PLOS Computational Biology

Prediction of the Potency and the Selectivity of Mammalian Cyclooxygenase Inhibitors with Ensemble Proteochemometric Modeling

--Manuscript Draft--

Manuscript Number:	
Full Title:	Prediction of the Potency and the Selectivity of Mammalian Cyclooxygenase Inhibitors with Ensemble Proteochemometric Modeling
Short Title:	Proteochemometric modeling of Cyclooxygenase/Ligands interactions
Article Type:	Research Article
Keywords:	cyclooxygenase; ensemble proteochemometric modeling; prediction of ligand potency and selectivity
Corresponding Author:	Thérèse E Malliavin Institut Pasteur Paris, FRANCE
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Institut Pasteur
Corresponding Author's Secondary Institution:	
First Author:	Isidro Cortes-Ciriano
First Author Secondary Information:	
Order of Authors:	Isidro Cortes-Ciriano
	Daniel S Murrell
	Gerard JP van Westen
	Andreas Bender
	Thérèse E Malliavin
Order of Authors Secondary Information:	
Abstract:	Cyclooxygenases (COX) are present in the body in two isoforms, namely: COX-1, constitutively expressed, and COX-2, induced in physiopathological conditions such as cancer or chronic inflammation. The inhibition of COX with non-steroideal anti-inflammatory drugs (NSAIDs) is the most widely used treatment for chronic inflammation despite the adverse effects associated to prolonged NSAIDs intake. Thus, capitalizing on bioactivity data from both isoforms simultaneously would contribute to develop COX inhibitors with better safety profiles. We applied ensemble proteochemometric modeling (PCM) for the simultaneous prediction of the potency and the selectivity of 3,228 distinct COX inhibitors on 11 mammalian cyclooxygenases. Ensemble PCM models (R²0test=0.65, and RMSEtest=0.71) outperformed on the test set models exclusively trained on compound (R20test=0.17, and RMSEtest=1.09) or protein descriptors (R20test=0.16 and RMSEtest=1.10), and predicted COX selectivity for 1,086 selective and nonselective COX inhibitors with R20test=0.59 and RMSEtest=0.76. These values are in agreement with the maximum and minimum achievable R20 test and RMSEtest values of approximately 0.68 for both metrics. Confidence intervals for individual predictions were calculated from the standard deviation of the predictions calculated by the individual models composing the ensembles. Finally, two substructure analysis pipelines singled out chemical substructures implicated in both potency and selectivity in agreement with the literature.
Suggested Reviewers:	Christian Kramer Centrum fuer Chemie und Biomedizin

	Christian.Kramer@uibk.ac.at Christian Kramer works on the proteochemometrics and on the drug design, using statistics. One of his works: http://www.ncbi.nlm.nih.gov/pubmed/24738976 is quite relevant for the discussion.
	Paul Sheridan Merck Research Laboratories sheridan@merck.com Paul Sheridan worked on the problem of the applicability domain of proteochemometrics models.
	Ulf Norinder H. Lundbeck A/S ulfn@lundbeck.com Ulf Norinder worked on the problem of the applicability domain of proteochemometrics models.
	Martin Eklund Uppsala biomedicinska centrum BMC martin.eklund@farmbio.uu.se Martin Eklund worked on the problem of the applicability domain of proteochemometrics models.
	Max Kuhn Pfizer Global R&D Max.Kuhn@pfizer.com Max Kuhn has developed the R package "caret" and he is thus the right person to evaluate the ensemble modeling approach proposed in the manuscript.
Opposed Reviewers:	
Additional Information:	
Question	**Response**
Data Availability PLOS journals require authors to make all data underlying the findings described in their manuscript fully available, without restriction and from the time of publication, with only rare exceptions to address legal and ethical concerns (see the PLOS Data Policy and FAQ for further details). When submitting a manuscript, authors must provide a Data Availability Statement that describes where the data underlying their manuscript can be found. Your answers to the following constitute your statement about data availability and will be included with the article in the event of publication. **Please note that simply stating 'data available on request from the author' is not acceptable.** ***If, however, your data are only available upon request from the author(s), you must answer "No" to the first question below, and explain your exceptional situation in the text box provided.*** Do the authors confirm that all data underlying the findings described in their manuscript are fully available without restriction?	Yes - all data are fully available without restriction
Please describe where your data may be	All relevant data are within the paper and its Supporting Information files.

found, writing in full sentences. **Your answers should be entered into the box below and will be published in the form you provide them, if your manuscript is accepted.** If you are copying our sample text below, please ensure you replace any instances of XXX with the appropriate details. If your data are all contained within the paper and/or Supporting Information files, please state this in your answer below. For example, “All relevant data are within the paper and its Supporting Information files.” If your data are held or will be held in a public repository, include URLs, accession numbers or DOIs. For example, “All XXX files are available from the XXX database (accession number(s) XXX, XXX)." If this information will only be available after acceptance, please indicate this by ticking the box below. If neither of these applies but you are able to provide details of access elsewhere, with or without limitations, please do so in the box below. For example: “Data are available from the XXX Institutional Data Access / Ethics Committee for researchers who meet the criteria for access to confidential data.” “Data are from the XXX study whose authors may be contacted at XXX.” * typeset	
Additional data availability information:	Tick here if the URLs/accession numbers/DOIs will be available only after acceptance of the manuscript for publication so that we can ensure their inclusion before publication.

Thérèse E. Malliavin
Unité de Bioinformatique Structurale
CNRS UMR 3825
Institut Pasteur
28 rue du Dr. Roux
75 015 Paris, France
terez@pasteur.fr

July 22nd, 2014

to Editorial Board
Plos Computational Biology

Dear Sir,

Please find enclosed a manuscript entitled "Prediction of the Potency and the Selectivity of Mammalian Cyclooxygenase Inhibitors with Ensemble Proteochemometric Modeling" by Isidro Cortes-Ciriano, Daniel S. Murrell, Gerard J.P. van Westen, Andreas Bender and myself, that we would like to submit for publication in *Plos Computational Biology*. This manuscript presents and benchmarks for the first time, to our knowledge, ensemble proteochemometric modeling of ligand bioactivity. The obtained model is successfully applied to the simultaneous prediction of the potency and the selectivity of 3,228 distinct inhibitors on 11 mammalian cyclooxygenases, which are key proteins in the process of inflammation.

Due to the originality of the approach applied, as well as the biological importance of cyclooxygenases, I feel that this work is suitable for submission to *Plos Computational Biology*.

Sincerely,

Thérèse E. Malliavin

Prediction of the Potency and the Selectivity of Mammalian Cyclooxygenase Inhibitors with Ensemble Proteochemometric Modeling

Isidro Cortes-Ciriano[1], Daniel S. Murrell[2], Gerard J.P. van Westen[3] Andreas Bender[2,*], Thérèse E. Malliavin[1,*]

1 Institut Pasteur, Unité de Bioinformatique Structurale; CNRS UMR 3825; Département de Biologie Structurale et Chimie; 25, rue du Dr Roux, 75015 Paris, France.

2 Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Cambridge, United Kingdom.

3 European Molecular Biology Laboratory European Bioinformatics Institute, Wellcome Trust Genome Campus, CB10 1SD, Hinxton, Cambridge, UK.

∗ **E-mail: ab454@cam.ac.uk, therese.malliavin@pasteur.fr**

Abstract

Cyclooxygenases (COX) are present in the body in two isoforms, namely: COX-1, constitutively expressed, and COX-2, induced in physiopathological conditions such as cancer or chronic inflammation. The inhibition of COX with non-steroideal anti-inflammatory drugs (NSAIDs) is the most widely used treatment for chronic inflammation despite the adverse effects associated to prolonged NSAIDs intake. Thus, capitalizing on bioactivity data from both isoforms simultaneously would contribute to develop COX inhibitors with better safety profiles. We applied ensemble proteochemometric modeling (PCM) for the simultaneous prediction of the potency and the selectivity of 3,228 distinct COX inhibitors on 11 mammalian cyclooxygenases. Ensemble PCM models ($R^2_{0\ test}$=0.65, and $RMSE_{test}$=0.71) outperformed on the test set models exclusively trained on compound ($R^2_{0\ test}$=0.17, and $RMSE_{test}$=1.09) or protein descriptors ($R^2_{0\ test}$=0.16 and $RMSE_{test}$=1.10), and predicted COX selectivity for 1,086 selective and non-selective COX inhibitors with $R^2_{0\ test}$=0.59 and $RMSE_{test}$=0.76. These values are in agreement with the maximum and minimum achievable $R^2_{0\ test}$ and $RMSE_{test}$ values of approximately 0.68 for both metrics. Confidence intervals for individual predictions were calculated from the standard deviation of the predictions calculated by the individual models composing the ensembles. Finally, two substructure analysis pipelines singled out chemical substructures implicated in both potency and selectivity in agreement with the literature.

Author Summary

Introduction

Cyclooxygenases (EC 1.14.99.1), also known as endoperoxidases, prostaglandin G/H synthases or simply COX, are involved in the biosynthesis of prostaglandin H_2 from arachidonic acid. (1) Prostaglandin H_2 is further converted into prostanoids which play a key role in inflammation. Thus, since the development of aspirin® in 1899, (2) the inhibition of the cyclooxygenase activity with non-steroidal anti-inflammatory drugs (NSAIDs) has been exploited to treat inflammation. Nonetheless, kidney failure and gastrointestinal side-effects, such as peptic ulcer, have been correlated to long-term intake of NSAIDs. (3) Until 1991, only one form of the enzyme (COX-1) was thought to be responsible for both the constitutive and the local biosynthesis of prostaglandins. In that year, (4) an inducible cyclooxygenase (COX-2) was discovered and the different roles of both isoenzymes were revealed, though there exists some overlap: COX-1 is constitutively expressed serving as the source of housekeeping prostaglandins,

whereas the expression of COX-2 increases in pathophysiological situations such as acute pain, inflammation or cancer, (5) giving rise to the hope that efficacy and side-effects can, to some extent, be delineated when blocking the prostaglandin synthesis pathway associated with inflammation and pain.

In the last two decades, research in both the pharmaceutical industry and academic laboratories has been driven by the hypothesis that selective COX-2 inhibitors would exhibit strong anti-inflammatory and analgesic properties without leading to the unwanted gastrointestinal side effects. (6) Nevertheless, a few organs, e.g. the brain cortex and renal glomeruli, express COX-2 constitutively. (1) The association between the inhibition of COX-2 in these organs with cardiovascular hazard (CVH) was ratified in 2004 and 2005. (7; 8) These findings led the US Food and Drug Agency (FDA) to retrieve rofecoxib (Vioxx) and valdecoxib (Bextra) from the market, and to include boxed warnings for all selective COX-2 inhibitors. Higher risk of heart attack and hypertension have also been reported for non-selective NSAIDs, thus highlighting that cardiovascular risk might not be related to the degree of COX selectivity. (9) In 2012, Yu *et al.* (10) demonstrated that the cardiovascular risk originates from COX-2 inhibition by selective and not selective NSAIDs and is taking place in blood vessels. These authors have shown that COX-2 inhibition leads to a decrease in prostaglandin (mainly PGI2) and to nitric oxide (NO) production which is sufficient to increase the risk of heart failure, hypertension and thrombosis. (10)

Nevertheless, there are still niche populations which can benefit from selective COX-2 inhibitors, e.g. patients who cannot afford to take non-selective COX inhibitors, due to an increased risk of peptic ulcers or cancer. In addition, selective COX-2 inhibitors continue to be the common treatment for chronic inflammatory and pain disorders. (3; 11) Moreover, NSAIDs are known to reduce the risk of (among others): (12; 13; 14; 15) colon cancer, (16; 17; 18; 19) Alzheimer's disease, and platelet aggregation. (5; 20) Overall, NSAIDs are still one of the most commonly prescribed drugs in the world. (21) and this trend is likely to increase owing to the aging of the population. Therefore, the administration of NSAIDs in clinics is currently subject to a benefit-risk assessment between the patients clinical profile and potential drugs side-effects, (22) always aiming at optimizing both the dosage and the duration of the drug regimen. (3)

The isoform selectivity of COX inhibitors stems from a structural difference in the binding site. The binding site of both cyclooxygenases is highly conserved except for the substitution of an isoleucine at position 523 in COX-1 with a valine in COX-2. (23) This substitution results in a larger binding site in COX-2, as the smaller size of valine allows access to a side-pocket. This structural difference has been exploited for the rational design of potent and selective COX-2 inhibitors by both medicinal and computational chemistry. (23; 24; 25) To date, a plethora of *in silico* studies have been published with the aim of better understanding and predicting the potency of COX inhibitors on either COX-1 or COX-2 using molecular docking and QSAR models. (26; 27; 28; 29; 30) Nonetheless, none of these studies was able to concomitantly predict potency and selectivity on both COX isoenzymes. Given that the bioactivity profiles of selective COX inhibitors on COX-1 and COX-2 are highly uncorrelated, (24; 25) only a predictive model trained on both the chemical and the target space would be able to predict potency and selectivity, as well as to predict the activity of a given compound on a yet untested isoform. In that way, new potent, selective and safe COX inhibitors could be discovered.

Proteochemometrics (PCM) has recently emerged as an approach capable to simultaneously relate the chemical and the target space in single machine learning models in order to predict the bioactivity for a given compound on a given biomolecular target. (31; 32) This integration of chemical and biological information enables the prediction of both compounds selectivity and potency. Moreover, the pair-input nature of PCM enables, within the limits of the data presented to the model, the extrapolation on both the chemical and the target space. Therefore, the bioactivity of new compounds on yet untested tar-

gets can be predicted. These features of PCM makes it different from both chemogenomics and QSAR, thus allowing, among others: (34; 35) (i) the inclusion of bioactivity information from orthologuous targets, (34), (ii) bioactivity prediction for emergent viral mutations, (35) or (iii) the design of personalized medicine for e.g. cancer treatment. (32)

In this contribution, we apply the principles of PCM to concomitantly predict the potency and selectivity of 3,228 compounds on 11 mammalian cyclooxygenases. To this aim, we have trained PCM models with different machine learning algorithms on public IC50 values from ChEMBL 16, (36) including data on human COX-1 and COX-2 and on 9 orthologues. In an attempt to increase model performance, these models have been combined in ensembles (ensemble modeling), thus constituting, to our knowledge, the first PCM study where ensemble PCM modeling is applied. Additionally, the description of compounds with keyed fingerprints has enabled the deconvolution of the chemical space to rationalize both the potency and the selectivity of COX inhibitors.

Materials and Methods

Dataset

IC50 values for 11 mammalian cyclooxygenases, listed in Table 1, were retrieved from ChEMBL 16. (36) To ensure the reliability of the bioactivity values, only IC50 values corresponding to small molecules and satisfying the following criteria were kept: (i) activity relationship equal to '=', (ii) assay score confidence ≥ 8, and (iii) activity unit equal to 'nM'. The average bioactivity was calculated when multiple IC50 values were annotated on the same compound-target combination. The application of these filters led to a final dataset composed of 3,228 distinct compounds and 11 sequences, being the total number of datapoints 4,937. The negative logarithm with base 10 of the IC50 values (pIC50) was used as the response variable to train all models. The crystallographic structure of the ovine COX-1 complexed with celecoxib (PDB (37) ID: 3KK6 (38)) was used to extract the residues in the binding site. Those residues within a sphere of radius equal to 10Å centered in the ligand were selected. The corresponding residues for the other 10 sequences were identified by sequence alignment. (39) The sequence alignment as well as the final residue selection are provided in the supplementary information.

