Artificial Intelligence in the Life Sciences 3 (2023) 100063

Contents lists available at ScienceDirect

Artificial Intelligence in the Life Sciences

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

Cover sheet

AI4DR: Development and implementation of an annotation system for high-throughput dose-response experiments

Marc Bianciotto[a,*], Lionel Colliandre[b], Kun Mi[c], Isabelle Schreiber[d], Cécile Delorme[d], Stéphanie Vougier[d], Hervé Minoux[e]

[a] *Integrated Drug Discovery, Computer Aided Drug Design, Sanofi, Vitry-sur-Seine, 94403, France*
[b] *Evotec (France) SAS, in-silico R&D, Integrated Drug Discovery, Toulouse, 31100, France*
[c] *Veeva Systems France, Neuilly-sur-Seine, 92200, France*
[d] *Integrated Drug Discovery, In-Vitro Biology, Sanofi, Vitry-sur-Seine, 94403, France*
[e] *Digital and Data Sciences, Sanofi, Chilly-Mazarin, 91380, France*

ARTICLE INFO

Keywords:
Dose-response curve
Artificial intelligence
Annotation system
High throughput screening

ABSTRACT

One of the common strategies to identify novel chemical matter in drug discovery consists in performing a High Throughput Screening (HTS). However, the large amount of data generated at the dose-response (DR) step of an HTS campaign requires a careful analysis to detect artifacts and correct erroneous datapoints before validating the experiments. This step which requires to review each DR experiment can be time consuming and prone to human errors or inconsistencies. AI4DR is a system that has been developed for the classification of DR curves based on a Convolutional Neural Network (CNN) acting on normalized images of the DR curves. AI4DR allows the annotation in minutes of thousands of curves among 14 categories to help the High Throughput Screening biologists in their analyses. Several categories are associated with active and inactive compounds, other categories correspond to features of interest such as the presence of noise, a weaker effect at high doses, or a suspiciously weak or strong slope at the inflexion point of the DR curves of actives. The classifier has been trained on an algorithmically generated dataset curated and refined by experts, tested using real screening campaigns and improved using thousands of annotations by experts. The solution is deployed using a MLFlow model server interfaced with the Genedata Screener data analysis software used by the end users. AI4DR improves the consistency, the robustness, and the speed of HTS data analysis as well as reducing the human effort to identify faster new medicines for patients.

1. Introduction

High Throughput Screening (HTS) is one of the major strategies used in the pharmaceutical industry for hit finding. Lately, screening technologies have become more sophisticated [1], leading to approaches like quantitative HTS [2] and strategies where more counter-screens or selectivity assays are used to qualify hits. In turn, these techniques have raised the volume of Dose-Response (DR) results generated. Other large Dose-Response datasets are obtained after the interrogation of protein libraries by selection techniques such as phage display, yeast display or fluorescence-activated cell sorting (FACS). The quality of the Dose-Response curves (DRCs, also known as Concentration-Response curves) is dependent on the screening conditions, on the protocols and overall assay robustness, and on the behavior of the compounds. The basic automatic analysis of these curves relies on a fitting algorithm which might be unreliable in suboptimal settings [2] because of the presence of outliers due to interference effects or other technical artifacts. In practice, DRCs need to be manually reviewed and acted upon in order to lead to a decision concerning the follow-up of the corresponding compound in the project. Thus, the visual inspection step is time consuming, even more when the hit rate is high, and the outcome of this step is dependent on the quality of the curves, the experience of the expert and the time available for the analysis. When dealing with large amounts of results, this approach delays the project and might lead, over time, to a lack of consistency and robustness in the analyses.

When an active compound has an IC50 (the concentration leading to 50% of the maximum response, be it inhibition or other activity measurement) within the concentration range of the assay, the ideal shape of its DR curve is either a full sigmoid, with its low and high asymptotes well defined, or a portion of it. Ideal DRCs can be flat, either because the compound shows its maximum activity, or no activity at all, in the whole concentration range. However, there are many reasons why a

[*] Corresponding author at: Sanofi Centre de recherche de Vitry-sur-Seine, 13, quai Jules Guesde, BP 14, 94403 Vitry-sur-Seine Cedex, France.
E-mail address: Marc.bianciotto@sanofi.com (M. Bianciotto).

https://doi.org/10.1016/j.ailsci.2023.100063
Received 28 September 2022; Received in revised form 24 December 2022; Accepted 6 January 2023
Available online 6 February 2023

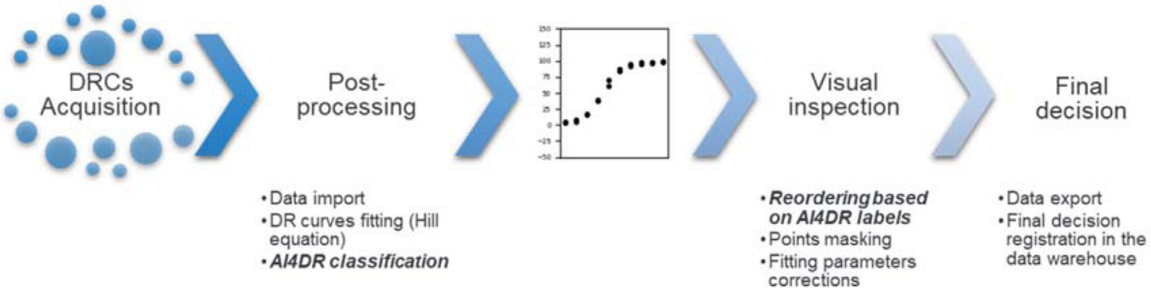
DRCs Acquisition
Post-processing
150
125
100
75
50
25
0
-25
-50
Visual inspection
Final decision
• Data import
• DR curves fitting (Hill equation)
• AI4DR classification
• Reordering based on AI4DR labels
• Points masking
• Fitting parameters corrections
• Data export
• Final decision registration in the data warehouse

Fig. 1. Typical analysis pipeline of Dose-Response Curves.

