as frequent hits (Donepudi, 2016). Some molecules may also be dyes, reactive species, or interfere with specific test technologies such as Fluorescent or Alpha Screen readouts. For removing such promiscuous chemicals from HTS hits, Baell and Holloway (2010) proposed a Pan Assay Interference Compounds (PAINS) filter.

The Blood-Brain-Barrier (BBB) is a crucial barrier in the central nervous system that separates blood from the brain (CNS) (Manavalan, 2019a). For the sake of avoiding undesired side effects, medications for CNS disease indications should traverse the BBB, while drugs for non-CNS disease indications should not. Passive diffusion permits water and lipid-soluble compounds to move through the BBB. Passive diffusion models based on logP and polar surface area (PSA) of chemicals (Egan et al., 2000), or active transport via a P-glycoprotein (P-gp) substrate probability of substances (Wang, 2015; Garg and Verma, 2006; Bynagari, 2020), are the two major estimations for BBB permeability. According to st, we filtered out all chemicals that do not have BBB permeability.

Deep neural networks (DNNs)

Deep Neural Networks (DNNs) were constructed, which are feed-forward artificial neural networks with multiple layers and a huge number of neurons (Mayr et al., 2016). DNNs are made up of layers of linked neurons that are stacked hierarchically. The neurons in the network's first layer (the input layer) receive an input vector containing the chemical compound's descriptors (Donepudi, 2017). The intermediate layers (hidden layers) are made up of hidden neurons with weighted connections to the neurons of the lower level layer, and can be thought of as abstract features constructed from the features below. The model's predictions are provided by the final layer (the output layer). A DNN may have thousands of

neurons in each layer, whereas previous networks used a modest number of neurons (Ciresan et al., 2012; Manojkumar et al., 2021). The general architecture of Deep Neural Networks is shown in Figure 4.
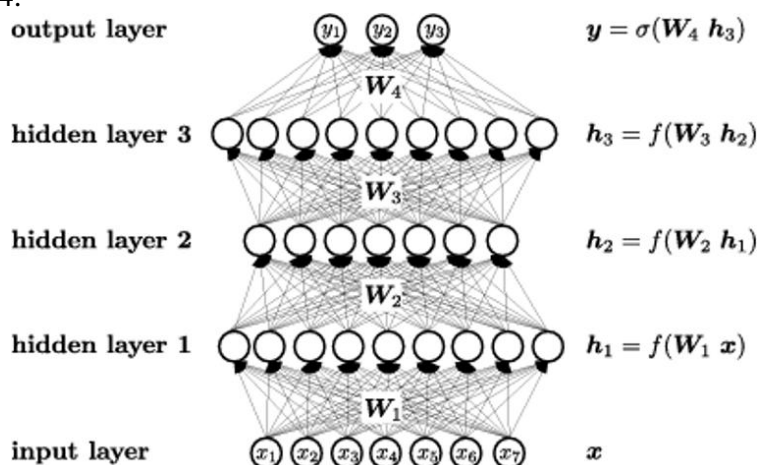

Figure 4: Architecture of Deep Neural Networks

As network inputs, HCI features are used.

Image attributes from Cell Profiler are averaged across several cells to represent each chemical component. These 841 characteristics were fed into the DNNs' input layer.

The function of activation: In the buried layers, we employed rectified linear units (ReLUs) as activation functions. Sigmoid activation functions are used in the output layer.

Regularization: A variety of regularization approaches was used, including Dropout (Srivastava et al., 2014) and early halting, to avoid overfitting. A validation set was used to calculate the dropout rate and the early-stopping parameter, which is the number of epochs after which learning stops (Donepudi, 2018).

Multitasking is a term that refers to the ability to do. The prediction job at hand is a multi-task situation since each chemical component can have several biological actions. Multi-task learning is a natural feature of deep learning (Caruana, 1997). These activities can help each other increase the representation of samples in the hidden layers, resulting in improved performance (Mayr et al 2016). Each assay was modeled using a separate output unit, resulting in DNNs with roughly 1,200 output units.

