Article:
Marchese Robinson, RL, Palczewska, A, Palczewski, JA orcid.org/0000-0003-0235-8746 et al. (1 more author) (2017) Comparison of the Predictive Performance and Interpretability of Random Forest and Linear Models on Benchmark Datasets. Journal of Chemical Information and Modeling, 57 (8). pp. 1773-1792. ISSN 1549-9596

https://doi.org/10.1021/acs.jcim.6b00753

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

Comparison of the Predictive Performance and Interpretability of Random Forest and Linear Models on Benchmark Datasets

Richard Liam Marchese Robinson, Anna Palczewska, Jan Palczewski, and Nathan Kidley

J. Chem. Inf. Model., **Just Accepted Manuscript** • DOI: 10.1021/acs.jcim.6b00753 • Publication Date (Web): 17 Jul 2017

Just Accepted

ACS Publications

Journal of Chemical Information and Modeling is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

Comparison of the Predictive Performance and Interpretability of Random Forest and Linear Models on Benchmark Datasets

Richard L. Marchese Robinson,[a,b,c,*] *Anna Palczewska,*[d,e] *Jan Palczewski;*[f] *Nathan Kidley*[a]

a. Syngenta Ltd, Jealott's Hill International Research Centre, Bracknell, Berkshire, RG42 6EY, United Kingdom

b. School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, James Parsons Building, Byrom Street, Liverpool, L3 3AF, United Kingdom

c. Present Address: School of Chemical and Process Engineering, University of Leeds, Leeds LS2 9JT, United Kingdom

d. Department of Computing, University of Bradford, BD7 1DP Bradford, United Kingdom

e. Present Address: School of Geography, Leeds Institute for Data Analytics, University of Leeds, Leeds LS2 9JT, United Kingdom

f. School of Mathematics, University of Leeds, Leeds LS2 9JT, United Kingdom

*Corresponding author. E-mail: R.L.MarcheseRobinson@leeds.ac.uk

KEYWORDS

quantitative structure-activity relationships; model interpretation; Machine Learning; Heat Map; Random Forest; Partial Least Squares; Support Vector Machines; Support Vector Regression

ABSTRACT

The ability to interpret the predictions made by quantitative structure activity relationships (QSARs) offers a number of advantages. Whilst QSARs built using non-linear modelling approaches, such as the popular Random Forest algorithm, might sometimes be more predictive than those built using linear modelling approaches, their predictions have been perceived as difficult to interpret. However, a growing number of approaches have been proposed for interpreting non-linear QSAR models in general and Random Forest in particular. In the current work, we compare the performance of Random Forest to two widely used linear modelling approaches: linear Support Vector Machines (SVM), or Support Vector Regression (SVR), and Partial Least Squares (PLS). We compare their performance in terms of their predictivity as well as the chemical interpretability of the predictions, using novel scoring schemes for assessing Heat Map images of substructural contributions. We critically assess different approaches to interpreting Random Forest models as well as for obtaining predictions from the forest. We assess the models on a large number of widely employed, public domain benchmark datasets corresponding to regression and binary classification problems of relevance to hit identification and toxicology. We conclude that Random Forest typically yields comparable or possibly better predictive performance than the linear modelling approaches and that its predictions may also be interpreted in a chemically and biologically meaningful way. In contrast to earlier work looking at interpreting non-linear QSAR models, we directly compare two methodologically distinct approaches for interpreting Random Forest models. The approaches for interpreting Random

Forest assessed in our article were implemented using Open Source programs, which we have made available to the community. These programs are the *rfFC* package [https://r-forge.r-project.org/R/?group_id=1725] for the R Statistical Programming Language, along with a Python program *HeatMapWrapper* [https://doi.org/10.5281/zenodo.495163] for Heat Map generation.

INTRODUCTION

The ability to interpret the predictions made by quantitative structure-activity relationships (QSARs), in terms of the size and sign of the influence of the underlying descriptor values, is valuable for a number of reasons. Mechanistic interpretation is viewed as particularly important in a regulatory context[1] and could yield novel mechanistic insight or consistency with established understanding could complement statistical validation, by conferring greater confidence in the predictions.[2] Furthermore, interpretable predictions could guide structural modifications required for desired changes in molecular properties.[2] Analysis of descriptor contributions may also help to rationalize outliers.[3]

Traditionally, QSARs were linear models.[2,4] For linear models, both the relative size and sign of the influence of an individual descriptor on the predictions is straightforward to obtain in terms of the corresponding coefficient.[4,5] The product of the coefficient and descriptor value for a given chemical will yield the component of an individual prediction arising from a particular descriptor.[6] In principle, although complications may arise in practice,[6,7] this makes the predictions obtained from linear QSARs relatively straightforward to interpret.

A number of algorithms for building linear QSARs, either for prediction of continuous (regression models) or categorical (classification models) measures of biological activities, are

used by modern QSAR practitioners. These include partial least squares (PLS),[8] for regression or classification[9] as well as, when used with a linear kernel function, Support Vector Machines (SVMs)[10,11] and Support Vector Regression (SVR)[12] for classification[13] and regression[14] respectively.

However, in recent decades, QSAR models have been increasingly developed using non-linear modelling approaches.[2,15] Interpreting the predictions made by these models is typically harder than for linear models. Indeed, it has been argued that there is an inherent trade-off between predictive performance and interpretability, with non-linear models suggested to generally be more predictive but less interpretable.[16] Nonetheless, in addition to approaches for estimating the overall "importance" of various descriptors in a QSAR model,[16] a growing number of studies has advocated approaches which evaluate how specific molecular characteristics contribute towards individual predictions made by a non-linear QSAR model.[2,17–25]

One non-linear modelling approach which is popular within the QSAR community[15,16] is Random Forest.[9,26] Approaches for estimating descriptor "importance" for Random Forest models are well established.[9,26–28] However, more recently, proposals for estimating the relative magnitude and sign of the influence of a given descriptor on a given prediction, as is the focus of our article, have been presented,[2,18,19,23,29–31] including those which are specifically designed for Random Forest models.[19,23,29–31]

In our current article, we consider QSAR models developed using Random Forest, PLS and SVM (or Support Vector Regression), using a linear kernel, for binary classification and regression modelling tasks which are relevant for hit identification or predictive toxicology. We compare the predictive performance of these algorithms, as well as the extent to which their

predictions for individual molecules can be interpreted in a chemically and biologically meaningful manner according to various established approaches for model interpretation, using well-known, publicly available benchmark datasets. As well as evaluating different algorithms for interpreting individual predictions, we further investigate the degree to which obtaining a chemically and biologically meaningful interpretation depends upon the quality of the corresponding prediction for a given molecule.

Specifically, we compare two different approaches for interpreting the same Random Forest prediction. The first approach[19,23,29–32] was originally proposed for Random Forest regression models[19] and was then extended to binary classification models.[29,30] The second approach is based on estimating local gradients (i.e. vectors of partial derivatives) of the prediction with respect to the descriptors and is analogous to algorithms recently used to interpret various kinds of non-linear QSAR models.[2,18,20,33] We evaluate these approaches via means of novel scoring schemes for assessing Heat Map images representing the influence of molecular substructures on model predictions. We extend the Heat Map approach introduced by Rosenbaum et al.[13] to work with any linear or non-linear classification or regression model, built using the same descriptors as per their earlier work, and present an Open Source implementation of our extension. Our extension of their approach includes novel "symmetrized" normalization schemes for translating descriptor contributions associated with substructural features into Heat Map molecular images. We present the first evaluation of our approach for interpreting binary classification Random Forest models[29,30] in cheminformatics, using suitable benchmark classification datasets. To the best of our knowledge, we present the first comparative assessment of methodologically distinct approaches to interpreting predictions obtained from a non-linear QSAR model.

METHODS AND DATA

Modelling approaches

More expansive descriptions of these modelling algorithms and, for those algorithms yielding linear models, the equations generating the predictions, are provided in the Supporting Information.

Partial least squares (PLS) regression[8,34] develops a linear regression model based upon so-called "latent variables", which correspond to a selected set of orthogonal, linear combinations of the original descriptors. A simple modification can transform PLS regression into a binary classifier: the two class labels are mapped onto two real numbers (1 and 0), PLS regression is used to build a model from the training set and the numeric predictions (or score) made by the regression model are converted into predicted class labels via comparison to some suitable threshold.[9] This so-called "probit" treatment yields *Probit-PLS*.[35] Note that this model differs from a classical probit regression, for which a random outcome is observed with a probability described by the model.[36]

Support Vector Machines (SVMs)[10,11] build a binary classification model with a functional form which is contingent upon the selected "kernel function". *Support Vector Regression (SVR)*[12,14] is an analogous approach to SVM. Using a linear "kernel function"[11,13] yields linear SVM and SVR models.

Random Forest classification[9,26] builds classification models, meaning binary classification models in our current article, via generating a forest of decision trees using the training data. Each tree is grown using independent random samples of the training data. Whilst "bootstrap" samples are commonly employed,[9,15,26] we employed sampling without replacement. This was

required in order to be able to calculate descriptor contributions via the "Kuz'min/Palczewska" approach (see below) using the *rfFC* software package.[37]

By default, binary classification predictions are made for a new molecule by passing it down each of the decision trees in the forest, each tree calculating a score corresponding to the fraction of molecules belonging to "class 1" in the terminal node to which the new molecule is assigned. If that fraction is greater than 0.5, i.e. if a majority of molecules belong to "class 1", the tree predicts the new molecule to belong to "class 1". The forest then returns an overall score corresponding to the proportion of trees predicting the molecule to belong to "class 1". If this overall score is greater (not greater) than the threshold of 0.5, i.e. if a majority of trees (do not) vote for "class 1", the predicted class is "class 1" ("class 0"). We also considered a non-default "averaged predictions" approach, which is analogous to Random Forest regression, where the overall score is calculated as the arithmetic mean of the scores generated by each tree and classification is based on the same threshold of 0.5.

Random Forest regression[9,26] model development is virtually identical to Random Forest classification. The bioactivity predicted for a molecule is the arithmetic mean of the bioactivity of all molecules assigned to the relevant terminal node, for a given tree, averaged across all trees in the forest.

Hyperparameters selection

A variety of "hyperparameters" can affect the performance of the modelling algorithms used for our current article. In addition, the degree of random "downsampling",[38] to address class imbalance, may be considered another hyperparameter for the binary classification datasets. Hence, a variety of hyperparameter combinations were evaluated, which are summarized in

Supporting Information Table S1. The best hyperparameter options were selected via iterative application of stratified, 2-fold cross-validation on the training set.[39] Different combinations of hyperparameters were evaluated in terms of the cross-validated Matthews Correlation Coefficient (*MCC*)[40,41] or the Coefficient of Determination (R^2)[42,43] for a binary classification or regression modelling approach respectively. Further details are provided in the Supporting Information.

Descriptors employed

The descriptors employed in our current work correspond to a set of "features" representing molecular substructures. The corresponding descriptor takes the value 1, or 0, if a feature encoded substructure is present, or not present, in a particular molecule. Specifically, the widely used extended connectivity fingerprint (ECFP)[44] features were employed. Here, we calculated these features as per Rosenbaum et al.[13] In brief, for each molecule, features are grown to represent circular substructures centered on each of the heavy atoms in the molecule. For each central atom, an initial identifier was assigned based on the Daylight invariant information (including atomic number and number of connections)[45] and ring type information. The assignment of this initial identifier is considered iteration number 0. In subsequent iterations, this initial identifier is supplemented with the identifiers of attached heavy atoms. Features are generated at each iteration by hashing the combinations of identifiers. The unique features, corresponding to molecular substructures of increasing size, obtained at each iteration are saved. This procedure stops when the maximum number of iterations is reached, which is controlled by the depth of the fingerprint: the longest path between atoms in the fingerprint. As per Rosenbaum et al., we used a depth of 4.[13]

Feature selection

Prior to selecting the most appropriate hyperparameters and building a final model on a given training set, the training set was used to select the most biologically relevant 200 features. Further details are provided in the Supporting Information.

Benchmark datasets

The datasets used in the current work were the protein binding regression datasets of Sutherland et al.,[46] along with several binary classification datasets containing molecules assessed with respect to various endpoints: protein binding (MUV846, MUV548, MUV713 datasets),[47] Ames test (Kazius dataset)[48] and Chromosome aberration test (CA dataset)[17] toxicity tests. (A summary of these datasets is provided in **Table 1**.) Whilst possible limitations of the Kazius[48] dataset have been noted and expanded Ames datasets described in the literature,[22,49] the Kazius,[48] Sutherland[46] and MUV[47] datasets are accepted QSAR benchmarks[13,50–53] and the Kazius, MUV and CA datasets have previously been used to illustrate approaches for interpreting QSAR predictions.[13,17]

Table 1. Benchmark datasets

Type	Dataset Name	Endpoint	Size[a]	Unique ECFP features before feature selection	Reference
Binary classification	MUV846	Coagulation factor Xia (protease) inhibition	15030 (30 actives, 15000 decoys)	175146	Rohrer and Baumann[47]
	MUV548	Protein kinase A	15030 (30 actives,	154311	

		inhibition	15000 decoys)		
	MUV713	estrogen receptor alpha coactivator binding inhibition	15030 (30 actives, 15000 decoys)	150857	
	Kazius	Mutagenicity (Ames test)	4337 (2401 mutagens, 1936 non-mutagens)	46282	Kazius et al.[48]
	CA	Chromosome aberration test	939 (351 positive, 588 negative)	15846	Mohr et al.[17]
Regression	DHFR	dihydrofolate reductase inhibition (pIC_{50} values)	397	4792	Sutherland et al.[46]
	COX2	cyclooxygenase-2 inhibition (pIC_{50} values)	322	3434	
	BZR	benzodiazepine receptor activity (pIC_{50} values)	163	2755	
	AchE	acetylcholinesterase inhibition (pIC_{50} values)	111	1587	
	ACE	angiotensin converting enzyme inhibition	114	1637	Sutherland et al. & Priest et

		(pIC_{50} values)			al.[46,54]
	THR	thrombin inhibition (pK_i values)	88	1084	Sutherland et al. & Böhm et al.[46,55]
	THERM	thermolysin inhibition (pK_i values)	76	1017	Sutherland et al. & Gohlke et al.[46,56]
	GPB	glycogen phosphorylase b inhibition (pK_i values)	66	817	

a. One CA dataset[17] entry ("CA847") was automatically removed during the preparation of datasets for calculations.

For all MUV datasets, the actives were treated as "class 1" and the decoys were "class 0". For the Kazius dataset, mutagens were "class 1" and non-mutagens were "class 0". For the CA dataset, positives, i.e. compounds determined to cause chromosome aberrations based on the chromosome aberration test,[17] were "class 1" and negatives were "class 0". Not all the MUV datasets[47] were modelled, due to the considerable computational overhead.

Preparation of datasets for calculations

All datasets were initially retrieved as SDF files and the structures processed using Pipeline Pilot[57] prior to calculating ECFP features. This was designed to generate neutral, single molecule representations, in standard tautomeric forms, with consistently represented functional groups

including explicit hydrogens, as well as remove any obviously erroneous structures. Further details are provided in the Supporting Information.

Statistical evaluation of predictive performance

The performance of all modelling approaches was assessed via *R* repetitions of *K*-fold "external cross-validation".[58] Feature selection was carried out, independently, using the entirety of each cross-validation training set and hyperparameters selection was carried out using "internal"[58] cross-validation on those training sets. This avoided optimistically biasing the results obtained on the "external cross-validation" test sets due to their having been used for model optimization.

For all datasets except the CA dataset, for which the single set of 10 folds defined by Mohr et al. was used,[17] the partitioning into folds was carried out randomly and independently for each repetition, with stratified sampling employed for the classification datasets. Two repetitions of five-fold cross-validation were used for the Kazius[48] and MUV[47] datasets, as per Rosenbaum et al.,[13] whilst the performance of the regression approaches on the Sutherland datasets[46] was evaluated using 20 repetitions of 10-fold cross-validation as per Hinselmann et al.[50]

Finally, it should be recalled that all model development protocols were dependent upon random selections. In the case of Random Forest models, this is intrinsic to the algorithm. In the case of binary classification models, this is true due to the random selection of subsets of the majority class when downsampling. In the case of all models, except for Random Forest regression models, this is further due to the random selection of internal cross-validation folds used to guide hyperparameter selection (see "Hyperparameters selection"). Hence, as well as averaging test set results across multiple repetitions of K-fold "external cross-validation", performance statistics

were averaged across five repetitions of the entire modelling and evaluation procedure: each repetition employed a different random number generator (RNG) seed for all random operations.

For binary classification datasets, the overall predictive performance of the models was evaluated in terms of Matthew's correlation coefficient (*MCC*),[40,41] Cohen's Kappa (*Kappa*),[35,59] the area under the ROC curve (*AUC*),[60,61] and balanced accuracy (*BA*).[62] For regression datasets, the overall predictive performance of the models was evaluated in terms of the coefficient of determination (R^2), Pearson's correlation coefficient (r) and root mean squared error (*RMSE*).[8,40,42,43] These were calculated, and may be interpreted, as explained in the Supporting Information.

Descriptor contributions

By "descriptor contributions", we mean a vector of estimates of signed influences of each descriptor value upon the prediction made by the model for a given molecule. Here, we first provide an overview of the different kinds of descriptor contributions which were calculated, and their interpretation, followed by a detailed explanation of how these different kinds of descriptor contributions were calculated.

For the linear modelling approaches (PLS regression, Probit-PLS, Support Vector Machines and Support Vector Regression with linear kernel functions), the coefficients were treated as descriptor contributions. For both Random Forest regression and Random Forest classification, two different kinds of methodology were employed to estimate descriptor contributions. The first approach was developed by Kuz'min et al.[19] for Random Forest regression and extended by Palczewska et al. to binary classification.[29,30] The second approach was to estimate the local gradient, i.e. a vector of partial derivatives of the predicted value with respect to each of the

descriptors. The local gradients approach is analogous to the approaches presented in a number of recent publications for interpreting non-linear QSARs.[2,18,20,33] N.B. For brevity, we refer to the first and second approach as the "Kuz'min/Palczewska" approach and the "local gradients" approach respectively.

Whilst this would not generally be the case, all analyzed descriptor contributions presented in our current work are estimates of the extent to which the corresponding molecular substructures promoted biological activity. This is because we employed binary descriptors for which the descriptors took values of 1 or 0, denoting the presence or absence of a corresponding substructure, and we only analyzed the descriptor contributions for molecules where the corresponding substructures were present. (See sections "Descriptors employed" and "Heat Map interpretation of individual predictions" for details.) However, for the linear models, the significance of given substructures in specific, local contexts (e.g. specific chemical classes), as opposed to their global significance, is not estimated.

For Random Forest classification, only the descriptor contributions calculated based on the "local gradients approach" were dependent upon the whether the default (majority voting) or "averaged predictions" approach was employed. However, as explained under "Kuz'min/Palczewska approach" below, only the predictions made using the "averaged predictions" approach can be guaranteed to be consistent with the descriptor contributions calculated according to the "Kuz'min/Palczewska" approach. Hence, we only consider the "averaged predictions" when evaluating the interpretability of Random Forest classification predictions based on the "Kuz'min/Palczewska" approach.

Linear coefficients approach

The output generated for all linear modelling approaches can be expressed using a common functional form. This allows coefficients to be obtained for the descriptors. The output generated by each model, for any given molecule, is either the predicted bioactivity, for regression models, or the score used, in combination with a threshold, to make predictions for classification models. The common functional form is presented in equation (1): o_i is the model output for the *i*th molecule, d_{ji} is the value of the *j*th descriptor (out of *J* descriptors in total) for the *i*th molecule, c_j is the coefficient for the *j*th descriptor and *b* is an offset constant. N.B. Consideration of Supporting Information equations (i)-(vi) demonstrates that the values of c_j in equation (1) are calculated from different model-specific parameters.

$$o_i = \sum_j^J c_j d_{ji} + b \quad (1)$$

In the current work, these coefficients (c_j) were treated as "descriptor contributions" for all linear models. In general, treating the coefficients as "descriptor contributions" does not quantify the total component of a given prediction arising from a given descriptor, which is given by the product of the descriptor and coefficient values (i.e. $c_j d_{ji}$).[6] Nonetheless, in the current work, the "descriptor contributions" analyzed for linear models were directly equivalent to the components of a given prediction arising from individual descriptors. (This is because, as described under "Descriptors employed", we employed binary descriptors, for which d_{ji} = 1 or 0, and we only analyzed the descriptor contributions where d_{ji} = 1 – corresponding to the presence of molecular substructures, as explained under "Heat Map interpretation of individual predictions".) Specifically, they corresponded to an estimate of the extent to which the corresponding molecular substructures promoted biological activity. Since the coefficients calculated for linear

models are constant for all molecules, the significance of given substructures in specific, local contexts (e.g. specific chemical classes), as opposed to their global significance, is not estimated.

Kuz'min/Palczewska approach

This approach starts with a trained Random Forest model for which the training set molecules, and their measured bioactivities, assigned to each node are known. Subsequent to treating the class labels as numeric bioactivity values (i.e. "class 1" is treated as 1 and "class 0" is treated as 0), the approaches considered in our current article are identical for both binary classification and regression i.e. binary classification "descriptor contributions" are calculated towards "class 1", such that negative values indicate contributions towards "class 0". The steps required to obtain the descriptor contribution for a single descriptor and for a single molecule (*X*) are as follows.

1. For a single tree in the forest, start at the "root node" corresponding to the complete "local training set": this "root node" is the first "parent node".

2. Obtain the arithmetic mean bioactivity for all training set molecules in the current "parent node". For binary classification, this is equivalent to calculating the fraction of "local training set" molecules belonging to "class 1" (see the preceding discussion).

3. Go to the "child node" into which the molecule *X* is directed based on its value for the descriptor used to partition the "parent node" compared to the "split value" (see "Modelling approaches").

4. *If* the descriptor used to split the "parent node" corresponds to the descriptor of interest, obtain the corresponding "local increment", as per equation (2). In equation (2), $LI_j^{C,T}$ denotes the "local increment" associated with the *j*th descriptor calculated for the *relevant*

"child node" *C* in tree *T*; $Y_{mean}^{C,T}$ ($Y_{mean}^{P,T}$) corresponds to the arithmetic mean bioactivity of the "local training set" molecules assigned to the "child node" *C* ("parent node" *P*) in tree *T*.

$$LI_j^{C,T} = Y_{mean}^{C,T} - Y_{mean}^{P,T} \quad (2)$$

5. Repeat steps [2]-[4], with each "child node" from the previous iteration becoming the "parent node" in the next iteration, until a "terminal node" is reached.