DR experiment can be perturbed, leading to curves that do not belong to these ideal shapes. Some are only related to the compound properties (e.g. colloidal aggregation or toxicity issues at high concentration) [3,4,5], others also depend on the assay type [6], the cell lines [7] or on the experiment protocol (e.g. colored and fluorescent compounds interfere with luminescence assays) [5]. The PAINS concept [8] has been promoted for the identification of structural characteristics associated with false positive compounds in assays, but PAINS compounds tested in DR assays do not necessarily show curves with defects, and vice-versa [9].

Experimental noise is also present, that can be related to many factors, up to the position of the wells in the plates, and can be considered for by a proper statistical treatment [8]. Overall, a few classifications of DRCs for active compounds have been reported [3,9] as well as specific curves defects and their origin [4,5,12] but we are not aware of a comprehensive ontology of the regular and defective DRCs that can happen in practice.

The standard post-processing workflow of DRCs consists in using the Hill equation to fit the inhibition percentage I(c) to a sigmoid and extracting several parameters, such as IC50 and its confidence interval, the slope at IC50, I(c) at the top and the bottom asymptotes (see [2] for a recent discussion of its parameters and associated uncertainties.) Anecdotal reports from HTS experts indicate that it is not always enough to examine only the fitted Hill's equation parameters to identify the problems that a DRC can display (this preconception is supported by models, as discussed in Section 3.2 "Models description"). The visual inspection of the DRCs is thus necessary, it involves masking outliers to generate a better fit (while it can be done automatically, as in [3]), confirming or adjusting the fitted parameters (top and bottom asymptotes, IC50, slope), spotting invalid experiments for re-testing, annotating valid curves displaying defects or specificities and finally tagging each curve with the final decision label: "Active" (A), "Non Active" (NA) or "Non Valid" (NV). This human validation is also necessary in terms of accountability, especially when it leads to tag a compound as active, as it might trigger costly follow-ups, such as secondary assays or batch resyntheses.

This curation step by experts goes with challenges: it is time-consuming, it can be expert-dependent, and even a single expert can face consistency issues when annotating border-line cases at different times. To alleviate these difficulties, we have developed two predictive models based on neural networks that we have integrated into AI4DR (Artificial Intelligence for Dose Response), a software infrastructure for the automated classification of DRC according to their characteristics. Overall, 14 different categories defined by experts can be given by the system, together with a classification probability. This number is similar to what is reported in other DRC classification systems [3,11]. This software solution allows experts to group together similar DRC in categories with interpretable labels for performing batch operations on them, and easily identify the less well predicted curves for an in-depth review. This procedure improves the speed of the inspection step and the robustness and consistency of the final decision. The DRC post-processing workflow, including the AI4DR-related steps, is outlined in Fig. 1.

In this paper, we present the process which led to develop AI4DR and put it in production. In this endeavor, two of the initial steps, the discussion of the scope of the classification model, and the strategy for building the training set, are to be highlighted. From the very beginning, our HTS experts have insisted on the diversity of the curves in each of the three umbrella categories (A, NA, NV), and on the practical interest to perform a fine-grained classification. This led to identify 13 different DRC shape categories to be classified, together with a category for DRC showing inter-replica dispersion. Then, as our focus was the identification of the most problematic curves, that end up being considered as NV and thus, are not stored in the corporate databases and not annotated as such, it was a challenge to obtain a good quality training set for building our predictors. We have resorted to build parametric equations for the curves of all categories and to generate the training set by random sampling of the parameter spaces in each category. This approach led to delineate the limits of the parameter space for each category. As in many cases, these limits are somewhat subjective (e.g. between a partial and a full inhibitor) and were learnt from multiple experts' annotations on a few thousands of examples.

Together with the integration of the solution into the standard users' workflow, these steps allowed to provide an efficient and consistent process for clustering and pre-annotating all the DRC from an HTS experiment, before validation by an expert.

2. Materials and methods

The setup of a robust Deep Learning model rests on several key points: 1. An appropriate data strategy leading to relevant datasets for the training and the validation of the Machine Learning (ML) models; 2. Optimized algorithms; 3. A deployment of the model which makes it easily accessible by the end user. Our approach of these three pillars is described below.

2.1. Data strategy

The initial discussions allowed to establish a list of 14 categories of DRCs that are commonly found, examples of which are shown Fig. 2. Each category has an associated label, and the two terms will be used indifferently. While the number of categories is roughly similar to what is described by other groups [10,11], our work include several categories for shape defects and DRC with inter-replica dispersion.

Six categories form a first group of flawless curves with different levels of activity:

- "Top" (CATOP) represents of a highly potent compound exhibiting full signal inhibition along the whole concentration range,
- "No Bottom" (CANB) corresponds to potent compounds with a sigmoid inhibition curve where the upper asymptote is visible but not the lower one,
- ”Sigmoid” (CASIG) is for the well-behaved sigmoid curve of active compounds, including lower and upper asymptotes,

