Artificial Intelligence in the Life Sciences 3 (2023) 100060

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

journal homepage: www.elsevier.com/locate/ailsci

Research Article

Combining molecular and cell painting image data for mechanism of action prediction

Guangyan Tian[a,1], Philip J Harrison[a,1,*], Akshai P Sreenivasan[a,b], Jordi Carreras-Puigvert[a], Ola Spjuth[a]

[a] Department of Pharmaceutical Biosciences, Uppsala University, Sweden
[b] Department of Medical Sciences, Uppsala University, Sweden

ARTICLE INFO

Keywords:
Bioinformatics
Convolutional neural networks
Cheminformatics
Deep learning
Machine learning

ABSTRACT

The mechanism of action (MoA) of a compound describes the biological interaction through which it produces a pharmacological effect. Multiple data sources can be used for the purpose of predicting MoA, including compound structural information, and various assays, such as those based on cell morphology, transcriptomics and metabolomics. In the present study we explored the benefits and potential additive/synergistic effects of combining structural information, in the form of Morgan fingerprints, and morphological information, in the form of five-channel Cell Painting image data. For a set of 10 well represented MoA classes, we compared the performance of deep learning models trained on the two datasets separately versus a model trained on both datasets simultaneously. On a held-out test set we obtained a macro-averaged F1 score of 0.58 when training on only the structural data, 0.81 when training on only the image data, and 0.92 when training on both together. Thus indicating clear additive/synergistic effects and highlighting the benefit of integrating multiple data sources for MoA prediction.

Introduction

Mechanism of action (MoA) refers to the biological interaction through which a potentially therapeutic small-molecule compound produces a pharmacological effect, such as the specific proteins that the compound targets and the pathways that it modulates. Uncovering the MoA of a compound, although a significant challenge in chemical biology [1], provides extremely useful information for lead compounds prior to clinical trials and for identifying possible toxicity or side-effects [2].

A variety of different data sources can be used to capture information on a compounds MoA, including structural information from the compound, gene expression from transcriptomics data, protein information from proteomics data, and metabolic enzyme activity from metabolomics data [2]. Recently, cell morphology data from high-content imaging has proven useful for this task [3]. A significant benefit of microscopy based image assays is that they can be scaled to high-throughput much more easily and less expensively than transcriptomics and metabolomics based assays [4]. Cell imaging also provides information at the single-cell resolution as opposed to condensing the output down to measures of population averages [5]. In terms of throughput and efficiency the L1000 [6] gene expression assay is perhaps currently the only feasible alternative to image-based assays [7] for large scale data generation to sustain predictive modeling.

Microscopy imaging can be used to capture the changes in cell morphology that arise when a cell culture is treated with a chemical compound [2]. The Cell Painting assay uses fluorescent dyes to paint the cells in multi-well plates as "richly as possible" to illuminate morphological changes in eight broadly relevant organelles and cellular sub-compartments (nuclei, mitochondria, cytoskeleton, Golgi apparatus, plasma membrane, cytoplasmic RNA, nucleoli and endoplasmic reticulum) using six fluorescent dyes imaged in five channels [7].

A comparative study for library enrichment reported better predictive power for High-throughput screening performance using Cell Painting as opposed to L1000 gene expression profiling [8]. Whereas, for predicting MoA, Way et al. [9] found that L1000 outperformed Cell Painting, but that there was complementarity, i.e. some MoAs were better predicted by one of the assays compared to the other. A related study by Lapins and Spjuth [10] compared Cell Painting, L1000 and chemical structure based predictors, and found MoA classes that were predicted better by each of the three predictors relative to the other two, supporting the idea of a likely benefit through combining these different data

* Corresponding author.
E-mail addresses: philip.harrison@farmbio.uu.se (P.J. Harrison), ola.spjuth@farmbio.uu.se (O. Spjuth).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.ailsci.2023.100060
Received 8 December 2022; Received in revised form 5 January 2023; Accepted 30 January 2023
Available online 17 February 2023
2667-3185/© 2023 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

sources. Another study predicting MoA [11], based on data from the ExCAPE database, which compared models built using image based features to those built using chemical structure descriptors, provided further support for the complementarity of these two data types, whereby the models performed somewhat differently at an individual class level.

Besides comparing models built using different types of data it is also possible to combine the datasets and analyze them simultaneously to search for additive or synergistic effects. For predicting cytotoxicity and proliferation Seal et al. [12] compared Random Forest models using Cell Painting image based features, molecular fingerprints, and combining both data sources. They found that the models based on image features outperformed those based on molecular fingerprints, but the combined models performed best in ten out of twelve cases. Another study predicting the bio-activity of approximately 16,000 compounds [13], found that models based on features derived from Cell Painting images outperformed those based on chemical structure profiles from graph convolutional networks (GCNs, [14]), but that the fusion of the two datasets gave a gain in performance.

Most traditional image analysis pipelines, including those mentioned above, first extract morphological features from the fluorescence stained images, including measures of size, shape, intensity and texture from the labeled cellular compartments, most often using the CellProfiler [15] software package, and subsequently apply machine learning methods to the extracted features for the predictive task at hand [16]. These methods require an accurate segmentation algorithm to identify the cellular compartments prior to feature extraction. However, when convolutional neural networks (CNNs) are used on the raw images, features are extracted in an automatic data-driven fashion, circumventing the need for cell segmentation and potentially providing better predictive performance [3,17]. For instance, Hofmarcher et al. [18] found that CNNs trained on Cell Painting image data, for predicting activity labels for over 10,000 compounds, performed significantly better than fully connected neural networks trained on pre-computed image features. The flexibility of the architectural choices for neural networks also provides a simple means of combining multiple data sources into the same modeling framework [19].

In the work presented in this manuscript we first compared a variety of traditional machine learning and deep learning models for the prediction of MoA based on chemical structure data for up to 20 MoA classes. Subsequently, based on a set of 10 MoA classes, we compared the performance of the best deep learning model at the compound structural level to a state-of-the-art CNN trained on Cell Painting image data for the same set of compounds. We selected the best deep learning based compound structure model so that we could finally train a joint model for the MoA prediction based on utilizing both structural and image data as input. Example Cell Painting images for the 10 MoA classes can be seen in Fig. 1. To the best of our knowledge our work represents the first combination of five channel Cell Painting image data and molecular fingerprint data trained in an end-to-end fashion to predict MoA, wherein the raw images, as opposed to features derived from the images, were used as input to the models.

