Maximizing Gain in HTS Screening Using Conformal Prediction

Ulf Norinder ULF.NORINDER@SWETOX.SE
Swetox, Karolinska Institutet, Unit of Toxicology Sciences, Forskargatan 20, SE-151 36 Södertälje, Sweden
Department of Computer and Systems Sciences, Stockholm University, Forum 100, SE-164 40 Kista, Sweden

Fredrik Svensson FS447@CAM.AC.UK
IOTA Pharmaceuticals, St Johns Innovation Centre, Cowley Road, Cambridge CB4 0WS, UK
Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK

Avid M. Afzal MAA76@CAM.AC.UK
Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK

Andreas Bender AB454@CAM.AC.UK
Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK

Editors: Alex Gammerman, Vladimir Vovk, Zhiyuan Luo, and Harris Papadopoulos

Abstract

Today, screening of large compound collections in high throughput screening campaigns form the backbone of early drug discovery. Although widely applied, this approach is resource and potentially labour intensive. Therefore, improved computational approaches to streamline screening is in high demand. In this study we introduce conformal prediction paired with a gain-cost function to make predictions in order to maximise the gain of screening campaigns on new screening sets. Our results indicate that using 20 % of the screening library as an initial screening set and using the data obtained together with a gain-cost function, the significance level of the predictor that maximise the gain can be identified. Importantly, the parameters for the predictor derived from the initial screening set was highly predictive of the maximal gain also on the remaining data. Using this approach, the gain of a screening campaign can be improved considerably.

Keywords: Drug discovery, Conformal Prediction, Virtual Screening, Gain-cost function.

1. Introduction

In order to maximize the chances of finding compounds with promising activities the aim of many high-throughput screening (HTS) campaigns became to screen as large compound collections as possible (Macarron, 2006). Although HTS represents a useful screening tool in many cases (Macarron et al., 2011), these approaches can be both resource- and cost-intensive and hence alternative approaches to screening have been explored (Bajorath, 2002).

Quantitative structure-activity relationship (QSAR) methods have long been used to develop in silico models for aiding HTS and virtual screening (Tropsha, 2009). Lately, conformal prediction (CP) has emerged as a promising alternative to traditional QSAR approaches for developing predictive models for use in drug discovery (Eklund et al., 2013; Norinder et al., 2014; Svensson et al., 2017a). CP has been shown to be particularly powerful for the modelling of highly imbalanced datasets, such as those typically encountered in screening campaigns (i.e. where the minority class is the class of interest) (Norinder and Boyer, 2016; Svensson et al., 2017b).

A conformal predictor is a type of confidence predictor that, for binary classification problems (such as active or inactive in a particular screen) provides prediction intervals (a set of labels) that are guaranteed to be valid in accordance with a user set significance level (Vovk et al., 2005). A Mondrian conformal predictor is guaranteed to be valid in respect to each of the predicted classes. This is especially valuable when predicting the outcome of HTS as it is important to accurately detect many active (minority class) compounds since these represent potentially important starting points for new drug discovery projects. At the same time, it is important to also limit the number of false positives, i.e. inactive compounds labelled as active, since the subsequent screening of these compounds will consume resources in terms of logistics, labour, screening consumables, and other screening related costs.

However, the efficiency, i.e. the percentage of single label (class) predictions, may vary considerably depending upon the underlying precision of the derived model and on the set significance level. This means that although all models should be valid, they can differ in their efficiency. When the models are used to prioritize compounds for tasks such as screening, they may therefore produce different outcomes if, for example, only compounds with a single label prediction as active are advanced for testing.

Virtual screening is typically evaluated retrospectively on datasets of known outcome. This is done using some form of performance metric based on how well the method enriches active compounds at the top of the hit list. The downside with this approach is that, although the average performance of a method over many datasets can be evaluated, there is a significant dataset dependence and it is difficult to assess the performance of a method on a particular target a priori. One way to tackle this problem is to screen part of the new dataset and use the information obtained to evaluate how to best proceed (Svensson et al., 2017a; Paricharak et al., 2016a,b).