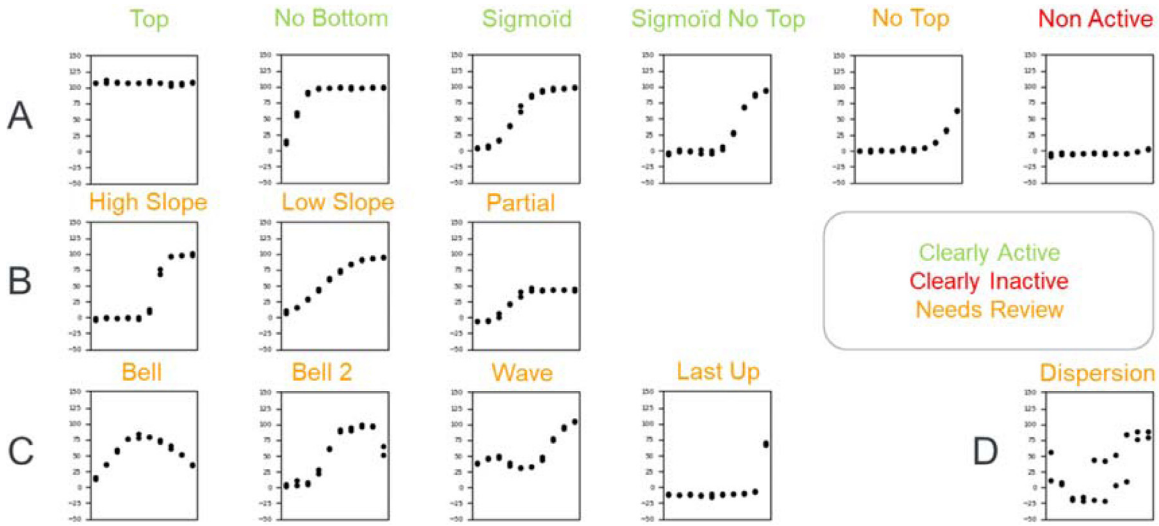

Fig. 2. DRC categories used in AI4DR. A: Flawless curves for compounds with different levels of activity. B: Sigmoid curves with extreme parameters. C: Curves with common shape defects. D: Curves showing dispersion between replicates. Categories which belong to the “Clearly Active” ensemble are in green, the categories from the “Needs Review” ensemble are in orange and the unique “Not Active” category forms the “Clearly Inactive” ensemble, in red. The “Bell” and the “Bell 2″ categories are merged in the final version of the model (see below).

- ”Sigmoid No Top” (CANT) stands for a DRC which includes the lower asymptote but not the upper one and reaches the 50% inhibition threshold,
- “No Top” (NT) is for weakly active compounds in the concentration range of the assay,
- “Non Active” (NA) DRCs are for compounds that are inactive in the assay.

A second group of 3 categories correspond to sigmoid DRC with some extreme parameters:

- “High Slope” (CAHS) stands for full sigmoid DRC with a high slope at the IC50 (nHill > 4),
- “Low Slope” (LS), conversely, describes full sigmoid DRC with a low slope at the IC50 (nHill < 0.5).

These patterns can be observed when cooperativity effects take place in the system.

- “Partial” (P) label is for full sigmoid DRC where the relative maximal inhibition (the difference between I(c) at the upper and lower asymptotes) is less than 70%, which is frequently observed in cellular assays.

The last group gathers DRCs which shows defects that are commonly observed in practice:

- In “Bell” (B) DRCs, I(c) decrease at the highest concentrations, leading to a bell-shape curve. This pattern could be due to a signal interference issue in a fluorescence assay format or to a compound aggregation issue [4,5].
- “Bell 2″ (B2, initially named “Toxicity”) label is also for full sigmoid inhibition curve with a sharply decreased inhibition at the highest concentrations. This can be due to several effects, such as the toxicity of the compound, as it is frequently observed in cellular assays, or to a hook effect [6].
- In “Wave” (W) DRCs there is an alternative increase and decrease of I(c). This could be due to various issues, such as interferences in the signal readout or a problem in the compound dilution series.
- The “Last up” (LU) label is for DRCs where no inhibition is observed in the titration curve except for the highest compound concentration, leading to a non-valid curve,
- “Dispersion” labels DRCs where noise is observed between replicates.

These categories belong to three higher order ensembles which are color-coded in Fig. 2, “Clearly Active”, “Clearly Inactive”, and “Needs Review”, this latter set being for all curves with specificities or defects. These ensembles have been used during the optimization of the algorithm to lower both the number of compounds classified in the “Needs Review” category and the number of false negatives: compounds that are misclassified as “Clearly Active” or “Clearly Inactive” while they would “Need Review”.

In terms of reference data, there are much DR data in our central data warehouse, but these data have two major drawbacks that prevent their use in a training set for our purpose. First, these are valid and curated data while we needed raw data displaying artifacts to learn these defaults. Second, these data are not annotated with the labels we were interested in. To circumvent this issue, we have generated algorithmically the training dataset before it has been reviewed by experts. Generally, the Hill equation was used for generating curves, together with a specific parameter space for each label: position of, and slope at the inflexion point, positions of the top and bottom asymptotes, before addition of noise and/or defects.

The quality of the datasets used for the training, the testing and the validation of a Machine Learning model have a strong influence on the model performance. It was thus important to build a high-quality validation dataset and to fine-tune the limits between categories, that are inherently shallow or ill-defined in some cases (e.g. between a too high, a too low, and an acceptable slope at IC50). To do so, we have organized two internal "AI4DR Challenges" with the help of Sanofi HTS experts. These crowd-sourced efforts have been performed to reach two important objectives: first, for validating the early classification models we developed, and second, for defining the limits of the parameter spaces used for generating the different categories of the training set DRCs. The experts were requested to review and annotate DRCs using a web interface for enriching the set of labelled examples. The first iteration of the Challenge, limited to HTS experts from one single site, was used to fine-tune a first incarnation of the classifier. The second AI4DR Challenge was proposed to Sanofi HTS experts worldwide to obtain a set of DRCs annotated by several experts and use the ones which make consensus for improving the initial classifier.

