Improving Screening Efficiency through Iterative Screening Using Docking and Conformal Prediction

Fredrik Svensson[†], *Ulf Norinder*[‡,§], *Andreas Bender*[†*]

[†] Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK

[‡] Swedish Toxicology Sciences Research Center, SE-151 36 Södertälje, Sweden

[§] Department of Computer and Systems Sciences, Stockholm University, Box 7003, SE-164 07 Kista, Sweden

Abstract

High throughput screening, where thousands of molecules rapidly can be assessed for activity against a protein, has been the dominating approach in drug discovery for many years. However, these methods are costly and require much time and effort. In order to suggest an improvement to this situation we, in this study, apply an iterative screening process where an initial set of compounds are selected for screening based on molecular docking. The outcome of the initial screen is then used to classify the remaining compounds through a conformal predictor. The approach was retrospectively validated using 41 targets from the Directory of Useful Decoys, Enhanced (DUD-E), ensuring scaffold diversity among the active compounds. The results show that 57% of the remaining active compounds could be identified while only screening 9.4% of the database. The overall hit rate (7.6%) was also higher than when using docking alone (5.2%). When limiting the search to the top scored compounds from docking, 39.6% of the active compounds could be identified compared to 13.5% when screening the same number of compounds solely based on docking. The use of conformal predictors also gives a clear indication of the number of compounds to screen in the next iteration. These results indicate that iterative screening based on molecular docking and conformal prediction can be an efficient way to find active compounds while screening only a small part of the compound collection.

Introduction

The introduction of high throughput screening (HTS) in drug discovery ushered in an era where the aim of many screening campaigns became to screen as large collection of compounds as possible to maximize the chances of finding compounds with promising activities.[1] Although HTS is a useful tool in many cases,[2] these approaches are resource-intensive, and hence alternative approaches to screening can be taken.[3]

Recently, iterative screening where the results of an initial smaller screen are used to guide further compound selection for the next round of screening, has been shown to increase screening efficiency.[4–7] Maciejewski *et al.* presented a iterative screening strategy based on selecting compound with a low but positive score from a Naïve Bayes model in order to increase the chemical diversity of the screening hits.[4] Also, Paricharak *et al.* demonstrated on Novartis in-house data that iterative screening can aid in the effective screening of bioactive molecules by selecting compounds for screening based on biological and chemical similarities to the hits from the previous iteration.[5] This approach consistently retrieved a high number of active compounds with only a small fraction of the total screening collection screened. Similarly, there have been attempts to define an 'informer set' of a small compound collection which can be screened initially, and which can be used to predict the activity against a wide variety of biological targets in turn, with the same rationale in mind.[7]

The downside of iterative screening is the difficulties associated with having to conduct screening at multiple occasions and using individually picked compounds. To some extent this has been mitigated by the recent improvements in screening technologies and atomization as it is now feasible to conduct screening also in smaller scale while maintaining reasonable efficiency.

On the other hand, the upside is that compared to traditional HTS a much smaller set of compounds might need to be screened in order to identify the same number of active compounds.

In order to use the results from the first screening iteration to determine what compounds to screen next, a number of active molecules needs to be identified. As virtual screening (VS) generally achieves higher hit rate than random screening,[8] VS is a promising approach for selecting the initial screening set.

A conformal predictor is a type of confidence predictor, generating prediction intervals that are guaranteed to be valid in accordance to a user set confidence level.[9] Conformal prediction paired with support vector machines and random forests have previously been shown to be promising approaches to model bioactivities.[10–13] Three features of conformal prediction makes it an attractive choice for modeling compound activities. Firstly, the guaranteed error rate gives the user a way to control the maximum number of false positives that can be accepted. Secondly, conformal prediction handles imbalanced data very well.[12,14,15] Lastly, the number of predicted active compounds will provide insight into how many compounds should be screened in the next iteration.

