



Universiteit
Leiden
The Netherlands

Transforming data into knowledge for intelligent decision-making in early drug discovery

Paricharak, S.A.

Citation

Paricharak, S. A. (2017, February 9). *Transforming data into knowledge for intelligent decision-making in early drug discovery*. Retrieved from <https://hdl.handle.net/1887/45874>

Version: Not Applicable (or Unknown)

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/45874>

Note: To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/45874> holds various files of this Leiden University dissertation

Author: Paricharak, S.A.

Title: Transforming data into knowledge for intelligent decision-making in early drug discovery

Issue Date: 2017-02-09

Chapter six

Proteochemometric Modelling Coupled to *in Silico* Target Prediction: an Integrated Approach for the Simultaneous Prediction of Polypharmacology and Binding Affinity/Potency of Small Molecules

Reproduced from Shardul Paricharak, Isidro Cortés-Ciriano, Adriaan P. IJzerman, Thérèse E. Malliavin, and Andreas Bender. (2015) *J. Cheminform.* 7, 15-25.

Abstract

The rampant increase of public bioactivity databases has fostered the development of computational chemogenomics methodologies to evaluate potential ligand-target interactions (polypharmacology) both in a qualitative and quantitative way. Bayesian target prediction algorithms predict the probability of an interaction between a compound and a panel of targets, thus assessing compound polypharmacology qualitatively, whereas structure-activity relationship techniques are able to provide quantitative bioactivity predictions. We propose an integrated drug discovery pipeline combining *in silico* target prediction and proteochemometric modelling (PCM) for the respective prediction of compound polypharmacology and potency/affinity. The proposed pipeline was evaluated on the retrospective discovery of *Plasmodium falciparum* DHFR inhibitors. The qualitative *in silico* target prediction model comprised 553084 ligand-target associations (a total of 262174 compounds), covering 3481 protein targets and used protein domain annotations to extrapolate predictions across species. The prediction of bioactivities for plasmodial DHFR led to a recall value of 79% and a precision of 100%, where the latter high value arises from the structural similarity of plasmodial DHFR inhibitors and *T. gondii* DHFR inhibitors in the training set. Quantitative PCM models were then trained on a dataset comprising 20 eukaryotic, protozoan and bacterial DHFR sequences, and 1505 distinct compounds (in total 3099 data points). The most predictive PCM model exhibited R^2_{test} and $\text{RMSE}_{\text{test}}$ values of 0.79 and 0.59 pIC₅₀ units respectively, which was shown to outperform models based exclusively on compound ($R^2_{\text{test}}/\text{RMSE}_{\text{test}} = 0.63/0.78$) and target information ($R^2_{\text{test}}/\text{RMSE}_{\text{test}} = 0.09/1.22$), as well as inductive transfer knowledge between targets, with respective R^2_{test} and $\text{RMSE}_{\text{test}}$ values of 0.76 and 0.63 pIC₅₀ units. Finally, both methods were integrated to predict the protein targets and the potency on plasmodial DHFR for the GSK TCAMS dataset, which comprises 13533 compounds displaying strong anti-malarial activity. 534 of those compounds were identified as DHFR inhibitors by the target prediction algorithm, while the PCM algorithm identified 25 compounds, and 23 compounds (predicted pIC₅₀ > 7) were identified by both methods. Overall, this integrated approach simultaneously provides target and potency/affinity predictions for small molecules.

Introduction

In recent years it has been demonstrated that drugs exert their therapeutic effect by modulating more than one target, in fact six on average.¹ Therefore, the early evaluation of the bioactivity profiles of lead compounds is essential for the success in developing new drugs, although efficacy is sometimes attained by the inhibition of single targets, e.g., viral proteins. Similarly, understanding drug polypharmacology can help in anticipating drug adverse effects.²

In parallel, the availability of public bioactivity databases has enabled the application of large-scale chemogenomics techniques to, among others, predict protein targets for small molecules, and to predict their affinity on therapeutically interesting targets.³ These techniques capitalize on bioactivity data to infer relationships between the compounds, encoded with numerical descriptors, and their targets, which can be represented as labels in a classification model or explicitly encoded by e.g., protein or amino acid descriptors.⁴

In silico target prediction algorithms assess potential compound polypharmacology through the computational evaluation of the (functionally unrelated) targets modulated by a given compound, or its selectivity to species-specific targets, as they predict the probability of interaction of that compound with a panel of targets.⁵ Initially, target prediction models were developed using Laplacian-modified Naïve Bayesian classifiers⁶ and the Winnow algorithm.⁷ Later, Keiser *et al.*⁸ developed a model which related biological targets based on ligand similarities and ranked the significance of the resulting similarity scores using the Similarity Ensemble Approach (SEA), followed by Wale and Karypis⁹ who applied SVM and ranking perceptron algorithms to rank targets for a given compound. More recently, Koutsoukas *et al.*¹⁰ compared the performance of both the Naïve Bayesian and Parzen-Rosenblatt Window classifiers, concluding that the overall performance of both methods is comparable though differences were found for certain target classes.

The ligand-target prediction methods described above generally predict the likelihood of interaction with a target, and they do not predict compound affinity or potency (e.g., K_i or IC_{50}). On the other hand, quantitative bioactivity prediction techniques, e.g., proteochemometric modeling (PCM),³ predict the potency or affinity for compound-target pairs, normally in the form of pIC_{50} or pK_i values. PCM combines information from compounds and related

targets, e.g., orthologs, in a single machine learning model,^{3,11} which enables the simultaneous modeling of chemical and biological information, and thus the prediction of compound affinity and selectivity across a panel of targets. Nonetheless, the effects of a compound at the cellular or the organism level are poorly understood in this case, as these methods cannot account for the interactions of a compound with other unrelated targets, which are not captured in the PCM model.

Given the limitations of both purely qualitative and purely quantitative bioactivity modeling approaches, in the current work, we propose an integrated drug discovery approach, combining *in silico* target prediction for the qualitative large-scale evaluation of compound bioactivity, and PCM for the quantitative prediction of compound potency. The proposed approach was evaluated on the discovery of DHFR inhibitors for *Plasmodium falciparum* (*P. falciparum*), the causative agent of the most dangerous form of malaria.¹² Whilst there are multiple anti-malarial drugs on the market, resistance to anti-malarial drugs is on the rise,^{13,14} and there are only 21 compounds in clinical or pre-clinical trials.¹⁵

In order to combat the lack of novel drugs for malaria, big pharmaceutical companies have generated a wealth of phenotypic data, namely the GlaxoSmithKline (GSK) TCAMS dataset, as well as the Novartis-GNF Malaria Box.^{16,17} Both datasets contain phenotypic readouts, describing how effective the compounds present in the datasets are in inhibiting the growth of *P. falciparum*. Nonetheless, none of them contain annotations about the *P. falciparum* target(s) involved, making it a challenge to elucidate the mode of action (MoA) of the compounds in the dataset, and hence, making the dataset difficult to interpret. This renders these datasets a very suitable case study for the algorithms we are presenting in this work.

In the context of malaria drug discovery, previous studies have applied machine learning algorithms to predict whether plasmodial proteins are secretory proteins based on their residue composition,¹⁸ and to predict the bioactivities of compounds against particular plasmodial targets.^{19,20} These approaches, though, did not account for the polypharmacology of anti-malarial compounds.