Computational Details

Descriptors

Chemical structures were standardized with the function *StandardiseMolecules* from the R package *camb* (40) with the following options: (i) inorganic molecules were removed, and (ii) molecules were selected irrespectively of the number of fluorines, chlorines, bromines or iodines present in their structure, or of their molecular mass. Morgan fingerprints (41; 42) were calculated using RDkit (release version 2013.03.02). (43; 44) For the calculation of unhashed Morgan fingerprints, (44) each compound substructure in the dataset, with a maximal diameter of four bonds, was assigned to an unambiguous identifier. Subsequently, substructures were mapped into an unhashed (keyed) array of counts. Physicochemical descriptors (PaDEL (45) were calculated with the function *GeneratePadelDescriptors* from the R package *camb*. The R package *vegan* was used to generate the distributions of pairwise compound similarities (Jaccard distance). (46)

The amino acids composing the binding site of the mammalian cyclooxygenases considered in this study (Table 1), were described with five amino acid extended principal property scales (5 z-scales). (47) Z-scales were calculated with the R package *camb*. (40)

Machine Learning Implementation

Machine learning models were built in R using the packages *caret* (48) and *camb* (40). Model ensembles were created with the help of the R package *caretEnsemble*. (49) Both the dataset and the modeling pipeline coded in R is available in the documentation of the R package *camb*. (40)

Model Generation

Descriptors with a variance close to zero were removed with the function *RemoveNearZeroVarianceFeatures* from the R package *camb* using a cut-off value equal to 30/1. (40) Subsequently, the remaining descriptors were centered to zero mean and scaled to unit variance with the function *PreProcess* from the R package *camb*.

The values of the models parameters were optimized by grid search and k-fold cross validation (CV). (50) The whole dataset was split into $k+1$ folds by stratified sampling of the pIC50 values. One fold, $1/k+1$, constituted the test set. The remaining folds, $k/k+1$, were used to optimize the values of the parameters in the following way. For each combination of parameters, a model was trained on k-$1/k$ folds, and the values for the remaining fold, $1/k$, were then predicted. This procedure was repeated k times, each time holding out of a different fold. The values of the parameters exhibiting the lowest average RMSE value along the k folds was considered as optimal. Subsequently, a model was trained on the whole training set, $k/k+1$, using the optimized values for the parameters. The predictive power of this model was assessed on the test set. To significantly compare the quality of the modeling with different machine learning algorithms, the same folds were used to train all models.

In order to assess if merging the chemical and the target space in a single PCM model enhanced models performance, we trained a Random Forest (RF) model using either: (i) only compound descriptors (Family QSAR), (51) or (ii) only target descriptors (Family QSAM). (51) Obtaining a high performance with a Family QSAR model indicates that the bioactivities of a given compound on different targets are correlated. Thus, target descriptors would not contribute to increase model performance. On the other hand, high performance observed for a Family QSAM model indicates that the bioactivity values only depend on the targets and not on the compounds, *i.e.* the bioactivities of a set of diverse compounds are correlated on a given target. In this case, compound descriptors would not be required to predict compounds affinity, as target descriptors alone would be sufficient.

Model Validation

Both internal and external validation were performed according to the criteria proposed by Tropsha *et al.* (52; 53; 54), and to the RMSE values (Equation 1). The formulae of the statistical metrics used in the internal ($\mathrm{RMSE_{int}}$ and q^2_{int}) and the external ($\mathrm{RMSE_{test}}$, q^2_{test} and $R^2_{ext\ 0}$) validation are:

$$RMSE = \frac{\sqrt{(y-\widetilde{y})^2}}{N} \tag{1}$$

$$q^2 = 1 - \frac{\sum_{i=1}^{N}(y_i-\widetilde{y}_i)^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2} \tag{2}$$

$$R_0^2 = 1 - \frac{\sum_{i=1}^{N}(y_i-\widetilde{y}_i^{r0})^2}{\sum_{i=1}^{N}(y_i-\bar{y})^2} \tag{3}$$

where N represents the size of the training or test set, y_i the observed bioactivity values, $\widetilde{y}_i$ the predicted bioactivity values, and $\bar{y}$ the average values of the response variable for those datapoints

included either into the internal or the test set, and $\widetilde{y}^{r0} = s\widetilde{y}$, with $s = \frac{\sum y_i \widetilde{y}_i}{\sum \widetilde{y}_i^2}$. To consider a model as statistically sound, the statistical metrics must satisfy the following criteria: (i) $q^2_{int} > 0.5$, and (ii) q^2_{test} and $R^2_{ext\ 0} > 0.6$. $R^2_{ext\ 0}$ imposes the regression line to pass through the origin (intercept equal to zero).

Assessment of Maximum Model Performance

To further assess the reliability of the models in the light of the uncertainty of the bioactivity values, (55; 56; 57) we established the maximum $R^2_{0\ test}$ and q^2_{test}, and minimum $\text{RMSE}_{\text{test}}$ values achievable given: (i) the uncertainty of public IC50 data, and (ii) the number of datapoints in both the training and the test set. The distributions of minimum $\text{RMSE}_{\text{test}}$, and maximum q^2_{test}, and $R^2_{0\ test}$ values were calculated in the following way.

Firstly, a random sample, A, was generated from the pCI50 values with a size equal to the test set. Secondly, the sample A_{noisy} was calculated by adding to A a random noise with mean zero and standard deviation equal to the experimental error. Then, the statistical metrics were calculated for A with respect to A_{noisy}. The calculation of the statistical metrics on 1,000 generations of random samples A and noisy samples A_{noisy} provided the distributions for the statistical metrics.

The maximum and minimum values of these distributions were then used to validate the values of the metrics obtained when evaluating the bioactivities predicted for the test set. If the obtained metrics were beyond the maximum values (for q^2_{test} and $R^2_{0\ test}$) or the minimum values (for $\text{RMSE}_{\text{test}}$) of the corresponding distributions, the model is likely to be over-optimistic. (50)

The experimental error required to define the random samples A_{noisy} was taken as 0.68 pIC50 units, which corresponds to the average standard deviation value for public IC50 datasets, as estimated by Kalliokoski *et al.* (56).

Ensemble Modeling

Gradient-boosting machines (GBM), (58) Random Forest (RF), (59) and Support Vector Machines (SVM) (60) were implemented to train a mode library. The resulting models were combined in ensembles using two techniques, namely: greedy optimization and model stacking. Depending on the models considered when training an ensemble, two types of model ensembles were defined: (i) homo-ensembles: the same algorithm was used to train all models of the ensemble, but the parameter values were different in each model, (ii) hetero-ensembles: the number of algorithms used to train the models combined in the ensemble was greater or equal than 2.

Greedy Optimization

Greedy optimization, based on the work of Caruana et al. (61), optimizes the RMSE between the bioactivities of the test set ($1/k+1$) and model predictions on the same set. These predictions were calculated with the model library trained on a training set with identical folds ($k/k+1$). Each model was assigned a weight in the following manner. Initially, all models had a weight equal to zero. Afterwards, the weight of a given model was repeatedly incremented by 1 if the subsequent normalized weight vector allowed a closer match between the weighted combination of model predictions and the observed values. This repetition was carried out n times, n = 1000 in the present work, and the resulting weight vector was normalized to obtain the final models weighting.

Model Stacking (MS)

The concept of model stacking is illustrated in Figure 1. In this case, the predictions on the training set ($k/k+1$) calculated with the model library during CV served as descriptors. Thus, a training matrix was

defined where rows were indexed by the datapoints in the training set used to train the model library, and columns by the models in the aforesaid library.

A machine learning model was trained on this matrix, irrespective of the algorithms used to generate the model library. If the selected algorithm has the inherent capability to determine the importance of each descriptor, as for Elastic Net, a vector of weights for the models can be defined. Given that each descriptor corresponds to a particular model, this vector will determine its contribution to the generated ensemble. In the present study we used the following algorithms: linear model, Elastic Net, SVM with linear and radial kernels, and RF.

Estimation of the Error of Individual Predictions

In order to estimate errors for individual predictions, we used the standard deviation of the predictions of the individual models composing a given model ensemble, *i.e.* ensemble standard deviation (E_{std}). Previous studies, (62; 63; 64; 65) have highlighted the usefulness of considering the ensemble standard deviation as a domain applicability (DA) measure, specially in the case of RF models, where the calculation of the standard deviation along the trees is straightforward. (62; 63) Here, we extend this idea to ensembles composed of models trained with different algorithms (hetero-ensembles). The ensemble standard deviations were scaled with the parameter β in order to obtain individual confidence intervals for each prediction, which are thus defined as:

$$IC_{\beta\ i} = \widetilde{y}_i \pm E_{std\ i}\ \beta\ \{\beta \in \mathbb{R} \mid \beta > 0\} \tag{4}$$

To assess the practical usefulness of the derived confidence intervals, the percentage of datapoints for which the predicted values lied within IC_β ($0 < \beta < 4$) was calculated. Both the predictions calculated during model training (using the optimal parameter values), *i.e.* cross-validated predictions, as well as the predictions on the test set were used.

Interpretation of Compound substructures

The contribution of chemical substructures to bioactivity on human cyclooxygenases was deconvoluted using a predictive and a statistical method:

Prediction of Bioactivity Values with and without each Compound Substructure (predictive method). (66; 67)
This first technique quantifies the contribution of each chemical substructure to bioactivity by calculating the distribution of differences between (i) the bioactivity of all compounds containing a given substructure, and (ii) the predicted bioactivity using PCM for these compounds, from which that substructure was virtually removed.
To virtually remove a substructure, we iteratively set its count equal to zero in all compound descriptors presenting it. The difference between the predicted bioactivity values in the presence or absence of a given substructure was then calculated. The average value of these differences, weighted by the number of counts of the feature in each compound, corresponds to the average contribution of that feature to bioactivity. (67) The contribution was estimated for all compound features considered in the model. The sign of the difference ({+/-}) indicates whether the feature is respectively beneficial or deleterious for compound bioactivity.

Statistical Significance between Bioactivity Distributions with and without each Compound Substructure (statistical method)
In order to identify chemical substructures that might not be recognized by the predictive method due to

moderate PCM model performance, we also deconvoluted the chemical space in a model-independent way. We created two bioactivity sets, each containing the pIC50 values for either human COX-1 or human COX-2. For each of these sets, we calculated the pIC50 distribution for all compounds either presenting or not a given substructure. The normality of these distributions was assessed with the Shapiro-Wilk test ($\alpha = 0.05$). If both distributions followed the Gaussian distribution, a two-tailed t-test for independent samples ($\alpha = 0.05$) was applied to statistically evaluate the difference between both distributions. If the difference was significant, we assumed the considered substructure to have an influence on bioactivity on the isoenzyme considered in each case. The sign of the difference between the mean value of the bioactivity distribution for compounds containing or not the substructure, indicates whether the presence of the substructure hampers or fosters compound bioactivity on that cyclooxygenase. Therefore, each substructure was assigned a label, 'deleterious' or 'beneficial', depending on its influence on bioactivity.

Finally, we intended to assess which substructures always increase or decrease compound bioactivity on human COX-1 and COX-2. In that way, substructures identified in the previous step are finally identified as: (i) increasing or decreasing bioactivity on human COX-1, (ii) increasing or decreasing bioactivity on human COX-2, and (iii) increasing or decreasing bioactivity on both human COX-1 and COX-2.

Results

Analysis of the Chemical and the Target Space

Target Space
The PCA analysis of the amino acid descriptors of the binding site of the 11 mammalian cyclooxygenases (Table 1) is shown in Figure 2. Orthologue sequences COX1 and COX2 define two distant clusters. As paralogues display more sequence variability than orthologues, and as small molecules tend to display similar binding within orthologues, (68) we hypothesize that merging bioactivities from orthologues and paralogues will lead to more predictive models. In addition, these results indicate that the amino acid descriptors account for structural differences between COX-1 and COX-2.

Chemical Space
The initial bioactivity selection from ChEMBL 16, (36) consisted of 6,804 datapoints. As previously highlighted, (55) a large number of target-compound combinations in ChEMBL are annotated with more than one bioactivity value, because the total number of different compound-target combinations after duplicate removal was 4,937. As can be seen in Figure S1A, the standard deviations for the bioactivity values annotated on the same compound-target combination are in less than 2% of the cases higher than two pIC50 units, whereas more than 90% of the repeated bioactivities exhibit a standard deviation close to zero (Figure S1B). Consequently, we decided to take the average of these repeated values instead of the median value: this latter value would be more suitable if only outliers were more aboundant.
As stated in the introduction, the main advantage of a PCM model applied on mammalian cyclooxygenases would be to predict selectivity. To ensure that our dataset covered chemical entities with diverse bioactivity profiles on COX-1 and COX-2, we selected all compounds annotated on both human cyclooxygenases. This resulted in a selection of 1,086 compounds, out of a total of 3,228 different inhibitors present in the dataset. The scatterplot of the bioactivities of these compounds on human COX-1 against human COX-2 (Figure 3A) reveals that the difference in bioactivity for some compounds depending on the isoenzyme is higher than 4 pIC50 units (upper left corner of Figure 3A). RMSE and $R^2_{0\,test}$ values for the bioactivities on COX-1 with respect to COX-2 are, respectively, 1.69 pIC50 units and -0.42. As the area above the diagonal of Figure 3A is more populated, there are more compounds with higher activity on COX-2 than on COX-1. Therefore, these data let us conclude that our bioactivity selection comprises

compounds exhibiting high selectivity towards COX-2. In addition, the overlap between the datapoints in the PCA of the compound descriptors (Figure S2) indicates that the compounds annotated on the COX targets cover the same regions of the chemical space.

PCM Validation

Overall, the models obtained with GBM, RF, and SVM (Table 2A) satisfied our model validation criteria, described in Materials and Methods (Equations (1) to (3)), namely: $q^2_{int} > 0.5$ and, q^2_{test} and $R^2_{ext\ 0} > 0.6$. The performance of the three algorithms is comparable since $R^2_{0\ test}$ values range from 0.60 to 0.61, and $RMSE_{test}$ from 0.76 to 0.79 pIC50 units between the different models. Interestingly, the predictive power did not vary when using hashed or unhashed fingerprints, being the $R^2_{0\ test}$ and $RMSE_{test}$ differences smaller than 0.01 in both cases (data not shown). Thus, we decided to rather use unhashed fingerprints as this choice enables an interpretation of the models according to chemical substructures.

To ensure that our modeling results did not arise from chance correlations, we trained models with an increasingly bigger fraction of randomized bioactivity values (y-scrambling). (69) The representation of model performance as a function of the percentage of randomized bioactivities is given in Figure S3. When approximately 35% of the bioactivity values are randomized, $R^2_{0\ test}$ become negative, which indicates that the relationships found by our models between both the chemical and the target space, and the bioactivity values are not spurious. (69)

PCM Models are in Agreement with the Maximum Achievable Performance

The distributions of the respectively maximum and minimum achievable $R^2_{0\ test}$ and $RMSE_{test}$ values are depicted in Figure 4. The maximum correlation values $R^2_{0\ test}$ are far from 1, which agrees with observations previously reported for public data. (67; 70) The mean of the minimum theoretical $RMSE_{test}$ values lies between 0.68 and 0.69, which is comparable to the level of uncertainty in public IC50 data reported by Kalliokoski *et al.*. (56) The mean of the distribution of theoretical $R^2_{0\ test}$ values is between 0.67 and 0.69. The minimum $RMSE_{test}$ and maximum $R^2_{0\ test}$ values obtained with the individual models, 0.76 and 0.61 respectively (Table 2A), thus appear consistent with the underlying uncertainty in the present dataset.