6. Repeat steps [2]-[5] for all trees in the forest.

7. Calculate the "descriptor contribution" from the "local increments" as per equation (3): dc_j^X denotes the descriptor contribution obtained for the *j*th descriptor for molecule *X*; *T* denotes a particular tree in the forest; n_{tree} denotes the total number of trees in the forest; C denotes a particular "child node" in a given tree. N.B. The set of "child nodes" for which "local increments" ($LI_j^{C,T}$) are summed in equation (3) will be, as explained for step [4], contingent upon the specific molecule and descriptor for which a "descriptor contribution" is calculated.

$$dc_j^X = \frac{1}{n_{tree}} \sum_T^{n_{tree}} \sum_C LI_j^{C,T} \quad (3)$$

It should be noted that these "descriptor contributions" are equivalent to the descriptor specific components of the prediction generated by a Random Forest regression model. This is also true for the score for "class 1" generated by a binary classification model under either of the following scenarios. The first scenario is that the "averaged predictions" predictions approach (see "Modelling approaches") of averaging the proportion of terminal node "local training set

molecules" in "class 1", across all trees in the forest, is employed. The second scenario is that the default approach of each tree predicting the terminal node majority class, with the model returning the proportion of trees voting for "class 1", is employed and, for each tree, all the terminal node molecules belong to the same class. (This default approach is equivalent to predicting the class with the majority of votes once this proportion is compared to the threshold of 0.50.) Indeed, under the second scenario, the "averaged predictions" and default predictions approaches are identical. Under these circumstances, the relationship between the Random Forest model output and the descriptor contributions is given in equation (4).[19,29,30] In equation (4), o_X is the model output for molecule *X* (either the predicted bioactivity for a regression model or the score for "class 1" for a binary classification model); dc_j^X denotes the descriptor contribution obtained for the *j*th descriptor for molecule *X*; A_0 is a constant, for a given model.

$$o_X = A_0 + \sum_j^J dc_j^X \quad (4)$$

Local gradients approach

For this approach, an estimate of the value, for a given molecule *X*, of the partial derivative of the model output with respect to a given descriptor was calculated as per equation (5). In equation (5), dc_j^X denotes the descriptor contribution obtained for the *j*th descriptor for molecule *X*; $o_X\left(d_{jX}, \{d_{j*X}\}_{j*\neq j}\right)$ denotes the model output calculated for molecule *X* based on its values for the *j*th descriptor (d_{jX}) and all other descriptors; $o_X\left(d_{jX} + \Delta, \{d_{j*X}\}_{j*\neq j}\right)$ denotes the model output when the value of the *j*th descriptor is adjusted by Δ and all other descriptor values are kept the same.

$$dc_j^X = \frac{o_X\left(d_{jX}+\Delta,\{d_{j*X}\}_{j*\neq j}\right) - o_X\left(d_{jX},\{d_{j*X}\}_{j*\neq j}\right)}{\Delta} \quad (5)$$

N.B. (1) The model output is the predicted bioactivity for a regression model, or score for "class 1" for a binary classification model. (2) Since binary descriptors were used for the work presented in the current article (see "Descriptors employed"), i.e. the only possible descriptor values were 1 or 0, Δ would be +1 if $d_{jX} = 0$ and -1 if $d_{jX} = 1$. (3) In practice, since only descriptor contributions for descriptors where $d_{jX} = 1$ were used as the basis for analysis in our current article (see "Heat Map interpretation of individual predictions"), $\Delta = -1$ for all calculations performed.

Heat Map interpretation of individual predictions

The high dimensionality (see Table 1) of ECFP fingerprints and the fact that their corresponding substructures overlap[13,44] means it makes sense[13,24] to aggregate the descriptor contributions obtained for all substructures containing a given bond (or atom) and color that bond (or atom) according to the aggregate descriptor contributions. This allows for the contributions of different regions of the molecule to the output generated by the model to be visualized in a manner that is transparent to chemists.

Here, we employed variations of the Heat Map coloring scheme of Rosenbaum et al.,[13] where bonds (or atoms) are colored via calculating raw scores for those bonds (or atoms), based on the descriptor contributions for the corresponding ECFP feature encoded substructures present in the molecule, and then normalizing the raw bond (or atom) scores between 0 and 1. The original, raw score for a bond (or atom) is computed by summing the descriptor contribution values for all substructures containing the bond (or atom). This means that bonds (or atoms) associated with

features making a positive (negative) contribution to the model prediction, according to the descriptor contributions, are assigned positive (negative) raw scores. The normalized bond (or atom) score is transformed to a color, which runs from strong red (normalized score of 0), to orange (normalized score above 0.25), through yellow (normalized score close to 0.5), to green (normalized score at least 0.54), and ultimately to strong green (normalized score of 1). The more positive the raw bond (or atom) score, the closer to 1 the normalized score will lie and, hence, the more green the assigned color will be.[13]

In this work, various approaches to normalizing the raw bond (or atom) scores were considered: (a) Single Molecule Normalization, (b) Full Dataset normalization, (c) Symmetrized Single Molecule Normalization and (d) Constant Symmetrized Normalization. Approaches (a – d) correspond to equations (6 – 9) respectively. Approaches (a) and (b) were proposed by Rosenbaum et al.,[13] whilst the symmetrized schemes (c) and (d) were introduced in the current work. In contrast to the previously proposed normalization schemes, Symmetrized Single Molecule Normalization and Constant Symmetrized Normalization ensure that negative and positive raw scores are matched to normalized scores less than 0.5 and greater than 0.5 respectively. This ensures that regions of the molecule contributing negatively (positively) to the prediction are colored from yellow, to orange, to red (from yellow to green). In principle, Symmetrized Single Molecule Normalization, as opposed to Constant Symmetrized Normalization, may result in weaker green coloring being assigned to equally positive regions of the molecule upon moving to a molecule with large magnitude negative raw scores. However, both Full Dataset Normalization and Constant Symmetrized Normalization may result in small magnitude positive or negative raw bond (or atom) scores being assigned normalized scores

close to 0.5, such that they appear yellow i.e. their sign is indeterminate from the Heat Map image.

$$s_n = \frac{\left(s_r - s_r^{min,mol}\right)}{\left(s_r^{max,mol} - s_r^{min,mol}\right)} \quad (6)$$

$$s_n = \frac{\left(s_r - s_r^{min,ds}\right)}{\left(s_r^{max,ds} - s_r^{min,ds}\right)} \quad (7)$$

$$s_n = \frac{\left(s_r + |s|_r^{max,mol}\right)}{\left(2 \times |s|_r^{max,mol}\right)} \quad (8)$$

$$s_n = \frac{\left(s_r + |s|_r^{max,set}\right)}{\left(2 \times |s|_r^{max,set}\right)} \quad (9)$$

In equations (6 – 9), s_r denotes a raw bond (or atom) score and s_n denotes a normalized score. $s_r^{min,mol}$ denotes the minimum (i.e. most negative if there are negative raw scores) raw score for the molecule, whilst $s_r^{max,mol}$ denotes the maximum (i.e. most positive if there are positive raw scores) raw score for the molecule. $s_r^{min,ds}$ and $s_r^{max,ds}$ denote the minimum and maximum raw bond (or atom) scores respectively across a specified dataset of molecules; in keeping with Rosenbaum et al.,[13] this dataset was the training set plus the molecule of interest. $|s|_r^{max,mol}$ denotes the largest raw bond (or atom) score *magnitude* for the molecule of interest. Finally, $|s|_r^{max,set}$ denotes the largest raw bond (or atom) score *magnitude* corresponding to a set of model predictions and associated descriptor contributions. In the current work, this set was the complete set of descriptor contributions calculated for all molecules of interest (see "Molecules considered for detailed analysis") for a given dataset with all combinations of modelling,

descriptor contribution and, where relevant to the descriptor contributions, predictions approaches.

Due to software implementation challenges, it was not possible to generate Heat Map images corresponding to Full Dataset Normalization, i.e. equation (7), for Random Forest models. (See "Computational details" in the Supporting Information.)

The interpretability of predictions was assessed via evaluating the biological plausibility of the Heat Map colors. Further details are provided under "Quality assessment of Heat Map images".

Molecules considered for detailed analysis

Heat Map images and corresponding "external, leave-one-out" predictions were generated and evaluated for specific molecules of interest. The models used to generate these Heat Maps and predictions were built upon the entirety of the corresponding dataset, save for the molecule of interest where applicable, using the same procedure employed to build models evaluated via "external cross-validation". For the Kazius and CA datasets, the same molecules subjected to detailed analysis by Rosenbaum et al.[13] and Mohr et al.[17] respectively were considered and removed, in turn, from the datasets. For all protein interaction datasets, corresponding Protein Data Bank (PDB)[63,64] protein-ligand crystal structures, with ligands not found in the original datasets, were identified. Further details are provided in the Supporting Information.

In addition to these molecules, for which manual evaluation of the Heat Map images was performed, "external, leave-one-out" predictions and corresponding Heat Map images were generated for an additional set of molecules from the Kazius dataset matching a set of toxicophore encoding SMARTS patterns. These SMARTS patterns are provided in the

Supporting Information file "Kazius_2005_SI_SA_SMARTS_subset_v2.xls" and were derived from Kazius et al.,[48] as explained therein.

Quality assessment of Heat Map images

The Heat Map images generated for the molecules of interest were assessed in terms of (a) available mechanistic knowledge concerning (potential) toxicophores and detoxifying groups or, for the models trained on the protein interaction datasets, (b) the corresponding PoseView[65] images of protein-ligand interactions available from the PDB.

Primarily, these images were assessed by considering how the molecular substructures highlighted as most positive (i.e. green), suggesting a significant positive contribution towards a prediction of membership of the active (toxic) class ("class 1") or the predicted continuous biological activity, corresponded to those substructures expected to be directly responsible for promoting biological activity. In the case of molecules from the Kazius[48] and CA datasets,[17] this entailed comparing the most positive regions of the molecule to the known (or hypothesized) toxicophores for Ames mutagenicity or chromosome aberration respectively.[13,17,48] In the case of those images which corresponded to PDB ligands, the most positive regions of the molecule were compared to the interactions indicated by the PoseView images[65] provided by the PDB.[63]

An important point to note is that, by assessing the correspondence between the Heat Map and the known influence of the biologically relevant substructures, this assessment is not necessarily equivalent to an assessment of how well the interpretation approach provides the basis for the model's prediction. Rather, this assessment evaluates the extent to which the Heat Map images can provide insights into the biological significance of different molecular substructures. For a prediction made for an untested chemical, this is useful information to obtain from a QSAR

model in its own right, hence an assessment of how well the Heat Map images can provide this information on chemicals for which the most biologically relevant substructures are known is valuable. (This assumes the ability of the Heat Map images to provide this information on those chemicals is representative of how well they would provide this information on chemicals for which the most biologically significant substructures are unknown.) Furthermore, if it can be assumed that the model makes correct predictions for the right reason, this assessment indicates how well a Heat Map image explains the basis for the model's prediction.

Assessment of the correspondence between green coloring, suggesting a significant positive contribution to predicted bioactivity, on the Heat Map and the substructures expected to be directly responsible for promoting biological activity took account of two important considerations. The first consideration was the extent to which regions of the molecule known to directly promote biological activity were highlighted as green. The second consideration was the extent to which biologically irrelevant and, for the minority of molecules for which this was applicable, known activity reducing regions of the molecule were highlighted as green. Both failing to (fully) identify those regions of the molecule directly responsible for promoting biological activity, as well as wrongly identifying biologically irrelevant or activity reducing regions of the molecule as promoting biological activity with respect to the endpoint of interest, would mislead scientists seeking to interpret the images, e.g. medicinal chemists seeking ideas for lead optimization. Hence, the following ranking scheme was initially used to assess the images (**Table 2**).

For some molecules for which Heat Map images were analyzed, detoxifying substructures were identified which were known to reduce (or even abolish) the biological activity expected to be conferred by the corresponding toxicophore.[13,48] For Heat Map images corresponding to

predictions for PDB ligands, hydrogen bond donors/acceptors (e.g. tertiary amines) in regions suggested to be involved in hydrophobic interactions by the corresponding PoseView[65] images were penalized when evaluating positive (green) coloring according to the ranking scheme (**Table 2**). A failure to satisfy ligand hydrogen bond acceptors and, possibly even more so, donors located within a hydrophobic pocket is enthalpically unfavorable.[66]

Table 2. Summary of the ranking scheme initially used to denote the quality of the Heat Map images based on considering the biological relevance of the most positively highlighted (i.e. green) regions of the molecule

Biological relevance (positive coloring): Rank[a]	Biological relevance (positive coloring): Label	Criteria used for assignment
1	Strongest	All regions of the molecule contributing directly towards biological activity were highlighted as most positive in their entirety, without any biologically irrelevant substructures, or substructures expected to reduce biological activity, highlighted as most positive.
2	Strong	The regions of the molecule contributing directly towards biological activity were partially highlighted as most positive, without any biologically irrelevant substructures, or substructures expected to reduce biological activity,

		highlighted as most positive.
3	Weak	All regions of the molecule contributing directly towards biological activity were highlighted as most positive in their entirety, but some biologically irrelevant substructures, or substructures expected to reduce biological activity, were also highlighted as most positive.
4	Weaker	The regions of the molecule contributing directly towards biological activity were partially highlighted as most positive, but some biologically irrelevant substructures, or substructures expected to reduce biological activity, were also highlighted as most positive.
5	None	No match between the molecular regions highlighted as most positive and those known to directly promote biological activity was observed.

a. A lower rank denotes greater biological relevance i.e. higher Heat Map image quality.

For the minority of molecules with known activity reducing substructures, an analogous analysis was performed based upon assessing the correspondence between those substructures highlighted as most negative (i.e. red) and the known activity reducing substructures.

Ultimately, these Heat Map images will be useful for chemists inspecting them to gain insights into the basis for a given model prediction and/or the indicated biological significance of molecular substructures, according to a given interpretation approach and QSAR model.

However, manual assessment of Heat Map images is not only time consuming but is also prone to inconsistent categorization, i.e. the scheme outlined in **Table 2** does not fully remove all potential subjectivity or inconsistency associated with manual Heat Map evaluation. Nonetheless, assessment according to a systematized scheme is necessary for effective evaluation of the degree to which different interpretation approaches yield chemically and biologically meaningful interpretations of QSAR predictions. To overcome the limitations of manual assessment, an automated, numeric "quality score" scheme was implemented. This automated scheme, as illustrated by **Figure 1**, was only implemented in the current work for Heat Map images based on atom coloring and either Symmetrized Single Molecule Normalization, or Constant Symmetrized Normalization. Furthermore, it was only applied to a subset of Kazius dataset molecules for which toxicity promoting substructures and, where applicable, corresponding detoxifying substructures were encoded using SMARTS patterns derived from the work of Kazius et al.[48]

Each of the relevant Heat Map images is assigned a single, numeric "quality score", as illustrated by **Figure 1**, designed to estimate the degree to which the image provides a chemically and biologically meaningful interpretation of the corresponding model prediction. For a molecule with one or more SMARTS pattern identified toxicophores, the quality score measures the degree to which the green coloring of atoms fully and precisely identifies the atoms lying inside the biologically active substructure(s). (Only atoms estimated to make a relatively significant positive contribution to the predicted bioactivity will be colored green if symmetrized normalization and atom coloring is used.) This is quantified in terms of the F-measure, which takes into account the proportion of true positives to false "negatives" and false positives.[67] Here, the true positives are all atoms with a normalized atom score >= 0.54, which was visually

estimated to be the point at which an atom starts to appear green, and which were matched by toxicophore SMARTS. False positives have normalized atom scores >= 0.54, yet are not matched by a toxicophore SMARTS. False "negative" atoms are those matched by the toxicophore SMARTS, but which are not selected based on this threshold. This calculation is illustrated, for an example molecule, on the LHS of **Figure 1**.

For a molecule which also contains substructures that are known to be deactivating, an analogous calculation is performed with regards to those atoms highlighted as sufficiently red. (This calculation is illustrated, for an example molecule, on the RHS of **Figure 1**.) For Heat Map images generated via symmetrized normalization and atom coloring, this means those atoms are known to make a significant negative contribution to the predicted bioactivity. The atoms indicated to significantly reduce toxicity, based on the Heat Map, were those with a normalized score <= 0.25, which was visually estimated to be the point at which an atom becomes discernably red. In this latter case, the overall quality score is calculated as the arithmetic mean of the two F-measures.

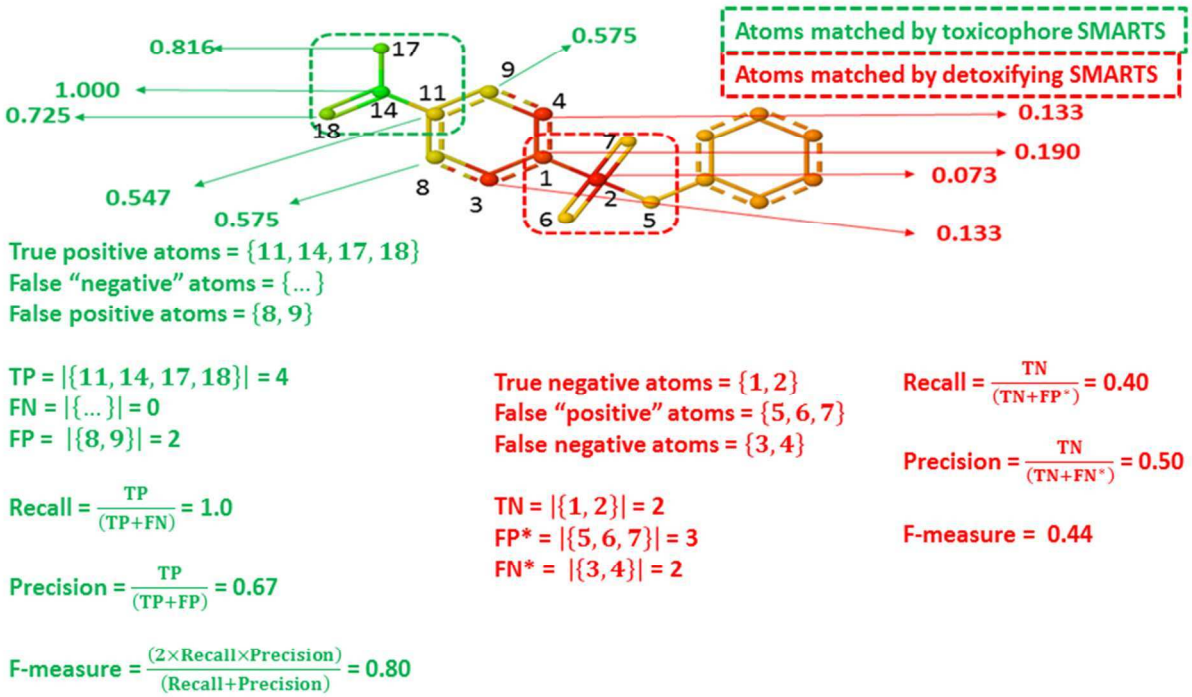


Figure 1 The procedure for calculating numeric Heat Map quality scores for an image obtained via symmetrized normalization (Symmetrized Single Molecule Normalization in this example) and atom coloring. The relevant normalized atom score estimates are shown matched to the atom IDs. In this case, the overall quality score would be the arithmetic mean of the two F-measures i.e. 0.62. Further details are provided in the associated text. SMARTS matching of toxicophores and detoxifying substructures was implemented using OpenBabel (version 2.3.2) & Pybel,[68,69] based on SMARTS adapted from the work of Kazius et al.[48] and reported in the Supporting Information. The atom IDs, used to identify the corresponding atom score estimates for this image, were identified using OpenBabel[68], via this command: "obabel [SDF containing molecule of interest] -O [name of image].svg -xi –xu".

Normalization of binary classification scores for molecules subjected to detailed analysis

The quality of the Heat Map images was compared to the quality of the corresponding predictions (see below). The raw prediction scores of the binary classification models were normalized to facilitate comparison across methods. The raw prediction scores for "class 1", the "toxic" class, are of different scales. Normalization entailed transforming these values, where required, to lie between 0 and 1. Subsequent to normalization, a score greater than 0.50 resulted in predicted assignment to "class 1" for all methods. Details of this normalization are provided in the Supporting Information.

Correlation between prediction quality and Heat Map image quality

Heat Map images have the potential to inform chemists in the analysis and design of molecules. The quality of those images, in terms of correctly highlighting active regions of the molecule, is therefore of paramount importance. It is useful to understand the extent to which the quality of a Heat Map image depends upon the quality/strength of the corresponding prediction. If a strong correlation between Heat Map quality and prediction quality is observed, this can help chemists judge whether a Heat Map image offers meaningful interpretation in the absence of knowing the true determinants of biological activity. Firstly, in the case that predictions are made for a biologically untested molecule, statistical estimates of the prediction quality (e.g. as obtained from cross-validation on the training set) may suggest whether the Heat Map image can be trusted. Secondly, in the case that the prediction is experimentally tested yet the mechanism of action remains unknown, the known quality of the prediction could be used to judge whether the Heat Map image offers useful insights.

In order to determine whether the quality of these Heat Map images, calculated robustly as per **Figure 1**, corresponded to the quality of the corresponding predictions, it was necessary to rank the quality of these predictions. Since the scoring scheme described in **Figure 1** could only be applied to external leave-one-out binary classification predictions for the Kazius dataset, high quality predictions were defined as correct predictions, with all incorrect predictions being considered of low quality.

Computational details

Scripts and software configuration files, used for evaluation of the modelling approaches considered in our article, have been made freely available.[70] These files are described, along with information about software and hardware dependencies used to generate our results, in the Supporting Information. Two key dependencies, which we have made freely available to the cheminformatics community, are the *rfFC* package[37] for the R Statistical Programming language[71] and the *HeatMapWrapper* Python program.[72] The *rfFC* package implements algorithms for interpreting predictions made using the *randomForest* package. The *HeatMapWrapper* program allows for any set of descriptor contributions, associated with binary descriptors corresponding to fingerprints calculated as per our work, to be used as the basis for Heat Map images. Hence, the former program supports interpretation of Random Forest predictions, whilst the latter program supports the interpretation of any QSAR prediction built using the same fingerprints as our models, as long as suitable descriptor contributions can be calculated. These dependencies were used to generate the Heat Map images analyzed in our current article, as well as provide estimates of the underlying normalized and corresponding raw atom scores. Additional details are presented in the Supporting Information.