Rather than evaluating screening predictions based on the enrichment of active compounds, a method can be evaluated based on the cost of the screening and the gain from the expected hits. This has the potential of providing the answer to how many compounds should be prosecuted in subsequent screening campaigns in order to maximise gain. This is a very different criteria from the above mentioned enrichment of actives. Although enrichment in a certain portion of the database can help differentiate between two methods that are being evaluated, it does not give an insight to what proportion of the collection to screen. For example, a very small selection could have a very high enrichment of active compounds but the gain can be much higher in a larger selection of compounds where the enrichment is lower since more active compounds in absolute numbers can be identified.

In this study we introduce the concept of evaluating virtual screening in regards to the cost and gain of screening. We use four different PubChem datasets of various imbalances and sizes, and show how conformal predictors coupled with a gain-cost function can be used to find the most appropriate significance settings in order to maximise the gain from screening.

2. Method

2.1. Data

Four datasets containing more than 40,000 compounds and corresponding assay outcome data were downloaded from the PubChem BioAssay database (Table 1).

Table 1: The number of compounds and actives for the datasets (PubChem AIDs) used in this study.

Dataset	#active cmpds	#total cmpds	%active	target
868	3,545	194,381	1.82	RAM network signaling
1460	1,189	47,025	2.53	tau fibrillization
2314	36,955	295,303	12.51	Stabilization of luciferase activity
2551	16,632	269,830	6.16	ROR gamma activity

The dataset structures were neutralized and salts removed using corina (Sadowski et al., 1994) followed by structure standardization using the IMI eTOX project standardizer (https://pypi.python.org/pypi/standardiser) in combination with the MolVS standardizer (https://pypi.python.org/pypi/MolVS) for tautomer standardization. Compound activities were based on the PubChem outcome annotation and records with missing or conflicting annotations were removed. Each dataset was randomly split into a training set (20 %) and an external test set (80 %).

2.2. Feature generation

97 different physicochemical/structural feature descriptors (physicochemical) as well as Morgan fingerprint descriptors (fingerprints) were calculated using RDKit (RDKit: Open-source cheminformatics; http://www.rdkit.org). The latter were subsequently hashed onto a binary feature vector of length 4,096.

2.3. Conformal Prediction

The performance of a conformal predictor is often measured by its validity. A conformal predictor is said to be valid if the frequency of errors does not exceed the user defined set significance level. New instances (compounds) are assigned a set of class labels through comparison to a calibration set with known labels and if the prediction outcome is similar enough (higher than the set cut-off) the new instance is assigned that class label. This process is performed for each label (class) in the data. Thus, there are four possible outcomes for a binary classification problem, a new instance can be labelled with either of the two classes, be assigned both labels (both classification) or neither one (empty classification). For a more detailed example of how conformal prediction is carried out we refer the reader to Norinder et al. (2014). A conformal prediction is considered correct if it includes the correct class label. Consequently, both predictions are always correct and empty predictions never are (i.e. always erroneous) and a trade off in conformal prediction is that between validity and efficiency of the model.

CP models were developed using Python, Scikit-learn (Pedregosa et al., 1994) version 0.17, and the nonconformist package (`https://github.com/donlnz/nonconformist`) version 1.2.5. The binary classification models were generated using the Scikit-learn RandomForestClassifier using 500 trees and all other options set at default. The ProbEstClassifierNC and IcpClassifier functions in the nonconformist package, with options for class conditional conformal predictions enabled, were used for the conformal predictions.

100 CP models were built for each dataset and the aggregated conformal prediction method described by Carlsson et al. (2014) was used for the final CP prediction outcome. For each of these models, the training set was randomly divided in proper training set and calibration set using 70 % and 30 % of the training data, respectively. The median predicted probability for each test compound was then calculated from the 100 models and used for class assignment in accordance with the set significance levels (Figure 1).

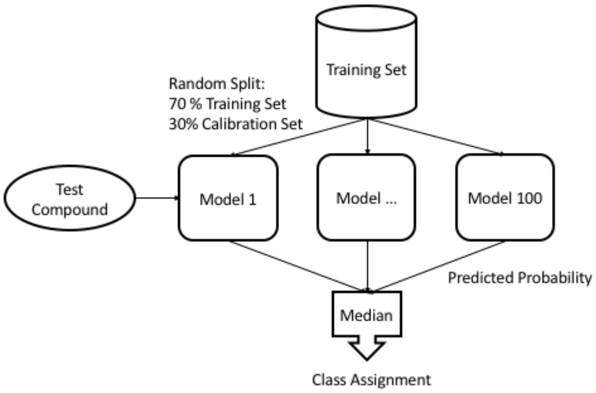

Figure 1: Aggregated conformal prediction scheme for test set prediction.