PCM Outperforms both Family QSAR and Family QSAM on this dataset

Interestingly, neither the Family QSAR nor the Family QSAM model alone could infer the relationships in the dataset, as the respective $R^2_{0\ test}$ and $RMSE_{test}$ values were: (i) for Family QSAR: 0.17 and 1.09 pIC50 units, and (ii) for Family QSAM: 0.16 and 1.10 pIC50 units (Table 2B). Taken together, these results suggest that: (i) compound bioactivities on different targets are not correlated, as indicated by the low performance of the Family QSAR model, and (ii) compound bioactivities depend on compounds structure, as highlighted by the low performance of the QSAM model.

PCM Outperforms Individual QSAR Models

We then evaluated on individual targets the usefulness of PCM in comparison with QSAR models. Independent QSAR models for those targets with more than 100 bioactivities, namely: human COX-1 and COX-2, ovine COX-1 and COX-2, and mouse COX-2. The human COX-2 model exhibits a $RMSE_{test}$ value of 0.78 pIC50 units, which is 0.03 pIC50 units larger than the $RMSE_{test}$ value for the datapoints annotated on human COX-2 averaged over ten PCM models, namely 0.76 +/- 0.04 pIC50 units. By contrast, the $R^2_{0\ test}$ value drops to 0.54. Better correlations are obtained for the individual QSAR models

corresponding to both the mouse and the ovine COX-2, for which the $R^2_{0\ test}$ values are 0.57 in both cases, whereas the $RMSE_{test}$ values are 0.81 and 0.79 pIC50 units. At contrary, the human and the ovine COX-1 QSAR models cannot relate the descriptor space to the bioactivity values in a statistically sound manner, as they exhibit respective $R^2_{0\ test}$ values of 0.30 and 0.36.

Altogether, these data evidence the versatility of PCM to integrate incomplete information from different sequences. Furthermore, PCM strongly outperforms one-target and one-space models (Family QSAR, individual QSAR, and Family QSAM). (32)

Model Ensembles Exhibit Higher Performance than Single PCM Models

As the most predictive PCM model exhibited moderately high $R^2_{0\ test}$ and q^2_{test} values, as well as moderately low $RMSE_{test}$ values (Table 2A), we explored the possibility of enhancing model performance by combining different models into a more predictive model ensemble (Table 2C and 2D). Two ensemble techniques were implemented, namely: greedy optimization and model stacking (MS), previously described in section "Ensemble Modeling". To gather a library of diverse models, we trained a total of 282 GBM, RF and SVM models. Each of these models was trained with different parameter values. Hence, the performance of single models ranged from very poor to that of the individual models described above (Table 2A).

Initially, we created ensembles using only the most predictive GBM, RF and SVM models (Table 2C). Overall, all model ensembles (Table 2C) exhibited higher predictive power than single models (Table 2A). The best $R^2_{0\ test}$ value, 0.63, was obtained with the greedy and the MS linear ensemble. The weights for the three models in the greedy ensemble were: (i) GBM: 0.35, (ii) RF: 0.12, and (iii) SVM: 0.53. The MS Elastic Net ensemble displayed the highest predictive power, with a $RMSE_{test}$ value of 0.72 (Table 2C). The small differences in performance observed between ensembles, with the exception of the RF ensemble are negligible, since, in the experience of the authors, (67) the standard deviation observed for the $R^2_{0\ test}$ and $RMSE_{test}$ values when using different samples during model training on datasets comprising several thousand of datapoints are between 0.1 and 0.3. The only model that led to worse results than the individual PCM models was the RF ensemble, with $R^2_{0\ test}$ and $RMSE_{test}$ values of 0.58 and 0.77 respectively.

In a second step, ensembles were optimized using all models in the model library, namely 282 (Table 2D). Interestingly, the values of the statistical metrics of all ensembles increased. This effect is more distinct for the MS RF ensemble, as the $R^2_{0\ test}$ value improved from 0.58 to 0.63 (Table 2). The greedy ensemble trained on the whole library was composed of 6 GBM, 1 RF and 3 SVM models. Model weights in this case ranged from 0.32 (SVM) to 0.0003 (GBM). The MS SVM ensemble with radial kernel displayed the highest predictive ability, with $R^2_{0\ test}$ and $RMSE_{test}$ of 0.65 and 0.71 pIC50 units. This $RMSE_{test}$ value differed of 0.02 pIC50 units from the average theoretical $RMSE_{test}$ of 0.68 achievable given the quality of public pIC50 data (Figure 4).

Worthy of mention is the lack of performance improvement (data not shown) of homo-ensembles (*i.e* ensembles created with models trained with a given algorithm but with different parameter values) with respect to the most predictive model trained with either GBM, RF or SVM (Table 2A), as the difference in $R^2_{0\ test}$ and $RMSE_{test}$ values was below 0.01 for both metrics. By contrast, the most peformant ensembles displaying were obtained when combining models with high and low predictive ability, instead of only considering the most predictive GBM, RF and SVM models. Overall, these data underline the highest predictive power of hetero-ensembles generated with a model library displaying a comprehensive range of predictive abilities.

The Ensemble Standard Deviation Enables the Definition of Informative Confidence Intervals

Figure 5 displays the percentage of datapoints which predicted value lies within confidence intervals calculated with increasingly larger β values (Equation 4). The ensemble model exhibiting the highest predictive power ($RMSE_{test}$: 0.71; $R^2_{0\ test}$: 0.65), namely MS SVM Radial Ensemble, was used to make the predictions and to calculate confidence intervals. Confidence intervals calculated for the cross-validated predictions (shown as squares in Figure 5) require larger β values to reach a given level of confidence when compared to those calculated on the test set (shown as triangles in Figure 5). This can be seen as the percentage of datapoints which true value is included within the confidence interval ($\beta = 1$) for the cross-validated predictions is 40%, whereas this value increases till 70% in the case of the test set. This difference was however expected, as the predictions on the test set are made with models trained on a larger fraction of the dataset. Overall, the percentage of true values lying within the confidence interval derived for a given β value is expected to increase with the number of datapoints available during model training. Figure 5 can be used to determine the β value corresponding to the confidence interval required by the user.

Ensemble Modeling Enables the Prediction of Human COX Inhibitors Selectivity

As previously stated, selectivity is a crucial aspect in the discovery and optimization of COX inhibitors. To assess whether PCM models were able to predict cyclooxygenase selectivity, we predicted the bioactivity values for the 1,086 compounds annotated on both COX-1 and COX-2. Figure 3B, which displays the observed against the predicted pIC50 values for these compounds, shows that PCM models are able to predict potency and selectivity. Indeed, the $R^2_{0\ test}$ and $RMSE_{test}$ values calculated for the observed pIC50 values with respect to those predicted by the PCM model are, respectively, 0.59 and 0.76 pIC50 units.

Subsequently, we analyzed the capability of PCM models to correctly predict the bioactivity for both selective and non-selective compounds. A compound was considered as selective or non selective if the absolute value of the difference between its bioactivity on COX-1 and COX-2 is larger or smaller than 1 pIC50 unit. On this basis, 592 compounds were considered as selective, and 494 as non selective. The error in prediction for the non selective compounds was lower than 1 pIC50 units in 85.1% of the cases, and lower than 0.5 pIC50 units for 56.8% thereof. Similarly, the error in prediction was lower than 1 pIC50 units for 80.3% of the selective compounds, and lower than 0.5 pIC50 units for 50.7% thereof. Consequently, these data indicate that PCM models are capable to predict the potency and selectivity for both selective and non selective compounds on human COX-1 and COX-2. In addition, we anticipate that model performance is likely to increase with the inclusion of more bioactivity data in the models.

Models Performance per Target is Related to Compounds Diversity

To further assess model performance on a *per* target basis, we generated 10 RF models each one trained on a different subset of the whole dataset. For all targets, a non-linear relationship can be established between average model performance, quantified by the standard deviation of the $RMSE_{test}$ over the ten models, and the amount of compounds with bioactivities annotated on a given target (Figure 6). The variation of performance between the protein targets can be also related to the compound diversity (Figure S4).

Human cyclooxygenases, with the highest number of annotated compounds (Table 1), exhibited average $RMSE_{test}$ values between 0.74 and 0.76 pIC50 units, and standard deviations within the range 0.05-0.10 pIC50 units. For these proteins, the distributions of compounds diversity, marked with an asterisk in

Figure S4, are skewed towards high similarity values, with mean values between 0.75 and 0.85.

On the other hand, mouse COX-2 and ovine COX-1 display average $RMSE_{test}$ values of 0.70 and 0.73 pIC50 units, respectively, with standard deviations within the range 0.10-0.20 pIC50 units, probably related to the smaller number of compounds annotated on these proteins (Table 1). High predictive ability on mouse COX-2 was expected given the high $R^2_{0\ test}$ value, 0.57, obtained with the individual QSAR model, whereas low performance was expected for ovine COX-1, as the individual QSAR model displayed a $R^2_{0\ test}$ value of 0.36. Unsurprisingly, skewed distributions in compounds diversity are observed for mouse COX-2 and ovine COX-1 (Figure 6).

Conversely, ovine COX-2, with 341 annotated targets, displayed a worse average $RMSE_{test}$ value, within the 0.80-0.85 range of pIC50 units (Figure 6). This decrease in performance for ovine COX-2 might be ascribed to the higher dispersion of the pairwise compound similarity distribution with respect to those observed for mouse COX-2 and ovine COX-1 (Figure S4).
The dependency of model performance on compounds diversity and the number of datapoints is even more contrasted for targets with less than 100 annotated bioactivities. Indeed, the average $RMSE_{test}$ value for mouse COX-1, with 50 compounds, lies within the 0.57-0.62 range of pIC50 units and the distribution of compunds diversity is skewed towards high similarity values (Figure S4). Conversely, the average $RMSE_{test}$ value increases till 0.80-0.90 pIC50 units for bovine COX-1 (Figure S4), annotated with 48 bioactivities and for which the pairwise compound similarity distribution presents several peaks, thus highlighting the structural diversity of the compounds. Finally, targets with less than 30 annotated compounds exhibit multimodal pairwise similarity distributions and, consequently, model performance is low, with standard deviations in the 0.50-1.00 range of pIC50 units (Figure 6).

Overall, chemical diversity in the training set contributes to enhance the applicability of a PCM model. Nonetheless, a balance needs to be established between this diversity and the number of datapoints, as targets with a small number of diverse compounds are poorly modeled.

Interpretation of compound substructure

Predictive Method

The usage of unhashed fingerprints permitted the deconvolution of the chemical space to determine the influence of compound substructure on bioactivity. Two substructure analysis methodologies were implemented, as described in the section "Interpretation of Compound substructures". The first approach, predictive method, relies on the PCM model to correctly predict the bioactivity for a compound when a given substructure is virtually removed from a compound descriptor. The second approach, statistical method, is a pipeline designed to statistically assess how the presence of a given substructure influences, on average, bioactivity on the compounds.

Figure 7 shows the contribution of each substructure to bioactivity on human COX-1 and COX-2 calculated with the predictive method. Red and blue areas correspond respectively to substructures that, on average, enhance or decrease compound bioactivity. Representative substructures either deleterious or beneficial for bioactivity are also shown. Generally, substructures shown to have an influence on bioactivity display an opposite behaviour depending on the isoenzyme type. For example, a pyrrole ring with aryl substituents in the 2,3-positions (substructure **c** in Figure 7) is predicted to have a high influence on bioactivity, increasing it on COX-2 and decreasing it on COX-1. This observation is in agreement with the literature as the 2,3-diarylpyrrole series with an halogen substituent in the 5-position acting as electron withdrawing group have been found as selective COX-2 inhibitors. (71; 72) The pyrrole moiety with a radical in the 1-position is also found as a selectivity feature (substructure **b** in Figure 7). This

agrees with the discovery by Khanna *et al.* (73) of a series of 1,2-diarylpyrroles as potent and selective COX-2 inhibitors.
On the other hand, substructures conferring a deleterious effect could also be identified. substructure **e** in Figure 7 is represented within compound 3-(1H-indol-5-yloxy)-5,5-dimethyl-4-(4-methylsulfonylphenyl)furan-2-one (CHEMBL322276). This compound is part of a series of 3-heteroaryloxy-4-phenyl-2(5H)-furanones reported as selective COX-2 inhibitors by Lau *et al.* (74). Its COX-1/COX-2 selectivity ratio is larger than 4.17, which agrees with the prediction of decreasing bioactivity on COX-1. In general, substructures decreasing bioactivity tend to be small and less informative (e.g. single atoms or substructures with two heavy atoms), than those fostering compounds potency.

Statistical Method

The implementation of the statistical method to deconvolute the chemical space, which evaluates the statistical significance between bioactivity distributions in the presence or absence of each compound substructure, led to the following observations: (i) 74 substructures increase bioactivity on COX-2, (ii) 64 substructures decrease bioactivity on COX-2, (iii) 9 substructures increase bioactivity on COX-1, (iv) 2 substructures decrease bioactivity on COX-1, (v) 1 substructure increases bioactivity on both COX-1 and COX-2, and (vi) 6 substructures decrease bioactivity on both COX-1 and COX-2.

Well-known chemical moieties conferring selectivity to COX-2 were present in this substructure selection. Figure S5 shows the 20 substructures predicted to have the highest influence to increase bioactivity on human COX-2. For instance, substructures containing thiazole, pyrrole, pyrazole and oxazole rings were enriched for COX-2. (24; 25) Likewise, tri-fluorometil and sulfonamide radicals were also enriched, which appear in *e.g.* celecoxib. (24) Substructures predicted to influence in the same way the compound bioactivity on both COX-1 and COX-2 are small, which makes difficult to extract medicinal chemistry knowledge therefrom (Figure S6).

It is nevertheless remarkable that the output of both methods is contradictory for some substructures. By way of example, substructure **d** in Figure 7 is considered as deleterious for bioactivity on COX-1 by the predictive method, whereas it is regarded as beneficial by the statistical method. Dannhardt *et al.* (75) highlighted the key role of the carbonyl moiety for the potency of a series of diarylmethanone compounds on both COX isoenzymes. Nonetheless, Scholz *et al.* (76) have recently reported a series of *ortho*-carbaborane derivatives of indomethacin as selective COX-2 inhibitors. Furthermore, substructure **d** also appears in a series of [2-[(4-substituted or 4,5-disubstituted)-pyridin-2-yl]carbonyl-(5- or 6-substituted or 5,6-disubstituted)-1H-indol-3-yl]acetic acid analogues identified as COX-2 inhibitors. (77) Plausible reasons for this divergence are analyzed in the Discussion section.
Overall, both substructure analysis pipelines have been proved to highlight chemical moieties conferring or decreasing potency and selectivity in agreement with the literature.

Discussion

In this contribution two ensemble modeling techniques, namely greedy optimization and model stacking, have been presented and benchmarked on a PCM dataset comprising the bioactivities of COX inhibitors on 11 mammalian cyclooxygenases (Table 1). PCM has been shown to relate the target and the chemical spaces to bioactivity in a statistically sound manner (Table 2). (52; 53; 54) Family QSAR as well as Family QSAM displayed poor performance (Table 2B).

Three machine learning algorithms (GBM, RF and SVM) have been implemented individually and combined in model ensembles. The application of ensemble modeling have been shown to outperform

single machine learning models, the improvement being increased if the three most predictive GBM, RF and SVM models are combined in the same ensemble (Table 2C). Nonetheless, the model stacking (MS) SVM radial kernel model trained on the predictions of a library of 282 single PCM models (Table 2D) displayed the lowest $RMSE_{test}$ and the highest $R^2_{0\ test}$ values. This non-linear model combination led to a $RMSE_{test}$ value comparable to the experimental uncertainty of public IC50 data. (56) It is noteworthy to mention that this ensemble, displaying the highest predictive power, was obtained when combining several hundreds of poor and highly predictive models, instead of only the most predictive models of each class, namely GBM, RF and SVM (Table 2C). Therefore, these results suggest that, if sufficient computing resources are available, higher predictive ability can be obtained with a large and diverse model library. Given that the ensemble concept is not restricted to any particular machine learning algorithm, the pipeline proposed in this study can be further explored.