RESULTS & DISCUSSION

Predictive performance

The overall performance estimates for the investigated modelling approaches, from "external cross-validation", are presented in **Table 3** for key binary classification datasets. The performance of all modelling approaches on all remaining binary classification datasets was worse, in terms of all figures of merit, than for the Kazius dataset and typically worse than for the CA dataset. Subsequent investigations of prediction interpretations, based on additional leave-one-out predictions, are discussed in detail for the Kazius dataset, with some illustrative examples also taken from the CA dataset. Since modelling results on the other datasets are not considered in greater detail in our article, further detailed performance estimates are only presented in the Supporting Information (Tables S5 and S6 under "Predictive performance: additional result summaries").

Nonetheless, in summary, Random Forest classification, with "averaged predictions", is the best performing approach, in terms of the arithmetic mean AUC, for all but one (MUV713) of the five classification datasets. (For MUV713, Random Forest classification using the default majority vote predictions is marginally better, with identical mean AUC at 3dp. However, the average performance of all models on the MUV713 dataset is close to that expected from random guessing.)[61] Random Forest classification, using the "averaged predictions" approach, also typically yields the highest predictive performance in terms of the arithmetic mean values for all other performance metrics. However, it should be noted that the differences in performance between methods are sometimes marginal and may not be robust.

Likewise, Random Forest regression is the best performing approach according to the arithmetic mean R^2 for four (ACE, DHFR, GPB, THR) out of the eight regression datasets. Again, it should be noted that the differences in performance are sometimes marginal and may not be robust.

Regarding the robustness of the results, an approximate assessment of statistical significance was carried out for the pairwise differences in mean performance metrics. For both the classification and regression datasets, over half of differences were statistically significant according to a standard paired, two-tail t-test, even upon applying multiple testing corrections.[73] However, this test suffers from elevated Type 1 error in our evaluation framework of averaged K-fold testing,[74,75] as explained in the discussion of "Statistical Significance" in the Supporting Information. Therefore, it is likely that the results are not trustworthy and, hence, no detailed analysis is presented.

Table 3. Overall performance estimates of binary classification modelling approaches obtained from "external cross-validation" for key datasets: arithmetic mean performance statistics (3dp)

Dataset	Modelling Approach	AUC	MCC	Kappa	BA
CA	Random Forest classification (AP)[a]	0.762	0.369	0.352	0.683
	Random Forest classification	0.748	0.358	0.333	0.676
	Support Vector Machine	0.747	0.346	0.327	0.670
	Probit-PLS	0.746	0.339	0.326	0.664
Kazius	Random Forest classification (AP)[a]	0.876	0.600	0.598	0.802

	Random Forest classification	0.862	0.595	0.592	0.799
	Probit-PLS	0.855	0.570	0.566	0.787
	Support Vector Machine	0.854	0.568	0.564	0.785

a. "AP" denotes Random Forest binary classification predictions made using the, non-default, "averaged predictions" approach described under the Random Forest classification subsection of "Modelling approaches".

Summary of Heat Map interpretations of individual predictions

A variety of Heat Map images were generated, corresponding to (a) leave-one-out binary classification predictions for the Kazius and CA datasets, for 39 and 15 molecules respectively, or (b) regression predictions made for eight PDB ligands corresponding to one of the Sutherland datasets. A summary of these images and the assessment made of the correspondence between the Heat Map and the known biologically relevant substructures is provided in the Supporting Information (see "FinalImagesAnalysis_withImages_resub.xlsx").

As well as different approaches to generating descriptor contributions, different Heat Map images were also obtained for the same molecule via translating these descriptor contributions into Heat Map images according to (a) different normalization schemes and (b) atom or bond coloring based on the descriptor contributions. These different schemes are explained in "Heat Map interpretation of individual predictions" under "Methods & Data" above. Due to software implementation limitations explained under "Computational details" in the Supporting Information, it was only possible to generate "Full Dataset Normalization" Heat Map images for

the linear modelling methods: Probit-PLS, PLS regression and, with a linear kernel, SVM and Support Vector Regression.

Systematic manual assessments, according to **Table 2**, were made for all Heat Map images corresponding to the examples from the Kazius and CA datasets (two and 15 molecules respectively) which were previously used for assessment of interpretation approaches,[13,17] along with all PDB ligands. In addition, it was possible to calculate automated numeric Heat Map quality scores, as per **Figure 1**, for all 39 molecules in the Kazius dataset, identified via toxicophore SMARTS pattern matches, for which Heat Map images were generated. Whilst 2026 Kazius dataset molecules were identified via toxicophore SMARTS pattern matches and, in principle, automated numeric Heat Map quality scores could have been calculated for all of these, the currently available software for generating Heat Map images[72] requires laborious manual intervention via a GUI to obtain each image. This limited the number of Heat Map images it was practical to generate for the current paper.

Effect of different coloring schemes upon Heat Map interpretations

The choice of atom or bond coloring, as well as the choice of normalization scheme, can yield discernably different Heat Map images. This is illustrated by **Figure 2**.

Only the symmetrized normalization schemes introduced in this paper (Symmetrized Single Molecule Normalization and Constant Symmetrized Normalization) are guaranteed to ensure that positive raw bond (atom) scores, calculated from the underlying descriptor contributions, are translated into yellow to green coloring, becoming increasingly green with increasing magnitude, and negative raw scores are translated into yellow to orange to red coloring. With the previously proposed normalization schemes,[13] Full Dataset Normalization and especially Single Molecule

Normalization, green or red coloring might be assigned which would give a misleading impression as to the sign of the raw bond (or atom) scores indicating whether a particular molecular substructure was responsible for a negative or positive effect on the model's prediction. This can be seen by comparing the images in **Figure 2** with the estimated corresponding raw atom scores for this molecule in **Figure 3**. N.B. All normalized and raw atom scores we present are estimates, as discussed under "Computational Details" in the Supporting Information.

As illustrated by **Figure 2**, both Full Dataset Normalization and Constant Symmetrized Normalization can make the magnitude of all normalized bond (or atom) scores close to 0.5. This either reduces the colors to yellow, which conveys no information about the sign of the raw bond (or atom) scores and allows for no visual interpretation, or makes them only weakly green or orange, which makes differences between different parts of the molecule harder to discern.

In principle, Constant Symmetrized Normalization ensures that only substructures contributing negatively to the prediction and substructures contributing positively would appear red and green respectively, in keeping with Symmetrized Single Molecule Normalization, as well as offering certain theoretical advantages. Firstly, it allows for the variation in the strength of negative and positive coloring between molecules to be evaluated, due to the use of consistent normalization across all Heat Map images. However, as long as the regions are sufficiently positive or negative to appear green and red respectively under Symmetrized Single Molecule Normalization or Constant Symmetrized Normalization, this would have no bearing on the outcome of the evaluation schemes applied in our work. Secondly, the use of Constant Symmetrized Normalization avoids a hypothetical problem with Symmetrized Single Molecule Normalization: the possibility that a positive region of the molecule may switch from green to yellow due to the

inclusion of a large magnitude negative raw atom score associated with a red substructure. However, this does not appear to have happened in practice, as consideration of both **Figure 4** and **Figure 5** indicates. Rather, as illustrated by **Figure 5**, the effect of Constant Symmetrized Normalization appears to typically be to reduce the normalized atom scores close to 0.5, making the corresponding atoms (or bonds) appear yellow and, hence, impossible to obtain any interpretation from. For this reason, the use of Symmetrized Single Molecule Normalization is recommended and Heat Map images generated according to this scheme were used to draw conclusions from our work.

Regarding the choice of atom or bond coloring, atom coloring is arguably more appropriate for discerning whether a hydrogen bond donor or acceptor atom[66] is making a significant contribution towards the prediction. Different bond coloring either side of this atom could lead to an ambiguous interpretation. However, it *may* be possible for some pi-bonds to act as weak hydrogen bond acceptors,[76] in which case bond coloring would be more appropriate, as was previously suggested in the literature on different grounds.[13] However, the Pybel software[69] used for SMARTS matching, hence automated numeric Heat Map quality scoring, for the current paper identified the parts of the molecule matched by the SMARTS patterns in terms of atom indices. Hence, this automated scoring was only applicable to Heat Map images based on atom coloring. For this reason, only Heat Map images corresponding to atom coloring are considered for the rest of our article.

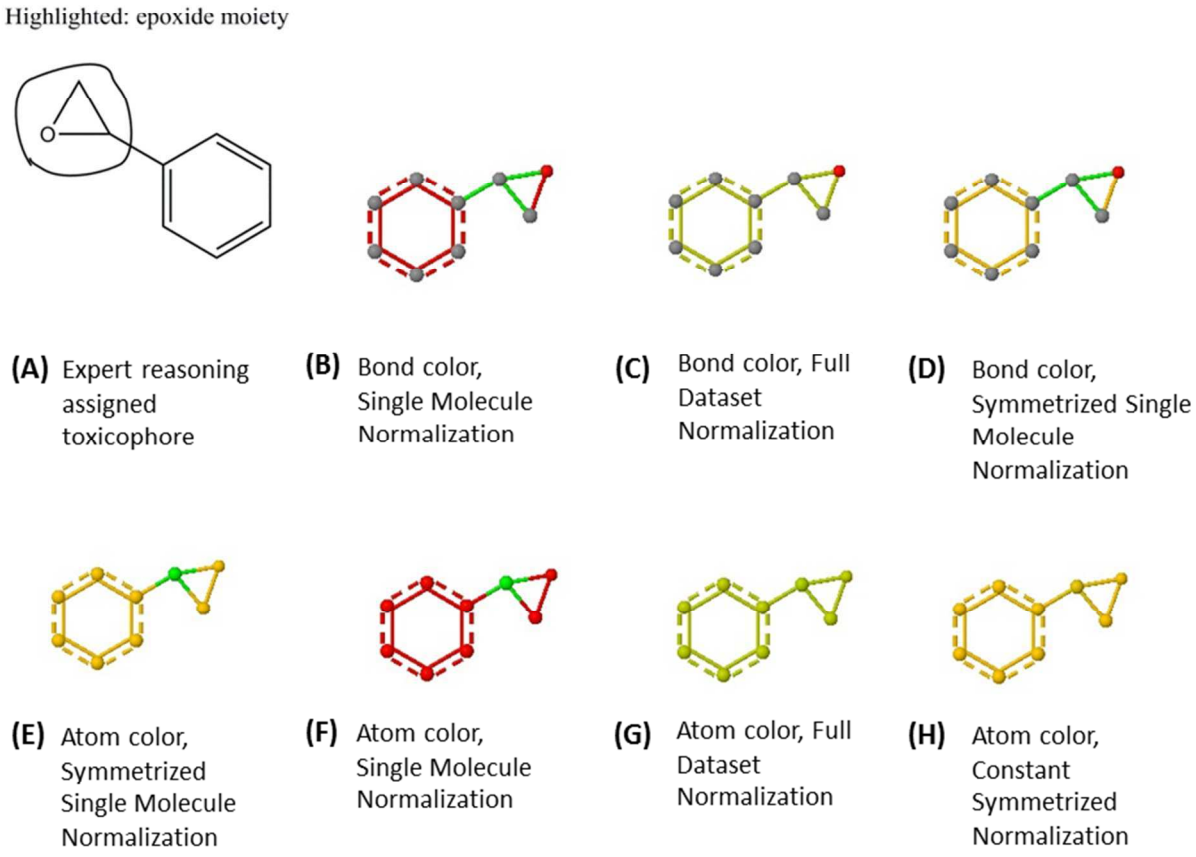


Figure 2 Heat Map analysis of a binary classification leave-one-out SVM prediction made for molecule "CA23" in the CA dataset (c.f. Figure 4(b) in Mohr et al.), determined to be toxic (positive result) in the chromosome aberration test.[17] (A) Expert reasoning assigned toxicophore (epoxide functionality), based on the rationale provided in Mohr et al.:[17] this substructure was flagged by an Ames test alerts knowledgebase, with 80% of Ames positives being chromosome aberration positives, and is known to exhibit electrophilic reactivity with DNA. (B – H): Heat Map images based on descriptor contributions associated with the same SVM prediction, generated according to either bond or atom coloring obtained via different normalization schemes.

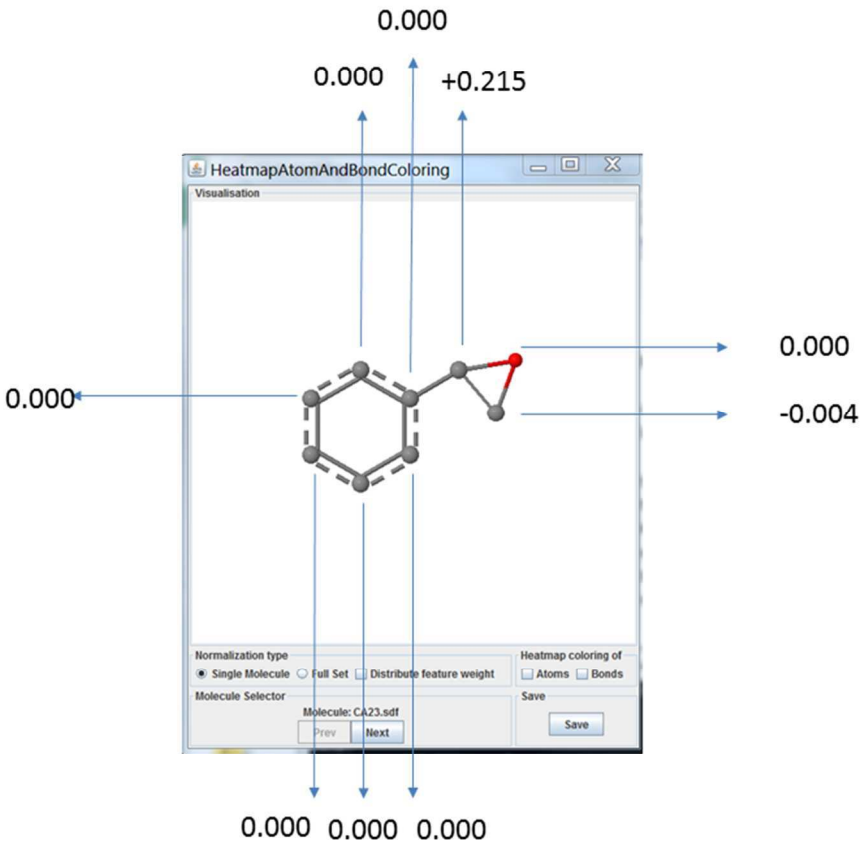


Figure 3 Image of the molecule "CA23" in the CA dataset[17] displayed in HeatmapViewer.jar GUI,[13] as generated by running the HeatMapWrapper tool,[72] with both atom and bond coloring switched off. The intermediate output generated via running the HeatMapWrapper tool in atom color, Symmetrized Single Molecule Normalization mode, based on leave-one-out SVM descriptor contributions, was processed to reveal estimated raw atom scores ranging from weakly negative in one case (-0.004), through zero in most cases, to positive (+0.215) in one case. The estimated raw atom scores are rounded to 3dp. This image corresponds to all Heat Map images shown in **Figure 2**, with different normalization schemes generating the different normalized scores used for coloring. Via switching off atom coloring, based on normalized scores, the epoxide oxygen (in red) and carbon atoms (in black) can readily be discerned. N.B. The specific

atom IDs, used to identify the corresponding atom score estimates, were identified as per **Figure 1**.

Strengths and weaknesses of different Heat Map quality assessment schemes

Ultimately, the Heat Map images are designed to provide chemically and biologically meaningful interpretations of model predictions to chemists, hence require manual consideration. However, a systematized evaluation scheme is required to allow for a reasonable assessment of the degree to which different modelling and prediction interpretation approaches, underpinning these Heat Map images, tend to yield chemically and biologically meaningful interpretations.

Whilst a manual, qualitative evaluation scheme (**Table 2**) was initially devised, this has a number of weaknesses. These prompted the development of an automated, numeric scoring scheme (**Figure 1**). The following figures (**Figure 4**, **Figure 5**, **Figure 6**) compare and contrast the application of both schemes to two Kazius dataset molecules, with Heat Map images generated according to various scenarios, for which external leave-one-out predictions were obtained.

Arguably the greatest weakness of the manual ranking scheme (**Table 2**), is the potential for subjectivity or inconsistency in assignments to be made. Inconsistency is likely to be greater if a larger number of images are considered, with the laborious requirement to visually consider each image restricting the number of images which can be considered and, hence, limiting the robustness of any observed trends. The application of the automated, numeric scoring scheme (**Figure 1**) allows for greater objectivity and for a large number of Heat Map images to be assessed with ease, allowing more robust conclusions to be drawn. Of course, this assumes a large number of Heat Map images have been generated and the molecular features directly

responsible for promoting (or reducing) activity have been suitably encoded as SMARTS patterns.

Nonetheless, it could be argued that the imposition of hard boundaries between those atoms involved in biologically relevant and irrelevant molecular substructures, via SMARTS matching, whilst removing subjectivity, might sometimes be problematic. Firstly, it should be acknowledged that this sharp distinction between biologically relevant and irrelevant molecular substructures is a simplification e.g. whole molecule physicochemical properties, such as logP, may affect biological activity and arise from molecular features not directly involved in interactions at the site of biological action. Secondly, the SMARTS patterns may fail to account for all biologically relevant knowledge. For example, consider the molecule shown in **Figure 7**. Whilst the putative toxicophore[17] encompasses the entire arylhydrazide substructure, it *may* be the case that biological activity arises from metabolism of this substructure to yield the aromatic amine substructure.[17,48] Hence, it can reasonably be argued that application of the qualitative, manual quality assessment scheme should only assign credit for green coloring of that substructure. A different example, which further illustrates the potential problem associated with applying hard boundaries between biologically relevant and irrelevant sets of atoms as supposed by the automated scoring scheme, is **Figure 4**. According to the SMARTS patterns used for application of the automated scoring scheme, **Figure 4** (B) and (E) should be penalized for highlighting the ortho positions, to the sulfonamide attached to the aromatic ring, as red. However, one might reasonably argue that, at least in this context, this still serves to indicate that the specific effect of having a sulfonamide attachment at this end of the ring is detoxifying. In principle, one could adjust the SMARTS patterns to accommodate these scenarios when applying the numeric scoring scheme e.g. by expanding the SMARTS patterns used for

automated assessment of the molecule in **Figure 4** to accommodate more of the aromatic ring. However, the most appropriate modifications would still need to be decided with care. For example, consider simply extending these SMARTS patterns to encompass a larger portion of the aromatic ring. This could risk giving credit to Heat Map images which highlighted a significant portion of the aromatic ring, or other attachment points, without highlighting the attachments which were actually responsible for promoting (or attenuating) biological activity. Likewise, with reference to this example, the exact point at which to stop the SMARTS pattern indicating that a sulfonamide attachment to the ring was detoxifying and a para-nitro attachment was a toxicophore is also debatable.

In spite of this, it should still be concluded that the numeric, automated scoring scheme is a more appropriate means for systematic assessment of Heat Map images than the qualitative, manual scheme. As well as the potential for inconsistency and difficulty of applying to large numbers of images already discussed, consider scenario in which there are two toxicophores that can be encoded as SMARTS patterns (**Figure 6**). The qualitative scheme would fail to assign more credit to a Heat Map image which fully highlighted one toxicophore and partially highlighted another than to a Heat Map image which only partially highlighted one of them. The numeric, automated scheme overcomes this limitation.

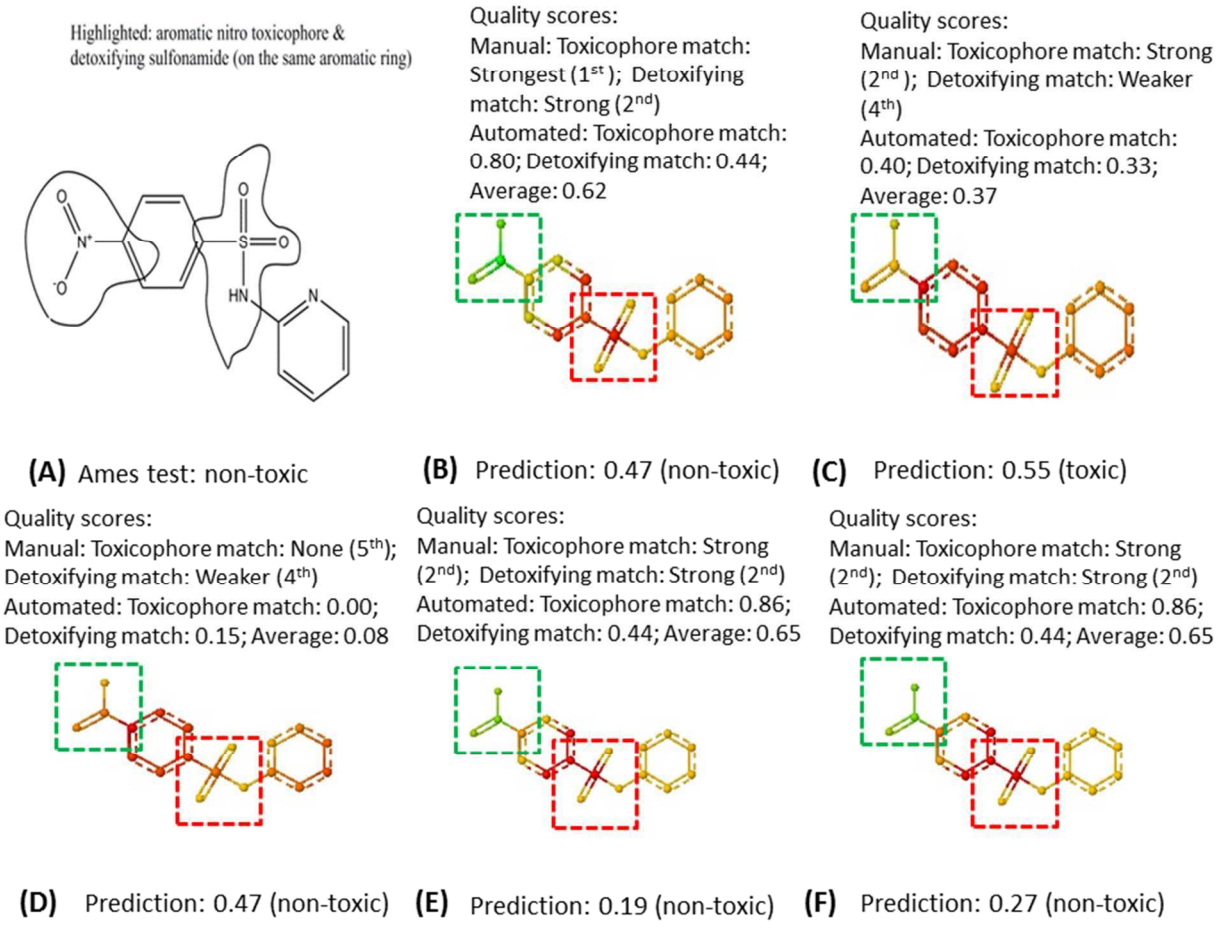


Figure 4 Heat Map (atom coloring & Symmetrized Single Molecule Normalization) analysis of binary classification leave-one-out predictions made for molecule "1028-11-1" in the Kazius dataset (c.f. Figure 3(A) in Rosenbaum et al.).[13] (A) Molecule with biologically significant substructures shown, based on a combination of mechanistic reasoning, expert knowledge and data analysis reported previously in the literature,[13,48] and experimental Ames test assignment reported.[13] (B – F) Heat Map images: (B) Random Forest classification (Kuz'min/Palczewska, averaged predictions); (C) Random Forest classification (local gradients, majority vote predictions); (D) Random Forest (local gradients, averaged predictions); (E) Probit-PLS; (F) SVM. The predictions shown are the normalized scores (transformed into values between 0 and 1) for the toxic class, with predicted assignment to the toxic class if this score was greater than

0.50. The green dotted lines show the atoms matched to toxicophore SMARTS and the red dotted lines the atoms matched to corresponding detoxifying SMARTS. SMARTS matching and identification of matched atoms were carried out as per **Figure 1**. The manual and automated assessments of Heat Map quality are based on the schemes explained under "Quality assessment of Heat Map images".