To overcome the limitations of these methods, we now integrate both *in silico* target prediction and PCM in a unified drug discovery approach. As illustrated in **Figure 29**, the target prediction algorithm used in this study, trained on approximately 553084 bioactivity data points spanning 3481

targets, used a domain-based similarity metric between targets to extrapolate target predictions from one species to another.

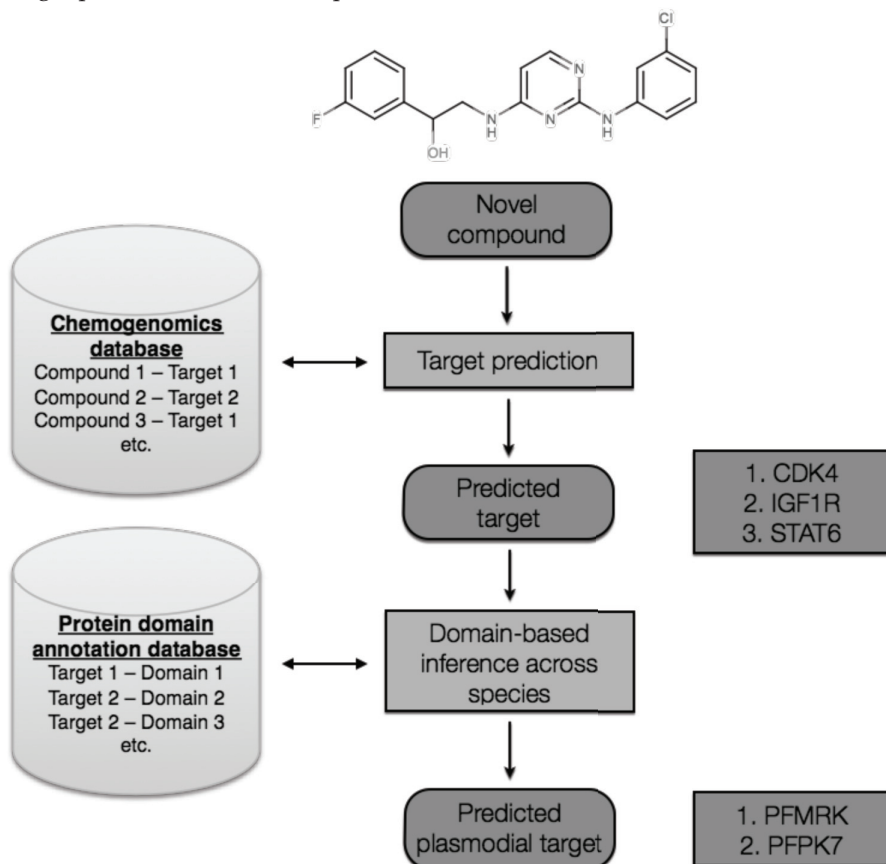


Figure 29. Schematic overview of *in silico* target prediction and domain-based extrapolation workflow. The conventional *in silico* target prediction approach¹⁰ is extended in this study by using protein domain annotations to extrapolate from non-plasmodial target predictions to protein target predictions in *P. falciparum*. This concept is generally applicable across organisms, in particular to those for which little bioactivity data is currently available.

Non-plasmodial targets were then extrapolated to plasmodial targets. Besides, the PCM model was trained on a dataset composed of 20 eukaryotic, protozoan and bacterial DHFR sequences, and of 1505 different DHFR inhibitors and a total of 3099 data points. To exploit the complementarity of the two prediction methods, *in silico* target prediction was used to predict MoA hypotheses for the anti-malarial compounds in the GSK TCAMS phenotypic dataset, whereas PCM was employed to quantify compound potency (pIC₅₀).

Methods

Exploratory Principal Component Analysis (PCA) of PCM and Target Prediction Datasets. A PCA was performed for compounds contained in the PCM dataset, as well as for those annotated on *P. falciparum* and *T. gondii* in the target prediction dataset. The Spearman's rank correlation coefficient was calculated for all pairs of compound descriptors, based on both physicochemical descriptors and Morgan fingerprints, thus defining a square correlation matrix. The PCA analysis was performed on this matrix in order to avoid the direct application of PCA on binary descriptors, i.e., Morgan fingerprints. Visualization was performed using R and Vortex.²¹

Target Prediction.

Training dataset. Bioactivity data were extracted from ChEMBL16²² according to the protocol described by Koutsoukas *et al.*¹⁰ The extracted data contained approximately 4 million bioactivities covering approximately 8000 biomolecular targets, of which approximately 4000 targets were proteins.^{22,23} Compound-target pairs were selected according to the following criteria: (i) K_i , K_d , IC_{50} or EC_{50} bioactivity values equal to or lower than 10 μM , and (ii) targets annotated with a confidence score of 8 (homologous single protein target assigned) or 9 (direct single protein target assigned). Subsequently, ligand structures were processed with the ChemAxon standardizer version 5.12.0,²⁴ with the following options: "Remove fragment", "Neutralize", "Aromatize", "Clean2D", "Tautomerize" and "Remove explicit hydrogens". After standardization, the entries with ligands annotated against multiple targets were detected based on their canonical SMILES and removed using custom Perl scripts, resulting in a training set of 553084 instances (262174 compounds) covering 3481 protein targets. The bioactivity data of *P. falciparum* (1513 instances – 1379 compounds covering 41 protein targets) was omitted from this dataset for training purposes. InterPro²⁵ domain annotations were retrieved for all protein targets using the Uniprot database.²⁶

P. falciparum dataset. The *P. falciparum* dataset was built using the same criteria as described above, resulting in a set comprising 41 *P. falciparum* targets and 1379 compounds. In addition, all annotated and reviewed *P. falciparum* targets from Uniprot were downloaded, resulting in a total of 148 *P. falciparum* protein targets. Finally, InterPro domain annotations were retrieved for all protein targets using the Uniprot database.

GSK TCAMS dataset. Approximately 2 million compounds present in GSK's screening collection have been tested in vitro by GSK for inhibitors of *P.*

falciparum's intraerythrocytic cycle based on growth inhibition assays.¹⁷ Briefly, assays were performed on both the reference laboratory strain 3D7, as well as on the multidrug resistant strain Dd2, where parasite growth was evaluated using LDH activity.¹⁷ 19451 compounds were identified as primary hits inhibiting the 3D7 strain growth by more than 80% at 2 μ M concentration, of which 13533 compounds displayed 80% or higher inhibition of parasite growth in at least 2 of the 3 assay runs in independent follow-up experiments. Hence, these 13533 compounds were considered as confirmed inhibitors (confirmation rate > 70%).