The evaluation of virtual screening techniques typically relies on retrospective analysis of benchmark datasets. DUD[16] and the improved version DUD-E[17] are among the most frequently used benchmark datasets used for the evaluation of virtual screening. By design DUD and DUD-E are intended to be used primarily for docking, but it has also been used to evaluate ligand-based techniques.[18]

In this work, we now combine iterative screening based on conformal prediction with an initial virtual screening step to identify a first screening set. The workflow is shown schematically in Figure 1. First the compound library is docked to the respective targets. For each target, based on

the docking score, the top one percent of all compounds are selected for a first iteration of screening. Based on these results a classification model is trained and used to predict the activities of the remaining compounds.

The suggested iterative screening approach was evaluated retrospectively on 41 targets from DUD-E. The results indicate that this approach can increase the hit rate in screening while also providing guidance on the number of compounds that should be screened in each iteration.

Methods

Iterative screening

The top one percent of each dataset selected by docking constitutes the first iteration of screening. Based on the compounds true activities conformal predictors were trained to classify the remaining compounds in each dataset. In order to be able to train a classification model on the data, a requirement was added that the active and inactive classes needed to be represented by at least ten compounds each. The compounds receiving a single label prediction as active are then advanced to the next iteration of screening. The cycle of predictions and screening is then repeated until the desired number of active compounds have been identified or the predictions indicate that there is little gain in additional screening.

Ligand preparation

Ligands, both active and decoys, were downloaded from the DUD-E webpage (http://dude.docking.org/) and prepared for docking using LigPrep[19] version 2.6 with the default settings.

Docking

Proteins were prepared for docking using the protein preparation wizard[20] in Maestro with the default settings and grids for docking were generated using the grid generation tool in Glide[21–23]

version 5.9 centering the grid (16 Å) on the co-crystallized ligand. Glide has previously been shown to have excellent performance compared to other docking software when evaluated on the DUD data sets.[24] Docking was performed using Glide in Standard Precision mode at the default settings. From the ligand preparation each compound can be present in the database in different chiral, protonation, and tautomeric states. When ranking compounds only the highest scored form of a compound was considered.

Feature generation

For machine learning the structures were neutralized and salts removed using CORINA[25]. Structure standardization was performed using the IMI eTOX project standardizer[26] in combination with tautomer standardization using the MolVS standardizer[27]. 97 different structural and physiochemical descriptors were calculated using RDKit[28] (complete list in Supporting Information). These descriptors have previously been used for successful modeling of activity data.[15]

Similarity ranking

Similarity ranking was performed by calculating the Euclidian distance in normalized RDKit space (scaled to zero mean and unit variance) to the active reference compounds. Compounds were ranked by their distance to the closest active compound.

Machine learning

A conformal predictor is a type of confidence predictor, i.e. it gives predictions with a guaranteed error rate. In this setting it is achieved by comparing new compounds to compounds of known outcome in a calibration set. The predictor then assigns a label to the new compound for each class where it is similar enough (according to a user set cut-off) to the calibration examples. A conformal predictor therefore outputs a set of predicted labels as opposed to

assigning only a single label to each compound. Thus for a binary classification problem, a compound can be classified as either of the two classes but also to both classes, *both*, or to none of the classes, *empty*. A conformal predictor is said to be valid if the frequency of errors does not exceed the set significance level (defined as 1-confidence level). When evaluating the validity of the predictor a *both* classification is always considered correct and an *empty* always incorrect. For a more in depth example of the conformal prediction algorithm we refer the reader to a recent paper by Norinder *et al.*[13]

The confidence in a prediction is evaluated by calculating the nonconformity score for the new compound and insert that number in a ordered list of nonconformity scores from the calibration set. This has the consequence that the resolution available in the assessment of the confidence is dependent of the number of compounds in the calibration set (as this will determine the number of positions available in the ranked list). In this studies some datasets have very few examples in the calibration set effectively limiting the resolution.

The training data was split into training (70%) and calibration set (30%) in a stratified manner to ensure a proportional distribution of the two classes. We used aggregated models, repeating the sampling 100 times and using the median prediction from the ensemble.[29] Performance was evaluated with respect to the accuracy of the predictions but also with respect to their validity.