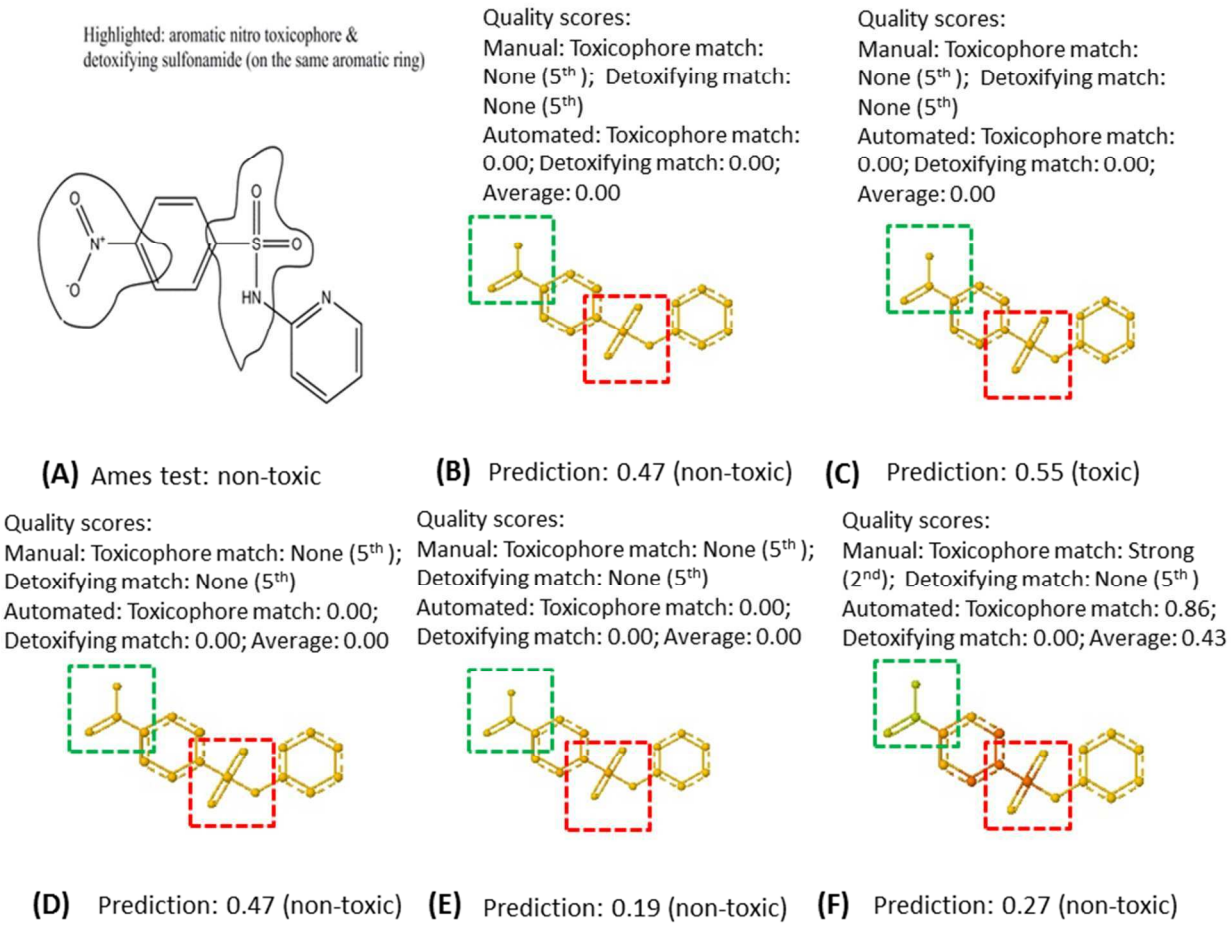


Figure 5 Heat Map (atom coloring & Constant Symmetrized Normalization) analysis of binary classification leave-one-out predictions made for molecule "1028-11-1" in the Kazius dataset (c.f. Figure 3(A) in Rosenbaum et al.).[13] (A) Molecule with biologically significant substructures shown, based on a combination of mechanistic reasoning, expert knowledge and data analysis

reported previously in the literature,[13,48] and experimental Ames test assignment reported.[13] (B – F) Heat Map images: (B) Random Forest classification (Kuz'min/Palczewska, averaged predictions); (C) Random Forest classification (local gradients, majority vote predictions); (D) Random Forest (local gradients, averaged predictions); (E) Probit-PLS; (F) SVM. Predictions are reported as per **Figure 4**. The green dotted lines show the atoms matched to toxicophore SMARTS and the red dotted lines the atoms matched to corresponding detoxifying SMARTS. SMARTS matching and identification of matched atoms were carried out as per **Figure 1**. The manual and automated assessments of Heat Map quality are based on the schemes explained under "Quality assessment of Heat Map images".

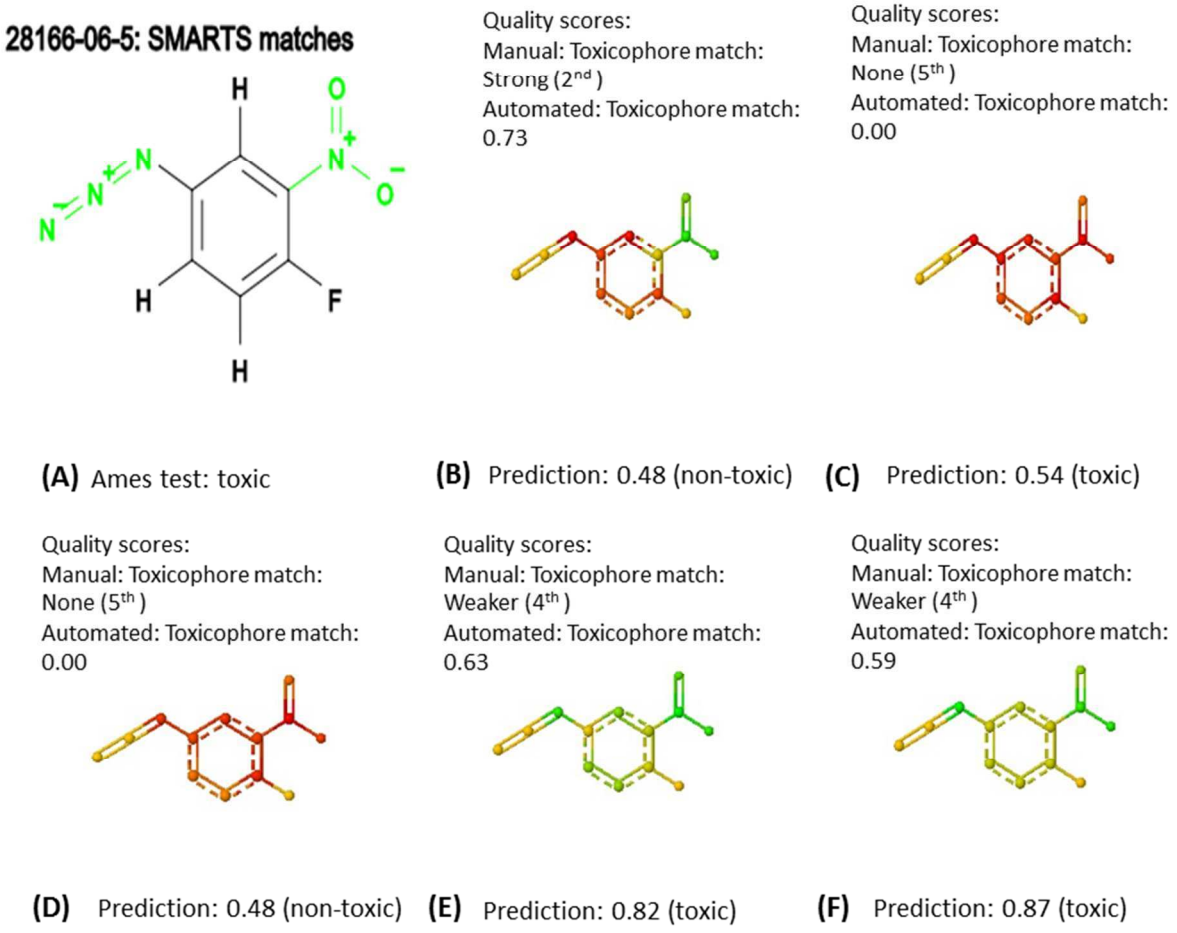


Figure 6 Heat Map (atom coloring & Symmetrized Single Molecule Normalization) analysis of binary classification leave-one-out predictions made for molecule "28166-06-5" in the Kazius dataset. (A) Molecule with SMARTS matches shown in green for specific aromatic nitro (RHS) and azide (LHS) toxicophores, derived from the work of Kazius et al.[48] based on a combination of mechanistic reasoning, expert knowledge and data analysis reported, and experimental Ames test assignment reported in the Kazius dataset SDF file.[13] The SMARTS matches (B – F) Heat Map images: (B) Random Forest classification (Kuz'min/Palczewska, averaged predictions); (C) Random Forest classification (local gradients, majority vote predictions); (D) Random Forest (local gradients, averaged predictions); (E) Probit-PLS; (F) SVM. Predictions are reported as per **Figure 4**. The SMARTS matches show the atoms matched to SMARTS patterns. SMARTS

matching and identification of matched atoms were carried out as per **Figure 1**. The manual and automated assessments of Heat Map quality are based on the schemes explained under "Quality assessment of Heat Map images". Quality scores, not to be confused with prediction scores, based on Heat Map correspondence to detoxifying groups were not assigned as no detoxifying groups were identified via SMARTS matching.

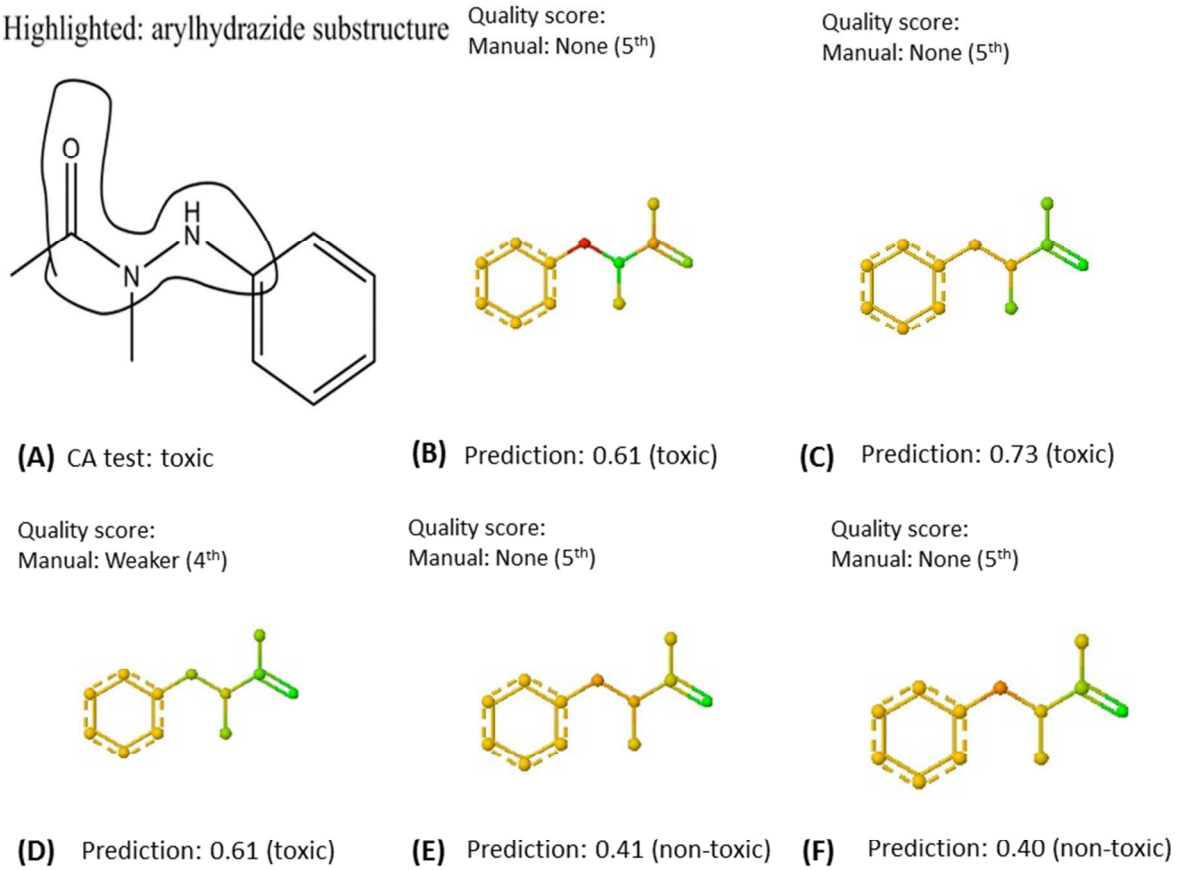


Figure 7 Heat Map (atom coloring & Symmetrized Single Molecule Normalization) analysis of binary classification leave-one-out predictions made for molecule CA31 in the CA dataset (c.f. Figure 6(A) in Mohr et al.).[17] (A) Molecule with biologically significant substructures shown, based on a combination of mechanistic reasoning and expert knowledge reported previously in

the literature,[17] and experimental chromosome aberration test assignment reported.[17] (B – F) Heat Map images: (B) Random Forest classification (Kuz'min/Palczewska, averaged predictions); (C) Random Forest classification (local gradients, majority vote predictions); (D) Random Forest (local gradients, averaged predictions); (E) Probit-PLS; (F) SVM. Predictions are reported as per **Figure 4**. The manual assessments of Heat Map quality are based on the scheme explained under "Quality assessment of Heat Map images". Here, the assignments were not simply based on matching the entire putative toxicophore,[17] but were informed by consideration of the likely toxic substructure revealed via metabolism, which is expected to be the arylhydrazine substructure and may even be the aromatic amine substructure arising from further metabolism.[17,48]

Distributions of numeric Heat Map quality scores for different interpretation approaches

The distributions of overall quality scores, calculated as the average of scores based on Heat Map correspondence to SMARTS pattern assigned toxicophores and detoxifying substructures as per **Figure 1**, for all Kazius dataset Heat Map images generated according to atom coloring and either Symmetrized Single Molecule Normalization or Constant Symmetrized Normalization are shown in **Figure 8** (A) and (B) respectively. As indicated by **Figure 5**, Constant Symmetrized Normalization tended to produce yellow coloring, yielding no information, which results in low quality scores. Hence, as previously noted, Symmetrized Single Molecule Normalization is recommended and no further consideration is given to the results obtained from Constant Symmetrized Normalization.

Furthermore, since only one of the molecules for which numeric Heat Map quality scores were assigned was associated with a detoxifying SMARTS pattern match, the use of the average of the scores calculated as per **Figure 1** is not representative of the distribution. Hence, this single

molecule was excluded from further analysis and all subsequent distributions reflect the quality scores calculated based on Heat Map correspondence to the identified toxciophores for the remaining 38 molecules, or relevant subsets thereof. As expected, removing this single molecule has little impact upon the overall distribution, as can be seen by comparing **Figure 8** (A) and **Figure 9** (A).

Consider the median quality scores presented in **Figure 9** (A). These suggest that the degree to which Heat Map images, based on atom coloring and Symmetrized Single Molecule Normalization, can be meaningfully interpreted was typically somewhat higher for those corresponding to linear SVM predictions compared to all other interpretation approaches. They also suggest the Kuz'min/Palczewska approach to calculating descriptor contributions for Random Forest classification predictions yielded Heat Map images with slightly higher median quality scores than the local gradients approach. The trend in arithmetic mean quality scores followed the same pattern.

However, it might be argued that the quality scores of Heat Map images associated with toxic predictions are of greater relevance, as these indicate the extent to which toxicophores are correctly identified when a toxic prediction is made. A Heat Map image which fully and precisely highlighted the toxicophores in combination with a prediction of toxicity, as per **Figure 11** (F), would be valuable to medicinal chemists seeking to design out a predicted toxic liability during lead optimization. (Of course, this is also reliant on the prediction of toxicity being reliable as per **Figure 11** (F).) Consider the distributions of quality scores, reflecting the degree to which Heat Map images associated with toxic predictions correctly identified the toxicophore(s), in **Figure 10** (A). These indicate roughly the same rank ordering as per **Figure 9** (A), except that the ordering of quality scores associated with Heat Map images corresponding to

different kinds of Random Forest classification local gradients descriptor contributions is reversed.

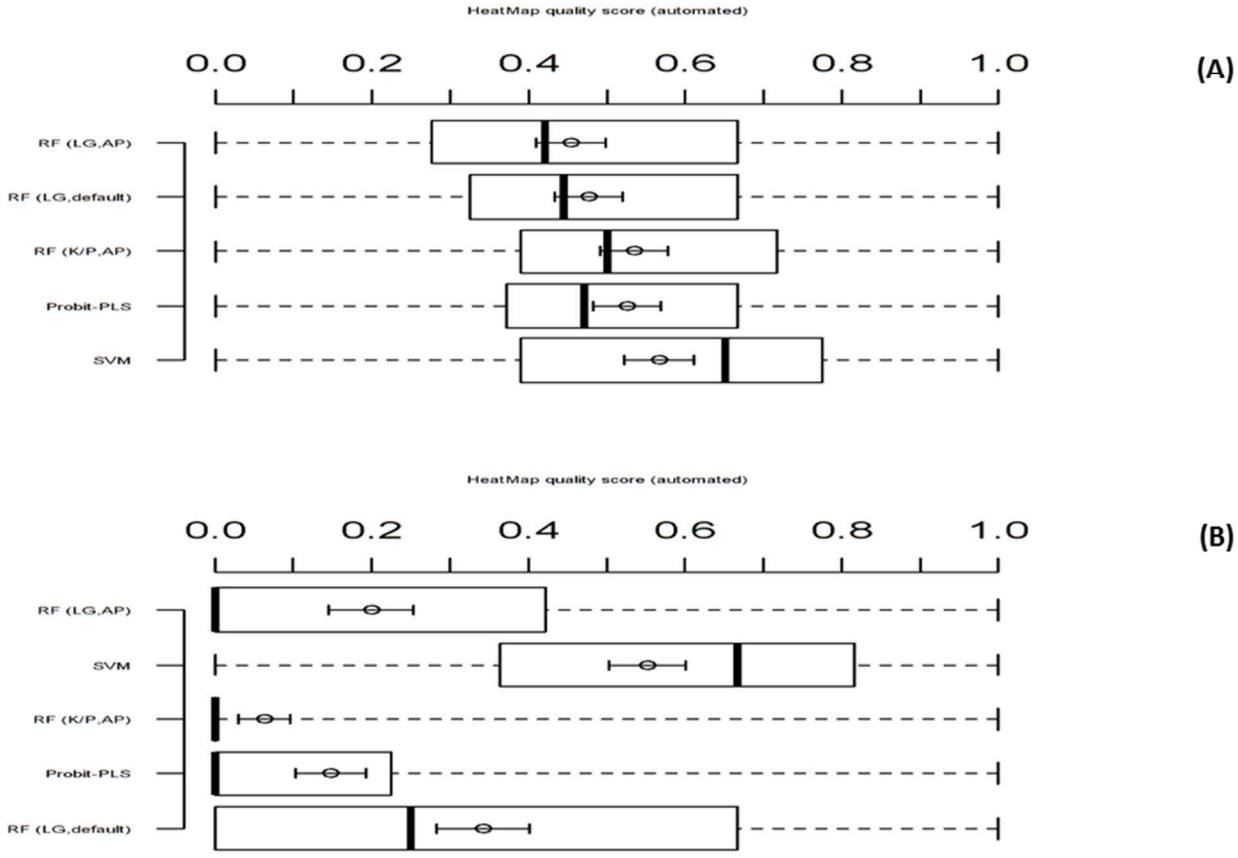


Figure 8 Automated Heat Map numeric quality scores, calculated as the average of F-measures for their correspondence to toxicophore and detoxifying substructure SMARTS matches as explained in **Figure 1**, for Heat Map images generated using atom coloring and (A) Symmetrized Single Molecule Normalization or (B) Constant Symmetrized Normalization. Quality score distributions correspond to all 39 leave-one-out predictions made for a subset of Kazius dataset molecules identified via toxicophore SMARTS matches. The different quality score distributions are annotated according to the different approaches used to obtain the

underlying descriptor contributions: RF (LG, AP) = Random Forest classification (local gradients, averaged predictions); RF (LG, default) = Random Forest classification (local gradients, majority votes predictions); RF (K/P, AP) = Random Forest classification (Kuz'min/Palczewska, averaged predictions); Probit-PLS (linear coefficients); SVM (linear coefficients). The distributions are represented as boxplots, with the bold lines showing the medians and the superimposed circle the arithmetic means, with error bars denoting the standard errors in the means.