Descriptors. A circular fingerprint implementation, Molprint2D^{27,28} was used for encoding molecular structures, since this method has previously been shown to capture structural aspects related to bioactivity better than most other descriptors in comparative studies.²⁹ This descriptor is based on count vectors of heavy atoms present at a topological distance from each heavy atom of a molecule.²⁸ For the present study, the pybel implementation was used.³⁰

Target prediction algorithm. A multiclass Laplacian-modified Naïve Bayesian classifier, as described by Nigsch *et al.*⁷ and later implemented by Koutsoukas *et al.*¹⁰ was implemented to classify the bioactivity dataset and to be able to predict targets for novel compounds. For the query molecule $\mathbf{x}$, consisting of a set of n Molprint2D features f_i , the likelihood to be active against a protein target ω_α was calculated using the following equation:

$$S_{\omega_\alpha}(\mathbf{x}) = \sum_{i=1}^n \log \left(\frac{N_{i,\omega_\alpha} + 1}{N_i \times p(\omega_\alpha) + 1} \right) + \log \left(\frac{\prod_{i=1}^d p(f_i)}{p(\mathbf{x})} \right) \quad (5)$$

where $S_{\omega_\alpha}(\mathbf{x})$ is the logarithmic likelihood score (proportional to the likelihood of bioactivity), N_{i,ω_α} is the total number of occurrences of feature f_i in protein class ω_α and N_i is the total number of occurrences of feature f_i in the entire training set. Furthermore, $p(\omega_\alpha)$ is the prior probability of protein class ω_α . The prior probability quantifies how likely a compound is active against protein target ω_α in the absence of any feature information. It can be calculated as follows:

$$p(\omega_\alpha) = \frac{N_{\omega_\alpha}}{N} \quad (6)$$

where N_{ω_α} is the number of instances (i.e., bioactivities) in class ω_α and N is the total number of instances. The predictive performance of this model was

assessed in terms of average class-specific recall and precision. Only target classes with 20 or more data points in the *P. falciparum* dataset were considered as suitable for testing due to a sufficient number of data points, resulting in a total of 16 target classes.

Domain-based extrapolation to P. falciparum targets. For each analyzed compound, the top *n* ranked predicted targets were compared to all 148 *P. falciparum* targets in terms of their InterPro domain composition. *P. falciparum* targets with an InterPro domain Tanimoto similarity above a variable cutoff were considered as predicted, but were not ranked. The cutoff value varied between 0.5 and 1, where 1 means that only orthologous proteins are considered. The target prediction and domain-based extrapolation pipeline are illustrated in **Figure 29**. The domain extrapolation extends the target prediction approach^{10,31} by using InterPro protein domain annotations to extrapolate from predicted non-plasmodial targets to *P. falciparum* targets. This is conceptually similar to a previously reported study for extrapolating bioactivities between species,³² and its application to *M. tuberculosis*.³³

The inclusion of plasmodial DHFR (ChEMBL1939) bioactivity data was expected to drastically improve the performance, and this was tested in the following way. A 2-fold cross validation (CV) was performed: the instances annotated on plasmodial DHFR were split into 2 half subsets, where one subset was added to the training set and the other half was used as a test set (and *vice versa*).

Proteochemometric Modeling.

Dataset. IC₅₀ values with a confidence score of 8 or 9 for 20 DHFR sequences were retrieved from ChEMBL16²² and this initial dataset comprised 5827 data points. In the cases where a compound-target combination had more than one annotated bioactivity value, the set of bioactivities was replaced by its mean value. This procedure is robust, because the standard deviation of the differences was smaller than 0.1 pIC₅₀ unit in more than 90% of the cases. This resulted in a dataset including 3099 distinct compound-target combinations. The matrix completeness of the dataset, calculated as the number of compound-target combinations present in the dataset over the total number of possible compound-target combinations, was 10.3%. Compounds included in the PCM dataset were not present in the target prediction dataset.

Descriptors. Chemical structures were standardized and cleaned with the function *StandardiseMolecules* of the R package *camb* using the default parameters³⁴ and PaDEL descriptors (1-D and 2-D). Morgan fingerprints were

calculated in the same environment. The function *AA_Descs* was used to calculate amino acid descriptors (3 Z-scales). To describe the target space, the residues in the binding site of human DHFR (PDB ID: 1OHJ)³⁵ within a sphere of 10 Å centered around the ligand were selected. The corresponding residues for the other 19 proteins were obtained from a sequence alignment realized with Clustal Omega.³⁶

Proteochemometric modeling. All descriptor values were centered to zero mean and scaled to unit variance. The dataset was split into six subsets, five of which were used to train models, and the sixth, test set, was withheld to assess the predictive ability of the models.³⁷ The hyperparameter values for all PCM models were optimized by 5-fold cross validation.³⁸ To assess both model predictive ability and performance, the pIC₅₀ values for the test set were predicted, thus providing the external validation by calculating RMSE_{test} and $R^2_{0\text{ test}}$ between the observed and the predicted pIC₅₀ values:

$$R^2_{0\text{ test}} = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i^{r0})^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (7)$$

$$RMSE = \sqrt{\frac{(y - \hat{y})^2}{N}} \quad (8)$$

where N represents the size of the test set, y_i the observed, $\hat{y}_i$ the predicted, and $\bar{y}$ the average pIC₅₀ values of those datapoints included in the test set, and $\hat{y}_i^{r0} = s\hat{y}$, with $s = \sum y_i \hat{y}_i / \sum \hat{y}_i^2$. Both internal (RMSE_{int} and R^2_{int}) and external validation (RMSE_{test} and $R^2_{0\text{ ext}}$) were assessed according to the criteria proposed by Tropsha *et al.*^{39,40} and calculated using the *Validation function* of the R package *camb*.³⁴

In order to assess whether the combination of compound and target information in a single PCM model constitutes an advantage with respect to one-space (ligand space and target space) models, two validation scenarios were explored. Firstly, a Family QSAR model⁴¹ was trained exclusively on compound descriptors. High performance of this model is expected in cases where the bioactivities of the same compound on different targets are highly correlated. Secondly, the Family QSAM⁴¹ model was trained on target descriptors only. In this case, high performance would indicate that the activities of a diverse set of compounds are correlated on a panel of targets. Thus, compound activities would largely depend on the target, and to a much lesser extent on the ligand structures.

Additionally, an inductive transfer PCM model (PCM IT) was trained to assess whether the performance of PCM models arises from explicit learning (EL), where the knowledge is extracted from target descriptors, or inductive transfer (IT). In IT the knowledge acquired when predicting compound bioactivities on a given target is exploited to predict the bioactivity of those compounds on another target.⁴¹ In the PCM IT model, targets were described with identity fingerprints (IFP), which are calculated as follows:

$$IFP(i, j) = \delta(i - j)(i, j \in 1, \dots, N_{targets}) \quad (9)$$

where δ is the Kronecker delta function and $N_{targets}$ the number of distinct targets. The performance of the models was assessed on a *per* target basis by training ten PCM models, each on a different subset of the whole dataset. Subsequently, $RMSE_{test}$ and $R^2_{0\ test}$ values were calculated on subsets of the test set grouped by target.

Machine learning implementation. Support Vector Machines (SVM),⁴² Gradient Boosting Machines (GBM),⁴³ Gaussian Processes (GP),⁴⁴ and Random Forest (RF)⁴⁵ models were built with the R package *camb*.^{34,46} The target prediction algorithm was implemented in Perl.

Results and discussion

Exploratory Analysis of PCM and Target Prediction Datasets. A PCA (Figure 30) was performed for the compounds annotated to be active against plasmodial DHFR and those active against *T. gondii* DHFR.