In this study we use the term coverage to describe the fraction of compounds with a single label prediction and accuracy to describe the fraction of correct classification for the compounds with single label predictions.

Models were developed using Python, Scikit-learn[30] version 0.17, and the nonconformist package[31] version 1.2.5. The underlying methods for the conformal predictors were binary random forest[32] classification models built using the Scikit-learn RandomForestClassifier with

500 trees and all other options set at default. Class conditional conformal predictions were performed using the ProbEstClassifierNC and IcpClassifier functions in the nonconformist package.

Results and Discussion

Docking to identify the first screening set

41 targets from DUD-E were selected for the study. The first step of our protocol was to dock all compounds against their respective targets. A summary of the datasets and the number of active compounds selected by docking is shown in Table 1.

For each target we chose the top one percent of the compounds based on the docking score. This selected an average of 31.9% active compounds across all the datasets, considerably better performance compared to the expected 1.6% if the compounds had been selected at random. This selection constitutes the first screening set and the activities of these compounds was used in the next steps to train conformal predictors.

Conformal prediction models for next iteration

32 datasets had a sufficient number of active and inactive compounds to allow for the training of a conformal predictor. We chose to use RDKit molecular descriptors as features for the machine learning. One motivation for this was to avoid artificially high enrichments that can be seen when using fingerprint based descriptors on the DUD-E datasets since the decoys are chosen to be diverse from the active compounds in ECFP space.[17] However, the method presented in this paper is not limited to the features applied here but can easily be extended to any input features desired. Once trained, we applied the respective model to classify all remaining compounds in the dataset at the 90% confidence level. The model statistics are shown in Table 2.

The models achieved an average validity of 85.5% for the inactive class and 60.0% for the active class. This is less than the 90% validity expected from the set significance level (see methods section). However, this is not surprising as a small subset of compounds is unlikely to cover the diversity of the full database. We would like to stress that this is not a limitation linked specifically to the conformal prediction algorithm but something that will pose a problem for any machine learning approach applied to this problem.

We next compared the number of active compound selected by the classification model as well as the number of active compounds selected by docking when screening the same number of compounds, the results of which are shown in Table 3. Overall, 57% of the active compounds and 8.8% of the inactive compounds were predicted as active by the models. This equaled a total average of 9.4% of all remaining compounds being predicted as active and screening of these compounds would have resulted in a hit rate of 7.6%. If instead, for each dataset, the corresponding number of compounds had been selected based on the docking scores the hit rate would have been 4.6% with only three datasets having higher hit rate compared to the conformal prediction (see Table 3). Also, selecting the top 10% from each dataset based on docking score (see Supporting Information) would have resulted in a hit rate of 5.2%. Due to the design of the DUD-E datasets where the active compounds are selected from clusters based on their Bemis-Murcko scaffolds, these results also reflects an excellent scaffold diversity.[17,33]

To generate a baseline for comparison we also calculated the number of actives located based on similarity to the initial active compounds in the same descriptor space used to train the predictors (Supporting Information). When allowed to select the same number of compounds as the conformal predictor this approach produced an average hit rate of 6.4%, lower than the average hit rate of the conformal predictor.

A major attraction with conformal prediction in this context is the guidance given with regards to the number of compounds to screen. In the conformal prediction framework the number of compounds to further screen can be derived from the number of single class predictions found at the, by the user, set confidence level. Hence, the models, based on the underlying data, are able to give guidance on how many compound to screen in a subsequent step based on the level of uncertainty that is acceptable to the user.

We also tried applying an approach not requiring extra rounds of experimental screening by training the models based on the docking results alone, considering top scoring compounds as active and bottom scoring compounds as inactive (see Supporting Information, Table S4). When set up in this way, these models did not improve the results compared to docking. It is also important to remember that in the context of conformal prediction training the models on computational data alone will remove the statistical guarantees of the predictor as the calibration set will include examples with the wrong label. However, the design of a purely computational pipeline using similar approaches represents an interesting area for future research.