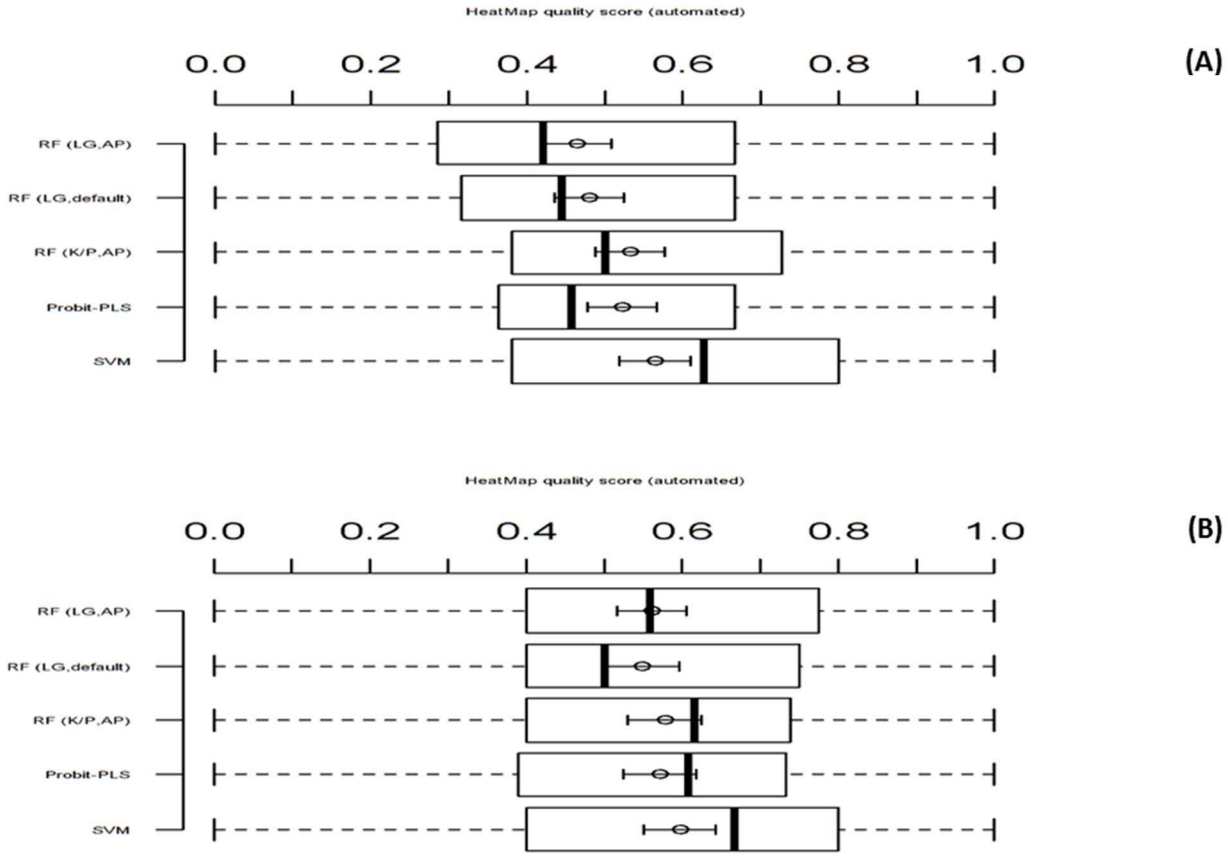


Figure 9 Automated Heat Map numeric quality scores, based on their correspondence to toxicophore substructure SMARTS matches as explained in **Figure 1**, for Heat Map images

generated using atom coloring and Symmetrized Single Molecule Normalization. In sub-figure (A), quality score distributions correspond to 38 leave-one-out predictions made for a subset of Kazius dataset molecules identified via toxicophore SMARTS matches, excluding the single molecule ("1028-11-1") for which a detoxifying substructure was also identified via SMARTS pattern matches. In sub-figure (B), the distributions correspond to the subsets of those 38 molecules which were correctly predicted. The number of molecules correctly predicted to be toxic or non-toxic varied across methods: Random Forest classification (averaged predictions) = 28; Random Forest classification (majority vote predictions) = 29; Probit-PLS = 28; SVM = 31. The different quality scores distributions are annotated according to the different approaches used to obtain the underlying descriptor contributions: RF (LG, AP) = Random Forest classification (local gradients, averaged predictions); RF (LG, default) = Random Forest classification (local gradients, majority votes predictions); RF (K/P, AP) = Random Forest classification (Kuz'min/Palczewska, averaged predictions); Probit-PLS (linear coefficients); SVM (linear coefficients). The distributions are represented as boxplots, with the bold lines showing the medians and the superimposed circle the arithmetic means, with error bars denoting the standard errors in the means.

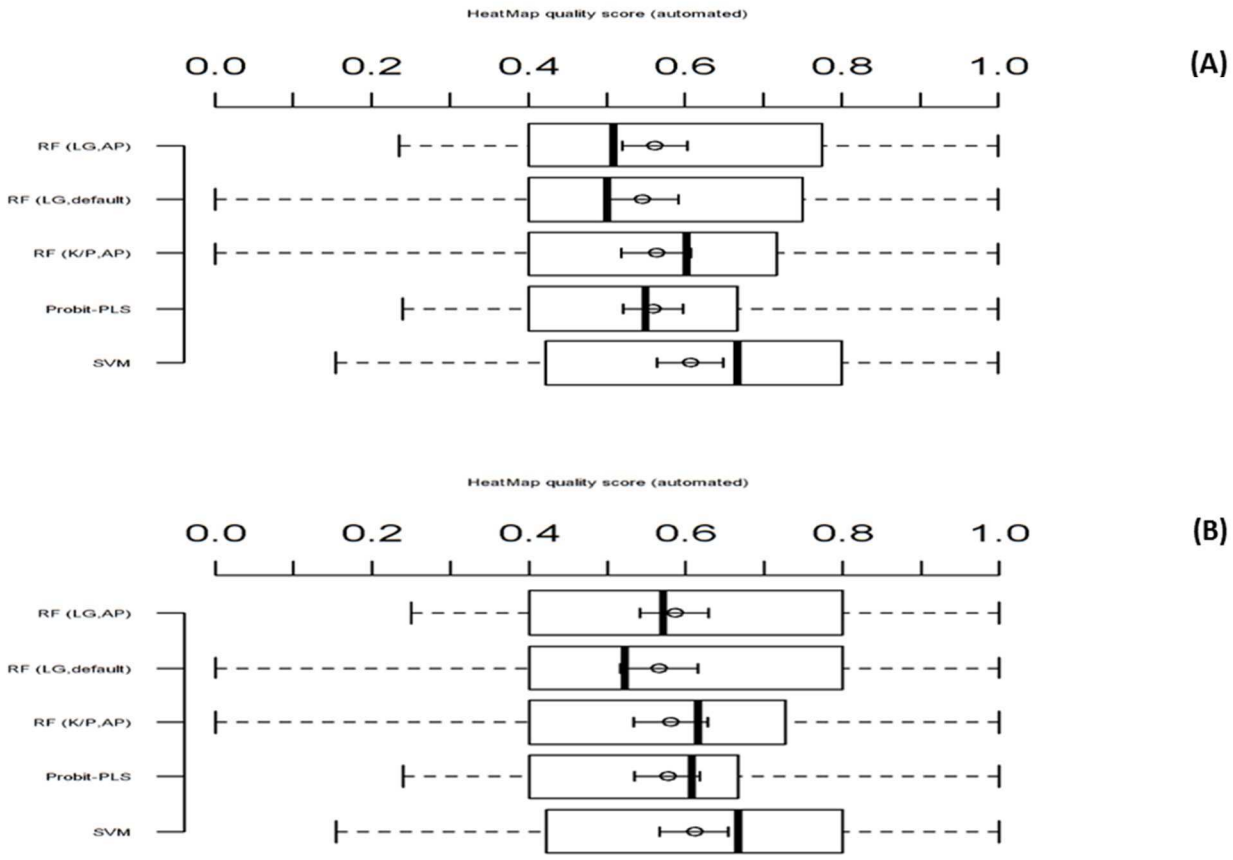


Figure 10 Automated Heat Map numeric quality scores, calculated based on their correspondence to toxicophore SMARTS matches as explained in **Figure 1**, for Heat Map images generated using atom coloring and Symmetrized Single Molecule Normalization. In sub-figure (A), the distributions of quality scores correspond to all toxic leave-one-out predictions made for a subset of Kazius dataset molecules identified via toxicophore SMARTS matches, excluding the single molecule ("1028-11-1") for which a detoxifying substructure was also identified via SMARTS pattern matches. In sub-figure (B), the distributions of quality scores correspond to the subsets of those toxic predictions which were correct. The number of molecules (correctly) predicted to be toxic, hence the number of Heat Map images for which these distributions were generated, varied across methods: Random Forest classification

(averaged predictions) = 28 (25 correct); Random Forest classification (majority vote predictions) = 29 (26 correct); Probit-PLS = 30 (26 correct); SVM = 31 (28 correct). The different quality scores distributions are annotated according to the different approaches used to obtain the underlying descriptor contributions: RF (LG, AP) = Random Forest classification (local gradients, averaged predictions); RF (LG, default) = Random Forest classification (local gradients, majority votes predictions); RF (K/P, AP) = Random Forest classification (Kuz'min/Palczewska, averaged predictions); Probit-PLS (linear coefficients); SVM (linear coefficients). The distributions are represented as boxplots, with the bold lines showing the medians and the superimposed circle the arithmetic means, with error bars denoting the standard errors in the means.

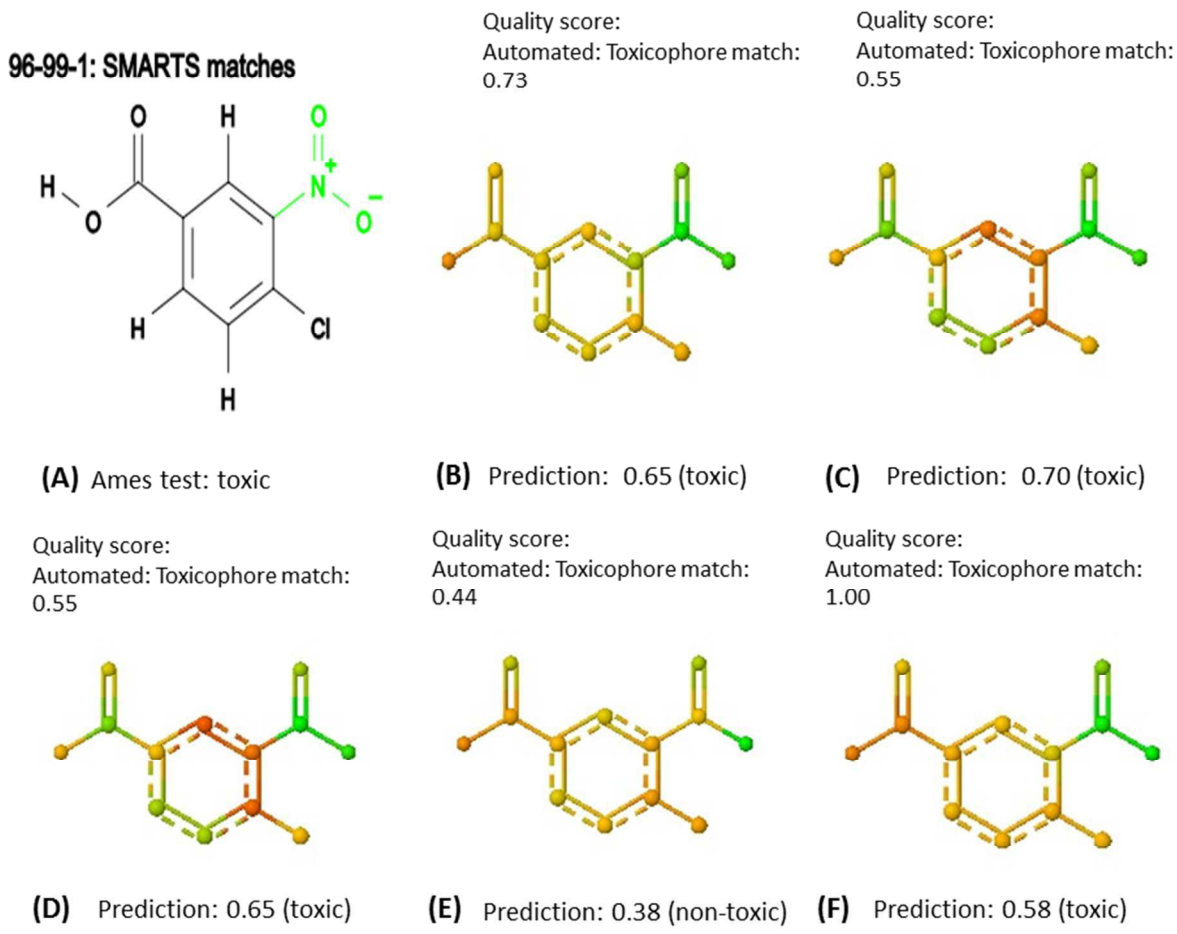


Figure 11 Heat Map (atom coloring & Symmetrized Single Molecule Normalization) analysis of binary classification leave-one-out predictions made for molecule "96-99-1" in the Kazius dataset. (A) Molecule with SMARTS matches shown in green for specific aromatic nitro toxicophore, derived from the work of Kazius et al.[48] based on a combination of mechanistic reasoning, expert knowledge and data analysis, and experimental Ames test assignment reported in the Kazius dataset SDF file.[13] The SMARTS matches (B – F) Heat Map images: (B) Random Forest classification (Kuz'min/Palczewska, averaged predictions); (C) Random Forest classification (local gradients, majority vote predictions); (D) Random Forest (local gradients, averaged predictions); (E) Probit-PLS; (F) SVM. Predictions are reported as per **Figure 4**. The SMARTS matches show the atoms matched to SMARTS patterns. SMARTS matching and

identification of matched atoms were carried out as per **Figure 1**. The automated assessments of Heat Map quality are based on the scheme explained in **Figure 1**. Quality scores, not to be confused with prediction scores, based on Heat Map correspondence to detoxifying groups were not assigned as no detoxifying groups were identified via SMARTS matching.

Effect of prediction quality on Heat Map quality

Figure 9 (B) and **Figure 10** (B) correspond to **Figure 9** (A) and **Figure 10** (A) respectively, after removing incorrect predictions. The corresponding average quality scores are summarized in Supporting Information tables S7 (corresponding to **Figure 9** (A) and **Figure 9** (B)) and S8 (corresponding to **Figure 10** (A) and **Figure 10** (B)). With the exception of SVM toxic predictions, there is an increase in the average numeric quality score, based on Heat Map correspondence to toxicophores, upon moving from all to correct and all toxic to correct toxic predictions. However, especially in the case of toxic predictions, this change is often quite small and the numbers of incorrect predictions are rather small, which may mean this finding is not robust. Nonetheless, these findings do suggest that in the case of an experimentally confirmed prediction, or in light of strong statistical grounds for believing a specific prediction is correct, the substructural cause of toxicity suggested by the (atom coloring, Symmetrized Single Molecule Normalization) Heat Map interpretation of the predictions may be considered more trustworthy. Hence, under these circumstances, the Heat Map image may be considered more useful for guiding lead optimization or suggesting the mechanism of toxic action.

Differences between Random Forest interpretation approaches

Our work revealed that different approaches for interpreting Random Forest predictions may sometimes yield different interpretations. This can be observed by contrasting the Heat Map images (atom coloring, Symmetrized Single Molecule Normalization) corresponding to the same Random Forest classification (averaged predictions) prediction using either the Kuz'min/Palczewska (sub-figure B) or local gradients approach (sub-figure D) in the following figures: **Figure 4**, **Figure 6**, **Figure 7**, **Figure 11**. Consideration of **Figure 12**, which presents raw atom score estimates corresponding to **Figure 11**, emphasizes that this is not merely an artefact of the Symmetrized Single Molecule Normalization scheme: the raw atom scores, derived from the different kinds of descriptor contributions, are clearly not identical.

Consideration of **Figure 12** shows that the estimated influence of specific atoms on the prediction may, indeed, change sign. This could simply reflect the fact that the local gradients approach is not as good as the Kuz'min/Palczewska approach for estimating descriptor influences on the prediction, as may be indicated by consideration of the somewhat lower median Heat Map quality scores shown in **Figure 8** (A), **Figure 9**, **Figure 10**. As to why this should be the case, we speculate that the manner in which local gradients are estimated in our work (see equation (5)) fails to take account of the possibility that switching the value of one binary descriptor might affect the manner in which the molecule of interest is passed down the trees of the forest. This could affect how other descriptors end up contributing to the prediction. This is a potential weakness compared to the Kuz'min/Palczewska approach, which merely takes account of the influence of the specific descriptor of interest on the prediction.

Nonetheless, in spite of these *potential* limitations of the local gradients approach compared to the Kuz'min/Palczweska approach, it should be acknowledged that the correlation between

estimated raw atom scores was typically higher than observed for the example presented in (**Figure 12**). This can be seen by comparing the Pearson's correlation coefficient (0.74) between these two sets of estimated raw atom scores and the distribution of correlation coefficients summarized in **Figure 13** (arithmetic mean = 0.79, median = 0.93).

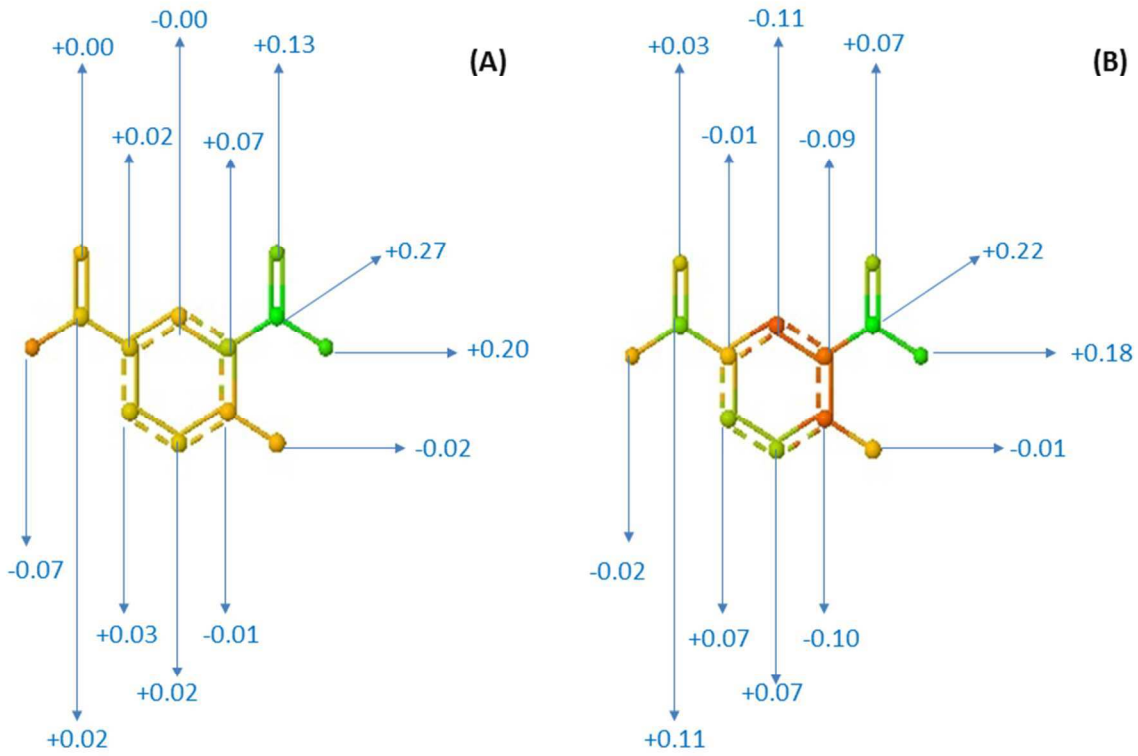


Figure 12 Comparison between the estimated raw atom scores (2dp), for the molecule "96-99-1" in the Kazius dataset, derived from the descriptor contributions corresponding to a single leave-one-out prediction made using Random Forest classification (averaged predictions) according to the Kuz'min/Palczewska (A) and local gradients (B) approaches. The raw scores were estimated

via processing output from the HeatMapWrapper tool[72] in atom coloring and Symmetrized Single Molecule Normalization mode. The correlation between the corresponding raw atom score estimates, in terms of the Pearson's correlation coefficient, is 0.74. The raw atom score estimates are used to annotate the corresponding atoms in the Heat Map (atom coloring & Symmetrized Single Molecule Normalization) image. N.B. The specific atom IDs, used to identify the corresponding atom score estimates, were identified as per **Figure 1**.