The predictive performance for each of the training sets was evaluated by a similar procedure where the training data set was randomly divided in new training set (80 %) and an internal test set (20 %). The new training set was then randomly divided into a proper training set (70 %) and a calibration set (30 %) and 100 models constructed. The median predicted probability for each internal test compound was then calculated from the 100 models and used for class assignment in accordance with the set significance level (Figure 2).

2.4. Gain-cost function

A gain-cost function needs to be established in order to investigate the outcome from the derived aggregated CP models both for the internally validated training sets as well as for

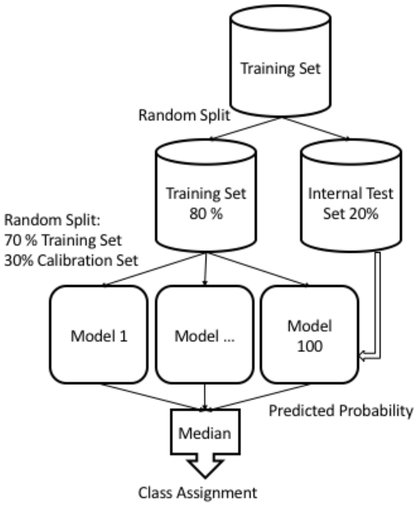

Figure 2: Aggregated conformal prediction scheme for internal training set validation.

the external test sets. With respect to the cost term of such a function, it is relatively easy to identify reasonable factors and levels from previously performed screens. For the gain term, the future value of finding active compounds during screening that may eventually generate some form of income, it is much more difficult to set reasonable values since these may differ considerably depending upon the structure of the identified active compound as well as the biological target in question. In this study we use three different rations between cost and gain by defining a gain of 400 per identified active compound and three levels of screening costs; low 4, medium 8, and high 12. These values were chosen to reflect realistic ratios based on the assumptions that a HTS on the investigated datasets should at least break even, but other values may be set at the discretion of the investigator.

To evaluate the screening gain-cost we defined the following setup:
1. CP significance levels 0.05 - 0.4 were investigated with increments of 0.05.
2. All compounds in the training sets were screened.
3. Only external test set compounds with a single label prediction as active were screened.
4. The potential gain for each screened active compound (gc) was set to 400 (arbitrary unit).
5. A fixed cost (fc) for each screened compound, covering development, personnel, storage etc., was set to 2.
6. A screen dependent cost (sdc) for each screened compound covering assay related costs, e. g. consumables and handling, was set to 4, 8 or 12 (low, medium and high cost screen).

The gain-cost function to be maximized can be written as:

$$gain = \sum_{i=1}^{ntra}(gc) - \sum_{i=1}^{ntr}(fc + sdc) + \sum_{i=1}^{ntesta}(gc) - \sum_{i=1}^{ntest}(fc + sdc) \tag{1}$$

where ntr and ntra are the number of screened training set compounds and the number of active training set compounds, respectively. Ntest and ntesta are the corresponding numbers for the test set compounds.

The overall workflow used in this study is presented in Figure 3.

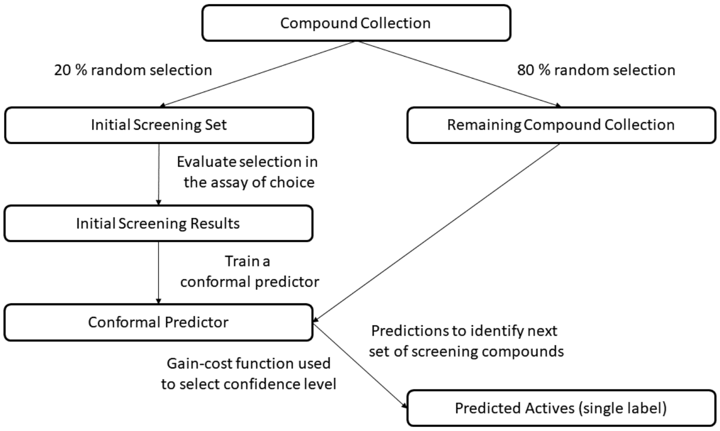
Compound Collection
20 % random selection
80 % random selection
Initial Screening Set
Remaining Compound Collection
Evaluate selection in the assay of choice
Initial Screening Results
Train a conformal predictor
Conformal Predictor
Predictions to identify next set of screening compounds
Gain-cost function used to select confidence level
Predicted Actives (single label)

Figure 3: Schematic of the workflow used for the screening.