Objective Purpose For our DNNs, we employed cross-entropy as a loss function:

$$\sum_i \sum_j m_{ij} \left(Y_{ij} \log(Y_{ij}) + (1 - Y_{ij}) \log(1 - Y_{ij}) \right)$$

where Y_{ij} is the chemical I and assay prediction, and Yij is the actual label. In this test, Yij indicates whether the chemical was active ($Y_{ij} = 1$) or inactive ($Y_{ij} = 0$). Furthermore, the activity of many chemicals has not been tested, which accommodated with a binary variable m_{ij} that is zero if a measurement is present and zero otherwise.

DNN training and software implementation

The DNNs were trained using mini batch Stochastic Gradient Descent (SGD). As a result, we used the CUDA parallel computing platform to develop the DNNs using NVIDIA Tesla K40 GPUs to achieve speedups of 20-100x over CPU implementations.

Search with hyper-parameters: In a nested cross-validation approach, we improved the DNN architecture and hyper-parameters such as learning rate, early-stopping parameter, and Dropout rate on a validation set (Baumann and Baumann, 2014; Neogy & Bynagari, 2018; Hochreiter

and Obremayer, 2004). By using hyper-parameter selection, this approach generates unbiased performance estimates.

Because only the top performing hyper-parameters are tried on the outer folds, and the hyper-parameters are optimized on the interior folds. With 1024, 2048, or 4096 units in each layer, we investigated 1, 2, or 3 layer networks. The learning rates that were examined were 0.01, 0.05, and 0.1. The dropout rates in the input layer and the hidden layers were either adjusted to zero or to 10% dropout in the input layer and 50% dropout in the hidden layers. Furthermore, we investigated whether the dropout rate should be increased arithmetically from 0 to the given dropout rate after each epoch ("dropout schedule"), or whether the dropout rates should remain constant ("no dropout schedule") during learning. The hyper-parameters and architecture design parameters that were employed for the DNNs, as well as their search ranges, are summarized in Table 1.

Table 1: Hype-rparameters considered for Deep Neural Networks

Hyperparameter	Considered values
Number of Hidden Units	{1024, 2048, 4096}
Number of Hidden Layers	{1, 2, 3}
Learning Rate	{0.01, 0.05, 0.1}
Dropout	{no , yes (50% Hidden Dropout, 10% Input Dropout)}
Dropout-schedule	{no, yes}

When averaged throughout the three cross-validation folds, the following hyper-parameters performed best: 3 layers with 2, 048 units, learning rate 0.05, Dropout: yes, Dropout-schedule: yes. 63 epochs was found to be the early stopping-parameter.

Random Forests (RF)

Random Forests perform well on a wide range of tasks using various types of descriptors (Breiman, 2001), and their performance is rather robust in terms of hyper-parameter choices (Polishchuk, 2009). To obtain a stable model with great performance, we used a large number of trees (Oshiro et al., 2012). The number of features examined at each split (Louppe, 2014), which we changed in the established nested cross-validation configuration, is the crucial parameter. In our constructed framework, we trained and assessed models for each assay individually using different hyper-parameters listed in Table 2 and the Random Forest implementation "ranger" (Wright and Ziegler, 2015).

Table 2: Hyper-parameters considered for Random Forests

Hyperparameter	Considered values
criterion	{Gini, cross-entropy}
number of trees	{250}
number of features considered at each split	{1, 2, 3} x √total number of features

Method Performance

For prediction performance using imaging features as inputs, we examined Macau, a regression approach based on Bayesian matrix factorization with side information (Online Methods), random forest classification (Breiman, 2001; Ganapathy et al., 2021), and deep neural networks (Mayr et al., 2016; Ganapathy et al., 2021a) (Online Methods). We discovered that the techniques were comparable in terms of properly predicting which assays could be performed. In particular, all three techniques predicted 11 assays with AUC greater than 0.9 (Venn diagram

in Figure 5). Similarly, all three methodologies showed that 182 assays had an AUC greater than 0.7. (Venn diagram in Figure 6). The number of tests in which only one or two techniques provided an AUC greater than 0.7 is comparable to the number of assays in which only one or two methods provided an AUC greater than 0.7. As a result, we conclude that image features, rather than the machine learning method, drive performance.
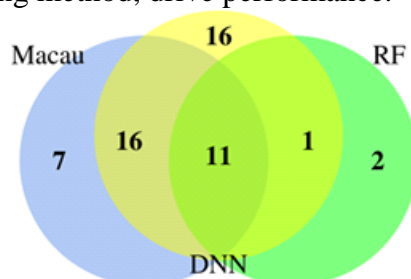