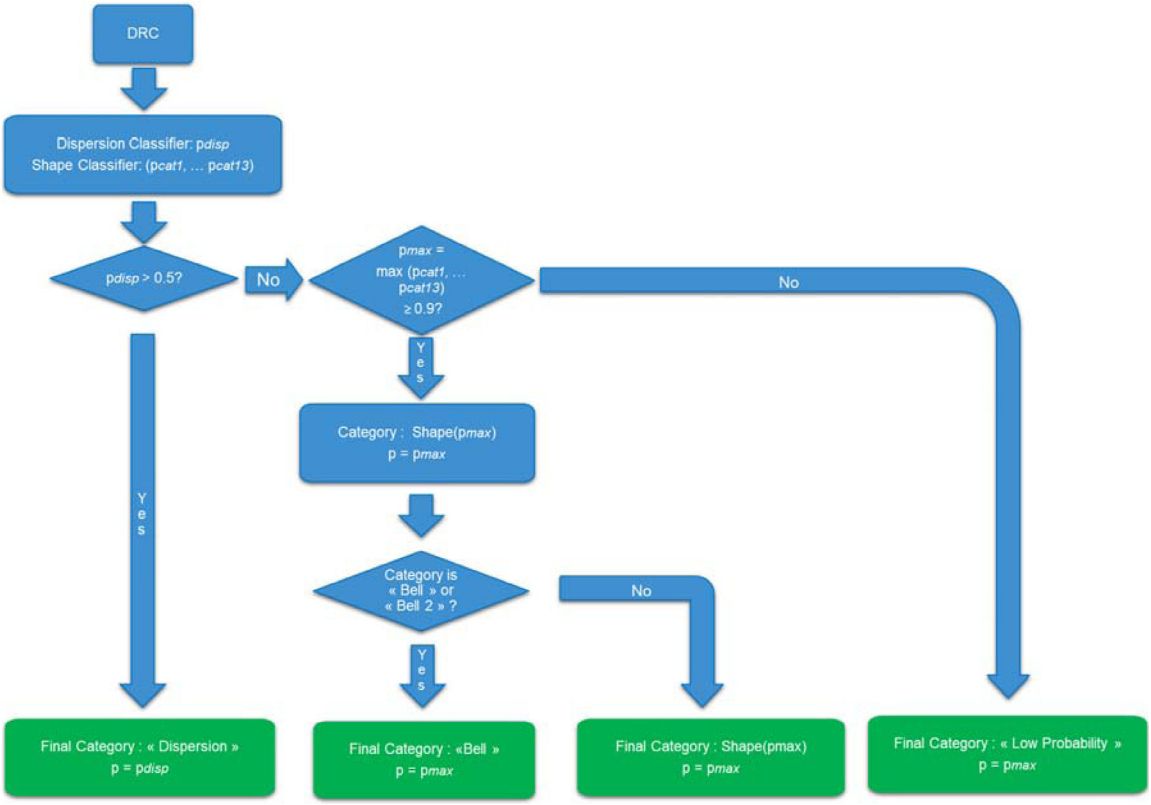
DRC
Dispersion Classifier: pdisp
Shape Classifier: (pcat1, ... pcat13)
pdisp > 0.5?
No
pmax = max (pcat1, … pcat13) ≥0.9?
No
Yes
Category : Shape(pmax)
p = pmax
Category is « Bell » or « Bell 2 » ?
No
Yes
Yes
Final Category : « Dispersion »
p = pdisp
Final Category : «Bell »
p = pmax
Final Category : Shape(pmax)
p = pmax
Final Category : « Low Probability »
p = pmax

Fig. 3. Overview of the AI4DR algorithm. The classifier for dispersion outputs a probability p_{disp} of being in the "Dispersion" category. The shape classifier output a probability for being in each of the 13 shape categories, and the category of highest probability is retained. The final label is adjusted if it is less than 0.9 or the category is "Bell" or "Bell 2″.

2.2. Classification pipeline

As the goal of our classifier is to mimic the visual inspection by experts who recognize the validity of the DR curve by its shape, we have setup a first model, the shape classifier. It is based on a Convolutional Neural Network [13] (CNN) which takes as input normalized images of the DRCs and outputs the most probable shape together with a probability estimation (which is the output of the last layer of the network for the most probable category). In an early attempt to evaluate this classifier over all the defined categories, the "Dispersion" class was found to be less well predicted than the others by the CNN. We figured out that DRCs showcasing dispersion had no specific shape so that the prediction would need to be approached differently than for the other categories. We built a "Dispersion Classifier" using the interquartile range between replicates at all concentrations and used the distribution of these ranges as the features to train a binary classifier of noisy curves implemented using a Multi-Layer Perceptron (MLP).

The whole classification pipeline is represented Fig. 3: first, the input data is normalized, and both the dispersion and the shape classifiers are applied. The DRC is annotated "Dispersion" if the associated probability is more than 0.5, otherwise the outcome of the shape classifier is considered. Curves assigned to a shape category with a probability lower than 0.9 are then reassigned to a specific "Low Probability" category associated to the "Needs Review" ensemble. Finally, if the shape category is "Bell" or "Bell 2″, it is assigned a common "Bell" label (see the rationale below).

2.3. Pipeline deployment

The availability and ease of use for the end user is of paramount importance for predictive models to be useful and have actual impact. We leveraged the possibility offered by Genedata Screener [14] server (GSS) to create custom plugins and integrated the AI4DR classifier results as a custom calculated column in the "Compounds table" view of the Genedata Screener web interface. These plugins are Python scripts that are executed on the GSS using its Jython library. For development and maintenance purposes, we chose to keep the model serving independent from Genedata infrastructure and used a MLFlow [15] server which manages and serves the model to web clients via RESTful web services [16]. The implementation of the pipeline is outlined Fig. 4: It starts when the user adds a calculated column called "AI4DR Curves Annotation" to their layer of interest. On the GSS side, the AI4DR plugin catches the compounds identifiers and the raw data (concentrations and related inhibition values for each replicate) and sends it to the model server. The server receives the REST call, performs the preprocessing of the raw data, applies the different models according to the algorithm described in Fig. 3. The final outcome, the most probable label and the associated probability, is sent back to the GSS. In the last step of the process, the results are received by the AI4DR plugin and included in the "Compounds table" view as two new sortable columns.

3. Results and discussion

3.1. Data strategy

The main classification model has been developed incrementally, starting with only two classes (A and NA). A second model was trained on 13 categories (Fig. 3, sets A to C) using 5000 computer-generated curves of each category in the training set. An early version of this model was confronted to the outcome of the first AI4DR Challenge: 1162 curves were each annotated by a single HTS expert using a combination of one to four labels among 14 categories (the 13 shape categories and the dispersion category). The discussion of the results with HTS experts lead to different improvements implemented in the second AI4DR challenge. As