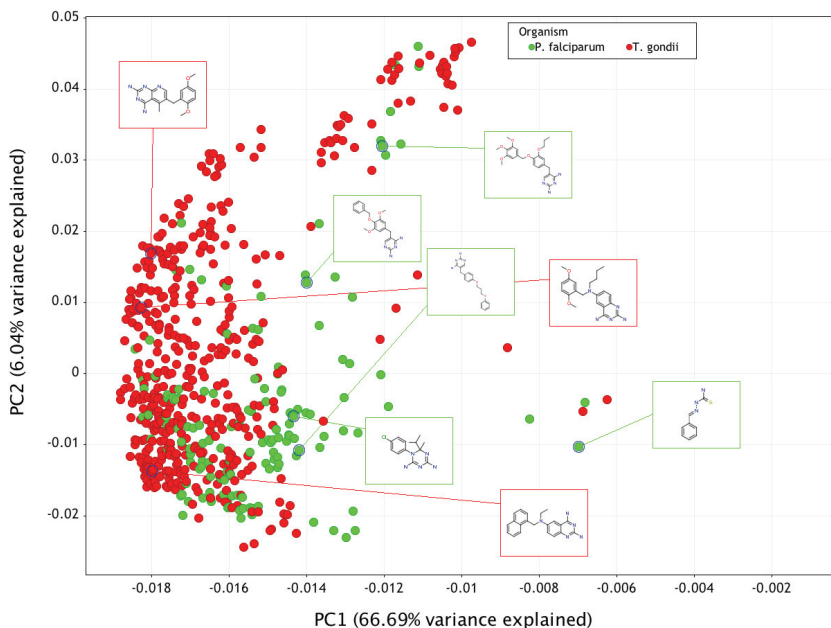


Figure 30. PCA of the compounds annotated as actives against plasmodial DHFR (green) as well as *T. gondii* DHFR (red). Overall, plasmodial DHFR inhibitors cover a substantial portion of the chemical space occupied by *T. gondii* DHFR inhibitors. However, some clusters of *T. gondii* DHFR inhibitors are located in additional chemical space not covered by the plasmodial inhibitors (red boxes). These clusters contain compounds with bicyclic ring systems. By contrast, plasmodial inhibitors only contain unfused rings (green boxes). These observations explain why recall is low (~35%) when plasmodial DHFR inhibitors are excluded from the training set: *T. gondii* inhibitors do not cover all relevant chemical space, particularly the space occupied by compounds with unfused ring systems.

The first two principal components explain 72.73% of the variance. In the two dimensions visualized for the descriptor space used here, the plasmodial inhibitors cover a substantial portion of the chemical space occupied by the *T. gondii* DHFR inhibitors. However, there are still a number of clusters of *T. gondii* DHFR inhibitors that occupy novel space not covered by plasmodial inhibitors. Compounds from these clusters contain bicyclic ring systems (shown in red boxes in **Figure 30**). On the other hand, there are also clusters of plasmodial inhibitors that occupy space not covered by *T. gondii* inhibitors: these plasmodial inhibitors do not contain bicyclic rings, but instead contain unfused rings (5 scaffolds identified shown in green boxes in **Figure 30**). In addition to the previous analysis, a PCA was also performed for the compounds present in the PCM dataset, where the first two principal

components explained 51.77% of the variance. Clusters contain compounds whose bioactivities on several targets are included in the dataset, thus indicating that compounds are overall structurally similar across the 20 DHFR sequences considered.

Application of Target Prediction for MoA prediction. The performance of the target prediction algorithm was assessed for varying values of n , which represents the top number of non-plasmodial predictions considered for extrapolation. It can be seen that performance varies widely across target classes: for most targets, including all aminopeptidases, calcium-dependent protein kinase 1, protein kinase Pfmrk, glucose-6-phosphate-1-dehydrogenase, dihydroorotate dehydrogenase, dUTP pyrophosphatase and enoyl-acyl-carrier protein reductase, performance is low, with both recall and precision values below 30%. For a small number of targets, however, the performance is much higher, with recall values up to ~60% and precision values up to 100%. Further investigation revealed that the targets for which the prediction algorithm performed well (plasmepsin 1 and 2, histone deacetylase, DHFR and to a lesser extent, falcipain 2) were plasmodial orthologs of non-plasmodial protein targets. This finding is in agreement with previous studies, which have used orthologous proteins to extrapolate the prediction of bioactivities between target classes across species such as *P. falciparum* and *M. tuberculosis*.^{47,48} However, these previous studies have not combined target prediction with PCM for MoA analysis, which is precisely the novelty of the approach presented here.

Target Prediction Performance for Plasmodial DHFR. The predictive performance of the target prediction algorithm was further investigated for the plasmodial target DHFR, where all 145 instances annotated on plasmodial DHFR were used as a test set. The top n predicted non-plasmodial targets were considered (n varied in the 1–12 range), after which these targets were extrapolated to plasmodial targets (section “Domain-based extrapolation to *P. falciparum* targets” in Materials and Methods). For n in the 1–3 range, the recall values are 0%, 2.8% and 14.5%, respectively, whereas for n in the 4–7 range, the recall values are around 35%. The 2-fold CV resulted in a recall value of 79%. These results indicate that despite the fact that the training set did not contain any plasmodial bioactivity data, the model is still able to predict compounds active against plasmodial DHFR with 100% precision, based on bioactivity data for orthologous proteins across other species. The high precision value arises from the structural similarity of plasmodial DHFR

inhibitors and *T. gondii* DHFR inhibitors in the training set (the average MOLPRINT2D pairwise similarity between the *T. gondii* inhibitors and the plasmodial inhibitors was 16%, whereas the average pairwise similarity within the plasmodial dataset and the *T. gondii* dataset was 19% and 18% respectively). These results show the added benefit of incorporating domain-based extrapolation for target prediction purposes.

In addition, we found that varying the domain Tanimoto similarity cutoff between 0.5 and 1 did not alter the performance. Hence, in order to maintain high precision, a stringent domain Tanimoto similarity cutoff of 1 (i.e., requiring a 100% overlap in domain presence and absence between two proteins) was chosen and the top *n* predicted non-plasmodial targets considered was set to 4 for further analysis. Further investigation of the extrapolation from non-plasmodial targets to plasmodial targets revealed that only one protein class (*T. gondii* DHFR) was responsible for the extrapolation of predicted activities to plasmodial DHFR. As described earlier, there are clusters of *T. gondii* DHFR inhibitors that do not contain any plasmodial DHFR inhibitors (scaffolds identified in these clusters are shown in red boxes – **Figure 30** and clusters of plasmodial inhibitors that occupy space not covered by *T. gondii* inhibitors (5 scaffolds identified shown in green boxes in **Figure 30**)). Hence, for these clusters there is no overlap in scaffolds between both datasets. These observations explain the low recall of the model at this stage: plasmodial DHFR inhibitors located outside the space covered by *T. gondii* DHFR inhibitors are not retrieved by the model, thereby increasing the number of false negatives, whereas the plasmodial DHFR inhibitors that are present in the chemical space shared by inhibitors from both species are predicted with very high precision.