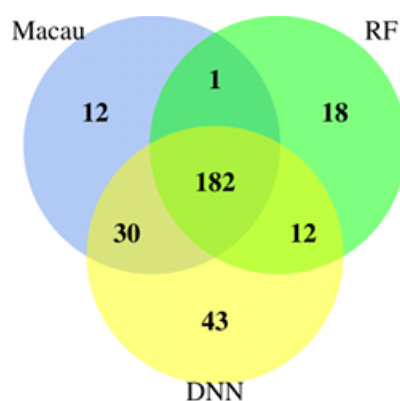Figure 5: Venn diagram: number of protein targets with high predictivity (AUC > 0:9).

Figure 5: Vienn diagram: number of protein targets with predictivity AUC >0:7.

4. METHODS

Data Set
There are 524,371 samples in the data set, 842 features, and 1,200 prediction tasks. Chemical compounds were supplied to cells, and the samples correspond to them. The features come from a combination of a biological imaging technique and the Cell Profiler program, which calculates morphological aspects of observed cells (Manavalan, 2021). This means that a chemical molecule is defined by the morphological changes it causes in cells. Compound activities were assessed in numerous separate biological studies, referred to as assays, for the prediction tasks. Typically, a single compound is detected in one or a few assays, resulting in a matrix with several unknowns that must be predicted. For the 1,200 prediction jobs, there are about 10 million activity values, resulting in a fill rate of around 1.6 percent. To enable nested cluster cross-validation, the samples were spread across three cross-validation folds (Mayr et al., 2016). (See Section "Cross-validation and Predictivity Measurement")

Experimental Approaches
We used a high-throughput imaging screen of 524,371 chemicals that was originally used to identify glucocorticoid receptor (GCR) nuclear translocation in our research. Each chemical was given to H4 brain neuroglioma cells at a concentration of 10M, incubated for one hour, and then 1M hydrocortisone was added to promote GCR translocation. Cells were frozen and photographed in 3-channel fluorescence after an additional hour of incubation, using a nuclear

stain (Hoechst), Cell Mask Deep Red (Invitrogen) to outline cell borders, and indirect immunofluorescence detection of GCR.

The photos were post processed with Cell Profiler software to repurpose the screen. We extracted unbiased maximum useful features from the photos using a process similar to Gustafsdottir and colleagues (2013). The pipeline created an image-based fingerprint with 841 features for each cell in the image. Each compound was then represented in our machine learning models by a vector of feature medians across all cells in the image (Vadlamudi et al., 2021; Manavalan, 2019b).

5. RESULTS AND DISCUSSION

Then, using Bayesian matrix factorization and the image-based fingerprint as side information, we developed a machine-learning model. Using more than ten million activity measurements, this model was tested for predictability across assays that map to more than 600 pharmacological target. A layered cluster cross-validation was used to evaluate the model's performance (Mayr et al., 2016). For 37.3 percent of the assays (AUC > 0.7 on 225 assays, Online Methods), the model was very predictive, and for 5.6 percent of the assays (AUC > 0.9 on 34 assays, Online Methods), the model was highly predictive.

The target for the oncology study was a kinase that had no known direct relationship to the glucocorticoid receptor (Ahmed & Ganapathy, 2021). We rated over 60,000 chemicals tested in the GR assay but for which no activity measurement was provided in the cancer screen using our matrix factorization algorithm. The 342 chemicals scored best by our matrix factorization method, about the number of noncontrol wells on a plate, were subsequently empirically confirmed. This yielded 141 sub-micromolar hits (41% hit rate), indicating a 60-fold enrichment over the first HTS results (0.725 percent hit rate).