Materials and methods

Data

Molecular data

Molecular data (Corsello et al. [20]), in the form of SMILES strings collected and processed by the Broad Institute, was used in this study. The cleansed dataset contains approximately 5500 compounds covering 1300 MoA classes, but most MoAs have very few compounds associated with them. The number of compounds that each MoA has is shown in Fig. 2. As our models should perform well at the compound level, namely to predict the MoA for unseen compounds, we used a subset of the data, the top 20 MoAs (i.e. the 20 MoAs having the most compounds associated with them).

Image data

The 5-channel Cell Painting image data was produced by the Pharmaceutical Bioinformatics Research Group at Uppsala University. We selected image data from 10 well-represented MoAs (MoAs that we presumed would be reasonably distinguishable and that had a sufficient number of compounds associated with them). The ten MoAs were ATPase inhibitors (ATPase-i, number of compounds, $n = 18$); Aurora kinase inhibitors (AuroraK-i, $n = 20$); HDAC inhibitors (HDAC-i, $n = 33$); HSP inhibitors (HSP-i, $n = 24$); JAK inhibitors (JAK-i, $n = 21$); PARP inhibitors (PARP-i, $n = 21$); protein synthesis inhibitors (Prot.Synth.-i, $n = 23$); retinoid receptor agonists (Ret.Rec.Ag, $n = 19$); topoisomerase inhibitors (Topo.-i, $n = 32$); and tubulin polymerization inhibitors (Tub.Pol.-i, $n = 20$). In total we had 12,582 images for 231 compounds. The compounds were administered to U2OS cells in 384 well plates at a dose of 10 micro-molar. Images at a resolution of 2160 x 2160 pixels were taken across 9 sites in each well and each compound was replicated 6 times. The compounds were distributed across 18 plates using PLAID (Plate Layouts using Artificial Intelligence Design, [21]).

Data preprocessing for models based on molecular data

We explored usin multi-layer perceptrons (MLPs), graph convolutional networks (GCNs), convolutional neural networks (CNNs), long short-term memory networks (LSTMs, with and without data augmentation), and traditional machine learning algorithms (those operating on tabular data) to predict MoA, and we pre-processed the data for each model. See the Modeling section below for further details on the models explored.

For the MLP and traditional machine learning algorithms, we used Morgan Fingerprints as input. Since SMILES strings are sequential they cannot be processed directly by these models. We used the RDKit package [22] to generate the Morgan fingerprints (binary vectors, 2048 bits) [23]. For the GCN, which requires the adjacency matrix and the node matrix as input, we applied Spektral [24] and NetworkX packages [25] to convert the SMILES strings into graphs. For the CNN we generated the feature matrix for each SMILES string based on the approach of Hirohara et al. [26]. Initially, we selected 42 chemical features to prepare the feature matrix of each SMILES string based on the selected chemical features. Secondly, due to the blank parts of the feature matrix resulting from inconsistent lengths of the SMILES, we applied zero padding to maintain uniform dimensions of the feature matrix. For the recurrent neural network, the LSTM, we utilized SMILES pair encoding [27] to tokenize the chemical structure data, so we obtained a series of numbers (tokens) representing the SMILES. Similarly to the CNN case, we also used zero padding to ensure that the lengths of all tokens were identical.

Data preprocessing for models based on image data

The 5 channels in the Cell Painting image data were standardized to remove plate-level effects based on the mean and standard deviation of the pixel intensities in the control DMSO wells in each plate. The images were resized from their original dimension down to 256 x 256 pixels. A quality control run on the data to detect saturation and blur in the images found no saturation issues (such as fibers across the field of view) but did detect some blurred images. However, given that a common data augmentation strategy for deep learning models is to purposefully blur the images we decided not to remove these blurred images from the dataset.

Data augmentation and data splitting

Data augmentation generates additional data based on the existing data and improves the generalizability of models. For the image based models we used flipping, 90 degree rotations and shift scale rotations to augment the data. However, for the compound structure based models data augmentation was only possible for the LSTM. As the LSTM is a sequence-based model that requires tokens as input, data augmentation