The variability in the predictions of the individual models composing model ensembles, quantified by the ensemble standard deviation, served to define informative confidence intervals. Previous studies highlighted the usefulness of this variability as a applicability domain metric. (62; 63; 64; 65) Here, we have extended this concept to ensembles of models trained on different algorithms (Figure 5). Overall, the application of ensemble modeling with a model library trained with either the same algorithm but different parameter values (homo-ensemble), or with different algorithms (hetero-ensemble) constitutes a promising alternative to single models in the context of predictive bioactivity modeling.

High predictive ability was not only attained for compounds potency, but also for selectivity, as both single models and model ensembles, were able to effectively predict cyclooxygenase selectivity on human cyclooxygenases (Figure 3B). Therefore, the present study illustrates how the combination of the target and the chemical spaces in a unique PCM model enables the concomitant prediction of compound potency and selectivity in the context of multi-target systems. The implications of COX-2 in widespread diseases, *e.g.* cancer, has prompted the design of potent and selective COX-2 inhibitors since the early 1990s. (24; 25) Thus, the suitability of PCM to concomitantly predict both potency and selectivity opens new avenues for the design of cyclooxygenase inhibitors.

The description of compounds with unhashed Morgan fingerprints enabled the deconvolution of the chemical space. Unhashed fingerprints enhance model interpretability, therefore enabling a chemically meaningful understanding of models behaviour. However, the amount of the chemical space that can be potentially covered is restricted by the chemical diversity present in the dataset on which models are trained, as the dataset defines the pool of compound substructures to be considered in the fingerprints. In that way, low chemical diversity in the training set might certainly decrease the application of any model trained thereon to unexplored chemical space. Therefore, the usage of hashed or unhashed fingerprints is to be determined by the application context of the PCM model. In those cases were the aim is to find inhibitors containing novel chemical moieties, it would be advisable not to describe compounds with unhashed fingerprints, but rather to use hashed fingerprints, as they enable the description of the whole chemical space. In addition, the number of datapoints and the diversity of the compounds annotated on a target needs to be balanced, as targets annotated with a small number of diverse compounds tend to be poorly modeled. (67)

The two approaches presented in this study for the deconvolution of the chemical space, namely: (i) bioactivity prediction with and without a given compound substructure (predictive method), and (ii) assessment of the statistical difference between the bioactivity distributions corresponding to compounds presenting or not a given compound substructure (statistical method), singled out chemical moieties responsible for COX-2 selectivity in agreement with the scientific literature.

The divergent results described for substructure **d** in Figure 7, plausibly arise from the following properties of the two methods. As in the predictive method the bioactivity is predicted by calculating the average difference between the predicted value for a compound with and without a given substructure, the (potentially non-linear) relationships between the substructures present in a molecule can be established, and the dependence of bioactivity on additional substructures or scaffolds present in the molecule accounted. On the other hand, the statistical method considers the substructures as independent. The two methods can thus give contrasted results for example in the following case. We can envision a compound, A, presenting a substructure, S_1, having no effect on bioactivity, and a second substructure, S_2, strongly fostering bioactivity on the studied biomolecular target. Additionaly, we consider compound B, which only harbors substructure S_2. Contradictory results would be given by the two methods with respect to the influence of substructure S_1 on bioactivity. The predictive method would predict a similar bioactivity value for compound A with and without substructure S_1, as the bioactivity depends on substructure S_2. By contrast, the statistical method would consider substructure S_1 as relevant for bioactivity given that the difference between the bioactivities of compounds A and B, *i.e.* either presenting or not substructure S_1, would be significant. It follows from the preceeding that the predictive method is best suited to give insight into the contribution of single substructures to the bioactivity of individual compounds, whereas the statistical method is more suited for the identification of the general relevance of the substructures to bioactivity. If the general influence of a substructure on bioactivity is assessed with the predictive method, both the mean value and the standard deviation of the differences between the predicted bioactivity values with and without a given substructure should be reported, as the standard deviation indicates whether the importance of that substructure to bioactivity depends or not on other substructures. (67)

In the statistical method, the pIC50 difference associated to a significant p-value might be negligible from a medicinal chemistry standpoint. In addition, the capability of the t-test to identify significant differences depends on the sample size. Thus, a small pIC50 difference can be detected as significant if the sample size is large, whereas it might not be detected for smaller samples. Therefore, the conclusions extracted from the application of the statistical method depend on the analyzed dataset, whereas the predictive method might be less dependent on the dataset composition if the models are applied within their applicability domain. For a recent and detailed discussion of the application of the student t-test to assess the statistical significance of bioactivity differences in the context of Matched Molecular Pair Analysis (MMPA), the reader is referred to Kramer *et al.* (78). In summary, the application of both methods can help to unravel whether the contribution of a given substructure to compound bioactivity depends exclusively on itself, or on the presence of other substructures or chemical scaffolds. (79)

Conclusions

Ensemble modeling has been introduced in the context of PCM to predict both the potency and the selectivity of mammalian cyclooxygenase inhibitors. The combination of single models in model ensembles has led to increased predictive ability, as well as to the definition of confidence intervals for individual predictions. PCM has been shown to enable the prediction of selectivity with high confidence. Finally, the implementation of two different substructure analysis pipelines, which reliability for different purposes has been pointed out, has permitted the recognition of chemical moieties implicated in potency and selectivity in agreement with the scientific literature.

Author Contributions

ICC, AB and TM conceived and designed the study. ICC gathered the dataset, trained the models, analyzed the results and produced the figures. DM provided analytical tools for ensemble modeling. ICC, DM, GvW, AB and TM wrote the paper.

Acknowledgements

ICC thanks the Paris-Pasteur International PhD Programme for funding. ICC and TM thank CNRS and Institut Pasteur for funding. DM thanks Unilever for funding. GvW thanks EMBL (EIPOD) and Marie Curie (COFUND) for funding. AB thanks Unilever and the European Research Commission (Starting Grant ERC-2013-StG 336159 MIXTURE) for funding.

References

1. Luo C, He Ml, Bohlin L (2005) Is COX-2 a perpetrator or a protector? selective COX-2 inhibitors remain controversial. Acta Pharm Sinic 26: 926–933.
2. Vane JR (1971) Inhibition of prostaglandin synthesis as a mechanism of action for aspirin-like drugs. Nature New Biol 231: 232–235.
3. Fine M (2013) Quantifying the impact of NSAID-associated adverse events. Am J Manag Care 19: 267–272.
4. Xie WL, Chipman JG, Robertson DL, Erikson RL, Simmons DL (1991) Expression of a mitogen-responsive gene encoding prostaglandin synthase is regulated by mRNA splicing. Proc Natl Acad Sci 88: 2692–2696.
5. Sostres C, Gargallo CJ, Lanas A (2014) Aspirin, cyclooxygenase inhibition and colorectal cancer. World J Gastrointest Pharmacol Ther 5: 40–49.
6. Warner TD, Giuliano F, Vojnovic I, Bukasa A, Mitchell JA, et al. (1999) Nonsteroid drug selectivities for cyclo-oxygenase-1 rather than cyclo-oxygenase-2 are associated with human gastrointestinal toxicity: a full in vitro analysis. Proc Natl Acad Sci 96: 7563–7568.
7. Bresalier RS, Sandler RS, Quan H, Bolognese JA, Oxenius B, et al. (2005) Cardiovascular events associated with rofecoxib in a colorectal adenoma chemoprevention trial. N Engl J Med 352: 1092–1102.
8. Nussmeier NA, Whelton AA, Brown MT, Langford RM, Hoeft A, et al. (2005) Complications of the COX-2 inhibitors parecoxib and valdecoxib after cardiac surgery. N Engl J Med 352: 1081–1091.
9. Howes LG (2007) Selective COX-2 inhibitors, NSAIDs and cardiovascular events - is celecoxib the safest choice? Ther Clin Risk Manag 3: 831–845.
10. Yu Y, Ricciotti E, Scalia R, Tang SY, Grant G, et al. (2012) Vascular COX-2 modulates blood pressure and thrombosis in mice. Sci Transl Med 4: 132ra54.
11. Crofford LJ (2013) Use of NSAIDs in treating patients with arthritis. Arthritis Res Ther 15: S2.
12. Hermanson DJ, Hartley ND, Gamble-George J, Brown N, Shonesy BC, et al. (2013) Substrate-selective COX-2 inhibition decreases anxiety via endocannabinoid activation. Nature Neurosci 16: 1291–1298.

13. Zhang S, Zhang XQ, Ding XW, Yang RK, Huang SL, et al. (2014) Cyclooxygenase inhibitors use is associated with reduced risk of esophageal adenocarcinoma in patients with barretts esophagus: a meta-analysis. Br J Cancer 110: 2378-2388.

14. Frolov RV, Singh S (2014) Celecoxib and ion channels: A story of unexpected discoveries. Eur J Pharmacol 730: 61–71.

15. Robak P, Smolewski P, Robak T (2008) The role of non-steroidal anti-inflammatory drugs in the risk of development and treatment of hematologic malignancies. Leuk Lymphoma 49: 1452–1462.

16. Moore BC, Simmons DL (2000) COX-2 inhibition, apoptosis, and chemoprevention by nonsteroidal anti-inflammatory drugs. Curr Med Chem 7: 1131–1144.

17. Chen L, He Y, Huang H, Liao H, Wei W (2008) Selective COX-2 inhibitor celecoxib combined with EGFR-TKI ZD1839 on non-small cell lung cancer cell lines: in vitro toxicity and mechanism study. Med Oncol 25: 161–171.

18. Thun MJ, Henley SJ, Patrono C (2002) Nonsteroidal anti-inflammatory drugs as anticancer agents: mechanistic, pharmacologic, and clinical issues. J Natl Cancer Inst 94: 252–266.

19. Soh JW, Kazi JU, Li H, Thompson WJ, Weinstein IB (2008) Celecoxib-induced growth inhibition in SW480 colon cancer cells is associated with activation of protein kinase g. Mol Carcinog 47: 519–525.

20. Jouzeau JY, Terlain B, Abid A, Ndlec E, Netter P (1997) Cyclo-oxygenase isoenzymes. how recent findings affect thinking about nonsteroidal anti-inflammatory drugs. Drugs 53: 563–582.

21. Jones R, Rubin G, Berenbaum F, Scheiman J (2008) Gastrointestinal and cardiovascular risks of nonsteroidal anti-inflammatory drugs. Am J Med 121: 464–474.

22. Curiel RV, Katz JD (2013) Mitigating the cardiovascular and renal effects of NSAIDs. Pain Med 14 Suppl 1: S23–28.

23. Blobaum AL, Marnett LJ (2007) Structural and functional basis of cyclooxygenase inhibition. J Med Chem 50: 1425–1441.

24. Dannhardt G, Laufer S (2000) Structural approaches to explain the selectivity of COX-2 inhibitors: is there a common pharmacophore? Curr Med Chem 7: 1101–1112.

25. de Leval X, Delarge J, Somers F, de Tullio P, Henrotin Y, et al. (2000) Recent advances in inducible cyclooxygenase (COX-2) inhibition. Curr Med Chem 7: 1041–1062.

26. Reddy RN, Mutyala R, Aparoy P, Reddanna P, Reddy MR (2007) Computer aided drug design approaches to develop cyclooxygenase based novel anti-inflammatory and anti-cancer drugs. Curr Pharm Des 13: 3505–3517.

27. Kim HJ, Chae CH, Yi KY, Park KL, Yoo Se (2004) Computational studies of COX-2 inhibitors: 3D-QSAR and docking. Bioorg Med Chem Lett 12: 1629–1641.

28. Dube PN, Bule SS, Mokale SN, Kumbhare MR, Dighe PR, et al. (2014) Synthesis and biological evaluation of substituted 5-methyl-2-phenyl-1H-pyrazol-3(2H)-one derivatives as selective COX-2 inhibitors: Molecular docking study. Chem Biol Drug Des .

29. Gupta GK, Kumar A (2012) 3D-QSAR studies of some tetrasubstituted pyrazoles as COX-II inhibitors. Acta Pol Pharm 69: 763–772.

30. Narsinghani T, Chaturvedi SC (2006) QSAR analysis of meclofenamic acid analogues as selective COX-2 inhibitors. Bioorg Med Chem Lett 16: 461–468.

31. van Westen GJP, Wegner JK, IJzerman AP, van Vlijmen HWT, Bender A (2011) Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. Med Chem Commun 2: 16-30.

32. Cortes Ciriano I, Ain QU, Subramanian V, Lenselink EB, Mendez Lucio O, et al. Polypharmacology modelling using proteochemometrics: Recent developments and future prospects. In revision at Med Chem Comm .

33. Park Y, Marcotte EM (2012) Flaws in evaluation schemes for pair-input computational predictions. Nature Methods 9: 1134–1136.

34. van Westen GJP, van den Hoven OO, van der Pijl R, Mulder-Krieger T, de Vries H, et al. (2012) Identifying novel adenosine receptor ligands by simultaneous proteochemometric modeling of rat and human bioactivity data. J Med Chem 55: 7010–7020.

35. van Westen GJ, Hendriks A, Wegner JK, Ijzerman AP, van Vlijmen HW, et al. (2013) Significantly improved HIV inhibitor efficacy prediction employing proteochemometric models generated from antivirogram data. PLoS Comput Biol 9: e1002899.

36. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, et al. (2011) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40: D1100–D1107.

37. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The protein data bank. Nucleic Acids Res 28: 235-242.

38. Rimon G, Sidhu RS, Lauver DA, Lee JY, Sharma NP, et al. (2010) Coxibs interfere with the action of aspirin by binding tightly to one monomer of cyclooxygenase-1. Proc Natl Acad Sci 107: 28-33.

39. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. Mol Syst Biol 7.

40. Murrell DS, Cortes-Ciriano I, van Westen GJ, Malliavin T, Bender A (2014) Chemistry Aware Model Builder (camb): an R package for predictive bioactivity modeling. http://githubcom/cambDI/camb .

41. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50: 742–754.

42. Glem RC, Bender A, Arnby CH, Carlsson L, Boyer S, et al. (2006) Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. IDrugs 9: 199–204.

43. Landrum G (2006). RDKit Open-source cheminformatics.

44. Cortes-Ciriano I (2014). Fingerprintcalculator: http://github.com/isidroc/fingerprintcalculator.

45. Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32: 1466–1474.

46. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, et al. (2013) vegan: Community Ecology Package. R package version 2.0-9.

47. Sandberg M, Eriksson L, Jonsson J, Sjstrm M, Wold S (1998) New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. J Med Chem 41: 2481–2491.

48. Kuhn M (2008) Building predictive models in r using the caret package. J Stat Soft 28: 1–26.

49. Mayer Z (2013) caretensemble: Framework for combining caret models into ensembles. [r package version 1.0] .

50. Hawkins DM, Basak SC, Mills D (2003) Assessing Model Fit by Cross-Validation. J Chem Inf Model 43: 579–586.

51. Brown J, Okuno Y, Marcou G, Varnek A, Horvath D (2014) Computational chemogenomics: Is it more than inductive transfer? J Comput Aided Mol Des : 1-22.