3. Results & Discussion

The validities of the aggregated CPs are presented in Table 2. Allowing for small statistical variations, that the class conditional validities are only compromised for the 1460 external test set active minority class when using physicochemical descriptors. Thus, as mentioned in the introduction, there seems to exist a number of models at various significance levels that, from a CP perspective, are valid both for the minority, active, class as well as for the majority, inactive, class for both training and external test sets.

The efficiency of the derived models differs considerably when using physicochemical or fingerprint descriptors, respectively. The former set of descriptors exhibits a similar pattern for all four datasets (Figure 4) while the latter type differs markedly between datasets (Figure 5). Most notably, fingerprint-based models for the two datasets with the fewest percentage of active compounds, 860 and 1460, show very low efficiencies at low significance levels. Dataset 868, with only 1.8 % actives, have low efficiencies for most of the investigated significance levels. This behaviour may be the result from relatively few active compounds in combination with a sparse structure representation (fingerprints) where the calibration set, especially for the actives, does not cover the dataset as a whole particularly well.

From a screening perspective it is of importance to identify which approach produce the most productive predictions in terms of screening efficiency, i.e. gain. For this purpose a gain-cost function was devised to describe the screening outcomes, considering the cost associated with the screening of compounds and the gain from finding hits. Using the gain-cost function described above we wanted to use the information from the initial compound set

Table 2. Validities of the derived models for both active and inactive predictions.

		Testset, signif. level								Trset internal validation, signif.level							
Dataset	**descriptor**	**0.05**	**0.1**	**0.15**	**0.2**	**0.25**	**0.3**	**0.35**	**0.4**	**0.05**	**0.1**	**0.15**	**0.2**	**0.25**	**0.3**	**0.35**	**0.4**
868 active	fingerprints	99.7	98.4	95.3	90.6	85.7	80.1	73.6	67.2	100	99.7	98.2	95.1	89.3	81.0	74.9	67.2
868 inactive	fingerprints	97.1	94.3	90.9	87.1	82.8	77.6	71.9	65.7	97.5	95.1	92.4	88.6	84.8	80.2	74.5	67.2
1460 active	fingerprints	99.9	94.9	90.2	84.5	78.0	72.2	65.7	58.9	100	96.2	86.4	83.3	78.0	75.0	70.5	66.7
1460 inactive	fingerprints	97.0	93.9	90.7	87.0	83.0	78.1	73.2	67.7	97.4	94.4	91.5	86.7	83.1	77.9	72.9	67.7
2314 active	fingerprints	96.4	91.8	87.2	82.1	77.2	72.1	66.8	61.7	97.7	93.5	89.4	84.2	78.8	73.9	68.4	62.8
2314 inactive	fingerprints	96.0	91.6	87.0	82.1	77.0	71.8	66.5	61.1	96.0	91.7	87.1	82.4	77.2	72.3	67.0	61.6
2551 active	fingerprints	94.9	90.7	85.1	78.4	73.6	68.5	64.0	58.8	97.9	92.9	87.5	81.4	76.5	71.3	66.4	61.5
2551 inactive	fingerprints	94.4	89.4	84.6	79.8	75.0	70.0	65.1	60.2	96.0	92.2	88.0	83.9	79.4	74.4	69.5	63.9
868 active	physchem	96.3	91.8	86.9	82.5	78.1	73.8	68.7	63.7	98.8	92.9	85.6	82.8	78.5	74.9	68.1	64.4
868 inactive	physchem	95.5	90.6	85.6	80.5	75.3	70.1	64.9	59.7	95.7	91.0	86.3	81.3	76.1	71.1	66.1	61.7
1460 active	physchem	92.9	85.2	79.6	74.8	69.9	65.3	61.5	58.0	99.2	94.7	92.4	86.4	80.3	75.0	68.9	65.2
1460 inactive	physchem	95.9	91.0	85.7	80.5	75.3	70.0	64.7	59.4	96.2	90.9	86.4	81.6	76.6	70.9	66.2	61.4
2314 active	physchem	95.5	90.5	85.3	80.1	75.4	70.5	65.7	60.7	95.7	91.0	86.0	81.3	76.0	71.1	66.0	61.0
2314 inactive	physchem	95.1	90.2	85.2	80.3	75.2	70.1	64.9	59.8	95.5	90.5	85.4	80.7	75.7	70.7	65.7	60.7
2551 active	physchem	95.8	91.3	86.1	81.2	76.1	71.3	66.6	61.4	96.4	91.5	86.6	81.9	76.9	71.3	65.9	61.2
2551 inactive	physchem	95.2	90.2	85.3	80.3	75.4	70.4	65.2	60.0	95.3	90.7	85.7	80.9	75.8	70.9	66.2	61.0