Using only top scored compounds from docking

In an effort to improve the model validity, we investigated the effects on prediction validity and model outcome when using the derived classification models to predict the remaining compounds from the top ten percent ranked by docking. The rationale behind this is that these compounds should be more similar to the top one percent of compounds used to train the classifier and the models should therefore have increased validities. The downside is that active compounds that are not in the top ten percent based on docking cannot be identified.

The model statistics from the predictions on the top ten percent of the database are shown in Table 4. Compared to predicting the full database the validities are higher, achieving an average validity of 89.6% and 70.9% for the inactive and active classes, respectively.

The number of active compounds selected by the predictive models as well as the corresponding results from docking and similarity search are shown in Table 5. Using this approach, screening all the compounds predicted to be active, would achieve a hit rate of 39.6% while selecting the same number of compounds based on docking would have resulted in a hit rate of 13.5%. The high hit rate can be attributed to the high validity of the predictions for the inactive class drastically limiting the number of false positives. As long as the predictions are valid, at the 90% confidence level, at most 10% of the inactive compounds can be wrongly classified as active.

One of the advantages with the approach presented in this paper is the inherent flexibility of the methods. The conformal prediction framework allows any machine learning algorithm to be used and the initial VS can be conducted using a VS method of choice. Thus, current methods already in place can easily be adapted to be applied within the presented framework.

Conclusions

In this study we present an iterative screening approach, which is based on initial compound ranking by molecular docking and subsequent compound selection by conformal prediction. By using docking to select the first compounds for screening a number of active molecules can be identified while screening only a small fraction of the database, thus allowing for the training of conformal predictors in order to select the compounds for the next iteration of screening. Using this approach, high hit rates can be achieved while screening only parts of the total compound collection. In this study 57% of the active compounds could be located by screening 9.4% of the

database. The average hit rate from the conformal predictors (7.6%) was also higher than using docking alone (5.2%). When classifying only compounds from the top ten percent based on their docking score, 39.6% of the available active compounds where selected while selecting the same number of compounds based on docking resulted in a hit rate of 13.5%. The conformal predictors also provide guidance on the number of compounds to screen in the next iteration based on a user defined confidence level.

Supporting Information

List of RDKit descriptors used for machine learning. Number of actives found when screening an additional 1%, and 9% of the database based on the docking score. Number of actives found when using similarity in RDKit descriptor space.

AUTHOR INFORMATION

Corresponding Author

* ab454@cam.ac.uk

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Funding

FS acknowledges the Swedish Pharmaceutical Society for financial support.

The research at Swetox (UN) was supported by Stockholm County Council, Knut & Alice Wallenberg Foundation, and Swedish Research Council FORMAS.

Notes

The authors declare no competing financial interest.

Abbreviations

DUD-E, Directory of Useful Decoys, Enhanced

References

(1) Macarron, R. Critical Review of the Role of HTS in Drug Discovery. *Drug Discov. Today* **2006**, *11* (7–8), 277–279.

(2) Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. a; Garyantes, T.; Green, D. V. S.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, S. Impact of High-Throughput Screening in Biomedical Research. *Nat. Rev. Drug Discov.* **2011**, *10* (3), 188–195.

(3) Bajorath, J. Integration of Virtual and High-Throughput Screening. *Nat. Rev. Drug Discov.* **2002**, *1* (11), 882–894.

(4) Maciejewski, M.; Wassermann, A. M.; Glick, M.; Lounkine, E. Experimental Design Strategy: Weak Reinforcement Leads to Increased Hit Rates and Enhanced Chemical Diversity. *J. Chem. Inf. Model.* **2015**, *55* (5), 956–962.

(5) Paricharak, S.; IJzerman, A. P.; Bender, A.; Nigsch, F. Analysis of Iterative Screening with Stepwise Compound Selection Based on Novartis In-House HTS Data. *ACS Chem. Biol.* **2016**, *11* (5), 1255–1264.

(6) Reker, D.; Schneider, G. Active-Learning Strategies in Computer-Assisted Drug Discovery. *Drug Discov. Today* **2015**, *20* (4), 458–465.