Adding plasmodial DHFR data to the training set drastically increased performance, more than doubling recall values to 79%, whereas precision values remained 100% (**Figure 31** – 2-fold CV).

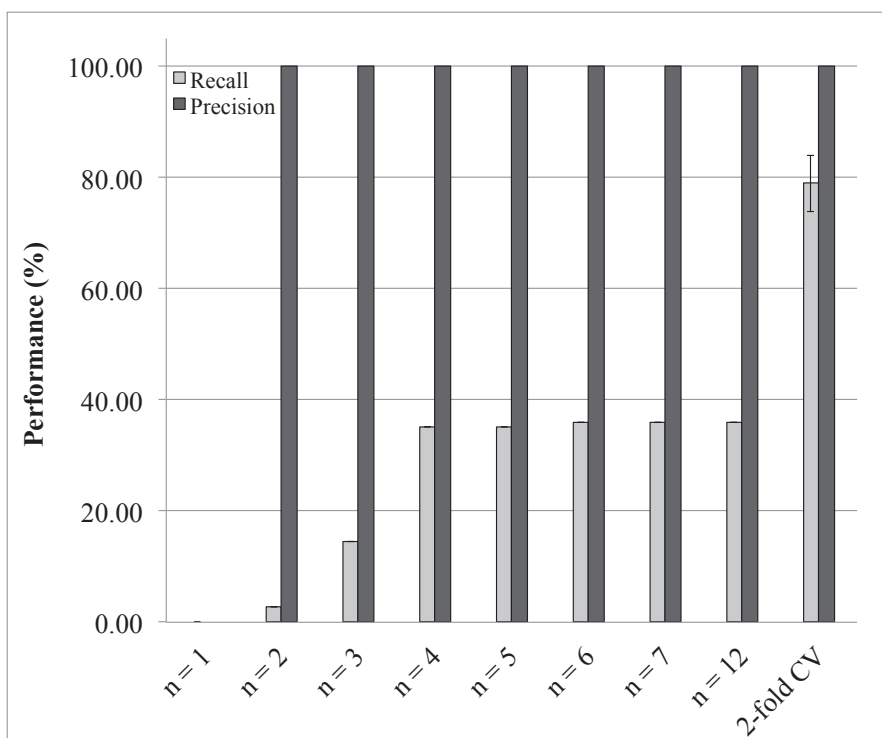


Figure 31. Performance of the DHFR target prediction model compared across a number of parameters. 145 data points annotated against plasmodial DHFR were used as a test set to assess the performance of the target prediction model. The top n predicted non-plasmodial targets were considered (n was varied for values between 1 and 12), after which these targets were extrapolated to plasmodial targets. When n increases, recall values rise up to 36% (with recall values of ~35% for $n = 3$ and $n = 4$). On the other hand, precision values are 100% for $n \geq 2$. The high precision values are likely to be explained by the fact that plasmodial DHFR inhibitors and *T. gondii* DHFR inhibitors occupy the same chemical space. In addition to varying the parameter n , we performed a 2-fold cross validation (averaged over 20 randomizations), which resulted in a drastic improvement as a recall value of 79% was achieved (with a standard deviation of 10.1%, which is shown as an error bar). These results show that domain-based extrapolations have added value to the prediction algorithm (correct predictions are made even when bioactivity data on plasmodial DHFR is not present in the training set) and that including plasmodial DHFR bioactivity data in the training set can drastically improve recall values.

Hence, this observation arises from the fact that the chemical space of the plasmodial DHFR inhibitors adds additional information corresponding to five new scaffolds (as highlighted in green boxes in **Figure 30**) to the model. However, despite the very high precision value achieved (100%), there is a drawback: given the great increase in recall value when novel scaffolds are

added to the dataset, the model is only able to correctly predict bioactivities for compounds with scaffolds that are already present in the training data. Hence, a diverse set of molecules is required in the training set in order to optimize recall values of the model. Given the benefit of both domain-based extrapolation and using plasmodial DHFR bioactivity data for model training, all plasmodial DHFR data were included in the training set for further MoA prediction of the GSK TCAMS phenotypic dataset in order to optimize recall values.

PCM Model Validation. The four algorithms used in this study (GBM, GP, RF and SVM) displayed similar performance on this dataset as the ranges of $RMSE_{test}$ and R^2_{test} differences are 0.04 pIC₅₀ and 0.02 units, respectively. The GBM model exhibited the highest predictive ability with R^2_{test} and $RMSE_{test}$ values of 0.79 and 0.59 pIC₅₀ units respectively. Both internal and external validation metrics are given in **Table 6**.

Table 6. PCM, Family QSAR and Family QSAM performance on the PCM dataset. Abbreviations: QSAM Quantitative Structure-Activity Modeling, QSAR Quantitative Structure-Activity Relationship, GBM Gradient Boosting Machine, GP Gaussian Process, RF Random Forest, SVM Support Vector Machine. PCM, with R^2_{test} and $RMSE_{test}$ values of 0.79 and 0.59 pIC₅₀ units, outperforms both Family QSAR, with R^2_{test} and $RMSE_{test}$ values of 0.63 and 0.78 pIC₅₀ units, respectively, and Family QSAM, with R^2_{test} and $RMSE_{test}$ values of 0.09 and 1.22 pIC₅₀ units, respectively.

	R^2_{cv}	$RMSE_{cv}$	R^2_{ext}	$RMSE_{ext}$
GBM PCM	0.75	0.64	0.79	0.59
GP PCM	0.75	0.65	0.76	0.63
RF PCM	0.74	0.66	0.77	0.62
SVM PCM	0.76	0.63	0.77	0.62
Family QSAM	0.07	1.24	0.09	1.22
Family QSAR	0.61	0.80	0.63	0.78
Inductive Transfer	0.72	0.68	0.76	0.63

To ensure that the model's predictive ability was not the consequence of spurious correlations in the data, we trained ten GBM models with an increasingly higher percentage of the pIC₅₀ values randomized. The performance of the ten models was assessed by examining the $RMSE_{test}$ and R^2_{test} values as a function of the level of randomization of the bioactivity values. The intercept was zero or negative when ~40% of the response variable was randomized. Therefore, the relationship established by the PCM

models between the descriptor space and the bioactivity values is not a consequence of chance correlations.⁴⁹

PCM Outperforms One-space Models and IT on This Dataset. The Family QSAM model, trained on target descriptors only, displayed poor predictive ability with $RMSE_{test}$ and $R^2_{O\ test}$ values of 1.22 pIC₅₀ units and 0.09, respectively (Table 6). By contrast, the Family QSAR model, trained on compound descriptors only, displayed satisfactory values for the statistical metrics according to our validation criteria, as the model exhibited $RMSE_{test}$ and $R^2_{O\ test}$ values of 0.78 pIC₅₀ units and 0.63, respectively (Table 6). Hence, compound descriptors explain a large proportion of the variance, which may stem from the high correlation of the bioactivities of identical compounds against orthologs.

Furthermore, better performance is obtained for the GBM PCM model trained on amino acid descriptors and compound fingerprints, than for the GBM model trained on target identity fingerprints and compound fingerprints, with $RMSE_{test}$ values of 0.59 vs. 0.63 pIC₅₀ units, respectively. This indicates that our selection of amino acid descriptors captured the binding site information of the different orthologs and thus allows explicit learning on this dataset (Table 6). Overall, these data suggest that the explicit inclusion of target information improves bioactivity prediction.