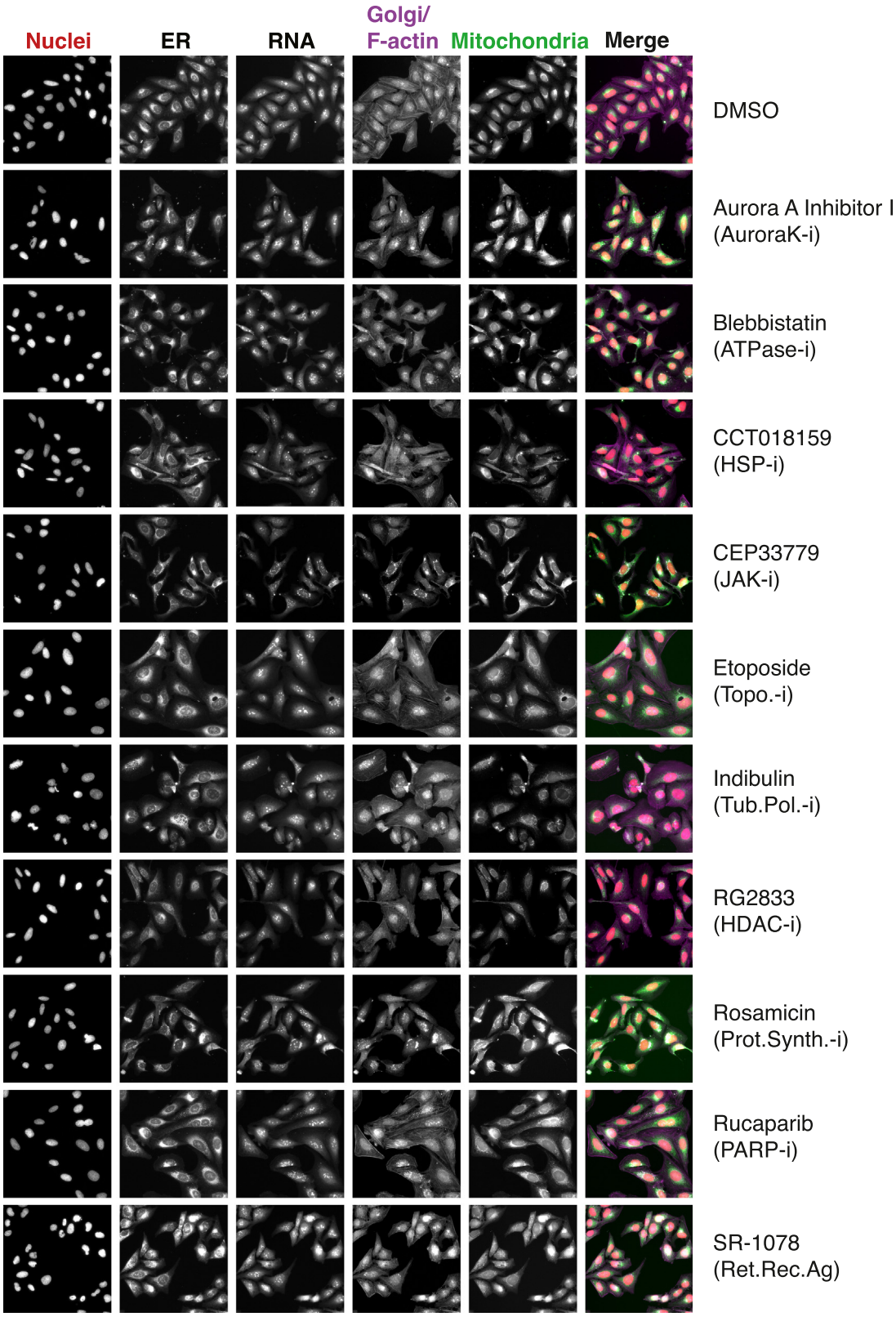

Fig. 1. Example Cell Painting images for the 10 MoA classes and the DMSO data used for standardization. The row titles give the compound names for the selected images with the MoA abbreviation in parenthesis, where i stands for inhibitor and Ag for agonist.

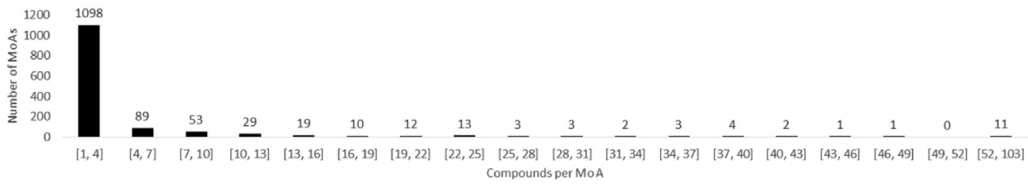

Fig. 2. Histogram representing the compound counts per MoA for different binning intervals. Note that the range of the interval for the final bin is larger than the others.

is feasible as slightly different tokens can be produced by randomized SMILES [28].

For splitting the data at the compound-level into training, validation and test sets we used stratification based on the proportion of compounds for each MoA. We split out 10% of the data for the final held-out test set. Stratified splitting for the remaining data was performed nine times (nine shuffles) for the SMILES data in the initial comparison and five times (five shuffles) for the image data and the corresponding SMILES subset. In each case 80% of the data was used for training and 10% for validation.

Modeling

Compound structure based models

We explored the following deep learning models for the prediction of MoA using chemical structure data: MLP, GCN, CNN, and LSTM with and without data augmentation. For the deep learning models we determined the optimal architectures and parameters through model exploration and parameter tuning on the validation sets. The MLP is a basic artificial neural network [29] that includes fully connected input, hidden, and output layers. Our MLP model contained one input layer, one hidden layer with dropout ($p = 0.85$), and one final prediction layer. GCNs are a subset of GNNs [30] that can process non-Euclidean data, such as graphs with nodes and edges [14]. Our GCN model included input layers for the adjacency matrix and the node matrix followed by three convolution layers with dropout ($p = 0.5$), one global attention pooling layer and one final prediction layer. Our CNN model contained one convolution layer, one max pooling layer with dropout ($p = 0.8$), one flattening layer with dropout ($p = 0.8$), and one final prediction layer. Our LSTM model included an embedding layer, a bidirectional LSTM layer, a dropout layer ($p = 0.96$) and a final prediction layer. For the LSTM with data augmentation, we adjusted the degree of augmentation to ensure that each MoA had approximately 1000 SMILES in the augmented training set.

We used the Adam optimizer [31], sparse categorical cross-entropy as the loss function, and validation loss as the metric for early stopping. To accommodate for imbalance of classes we applied class weighting in the loss functions to train the models.

We also explored machine learning algorithms that operate on tabular data (in contrast to the deep neural networks described above). The more traditional machine learning models have shown competitive performance with deep learning models when dataset sizes are relatively small [32]. For instance, Jiang et al. [33] showed that four descriptor-based models outperformed four graph-based models on several benchmark datasets. We examined five individual machine learning algorithms and four ensemble algorithms. The individual algorithms included random forests [34], light gradient boosting machines [35], cat boost [36], k-nearest neighbors classifiers [37], and logistic regression [38]. The ensemble algorithms included bagging [39], stacking [40], voting [41], and adaboost [42].

Cell morphology based model

We applied the state-of-the-art CNN model EfficientNet [43] to predict MoA based on the 5-channel Cell Painting image data. EfficientNet applies a compound scaling method to adjust width, depth, and resolution simultaneously, achieving competitive performance in image-based tasks with less training time and fewer parameters. We adopted the EfficientNetB1 architecture and used the AdamW optimizer [44] with weighted sparse categorical cross entropy as the loss function.