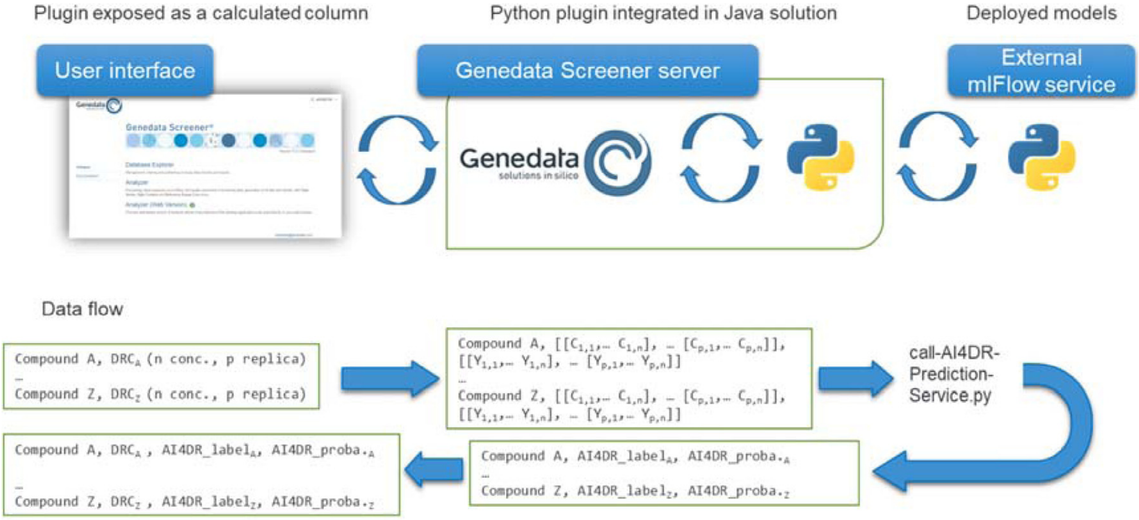
Plugin exposed as a calculated column
Python plugin integrated in Java solution
Deployed models
User interface
Genedata Screener server
External mlFlow service
Genedata Screener
Genedata
solutions in silico
Data flow
Compound A, DRC_A (n conc., p replica)
…
Compound Z, DRC_Z (n conc., p replica)
Compound A, [[C1,1,… C1,n], … [Cp,1,… Cp,n]], [[Y1,1,… Y1,n], … [Yp,1,… Yp,n]]
…
Compound Z, [[C1,1,… C1,n], … [Cp,1,… Cp,n]], [[Y1,1,… Y1,n], … [Yp,1,… Yp,n]]
call-AI4DR-Prediction-Service.py
Compound A, AI4DR_label_A, AI4DR_proba._A
…
Compound Z, AI4DR_label_Z, AI4DR_proba._Z
Compound A, DRC_A , AI4DR_label_A, AI4DR_proba._A
…
Compound Z, DRC_Z , AI4DR_label_Z, AI4DR_proba._Z

Fig. 4. AI4DR implementation architecture and data flow. Top: Implementation architecture. AI4DR is exposed to the Genedata Screener user as a calculated column using a Python plugin installed in the GSS which interacts with an external MLFlow server. Bottom: Data flow. For all DRC showed in the user interface, arrays of concentrations and responses are extracted by the AI4DR python plugin and passed over to the MLFlow server which hosts the models. It sends back a table of ids, labels and probabilities that are displayed as calculated columns in the user interface.

Table 1
Legend: Comparison between AI4DR classification and expert annotations for the dataset annotated during the AI4DR Challenge 2. The “Annotated Once” subset is for the DRC that have been annotated only once by experts, and the other two subsets have been annotated twice or three times, with or without reaching a consensus annotation.

Shape classifier performance on datasets annotated in AI4DR Challenge 2	DRC annotated once	DRC with consensus annotation	DRC without a consensus annotation
Accuracy (3 categories)	87.20%	89.81%	–
Accuracy (12 categories)	66.54%	79.57%	–
Classification agrees with one of the annotations (3 categories)	–	95.30%	96.59%
Classification agrees with one of the annotations (12 categories)	–	87.45%	80.98%
Classification errors: “Needs Review” DRCs classified as “Clearly Active/Inactive” (3 categories)	3.11%	3.88%	–

in the first one, the experts have annotated using a simple web interface 1491 DRCs coming from experimental data or computer-generated.

In total, 2965 unique annotations by 33 users were recorded: 558 curves were annotated once, 392 twice, and 541 three times. For the DRCs annotated two or three times, we defined a consensus annotation when a unique label combination (shape and defects) was found to have a majority. It led to a first dataset of 448 curves with a consensus annotation, and second one of 485 curves annotated two or three times without a consensus annotation. The main results of the AI4DR Challenge 2 on those three datasets are summarized in Table 1 and in Table S4 (that details the breakdown of the classifications by the shape and dispersion classifiers). There is a relatively low ratio of consensus annotations. It shows the difficulty of the visual inspection step for borderline cases, the difficulty to standardize this process, and the effect of the number of options offered to the experts in the challenge. More importantly, it sets the bar for the upper limit of the classification model performance, which cannot be better than the human judgement it has been trained on.

When considering the investment in HTS experts’ time necessary to setup our solution, it is interesting to note that only 4k-4.5k unique DRC annotations were needed. It is the same order of magnitude as the number of hits (considering a hit rate of c.a. 1%) submitted to the DR step after only one HTS campaign starting with 500k compounds, so that the effort is comparable to the review of DRCs from one HTS campaign only.

3.2. Models description

We will give here an overview of our models, which are described in detail in the Supplementary Material. Our main model, the shape classifier, is a CNN built using the Keras [17] API on top of the Deep Learning library Tensorflow [18] which takes gray scale DRCs images of 150 × 150 pixels as input. For generating these normalized images from the DR data, we omitted any concentration label or tick in abscissa, but we kept those in ordinate in order to differentiate between e.g. partial and full inhibition. Other featurizations of the DRCs and other Machine Learning approaches could also lead to good quality classifiers and indeed, using a CNN was not the most obvious choice. However, it allowed not only to benefit from all the recent improvements of this very popular algorithm, but also to seamlessly manage the classification of DRCs measured with a different number of doses than our training set or with missing points: control experiments showed that the performance of the classifier is not much affected by these differences in the input data (see below). Overall, the architecture and the parametrizations of the CNN were optimized to extract the general shape of the DRC regardless of its details and the presence of outliers.

We have evaluated the merits of this modeling approach by comparing the performance of this shape model with a Random Forest classifier built using the parameters extracted from the Hill’s equation fit of the DRC (R2, position of the top and bottom asymptotes, EC50, nH).