To assess the chemical diversity of the hits, we calculated the Tanimoto similarity of each hit to the nearest hit discovered by the original HTS (based on extended-connectivity fingerprints (ECFP) (Rogers and Hahn, 2010). When compared to the initial finds (red distribution in Fig. 6a), the distribution of these similarities indicates that the image-based hits have significantly improved chemical structural diversity (shift to the left) (green).
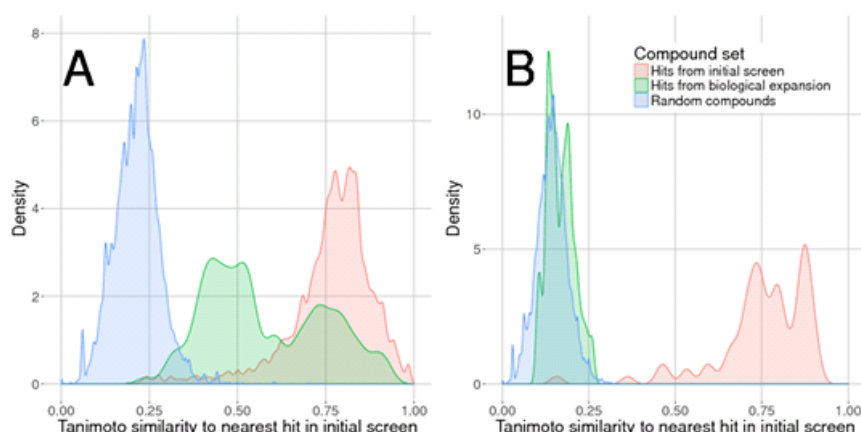

Figure 6: For each compound in a set, the ECFP (radius 4) based Tanimoto similarity to the nearest hit from the initial screen was considered

The image also shows the distribution of substances chosen at random as a point of comparison (blue). As a result, the HTI matrix factorization model identified candidate compounds with a

high hit rate, and the hit space was diversified. The target for the CNS project was a non-kinase enzyme with no evident relationship to the glucocorticoid receptor.

The activity of all 500,000 image-annotated chemicals was predicted using our matrix factorization model. Unfavorable properties (Baell and Holloway, 2010), such as autofluorescence and low predicted central nervous system availability, were filtered out of compounds with predicted submicromolar activity, and the remaining compounds were grouped into chemical clusters from which we randomly selected a handful of representatives from each cluster.

The 140 compounds that resulted from this technique were evaluated in the lab, and submicromolar activity was confirmed in 37 of them, resulting in a 22.7 percent hit rate, or a 250-fold enrichment over the initial HTS hit rate (0.088 percent). The 37 hits mapped to 32 Murcko scaffolds (Bynagari & Amin, 2019) that were not present in the initial list of hits, which is significant. That conclusion is supported by the distribution of Tanimoto similarities to the closest hit in the initial screen (Figure 6b). These findings suggested that using an image-based compound selection technique can improve hit rate and diversity.

We used three different machine learning approaches to see if the success of our approach is due to the machine learning method or the description of chemical compounds by image features (Ahmed, 2021b). We employed Macau (Simm, 2015), a Bayesian matrix factorization with side information regression approach (Online Methods), random forest classification (Breiman, 2001), and deep neural networks (Mayr et al., 2016; Khan et al., 2021). In terms of the tests that could be reliably predicted, these machine learning models performed similarly (Manavalan, 2018). Our cross-validation setup further ensures that the predicted performance is not based on the same compound's activity data across many targets. As a result, we conclude that the description of compounds using imaging features is critical to the success of our method.

The method is a supervised machine learning method, which means that output labels (in this case, activity measures) are required to train the model. This necessitates the acquisition of activity measurements for a sufficiently large library of chemicals (Manavalan, 2020).

Our findings suggest that photos from HTI screening initiatives undertaken in a variety of institutions can be repurposed to boost hit rates in other projects, even those that appear unrelated to the HTI screen's initial aim.

As a result, it may be able to replace certain assays with imaging technologies and machine learning models, which could be more cost-effective (Ahmed, 2021a). Imaging screens provide a view of multiple cellular activities by accessing rich morphological aspects of the cell, resulting in a fingerprint of biological action. This raises the intriguing question of how many targets imaging screens could reach if alternative cell lines, culture conditions, organelle labeling, and/or incubation durations are used.