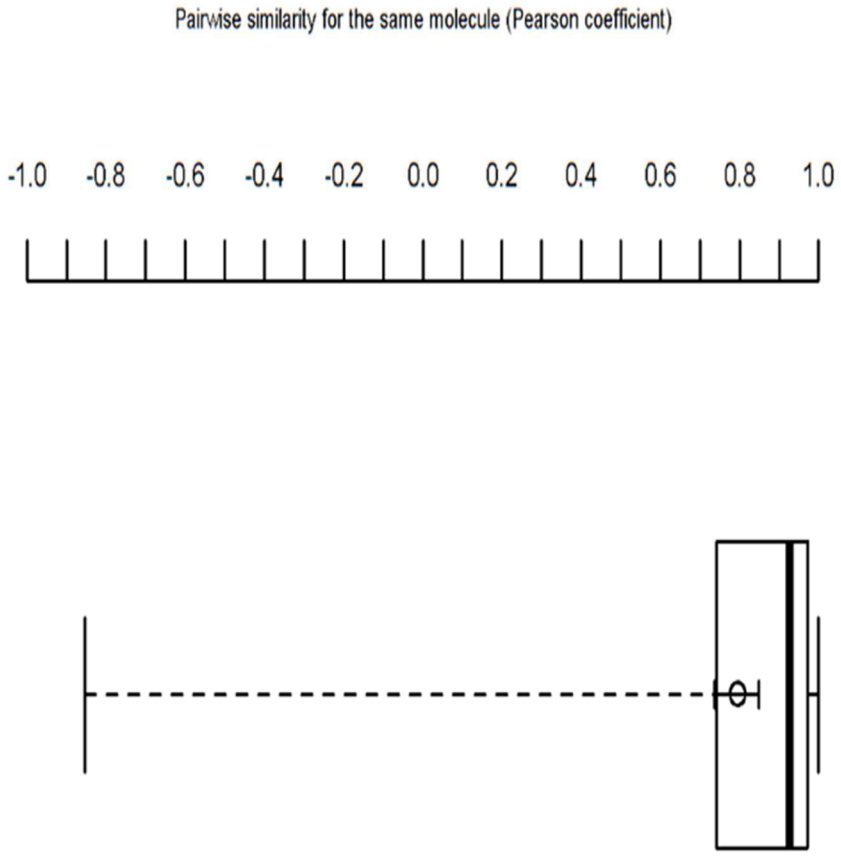


Figure 13 The distribution of pairwise similarities, calculated via the Pearson's correlation coefficient using estimates of the raw atom scores derived from descriptor contributions generated via the Kuz'min/Palczewska and local gradients approaches, for all leave-one-out

Random Forest classification predictions (averaged predictions approach) made for a subset of molecules in the Kazius dataset identified via toxicophore SMARTS pattern matches. (The raw atom score estimates were obtained via processing the output of the HeatMapWrapper program[72] in atom coloring and Symmetrized Single Molecule Normalization mode.) The distribution is represented as a boxplot, with the bold line showing the median and the superimposed circle the arithmetic mean, with error bars denoting the standard error in the mean.

CONCLUSIONS

In this paper, a comparison was presented between the widely used non-linear Random Forest algorithm and well-known linear modelling approaches: Probit-PLS, PLS-regression, Support Vector Machines and Support Vector Regression, with the latter two approaches yielding linear models when employed with a linear kernel function. Specifically, these algorithms were compared in terms of their ability to build predictive models, based on commonly employed extended connectivity fingerprints, for numerical and categorical measures of biological activity on a variety of established benchmark datasets. As well as comparing the predictive performance of the models, using external cross-validation, the interpretability of individual predictions was also assessed.

For Random Forest, different approaches for generating binary classification predictions and for interpreting individual predictions were investigated. To the best of our knowledge, our work represents the first time that methodologically distinct approaches for interpreting predictions obtained from a non-linear QSAR model have been compared. Predictions were interpreted via converting the corresponding substructural descriptor contributions, obtained via different algorithms, into colored Heat Map molecular images, as previously proposed. We demonstrated

that different ways of translating descriptor contributions into Heat Map images can yield different interpretations and we recommend the use of a novel approach to Heat Map generation: atom coloring combined with Symmetrized Single Molecule Normalization. Another novelty of our work is the introduction of systematic assessment schemes for assessing the quality of Heat Map molecular images as a means of assessing the extent to which the corresponding predictions can be interpreted in a chemically and biologically meaningful fashion. These images are ultimately intended to provide information to chemists, hence require manual inspection. However, we advocate the use of our automated, numeric scoring scheme for systematic evaluation of trends in the degree to which Heat Maps generated according to different protocols can be meaningfully interpreted. A manual ranking scheme, applied to an initial set of Heat Map prediction interpretations corresponding to models based upon a variety of datasets, was determined to be too problematic for systematic evaluation of these images. However, due to implementation challenges and the requirement for molecular substructures responsible for (attenuating) biological activity to be encoded as SMARTS patterns, it was only practical to apply the automated scoring scheme to a subset of leave-one-out binary classification predictions obtained for an Ames mutagenicity dataset.

Random Forest classification was observed to produce better predictive performance, on average, for most of the analyzed datasets. The non-default “averaged predictions” approach to generating classification predictions was typically observed to produce better predictive performance, on average, compared to the default “majority votes” approach. Random Forest regression was observed to less consistently outperform the other regression approaches. However, differences in performance were sometimes marginal and may not be robust.

This work emphasizes that predictions obtained from both the non-linear (Random Forest) and linear modelling algorithms considered here *can* be meaningfully interpreted via identifying regions of the molecule expected to be responsible for biological activity. A detailed evaluation was performed of Heat Map images obtained via a variety of interpretation approaches, each interpretation approach corresponding to a different combination of modelling algorithm and, as applicable, method for generating predictions and corresponding descriptor contributions from the model. Our evaluation was based on the distribution of numeric Heat Map quality scores obtained for leave-one-out predictions of Ames mutagenicity, denoting the extent to which those images correspond to the known toxicophores. Considering those distributions, the usefulness of Heat Map images corresponding to linear SVM predictions appeared somewhat higher on average than all other interpretation approaches and the Kuz'min/Palczewska approach to calculating descriptor contributions for Random Forest classification predictions (averaged predictions) yielded Heat Map images with slightly higher average quality scores than the local gradients approach. At least in terms of the median quality scores, a similar pattern was observed upon consideration of toxic predictions, for which Heat Map ability to identify toxicophores would be useful for medicinal chemists seeking to design out a toxic liability during lead optimization. These trends were also observed when only correct, or correct toxic, predictions were considered.

Improved prediction quality seemed to typically correspond to improved interpretability, in terms of the correspondence between the Heat Map images and known toxicophores, based upon considering the same set of leave-one-out predictions of Ames mutagenicity. However, increases in average Heat Map quality scores were typically small, upon moving from all (toxic) to correct (toxic) predictions and only a small number of incorrect (toxic) predictions were considered.

Nonetheless, these findings suggest that, if the corresponding predictions were experimentally verified, there were even stronger grounds for believing the Heat Map images could be trusted to offer at least some insight into an unknown mechanism of action or opportunities for lead optimization. This also suggests that, where there are statistical grounds for high confidence in the prediction of an untested molecule, there are stronger grounds for drawing conclusions from the corresponding Heat Map image.

Overall, we can advocate Random Forest as a suitable option for QSAR modelers interested in both accurate and interpretable predictions. In keeping with the literature, our results suggest Random Forest may yield better predictive performance than linear methods. They also indicate that our Open Source *rfFC* and *HeatMapWrapper* software tools may allow for chemically and biologically meaningful interpretations of Random Forest predictions. For the purpose of generating Heat Map molecular interpretations of those predictions, we advocate the use of atom coloring, in combination with our Symmetrized Single Molecule Normalization scheme, and tentatively suggest the Kuz'min/Palczewska approach may be preferable to the local gradients approach considered in our work.

Supporting Information

Supporting Information Available: (1) "SupportingInformation_doc_JCIM_rev.2.ret.pdf" contains three sections: (A) describes the methods used and computational analysis in greater detail; (B) explains how we generated our results; (C) summarizes additional results, including raw results provided in "additional.results.resub.fixed.zip". (2) "FinalImagesAnalysis_withImages_resub.xlsx" presents the full analysis of all Heat Map images and corresponding predictions. (3) "repeating_ligand_bioactivity_searches.xlsx" documents how

experimental pK_i and pIC_{50} values were obtained to derive the values reported in "FinalImagesAnalysis_withImages_resub.xlsx". (4) "ligands_released.zip" presents copies of the ligand files used to generate Heat Map images. (5) "additional.results.resub.fixed.zip" provides additional raw results files. (6) "Kazius_2005_SI_SA_SMARTS_subset_v2.xls" provides the SMARTS patterns for toxicophores and detoxifying moieties used for automated, numeric scoring of Kazius LOO Heat Map images. This material is available free of charge via the Internet at http://pubs.acs.org.

Author Contributions

RLMR conceived of the project, designed the computational experiments, wrote and executed the necessary scripts, analyzed the results and wrote the first draft of the manuscript. NK helped design the Pipeline Pilot workflow for standardizing molecular structures and prepared all ligand SDF files from PDB files for analysis using Schrödinger's Protein Preparation Wizard. AP and JP provided feedback which improved the analysis carried out by RLMR and assisted with writing some scripts. AP developed the *rfFC* package, with some conceptual input from RLMR and JP. RLMR wrote the *HeatMapWrapper* Python tool used for generating Heat Map images analyzed in the current publication. JP assisted with the design and implementation of the Symmetrized Single Molecule Normalization scheme. The final manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding Sources

N/A

Notes

N/A

ACKNOWLEDGEMENT

RLMR thanks Kim Travis (Syngenta Ltd.), Mark Earll (Syngenta Ltd.) and Mark Forster (Syngenta Ltd.) for useful discussions. RLMR thanks Lars Rosenbaum (University of Tübingen) for providing guidance on how to generate the Heat Map images reported in this publication. RLMR also thanks Andreas Zell (University of Tübingen) for further helpful correspondence regarding the Heat Map software. RLMR thanks Johannes Mohr (Technische Universität Berlin) for valuable correspondence regarding the CA dataset. RLMR thanks Michael Mackay (Liverpool John Moores University) for providing access to a virtual machine used to run cross-validation calculations. RLMR thanks Rachel Kramer Green (RCSB PDB Leadership Team), Heather Carlson (University of Michigan) and Renxiao Wang (Shanghai Institute of Organic Chemistry) for helpful correspondence regarding the PDB, BindingMOAD and PDBbind databases respectively. RLMR thanks Knut Baumann (Technische Universität Braunschweig) for providing access to the MUV datasets. The authors thank Lucy Entwistle (Syngenta Ltd.) for writing the Pipeline Pilot workflow used to process all molecules prior to fingerprint calculations.

ABBREVIATIONS

N/A

REFERENCES

(1) *Guidance Document on the Validation of (Quantitative)Structure-Activity Relationships [(Q)SAR] Models (ENV/JM/MONO(2007)2)*; OECD Series on Testing and Assessment; 69; Organisation for Economic Co-operation and Development, 2007.

(2) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; et al. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57*, 4977–5010.

(3) Gavaghan, C. L.; Arnby, C. H.; Blomberg, N.; Strandlund, G.; Boyer, S. Development, Interpretation and Temporal Evaluation of a Global QSAR of hERG Electrophysiology Screening Data. *J. Comput.-Aided Mol. Des.* **2007**, *21* (4), 189–206.

(4) Hansch, C.; Fujita, T. P-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86* (8), 1616–1626.

(5) Baskin, I. I.; Ait, A. O.; Halberstam, N. M.; Palyulin, V. A.; Zefirov, N. S. An Approach to the Interpretation of Backpropagation Neural Network Models in QSAR Studies. *SAR QSAR Environ. Res.* **2002**, *13* (1), 35–41.

(6) Johnson, S. R.; Chen, Q. X.; Murphy, D.; Gudmunsson, O. A Computational Model for the Prediction of Aqueous Solubility That Includes Crystal Packing, Intrinsic Solubility, and Ionization Effects. *Mol. Pharmaceutics* **2007**, *4* (4), 513–523.

(7) Dearden, J. C.; Cronin, M. T. D.; Kaiser, K. L. E. How Not to Develop a Quantitative Structure–activity or Structure–property Relationship (QSAR/QSPR). *SAR QSAR Environ. Res.* **2009**, *20* (3–4), 241–266.

(8) Jensen, F. Chapter 17: Statistics and QSAR. In *Introduction to Computational Chemistry*; John Wiley & Sons Ltd, 2007; pp 547–561.

(9) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 1947–1958.

(10) Muller, K.; Mika, S.; Ratsch, G.; Tsuda, K.; Scholkopf, B. An Introduction to Kernel-Based Learning Algorithms. *IEEE Transactions on Neural Networks* **2001**, *12* (2), 181–201.

(11) Hsu, C.-W.; Chang, C.-C.; Lin, C.-J. *A Practical Guide to Support Vector Classification*; Department of Computer Science, National Taiwan University: Taiwan, 2010.

(12) Smola, A. J.; Schölkopf, B. A Tutorial on Support Vector Regression. *Statistics and Computing* **2004**, *14* (3), 199–222.

(13) Rosenbaum, L.; Hinselmann, G.; Jahn, A.; Zell, A. Interpreting Linear Support Vector Machine Models with Heat Map Molecule Coloring. *J. Cheminf.* **2011**, *3* (1), 11.

(14) Song, M.; Clark, M. Development and Evaluation of an in Silico Model for hERG Binding. *J. Chem. Inf. Model* **2006**, *46* (1), 392–400.

(15) Mitchell, J. B. O. Machine Learning Methods in Chemoinformatics. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2014**, *4*, 468–481.

(16) Guha, R. On the Interpretation and Interpretability of Quantitative Structure–activity Relationship Models. *J. Comput.-Aided Mol. Des.* **2008**, *22* (12), 857–871.

(17) Mohr, J.; Jain, B.; Sutter, A.; Laak, A. T.; Steger-Hartmann, T.; Heinrich, N.; Obermayer, K. A Maximum Common Subgraph Kernel Method for Predicting the Chromosome Aberration Test. *J. Chem. Inf. Model.* **2010**, *50* (10), 1821–1838.

(18) Carlsson, L.; Helgee, E. A.; Boyer, S. Interpretation of Nonlinear QSAR Models Applied to Ames Mutagenicity Data. *J. Chem. Inf. Model.* **2009**, *49* (11), 2551–2558.

(19) Kuz'min, V. E.; Polishchuk, P. G.; Artemenko, A. G.; Andronati, S. A. Interpretation of QSAR Models Based on Random Forest Methods. *Mol. Inf.* **2011**, *30* (6-7), 593–603.
(20) An, Y.; Sherman, W.; Dixon, S. L. Kernel-Based Partial Least Squares: Application to Fingerprint-Based QSAR with Model Visualization. *J. Chem. Inf. Model.* **2013**, *53*, 2312–2321.
(21) Polishchuk, P. G.; Kuz'min, V. E.; Artemenko, A. G.; Muratov, E. N. Universal Approach for Structural Interpretation of QSAR/QSPR Models. *Mol. Inf.* **2013**, *32* (9–10), 843–853.
(22) Webb, S. J.; Hanser, T.; Howlin, B.; Krause, P.; Vessey, J. D. Feature Combination Networks for the Interpretation of Statistical Machine Learning Models: Application to Ames Mutagenicity. *J. Cheminf.* **2014**, *6* (1), 8.
(23) Welling, S. H.; Clemmensen, L. K. H.; Buckley, S. T.; Hovgaard, L.; Brockhoff, P. B.; Refsgaard, H. H. F. In Silico Modelling of Permeation Enhancement Potency in Caco-2 Monolayers Based on Molecular Descriptors and Random Forest. *Eur. J. Pharm. Biopharm.* **2015**, *94*, 152–159.
(24) Balfer, J.; Bajorath, J. Visualization and Interpretation of Support Vector Machine Activity Predictions. *J. Chem. Inf. Model.* **2015**, *55* (6), 1136–1147.
(25) Polishchuk, P.; Tinkov, O.; Khristova, T.; Ognichenko, L.; Kosinskaya, A.; Varnek, A.; Kuz'min, V. Structural and Physico-Chemical Interpretation (SPCI) of QSAR Models and Its Comparison with Matched Molecular Pair Analysis. *J. Chem. Inf. Model.* **2016**, *56* (8), 1455–1469.
(26) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32.
(27) Marchese Robinson, R. L.; Glen, R. C.; Mitchell, J. B. O. Development and Comparison of hERG Blocker Classifiers: Assessment on Different Datasets Yields Markedly Different Results. *Mol. Inf.* **2011**, *30*, 443–458.
(28) Strobl, C.; Boulesteix, A.-L.; Zeileis, A.; Hothorn, T. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinf.* **2007**, *8* (1), 25.
(29) Palczewska, A.; Palczewski, J.; Robinson, R. M.; Neagu, D. Interpreting Random Forest Models Using a Feature Contribution Method. In *2013 IEEE 14th International Conference on Information Reuse and Integration (IRI)*; 2013; pp 112–119.
(30) Palczewska, A.; Palczewski, J.; Robinson, R. M.; Neagu, D. Interpreting Random Forest Classification Models Using a Feature Contribution Method. In *Integration of Reusable Systems*; Bouabana-Tebibel, T., Rubin, S. H., Eds.; Advances in Intelligent Systems and Computing; Springer International Publishing, 2014; pp 193–218.
(31) Welling, S. H.; Refsgaard, H. H. F.; Brockhoff, P. B.; Clemmensen, L. H. Forest Floor Visualizations of Random Forests. *arXiv.org, e-Print Arch., Statistics* **2016**.
(32) Cassano, A.; Marchese Robinson, R. L.; Palczewska, A.; Puzyn, T.; Gajewicz, A.; Tran, L.; Manganelli, S.; Cronin, M. T. D. Comparing the CORAL and Random Forest Approaches for Modelling the in Vitro Cytotoxicity of Silica Nanomaterials. *ATLA, Altern. Lab. Anim.* **2016**, *44* (6), 533–556.
(33) Stålring, J.; Almeida, P. R.; Carlsson, L.; Helgee Ahlberg, E.; Hasselgren, C.; Boyer, S. Localized Heuristic Inverse Quantitative Structure Activity Relationship with Bulk Descriptors Using Numerical Gradients. *J. Chem. Inf. Model.* **2013**, *53* (8), 2001–2017.
(34) Mevik, B.-H.; Wehrens, R. The Pls Package: Principal Component and Partial Least Squares Regression in R. *J Stat Softw* **2007**, *18* (2).
(35) Obrezanova, O.; Segall, M. D. Gaussian Processes for Classification: QSAR Modeling of ADMET and Target Activity. *J. Chem. Inf. Model.* **2010**, *50* (6), 1053–1061.

(36) Agresti, A. Generalized Linear Models for Binary Data. In *An Introduction to Categorical Data Analysis*; Wiley Series in Probability and Statistics; WILEY-INTERSCIENCE, 2007; pp 68–72.
(37) R-Forge: randomForest Feature Contribution: R Development Page (rfFC package repository) https://r-forge.r-project.org/R/?group_id=1725 (accessed Apr 30, 2016).
(38) Liu, Y.; Chawla, N. V.; Harper, M. P.; Shriberg, E.; Stolcke, A. A Study in Machine Learning from Imbalanced Data for Sentence Boundary Detection in Speech. *Comput. Speech Lang.* **2006**, *20* (4), 468–494.
(39) Braga-Neto, U. M.; Dougherty, E. R. Is Cross-Validation Valid for Small-Sample Microarray Classification? *Bioinformatics* **2004**, *20* (3), 374–380.
(40) Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A. F.; Nielsen, H. Assessing the Accuracy of Prediction Algorithms for Classification: An Overview. *Bioinformatics*. **2000**, *16* (5), 412–424.
(41) Gorodkin, J. Comparing Two K-Category Assignments by a K-Category Correlation Coefficient. *Comput. Biol. Chem.* **2004**, *28* (5–6), 367–374.
(42) Konovalov, D. A.; Llewellyn, L. E.; Vander Heyden, Y.; Coomans, D. Robust Cross-Validation of Linear Regression QSAR Models. *J. Chem. Inf. Model.* **2008**, *48* (10), 2081–2094.
(43) Alexander, D. L. J.; Tropsha, A.; Winkler, D. A. Beware of R2: Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *J. Chem. Inf. Model.* **2015**, *55* (7), 1316–1322.
(44) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model* **2010**, *50* (5), 742–754.
(45) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97–101.
(46) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. A Comparison of Methods for Modeling Quantitative Structure−Activity Relationships. *J. Med. Chem.* **2004**, *47* (22), 5541–5554.
(47) Rohrer, S. G.; Baumann, K. Maximum Unbiased Validation (MUV) Data Sets for Virtual Screening Based on PubChem Bioactivity Data. *J. Chem. Inf. Model.* **2009**, *49* (2), 169–184.
(48) Kazius, J.; McGuire, R.; Bursi, R. Derivation and Validation of Toxicophores for Mutagenicity Prediction. *J. Med. Chem.* **2005**, *48* (1), 312–320.
(49) Przybylak, K. R.; Madden, J. C.; Cronin, M. T. D.; Hewitt, M. Assessing Toxicological Data Quality: Basic Principles, Existing Schemes and Current Limitations. *SAR QSAR Environ. Res.* **2012**, *23* (5–6), 435–459.
(50) Hinselmann, G.; Rosenbaum, L.; Jahn, A.; Fechner, N.; Zell, A. jCompoundMapper: An Open Source Java Library and Command-Line Tool for Chemical Fingerprints. *J Cheminf* **2011**, *3* (1), 3.
(51) Hinselmann, G.; Rosenbaum, L.; Jahn, A.; Fechner, N.; Ostermann, C.; Zell, A. Large-Scale Learning of Structure−Activity Relationships Using a Linear Support Vector Machine and Problem-Specific Metrics. *J. Chem. Inf. Model* **2011**, *51* (2), 203–213.
(52) Maunz, A.; Gütlein, M.; Rautenberg, M.; Vorgrimmler, D.; Gebele, D.; Helma, C. Lazar: A Modular Predictive Toxicology Framework. *Front. Pharmacol* **2013**, *4*, 38.
(53) Manchester, J.; Czerminski, R. CAUTION: Popular "Benchmark" Data Sets Do Not Distinguish the Merits of 3D QSAR Methods. *J. Chem. Inf. Model.* **2009**, *49* (6), 1449–1454.

(54) DePriest, S. A.; Mayer, D.; Naylor, C. B.; Marshall, G. R. 3D-QSAR of Angiotensin-Converting Enzyme and Thermolysin Inhibitors: A Comparison of CoMFA Models Based on Deduced and Experimentally Determined Active Site Geometries. *J. Am. Chem. Soc.* **1993**, *115* (13), 5372–5384.

(55) Böhm, M.; Stürzebecher, J.; Klebe, G. Three-Dimensional Quantitative Structure−Activity Relationship Analyses Using Comparative Molecular Field Analysis and Comparative Molecular Similarity Indices Analysis To Elucidate Selectivity Differences of Inhibitors Binding to Trypsin, Thrombin, and Factor Xa. *J. Med. Chem.* **1999**, *42* (3), 458–477.

(56) Gohlke, H.; Klebe, G. DrugScore Meets CoMFA: Adaptation of Fields for Molecular Comparison (AFMoC) or How to Tailor Knowledge-Based Pair-Potentials to a Particular Protein. *J. Med. Chem.* **2002**, *45* (19), 4153–4170.

(57) BIOVIA Pipeline Pilot | Scientific Workflow Authoring Application for Data Analysis http://accelrys.com/products/collaborative-science/biovia-pipeline-pilot/ (accessed May 2, 2016).