Several High-affinity DHFR Inhibitors Are Identified by Both Target Prediction and PCM. The targets for which the target prediction model had a class-specific F-measure higher than 40% were selected, leading to a shortlist of five proteins, namely: plasmepsin 1 and 2, histone deacetylase, DHFR and falcipain 2. Overall, a total of 1291 plasmodial predictions were made for 1017 compounds. DHFR is the most commonly predicted target, which represents 534 (41%) of the total predictions, followed by plasmepsin 1 (280 predictions – 22%) and plasmepsin 2 (273 predictions – 21 histone deacetylase (184 predictions – 14%) and falcipain 2 (20 predictions – 2%). Plasmodial DHFR has previously been proposed as a candidate target against resistant plasmodial strains.⁵⁰ In addition, the plasmepsin (1 and 2) and falcipain targets have previously been proposed as potential targets for anti-malarial therapy,⁵¹ due to their involvement in the hemoglobin catabolism that occurs during the erythrocytic stage of the malarial parasite life cycle (plasmepsin proteins and falcipain proteins), and to their involvement in erythrocyte invasion and erythrocyte rupture (falcipain proteins).⁵² Finally, plasmodial histone deacetylase has been proposed as a promising target for anti-malarial

therapy due to its key role in regulating gene transcription, and it has been shown that histone deacetylase inhibitors are potent inhibitors of the growth of *P. falciparum*.⁵³ Hence, there is sufficient evidence for all five predicted proteins for being a potential target.

In total, 534 compounds of the GSK TCAMS dataset were predicted to interact with DHFR, representing 3.95% of the total number of compounds in this dataset. Out of these 534 compounds, the predicted pIC₅₀ values using PCM was 7 or greater for 25 compounds, between 6 and 7 for 92 compounds, and between 5 and 6 for 420. None of the 534 compounds was predicted to be inactive on DHFR (**Figure 32**).

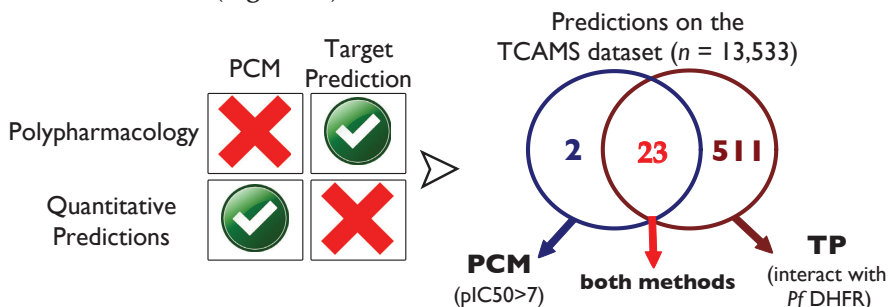


Figure 32. Complementarity between *in silico* target prediction and PCM. The target prediction algorithm predicted 534 compounds of the GSK TCAMS dataset to interact with DHFR, representing 3.95% of the total number of compounds in this dataset. Out of these 534 compounds, the PCM model predicted 23 compounds to have a pIC₅₀ value of 7 or greater. Therefore, the combination of both methods permits the assessment of compound polypharmacology and provides quantitative bioactivity predictions.

Given that many of the compounds in ChEMBL are active in the low micromolar range, it is thus not surprising to obtain most of the predictions in this range.⁵⁴

Interestingly, 23 of the 25 compounds with a predicted pIC₅₀ value higher than 7 were already predicted to interact with DHFR by the target prediction algorithm (**Figure 32**) at the exclusion of any other target. The analysis of chemical scaffolds in the 25 compounds shows that only 2 scaffolds were identified, as 22 out of the 25 compounds (**Figure 33** – excluding compounds 137850, 123550 and 125380), share a common scaffold, namely: a 5-methylpyrido[2,3-d]pyrimidine-2,4-diamine ring with an aryl substituent in the 6-position.

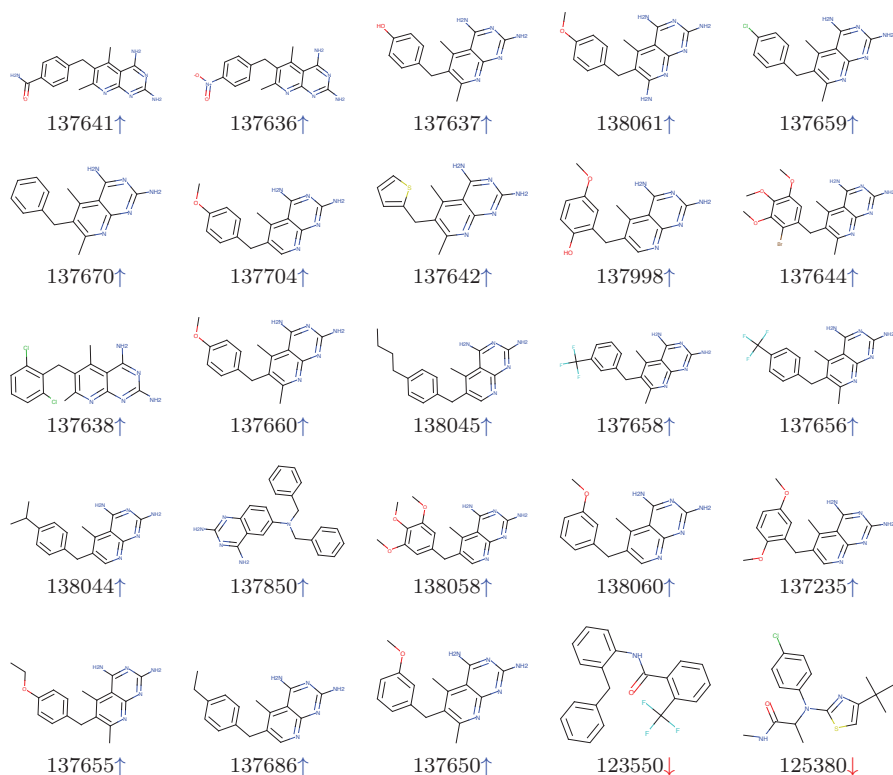


Figure 33. Compounds predicted to interact with DHFR by the target prediction algorithm, and predicted by the PCM model to have a pIC_{50} value higher than 7 pIC_{50} units. Compound IDs correspond to the TCMDC identifier given in the original dataset. The 23 compounds for which the IDs are accompanied by an upward-pointing arrow were identified by the two methods. The two compounds predicted to have a pIC_{50} value higher than 7 by the PCM model, but not predicted to interact with DHFR by the target prediction algorithm, are accompanied by a downward-pointing arrow. The 23 compounds predicted to be high-affinity DHFR inhibitors (upward-pointing arrows) share a common scaffold: a 5-methylpyrido [2,3-d]pyrimidine-2,4-diamine ring with an aryl substituent in the 6-position. Overall, it can be seen that these data indicate a high agreement between the target prediction algorithm and the PCM model to identify high-affinity DHFR inhibitors.

