KekuleScope: improved prediction of cancer cell line sensitivity using convolutional neural networks trained on compound images

Isidro Cortés-Ciriano\textsuperscript{1,*} and Andreas Bender\textsuperscript{1}

\textsuperscript{1}Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB2 1EW, United Kingdom.

*Corresponding author: isidrolauscher@gmail.com
Abstract

The application of convolutional neural networks (ConvNets) to harness high-content screening images or 2D compound representations is gaining increasing attention in drug discovery. However, existing applications often require large data sets for training, or sophisticated pretraining schemes for the networks. Here, we show on eight cytotoxicity IC\textsubscript{50} data sets from ChEMBL 23 that the \textit{in vitro} activity of compounds on cancer cell lines can be accurately predicted on a continuous scale from their Kekulé structure representations alone by extending existing architectures (AlexNet, DenseNet-201, ResNet152 and VGG-19), which were pretrained on unrelated image data sets. We show that the predictive power of the generated models, which just require standard 2D compound representations as input, is comparable to that of Random Forest (RF) models trained on circular (Morgan) fingerprints, a combination which is considered to be the state of the art. Notably, including additional fully-connected layers further increases the predictive power of the networks by up to 10%. Analysis of the predictions generated by RF models and ConvNets shows that by simply averaging the output of the RF models and ConvNets we constantly obtain significantly lower errors in prediction (4-12% decrease in RMSE on the test set) than those obtained with either model alone, indicating that the features extracted by the convolutional layers of the ConvNets provide complementary predictive signal to Morgan fingerprints. Overall, in this work we present a set of ConvNet architectures for the prediction of compound activity from their Kekulé structure representations with state-of-the-art performance, that require no generation of compound descriptors or use of sophisticated image processing techniques. The data sets and all the code needed to reproduce the results presented in this study are provided at https://github.com/isidroc/kekulescope.
Introduction

Cultured cancer cell lines are limited disease models in that they do not recapitulate the tumor microenvironment nor interactions with the immune system\textsuperscript{1–6}, fundamental properties of cellular organization are altered in culture\textsuperscript{7}, and their response to anticancer drugs is affected by both assay heterogeneity\textsuperscript{8} and genomic alterations acquired in vitro\textsuperscript{9}. However, cancer cell lines still represent versatile models to study fundamental aspects of cancer biology\textsuperscript{10,11}, and the genomic determinants of drug response\textsuperscript{3,12–14}. Hence, the development of computational methods to harness the large amount of in vitro sensitivity data collected to date to unravel the underlying molecular mechanisms mediating drug activity and identify novel biomarkers for drug response is an area of intense research\textsuperscript{14–20}.

Whereas existing computational tools to model in vitro compound activity mostly rely on established algorithms (e.g., Random Forest or Support Vector Machines), the utilization of deep learning in drug discovery is gaining momentum, a trend that is only expected to increase in the coming years\textsuperscript{21}. Deep learning techniques have been already applied in numerous drug discovery tasks, including toxicity modelling\textsuperscript{22,23}, bioactivity prediction\textsuperscript{24–30}, and de novo drug design\textsuperscript{31–34}, among others. Most of these studies have utilized feedforward neural networks consisting of multiple fully-connected layers trained on one of the many compound descriptors developed over the last >30 years in the chemoinformatics field\textsuperscript{27,35}. However, the high performance of convolutional neural networks (ConvNets)\textsuperscript{36–38}, a type of neural networks developed for image recognition tasks, in finding complex high-dimensional relationships in diverse image data sets is fostering their application in drug discovery\textsuperscript{21,39}.

ConvNets consist of two sets of layers (Figure 1): (i) the convolutional layers, which extract features from the input images, and (ii) the classification/regression layers, which are generally fully-connected layers that output one value for each of the tasks being modelled. A major advantage of ConvNets is that the extraction of features is performed on a fully automatic and data-driven fashion, thus not requiring to engineer feature selection or image preprocessing filters beforehand\textsuperscript{39}. Today, convolutional neural networks are applied to diverse image recognition tasks in healthcare and biomedicine\textsuperscript{40–43}. An obvious critical element for the application of ConvNets is the availability of images for training, or the ability to formulate the modelling task of interest as an image classification problem. An illustrative example of the latter is DeepVariant\textsuperscript{44}, a recently proposed algorithm that uses images of sequencing read pileups as
input to detect small indels and single-nucleotide polymorphisms, instead of statistical modelling of the genotypes supported by the data, as has been the standard approach for years.

In drug discovery, applications of ConvNets include elucidation of the mechanism of action of small molecules from high-content screening images\textsuperscript{45,46}, and modelling \textit{in vitro} assay endpoints using 2D representations of compound structures, termed “compound images”, as input\textsuperscript{23,47–50}. Efforts to model compound activity using ConvNets trained on compound images were spearheaded by Goh \textit{et al}., who developed \textit{Chemception}\textsuperscript{47,51}, a ConvNet based on the Inception-ResNet v2 architecture\textsuperscript{52}. The performance of \textit{Chemception} was compared to multi-layer perceptron deep neural networks trained on circular fingerprints in three tasks: prediction of free energy of solvation (632 compounds; regression), inhibition of HIV replication (41,193; binary classification), and compound toxicity using data from the “Toxicology in the 21st Century” (Tox21) project (8014; multi-task binary classification)\textsuperscript{53}. \textit{Chemception} slightly outperformed the multi-layer perceptron networks except for the TOX21 task. In a follow-up study, the same group introduced \textit{ChemNet}\textsuperscript{47}, a learning strategy that consists of pre-training a ConvNet (\textit{e.g}., \textit{Chemception}\textsuperscript{51}) on a large set of compounds (1.7M) to predict their physicochemical properties (\textit{e.g}., logP, which represents an easy task) in order to learn general aspects of the images related to chemistry. Subsequently, the trained networks are applied to model smaller data sets using transfer learning. Although such an approach led to higher performance than \textit{Chemception}, a major disadvantage thereof is that it requires the initial training of the network on a large set of compounds, which is computationally demanding. More recently, Fernández \textit{et al}.

proposed \textit{Toxic Colors}, a framework to classify toxic compounds from the TOX21 data set using compound images as input\textsuperscript{23}. Although these studies have paved the way for the application of ConvNets to model the bioactivity of compounds using their images as input, a comprehensive analysis of ConvNet architectures with a reduced computational footprint to model cancer cell line sensitivity on a continuous scale and comparison against the state of the art is still missing. Moreover, whether the combination of models trained on widely-used compound descriptors (\textit{e.g}., circular fingerprints) and ConvNets trained using compound images leads to increased predictive power remains to be studied.