52. Golbraikh A, Tropsha A (2002) Beware of q2! J Mol Graphics Modell 20: 269–276.

53. Tropsha A, Golbraikh A (2010) Predictive Quantitative Structure-Activity Relationships Modeling. Handbook of Chemoinformatics Algorithms 33: 211.

54. Tropsha A, Gramatica P, Gombar VK (2003) The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. QSAR Comb Sci 22: 69-77.

55. Kramer C, Kalliokoski T, Gedeck P, Vulpetti A (2012) The Experimental Uncertainty of Heterogeneous Public Ki Data. J Med Chem 55: 5165–5173.

56. Kalliokoski T, Kramer C, Vulpetti A, Gedeck P (2013) Comparability of mixed IC50 data - a statistical analysis. PloS ONE 8: e61007.

57. Kramer C, Lewis R (2012-09-01T00:00:00) Qsars, data and error in the modern age of drug discovery. Curr Top Med Chem 12: 1896-1902.

58. Friedman JH (2000) Greedy function approximation: A gradient boosting machine. Annals of Statistics 29: 1189–1232.

59. Breiman L (2001) Random forests. Mach Learn 45: 5–32.

60. Ben-Hur A, Ong CS, Sonnenburg S, Schlkopf B, Rtsch G (2008) Support Vector Machines and Kernels for Computational Biology. PLoS Comput Biol 4: e1000173.

61. Caruana R, Niculescu-Mizil A, Crew G, Ksikes A (2004) Ensemble Selection from Libraries of Models. New York, NY, USA: ACM. Banff, Alberta, Canada, 18 pp.

62. Sheridan RP (2012) Three useful dimensions for domain applicability in QSAR models using random forest. J Chem Inf Model 52: 814–823.

63. Sheridan RP (2013) Using random forest to model the domain applicability of another random forest model. J Chem Inf Model 53: 2837-2850.

64. Wood DJ, Carlsson L, Eklund M, Norinder U, Stå lring J (2013) QSAR with experimental and predictive distributions: an information theoretic approach for assessing model quality. J Comput Aided Mol Des 27: 203–219.

65. Dragos H, Gilles M, Alexandre V (2009) Predicting the predictability: a unified approach to the applicability domain problem of QSAR models. J Chem Inf Model 49: 1762–1776.

66. van Westen GJ, Wegner JK, Geluykens P, Kwanten L, Vereycken I, et al. (2011) Which compound to select in lead optimization? Prospectively validated proteochemometric models guide preclinical development. PLoS ONE 6: e27518.

67. Cortes Ciriano I, van Westen GJ, Lenselink EB, Murrell DS, Bender A, et al. (2014) Proteochemometrics modeling in a Bayesian framework. J Cheminf 6: 35.

68. Kruger FA, Overington JP (2012) Global Analysis of Small Molecule Binding to Related Protein Targets. PLoS Comput Biol 8: e1002333.

69. Clark R, Fox P (2004) Statistical variation in progressive scrambling. Journal of Computer-Aided Molecular Design 18: 563-576.

70. Brown SP, Muchmore SW, Hajduk PJ (2009) Healthy skepticism: assessing realistic model performance. Drug Discov Today 14: 420–427.

71. Wilkerson WW, Galbraith W, Gans-Brangs K, Grubb M, Hewes WE, et al. (1994) Antiinflammatory 4,5-diarylpyrroles: Synthesis and qsar. J Med Chem 37: 988-998.

72. Wilkerson WW, Copeland RA, Covington M, Trzaskos JM (1995) Antiinflammatory 4,5-diarylpyrroles. 2. activity as a function of cyclooxygenase-2 inhibition. J Med Chem 38: 3895-3901.

73. Khanna IK, Weier RM, Yu Y, Collins PW, Miyashiro JM, et al. (1997) 1,2-diarylpyrroles as potent and selective inhibitors of cyclooxygenase-2. J Med Chem 40: 1619-1633.

74. Lau CK, Brideau C, Chan CC, Charleson S, Cromlish WA, et al. (1999) Synthesis and biological evaluation of 3-heteroaryloxy-4-phenyl-2(5h)-furanones as selective cox-2 inhibitors. Bioorg Med Chem Lett 9: 3187 - 3192.

75. Dannhardt G, Fiebich BL, Schweppenhauser J (2002) COX-1/COX-2 inhibitors based on the methanone moiety. Eur J Med Chem 37: 147–161.

76. Scholz M, Blobaum AL, Marnett LJ, Hey-Hawkins E (2012) ortho-carbaborane derivatives of indomethacin as cyclooxygenase (COX)-2 selective inhibitors. Bioorg Med Chem Lett 20: 4830–4837.

77. Hayashi S, Ueno N, Murase A, Nakagawa Y, Takada J (2012) Novel acid-type cyclooxygenase-2 inhibitors: Design, synthesis, and structureactivity relationship for anti-inflammatory drug. Eur J Med Chem 50: 179–195.

78. Kramer C, Fuchs JE, Whitebread S, Gedeck P, Liedl KR (2014) Matched molecular pair analysis: Significance and the impact of experimental uncertainty. J Med Chem 57: 3786-3802.

79. Klekota J, Roth FP (2008) Chemical substructures that enrich for biological activity. Bioinformatics 24: 2518–2525.

Figure Legends

Figure 1. Ensemble modeling with model stacking. A A set of models are trained with diverse machine learning algorithms (*Model1 .. Model n* in the Figure). The predictions of these models on each datapoint in the training set calculated during cross validation, are used as descriptors to create a new training matrix, which rows are indexed by the datapoints in the training set and columns by the models in the library. A machine learning model is trained on this matrix. The resulting model is the model ensemble. **B**. The model ensemble is then applied on the test set.

Figure 2. PCA analysis of the target space. PCA analysis was applied on the binding site descriptors used to train the models. The first two principal components explained more than 80% of the variance, thus indicating that there are mainly two sources of variability in the descriptor space, namely the isoenzyme type. This fact can be seen as COX-1 (triangles) and COX-2 (squares) define two distant clusters. Overall, the binding sites of orthologue cyclooxygenases are more similar than those of paralogue sequences. These results also indicate that the amino acid descriptors account for structural differences between COX-1 and COX-2. Thus, it is expected that merging orthologues and paralogues will lead to more predictive models.

Figure 3. COX inhibitors selectivity on human COX-1 and COX-2. A. Scatterplot corresponding to the comparison of bioactivities against human COX-1 and COX-2 for 1,288 compounds. A large proportion of the compounds present a COX-2/COX-1 selectivity ratio between 2 and 4 pIC50 units. Therefore, the present dataset includes COX inhibitors with highly divergent bioactivity profiles for COX-1 and COX-2 ($R_0^2 = -0.420$). **B.** Scatterplot of the observed against the predicted pIC50 values for the compounds described in **A.**. Blue squares correspond to the activity on COX-1, whereas orange squares correspond to the activity on COX-2. The PCM models explain more than 59% of the variance ($R_0^2 = 0.593$), thus highlighting the ability of the PCM models to concomitantly predict potency and selectivity.

Figure 4. Distribution of theoretical $R_{0\ test}^2$ (A) and $\mathbf{RMSE_{test}}$ (B) values. The mean of the $R_{0\ test}^2$ distribution, 0.68, highlights that the uncertainty in public bioactivity data does not permit the obtention of models displaying $R_{0\ test}^2$ values close to 1. Similar results were obtained for q_{test}^2. The minimum $\mathrm{RMSE_{test}}$ value that a model can achieve without exhibiting overfitting is close to the experimental uncertainty.

Figure 5. Confidence intervals calculated from the ensemble standard deviation of the models present in the model ensembles. The percentage of datapoints which predicted bioactivities lie within confidence intervals calculated with increasingly larger β values (Equation 4), is shown for: (i) the cross validated predictions calculated during model training (*Training* in the Figure), and (ii) for the predictions on the test set (*Test* in the Figure) calculated with the most predictive model ensemble, namely "Stacking SVM Radial Ensemble". The percentage of true values lying within the confidence interval derived for a given β value increases with the number of datapoints available during model training. Overall, the confidence intervals derived from the ensemble standard deviation provide an estimation of the reliability of individual predictions, as in practice, this plot can be used to determine the β value corresponding to a given confidence level.

Figure 6. Target-averaged model performance. The number of datapoints is displayed through the size of the squares and has been converted to the logarithmic scale for the ease of representation. A correlation can be established between the number of datapoints and model performance, quantified by the standard deviation of the $RMSE_{test}$ values. Targets annotated with less than 30 compounds or with chemical structures displaying high structural diversity (*Oryctolagus cuniculus* COX-1, *Rattus norvegicus* COX-1, *Bos taurus* COX-1, and *Bos taurus*) are produced with high mean $RMSE_{test}$ values. These observations indicate that PCM models are not always able to extrapolate in the chemical or the target space if a given target or compound family is not sufficiently represented in the dataset.

Figure 7. Influence of compound substructures on potency and selectivity on human COX-1 and COX-2. Rows in the heatmap are indexed by the isoenzyme type whereas columns correspond to compound substructures. Substructures are depicted in red within arbitrary molecules presenting it. The color represents the average influence (pIC50 units) of each substructure on bioactivity. Red corresponds to an average increase in bioactivity, whereas blue indicates the a deleterious effect. Well-known chemical moieties, e.g. pyrrole rings (**c**), were singled out as selectivity determinants. For instance, substructure **d** is present in sulfonamides such as diflumidone, and substructure **b** in selective 1,2-diarylpyrroles COX-2 inhibitors.

Tables

UniProt ID	Type	Organism	Number of Bioactivities	% Compounds
P23219	1	*Homo sapiens*	1,346	41.7
O62664	1	*Box taurus*	48	1.5
P22437	1	*Mus musculus*	50	1.5
O97554	1	*Oryctolagus cuniculus*	11	0.3
P05979	1	*Ovis aries*	442	13.7
Q63921	1	*Rattus Norvegicus*	23	0.7
P35354	2	*Homo sapiens*	2,311	71.6
O62698	2	*Bos taurus*	21	0.7
Q05769	2	*Mus musculus*	305	9.4
P79208	2	*Ovis aries*	341	10.6
P35355	2	*Rattus Norvegicus*	39	1.2

Table 1. Composition of the COX dataset. The total number of bioactivities, after duplicate removal and selected from ChEMBL as described in Materials and Methods, and of distinct compounds are 4,937 and 3,228 respectively. The last column indicates the percentage of the total number of distinct compounds (3,228) annotated on each target.

		q^2_{int}	$RMSE_{int}$	$R^2_{0\ test}$	$RMSE_{test}$	q^2_{test}	
A	GBM	0.59	0.77	0.60	0.76	0.60
	RF	0.60	0.78	0.61	0.79	0.61
	SVM	0.61	0.75	0.60	0.76	0.60
B	Family QSAR	0.17	1.13	0.17	1.09	0.17
	Family QSAM	0.16	1.10	0.16	1.10	0.16
C	Greedy Ensemble Best	-	0.73	0.63	0.73	0.63
	MS Linear Ensemble Best	0.63	0.73	0.63	0.73	0.63
	MS EN Ensemble Best	0.63	0.72	0.62	0.72	0.62
	MS SVM Linear Ensemble Best	0.63	0.73	0.62	0.73	0.63
	MS SVM Radial Ensemble Best	0.63	0.73	0.63	0.73	0.63
	MS RF Ensemble Best	0.61	0.76	0.58	0.77	0.58
D	Greedy Ensemble	-	0.73	0.64	0.72	0.64
	MS Linear Ensemble	0.63	0.73	0.64	0.72	0.64
	MS EN Ensemble	0.64	0.73	0.63	0.73	0.63
	MS SVM Linear Ensemble	0.64	0.73	0.64	0.71	0.64
	MS SVM Radial Ensemble	0.64	0.73	0.65	0.71	0.65
	MS RF Ensemble	0.64	0.73	0.63	0.73	0.63

Table 2. Internal and external validation metrics for the PCM, QSAM and QSAR models. "Best" refers to the ensembles trained on only the three most predictive RF, GBM and SVM models. MS of models trained with different algorithms in a models ensemble allows to increase predictive ability, as the highest $R^2_{0\ test}$ and $RMSE_{test}$ values, 0.652 and 0.706 pIC50 units respectively, were obtaind with the "MS SVM Radial Ensemble".
Abbreviations. CV: cross-validation; EN: Elastic Net; Ext.: external; GBM: Gradient Boosting Machine; MS: Models Stacking; RF: Random Forest; RMSEP: root mean square error in prediction; SVM: Support Vector Machines.

Supporting Information Legends

Supplementary Material, Figure S1. Statistiscs of the repeated bioactivities for each compound-target combination. A. The abcissa represents the mean value for the bioactivities repeated for each compound-target combination with more than one annotated bioactivity. The ordinate represents their standard deviations. Repeated bioactivities are equally distributed for low, moderate and high affinity COX inhibitors. **B.** Histogram of the standard deviation of the repeated bioactivities. The distribution is strongly skewed towards 0, thus indicating that the differences between repeated bioactivities are generally negligible.

Supplementary Material, Figure S2. PCA of the compound descriptors used to train the PCM models. The PCA was performed on the pairwise Pearson rank correlation matrix calculated with the compound descriptors used to train the models. The two first principal components (PC) explain 58.03% of the variance. COX-1 and COX-2 are represented with squares and triangles respectively. Overall, the overlap between the datapoints indicate that the compounds annotated on different targets cover the same regions of the chemical space.

Supplementary Material, Figure S3. Y-scrambling. Scatterplots corresponding to the percentage of bioactivities randomized, against (**A**) $R^2_{0\ test}$ and (**B**) $\mathrm{RMSE}_{\mathrm{test}}$ values. The intercept in **A** becomes negative when 25-50% of the bioactivity variable is randomized. This finding indicates that PCM performance is not the consequency of spurious correlations in the descriptor space.

Supplementary Material, Figure S4. Jaccard pairwise similarity distributions for the compounds annotated on each target. Compounds annotated on the human cyclooxygenases (annotated with a star in the plots) display compound similarity distributions with mean values skewed towards 1. By contrast, compounds annotated on targets with less than 30 annotated bioactivities display multimodal similarity distributions. A correlation between model performance and both the number of datapoints and chemical diversity was established (see main text). Distributions were calculated with the same descriptors than the ones used to train the PCM models.

Supplementary Material, Figure S5. Compound substructures predicted to increase the bioactivity on human COX-2. The 20 substructures predicted to have the highest influence on bioactivity on human COX-2 (P35354) are plotted. Known chemical moieties such as pyrrole rings (1), aryl substituents (*e.g.* 4 and 5) or benzylsulfonamide (17) are represented. These substructures appear in diverse NSAIDs such as rofecoxib or etericoxib, as well as in chemical families of COX-2 inhibitors based on *e.g.* 1,5-diarylpyrazoles or 3,4-diaryl-substituted furans. (23; 24; 25)

Supplementary Material, Figure S6. Compound substructures predicted to have the same influence on human COX-1 and COX-2 Sub-strucutures predicted to decrease bioactivity are accompanied by a blue arrow, whereas that predicted to increase bioactivity are followed by a red arrow. Smaller substructures are found in this case, predominating substituents on the benzene ring. Therefore, substructure-activity relationships are difficult to be determined.