(7) Paricharak, S.; Méndez-Lucio, O.; Ravindranath, A. C.; Bender, A.; IJzerman, A. P.; van Westen, G. J. P. Data-Driven Approaches Used for Compound Library Design, Hit Trage and Bioactivity Modeling in High-Throughput Screening.

(8) Ripphausen, P.; Nisius, B.; Peltason, L.; Bajorath, J. Quo Vadis, Virtual Screening? A Comprehensive Survey of Prospective Applications. *J. Med. Chem.* **2010**, *53* (24), 8461–8467.

(9) Vovk, V.; Gammerman, A.; Shafer, G. *Algorithmic Learning in a Random World*; Springer: New York, 2005.

(10) Eklund, M.; Norinder, U.; Boyer, S.; Carlsson, L. Application of Conformal Prediction in QSAR. In *IFIP Advances in Information and Communication Technology*; 2012; Vol. 382 AICT, pp 166–175.

(11) Eklund, M.; Norinder, U.; Boyer, S.; Carlsson, L. The Application of Conformal Prediction to the Drug Discovery Process. *Ann. Math. Artif. Intell.* **2013**, *74* (1), 117–132.

(12) Svensson, F.; Norinder, U.; Bender, A. Modelling Compound Cytotoxicity Using Conformal Prediction and PubChem HTS Data. *Toxicol. Res.* **2017**, *6*, 73–80.

(13) Norinder, U.; Carlsson, L.; Boyer, S.; Eklund, M. Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination. *J. Chem. Inf. Model.* **2014**, *54* (6), 1596–1603.

(14) Löfström, T.; Boström, H.; Linusson, H.; Johansson, U. Bias Reduction through Conditional Conformal Prediction. *Intell. Data Anal.* **2015**, *19*, 1355–1375.

(15) Norinder, U.; Boyer, S. Conformal Prediction Classification of a Large Data Set of Environmental Chemicals from ToxCast and Tox21 Estrogen Receptor Assays. *Chem. Res. Toxicol.* **2016**, *29*, 1003–1010.

(16) Huang, N.; Scoichet, B. K.; Irwin, J. J. Benchmarking Sets for Molecular Docking. *J. Med. Chem.* **2006**, *49*, 6789–6801.

(17) Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55* (14), 6582–6594.

(18) Venkatraman, V.; Pérez-Nueno, V. I.; Mavridis, L.; Ritchie, D. W. Comprehensive Comparison of Ligand-Based Virtual Screening Tools Against the DUD Data Set Reveals Limitations of Current 3D Methods. *J. Chem. Inf. Model.* **2010**, *50* (12), 2079–2093.

(19) LigPrep, version 2.6, Schrödinger, LLC, New York, NY 2015.

(20) Madhavi Sastry, G.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and Ligand Preparation: Parameters, Protocols, and Influence on Virtual Screening Enrichments. *J. Comput. Aided. Mol. Des.* **2013**, *27* (3), 221–234.

(21) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelly, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.

(22) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks,

J. L. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **2004**, *47*, 1750–1759.

(23) Glide, version 5.9, Schrödinger, LLC, New York, NY 2015.

(24) Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy. *J. Chem. Inf. Model.* **2009**, *49* (6), 1455–1474.

(25) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-Ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34* (4), 1000–1008.

(26) IMI eTOX project standardizer, https://pypi.python.org/pypi/standardiser

(27) MolVS standardizer, https://pypi.python.org/pypi/MolVS

(28) RDKit: Open-source cheminformatics, http://www.rdkit.org

(29) Carlsson, L.; Eklund, M.; Norinder, U. Aggregated Conformal Prediction. In *Artificial Intelligence Applications and Innovations: AIAI 2014 Workshops: CoPA, MHDW, IIVC, and MT4BD, Rhodes, Greece, September 19-21, 2014. Proceedings*; Iliadis, L., Maglogiannis, I., Papadopoulos, H., Sioutas, S., Makris, C., Eds.; Springer International Publishing: Berlin, Heidelberg, 2014; pp 231–240.

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

Learn. Res. **2011**, *12*, 2825–2830.