Here, we introduce \textit{KekuleScope}, a flexible framework for modelling the bioactivity of compounds on a continuous scale from their Kekulé structure representation using ConvNets pretrained to model unrelated image classification tasks. We demonstrate using eight
cytotoxicity data sets from ChEMBL version 23 (Table 1) that this type of compound images convey enough predictive power to build robust models using ConvNets. Instead of using networks pretrained on compound images\textsuperscript{47}, we show that widely-used architectures for image classification tasks (AlexNet\textsuperscript{54}, DenseNet-201\textsuperscript{55}, ResNet152\textsuperscript{56} and VGG-19\textsuperscript{57}) are versatile enough to generate robust predictions across a dynamic range of bioactivity values using compound images as input. Moreover, comparison with Random Forest models trained on circular fingerprints (Morgan fingerprints\textsuperscript{58,59}) reveals that ConvNets trained using compound images lead to comparable or higher predictive power on the test set. In addition, combining RF and ConvNet predictions into model ensembles constantly leads to increased model performance, suggesting that the features extracted by the convolutional layers of the networks provide complementary information to Morgan fingerprints. Therefore, our work presents a novel framework for the prediction of compound activity that requires minimal processing of chemical structures and no descriptor choice, and that leads to improved predictive power over the state of the art in our validation on eight cancer cell line sensitivity data sets.
Methods

Data Collection and Curation

We gathered cytotoxicity IC$_{50}$ data for eight cancer cell lines from ChEMBL database version 23 using the chembl_webresource_client python module$^{60-62}$. To gather high-quality cytotoxicity data sets, we only kept IC$_{50}$ values for small molecules that satisfied the following stringent filtering criteria$^8$: (i) activity unit equal to “nM”, (ii) activity relationship equal to ‘=’. The average pIC$_{50}$ value was calculated when multiple IC$_{50}$ values were annotated for the same compound-cell line pair. IC$_{50}$ values were modeled in a logarithmic scale (pIC$_{50}$ = −log$_{10}$ IC$_{50}$ [M]). Further information about the data sets is given in Table 1. All data sets used in this study are available at https://github.com/isidroc/kekulescope.

| Cell line | Description                  | ChEMBL Cell ID | Cellosaurus ID | Organism of origin | Number of bioactivity data points |
|-----------|------------------------------|----------------|----------------|--------------------|----------------------------------|
| A2780     | Ovarian carcinoma cells      | CHEMBL3308421  | CVCL_0134      | Homo sapiens       | 2,255                            |
| CCRF-CEM  | T-cell leukemia              | CHEMBL3307641  | CVCL_0207      | Homo sapiens       | 3,047                            |
| DU-145    | Prostate carcinoma           | CHEMBL3308034  | CVCL_0105      | Homo sapiens       | 2,512                            |
| HCT-15    | Colon adenocarcinoma cells   | CHEMBL3307945  | CVCL_0292      | Homo sapiens       | 994                              |
| KB        | Squamous cell carcinoma      | CHEMBL3307959  | CVCL_0372      | Homo sapiens       | 2,731                            |
| LoVo      | Colon adenocarcinoma cells   | CHEMBL3307691  | CVCL_0399      | Homo sapiens       | 1,120                            |
| PC-3      | Prostate carcinoma cells      | CHEMBL3307570  | CVCL_0035      | Homo sapiens       | 4,294                            |
| SK-OV-3   | Ovarian carcinoma cells      | CHEMBL3307746  | CVCL_0532      | Homo sapiens       | 1,589                            |

Molecular Representation

We standardized all chemical structures to a common representation scheme using the python module standardizer (https://github.com/flatkinson/standardiser). Inorganic molecules were removed, and the largest fragment was kept in order to filter out counterions. We note that, although imperfect, removing counterions is a standard procedure in the field$^{63,64}$. In addition, salts are not generally well-handled by descriptor calculation software, and hence, filtering them out is generally preferred$^{65}$. 
Kekulé structure representations for all compounds (i.e., ‘compound images’) in Scalable Vector Graphics (SVG) format were generated from the compound structures in SDF format using the RDkit function `MolsToGridImage` and default parameter values. SVG images were then converted to Portable Network Graphics (PNG) format using the programme `convert` (version ImageMagick 6.7.8-9 2016-03-31 Q16; http://www.imagemagick.org) and resized to 224 x 224 pixels using a density (-d argument) of 800. The code needed to reproduce the results presented in this study are provided at https://github.com/isidroc/kekulescope.

To represent molecules for subsequent model generation based on fingerprints, we computed circular Morgan fingerprints\(^{58}\) for all compounds using RDkit (release version 2013.03.02)\(^{66}\). The radius was set to 2 and the fingerprint length to 256.

**Machine Learning**

- **Data Splitting**

  The data sets were randomly split into a training (70% of the data), validation (15%), and test set (15%). For each data set, the training set was used to train a given network, the validation set served to monitor its predictive power during the training phase, and the test set served to assess the predictive power after the ConvNets were trained.

- **Network Architecture and Training**

  ConvNets pretrained on the ImageNet\(^{67}\) data set were downloaded using the python library Pytorch\(^{68}\). The structure of the classification layer(s) in each of the architectures used was modified to output a single value, corresponding to compound pIC\(_{50}\) values in this case, by removing the softmax transformation of the last fully connected layer (which is used in classification tasks to output class scores in the 0-1 range). The Root Mean Squared Error (RMSE) value on the validation set was used as the loss function during the training phase of the ConvNets, and to compare the predictive power of RF and ConvNets on the test set. We performed grid search to find the optimal combination of parameters for all networks. The parameter values considered are listed in Table 2.

We generated an extended version of each architecture by including five fully-connected layers (Figure 1C), consisting of 4,096, 1000, 200 and 100 neurons (Figure 1). Thus, for each architecture we implemented two regression versions, one containing one fully-connected layer, and a second one containing five fully-connected layers (abbreviated from now on as “extended”). The feature extraction layers were not modified.
Table 2. Parameters tuned during the training phase using grid search. The names in parentheses indicate the parameter name abbreviation used in the main text and figures.

| Parameter                        | Values evaluated  |
|----------------------------------|-------------------|
| Learning rate (Lr)               | {0.1, 0.01, 0.001, 0.005, 0.001, 0.0001} |
| Decay rate                       | {0.1, 0.6}        |
| Annealing rate step              | {10, 25}          |
| Data augmentation (Augmentation) | {Yes: 1, No: 0}   |
| Batch size (Batch)               | {4, 16, 32}       |

In cases where the data sets were augmented, the following transformations were applied (as implemented in the Pytorch library): (i) 180° rotation about the vertical axis (function `transforms.RandomHorizontalFlip`); (ii) 180° rotation about the horizontal axis (function `transforms.RandomVerticalFlip`); and (iii) random 90° rotation (function `transforms.RandomRotation`). In the three cases, each transformation was applied at every epoch during the training phase with a 50% chance. Thus, in some cases a set of the images might remain intact depending on this sampling step during a given epoch.