Table 1
F1 scores on the test set for the main three models explored for predicting the 10 selected MoAs. MLP used the chemical structure data, EfficientNet used the image data, and the Global model (see Fig. 3) used both data formats. The results are based on the averages across the five shuffles of the training and validation data.

	MLP	EfficientNet	Global model
ATPase-i	0.64	0.58	0.68
AuroraK-i	0.55	0.48	0.71
HDAC-i	0.98	0.89	0.99
HSP-i	0.45	0.83	0.95
JAK-i	0.08	0.85	0.94
PARP-i	0.51	0.96	0.98
Prot.Synth.-i	0.54	0.98	0.99
Ret.Rec.Ag	1.00	0.98	1.00
Topo.-i	0.68	0.64	0.97
Tub.Pol.-i	0.37	0.92	0.97
Accuracy	0.62	0.81	0.93
Macro average F1	0.58	0.81	0.92
Weighted average F1	0.61	0.81	0.93

Global model

For our global model trained on the data for the 10 selected MoA classes, we integrated the MLP (our best performing deep learning model based on the compound structure data) and EfficientNet (for the image data). The models were first trained separately and then combined and their weights finetuned. The architecture of our global model is shown in Fig. 3.

Results

Summaries of the performance of the traditional and deep learning models for the compound structure based models for the 20 MoA subsets are shown in Fig. 4. These figures present the average F1 scores across the nine shuffles of the training and validation data as well as the results of randomization tests performed to assess the level of significance in the performance differences. We applied Bonferroni corrections [45] to the p-values to account for the fact that we were performing several comparisons. The performances of the traditional machine learning algorithms were all quite comparable, however there were larger differences for the deep learning models compared. The best performing deep learning model was the MLP and the worst was the CNN; we note that the MLP performed on par with the best traditional machine learning models.

Test set F1 scores for the 10 selected MoAs (averaged across five shuffles of the training and validation data) comparing the MLP, trained on the compound structure data, EfficientNet, trained on the Cell Painting image data, and the global model, trained on both data sources are shown in Table 1. Our test set contained 24 compounds. This test set was the same for each of the shuffles of the training and validation data. For the MLP the F1 scores were very variable across the MoA classes, ranging from 0.08 for the JAK inhibitor test compounds to 1.00 for the Retinoid

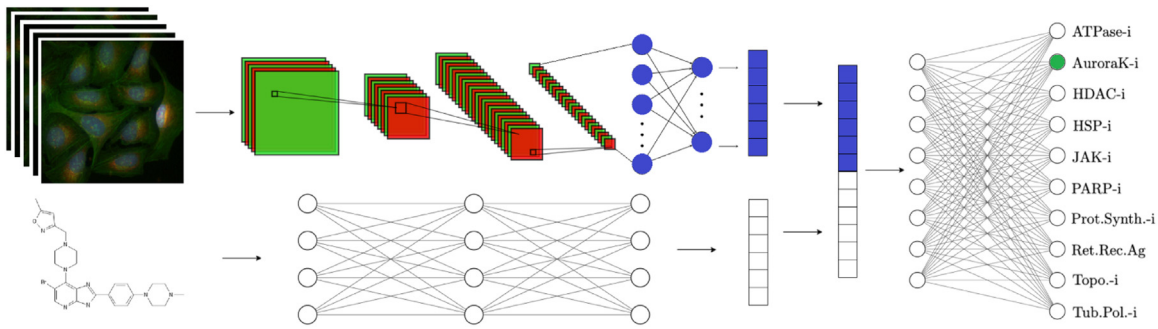

Fig. 3. The architecture of the global model with two input paths, one for the Cell Painting image data and one for the chemical structure data.