(31) nonconformist package, https://github.com/donlnz/nonconformist

(32) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5–32.

(33) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887–2893.

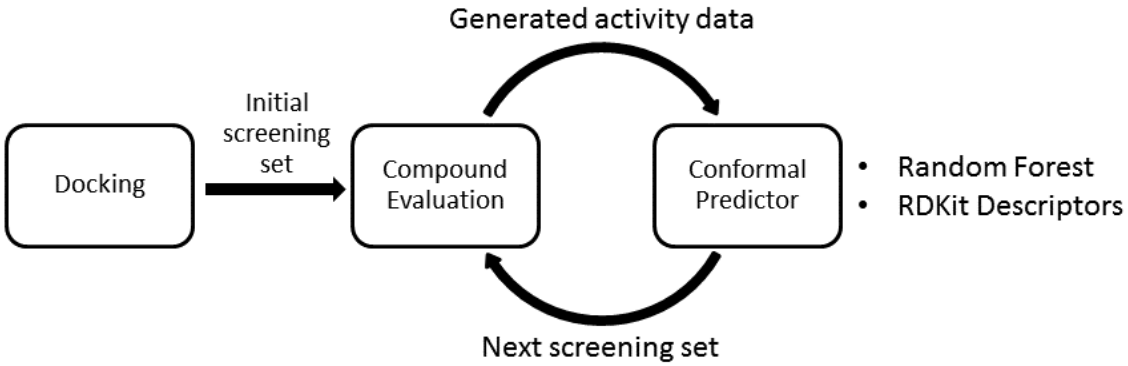

Figure 1. Overview of the workflow presented in this study. An initial docking study is used to select compounds for a first screening set. The activities of the selected compounds are then experimentally evaluated and the results are used to train a conformal predictor. This predictor is then used to classify the remaining compounds in order to select the next set for screening. This iteration, testing and prediction, can be repeated until the desired number of active compounds has been identified.

Table 1. Number of compounds and number of active compounds in the whole datasets as well as in the top 1 % based on docking score. It can be seen that for most datasets docking finds a high number of active compounds in the top 1 %.

Dataset	Inactive compounds	Active compounds	Compounds in top 1 %	Active in top 1 %
ABL1	10,746	182	109	37
ACES	26,233	453	267	64
ADA	5,449	93	55	6
AMPC	2,832	48	29	9
ANDR	14,343	269	146	54
BRAF	9,942	152	101	44
CASP3	10,692	199	109	56
CSF1R	12,143	166	123	44
CXCR4	3,406	40	33	1
DPP4	40,915	533	414	15
DRD3	34,022	480	345	28
DYR	17,170	231	174	48
EGFR	35,020	542	356	106
FNTA	51,430	592	520	54
GCR	14,986	258	152	44
GRIK1	6,546	101	66	27
HDAC8	10,448	170	106	6
HIVRT	18,879	338	192	61
HS90A	4,802	88	49	3
JAK2	6,495	107	66	33
KIF11	6,848	116	70	57
KITH	2,850	57	29	28
KPCB	8,692	135	88	21
MAPK2	6,147	101	62	37

MK01	4,548	79	46	17
MK14	35,810	578	364	85
MP2K1	8,147	121	83	17
NOS1	8,050	100	82	11
PA2GA	5,146	99	52	33
PARP1	30,035	508	305	210
PDE5A	27,520	398	279	61
PGH2	23,135	435	236	133
PNPH	6,950	103	71	32
PPARA	19,356	373	197	61
PPARG	25,256	484	257	98
PUR2	2,694	50	27	27
PYRD	6,446	111	66	44
RENI	6,955	104	71	25
SAHH	3,450	63	35	35
TRY1	25,914	449	264	132
WEE1	6,148	102	63	63

Table 2. Validity, coverage and accuracy of the conformal predictor at 90 % confidence level for all datasets when applied to all the remaining compounds. It can be seen that the validities, especially for the active class, are below the expected value (0.9).