We used stochastic Gradient Descent algorithm with Nesterov momentum to train all networks, which was set to 0.9 and kept constant during the training phase. The parameters for all layers, including the convolutional and regression layers, were optimized during the training phase. Networks were allowed to evolve over 600 epochs. To reduce the chance of overfitting, we used early stopping, i.e., the training phase was stopped if the validation loss did not decrease after 250 epochs, and 50% dropout in the five fully-connected layers (labelled as “regression layers” in Figure 1) in the extended versions of the architectures considered. The training phase was divided into cycles, throughout which the learning rate was annealed and set back to its original value at the beginning of the next cycle. The learning rate was decreased by 90 or 40% every 10 or 25 epochs (decay rates of 0.1 and 0.6, respectively; Table 2).

- Random Forest (RF)

RF models were generated using the python library scikit learn based on Morgan fingerprint representations, which were calculated as described above. Default parameter values were used except for the number of trees, which was set to 100 because higher values do not generally increase model performance when modelling bioactivity data sets. Identical data splits were used to train the ConvNets and the RF models.

Experimental Design
To compare the predictive power of the ConvNets and RF models in a robust statistical manner, we designed a balanced fixed-effect full-factorial experiment with replications\textsuperscript{74}. The following factors were considered:

(i) \textit{Data set}: 8 data sets (Table 1).

(ii) \textit{Model}: 7 convolutional network architectures.

(iii) \textit{Batch size (Batch)}: number of compound images processed in each batch during the training phase.

(iv) \textit{Data Augmentation (Augmentation)}: binary variable indicating whether data augmentation was applied during the training phase.

We implemented the following linear model to study this factorial design:

\textit{Equation 1:}

\[ pIC_{50} = \text{Data set}_i + \text{Model}_j + \text{Batch}_k + \text{Augmentation}_l + \mu_0 + \epsilon_{i,j,k,l,m} \]

\((i \in \{1, ..., N_{\text{data sets}} = 8\}; j \in \{1, ..., N_{\text{models}} = 7\}; k \in \{1, ..., N_{\text{batch sizes}} = 3\};

l \in \{1, ..., N_{\text{augmentation}} = 2\}; m \in \{1, ..., N_{\text{repetitions}} = 10\})\)

where the factors \textit{Data set}, \textit{Model}, \textit{Batch}, \textit{Augmentation}, are the main effects considered in the model. The levels “A2780” (\textit{Data set}), “AlexNet” (\textit{Model}), “4” (\textit{Batch}), and “0” (\textit{Augmentation}) were used as reference factor levels to calculate the intercept term of the linear model, \(\mu_0\), which corresponds to the mean \(pIC_{50}\) value for this combination of factor levels. The coefficients (\textit{i.e.}, slopes) for the other combinations of factor levels correspond to the difference between their mean \(pIC_{50}\) value and the intercept. The error term, \(\epsilon_{i,j,k,l,m}\), corresponds to the random error of each \(pIC_{50}\) value, defined as \(\epsilon_{i,j,k,l,m} = pIC_{50i,j,k,l,m} - \text{mean}(pIC_{50i,j,k,l})\). These errors are assumed to (i) be mutually independent, (ii) have zero expectation value, and (iii) have constant variance.

We trained ten models for each combination of factor levels, each time randomly assigning different sets of data points to the training, validation and test sets. The normality and homoscedasticity assumptions of the linear models were respectively assessed with (i) quantile–quantile (Q-Q) plots and (ii) by plotting the fitted values against the residuals\textsuperscript{74}. Homoscedasticity means that the residuals are equally dispersed across the range of the dependent variable used in the linear model. A systematic bias of the residuals would indicate
that the errors are not random and that they contain predictive information that should be included in the model\textsuperscript{75,76}.

To compare the performance of RF, the most predictive ConvNet for each data set and replication, and the Ensemble models, we also used a linear model with two factors, namely \textit{Data set} and \textit{Model}. In this case, we only considered the results of the ConvNet architecture leading to the lowest RMSE value on the test set for each data set and replication.

\textit{Equation 2:}
\[ pIC_{50} = Data\ set_i + Model_j + (Data\ set \ast Model)_{i,j} + \mu_0 + \varepsilon_{i,j,k} \]
\[(i \in \{1, \ldots, N_{\text{data sets}} = 8\}; j \in \{1, \ldots, N_{\text{models}} = 3\}) \]
Results and Discussion

We initially evaluated the performance of ConvNets to predict the activity of compounds from their Kekulé structure representations. To this aim, we generated models for all data sets using four widely-used architectures, namely AlexNet, DenseNet 201, ResNet-152, and VGG-19 with batch normalization (VGG-19-bn), and the extended versions thereof that we implemented by including four additional fully-connected layers after the convolutional layers (see Methods and Figure 1). We obtained high performance on the test set for all networks, with mean RMSE values in the 0.65-0.96 pIC$_{50}$ range (Figure 2). These errors in prediction are comparable to the uncertainty of heterogeneous IC$_{50}$ measurements in ChEMBL$^8$, and to the performance of drug sensitivity prediction models previously reported$^{15, 18, 77}$. Notably, high performance was also obtained for data sets containing few hundred compounds (e.g., LoVo or HCT-15), suggesting that the framework proposed here is applicable to model small data sets.

In order to study the relative performance of the networks in a robust manner, we implemented a factorial design that we evaluated using a linear model (Equation 1). The linear model displayed an $R^2$ value adjusted for the number of parameters of 0.68 ($P < 10^{-12}$), thus indicating that the variables considered in our factorial design explain a large proportion of the variation observed in model performance, and hence, its coefficients provide valuable information to study their relative performance in a statistically sound manner. Analysis of the model coefficients revealed that the performance of the extended versions of the architectures considered constantly led to a decrease in the RMSE values of ~5-10% ($P < 10^{-12}$; Figure 2), with ResNet-152, and VGG-19-bn constantly leading to the highest predictive models. Together, these results thus suggest that the four additional fully-connected layers we included in the architectures and the use of dropout regularization help palliate overfitting (Figure 2), and hence, increase the generalization capabilities of the networks.

We previously showed that data augmentation represents a versatile approach to increase the predictive power of Random Forest models trained on compound fingerprints$^{78}$. Similarly, we here find a significant increase in performance for ConvNets trained on augmented data sets ($P = 0.02$). In fact, the utilization of data augmentation during training led to the most predictive models in 68% of the cases; when considering the most predictive network for each data set and run only, we find that data augmentation was used in 91% of the cases. Overall, these
results indicate that the extraction of chemical information by the ConvNets is robust against rotations of the compound images, and that data augmentation helps improve chemical-structure activity modelling based on compound images\textsuperscript{78}.

In addition to assessing the predictive power on the test set, we performed Y-scrambling experiments to ensure that the predictive power obtained by the ConvNets did not arise by chance. With this aim in mind, the bioactivity values for the training and validation set instances were shuffled before training. We observed $R^2$ values around 0 ($P < 0.001$) for the observed against the predicted values on the test set for all the Y-scrambling experiments we performed. Therefore, these results indicate that the features extracted by the convolutional layers capture chemical information related to bioactivity.