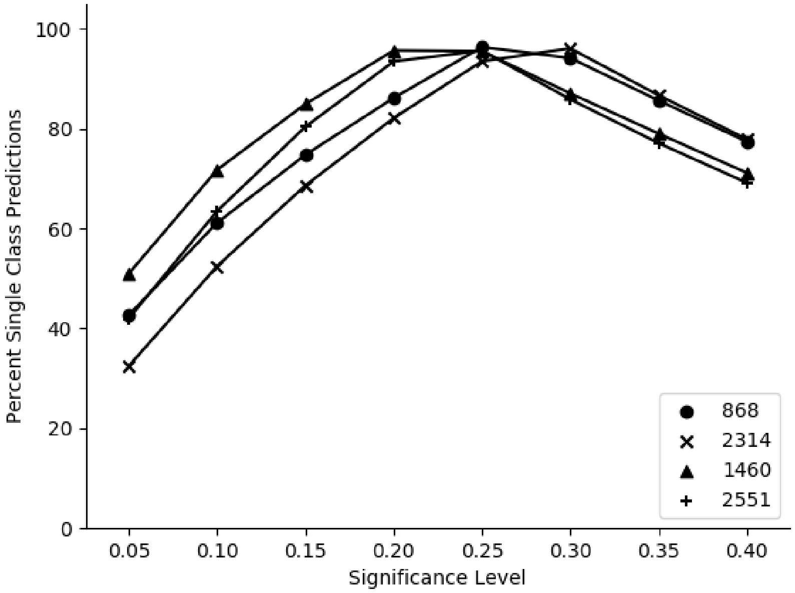

Figure 4: Efficiency of the physicochemical descriptor-based models.

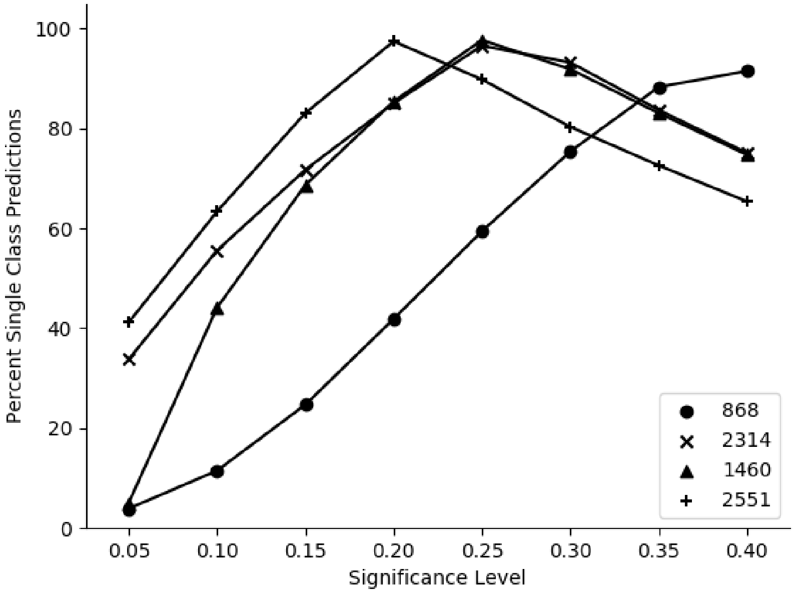

Figure 5: Efficiency of the fingerprint-based models.