Figure
Click here to download Figure: Figure1.eps

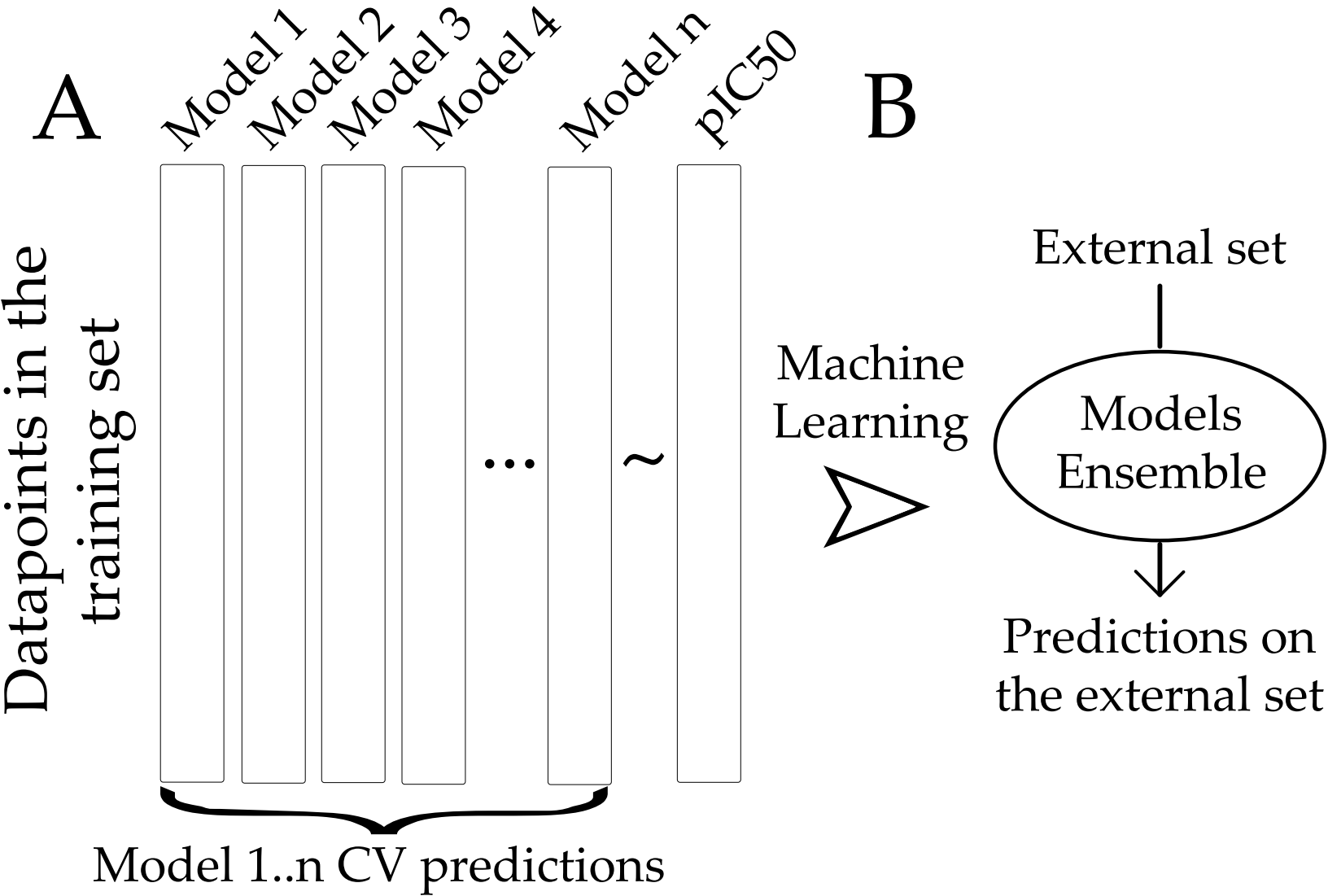


Figure
Click here to download Figure: Figure2.eps

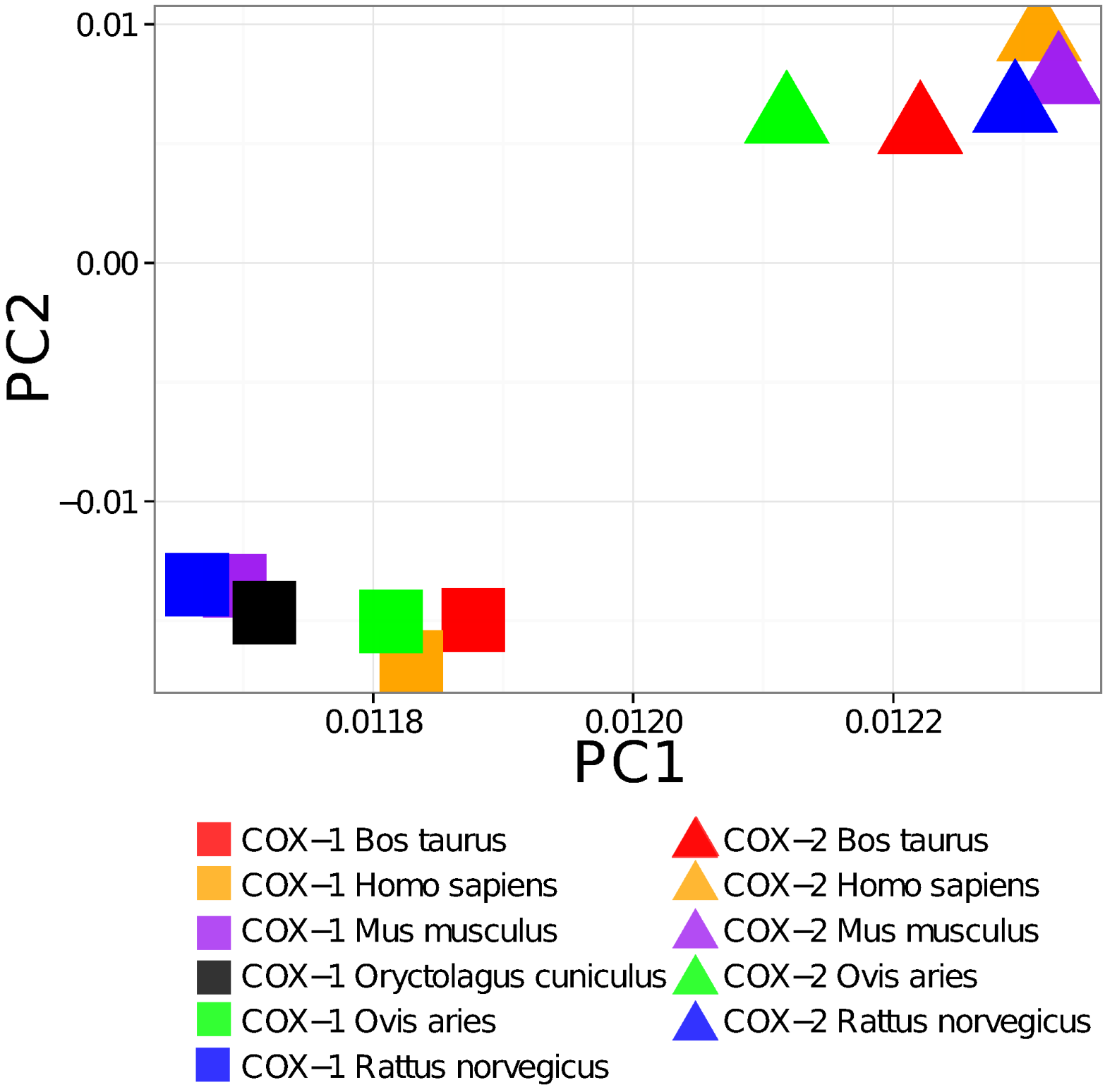


Figure
Click here to download Figure: Figure3.eps

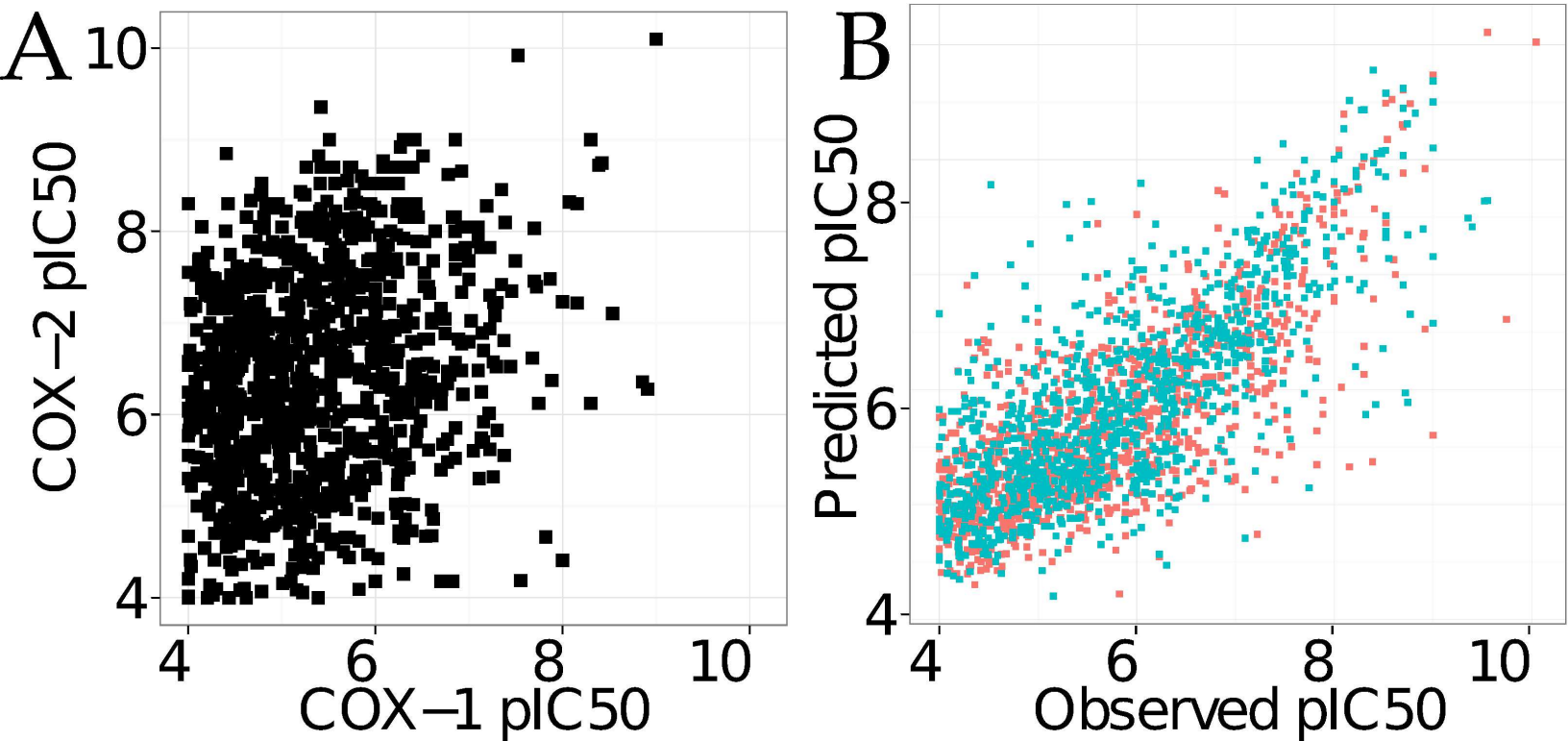


Figure
Click here to download Figure: Figure4.eps

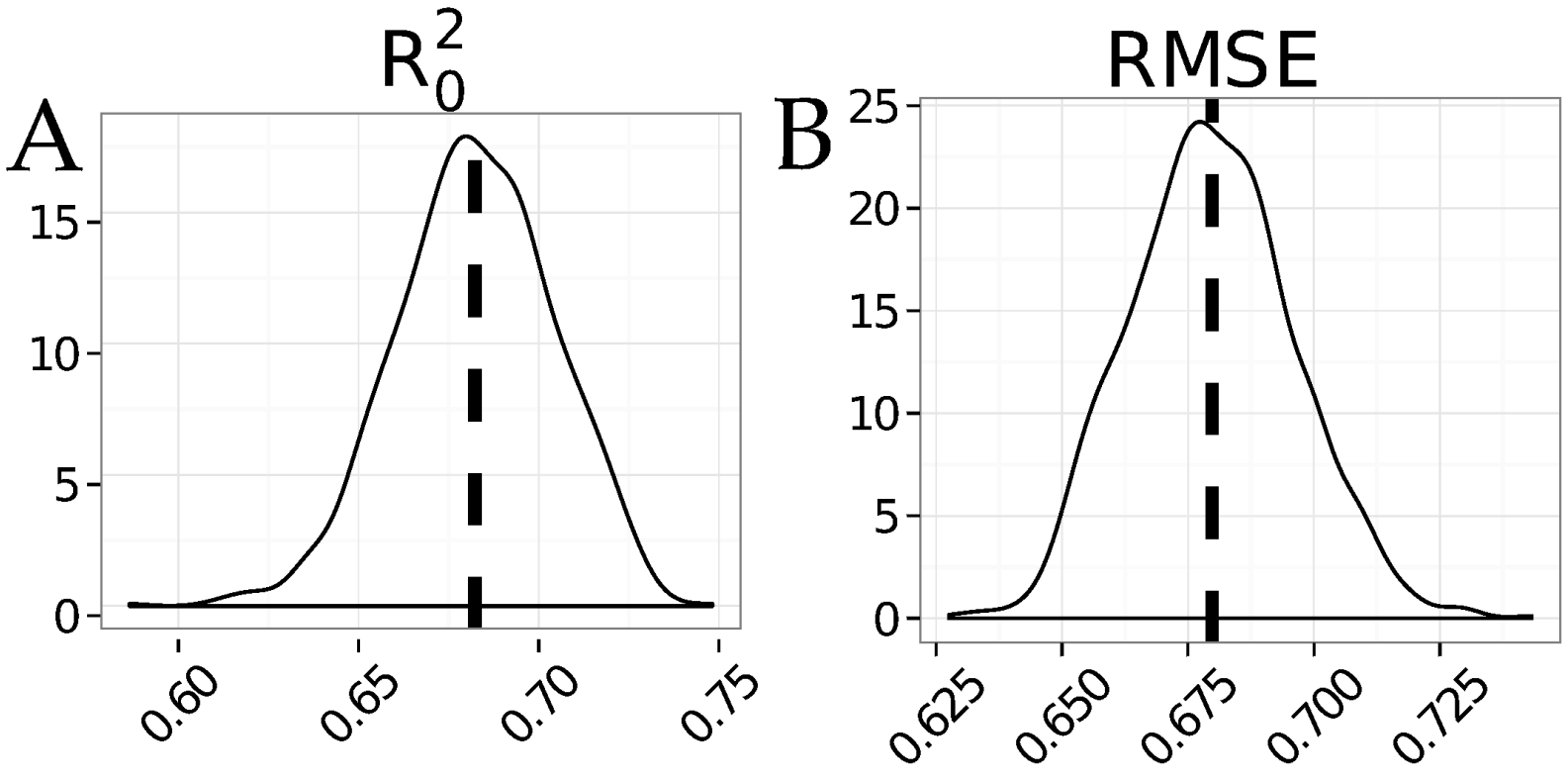


Figure
Click here to download Figure: Figure5.eps

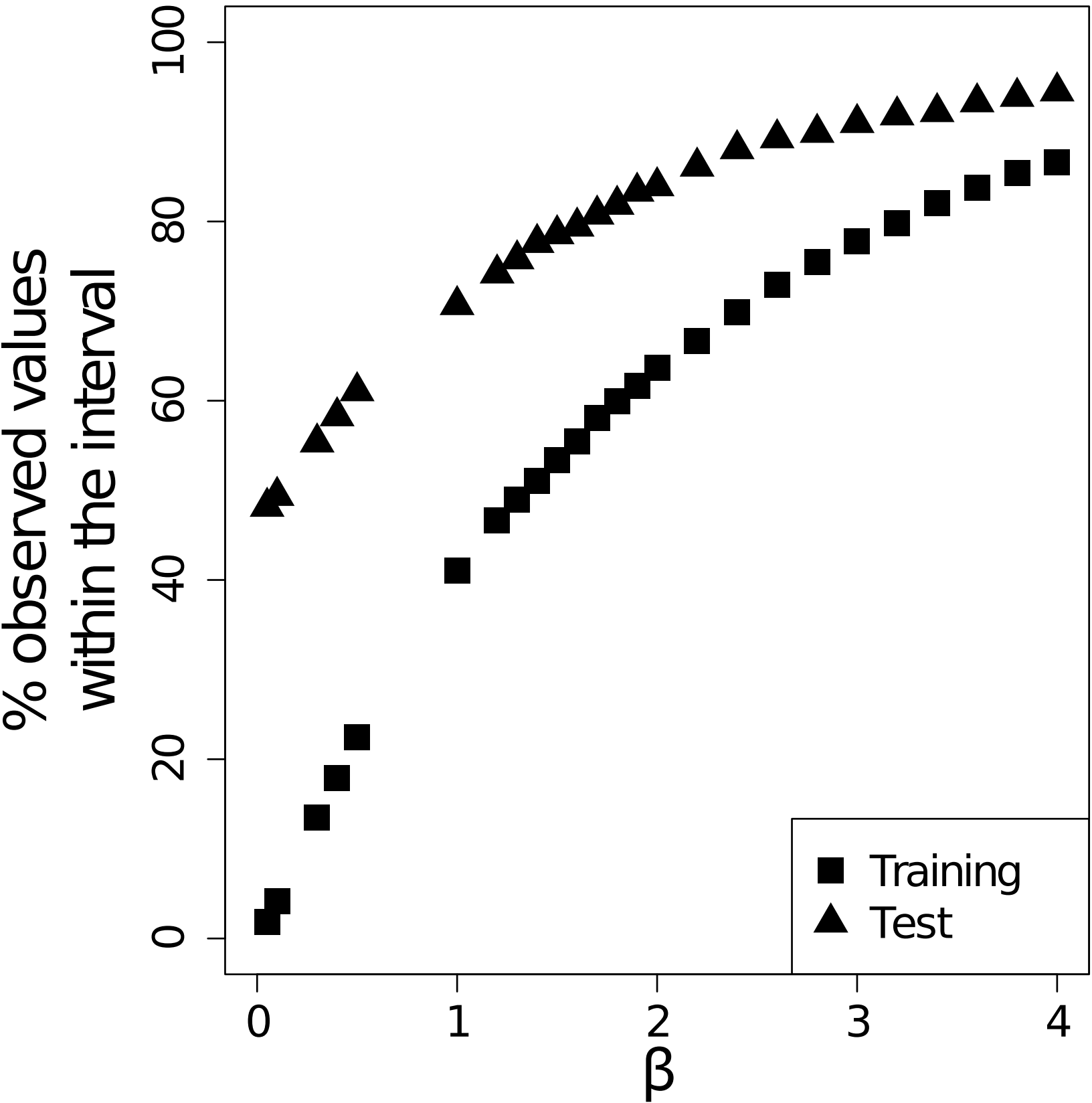


Figure
Click here to download Figure: Figure6.eps

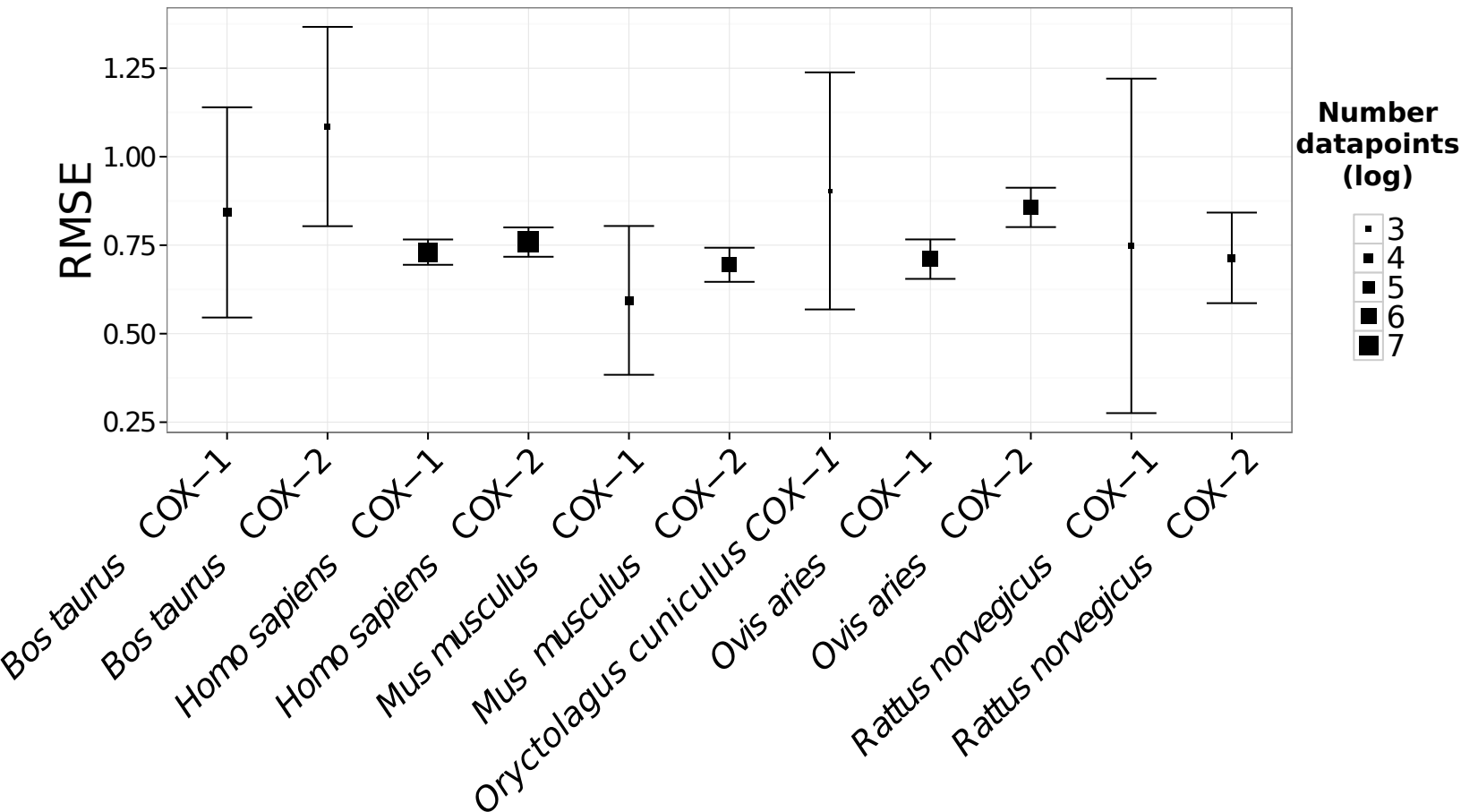


Figure
Click here to download Figure: Figure7.eps

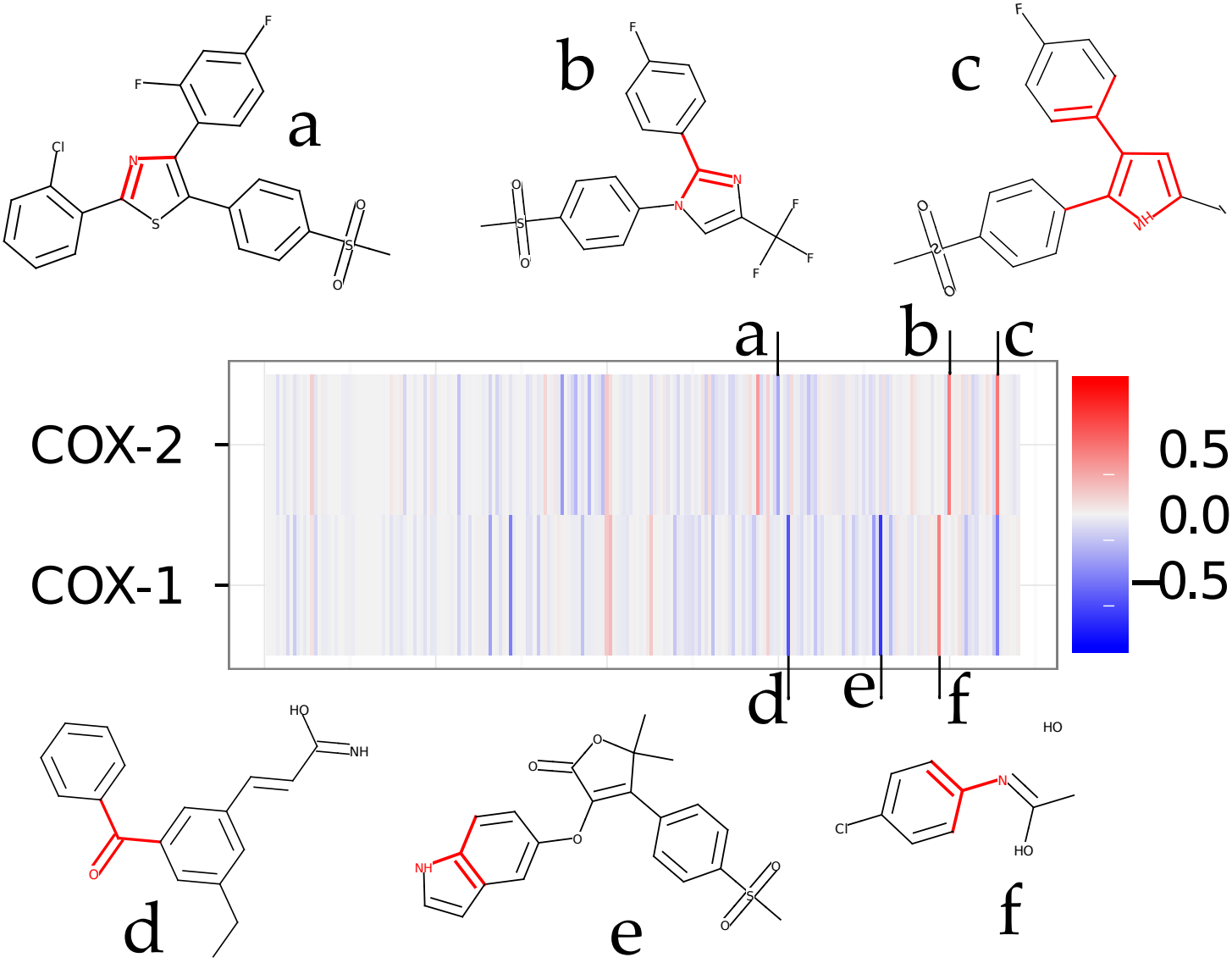


Figure
Click here to download Figure: FigureS1.eps

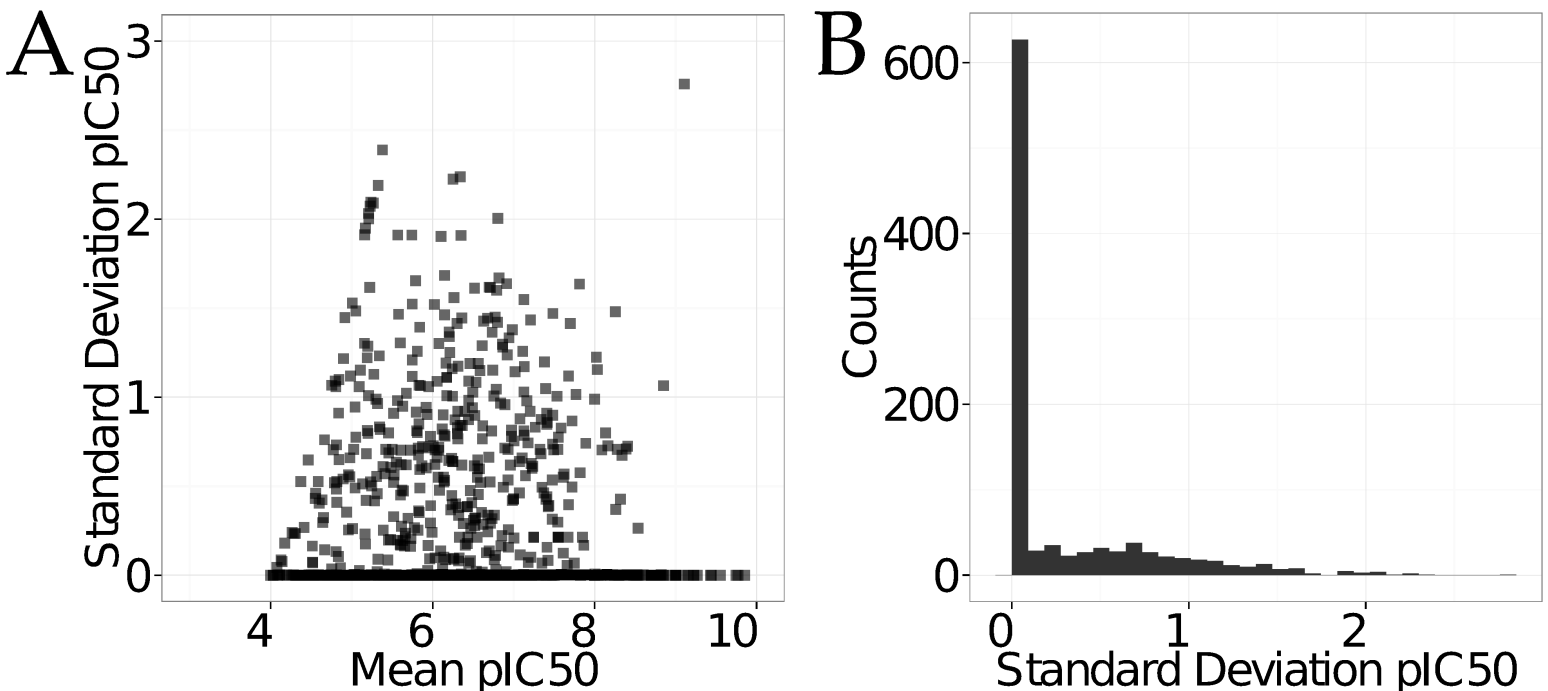


Figure
Click here to download Figure: FigureS2.eps

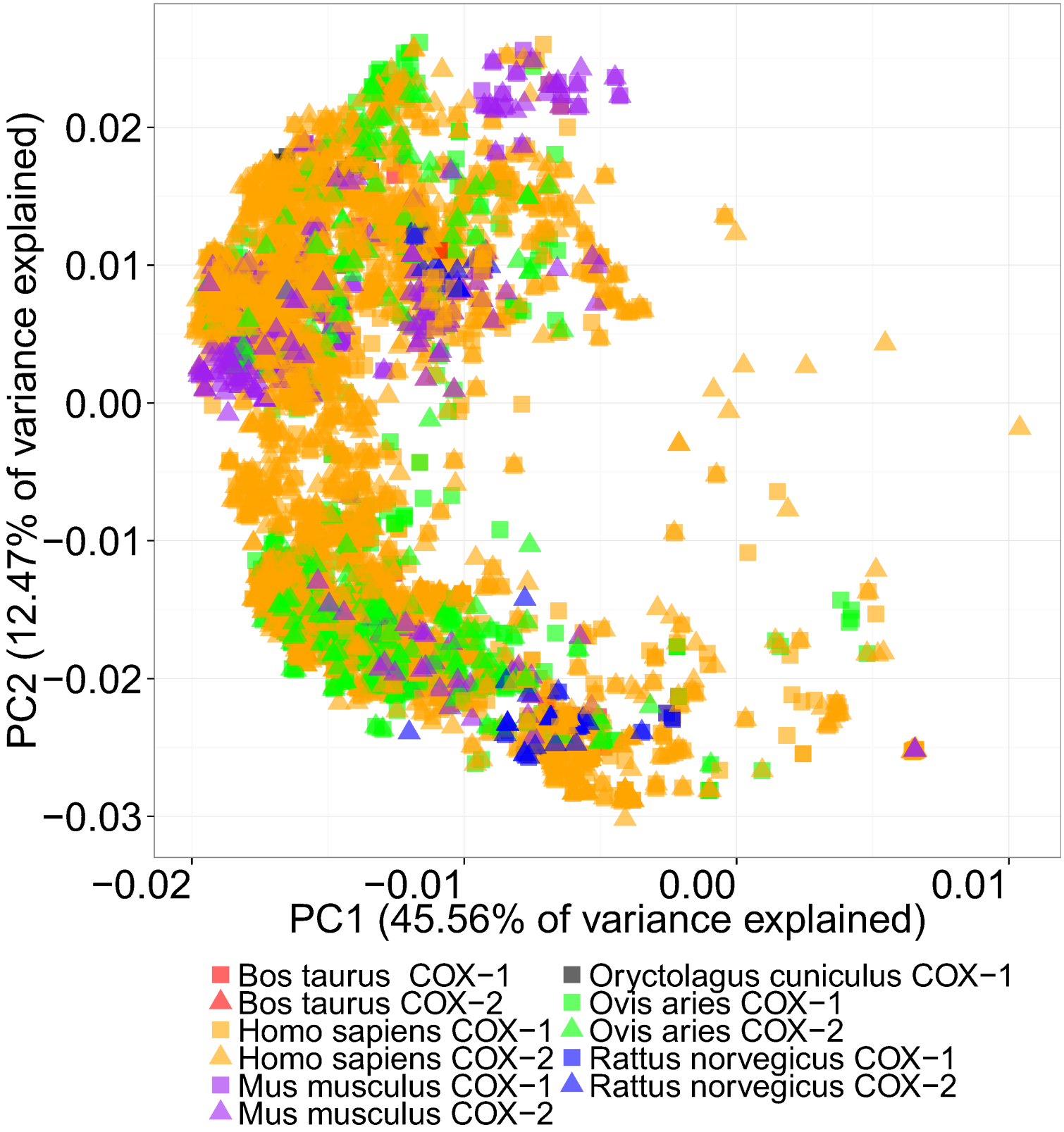


Figure
Click here to download Figure: FigureS3.eps