(58) Low, Y.; Uehara, T.; Minowa, Y.; Yamada, H.; Ohno, Y.; Urushidani, T.; Sedykh, A.; Muratov, E.; Kuz'min, V.; Fourches, D.; et al. Predicting Drug-Induced Hepatotoxicity Using QSAR and Toxicogenomics Approaches. *Chem. Res. Toxicol.* **2011**, *24*, 1251–1262.

(59) Cohen, J. Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. *Psychol. Bull.* **1968**, *70* (4), 213–220.

(60) Sing, T.; Sander, O.; Beerenwinkel, N.; Lengauer, T. ROCR: Visualizing Classifier Performance in R. *Bioinformatics* **2005**, *21* (20), 3940–3941.

(61) Berrar, D.; Flach, P. Caveats and Pitfalls of ROC Analysis in Clinical Microarray Research (and How to Avoid Them). *Briefings Bioinf.* **2012**, *13* (1), 83–97.

(62) Anger, L. T.; Wolf, A.; Schleifer, K.-J.; Schrenk, D.; Rohrer, S. G. Generalized Workflow for Generating Highly Predictive in Silico Off-Target Activity Models. *J. Chem. Inf. Model.* **2014**, *54* (9), 2411–2422.

(63) RCSB Protein Data Bank - RCSB PDB http://www.rcsb.org/pdb/home/home.do (accessed Feb 4, 2016).

(64) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucl. Acids Res.* **2000**, *28* (1), 235–242.

(65) Stierand, K.; Rarey, M. Drawing the PDB: Protein−Ligand Complexes in Two Dimensions. *ACS Med. Chem. Lett.* **2010**, *1* (9), 540–545.

(66) Bissantz, C.; Kuhn, B.; Stahl, M. A Medicinal Chemist's Guide to Molecular Interactions. *J. Med. Chem.* **2010**, *53* (14), 5061–5084.

(67) Cannon, E. O.; Nigsch, F.; Mitchell, J. B. A Novel Hybrid Ultrafast Shape Descriptor Method for Use in Virtual Screening. *Chem. Cent. J.* **2008**, *2*, 3–3.

(68) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminf.* **2011**, *3* (1), 33.

(69) O'Boyle, N.; Morley, C.; Hutchison, G. Pybel: A Python Wrapper for the OpenBabel Cheminformatics Toolkit. *Chem. Cent. J.* **2008**, *2* (1), 5.

(70) Marchese Robinson, Richard. Scripts for "Comparison of the predictive performance and interpretability of Random Forest and linear models on benchmark datasets" [revision 2.0] https://doi.org/10.5281/zenodo.824245 (accessed Jul 7, 2017).

(71) The R Project for Statistical Computing http://www.r-project.org/ (accessed Oct 18, 2016).

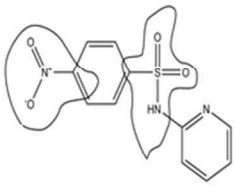
(72) Marchese Robinson, R. L. HeatMapWrapper (version 0.16) https://doi.org/10.5281/zenodo.495163 (accessed Apr 5, 2017).
(73) Benjamini, Y.; Yekutieli, D. The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics* **2001**, *29* (4), 1165–1188.
(74) Bengio, Y.; Grandvalet, Y. No Unbiased Estimator of the Variance of K-Fold Cross-Validation. *J. Mach. Learn. Res.* **2004**, *5*, 1089–1105.
(75) Nadeau, C.; Bengio, Y. Inference for the Generalization Error. *Mach. Learn.* **2003**, *52* (3), 239–281.
(76) Steiner, T. The Hydrogen Bond in the Solid State. *Angew. Chem., Int. Ed.* **2002**, *41* (1), 48–76.

for Table of Contents use only

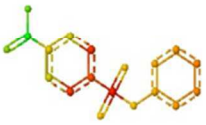
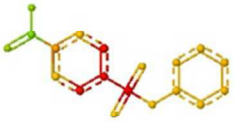
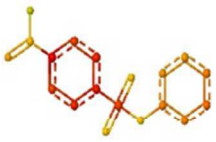
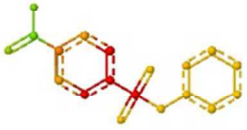
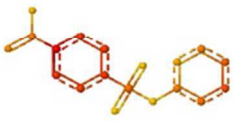
Comparison of the predictive performance and interpretability of Random Forest and linear models on benchmark datasets

Richard L. Marchese Robinson, Anna Palczewska, Jan Palczewski, Nathan Kidley

Non-toxic molecule with toxicophore (LHS) and detoxifying substructure (RHS) highlighted



How good are these QSAR predictions?
How meaningful are their interpretations?

Random Forest		***Linear models***	
Interpretation approach 1		Probit-PLS	
Interpretation approach 2(a)		SVM (linear kernel)	
Interpretation approach 2(b)			

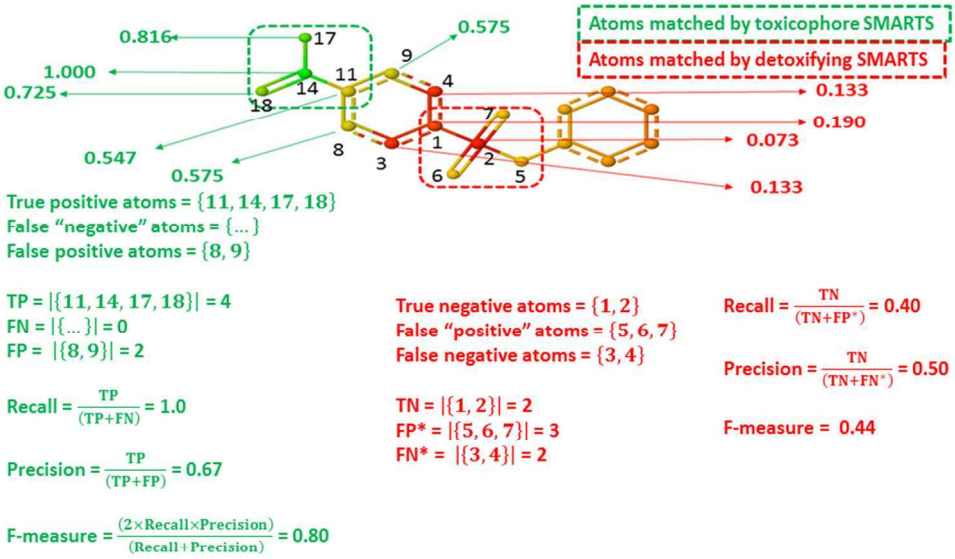


Figure 1 The procedure for calculating numeric Heat Map quality scores for an image obtained via symmetrized normalization (Symmetrized Single Molecule Normalization in this example) and atom coloring. The relevant normalized atom score estimates are shown matched to the atom IDs. In this case, the overall quality score would be the arithmetic mean of the two F-measures i.e. 0.62. Further details are provided in the associated text. SMARTS matching of toxicophores and detoxifying substructures was implemented using OpenBabel (version 2.3.2) & Pybel,68,69 based on SMARTS adapted from the work of Kazius et al.48 and reported in the Supporting Information. The atom IDs, used to identify the corresponding atom score estimates for this image, were identified using OpenBabel68, via this command: "obabel [SDF containing molecule of interest] -O [name of image].svg -xi –xu".

254x190mm (96 x 96 DPI)

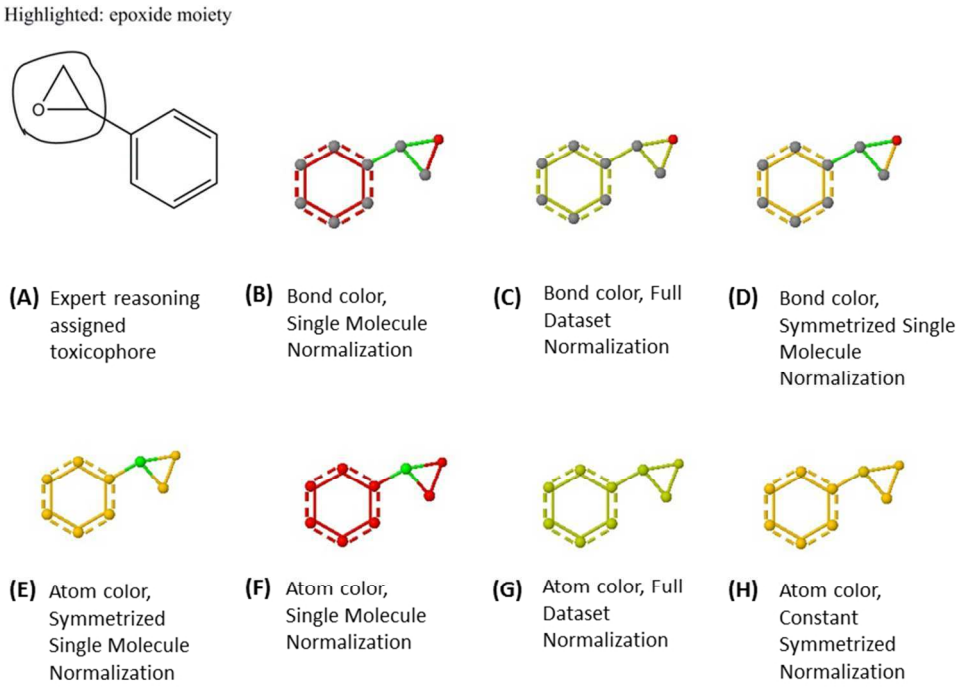


Figure 2 Heat Map analysis of a binary classification leave-one-out SVM prediction made for molecule "CA23" in the CA dataset (c.f. Figure 4(b) in Mohr et al.), determined to be toxic (positive result) in the chromosome aberration test.17 (A) Expert reasoning assigned toxicophore (epoxide functionality), based on the rationale provided in Mohr et al.:17 this substructure was flagged by an Ames test alerts knowledgebase, with 80% of Ames positives being chromosome aberration positives, and is known to exhibit electrophilic reactivity with DNA. (B – H): Heat Map images based on descriptor contributions associated with the same SVM prediction, generated according to either bond or atom coloring obtained via different normalization schemes.

254x190mm (96 x 96 DPI)

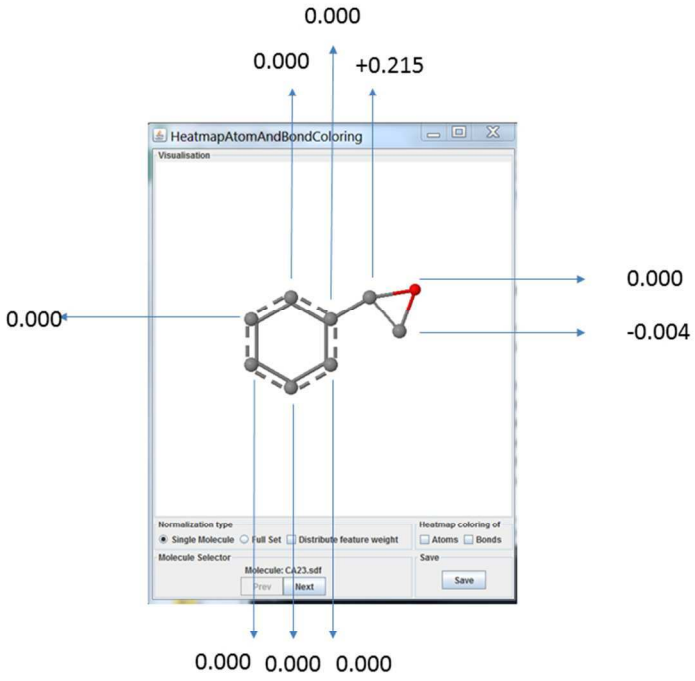


Figure 3 Image of the molecule "CA23" in the CA dataset17 displayed in HeatmapViewer.jar GUI,13 as generated by running the HeatMapWrapper tool,72 with both atom and bond coloring switched off. The intermediate output generated via running the HeatMapWrapper tool in atom color, Symmetrized Single Molecule Normalization mode, based on leave-one-out SVM descriptor contributions, was processed to reveal estimated raw atom scores ranging from weakly negative in one case (-0.004), through zero in most cases, to positive (+0.215) in one case. The estimated raw atom scores are rounded to 3dp. This image corresponds to all Heat Map images shown in Figure 2, with different normalization schemes generating the different normalized scores used for coloring. Via switching off atom coloring, based on normalized scores, the epoxide oxygen (in red) and carbon atoms (in black) can readily be discerned. N.B. The specific atom IDs, used to identify the corresponding atom score estimates, were identified as per Figure 1.

254x190mm (96 x 96 DPI)

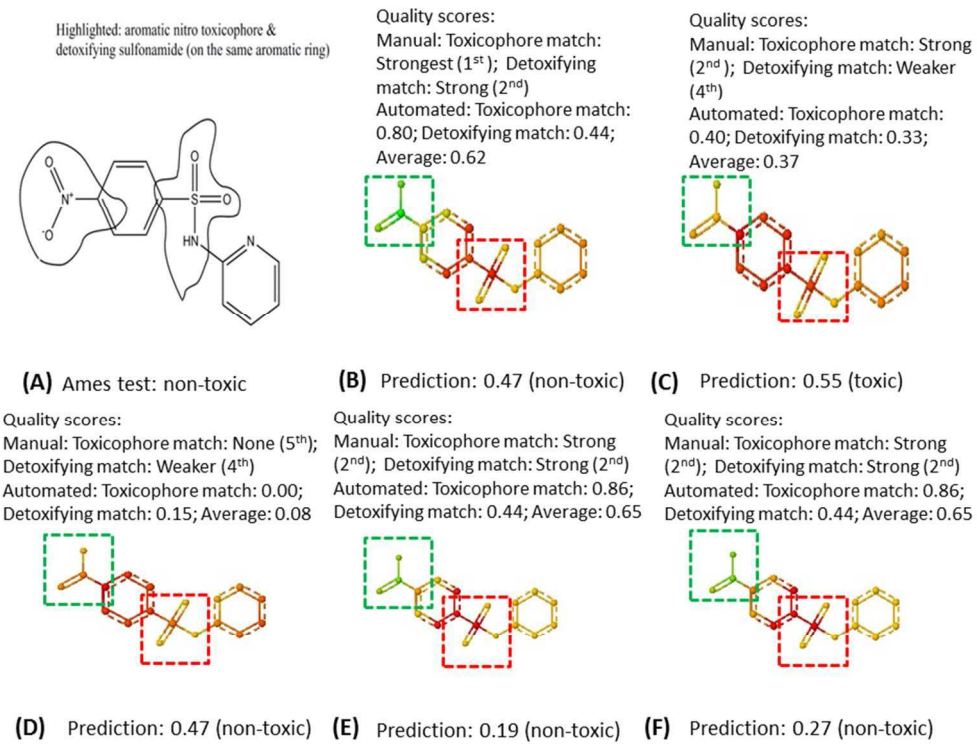


Figure 4 Heat Map (atom coloring & Symmetrized Single Molecule Normalization) analysis of binary classification leave-one-out predictions made for molecule "1028-11-1" in the Kazius dataset (c.f. Figure 3(A) in Rosenbaum et al.).13 (A) Molecule with biologically significant substructures shown, based on a combination of mechanistic reasoning, expert knowledge and data analysis reported previously in the literature,13,48 and experimental Ames test assignment reported.13 (B – F) Heat Map images: (B) Random Forest classification (Kuz'min/Palczewska, averaged predictions); (C) Random Forest classification (local gradients, majority vote predictions); (D) Random Forest (local gradients, averaged predictions); (E) Probit-PLS; (F) SVM. The predictions shown are the normalized scores (transformed into values between 0 and 1) for the toxic class, with predicted assignment to the toxic class if this score was greater than 0.50. The green dotted lines show the atoms matched to toxicophore SMARTS and the red dotted lines the atoms matched to corresponding detoxifying SMARTS. SMARTS matching and identification of matched atoms were carried out as per Figure 1. The manual and automated assessments of Heat Map quality are based on the schemes explained under "Quality assessment of Heat Map images".

254x190mm (96 x 96 DPI)

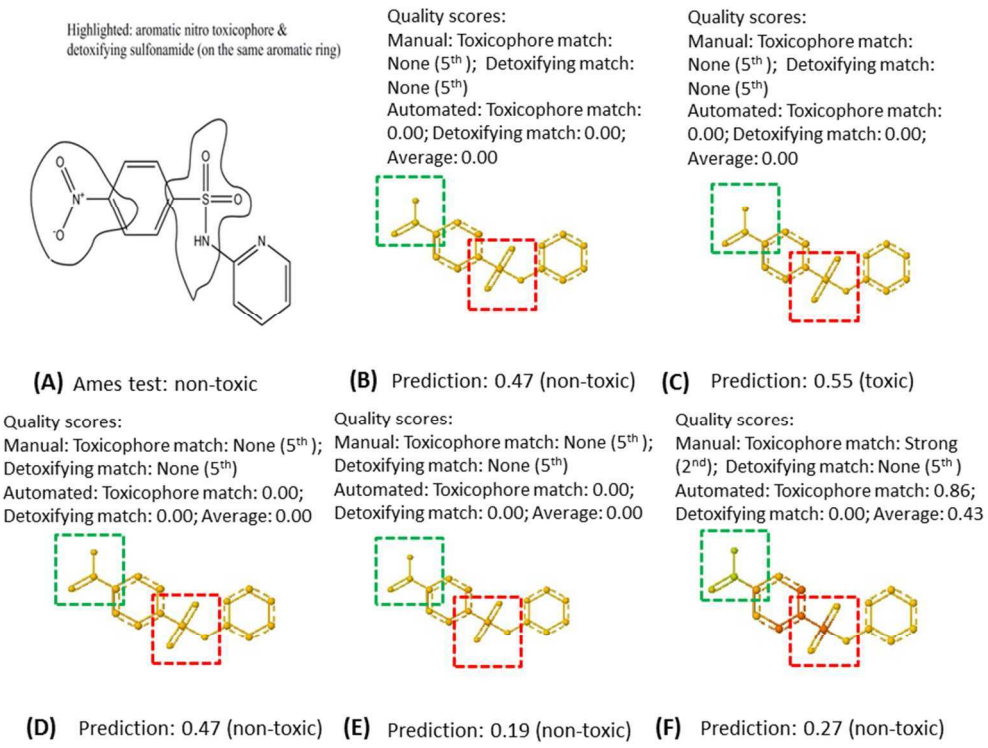


Figure 5 Heat Map (atom coloring & Constant Symmetrized Normalization) analysis of binary classification leave-one-out predictions made for molecule "1028-11-1" in the Kazius dataset (c.f. Figure 3(A) in Rosenbaum et al.).13 (A) Molecule with biologically significant substructures shown, based on a combination of mechanistic reasoning, expert knowledge and data analysis reported previously in the literature,13,48 and experimental Ames test assignment reported.13 (B – F) Heat Map images: (B) Random Forest classification (Kuz'min/Palczewska, averaged predictions); (C) Random Forest classification (local gradients, majority vote predictions); (D) Random Forest (local gradients, averaged predictions); (E) Probit-PLS; (F) SVM. Predictions are reported as per Figure 4. The green dotted lines show the atoms matched to toxicophore SMARTS and the red dotted lines the atoms matched to corresponding detoxifying SMARTS. SMARTS matching and identification of matched atoms were carried out as per Figure 1. The manual and automated assessments of Heat Map quality are based on the schemes explained under "Quality assessment of Heat Map images".

254x190mm (96 x 96 DPI)

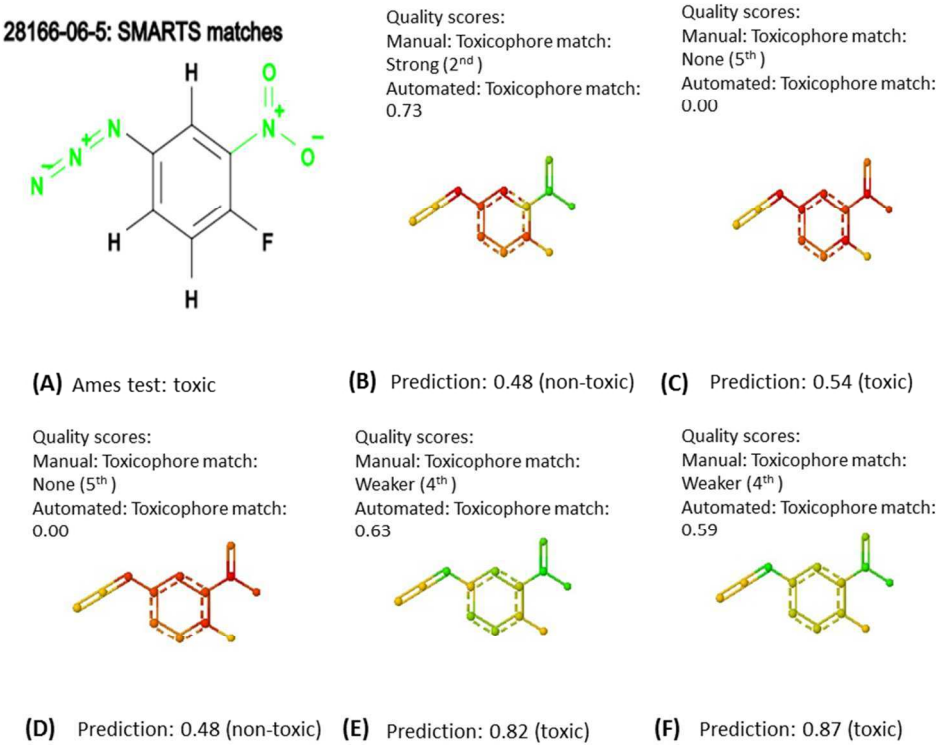


Figure 6 Heat Map (atom coloring & Symmetrized Single Molecule Normalization) analysis of binary classification leave-one-out predictions made for molecule "28166-06-5" in the Kazius dataset. (A) Molecule with SMARTS matches shown in green for specific aromatic nitro (RHS) and azide (LHS) toxicophores, derived from the work of Kazius et al.48 based on a combination of mechanistic reasoning, expert knowledge and data analysis reported, and experimental Ames test assignment reported in the Kazius dataset SDF file.13 The SMARTS matches (B – F) Heat Map images: (B) Random Forest classification (Kuz'min/Palczewska, averaged predictions); (C) Random Forest classification (local gradients, majority vote predictions); (D) Random Forest (local gradients, averaged predictions); (E) Probit-PLS; (F) SVM. Predictions are reported as per Figure 4. The SMARTS matches show the atoms matched to SMARTS patterns. SMARTS matching and identification of matched atoms were carried out as per Figure 1. The manual and automated assessments of Heat Map quality are based on the schemes explained under "Quality assessment of Heat Map images". Quality scores, not to be confused with prediction scores, based on Heat Map correspondence to detoxifying groups were not assigned as no detoxifying groups were identified via SMARTS matching.