Table 3. Maximum gains for the datasets and corresponding significance levels.

Dataset	descriptor	signif. level test-set	signif. level trset	cost/ cmpd (sdc)	% ac-tives	% ac-tives found	% screened cmpds	total gain@ trset max	total gain@ testset max	% trset max loss	gain entire dataset	trset gain	gain en-tire trset
868	fingerprints	0.4		4	1.82	67.81	36	541786		0	251714	51404	13670
868	fingerprints	0.25	0.35	8	1.82	66.18	35.1	256150	262590	2.45	-525810	30010	-64150
868	fingerprints	0.2		12	1.82	46.94	22.06	65378		0	-1303334	17688	-141970
1460	fingerprints	0.25		4	2.53	78.55	26.6	298540		0	193450	35174	24186
1460	fingerprints	0.25		8	2.53	78.55	26.6	248500		0	5350	31690	5110
1460	fingerprints	0.2	0.15	12	2.53	68.71	19.79	196530	201926	2.67	-182750	29210	-13966
2314	fingerprints	0.25	0.3	4	12.51	74.91	34.88	10455236	10657424	1.9	13010182	1050324	1315000 *
2314	fingerprints	0.25	0.3	8	12.51	74.91	34.88	10043260	10227440	1.8	11828970	1015340	1197000 *
2314	fingerprints	0.25	0.3	12	12.51	74.91	34.88	9631284	9797456	1.7	10647758	980356	1079000 *
2551	fingerprints	0.2		4	6.16	79.59	30.46	4801708		0	5033820	493976	509656 *
2551	fingerprints	0.2		8	6.16	79.59	30.46	4472980		0	3954500	472360	401160
2551	fingerprints	0.2		12	6.16	79.59	30.46	4144252		0	2875180	450744	292664
868	physchem	0.25	0.3	4	1.82	73.81	24.98	701882	703962	0.3	251714	65676	13670
868	physchem	0.3	0.4	8	1.82	63.68	18.48	432660	449270	3.7	-525810	45580	-64150
868	physchem	0.15	0.4	12	1.82	63.68	18.48	225564	225994	0.19	-1303334	30212	-141970
1460	physchem	0.25		4	2.53	69.91	12.92	257084		0	193450	34604	24186
1460	physchem	0.2	0.25	8	2.53	69.91	21.47	209980	210090	0.05	5350	29940	5110
1460	physchem	0.2	0.25	12	2.53	69.91	21.47	154612	155886	0.82	-182750	25276	-13966
2314	physchem	0.3		4	12.51	70.51	31.54	10183356		0	13010182	1003746	1315006 *
2314	physchem	0.3		8	12.51	70.51	31.54	9730660		0	11828970	965710	1197010 *
2314	physchem	0.3		12	12.51	70.51	31.54	9277964		0	10647758	927674	1079014 *
2551	physchem	0.25		4	6.16	76.12	23.6	4718418		0	5033820	475072	509656 *
2551	physchem	0.25		8	6.16	76.12	23.6	4380830		0	3954500	447520	401160
2551	physchem	0.25		12	6.16	76.12	23.6	4043242		0	2875180	419968	292664

% actives: percentage actives in data set; % screened cmpds: percentage screened compounds (all training set compounds + optimal number of test set compounds to maximize gain); total gain@trset max: total gain from screening (all training set compounds + optimal number of test set compounds) at significance level indicated to give training set maximum gain; total gain@testset max: total gain from screening (all training set compounds + optimal number of test set compounds) at significance level indicated to give test set maximum gain; % trset max loss: percentage loss in total gain (trset + test set) when using significance level indicated to give training set maximum gain instead of significance level indicated to give test set maximum gain (0 = identical significance levels for training and test set maximum gain); gain entire dataset: gain if the entire test set and training set were screened; trset gain: gain from internal validation of the training set at found significance level optimum; gain entire trset: gain if the entire training set was screened to find the optimal significance level of the CP to predict the remaining 80 % of compounds for screening in order to obtain the maximum gain. The results of such gain-cost function CP usage with the aim of maximizing the gain for each assay are shown in Table 3.