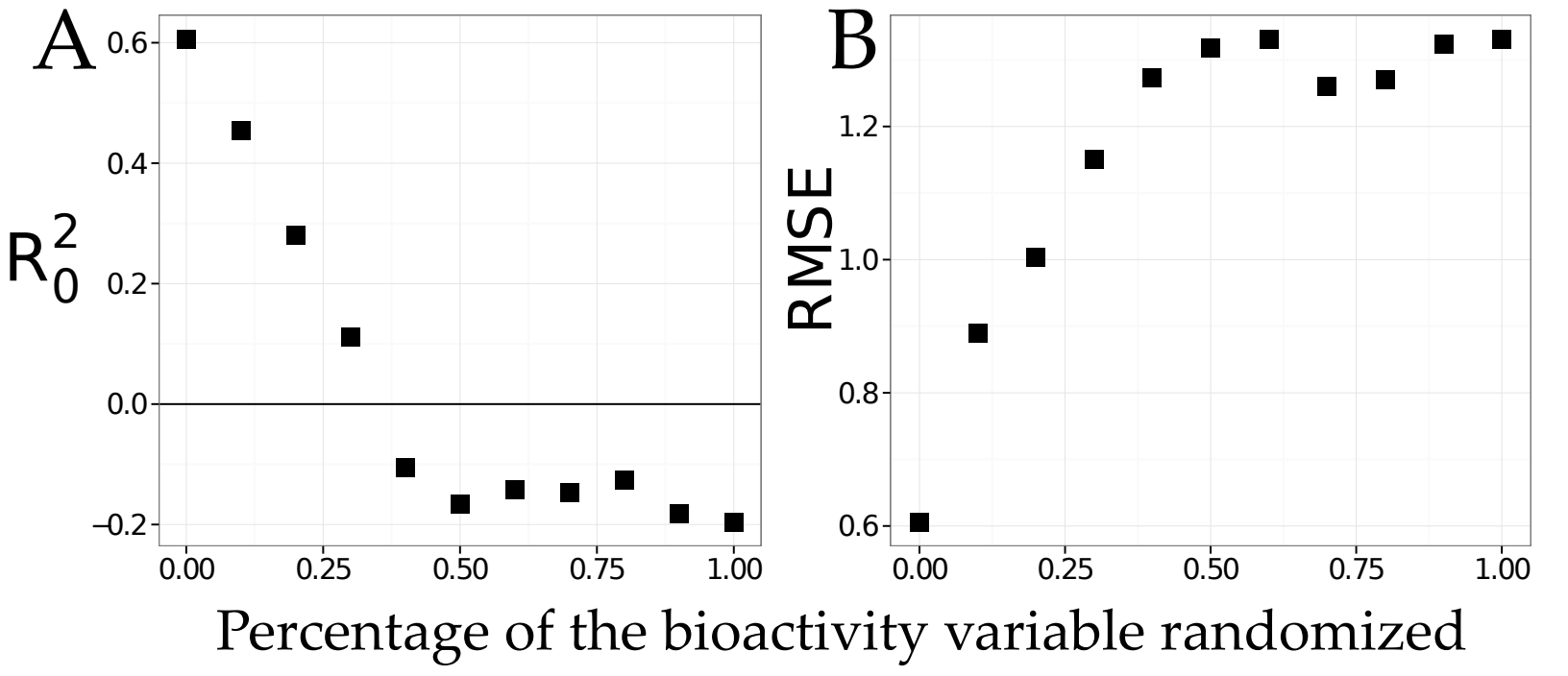
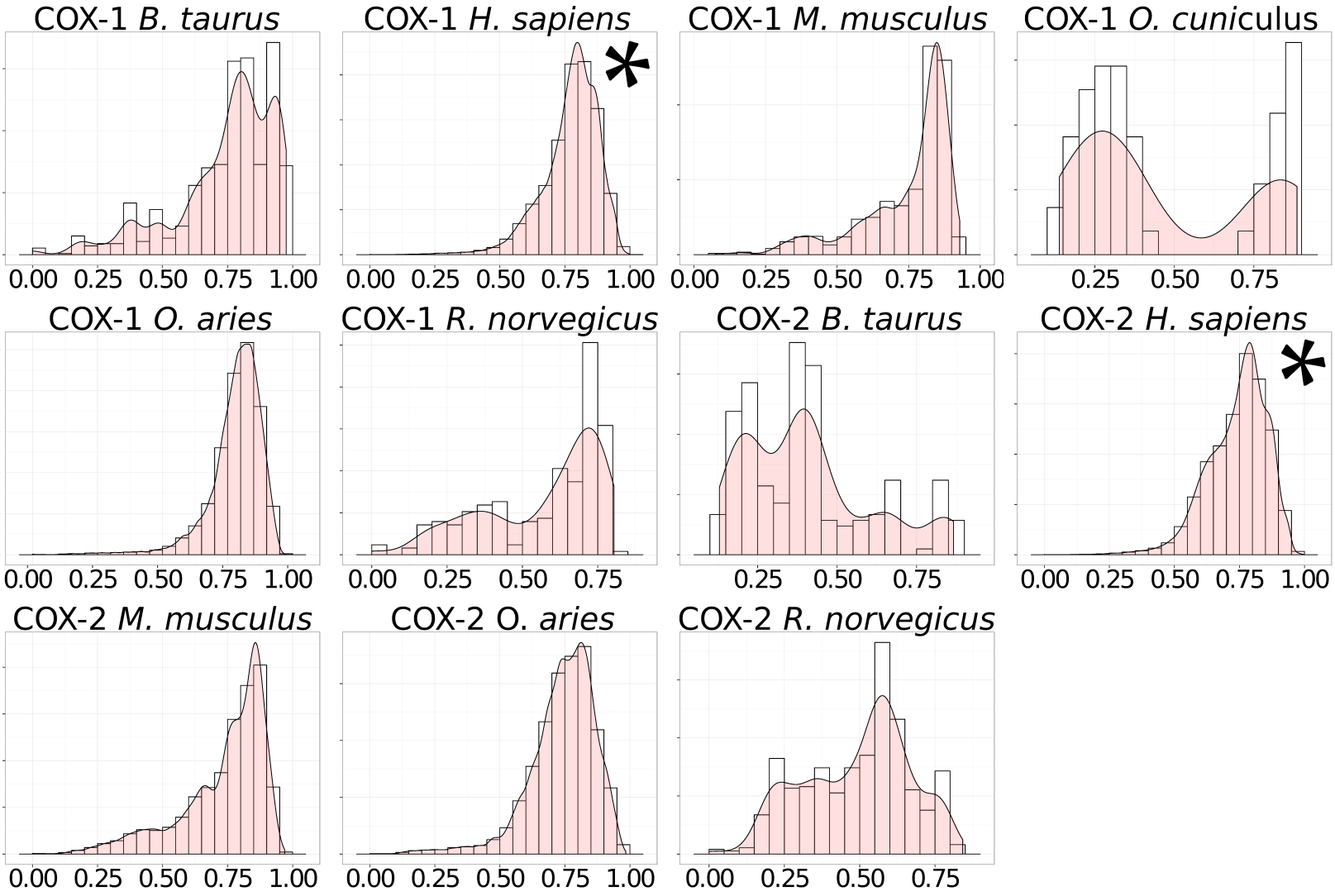
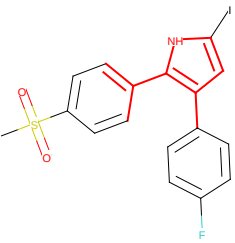


Figure
Click here to download Figure: FigureS4.eps

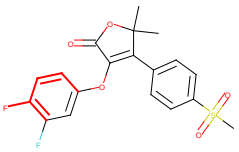


Figure

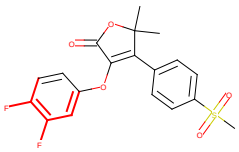
Click here to download Figure: FigureS5.eps



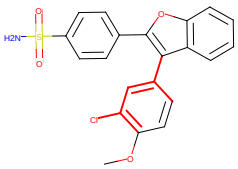
1



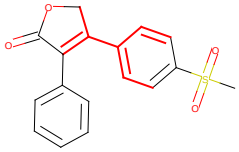
2



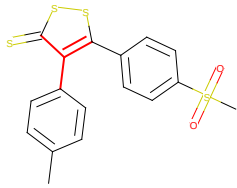
3



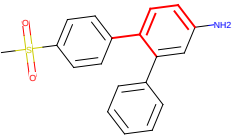
4



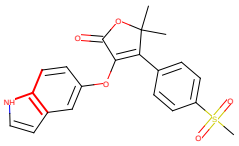
5



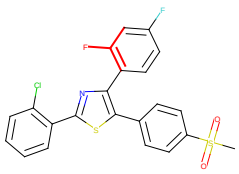
6



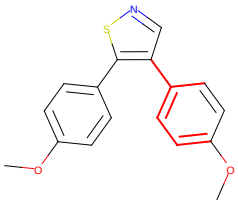
7



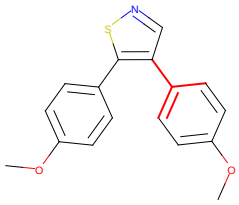
8



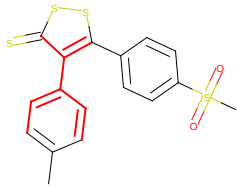
9



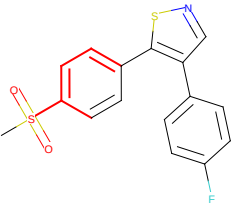
10



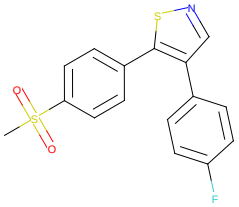
11



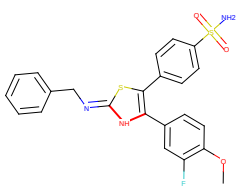
12