Next, we compared the predictive power of the ConvNets to that of RF models trained on Morgan fingerprints using the factorial design described in Equation 2. The linear model in this case showed an adjusted $R^2$ value of 0.97, suggesting that the covariates we considered account for most of the variability in model performance. Overall, we did not find significant differences in performance between RF models trained on circular fingerprints and ConvNets trained on compound images ($P = 0.76$; Figures 3-4), although the average performance of ConvNets was slightly higher than RF for several data sets. These models are using Morgan FP and RF, which have previously been shown to generate models with high predictive power in benchmarking studies of compound descriptors and algorithms\textsuperscript{78–80}. Taken together, these results suggest that compound images provide sufficient predictive signal to generate ConvNets with comparable predictive power to state-of-the-art methods, even for small data sets of few hundred compounds.

To further characterize the differences between RF and ConvNets, we firstly assessed the correlation between the predicted values calculated for the same test set instances using models trained on the same data splits. We found (as shown in Figure 5) that the predictions of both models are highly correlated for all data sets, with $R^2$ values in the 0.80-0.89 range (Pearson’s correlation coefficient; $P < 0.05$), thus indicating that the predictions calculated with the RF models explain a large fraction of the variance observed for the predictions calculated with the ConvNets, and \textit{vice versa}. Analysis of the correlation of the absolute error in prediction for each test set instance however revealed that the error profiles of RF and ConvNets are only
moderately correlated ($R^2$ in the 0.58-0.65 range, $P < 0.05$; Figure 6). From the latter, we hypothesized that combining the predictions generated by each modelling technique into a model ensemble might lead to increased predictive power$^{32}$. In fact, ensemble models built by averaging the predictions generated by RF and ConvNet models constantly displayed higher predictive power, leading to 4-12% and 5-8% decrease in RMSE values with respect to RF and ConvNet models, respectively ($P < 10^{-5}$; green bars in Figure 3). In contrast to previous analyses$^{47}$, where compound fingerprints and related representations were often thought to contain most information related to bioactivity$^{81}$, our results indicate that Morgan fingerprints and the features extracted from compound images with the ConvNets convey complementary predictive signal, thus permitting to obtain more accurate predictions than either model alone by combining them into a model ensemble.

Finally, we examined the distributions of the residuals for the ConvNet and RF models. This analysis serves to ensure that the low RMSE values observed are not a consequence of simply predicting the mean value of the response variable, because, as we have previously shown for protein-ligand systems$^{83}$, networks that fail to converge often simply predict the mean value of the dependent variable. Overall, we observed similar patterns for both modelling approaches (as shown in Figure 7), with residuals centered around zero and generally showing homoscedasticity, i.e., displaying comparable variance across the entire bioactivity range. Examination of the residuals is also important when modelling imbalanced data sets, which is generally the case for data sets extracted from ChEMBL, because a large fraction of instances are annotated with pIC$_{50}$ values in the low micromolar range (4-5 pIC$_{50}$ units), and by simply predicting the mean value of the response variable one might already obtain low RMSE values. In such cases, the residuals would be heteroscedastic, displaying increasingly higher variances towards the low-nanomolar range (i.e., pIC$_{50}$ values of 8-9), which however was not the case for the models generated here. Together, these results thus indicate that compound images convey sufficient chemical information to model compound bioactivities across a wide dynamic range of pIC$_{50}$ values.

In this work, we show that a proper design and parametrization of ConvNets is sufficient to generate highly predictive models trained on images of structural compound representations sketched using standard functionalities of commonly used software packages (e.g., RDkit). Therefore, exploiting such networks which were designed for general image recognition tasks,
and pre-trained on unrelated image data sets, represents a versatile approach to model compound activity directly from Kekulé structure representations in a purely data-driven fashion. However, it is paramount to note that the computational footprint of ConvNets still represents a major limitation of this approach: whereas training the Random Forest models for these data sets required 6-14 seconds per model using 16 CPU cores and no parameter optimization, training times per epoch for the ConvNets were in the 15-64 seconds range (i.e., 150-640 minutes per model using one GPU card and 16 CPU cores for image processing).

While the computation of compound descriptors has traditionally relied on predefined rules or prior knowledge of chemical properties, bioactivity profiles or topological information of compounds, among others, the descriptors calculated by the convolutional layers of ConvNets represent an automatic and data-driven approach to derive features from chemical structures. As we show in this study, these compound features permit to model compound activity with high accuracy even on a continuous scale. However, image-derived features are still harder to interpret than more traditional descriptors e.g., unhashed Morgan fingerprints. We anticipate that extending the work presented here by including 3D representations of compounds and binding sites using 3D convolutional neural networks to account for conformational changes of small molecules and protein dynamics, respectively, will likely improve compound activity modelling. Finally, future work will also be required to evaluate whether ConvNets trained on both compound and cellular images lead to more accurate modelling of compound activity on cancer cell lines, as well as other output variables (i.e., toxicity), than current modelling approaches based on gene expression or mutation profiles.
Conclusions

In this contribution, we introduce *KekuleScope*, a framework to model compound activity on a continuous scale using extended versions of four widely-used architectures trained on Kekulé structure representations without requiring any image preprocessing steps. The generated models achieve comparable performance to RF models trained on circular fingerprints, and to the estimated experimental uncertainty of the input data. Our work shows that Kekulé representations can be harnessed to derive robust models without requiring any additional descriptor calculation. In addition, we show that the chemical information extracted by the convolutional layers of the ConvNets is complementary to that provided by Morgan fingerprints, which enables the generation of model ensembles with significantly higher predictive power than either RF models or ConvNets alone. The framework proposed here is generally applicable across endpoints, and it is expected that also on other datasets the combination of models will lead to increases in performance.
Author Contributions

I.C.-C. conceived and designed the study. I.C.-C. implemented the models, and interpreted and analyzed the results. I.C.-C. generated the figures and wrote the paper with substantial input from A.B.

Acknowledgements

This project has received funding from the European Union’s Framework Programme For Research and Innovation Horizon 2020 (2014-2020) under the Marie Curie Sklodowska-Curie Grant Agreement No. 703543 (I.C.C.).

Conflicts of Interest

The authors declare no conflict of interests.

Supporting Information Available: The code and data sets used in this study is available free of charge at https://github.com/isidroc/kekulescope.
Figures

**Figure 1 KekuleScope framework.** (A) We collected and curated a total of eight cytotoxicity data sets from ChEMBL version 23. (B) Compound Kekulé representations were generated for all compounds and used as input to the ConvNets. (C) We implemented an extended version of commonly used architectures (e.g., VGG-19-bn shown in the Figure) by including five additional fully-connected layers to predict pIC$_{50}$ values on a continuous scale. (D) The generalization power of the networks was assessed on the test set, and compared to Random Forest models trained using Morgan fingerprints as covariates.