Interesting to note is the excellent correspondence in terms of maximizing the gain, between the optimal significance level identified by the internal validation of the training set and the same setting for the external test set for both fingerprints and models based on physicochemical descriptors (7 out of 12 cases, Table 3). Furthermore, for those cases where the optimal significance level identified during the internal validation of the training set does not correspond to the same setting for the external test set (both fingerprints and physicochemical: 5 out of 12 cases) the decrease in gain (% trset max loss) is minimal with an average and maximum decrease in gain for the fingerprint models of only 2.1 % and 2.7 %, respectively, and 1.0 % and 3.7 %, respectively, for the physicochemical descriptor-based models.

Also important to note is that for some datasets, e.g. 2314 and 2551, the maximum gain is achieved when screening the entire library. Gratifyingly, this was also indicated by the internal validation procedure of the respective training set (indicated by asterisk in Table 3). This means that the models developed for the internal validation are predictive also of this behaviour.

Overall, our results indicate a very robust framework where the training set can be used to identify the settings for the subsequent screening that generates the maximum gain. This robustness is depicted in Figures 6 - 9 where the largest deviations between gains are seen for assays 868 and 1460 with the fewest percentage of active compounds. The same trends with respect to gain is observed for both the training set as well as the external test set (the remainder of the chemical library) in all four investigated datasets.

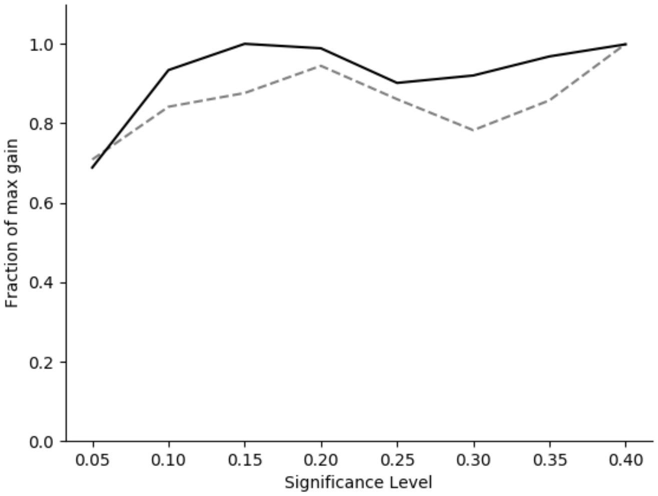

Figure 6: Fraction of maximum gain at different significance levels for PubChem Assay ID 868 (cost = 12, solid line = test set, dashed line = training set)

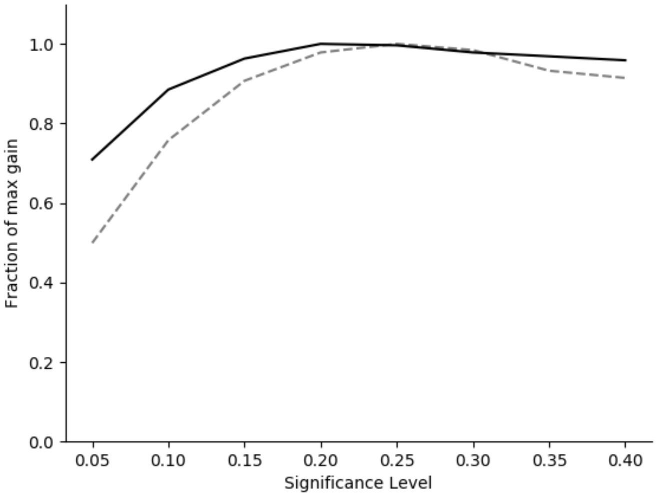

Figure 7: Fraction of maximum gain at different significance levels for PubChem Assay ID 1460 (cost = 12, solid line = test set, dashed line = training set)