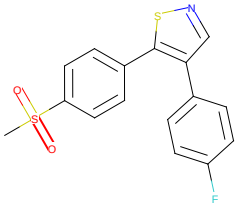
13



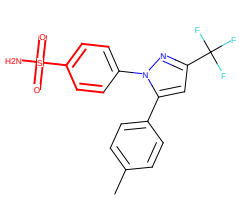
14



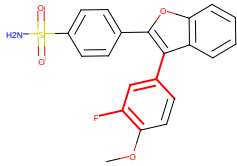
15



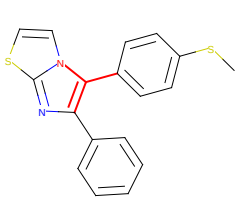
16



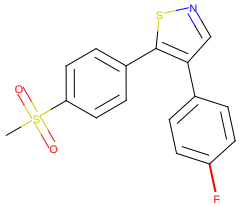
17



18

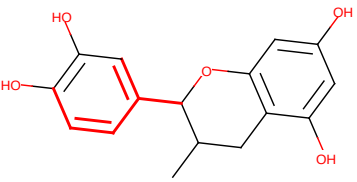


19

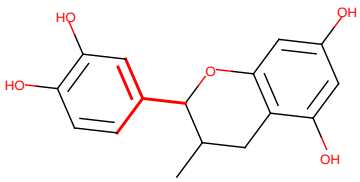


20

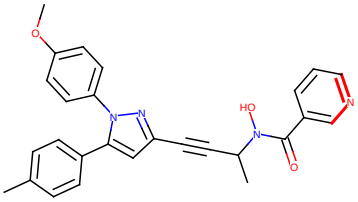
Figure
Click here to download Figure: FigureS6.eps



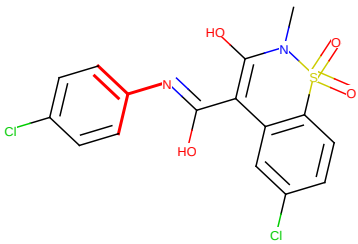
1 ↓



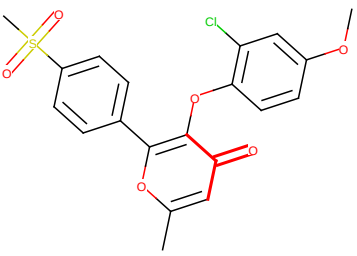
2 ↓



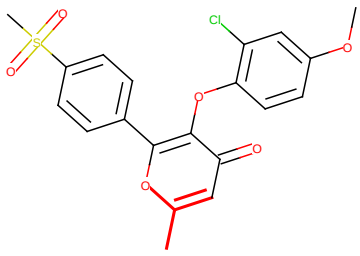
3 ↓



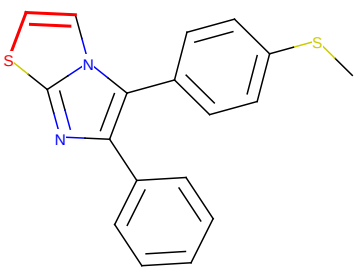
4 ↓



5 ↓



6 ↓



7 ↑

Supporting Information

Click here to download Supporting Information: final_alignment_10_paper.xls

Supporting Information - Compressed/ZIP File Archive

Click here to download Supporting Information - Compressed/ZIP File Archive: dataset_paper.csv.zip