**Figure 2 Benchmarking the predictive power of ConvNet architectures.** Mean RMSE values (+/- standard deviation) on the test set across ten runs for each of the ConvNet architectures explored in this study (AlexNet$^{54}$, DenseNet-201$^{55}$, ResNet152$^{56}$ and VGG-19$^{57}$). Overall, all architectures enabled the generation of models with high predictive power on the test set, with RMSE values in the 0.65-0.96 pIC$_{50}$ range. However, the extended versions of these architectures that we designed by including 5 fully-connected layers (see Figure 1) constantly led to increased predictive power on the test set.

**Figure 3 Comparing the predictive power of ConvNets and RF.** Mean RMSE values (+/- standard deviation) on the test set across ten runs for (i) the ConvNet showing the highest predictive power for each data set and run combination, (ii) RF models trained on Morgan fingerprints, and (iii) the ensemble models built by averaging the predictions generated with the RF and ConvNet models. Overall, it can be seen that ConvNets lead to comparable predictive power than RF models (or higher in some cases, although the effect size is small and not significant). Ensemble models constantly displayed higher predictive power than either model alone ($P < 10^{-5}$), leading to 4-12% and 5-8% decrease in RMSE values with respect to RF and ConvNet models.

**Figure 4 Predictions for the test set molecules.** Observed against predicted pIC$_{50}$ values for the test set compounds calculated using ConvNets (top panels) or RF models (bottom panels). The results for the ten repetitions are shown (in each repetition the molecules in the training, validation and test sets were different). Overall, both RF models and ConvNets generated comparable error profiles across the entire bioactivity range considered, showing Pearson’s correlation coefficient values in the 0.72-0.84 range.

**Figure 5 RF and ConvNet predictions on the test set.** Correlation between the predictions for the test set compounds calculated with RF models and ConvNets trained on the same training set instances. Overall, the predictions show a positive and significant correlation ($P < 0.05$; Pearson’s correlation coefficient values in the 0.72-0.84 range). The predictions for the ten runs are shown.

**Figure 6 Absolute errors in prediction.** Relationship between the absolute errors in prediction for the same test set instances calculated with ConvNets (x-axis) and RF (y-axis) models.
trained on the same training set instances. The predictions generated by each model differ in >2 pIC$_{50}$ units in some cases, and are moderately correlated ($R^2$ in the 0.58-0.65 range; $P < 0.05$).

**Figure 7** Analysis of the residuals. Residuals for the ConvNets (top panels) and RF (bottom panels) models. Overall, the residuals are centered around zero and show comparable variance across the bioactivity range for both types of models, indicating that compound images permit to model the activity of small molecules across a dynamic range of pIC$_{50}$ values.
References

(1) Basu, A.; Bodycombe, N. E.; Cheah, J. H.; Price, E. V.; Liu, K.; Schaefer, G. I.; Ebright, R. Y.; Stewart, M. L.; Ito, D.; Wang, S.; Bracha, A. L.; Liefeld, T.; Wawer, M.; Gilbert, J. C.; Wilson, A. J.; Stransky, N.; Kryukov, G. V.; Dancik, V.; Barretina, J.; Garraway, L. A.; Hon, C. S.-Y.; Munoz, B.; Bittker, J. A.; Stockwell, B. R.; Khabele, D.; Stern, A. M.; Clemons, P. A.; Shamji, A. F.; Schreiber, S. L. An Interactive Resource to Identify Cancer Genetic and Lineage Dependencies Targeted by Small Molecules. *Cell* 2013, 154, 1151–1161.

(2) Seashore-Ludlow, B.; Rees, M. G.; Cheah, J. H.; Cokol, M.; Price, E. V; Coletti, M. E.; Jones, V.; Bodycombe, N. E.; Soule, C. K.; Gould, J.; Alexander, B.; Li, A.; Montgomery, P.; Wawer, M. J.; Kuru, N.; Kotz, J. D.; Hon, C. S.-Y.; Munoz, B.; Liefeld, T.; Dančík, V.; Bittker, J. A.; Palmer, M.; Bradner, J. E.; Shamji, A. F.; Clemons, P. A.; Schreiber, S. L. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discov.* 2015, 5, 1210–1223.

(3) Shoemaker, R. H. The NCI60 Human Tumour Cell Line Anticancer Drug Screen. *Nat. Rev. Cancer.* 2006, 6, 813–823.

(4) Yang, W.; Soares, J.; Greninger, P.; Edelman, E. J.; Lightfoot, H.; Forbes, S.; Bindal, N.; Beare, D.; Smith, J. A.; Thompson, I. R.; Ramaswamy, S.; Futreal, P. A.; Haber, D. A.; Stratton, M. R.; Benes, C.; McDermott, U.; Garnett, M. J. Genomics of Drug Sensitivity in Cancer (GDSC): A Resource for Therapeutic Biomarker Discovery in Cancer Cells. *Nucleic Acids Res.* 2013, 41, D955-961.

(5) Barretina, J.; Caponigro, G.; Stransky, N.; Venkatesan, K.; Margolin, A. A.; Kim, S.; Wilson, C. J.; Lehár, J.; Kryukov, G. V; Sonkin, D.; Reddy, A.; Liu, M.; ... Garraway, L. A. The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity. *Nature* 2012, 483, 603–607.

(6) Garnett, M. J.; McDermott, U. The Evolving Role of Cancer Cell Line-Based Screens to Define the Impact of Cancer Genomes on Drug Response. *Curr. Opin. Genet. Dev.* 2014, 24, 114–119.

(7) Knouse, K. A.; Lopez, K. E.; Bachofner, M.; Amon, A. Chromosome Segregation Fidelity in Epithelia Requires Tissue Architecture. *Cell* 2018, 175, 200–211.e13.

(8) Cortés-Ciriano, I.; Bender, A. How Consistent Are Publicly Reported Cytotoxicity Data? Large-Scale Statistical Analysis of the Concordance of Public Independent Cytotoxicity Measurements. *ChemMedChem* 2015, 11, 57–71.

(9) Ben-David, U.; Siranosian, B.; Ha, G.; Tang, H.; Oren, Y.; Hinohara, K.; Strathdee, C. A.; Dempster, J.; Lyons, N. J.; Burns, R.; Nag, A.; Kugener, G.; Cimini, B.; Tsvelkov, P.; Maruvka, Y. E.; O'Rourke, R.; Garrity, A.; Tubelli, A.; Bandopadhayay, P.; Tsherniak, A.; Vazquez, F.; Wong, B.; Birger, C.; Ghandi, M.; Thorner, A. R.; Bittker, J. A.; Meyerson, M.; Getz, G.; Beroukhim, R.; Golub, T. R. Genetic and Transcriptional Evolution Alters Cancer Cell Line Drug Response. *Nature* 2018, 560, 325–330.