The performance metrics of the two models on the AI4DR shape classifier test set are compared in Table S1. While the RF model displays a slightly better precision than the CNN for the categories related to flawless curves (CATOP, CANB, CASIG, CANT, CAHS, CAN), it is slightly worse than the CNN for the other DRC categories, including the ones related to curves with defects (LS, B, B2 and W). The CNN also displays an equal or better recall than the RF in all but one category. It is also to note that the algorithm for fitting the Hill's equation has not converged for all DRC of the test set, leading to a different support between CNN and RF for 3 categories.

Both in terms of featurization and in terms of classification approaches, several simple approaches could be followed for estimating the presence of noise in the DRCs, e.g. based on distance metric between the Hill's function best fit and the actual datapoints. However, the best model we found for dispersion detection is a Multi-Layer Perceptron (MLP) classifier based on the normalized quartile values over the difference in I(c) between replicates (See Supp. Mat. for more details).

3.3. Models optimization and validation

In the optimization of the classification pipeline, we aimed at keeping a good balance between lowering the proportion of curves considered as false negatives (misclassified as “Clearly Active” or “Clearly Inactive”) without having too many of them in the "Needs Review" category. We found that the probability associated to the classifications was indeed a useful metric to evaluate the reliability of classifications, so we defined a probability threshold below which the DRCs were reannotated as “Low Probability” to warrant expert review. We found that setting the threshold value to 0.9 was a good compromise to balance between the classification accuracy and limitation in size of the “Needs Review” category.

The results of the first Challenge allowed us to fine-tune the boundaries between categories, which was found to have more influence on the classifier performance than the optimization of the CNN hyperparameters. During the second Challenge, we found that the "Bell" and the "Bell 2″ categories were in fact not distinguished by the HTS experts and we decided to fuse them together in a post-processing step, while keeping the two categories separated during the model training for it learn independently the slightly different visual patterns of each class. It is still to be determined if the distinction between these two categories by AI4DR might be of interest for going beyond the human-level performance.

The analysis of AI4DR performances on the 448 curves with consensus annotations from the Challenge 2 has given actionable insights on how to improve it further. The model for noise detection has flagged 48 of the consensus curves (see Table S4), among which it detected 8 curves over the 9 that were annotated “Very Disperse” (the remaining curve showed noise only for one concentration). It has also detected 27 of the 70 curves annotated “Some Noise”, and only 13 false positives were found among the 366 curves without any noise-related consensus annotation. Over the 400 remaining curves submitted to the shape classifier, 166 ended up in the “Low Probability” category. The accuracy of the classifier for the best predicted curves is 79,6% over 12 shape categories (with the “Bell” and “Bell 2″ categories merged). The confusion matrix revealed that only one type of misclassification was frequent, as 7 curves annotated as “Sigmoid” were predicted “Sigmoid No Top” with high probability.

We investigated on this consensus dataset what could be the consequence of the classifications errors on the final “Clearly Active”/”Clearly Inactive”/”Needs Review” ensembles, as each of the 12 shape categories can be related to one of these labels (see color code on Fig. 3). The accuracy of the classifier over the three final labels on the curves classified with a high probability is 89.8%, and 9 curves (3.9%) are actually false negatives: these curves ”Need Review” while they are classified either to be “Clearly Active” or “Clearly Inactive”, which is the worst-case scenario. These 9 curves are listed in Table S2. This analysis also led us to reassign the “High Slope” category, which we initially associated with the “Clearly Active” ensemble, to “Needs Review”, as we reviewed several DRCs which fall into this category that clearly belong to this ensemble.

In many cases, the expert's annotations of these false negative curves match with the category assigned by the classifier. However, the classifier misses the defects reported by experts over the main DRC shape. In all cases, the parameters obtained from the standard fitting algorithm of these curves would have drawn an alert on the need for an in-depth review.

As we have used the dataset with consensus annotations for refining limits between some categories and generally optimize the model, it could not be used as an external test of the optimized model. Therefore, to get the most from the Challenge results, we evaluated it on the set of curves annotated once and on the set of curves annotated without consensus. For the curves annotated once, the proportion of false negatives (3.1%, 9 DRCs) and the accuracy over the three final labels (87.2%) are comparable to the consensus dataset. Again, the examination of the 9 false negatives listed in Table S3 reveals that these “Need Review” examples were misclassified as “Clearly Active”, a category which is to be examined more closely by the experts in a real setting than the overwhelming set of “Clearly Inactives”. In some cases, these false negatives come from a questionable annotation from the expert or annotations with multiple defects, a possibility that was offered in the Challenge 2 but that is beyond the current capabilities of the classifier. In general, we found that these misclassifications were harmless in terms of final decision making. Finally, the accuracy of the classifier on the 12-categories shape model is notably lower (66.5% vs 79.6%) in this set than on the consensus set.

For obvious reasons, the metrics had to be modified for evaluating the non-consensual dataset, so we examined if the classifier agreed with at least one of the annotations given by the experts. Interestingly, the model performance is high over the three final labels, and comparable to the one in the consensus dataset (96.5% vs 95.3%) but is markedly lower (81.0% vs 87.5%) on the 12-categories shape model. Indeed, the classifier is more easily confused by examples that are borderline enough for experts to disagree on them.

Overall, the models were trained using 5000 generated curves for each of the 14 categories. The limits of the parameter spaces for each curve category were fine-tuned using the ground truth provided by the two challenges, and the experimental curves annotated in the Challenge 2 were used to test the AI4DR models.