6. CONCLUSION

The goal of this paper was to show how HTI data may be used to identify a variety of hits without having to retest the entire library in the target assay. Our models may also aid target deconvolution for phenotypic screens by prioritizing targets with anticipated activities that correspond to phenotypic findings. Raw photos could also be used directly in the activity prediction pipeline, thanks to recent developments in convolutional neural networks. This would enable the model to learn the optimal image attributes for the task at hand, thereby strengthening our strategy. Furthermore, our findings are based on a single HTI screen, and we believe that a combination of many HTI screens could be even more effective in predicting

assay activity. Finally, the median values across all cells in an image are our imaging characteristics. Models based on the distribution of feature values (e.g. quantiles) or even single cell analysis, on the other hand, may have greater predictive potential and will be studied in the near future.

7. REFERENCES

[1] Ahmed, A. A. A. (2021a). Perception of the audience of interests on the qualitative characteristics of financial reporting. International Journal of Intellectual Property Management. https://doi.org/10.1504/IJIPM.2021.10038267

[2] Ahmed, A. A. A. (2021b). Event Ticketing Accounting Information System using RFID within the COVID-19 Fitness Etiquettes. Academia Letters, Article 1379. https://doi.org/10.20935/AL1379

[3] Ahmed, A. A. A., & Ganapathy, A. (2021). Creation of Automated Content with Embedded Artificial Intelligence: A Study on Learning Management System for Educational Entrepreneurship. *Academy of Entrepreneurship Journal, 27*(3), 1-10, https://doi.org/10.5281/zenodo.4973057

[4] Ahmed, A. A. A.; Paruchuri, H.; Vadlamudi, S.; & Ganapathy, A. (2021). Cryptography in Financial Markets: Potential Channels for Future Financial Stability. Academy of Accounting and Financial Studies Journal, 25(4), 1–9. https://doi.org/10.5281/zenodo.4774829

[5] 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. *Journal of medicinal chemistry, 53*(7), 2719–2740.

[6] Baumann, D. and Baumann, K. (2014). Reliable estimation of prediction errors for qsar models under model uncertainty using double cross-validation. *Journal of cheminformatics,* 6, 1.

[7] Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32

[8] Bynagari, N. B. (2018). On the ChEMBL Platform, a Large-scale Evaluation of Machine Learning Algorithms for Drug Target Prediction. *Asian Journal of Applied Science and Engineering*, *7*, 53–64. Retrieved from https://upright.pub/index.php/ajase/article/view/31

[9] Bynagari, N. B. (2019). GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *Asian Journal of Applied Science and Engineering*, *8*, 25–34. Retrieved from https://upright.pub/index.php/ajase/article/view/32

[10] Bynagari, N. B. (2020). The Difficulty of Learning Long-Term Dependencies with Gradient Flow in Recurrent Nets. *Engineering International*, *8*(2), 127-138. https://doi.org/10.18034/ei.v8i2.570

[11] Bynagari, N. B. (2021). Cost Optimization at Early Stages of Design Using Deep Reinforcement Learning (DRL). *Design Engineering, 2021*(6), 5238-5247. Retrieved from http://www.thedesignengineering.com/index.php/DE/article/view/3922

[12] Bynagari, N. B., & Amin, R. (2019). Information Acquisition Driven by Reinforcement in Non-Deterministic Environments. *American Journal of Trade and Policy*, *6*(3), 107-112. https://doi.org/10.18034/ajtp.v6i3.569

[13] Bynagari, N. B., & Fadziso, T. (2018). Theoretical Approaches of Machine Learning to Schizophrenia. *Engineering International*, *6*(2), 155-168. https://doi.org/10.18034/ei.v6i2.568

[14] Bynagari, N. B., and Asadullah, A. B. M. (2021). Artificial Intelligence for Image Processing and Analysis. *Design Engineering*, *2021*(7), 11023 - 11036. Retrieved from http://thedesignengineering.com/index.php/DE/article/view/4099

[15] Caicedo, J. C., Singh, S., Carpenter, A. E. (2016). Applications in image-based profiling of perturbations. *Current opinion in biotechnology, 39*, 134–142.