(10) Iorio, F.; Knijnenburg, T. A.; Vis, D. J.; Bignell, G. R.; Menden, M. P.; Schubert, M.; Aben, N.; Gonçalves, E.; Barthorpe, S.; Lightfoot, H.; Cokelaer, T.; Greninger, P.; van Dyk, E.; Chang, H.; de Silva, H.; Heyn, H.; Deng, X.; Egan, R. K.; Liu, Q.; Mironenko, T.; Mitropoulos, X.; Richardson, L.; Wang, J.; Zhang, T.; Moran, S.; Sayols, S.; Soleimani, M.; Tamborero, D.; Lopez-Bigas, N.; Ross-Macdonald, P.; Esteller, M.; Gray, N. S.; Haber, D. A.; Stratton, M. R.; Benes, C. H.; Wessels, L. F. A.; Saez-Rodriguez, J.; McDermott, U.; Garnett, M. J. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* 2016, 168, 740–754.

(11) Najgebauer, H.; Yang, M.; Francies, H.; Stronach, E. A.; Garnett, M. J.; Saez-Rodriguez, J.; Iorio, F. CELLector: Genomics Guided Selection of Cancer in Vitro Models. *bioRxiv* 2018, 275032.
(12) Gorthi, A.; Romero, J. C.; Loranc, E.; Cao, L.; Lawrence, L. A.; Goodale, E.; Iniguez, A. B.; Bernard, X.; Masamsetti, V. P.; Roston, S.; Lawlor, E. R.; Toretsky, J. A.; Stegmaier, K.; Lessnick, S. L.; Chen, Y.; Bishop, A. J. R. EWS–FLI1 Increases Transcription to Cause R-Loops and Block BRCA1 Repair in Ewing Sarcoma. *Nature* 2018, 555, 387–391.

(13) Menden, M. P.; Casale, F. P.; Stephan, J.; Bignell, G. R.; Iorio, F.; McDermott, U.; Garnett, M. J.; Saez-Rodriguez, J.; Stegle, O. The Germline Genetic Component of Drug Sensitivity in Cancer Cell Lines. *Nat. Commun.* 2018, 9, 3385.

(14) Rees, M. G.; Seashore-Ludlow, B.; Cheah, J. H.; Adams, D. J.; Price, E. V.; Gill, S.; Javaid, S.; Coletti, M. E.; Jones, V. L.; Bodycombe, N. E.; Soule, C. K.; Alexander, B.; Li, A.; Montgomery, P.; Kotz, J. D.; Hon, C. S.-Y.; Munoz, B.; Liefeld, T.; Dancik, V.; Haber, D. A.; Clish, C. B.; Bittker, J. A.; Palmer, M.; Wagner, B. K.; Clemons, P. A.; Shamji, A. F.; Schreiber, S. L. Correlating Chemical Sensitivity and Basal Gene Expression Reveals Mechanism of Action. *Nat Chem Biol* 2016, 12, 109–116.

(15) Cortés-Ciriano, I.; van Westen, G. J. P.; Bouvier, G.; Nilges, M.; Overington, J. P.; Bender, A.; Malliavin, T. E. Improved Large-Scale Prediction of Growth Inhibition Patterns Using the NCI60 Cancer Cell Line Panel. *Bioinformatics* 2016, 32, 85–95.

(16) Cortes-Ciriano, I.; Mervin, L.; Bender, A. Current Trends in Drug Sensitivity Prediction. *Curr. Pharm. Des.* 2017, 22, 6918–6927.

(17) Altman, R. B. Predicting Cancer Drug Response: Advancing the DREAM. *Cancer Discov.* 2015, 5, 237–238.

(18) Menden, M. P.; Iorio, F.; Garnett, M.; McDermott, U.; Benes, C. H.; Ballester, P. J.; Saez-Rodriguez, J. Machine Learning Prediction of Cancer Cell Sensitivity to Drugs Based on Genomic and Chemical Properties. *PLoS One* 2013, 8, e61318.

(19) Geeleher, P.; Cox, N. J.; Huang, R. S. Clinical Drug Response Can Be Predicted Using Baseline Gene Expression Levels and in Vitro Drug Sensitivity in Cell Lines. *Genome Biol.* 2014, 15, R47.

(20) Naulaerts, S.; Dang, C. C.; Ballester, P. J. Precision and Recall Oncology: Combining Multiple Gene Mutations for Improved Identification of Drug-Sensitive Tumours. *Oncotarget* 2017, 8, 97025–97040.

(21) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discov. Today* 2018, 23, 1241–1250.

(22) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction Using Deep Learning. *Front. Environ. Sci.* 2016, 3, 80.

(23) Fernandez, M.; Ban, F.; Woo, G.; Hsing, M.; Yamazaki, T.; LeBlanc, E.; Rennie, P. S.; Welch, W. J.; Cherkasov, A. Toxic Colors: The Use of Deep Learning for Predicting Toxicity of Compounds Merely from Their Graphic Images. *J. Chem. Inf. Model.* 2018, 58, 1533–1543.

(24) Ramsundar, B.; Kearnes, S.; Riley, P.; Webster, D.; Konerding, D.; Pande, V.; Edu, P. Massively Multitask Networks for Drug Discovery. 2015, arXiv1502.02072 arXiv.org ePrint Arch. https://arxiv.org/abs/1502.02072 (accessed Jul 20, 2018).

(25) Dahl, G. E.; Jaitly, N.; Salakhutdinov, R. Multi-Task Neural Networks for QSAR Predictions. 2014, arXiv1406.1231 arXiv.org ePrint Arch. http://arxiv.org/abs/1406.1231 (accessed Jul 19, 2018).

(26) Preuer, K.; Lewis, R. P. I.; Hochreiter, S.; Bender, A.; Bulusu, K. C.; Klambauer, G.; Wren, J. DeepSynergy: Predicting Anti-Cancer Drug Synergy with Deep Learning. *Bioinformatics* 2018, 34, 1538–1546.

(27) Koutsoukas, A.; Monaghan, K. J.; Li, X.; Huan, J. Deep-Learning: Investigating Deep Neural Networks Hyper-Parameters and Comparison of Performance to Shallow Methods for Modeling Bioactivity Data. *J. Cheminform.* 2017, 9, 42.

(28) Altae-Tran, H.; Ramsundar, B.; Pappu, A. S.; Pande, V. Low Data Drug Discovery with
One-Shot Learning. *ACS Cent. Sci.* **2017**, *3*, 283–293.

(29) Subramanian, G.; Ramsundar, B.; Pande, V.; Denny, R. A. Computational Modeling of β-Secretase 1 (BACE-1) Inhibitors Using Ligand Based Approaches. *J. Chem. Inf. Model.* **2016**, *56*, 1936–1949.

(30) Korotcov, A.; Tkachenko, V.; Russo, D. P.; Ekins, S. Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. *Mol. Pharm.* **2017**, *14*, 4462–4475.

(31) Putin, E.; Asadulaev, A.; Vanhaelen, Q.; Ivanenko, Y.; Aladinskaya, A. V.; Aliper, A.; Zhavoronkov, A. Adversarial Threshold Neural Computer for Molecular de Novo Design. *Mol. Pharm.* **2018**, *15*, 4386–4397.