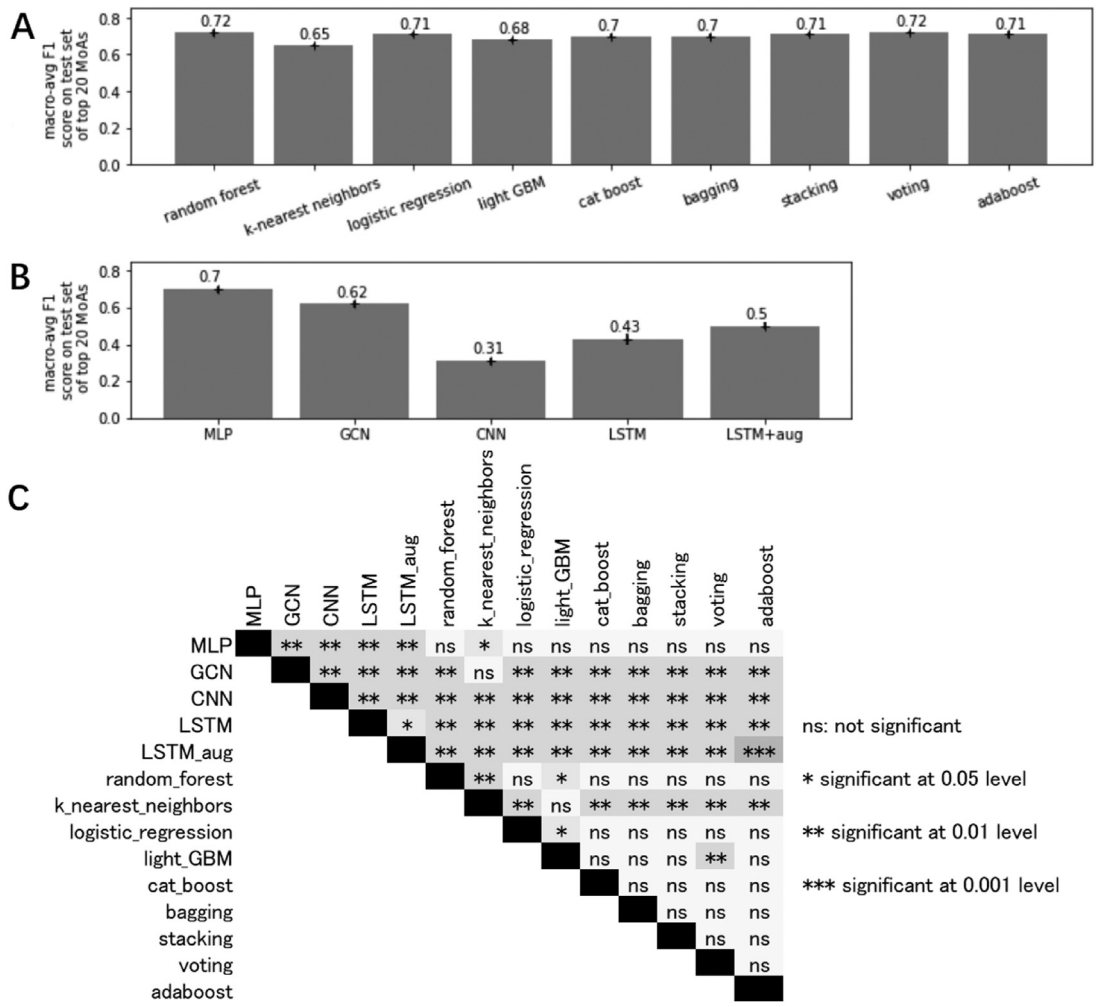

Fig. 4. A). Comparison of macro-averaged F1 scores on the test set of the traditional machine learning models for the top 20 MoAs (i.e. the MoAs that were best represented in the data in terms of the number of compounds they had). B). Comparison of macro-averaged F1 scores on the test set of the deep learning models for the top 20 MoAs. C). Randomization test with Bonferroni correction of macro-averaged F1 scores on the test set of top 20 MoAs. The results are based on the averages across the nine shuffles of the training and validation data.

receptor agonist compounds. For EfficientNet the results were somewhat more stable, ranging from 0.48 for the Aurora kinase inhibitors to 0.98 for both the Protein synthesis inhibitors and the Retinoid receptor agonists. For the global model the results were even more stable, ranging from 0.68 for the ATPase inhibitors to 1.00 for the Retinoid receptor agonists. Our global model, achieving a macro-averaged F1 score of 0.92, revealed a clear additive/synergistic effect with an increase in F1 score of 0.11. The three different models were all significantly different from one another at the 5% significance-level based on randomization tests with Bonferroni corrected p-values. The predictive performance of the models for each of the test compounds is summarized in Table S1 (Supplementary materials). We also show this comparison graphically in Fig. 5 in which we have highlighted the compounds NKP-1339 and amonafide, showing a very pronounced synergistic effect.

In light of the high variability in the test performances for the MLP fitting to the compound structure data we further explored the structural properties of the compounds in our dataset using DataWarrior [46]. The compounds' SMILES strings were used as input to DataWarrior and for