[16] Caruana, R. (1997). Multitask Learning. *Machine Learning*, 28, 41–75.

[17] Ciresan, D. C., Meier, U., Gambardella, L. M. and Schmidhuber, J. (2012). Deep Big Multilayer Perceptrons for Digit Recognition. In Montavon, G., Orr, G. B. & M¨uller, K.-R. (eds.). *Neural Networks: Tricks of the Trade*, 581–598

[18] Donepudi, P. K. (2015). Crossing Point of Artificial Intelligence in Cybersecurity. American Journal of Trade and Policy, 2(3), 121-128. https://doi.org/10.18034/ajtp.v2i3.493

[19] Donepudi, P. K. (2016). Influence of Cloud Computing in Business: Are They Robust?. Asian Journal of Applied Science and Engineering, 5(3), 193-196. Retrieved from https://journals.abc.us.org/index.php/ajase/article/view/1181

[20] Donepudi, P. K. (2017). Machine Learning and Artificial Intelligence in Banking. Engineering International, 5(2), 83- 86. https://doi.org/10.18034/ei.v5i2.490

[21] Donepudi, P. K. (2018). Application of Artificial Intelligence in Automation Industry. Asian Journal of Applied Science and Engineering, 7, 7–20. Retrieved from https://upright.pub/index.php/ajase/article/view/23

[22] Donepudi, P. K., and Bynagari, N. B. (2021). Prediction of Transfusion Based on Machine Learning. *International Journal of Aquatic Science, 12*(3), 2168-2180. http://www.journal-aquaticscience.com/article_136518.html

[23] Egan, W. J., Merz, K. M. and Baldwin, J. J. (2000). Prediction of drug absorption using multivariate statistics. *Journal of medicinal chemistry,* 43, 3867–3877.

[24] Ganapathy, A., Ahmed, A. A. A, Siddique, M. NEA., (2021). Easy URLs in the Content Management System with Crawlers for Added Security. Academy of Marketing Studies Journal, 25(4), 1-10. https://doi.org/10.5281/zenodo.5002945

[25] Ganapathy, A., Ahmed, A. A. A, Siddique, M. NEA., (2021a). Easy URLs in the Content Management System with Crawlers for Added Security. Academy of Marketing Studies Journal, 25(4), 1-10. https://doi.org/10.5281/zenodo.5002945

[26] Ganapathy, A., Vadlamudi, S., Ahmed, A. A. A., Hossain, M. S., Islam, M. A. (2021b). HTML Content and Cascading Tree Sheets: Overview of Improving Web Content Visualization. *Turkish Online Journal of Qualitative Inquiry, 12*(3), 2428-2438. https://www.tojqi.net/index.php/journal/article/view/1724

[27] Garg, P. and Verma, J. (2006). In silico prediction of blood brain barrier permeability: an artificial neural network model. *Journal of chemical information and modeling*, 46, 289–297.

[28] Gustafsdottir, S. M. (2013). Multiplex cytological profiling assay to measure diverse cellular states. *PloS one*, 8(12):e80999.

[29] Held, M. (2010). Cellcognition: time-resolved phenotype annotation in high-throughput live cell imaging. *Nature methods, 7*(9), 747–754.

[30] Hochreiter, S. and Obermayer, K. (2004). Gene selection for microarray data. *Kernel methods in computational biology*, 319.

[31] Khan, W., Ahmed, A. A. A., Vadlamudi, S., Paruchuri, H., Ganapathy, A. (2021). Machine Moderators in Content Management System Details: Essentials for IoT Entrepreneurs. Academy of Entrepreneurship Journal, 27(3), 1-11. https://doi.org/10.5281/zenodo.4972587

[32] Louppe, G. (2014). Understanding Random Forests: From Theory to Practice. arXiv preprint arXiv:1407.7502.