Dataset	Validity inactive	Validity active	Coverage inactive	Coverage active	Accuracy inactive	Accuracy active
ABL1	77.6	86.2	59.1	88.3	62.1	84.4
ACES	87.0	52.2	94.8	95.9	86.3	50.1
ANDR	88.9	72.1	96.7	95.8	92.0	75.2
BRAF	95.1	73.1	98.4	95.4	95.4	73.8
CASP3	93.8	17.5	95.3	92.3	98.5	18.9
CSF1R	79.6	71.3	87.9	82.8	76.8	65.3
DPP4	80.0	96.5	45.4	66.2	55.9	94.8
DRD3	73.7	64.4	88.7	85.6	83.0	75.2
DYR	91.9	79.2	92.1	87.4	99.8	90.6
EGFR	81.2	44.0	81.3	54.8	99.8	80.3
FNTA	93.6	55.9	96.0	86.1	97.5	65.0
GCR	86.7	14.0	87.5	57.9	99.0	24.2
GRIK1	96.0	10.8	96.6	87.8	99.4	12.3
HIVRT	77.8	69.0	54.4	64.6	59.2	52.0
JAK2	98.5	89.2	74.9	58.1	98.0	81.4
KIF11	99.8	28.8	88.7	74.6	99.8	4.5
KPCB	88.5	47.4	88.7	84.2	99.7	56.3
MAPK2	99.2	42.2	99.5	95.3	99.2	39.3
MK01	96.1	46.8	98.0	88.7	96.8	50.9
MK14	82.1	60.9	89.4	89.5	91.9	68.0
MP2K1	78.5	16.3	80.1	75.0	98.0	21.8
NOS1	56.1	96.6	57.0	87.6	22.9	96.2
PA2GA	86.6	47.0	87.0	87.9	99.6	53.4
PARP1	76.4	80.5	90.2	92.6	84.7	87.0

PDE5A	86.5	96.7	59.8	82.5	77.4	96.0
PGH2	86.2	27.2	86.6	75.8	99.5	35.8
PNPH	94.4	98.6	97.5	98.6	96.8	100
PPARA	72.2	79.2	83.3	88.5	86.7	89.5
PPARG	67.9	69.4	79.3	85.0	85.6	81.7
PYRD	97.8	34.3	98.5	98.5	99.3	34.8
RENI	90.8	83.5	97.2	98.7	93.4	84.6
TRY1	74.9	70.3	76.0	76.0	98.5	92.5

Table 3. Number of compounds predicted to be active by the conformal predictors for each dataset, the number of true active compounds from the predictions as well as the number of active compounds that would have been identified when screening the same number of compounds based on docking score. The highest number of active compound for each dataset is indicated in bold. Overall, the conformal predictor locates more active compounds compared to docking.

Dataset	Compounds evaluated	Conformal active	Docking active
ABL1	2,499	**108**	80
ACES	3,570	**187**	159
ANDR	1,260	**155**	66
BRAF	526	**76**	50
CASP3	179	**25**	15
CSF1R	2,523	66	**102**
DPP4	8,439	**325**	232
DRD3	5,367	**291**	219
DYR	175	**145**	21
EGFR	254	**192**	17
FNTA	1,518	**301**	68
GCR	155	**30**	5
GRIK1	47	**8**	4
HIVRT	4,257	93	**168**
JAK2	134	**35**	26
KIF11	16	2	**7**
KPCB	76	**54**	4
MAPK2	70	**24**	18
MK01	169	**28**	14
MK14	2,864	**300**	70

MP2K1	149	**17**	8
NOS1	3,580	**75**	60
PA2GA	51	**31**	8
PARP1	4,360	**240**	230
PDE5A	3,949	**267**	176
PGH2	175	**82**	32
PNPH	286	**70**	39
PPARA	2,373	**247**	174
PPARG	3,124	**268**	232
PYRD	69	**23**	9
RENI	506	**66**	34
TRY1	516	**223**	80

Table 4. Validity, coverage and accuracy of the conformal predictor at 90 % confidence level for all datasets when applied to the top ten percent of compounds ranked by docking. This approach achieves validities closer to the expected (0.9) compared to when predicting the full dataset (Table 2).