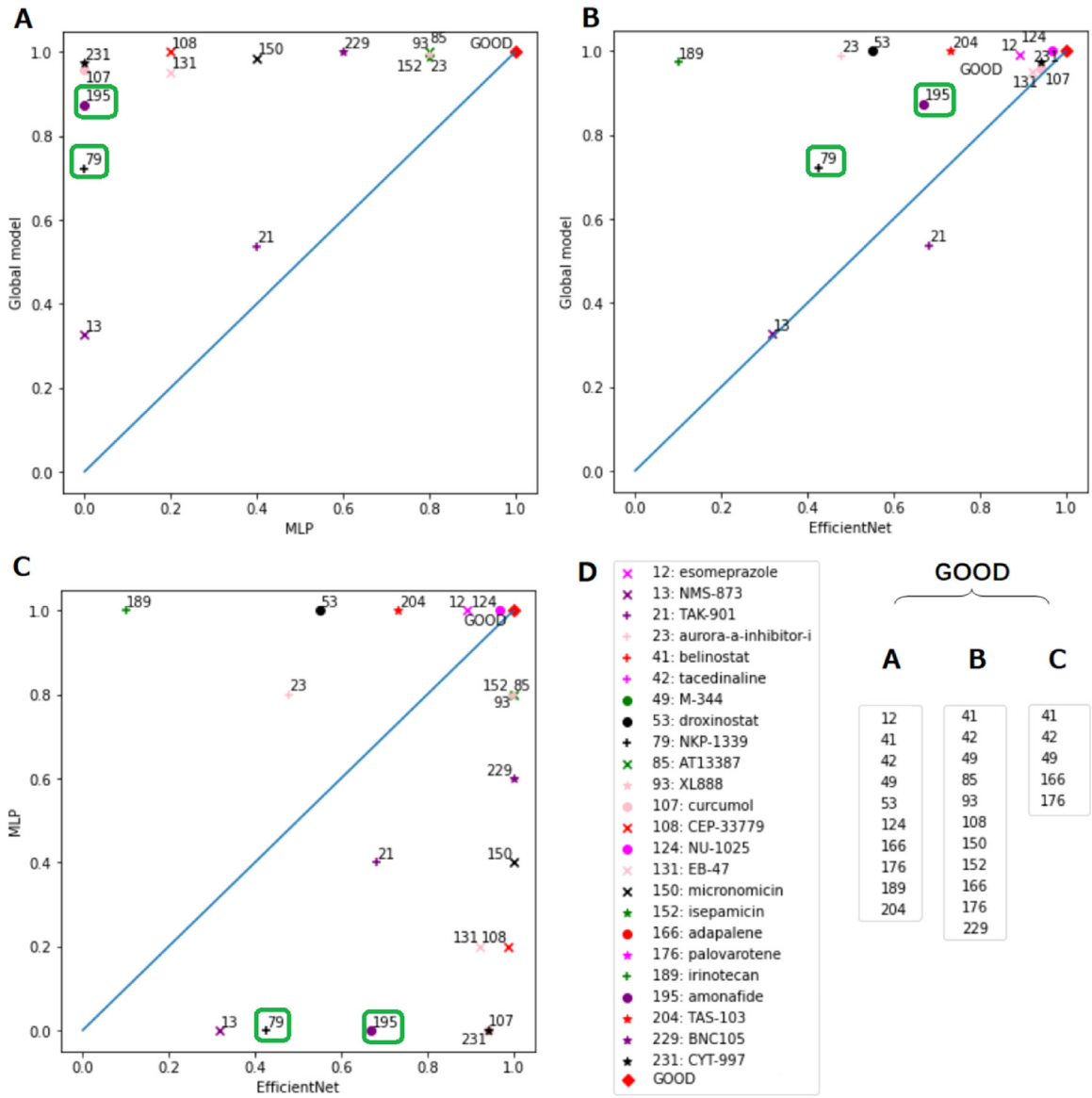

Fig. 5. Comparison of the prediction rates for the three models for each compound: A). MLP, trained solely on the chemical structure data, versus the Global model, trained on both the chemical structure and Cell Painting image data; B). EfficientNet, trained solely on the image data, versus the Global model; C). EfficientNet versus MLP; D). 'GOOD' cluster contained compounds that possessed a prediction rate above 0.97 in panels A-C. The compounds NKP-1339 and amonafide have been highlighted in green boxes in panels A-C as they showed a greater synergistic effect than the other compounds. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the analyses we performed we used the default parameter settings. In Figure S1A (Supplementary materials) we show a neighborhood analysis where compounds with at least one structurally similar neighbor have connecting lines. These structurally connected compounds are isolated in Figure S1B (Supplementary materials) and their compound ID numbers are shown. The names for these compounds and their SMILES strings are available on our GitHub repository (see Data and code availability). Although the groupings tend to show compounds belonging to the same MoA class, there are several unconnected compounds. In Figure S2 (Supplementary materials) we show the distributions for the compounds, grouped by their MoA classes, for a variety of physico-chemical properties predicted by DataWarrior. We also highlight in this figure the six test set compounds that were poorly predicted by our MLP model (see Table S1 in the Supplementary materials). From this we can see that in some cases these poorly predicted compounds have physico-chemical properties in the tails of the distributions for their MoA classes. This is most evident for the property cLogS, a measure of aqueous solubility. The two JAK inhibitor compounds in our test set (curcumol and CEP-33779), both of which were predicted poorly, are each at opposite extremes of the distribution for cLogS. The compounds NMS-873, NKP-1339 and CYT-997 were also the least soluble for their MoA classes.

Discussion

We have introduced a novel and efficient approach for MoA prediction, which combines both Cell Painting image data and chemical structure data (in the form of Morgan fingerprints). Similarly to the study predicting cytotoxicity and proliferation [12] mentioned earlier we found that image based models outperformed those based on Morgan fingerprints and that integrating both data sources further boosted the performance. However, Lapins and Spjuth [10] found that chemical structure based models for MoA prediction were generally better than either L1000 or Cell Painting based models (note however that they used features derived from the images as opposed to the raw images them-

selves as model input). Chemical structure data (such as that obtained from Morgan fingerprints) can suffer from "Activity Cliffs", whereby a small change in structure can result in a large difference in bio-activity, highlighting the need to supplement chemical structure data with additional sources of information [2], such as the Cell Painting image data used in the current study.

It should be noted that the Cell Painting assay was performed on a single cancer cell line (U2OS). Although performing MoA and target identification studies using cancer cell lines as proxy is well-accepted [47,48], there is the possibility that using a different cancer (or non-cancer) cell line, could yield slightly different results. Nevertheless, given that the experiment was executed in a controlled manner, and that all cells were treated equally, we are confident in the robustness of our results.

Similarly, the drug screen in this study was performed at 10uM for all compounds. This is a concentration typically used in drug screens to ensure a cellular response. However, given that the compounds in this work could have different potency, a dose response would have perhaps aided a more accurate identification of the compounds' activity.

Due to data limitations our test set had only two to four compounds per MoA. The test-level predictive performance under such conditions may suffer if by chance any of these compounds happened to be outliers for their class. Based on our compound level analysis using DataWarrior it appears that this may have been the case for our two JAK inhibitor test compounds, thus potentially explaining the low F1 score for the MLP for this MoA.

However, a few compounds were better predicted by the chemical structure based model, relative to the image based model. Perhaps the main disadvantage of image data such as Cell Painting is that not all compounds will necessarily produce a morphological change or the morphological effects may be very subtly and potentially masked by unaccounted for technical variations within and between plates during image capturing [2]. However, in the current study, to reduce potential bias caused by positional effects in the micro-well plates, the compounds and controls were distributed over the plates using PLAID and we standardized the images across the plates based on the control/DMSO wells. It is also possible that the compound does produce a morphological change but not in any of the cellular compartments or organelles captured using the Cell Painting assay. Another possibility is that the dose applied was not sufficient to produce a morphological change.

Concerning the traditional machine learning methods we explored for the chemical structure data, the ensemble methods outperformed the individual methods. Similarly, when combining a model based on multiple inputs, with separate modeling paths that come together to make a final prediction, we can potentially achieve better results than the models built on just one of the data categories. In this study, we showed this type of additive/synergistic effect by combining MLP for chemical structure data and EfficientNet for image data for MoA prediction.

Although it was somewhat surprising that for our models based on only the chemical descriptors, the simplest deep learning architecture, the MLP, outperformed the more complex networks architectures explored, a similar result has been obtained in a previous study [49] performing drug target prediction on a large benchmark dataset from the ChEMBL database. In our MLP architecture, we used an unconventionally high dropout rate to alleviate the overfitting problem as a result of the scarcity of chemical structure data. We also tested other possible architectures, such as reducing the dropout rate and increasing the number of hidden layers, with fewer neurons in each layer. However, these modifications did not improve the model performance.

It should also be mentioned that the purpose of the current study was to compare the accuracy of models trained on morphological and structural data, and that the domain of the models hence limit their applicability outside of this scope (i.e. for making predictions when a test compound's MoA does not belong to any of those on which the model was trained). This is due to the fact that the predictions of neural network models are neither probabilistic nor well-calibrated [50], and will always make a prediction of the most likely class even if all classes should be predicted with low probability. We are currently developing methods to fill these gaps using what we refer to as Predictive Target Profiles (PTP) using well-calibrated conformal predictors [51,52].