A methyl group or an amine group in the 7-position are also present in some compounds, such as 137637 and 138061, respectively. In all compounds with the common scaffold the aryl substituent is a phenyl ring with different substituents in the 3,4,5-positions, e.g., methoxy, hydroxy and carboxamide, except for compound 137642, which has 2-methyl-thiophene as aryl substituent.

Two additional compounds, 123550 and 125380 (**Figure 33**), predicted by PCM to display pIC_{50} values of 7.11 and 7.07, respectively, represent new

scaffolds. Remarkably, these two scaffolds were neither present in the PCM nor in the target prediction training set. Taken together, our results indicate a high agreement between the target prediction algorithm and the PCM model to identify high-affinity DHFR inhibitors. Using both methods simultaneously, it is possible to give higher priority to the compounds that are identified by both methods.

Conclusions

In this study, the complementarity of *in silico* target predictions and proteochemometric modelling (PCM) was evaluated for the retrospective identification of *P. falciparum* DHFR inhibitors. The target prediction algorithm exhibited respective recall and precision values of 79% and 100% for plasmodial DHFR. The high precision value is explained by the structural similarity of plasmodial and the *T. gondii* DHFR inhibitors, which were part of the training set and were found to be relevant for extrapolation (the average MOLPRINT2D pairwise similarity between the *T. gondii* inhibitors and the plasmodial inhibitors was 16%, whereas the average pairwise similarity within the plasmodial dataset and the *T. gondii* dataset was 19% and 18% respectively).

We showed that high-affinity inhibitors from the GSK TCAMS phenotypic dataset are independently identified by both methods: 534 compounds from the GSK TCAMS dataset were identified as DHFR inhibitors by the target prediction algorithm, whereas the PCM algorithm identified 25 high affinity compounds, 23 of which were already identified by the target prediction algorithm. The combination of both methods permits the assessment of compound polypharmacology and provides insight into the potency/affinity of small molecules.

We presented an approach that can be potentially extended to other human, bacterial or plasmodial targets. The inherent capability of PCM to combine bioactivity data for related targets, even for targets spanning distant phyla, is likely to improve the mining of currently available multi-target bioactivity databases. Similarly, domain-based extrapolation permits *in silico* target predictions to be extended to non-mammalian orthologous proteins for which less bioactivity data is usually available.

References

- (1) Jalencas, X., and Mestres, J. (2013) On the origins of drug polypharmacology. *Med. Chem. Commun.* 4, 80–87.
- (2) Lounkine, E., Keiser, M. J., Whitebread, S., Mikhailov, D., Hamon, J., Jenkins, J. L., Lavan, P., Weber, E., Doak, A. K., Cote, S., Shoichet, B. K., and Urban, L. (2012) Large-scale prediction and testing of drug activity on side-effect targets. *Nature* 486, 361–367.
- (3) Cortés-Ciriano, I., Ain, Q. U., Subramanian, V., Lenselink, E. B., Mendez-Lucio, O., IJzerman, A. P., Wohlfahrt, G., Prusis, P., Malliavin, T., Van Westen, G. J. P., and Bender, A. (2015) Polypharmacology modelling using proteochemometrics (PCM): recent methodological developments, applications to target families, and future prospects. *Med. Chem. Commun.* 6, 24–50.
- (4) Van Westen, G. J., Swier, R. F., Cortés-Ciriano, I., Wegner, J. K., Overington, J. P., IJzerman, A. P., Van Vlijmen, H. W., and Bender, A. (2013) Benchmarking of protein descriptor sets in proteochemometric modeling (part 2): modeling performance of 13 amino acid descriptor sets. *J. Cheminform.* 5, 42–61.
- (5) Poroikov, V., Filimonov, D., Lagunin, A., Glorizova, T., and Zakharov, A. (2007) PASS: identification of probable targets and mechanisms of toxicity. *SAR QSAR Environ. Res.* 18, 101–110.
- (6) Nidhi, Glick, M., Davies, J. W., and Jenkins, J. L. (2006) Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J. Chem. Inf. Model.* 46, 1124–1133.
- (7) Nigsch, F., Bender, A., Jenkins, J. L., and Mitchell, J. B. O. (2008) Ligand-Target Prediction Using Winnow and Naive Bayesian Algorithms and the Implications of Overall Performance Statistics. *J. Chem. Inf. Model.* 48, 2313–2325.
- (8) Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J., and Shoichet, B. K. (2007) Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* 25, 197–206.
- (9) Wale, N., and Karypis, G. (2010) Target Fishing for Chemical Compounds using Target-Ligand Activity data and Ranking based Methods. *J. Chem. Inf. Model.* 49, 2190–2201.
- (10) Koutsoukas, A., Lowe, R., KalantarMotamedi, Y., Mussa, H. Y., Klaffke, W., Mitchell, J. B., Glen, R. C., and Bender, A. (2013) In silico target predictions: defining a benchmarking dataset and comparison of performance of the multiclass Naïve Bayes and Parzen-Rosenblatt Window. *J. Chem. Inf. Model.* 53, 1957–1966.
- (11) Van Westen, G. J. P., Wegner, J. K., IJzerman, A. P., Van Vlijmen, H. W. T., and Bender, A. (2011) Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *Med. Chem. Commun.* 2, 16–30.
- (12) Perlmann, P., and Troye-Blomberg, M. (2000) Malaria blood-stage infection and its conyrol by the immune system. *Folia Biol.* 46, 210–218.
- (13) Olhio, P. (2001) Mode of action and mechanisms of resistance for antimalarial drugs. *Pharmacol. Ther.* 89, 207–219.
- (14) Hecht, D., and Fogel, G. B. (2012) Modeling the evolution of drug resistance in malaria. *J. Comput. Aided Mol. Des.* 26, 1343–1353.
- (15) Moran, M., Guzman, J., and Ropars, A.-L. (2007) The malaria product pipeline: planning for the future. *Georg. Inst. Int. Heal.*
- (16) ChEMBL - Neglected Tropical Disease.
- (17) Gamo, F.-J., Sanz, L. M., Vidal, J., de Cozar, C., Alvarez, E., Lavandera, J.-L., Vanderwall, D. E., Green, D. V. S., Kumar, V., Hasan, S., Brown, J. R., Peishoff, C. E., Cardon, L. R., and Garcia-Bustos, J. F. (2010) Thousands of chemical starting points for antimalarial lead identification. *Nature* 465, 305–310.
- (18) Verma, R., Tiwari, A., Kaur, S., Varshney, G. C., and Raghava, G. P. S. (2008) Identification of proteins secreted by malaria parasite into erythrocyte using SVM and PSSM profiles. *BMC Bioinformatics* 9, 201–211.