Dataset	Validity inactive	Validity active	Coverage inactive	Coverage active	Accuracy inactive	Accuracy active
ABL1	91.2	75.7	79.4	81.1	89.0	70.0
ACES	91.9	72.7	98.0	95.3	91.7	71.3
ANDR	91.3	70.1	97.1	91.0	94.0	77.0
BRAF	90.8	79.4	98.1	100	91.0	79.4
CASP3	92.7	62.5	94.7	87.5	97.9	71.4
CSF1R	84.4	77.3	88.7	78.8	82.4	71.2
DPP4	87.0	96.5	44.6	69.9	70.7	94.9
DRD3	87.3	80.6	94.9	91.7	92.0	87.9
DYR	91.1	90.8	91.3	92.9	99.8	97.8
EGFR	85.8	50.4	86.2	63.0	99.6	80.0
FNTA	93.2	58.8	95.5	82.9	97.6	70.9
GCR	91.1	23.1	91.5	50.0	99.6	46.2
GRIK1	97.7	17.5	98.0	92.5	99.6	18.9
HIVRT	86.3	83.9	57.9	66.1	76.3	75.7
JAK2	96.5	85.4	77.8	75.0	95.5	80.6
KIF11	99.5	40.5	90.0	62.2	99.4	4.3
KPCB	91.3	37.9	91.3	79.3	100	47.8
MAPK2	98.6	47.1	99.0	96.1	98.6	44.9
MK01	92.4	78.3	94.4	91.3	92.2	76.2
MK14	86.3	73.3	90.8	88.0	94.9	83.3
MP2K1	83.9	46.4	84.6	78.6	99.2	59.1
NOS1	71.4	95.7	46.8	93.6	38.9	95.5

PA2GA	92.4	91.7	94.0	100	98.3	91.7
PARP1	82.1	88.0	95.5	95.9	86.1	91.8
PDE5A	94.1	94.9	69.8	87.0	91.6	94.2
PGH2	88.8	51.8	88.9	84.7	99.8	61.1
PNPH	94.5	100	97.4	100	97.0	100
PPARA	82.7	84.1	89.8	90.7	92.1	92.7
PPARG	82.4	80.1	89.3	91.9	92.3	87.1
PYRD	96.6	62.1	97.7	93.1	98.9	66.7
RENI	91.3	91.9	98.5	100	92.7	91.9
TRY1	81.2	80.0	82.2	83.3	98.8	96.0

Table 5. The number of compounds selected for screening by the predictive models when applied to the top ten percent of the database selected by docking, as well as the number of active compounds identified by the model and docking. The highest number of active compound for each dataset is indicated in bold. The conformal predictors perform better or equally good as docking across all data sets.

Dataset	Compounds evaluated	Conformal active	Docking active
ABL1	104	**21**	5
ACES	271	**87**	27
ANDR	120	**47**	17
BRAF	125	**50**	19
CASP3	39	**20**	6
CSF1R	200	**37**	27
DPP4	547	**75**	25
DRD3	340	**116**	15
DYR	92	**89**	11
EGFR	71	**60**	7
FNTA	219	**117**	13
GCR	11	**6**	0
GRIK1	9	**7**	1
HIVRT	278	**56**	39
JAK2	48	**29**	9
KIF11	4	1	1
KPCB	11	**11**	0
MAPK2	29	**22**	10
MK01	45	**16**	7
MK14	202	**55**	18
MP2K1	18	**13**	1

NOS1	238	**42**	30
PA2GA	29	**22**	6
PARP1	528	**191**	132
PDE5A	252	**113**	15
PGH2	48	**44**	14
PNPH	64	**47**	16
PPARA	243	**127**	48
PPARG	314	**169**	66
PYRD	24	**18**	6
RENI	77	**34**	13
TRY1	166	**144**	38

For Table of Content Use Only

Improving Screening Efficiency through Iterative Screening Using Docking and Conformal Prediction

Fredrik Svensson, Ulf Norinder, Andreas Bender