254x190mm (96 x 96 DPI)

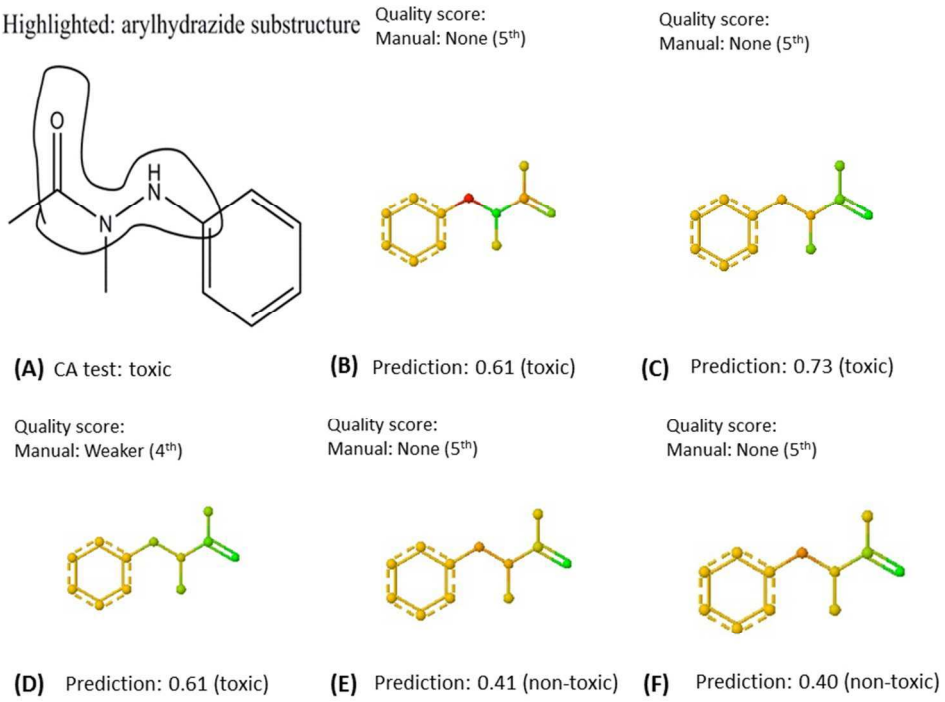


Figure 7 Heat Map (atom coloring & Symmetrized Single Molecule Normalization) analysis of binary classification leave-one-out predictions made for molecule CA31 in the CA dataset (c.f. Figure 6(A) in Mohr et al.).17 (A) Molecule with biologically significant substructures shown, based on a combination of mechanistic reasoning and expert knowledge reported previously in the literature,17 and experimental chromosome aberration test assignment reported.17 (B – F) Heat Map images: (B) Random Forest classification (Kuz'min/Palczewska, averaged predictions); (C) Random Forest classification (local gradients, majority vote predictions); (D) Random Forest (local gradients, averaged predictions); (E) Probit-PLS; (F) SVM. Predictions are reported as per Figure 4. The manual assessments of Heat Map quality are based on the scheme explained under "Quality assessment of Heat Map images". Here, the assignments were not simply based on matching the entire putative toxicophore,17 but were informed by consideration of the likely toxic substructure revealed via metabolism, which is expected to be the arylhydrazine substructure and may even be the aromatic amine substructure arising from further metabolism.17,48

254x190mm (96 x 96 DPI)

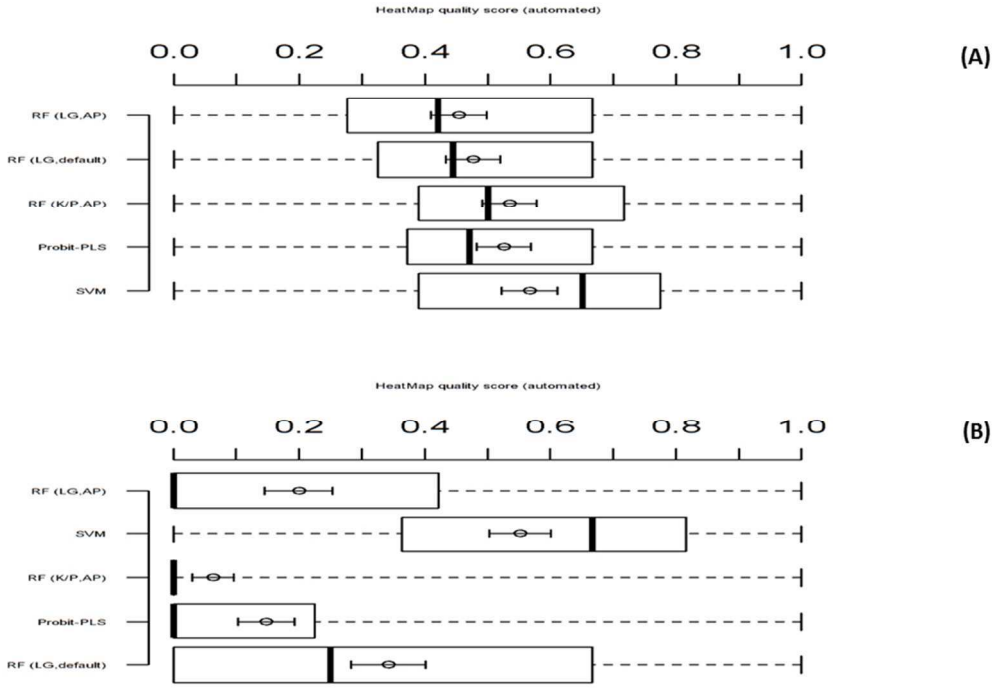


Figure 8 Automated Heat Map numeric quality scores, calculated as the average of F-measures for their correspondence to toxicophore and detoxifying substructure SMARTS matches as explained in Figure 1, for Heat Map images generated using atom coloring and (A) Symmetrized Single Molecule Normalization or (B) Constant Symmetrized Normalization. Quality score distributions correspond to all 39 leave-one-out predictions made for a subset of Kazius dataset molecules identified via toxicophore SMARTS matches. The different quality score distributions are annotated according to the different approaches used to obtain the underlying descriptor contributions: RF (LG, AP) = Random Forest classification (local gradients, averaged predictions); RF (LG, default) = Random Forest classification (local gradients, majority votes predictions); RF (K/P, AP) = Random Forest classification (Kuz'min/Palczewska, averaged predictions); Probit-PLS (linear coefficients); SVM (linear coefficients). The distributions are represented as boxplots, with the bold lines showing the medians and the superimposed circle the arithmetic means, with error bars denoting the standard errors in the means.

254x190mm (96 x 96 DPI)

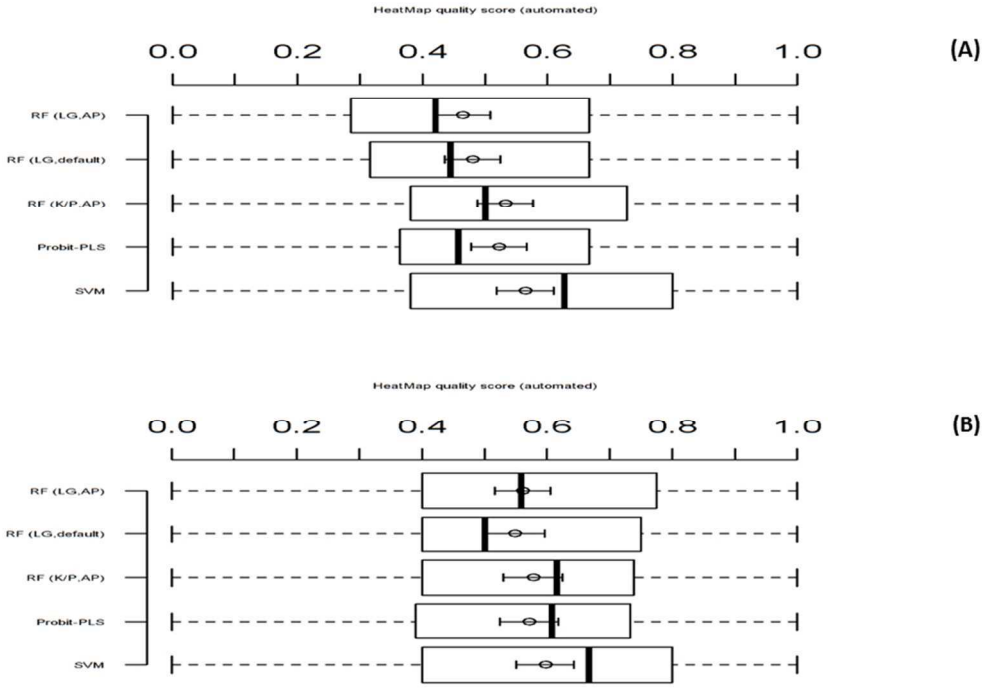


Figure 9 Automated Heat Map numeric quality scores, based on their correspondence to toxicophore substructure SMARTS matches as explained in Figure 1, for Heat Map images generated using atom coloring and Symmetrized Single Molecule Normalization. In sub-figure (A), quality score distributions correspond to 38 leave-one-out predictions made for a subset of Kazius dataset molecules identified via toxicophore SMARTS matches, excluding the single molecule ("1028-11-1") for which a detoxifying substructure was also identified via SMARTS pattern matches. In sub-figure (B), the distributions correspond to the subsets of those 38 molecules which were correctly predicted. The number of molecules correctly predicted to be toxic or non-toxic varied across methods: Random Forest classification (averaged predictions) = 28; Random Forest classification (majority vote predictions) = 29; Probit-PLS = 28; SVM = 31. The different quality scores distributions are annotated according to the different approaches used to obtain the underlying descriptor contributions: RF (LG, AP) = Random Forest classification (local gradients, averaged predictions); RF (LG, default) = Random Forest classification (local gradients, majority votes predictions); RF (K/P, AP) = Random Forest classification (Kuz'min/Palczewska, averaged predictions); Probit-PLS (linear coefficients); SVM (linear coefficients). The distributions are represented as boxplots, with the bold lines showing the medians and the superimposed circle the arithmetic means, with error bars denoting the standard errors in the means.

254x190mm (96 x 96 DPI)

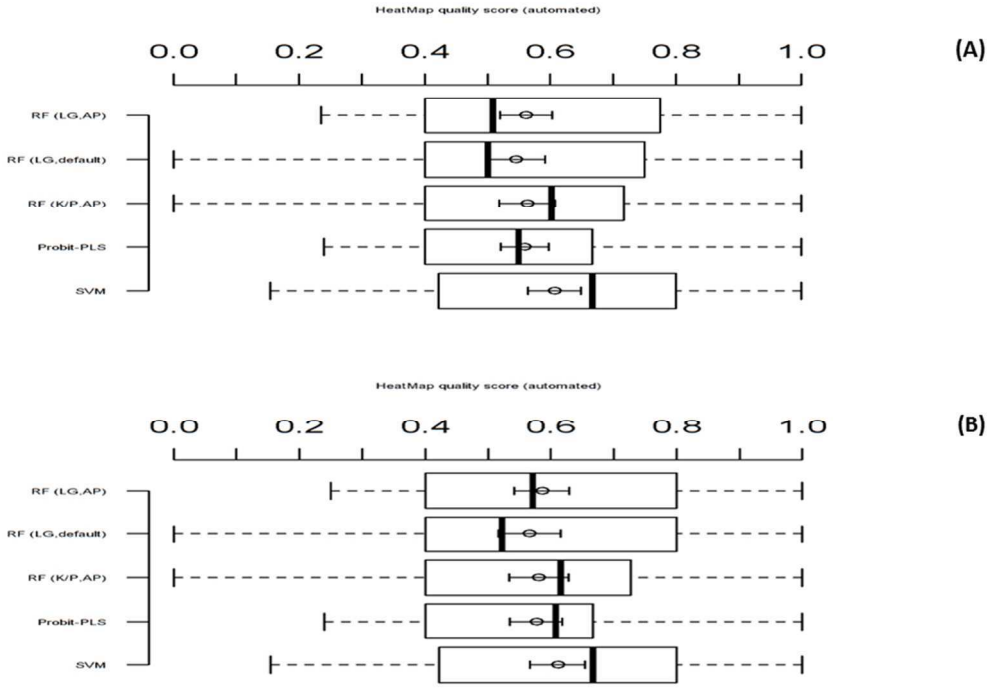


Figure 10 Automated Heat Map numeric quality scores, calculated based on their correspondence to toxicophore SMARTS matches as explained in Figure 1, for Heat Map images generated using atom coloring and Symmetrized Single Molecule Normalization. In sub-figure (A), the distributions of quality scores correspond to all toxic leave-one-out predictions made for a subset of Kazius dataset molecules identified via toxicophore SMARTS matches, excluding the single molecule ("1028-11-1") for which a detoxifying substructure was also identified via SMARTS pattern matches. In sub-figure (B), the distributions of quality scores correspond to the subsets of those toxic predictions which were correct. The number of molecules (correctly) predicted to be toxic, hence the number of Heat Map images for which these distributions were generated, varied across methods: Random Forest classification (averaged predictions) = 28 (25 correct); Random Forest classification (majority vote predictions) = 29 (26 correct); Probit-PLS = 30 (26 correct); SVM = 31 (28 correct). The different quality scores distributions are annotated according to the different approaches used to obtain the underlying descriptor contributions: RF (LG, AP) = Random Forest classification (local gradients, averaged predictions); RF (LG, default) = Random Forest classification (local gradients, majority votes predictions); RF (K/P, AP) = Random Forest classification (Kuz'min/Palczewska, averaged predictions); Probit-PLS (linear coefficients); SVM (linear coefficients). The distributions are represented as boxplots, with the bold lines showing the medians and the superimposed circle the arithmetic means, with error bars denoting the standard errors in the means.

254x190mm (96 x 96 DPI)

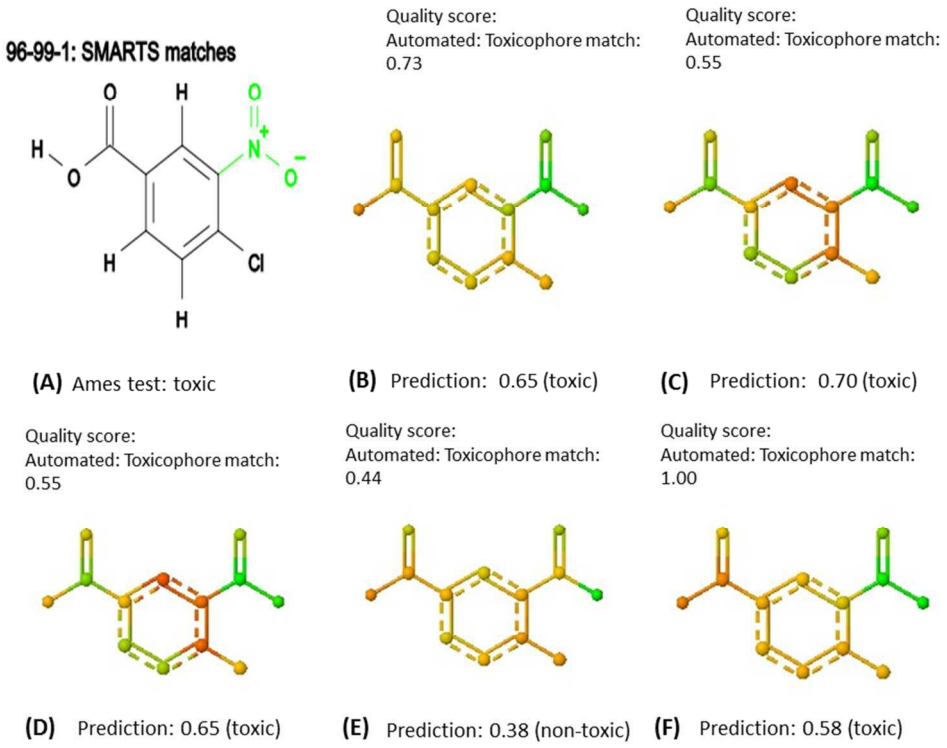


Figure 11 Heat Map (atom coloring & Symmetrized Single Molecule Normalization) analysis of binary classification leave-one-out predictions made for molecule "96-99-1" in the Kazius dataset. (A) Molecule with SMARTS matches shown in green for specific aromatic nitro toxicophore, derived from the work of Kazius et al.48 based on a combination of mechanistic reasoning, expert knowledge and data analysis, and experimental Ames test assignment reported in the Kazius dataset SDF file.13 The SMARTS matches (B – F) Heat Map images: (B) Random Forest classification (Kuz'min/Palczewska, averaged predictions); (C) Random Forest classification (local gradients, majority vote predictions); (D) Random Forest (local gradients, averaged predictions); (E) Probit-PLS; (F) SVM. Predictions are reported as per Figure 4. The SMARTS matches show the atoms matched to SMARTS patterns. SMARTS matching and identification of matched atoms were carried out as per Figure 1. The automated assessments of Heat Map quality are based on the scheme explained in Figure 1. Quality scores, not to be confused with prediction scores, based on Heat Map correspondence to detoxifying groups were not assigned as no detoxifying groups were identified via SMARTS matching.

254x190mm (96 x 96 DPI)

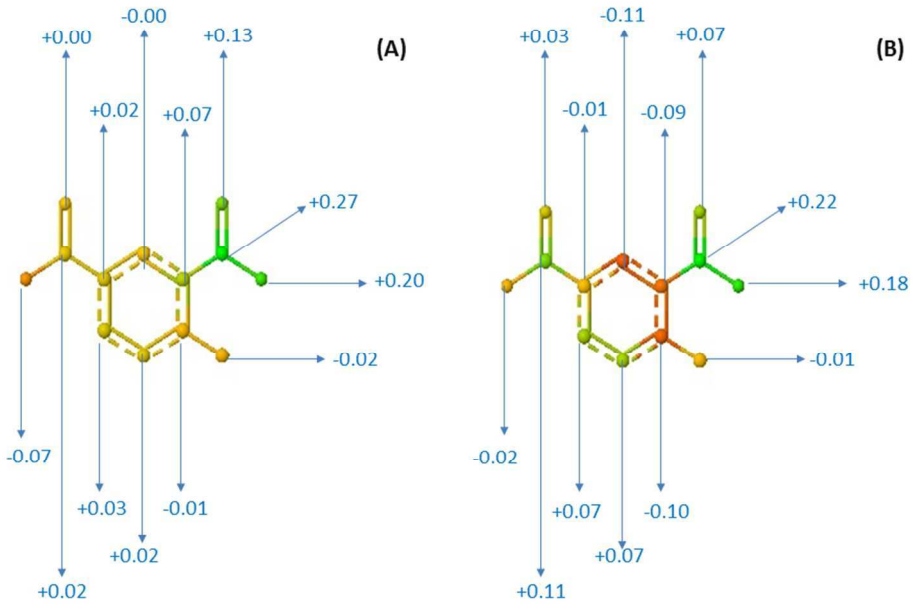


Figure 12 Comparison between the estimated raw atom scores (2dp), for the molecule "96-99-1" in the Kazius dataset, derived from the descriptor contributions corresponding to a single leave-one-out prediction made using Random Forest classification (averaged predictions) according to the Kuz'min/Palczewska (A) and local gradients (B) approaches. The raw scores were estimated via processing output from the HeatMapWrapper tool72 in atom coloring and Symmetrized Single Molecule Normalization mode. The correlation between the corresponding raw atom score estimates, in terms of the Pearson's correlation coefficient, is 0.74. The raw atom score estimates are used to annotate the corresponding atoms in the Heat Map (atom coloring & Symmetrized Single Molecule Normalization) image. N.B. The specific atom IDs, used to identify the corresponding atom score estimates, were identified as per Figure 1.

254x190mm (96 x 96 DPI)

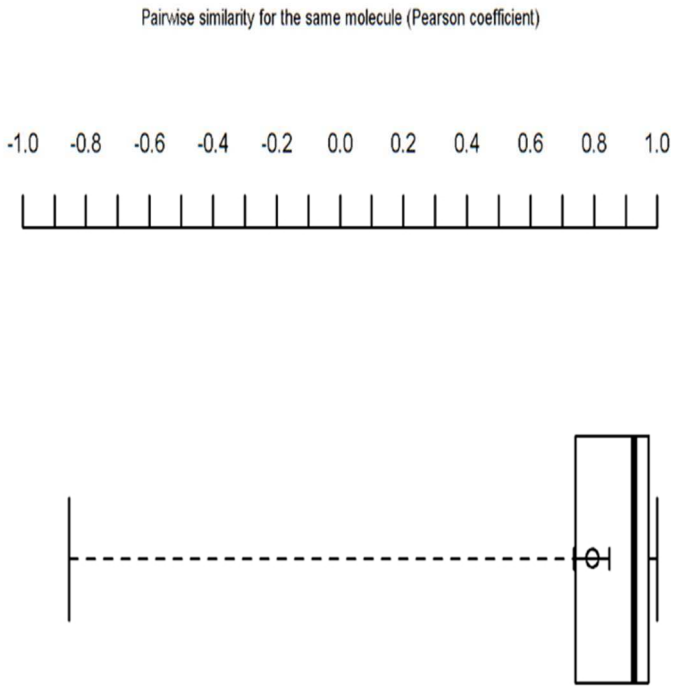


Figure 13 The distribution of pairwise similarities, calculated via the Pearson's correlation coefficient using estimates of the raw atom scores derived from descriptor contributions generated via the Kuz'min/Palczewska and local gradients approaches, for all leave-one-out Random Forest classification predictions (averaged predictions approach) made for a subset of molecules in the Kazius dataset identified via toxicophore SMARTS pattern matches. (The raw atom score estimates were obtained via processing the output of the HeatMapWrapper program72 in atom coloring and Symmetrized Single Molecule Normalization mode.) The distribution is represented as a boxplot, with the bold line showing the median and the superimposed circle the arithmetic mean, with error bars denoting the standard error in the mean.

254x190mm (96 x 96 DPI)