(32) Olivecrona, M.; Blaschke, T.; Engkvist, O.; Chen, H. Molecular De-Novo Design through Deep Reinforcement Learning. *J. Cheminform.* **2017**, *9*, 48.

(33) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268–276.

(34) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4*, 120–131.

(35) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships. *J. Chem. Inf. Model.* **2015**, *55*, 263–274.

(36) Lecun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature*. Nature Publishing Group May 28, 2015, pp 436–444.

(37) Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, *86*, 2278–2324.

(38) Le Cun, Y.; Boser, B.; Denker, J. S.; Henderson, D.; Howard, R. E.; Hubbard, W.; Jackel, L. D. Handwritten Digit Recognition with a Back-Propagation Network. *Proceedings of the 2nd International Conference on Neural Information Processing Systems*. MIT Press 1989, pp 396–404.

(39) Wainberg, M.; Merico, D.; Delong, A.; Frey, B. J. Deep Learning in Biomedicine. *Nat. Biotechnol.* **2018**, *36*, 829–838.

(40) Cooper, L. A.; Demicco, E. G.; Saltz, J. H.; Powell, R. T.; Rao, A.; Lazar, A. J. PanCancer Insights from The Cancer Genome Atlas: The Pathologist’s Perspective. *J. Pathol.* **2018**, *244*, 512–524.

(41) Coudray, N.; Ocampo, P. S.; Sakellaropoulos, T.; Narula, N.; Snuderl, M.; Fenyö, D.; Moreira, A. L.; Razavian, N.; Tsirigos, A. Classification and Mutation Prediction from Non–Small Cell Lung Cancer Histopathology Images Using Deep Learning. *Nat. Med.* **2018**, *1*, .

(42) Yu, K.-H.; Zhang, C.; Berry, G. J.; Altman, R. B.; Ré, C.; Rubin, D. L.; Snyder, M. Predicting Non-Small Cell Lung Cancer Prognosis by Fully Automated Microscopic Pathology Image Features. *Nat. Commun.* **2016**, *7*, 12474.

(43) Yu, K.-H.; Beam, A. L.; Kohane, I. S. Artificial Intelligence in Healthcare. *Nat. Biomed. Eng.* **2018**, *2*, 719–731.

(44) Poplin, R.; Chang, P.-C.; Alexander, D.; Schwartz, S.; Colthurst, T.; Ku, A.; Newburger, D.; Dijamco, J.; Nguyen, N.; Afshar, P. T.; Gross, S. S.; Dorfman, L.; McLean, C. Y.; DePristo, M. A. A Universal SNP and Small-Indel Variant Caller Using Deep Neural Networks. *Nat. Biotechnol.* **2018**, *36*, 983.

(45) Scheeder, C.; Heigwer, F.; Boutros, M. Machine Learning and Image-Based Profiling in Drug Discovery. *Curr. Opin. Syst. Biol.* **2018**, *10*, 43–52.

(46) Kraus, O. Z.; Ba, J. L.; Frey, B. J. Classifying and Segmenting Microscopy Images with Deep Multiple Instance Learning. *Bioinformatics* **2016**, *32*, i52–i59.
(47) Goh, G. B.; Siegel, C.; Vishnu, A.; Hodas, N. O. Using Rule-Based Labels for Weak Supervised Learning: A ChemNet for Transferable Chemical Property Prediction. 2017, arXiv1712.02734 arXiv.org ePrint Arch. https://arxiv.org/abs/1712.02734 (accessed Jul 8, 2018).

(48) Goh, G. B.; Siegel, C.; Vishnu, A.; Hodas, N.; Baker, N. How Much Chemistry Does a Deep Neural Network Need to Know to Make Accurate Predictions? In Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018; 2018; vol. 2018-Janua, pp 1340–1349.

(49) Simm, J.; Klambauer, G.; Arany, A.; Steijaert, M.; Wegner, J. K.; Gustin, E.; Chupakhin, V.; Chong, Y. T.; Vialard, J.; Buijningers, P.; Velter, I.; Vapirev, A.; Singh, S.; Carpenter, A. E.; Wuys, R.; Hochreiter, S.; Moreau, Y.; Ceulemans, H. Repurposing High-Throughput Image Assays Enables Biological Activity Prediction for Drug Discovery. Cell Chem. Biol. 2018, 25, 611–618.e3.

(50) Duvenaud, D.; Maclaurin, D.; Aguilera-Iparraguirre, J.; Gómez-Bombarelli, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints.

(51) Goh, G. B.; Siegel, C.; Vishnu, A.; Hodas, N. O.; Baker, N. Chemception: A Deep Neural Network with Minimal Chemistry Knowledge Matches the Performance of Expert-Developed QSAR/QSPR Models. 2017, arXiv1706.06689 arXiv.org ePrint Arch. https://arxiv.org/abs/1706.06689 (accessed Jul 8, 2018).

(52) Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. 2016.

(53) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: A Benchmark for Molecular Machine Learning. 2017.

(54) Krizhevsky, A.; Krizhevsky, A.; Sutskever, I.; Hinton, G. E. Imagenet Classification with Deep Convolutional Neural Networks. Adv. Neural Inf. Process. Syst. 2012.

(55) Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K. Q. Densely Connected Convolutional Networks. 2016.

(56) He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. 2015.

(57) Simonyan, K.; Andrew Zisserman; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. ICLR 2015, 1–14.

(58) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. J. Chem. Inf. Model. 2010, 50, 742–754.

(59) Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. J. Chem. Doc. 1965, 5, 107–113.

(60) Nowotka, M.; Papadatos, G.; Davies, M.; Dedman, N.; Hersey, A. Want Drugs? Use Python. 2016, arXiv1607.00378 arXiv.org ePrint Arch. https://arxiv.org/abs/1607.00378 (accessed Jul 10, 2018).

(61) Davies, M.; Nowotka, M.; Papadatos, G.; Dedman, N.; Gaulton, A.; Atkinson, F.; Bellis, L.; Overington, J. P. ChEMBL Web Services: Streamlining Access to Drug Discovery Data and Utilities. Nucleic Acids Res. 2015, 43, W612-20.

(62) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. Nucleic Acids Res. 2011, 40, 1100–1107.

(63) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? J. Med. Chem. 2014, 57, 4977–5010.

(64) O’Boyle, N. M.; Sayle, R. A. Comparing Structural Fingerprints Using a Literature-Based Similarity Benchmark. J. Cheminform. 2016, 8, 36.
(65) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but Verify: On the Importance of Chemical Structure Curation in Cheminformatics and QSAR Modeling Research. *J. Chem. Inf. Model.* **2010**, *50*, 1189–1204.

(66) Landrum, G. RDKit: Open-Source Cheminformatics. [https://www.rdkit.org/](https://www.rdkit.org/) (accessed Jan 12, 2017).