3.4. Evaluation on a publicly available dataset

We also evaluated the AI4DR pipeline on several assays from Tox21 dataset [19], which is one of the very few publicly available large datasets of assigned dose-response experiments. We will describe the classification of tox21-luc-biochem-p1 assay results, which is the only biochemical assay of the Tox21 dataset as it is a counterscreen of the luciferase-based Tox21 cellular assays. It contains results for 9477 compounds tested in triplicate at 15 doses. An annotation of the curve, the “curve class”, is obtained from the fit of the data to a four-parameters Hill equation. It is provided together with an annotation of the compound's activity on the assay, the “activity category”, which can be “active agonist”, “active antagonist”, “inactive”, “inconclusive”, “inconclusive agonist” or “inconclusive antagonist” [11]. As the raw inhibition data typically ranges between −100 and 0, we preprocessed it before generating the normalized plots by subtracting to the raw values the of the minimum response for each replica, and we present in Tables S6 to S8 not only the normalized DRCs used for AI4DR classification, but also the raw DRCs. As the AI4DR classifiers have been trained with DRCs with 2 replica and 10 concentrations, and the Tox21 annotations on 15 concentrations DRCs in triplicate, we have first checked the robustness of the annotation with respect to the number of replicas. Table S5 compares the AI4DR category obtained when considering the DRCs obtained with 3 replica and those obtained only with replica 1 and 2. While in principle, there might be good reasons for the annotation to change because of the information added by the third replica, it is reassuring to note that it is the case only for c.a. 2% of the curves: the neural network predictions are stable with respect to the addition of one more replica than what was used for network training. We then examined the robustness of the annotation with respect to the number of concentrations considered to build the DRCs. Table S6 compares the AI4DR annotation obtained on 15 concentrations DRCs with those obtained on 8 concentrations DRCs by removing datapoints for the 2nd, 4th,6th … concentrations. We note that discarding nearly half of the information only leads c.a. 1.5% of the AI4DR annotations to change. We also compared the AI4DR annotations obtained using 15 concentrations DRCs in triplicate with the corresponding Tox21 assay outcomes, and the results are listed in Table S7. It is first interesting to note that all curves annotated CASIG, CANB or CAHS by AI4DR are associated to the "active antagonist" assay outcome. However, some less obvious combinations of AI4DR annotations and Tox21 assay outcomes required to examine the actual curves. For example, Table S8 lists DRCs with an "active antagonist" assay outcome that are classified as "CNA" by AI4DR, three with maximum (1.0) probability, and three with the minimum (0.9–0.91) probability. Whereas there is indeed an increase of I% with concentration in these curves, their top asymptote is located around 25% inhibition, which is below what is generally considered of interest in the industry for an antagonist, even a partial one. The annotated DRCs listed in Table S9 offers other insights on AI4DR's merits and limits: they have an "active antagonist" assay outcome and are classified as "P", "W", "LS" or "Dispersion" by AI4DR. On the first row are compared the two "active antagonist" curves annotated as "P" with the highest and the lowest probability in the dataset. While the curve with the lowest probability is similar in shape with the "CNA" curves of Table S8, albeit with a higher top asymptote, the curve with the highest probability has the typical sigmoid shape of a well-behaved partial antagonist. The two curves classified as "W" show some dispersion on the measurements for the two firsts concentrations. While the "W" classification itself might not be the most appropriate for these DRCs, it tags these curves in the "Needs Review" ensemble, which meets the objective of focusing the HTS expert attention on the cases where it is needed the most. The same is true for the three last examples of the table: while the "LS" DRC shows inhibitory behavior, in the context of an HTS analysis it would be advisable to tag the corresponding compound as it might not be the best suited for a follow-up. The "active antagonist" compound whose curve is classified "Dispersion" with the lowest probability in the dataset might be worth retesting to determine its IC50 with enough precision, as the I% measurements are spread for the two or three concentrations around IC50. We have checked examples of "active antagonists" classified as "Clearly Active" for false positives. In Table S10 are listed the DRCs with the lowest classification probability in the "CASIG", "CAHS" or "CANB" categories in the dataset. Despite the low probabilities associated to the classifications, these examples do not show obvious misclassifications. Finally, we compared the DRC that were not well predicted either with AI4DR (297 DRC in the "Low Probability" category) or with the RF classifier (248 DRC associated with a probability less than 0.5 in all categories). Two-thirds of the DRC that are not well predicted by the RF model are classified with $p>0.9$ by AI4DR, while only 10 DRC in the "Low Probability" category are classified with $p>0.9$ by the RF classifier. Table S11 shows examples of the two sets, and while the AI4DR categories are reasonable with respect to the curve's characteristics, it is less true for the RF categories of the "Low Probability" DRC: they are labelled as "B" or "P" while they show weakly active compounds with a top asymptote is located around 25% inhibition.

Overall, we find that the classification of this Tox21 dataset by AI4DR is relevant despite several significant differences (number of replica, number of concentrations) with the DRCs used for model training. This analysis confirms the robustness and the predictivity of AI4DR for classifying DRC with different parameters. As it shows several weakly active compounds annotated as active antagonists, it might also be useful for questioning studies which use the Tox21 datasets for building models [20]. We provide a github repository hosting the AI4DR shape and dispersion models, the RF model, scripts for performing the DRC annotation by both methods on the Tox21 dataset together with two analysis Jupyter notebooks for the AI4DR and the RF shape classifiers.

3.5. Working with the AI4DR plugin

The standard workflow for analyzing DR experiments in Genedata Screener is not much changed when using AI4DR: after importing the raw data, having them pre-processed to match plate wells, concentrations, and compound ID, and performing the usual curves fitting, the AI4DR annotations are added as a "calculated column", with a typical throughput of 350–400 DRCs/minute. It is then straightforward to group the curves by their AI4DR category and/or to sort them by their prediction probability to ease the review process. After the fitting parameters are adjusted and the final decision is made, the data and the final label can be exported in the data warehouse.

For the HTS expert, the main benefits of using the AI4DR plugin are related to speed and robustness: having the possibility to review similar curves together allows if needed to perform the same adjustments once for the whole set, to end up with the same final annotation. Furthermore, it allows to spot more easily specific profiles that might be related to the assay biology, as it can happen for compounds with complex mode of action identified from cellular screening. Considering DRC specificities could also benefit for the understanding of Structure-Activity Relationships by enriching the number of hits among the chemical series, revealing series-specific assay interference issues, and rescuing singleton compounds that would otherwise have probably been discarded.

Finally, building a list of common visual defects gives the opportunity in a global company to increase the robustness and consistency of the DR analysis in the different groups and move towards more standardized protocols for this time-consuming step of the value chain.