Predicting the MoA of a compound can benefit greatly from the interaction of multiple sources of data [2]. Various studies [9,10,13,53] have shown that image and transcriptomics assays contain both overlapping and distinct cell state information. Thus, an even better predictive model than our final one could potentially be achieved with the additional integration of transcriptomics data. Further, other types of chemical structure representations such as 3D descriptors could be explored. In the present study, for the image based model and the combined model, we used a set of ten well represented MoA classes and a widely established 2D descriptor for representing chemistry. In future work we will explore the predictive ability of these models across a wider range of classes whilst accounting for potential polypharmacological effects. Again the conformal prediction method mentioned above will likely prove useful for this purpose. Contrary to previous belief, polypharmacology, where a compound concurrently engages with multiple targets or processes, is the rule rather than the exception in biology [19].

Conclusions

In this work we explored the fusion of chemical structure and cell morphology data for the purpose of mechanism of action prediction. To the best of our knowledge our work represents the first combination of molecular fingerprint data and five-channel fluorescence Cell Painting image data, trained using deep learning in an end-to-end fashion, to predict mechanism of action. Furthermore, for improved model flexibility and performance, we used the raw input images, as opposed to featured derived from them, as input to the models. We found a clear and significant improvement in predictive performance for models trained on both input types simultaneously, as opposed to in isolation, with an increase in F1 score of 0.11, highlighting the benefit of combining data sources for mechanism of action prediction.

Data and code availability

The image data to accompany this paper has been uploaded to Figshare (https://doi.org/10.17044/scilifelab.21378906). The python code to accompany the paper is available on GitHub (https://github.com/pharmbio/CP-Chem-MoA). Also in the GitHub repository is the csv file *compound_list_10_MoAs* which gives the compound names, their SMILES strings and their reference ID numbers (1:231) used in our analyses.

Funding

This project was supported by the Swedish Research Council (grants 2020-03731 and 2020-01865), FORMAS (grant 2018-00924), the Swedish Foundation for Strategic Research (grant BD15-0008SB16-0046), and the Swedish strategic research program eSSENCE.

Data availability

Links to the data and code are provided at the end of our paper.

Acknowledgments

This project was supported by the Swedish Research Council (grants 2020-03731 and 2020-01865),FORMAS (grant 2018-00924), the Swedish Foundation for Strategic Research (grant BD15-0008SB16-0046), and the Swedish strategic research program eSSENCE. We also thank Jonne Rietdijk and Polina Georgiev for performing the Cell Painting assay and Anders Larsson for IT infrastructure assistance.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ailsci.2023.100060.

References

[1] Hight SK, Clark TN, Kurita KL, McMillan EA, Bray W, Shaikh AF, Haeckl FPJ, Carnevale-Neto F, La S, Lohith A, et al. High-throughput functional annotation of natural products by integrated activity profiling. BioRxiv 2022:748129. doi:10.1101/748129.

[2] Trapotsi M-A, Hosseini-Gerami L, Bender A. Computational analyses of mechanism of action (MoA): data, methods and integration. RSC Chem Biol 2022;3(2):170–200.

[3] Kensert A, Harrison PJ, Spjuth O. Transfer learning with deep convolutional neural networks for classifying cellular morphological changes. SLAS Discov Adv Life Sci R&D 2019;24(4):466–75.

[4] Caicedo JC, Singh S, Carpenter AE. Applications in image-based profiling of perturbations. Curr Opin Biotechnol 2016;39:134–42.

[5] Feng Y, Mitchison TJ, Bender A, Young DW, Tallarico JA. Multi-parameter phenotypic profiling: using cellular effects to characterize small-molecule compounds. Nat Rev Drug Discov 2009;8(7):567–78.

[6] Peck D, Crawford ED, Ross KN, Stegmaier K, Golub TR, Lamb J. A method for high--throughput gene expression signature analysis. Genome Biol 2006;7(7):1–6.

[7] Bray M-A, Singh S, Han H, Davis CT, Borgeson B, Hartland C, et al. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nat Protoc 2016;11(9):1757–74.

[8] Wawer MJ, Li K, Gustafsdottir SM, Ljosa V, Bodycombe NE, Marton MA, et al. Toward performance-diverse small-molecule libraries for cell-based phenotypic screening using multiplexed high-dimensional profiling. Proc Natl Acad Sci 2014;111(30):10911–16.

[9] Way GP, Natoli T, Adeboye A, Litichevskiy L, Yang AX, Lu X, Caicedo J, Cimini BA, Karhohs K, Logan DJ, et al. Morphology and gene expression profiling provide complementary information for mapping cell state. bioRxiv 2021. doi:10.1101/2021.10.21.465335.

[10] Lapins M, Spjuth O. Evaluation of gene expression and phenotypic profiling data as quantitative descriptors for predicting drug targets and mechanisms of action. Biorxiv 2019:580654. doi:10.1101/580654.

[11] Trapotsi M-A, Mervin LH, Afzal AM, Sturm N, Engkvist O, Barrett IP, et al. Comparison of chemical structure and cell morphology information for multitask bioactivity predictions. J Chem Inf Model 2021;61(3):1444–56.

[12] Seal S, Yang H, Vollmers L, Bender A. Comparison of cellular morphological descriptors and molecular fingerprints for the prediction of cytotoxicity-and proliferation-related assays. Chem Res Toxicol 2021;34(2):422–37.

[13] Moshkov N, Becker T, Yang K, Horvath P, Dancik V, Wagner BK, Clemons PA, Singh S, Carpenter AE, Caicedo JC. Predicting compound activity from phenotypic profiles and chemical structures. bioRxiv 2022:2020–112. doi:10.1101/2020.12.15.422887.

[14] Scarselli F, Gori M, Tsoi AC, Hagenbuchner M, Monfardini G. The graph neural network model. IEEE Trans Neural Netw 2008;20(1):61–80.

[15] Carpenter AE, Jones TR, Lamprecht MR, Clarke C, Kang IH, Friman O, et al. CellProfiler: image analysis software for identifying and quantifying cell phenotypes. Genome Biol 2006;7(10):1–11.

[16] Grys BT, Lo DS, Sahin N, Kraus OZ, Morris Q, Boone C, et al. Machine learning and computer vision approaches for phenotypic profiling. J Cell Biol 2017;216(1):65–71.