[33] Manavalan, M. (2018). Do Internals of Neural Networks Make Sense in the Context of Hydrology?. *Asian Journal of Applied Science and Engineering*, *7*, 75–84. Retrieved from https://upright.pub/index.php/ajase/article/view/41

[34] Manavalan, M. (2019a). P-SVM Gene Selection for Automated Microarray Categorization. *International Journal of Reciprocal Symmetry and Physical Sciences*, *6*, 1–7. Retrieved from https://upright.pub/index.php/ijrsps/article/view/43

[35] Manavalan, M. (2019b). Using Fuzzy Equivalence Relations to Model Position Specificity in Sequence Kernels. *Asian Journal of Applied Science and Engineering*, *8*, 51–64. Retrieved from https://upright.pub/index.php/ajase/article/view/42

[36] Manavalan, M. (2020). Intersection of Artificial Intelligence, Machine Learning, and Internet of Things – An Economic Overview. *Global Disclosure of Economics and Business*, *9*(2), 119-128. https://doi.org/10.18034/gdeb.v9i2.584

[37] Manavalan, M. (2021). Molecular Generators and Optimizers Failure Modes. *Malaysian Journal of Medical and Biological Research*, *8*(2), 53-62. https://doi.org/10.18034/mjmbr.v8i2.583

[38] Manojkumar, P., Suresh, M., Ahmed, A. A. A., Panchal, H., Rajan, C. C. A., Dheepanchakkravarthy, A., Geetha, A., Gunapriya, B., Mann, S., & Sadasivuni, K. K. (2021). A novel home automation distributed server management system using Internet of Things. International Journal of Ambient Energy, https://doi.org/10.1080/01430750.2021.1953590

[39] Mayr, A., Klambauer, G., Unterthiner, T., and Hochreiter, S. (2016). Deeptox: toxicity prediction using deep learning. *Frontiers in Environmental Science*, 3:80.

[40] Neogy, T. K., & Bynagari, N. B. (2018). Gradient Descent is a Technique for Learning to Learn. *Asian Journal of Humanity, Art and Literature*, *5*(2), 145-156. https://doi.org/10.18034/ajhal.v5i2.578

[41] Oshiro, T. M., Perez, P. S. and Baranauskas, J. A. (2012). How many trees in a random forest? *In International Workshop on Machine Learning and Data Mining in Pattern Recognition*, 154–168.

[42] Pepperkok, R and Ellenberg, J. (2006). High-throughput fluorescence microscopy for systems biology. *Nature Reviews Molecular Cell Biology*, 7(9), 690–696.

[43] Polishchuk, P. G. (2009). Application of random forest approach to qsar prediction of aquatic toxicity. *Journal of chemical information and modeling*, 49, 2481–2488.

[44] Rogers, D., and Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5), 742–754.

[45] Simm, J. (2015). Macau: scalable bayesian multi-relational factorization with side information using mcmc. arXiv preprint arXiv:1509.04610.

[46] Singh, S, Carpenter, A. E. and Genovesio, A. (2014). Increasing the content of high-content screening an overview. *Journal of biomolecular screening,* *19*(5), 640–650.

[47] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research,* 15, 1929–1958.

[48] Starkuviene, V. and Pepperkok, R. (2007). The potential of high-content high-throughput microscopy in drug discovery. *British journal of pharmacology*, *152*(1), 62–71.

[49] Vadlamudi, S., Islam, M. A., Hossain, M. S., Ahmed, A. A. A., Asadullah, ABM. (2021). Watermarking Techniques for Royalty Accounts in Content Management

Websites for IoT Image Association. Academy of Marketing Studies Journal, 25(4), 1-9. https://doi.org/10.5281/zenodo.5002840

[50] Walter, T. (2010). Visualization of image data from cells to organisms. *Nature methods*, 7, S26–S41.

[51] Wang, Y. (2015). In silico adme/t modelling for rational drug design. *Quarterly reviews of biophysics,* 48, 488–515.

[52] Wawer, M. J. (2014). Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling. *Proceedings of the National Academy of Sciences*, *111*(30), 10911–10916.

[53] Wright, M. N. and Ziegler, A. (2015). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. arXiv preprint arXiv:1508.04409.

[54] Yarrow, J., Feng, Y., Perlman, Z., Kirchhausen, T., and Mitchison, T. (2003). Phenotypic screening of small molecule libraries by high throughput cell imaging. *Combinatorial chemistry & high throughput screening, 6*(4), 279–286.

[55] --0--