- (19) Jamal, S., Periwal, V., and Scaria, V. (2013) Predictive modeling of anti-malarial molecules inhibiting apicoplast formation. *BMC Bioinformatics* 14, 2105–2114.
- (20) Subramaniam, S., Mehrotra, M., and Gupta, D. (2011) Support Vector Machine Based Prediction of *P. falciparum* Proteasome Inhibitors and Development of Focused Library by Molecular Docking. *Comb. Chem. High Throughput Screen.* 14, 898–907.
- (21) *Dotmatics Vortex*, version 2013.03.20719, Dotmatics: The Old Monastery, Windhill, Bishops Stortford, Herts, U.K., 2013.
- (22) Bender, A. (2010) Databases: Compound bioactivities go public. *Nat. Chem. Biol.* 6, 309–309.
- (23) Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., and Overington, J. P. (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucl. Acids Res.* 40, D1100–1107.
- (24) *ChemAxon Standardizer*, version 5.12; ChemAxon, Ltd: Budapest, Hungary, 2012.
- (25) Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T. K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S., de Castro, E., Coghill, P., Corbett, M., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Fraser, M., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., McMenamin, C., Mi, H., Mutowo-Muellenet, P., Mulder, N., Natale, D., Orengo, C., Pesseat, S., Punta, M., Quinn, A. F., Rivoire, C., Sangrador-Vegas, A., Selengut, J. D., Sigrist, C. J. A., Scheremetjew, M., Tate, J., Thimmajananathan, M., Thomas, P. D., Wu, C. H., Yeats, C., and Yong, S.-Y. (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucl. Acids Res.* 40, D306–D312.
- (26) The Uniprot Consortium. (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. *Nucl. Acids Res.* 41, D43–D47.
- (27) Bender, A., Mussa, H. Y., and Glen, R. C. (2004) Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier. *J. Chem. Inf. Model.* 44, 170–178.
- (28) Bender, A., Mussa, H. Y., and Glen, R. C. (2004) Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance. *J. Chem. Inf. Model.* 44, 1708–1718.
- (29) Sastry, M., Lowrie, J. F., Dixon, S. L., and Sherman, W. (2010) Large-Scale Systematic Analysis of 2D Fingerprint Methods and Parameters to Improve Virtual Screening Enrichments. *J. Chem. Inf. Model.* 50, 771–784.
- (30) O’Boyle, N. M., Morley, C., and Hutchison, G. R. (2008) Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chem. Cent. J.* 2, 5–11.
- (31) Crisman, T. J., Parker, C. N., Jenkins, J. L., Scheiber, J., Thoma, M., Kang, Z. Bin, Bender, A., Nettles, J. H., Davies, J. W., and Glick, M. (2007) Understanding false positives in reporter gene assays: in silico chemogenomics approaches to prioritize cell-based HTS data. *J. Chem. Inf. Model.* 47, 1319–1327.
- (32) Bender, A., Mikhailov, D., Glick, M., Scheiber, J., Davies, J. W., Cleaver, S., Marshall, S., Tallarico, J., Harrington, E., Cornella-Taracido, I., and Jenkins, J. (2009) Use of Ligand Based Models for Protein Domains To Predict Novel Molecular Targets and Applications To Triage Affinity Chromatography Data. *J. Proteome Res.* 8, 2575–2585.
- (33) Prathipati, P., Ma, N., Manjunatha, U. H., and Bender, A. (2009) Fishing the target of antitubercular compounds: in silico target deconvolution model development and validation. *J. Proteome Res.* 8, 2788–2798.
- (34) Murrell, D. S., Cortés-Ciriano, I., Van Westen, G. J. P., Malliavin, T., and Bender, A. (2014) Chemistry Aware Model Builder (camb): an R Package for Predictive Bioactivity Modeling.
- (35) Cody, V., Galitsky, N., Luft, J. R., Pangborn, W., Rosowsky, A., and Blakley, R. L. (1997) Comparison of two independent crystal structures of human dihydrofolate reductase ternary

complexes reduced with nicotinamide adenine dinucleotide phosphate and the very tight-binding inhibitor PT523. *Biochemistry* 36, 13897–13903.

(36) Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., and Higgins, D. G. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7, 539–544.

(37) Consonni, V., Ballabio, D., and Todeschini, R. (2010) Evaluation of model predictive ability by external validation techniques. *J. Chemom.* 24, 194–201.

(38) Hawkins, D. M., Basak, S. C., and Mills, D. (2003) Assessing Model Fit by Cross-Validation. *J. Chem. Inform. Comput. Sci.* 43, 579–586.

(39) Tropsha, A., Gramatica, P., and Gombar, V. (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.

(40) Golbraikh, A., and Tropsha, A. (2002) Beware of q^2 ! *J. Mol. Graph. Model.* 20, 269–276.

(41) Brown, J. B., Okuno, Y., Marcou, G., Varnek, A., and Horvath, D. (2014) Computational chemogenomics: Is it more than inductive transfer? *J. Comput. Aided Mol. Des.* 28, 597–618.

(42) Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., and Rätsch, G. (2008) Support vector machines and kernels for computational biology. *PLoS Comput. Biol.* 4, e1000173.

(43) Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* 29, 1189–1232.

(44) Rasmussen, C. E., and Williams, C. K. I. (2006) Gaussian Processes for Machine Learning. The MIT Press.

(45) Breiman, L. (2001) Random Forests. *Mach. Learn.* 45, 5–32.

(46) Kuhn, M. (2008) Building Predictive Models in R Using the caret Package. *J. Stat. Softw.* 28, 1–26.

(47) Spitzmüller, A., and Mestres, J. (2013) Prediction of the *P. falciparum* target space relevant to malaria drug discovery. *PLoS Comput. Biol.* 9, e1003257.

(48) Martínez-Jiménez, F., Papadatos, G., Yang, L., Wallace, I. M., Kumar, V., Pieper, U., Sali, A., Brown, J. R., Overington, J. P., and Marti-Renom, M. A. (2013) Target prediction for an open access set of compounds active against *Mycobacterium tuberculosis*. *PLoS Comput. Biol.* 9, e1003253.

(49) Clark, R. D., and Fox, P. C. (2004) Statistical variation in progressive scrambling. *J. Comput. Aided Mol. Des.* 18, 563–576.

(50) Yuthavong, Y., Tarnchompoo, B., Vilaivan, T., Chitnumsub, P., Kamchonwongpaisan, S., Charman, S. A., McLennan, D. N., White, K. L., Vivas, L., Bongard, E., Thongphanchang, C., Tawechai, S., Vanichtanankul, J., Rattanajak, R., Arwon, U., Fantauzzi, P., Yuvaniyama, J., Charman, W. N., and Matthews, D. (2012) Malarial dihydrofolate reductase as a paradigm for drug development against a resistance-compromised target. *Proc. Natl. Acad. Sci. U.S.A.* 109, 16823–16828.

(51) Ersmark, K., Samuelsson, B., and Hallberg, A. (2006) Plasmepsins as Potential Targets for New Antimalarial Therapy. *Med. Res. Rev.* 26, 626–666.

(52) Marco, M., and Coterón, J. M. (2012) Falcipain inhibition as a promising antimalarial target. *Curr. Top Med. Chem.* 12, 408–444.

(53) Andrews, K. T., Tran, T. N., Wheatley, N. C., and Fairlie, D. P. (2009) Targeting histone deacetylase inhibitors for anti-malarial therapy. *Curr. Top Med. Chem.* 9, 292–308.

(54) Cortés-Ciriano, I., Koutsoukas, A., Abian, O., Glen, R. C., Velazquez-Campoy, A., and Bender, A. (2012) Experimental validation of in silico target predictions on synergistic protein targets. *Med. Chem. Commun.* 4, 278–288.