(67) Deng, J.; Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Fei-Fei, L. Imagenet: A Large-Scale Hierarchical Image Database. *CVPR* **2009**.

(68) Paszke, A.; Chanan, G.; Lin, Z.; Gross, S.; Yang, E.; Antiga, L.; Devito, Z. Automatic Differentiation in PyTorch. In *Advances in Neural Information Processing Systems 30*; 2017; pp 1–4.

(69) Sutskever, I.; Martens, J.; Dahl, G.; Hinton, G. On the Importance of Initialization and Momentum in Deep Learning. In *Proceedings of the 30th International Conference on Machine Learning, PMLR 28*; 2013; pp 1139–1147.

(70) Srivastava, N.; Hinton, G.; Krizhevsky, A.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.

(71) Lenselink, E. B.; ten Dijke, N.; Bongers, B.; Papadatos, G.; van Vlijmen, H. W. T.; Kowalczyk, W.; IJzerman, A. P.; van Westen, G. J. P. Beyond the Hype: Deep Neural Networks Outperform Established Methods Using a ChEMBL Bioactivity Benchmark Set. *J. Cheminform.* **2017**, *9*, 45.

(72) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(73) Sheridan, R. P. Three Useful Dimensions for Domain Applicability in QSAR Models Using Random Forest. *J. Chem. Inf. Model.* **2012**, *52*, 814–823.

(74) Winer, B.; Brown, D.; Michels, K. *Statistical Principles in Experimental Design*, 3rd ed.; Psychology, M.-H. series in, Ed.; McGraw-Hill: New York, NY, USA, 1991.

(75) Roy, K.; Ambure, P.; Aher, R. B. How Important Is to Detect Systematic Error in Predictions and Understand Statistical Applicability Domain of QSAR Models? *Chemom. Intell. Lab. Syst.* **2017**, *162*, 44–54.

(76) Roy, K.; Das, R. N.; Ambure, P.; Aher, R. B. Be Aware of Error Measures. Further Studies on Validation of Predictive QSAR Models. *Chemom. Intell. Lab. Syst.* **2016**, *152*, 18–33.

(77) Ammad-ud-din, M.; Georgii, E.; Gönen, M.; Laitinen, T.; Kallioniemi, O.; Wennerberg, K.; Poso, A.; Kaski, S. Integrative and Personalized QSAR Analysis in Cancer by Kernelized Bayesian Matrix Factorization. *J. Chem. Inf. Model.* **2014**, *54*, 2347–2359.

(78) Cortes-Ciriano, I.; Bender, A. Improved Chemical Structure–Activity Modeling Through Data Augmentation. *J. Chem. Inf. Model.* **2015**, *55*, 2682–2692.

(79) Polishchuk, P. G.; Muratov, E. N.; Artemenko, A. G.; Kolumbin, O. G.; Muratov, N. N.; Kuz’min, V. E. Application of Random Forest Approach to QSAR Prediction of Aquatic Toxicity. *J. Chem. Inf. Model.* **2009**, *49*, 2481–2488.

(80) Marchese Robinson, R. L.; Palczewska, A.; Palczewski, J.; Kidley, N. Comparison of the Predictive Performance and Interpretability of Random Forest and Linear Models on Benchmark Data Sets. *J. Chem. Inf. Model.* **2017**, *57*, 1773–1792.

(81) Koutsoukas, A.; Paricharak, S.; Galloway, W. R. J. D.; Spring, D. R.; IJzerman, A. P.; Glen, R. C.; Marcus, D.; Bender, A. How Diverse Are Diversity Assessment Methods? A Comparative Analysis and Benchmarking of Molecular Descriptor Space. *J. Chem. Inf. Model.* **2013**, *54*, 230–242.

(82) Cortes-Ciriano, I.; Murrell, D. S.; van Westen, G. J. P.; Bender, A.; Malliavin, T. Prediction of the Potency of Mammalian Cyclooxygenase Inhibitors with Ensemble Proteochemometric Modeling. *J. Cheminf.* **2014**, *7*, 1.
(83) Cortes-Ciriano, I.; Bender, A. Deep Confidence: A Computationally Efficient Framework for Calculating Reliable Errors for Deep Neural Networks. *J. Chem. Inf. Model.* **2018**, *In press*, 1809.09060.

(84) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; 2008.

(85) Wallach, I.; Dzamba, M.; Heifets, A. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-Based Drug Discovery. 2015, *arXiv1510.02855 arXiv.org ePrint Arch.* [https://arxiv.org/abs/1510.02855 (accessed Jul 8, 2018)].

(86) Amidi, A.; Amidi, S.; Vlachakis, D.; Megalooikonomou, V.; Paragios, N.; Zacharakis, E. I. EnzyNet: Enzyme Classification Using 3D Convolutional Neural Networks on Spatial Representation. *PeerJ* **2018**, 6, e4750.

(87) Derevyanko, G.; Grudinin, S.; Bengio, Y.; Lamoureux, G. Deep Convolutional Networks for Quality Assessment of Protein Folds. *Bioinformatics* **2018**.

(88) Torng, W.; Altman, R. B. 3D Deep Convolutional Neural Networks for Amino Acid Environment Similarity Analysis. *BMC Bioinformatics* **2017**, 18, 302.

(89) Bansal, M.; Yang, J.; Karan, C.; Menden, M. P.; Costello, J. C.; Tang, H.; Xiao, G.; Li, Y.; Allen, J.; Zhong, R.; Chen, B.; Kim, M.; Wang, T.; Heiser, L. M.; Realubit, R.; Mattioli, M.; Alvarez, M. J.; Shen, Y.; Community, N.-D.; Gallahan, D.; Singer, D.; Saez-Rodriguez, J.; Xie, Y.; Stolovitzky, G.; Califano, A. A Community Computational Challenge to Predict the Activity of Pairs of Compounds. *Nat Biotech* **2014**, 32, 1213–1222.
Figure 1

A Data Sets

B Image generation

C Deep learning

D Evaluation of predictive power

ChEMBL

8 data sets

70% training

15% validation

15% testing

Image dimensions: 224x224x3

Adapted VGG 19 network architecture

Feature extraction layers

Regression layers

In vitro cytotoxicity

Observed pIC50

Predicted pIC50

Kekulé representations

3x3 Convolutional + batch norm. + ReLU

Maxpool

Fully-connected layers

25088

4096

1000

200

100

3x3 Convolutional + batch norm. + ReLU

Maxpool

3x3 Convolutional + batch norm. + ReLU

Maxpool

Fully-connected layers

25088

4096

1000

200

100
Figure 3

RMSE test set (pIC50)

Data sets

A2780  CCRF−CEM  DU−145  HCT−15  KB  LoVo  PC−3  SK−OV−3

ConvNet  RF  Ensemble
Figure 4
Figure 5

![Scatter plots showing predictions for different cell lines.](image)
Figure 6
Figure 7