[17] Gupta A, Harrison PJ, Wieslander H, Pielawski N, Kartasalo K, Partel G, Solorzano L, Suveer A, Klemm AH, Spjuth O, et al. Deep learning in image cytometry: a review. Cytometry Part A 2019;95(4):366–80.

[18] Hofmarcher M, Rumetshofer E, Clevert D-A, Hochreiter S, Klambauer G. Accurate prediction of biological assays with high-throughput microscopy images and convolutional networks. J Chem Inf Model 2019;59(3):1163–71.

[19] Chandrasekaran SN, Ceulemans H, Boyd JD, Carpenter AE. Image-based profiling for drug discovery: due for a machine-learning upgrade? Nat Rev Drug Discov 2021;20(2):145–59.

[20] Corsello SM, Bittker JA, Liu Z, Gould J, McCarren P, Hirschman JE, et al. The drug repurposing hub: a next-generation drug library and information resource. Nat Med 2017;23(4):405–8.

[21] Rodríguez MAF, Carreras-Puigvert J, Spjuth O. Designing microplate layouts using artificial intelligence. bioRxiv 2022. doi:10.1101/2022.03.31.486595.

[22] Landrum G. Rdkit documentation. Release 2013;1(1–79):4.

[23] Cereto-Massagué A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallvé S, Pujadas G. Molecular fingerprint similarity search in virtual screening. Methods 2015;71:58–63.

[24] Grattarola D, Alippi C. Graph neural networks in TensorFlow and keras with spektral [application notes]. IEEE Comput Intell Mag 2021;16(1):99–106.

[25] Kaur M, Kaur H. Implementation of enhanced graph layout algorithm for visualizing social network data using NetworkX library. Int J Adv ResComput Sci 2017;8(3).

[26] Hirohara M, Saito Y, Koda Y, Sato K, Sakakibara Y. Convolutional neural network based on SMILES representation of compounds for detecting chemical motif. BMC Bioinformatics 2018;19(19):83–94.

[27] Li X, Fourches D. SMILES pair encoding: a data-driven substructure tokenization algorithm for deep learning. J Chem Inf Model 2021;61(4):1560–9.

[28] Bjerrum E.J. Smiles enumeration as data augmentation for neural network modeling of molecules. arXiv preprint arXiv:1703.07076 2017.

[29] Murtagh F. Multilayer perceptrons for classification and regression. Neurocomputing 1991;2(5-6):183–97.

[30] Kipf T.N., Welling M.. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 2016.

[31] Kingma D.P., Ba J. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 2014.

[32] Wang P, Fan E, Wang P. Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. Pattern Recognit Lett 2021;141:61–7.

[33] Jiang D, Wu Z, Hsieh C-Y, Chen G, Liao B, Wang Z, et al. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. J Cheminform 2021;13(1):1–23.

[34] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.

[35] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: a highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 2017;30.

[36] Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Adv Neural Inf Process Syst 2018;31.

[37] Keller JM, Gray MR, Givens JA. A fuzzy k-nearest neighbor algorithm. IEEE Trans Syst Man Cybern 1985(4):580–5.

[38] LaValley MP. Logistic regression. Circulation 2008;117(18):2395–9.

[39] Breiman L. Bagging predictors. Mach Learn 1996;24(2):123–40.

[40] Pavlyshenko B. Using stacking approaches for machine learning models. In: 2018 IEEE Second international conference on data stream mining & processing (DSMP). IEEE; 2018. p. 255–8.

[41] Dietterich TG. Ensemble methods in machine learning. In: International workshop on multiple classifier systems. Springer; 2000. p. 1–15.

[42] Schapire RE. Explaining adaboost. In: Empirical inference. Springer; 2013. p. 37–52.

[43] Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks. In: International conference on machine learning. PMLR; 2019. p. 6105–14.

[44] Loshchilov I., Hutter F.. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 2017.

[45] Bland JM, Altman DG. Multiple significance tests: the Bonferroni method. BMJ 1995;310(6973):170.

[46] Sander T, Freyss J, von Korff M, Rufener C. DataWarrior: an open-source program for chemistry aware data visualization and analysis. J Chem Inf Model 2015;55(2):460–73. doi:10.1021/ci500588j.

[47] Lin A, Giuliano CJ, Palladino A, John KM, Abramowicz C, Yuan ML, et al. Off-target toxicity is a common mechanism of action of cancer drugs undergoing clinical trials. Sci Transl Med 2019;11(509):eaaw8412. doi:10.1126/scitranslmed.aaw8412.

[48] GonĂalves E, Segura-Cabrera A, Pacini C, Picco G, Behan FM, Jaaks P, et al. Drug mechanism-of-action discovery through the integration of pharmacological and CRISPR screens. Mol Syst Biol 2020;16(7):e9405. doi:10.15252/msb.20199405.

[49] Mayr A, Klambauer G, Unterthiner T, Steijaert M, Wegner JK, Ceulemans H, et al. Large-scale comparison of machine learning methods for drug target prediction on chEMBL. Chem Sci 2018;9(24):5441–51.

[50] Wieslander H, Harrison PJ, Skogberg G, Jackson S, FridĘn M, Karlsson J, et al. Deep learning with conformal prediction for hierarchical analysis of large-scale whole-slide tissue images. IEEE J Biomed Health Inform 2021;25(2):371–80. doi:10.1109/JBHI.2020.2996300.

[51] Lampa S, Alvarsson J, Arvidsson Mc Shane S, Berg A, Ahlberg E, Spjuth O. Predicting off-target binding profiles with confidence using conformal prediction. Front Pharmacol 2018;9. doi:10.3389/fphar.2018.01256.

[52] Vovk V, Gammerman A, Shafer G. Algorithmic learning in a random world. 1st ed. Springer Publishing Company, Incorporated; 2010. ISBN 1441934715, 9781441934710

[53] Haghighi M, Singh S, Caicedo JC, Carpenter AE. High-dimensional gene expression and morphology profiles of cells across 28,000 genetic and chemical perturbations. bioRxiv 2021. doi:10.1101/2021.09.08.459417.