4. Conclusions

In summary, we have developed and implemented a novel approach to ease the analysis of high-throughput DR experiments using a combination of ML models. Indeed, if a few approaches have been described for comparing and clustering DRC [3,21], the use of Deep Learning approaches for the classification of DR experiments has to our knowledge not been reported yet. To build this solution, we had to solve three main issues: first, getting to a deep understanding of the business needs in order to tailor correctly the solution.

For example, an important milestone has been to agree that the objective was to assist the expert by looking for a reasonable annotation that would be further reviewed, rather than looking for a very-high quality model of the final annotation. This decision had profound consequences not only on the buyout from the users, but also on the whole modeling approach, from the generation and curation of the datasets to the optimization of the different models used in the solution. The second issue was related to the size and the quality of the dataset on which AI approaches are based, that we tackled by generating the training set algorithmically and by organizing the two AI4DR Challenges. While typical image recognition problems generally require large training sets to build a well-performing AI-based classifier, a very decent performance is obtained here with only a few thousands curated samples, that can be obtained with a reasonable effort from HTS experts. Finally, the third issue, which was to maximize the impact and the use of AI4DR by making it easy to use, has been met by leveraging the possibilities offered by GSS to integrate custom solutions within our standard HTS processing toolchain.

While AI4DR has been developed with the Sanofi context in mind, its application to an external dataset from Tox21 has demonstrated its robustness and usefulness in quickly correctly classifying a large set of DRC. In addition, during this exercise, we have identified interesting cases where questionable assignments could be easily identified and revised by experts. This could be important in two contexts. First, when using experimental datasets to develop a predictive model using machine learning, where the quality of data is crucial for the predictivity of the model made. In this context, AI4DR could serve to filter out questionable results and keep only clearly active compounds. Second, when large dataset of DRC is analyzed by several scientists (e.g. in the context of collaborations or consortia), AI4DR could remove some inherent bias between experts to have a homogeneous primary assignment and clustering before validation.

Overall, we believe this tool that is being used routinely inhouse for the last two years is a first step in the use of AI approaches to improve the quality and the speed in analyzing in-vitro results in drug discovery that could lead to new applications eventually beyond this original scope and to future improvements. A patent application has been filed for the methods used in AI4DR [22].

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Marc Bianciotto, Isabelle Schreiber, Cécile Delorme, Stéphanie Vougier, and Hervé Minoux are employees of Sanofi.

Kun Mi is a former employee of Sanofi.

Lionel Colliandre was consultant at Sanofi when the research was performed.

Marc Bianciotto and Kun Mi are inventors of the patent application filed under number WO2022112568A1 relative to the research presented in the manuscript.

Data availability

We make the data and code available on a github repository indicated in the manuscript

Acknowledgments

We gratefully acknowledge the Sanofi HTS experts who have participated in the AI4DR Challenges, and Tamim Al-Ahmad for his contributions to the initial proof of concept.

Funding

All authors were Sanofi employees or consulting for Sanofi at the time the research was performed.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ailsci.2023.100063.

References

[1] Busby SA, Carbonneau S, Concannon J, et al. Advancements in assay technologies and strategies to enable drug discovery. ACS Chem Biol 2020;15:2636–48.
[2] Shockley KR. Quantitative high-throughput screening data analysis: challenges and recent advances. Drug Discov Today 2015;20:296–300.
[3] Davies G, Vincent J, Packer MJ, Murray D. Grouping concentration response curves by features of their shape to aid rapid and consistent analysis of large data sets in high throughput screens. Slas Discov 2022;27:272–7.
[4] Owen SC, Doak AK, Ganesh AN, et al. Colloidal drug formulations can explain "Bell-shaped" concentration–response curves. ACS Chem Biol 2014;9:777–84.
[5] Jadhav A, Ferreira RS, Klumpp C, et al. Quantitative analyses of aggregation, autofluorescence, and reactivity artifacts in a screen for inhibitors of a thiol protease. J Med Chem 2010;53:37–51.
[6] Butch AW. Dilution protocols for detection of hook effects/prozone phenomenon. Clin Chem 2000;46(10):1719–20.
[7] Fallahi-Sichani M, Honarnejad S, Heiser LM, et al. Metrics other than potency reveal systematic variation in responses to cancer drugs. Nat Chem Biol 2013;9:708–14.
[8] The Ecstasy and Agony of Assay Interference Compounds. J Chem Inf Model 2017;57:387–90.
[9] Kenny PW. Comment on the ecstasy and agony of assay interference compounds. J Chem Inf Model 2017;57:2640–5.
[10] Parham F, Austin C, Southall N, et al. Dose-response modeling of high-throughput screening data. J Biomol Screen 2009;14:1216–27.
[11] Huang R. A quantitative high-throughput screening data analysis pipeline for activity profiling. In: Zhu H, Xia M, editors. High-throughput screening assays in toxicology. New York, NY: Springer New York; 2016. p. 111–22.
[12] Shoichet BK. Interpreting steep dose-response curves in early inhibitor discovery. J Med Chem 2006;49:7274–7.
[13] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44.
[14] Genedata. Genedata Screener. Accessed December 19th, 2022. https://www.genedata.com/products-services/screener
[15] Zaharia, M.; Chen, A.; Davidson, A.; et al. Accelerating the Machine Learning Lifecycle with MLflow. 2018, 41, 39–45.
[16] Richardson L, Ruby S. RESTful web services. O'Reilly Media, Inc; 2007.
[17] Chollet, F. Keras. https://github.com/fchollet/keras
[18] Abadi, M.; Agarwal, A.; Barham, P.; et al. TensorFlow: large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/
[19] Huang R, Xia M, Cho M-H, et al. Chemical genomics profiling of environmental chemical modulation of human nuclear receptors. Environ Health Persp 2011;119:1142–8.
[20] Borrel A, Huang R, Sakamuru S, et al. High-throughput screening to predict chemical-assay interference. Sci Rep 2020;10:3986.
[21] Sedykh A. CurveP method for rendering high-throughput screening dose-response data into digital fingerprints. In: Zhu H, Xia M, editors. High-throughput screening assays in toxicology. New York, NY: Springer New York; 2016. p. 135–41.
[22] Bianciotto, M.; Mi, K. Classifying Images Of Dose-Response Graphs. WO2022112568A1, 2020.