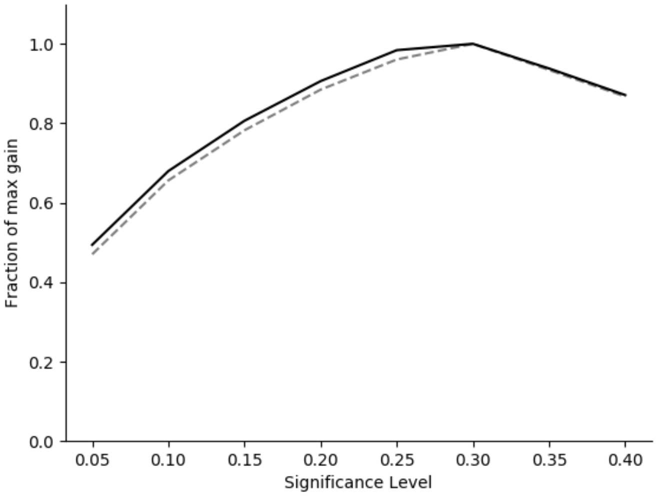

Figure 8: Fraction of maximum gain at different significance levels for PubChem Assay ID 2314 (cost = 12, solid line = test set, dashed line = training set)

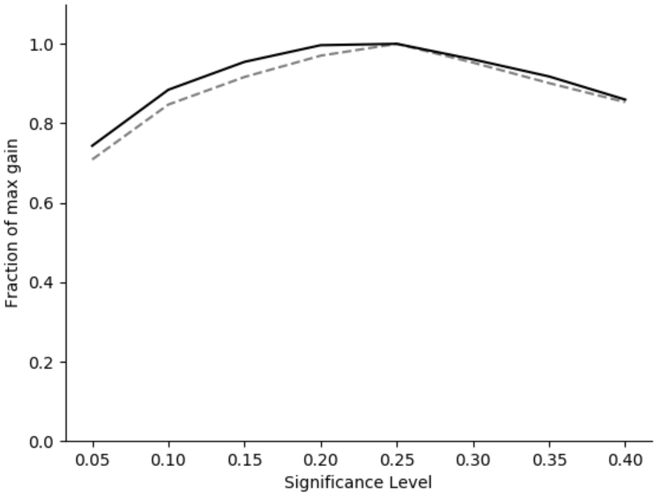

Figure 9: Fraction of maximum gain at different significance levels for PubChem Assay ID 2551 (cost = 12, solid line = test set, dashed line = training set)

Comparing the models based on fingerprints with those based on physicochemical descriptors the former retrieve more active test set compounds but, at the same time, predict more test compounds as active for subsequent screening (Figures 10 - 11). For the four datasets, screening between 20 % and 35 % of the remaining compounds identified between 60 % and 80 % of the active compounds. However, differences in total gain between the two approaches, fingerprints and physicochemical, are generally small except for the 868 dataset where the latter approach is much more effective and retrieves both more actives as well as shows a higher gain.

The main limitation associated with the presented strategy is the determination of the gain-cost ratios to use in the evaluation. We opted to use a ratio that roughly corresponded to cost neutrality, i.e. no gain or loss, of the full screening data, clearly this is a ery conservative measure since the produced hits from a screening campaign probably are associated with a future gain. However, the exact size of the gain to cost ratio is dependent on many factors and must be calibrated for each project.

To further study and substantiate the findings in this investigation, assessment of additional datasets are underway.

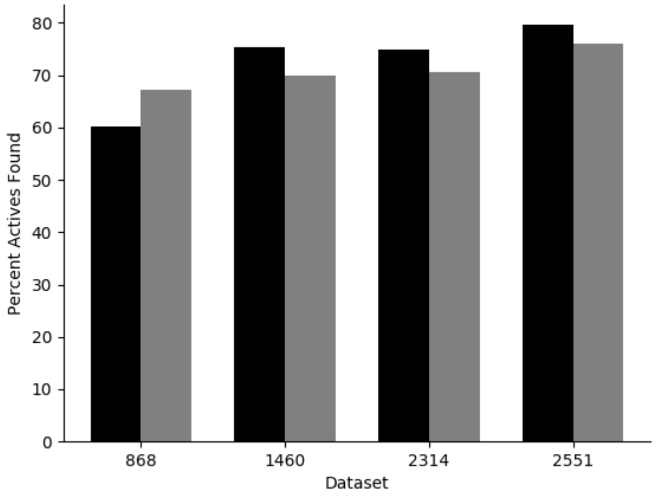

Figure 10: Average percentage retrieved active compounds (black bars: fingerprints and grey bars: physicochemical descriptors)

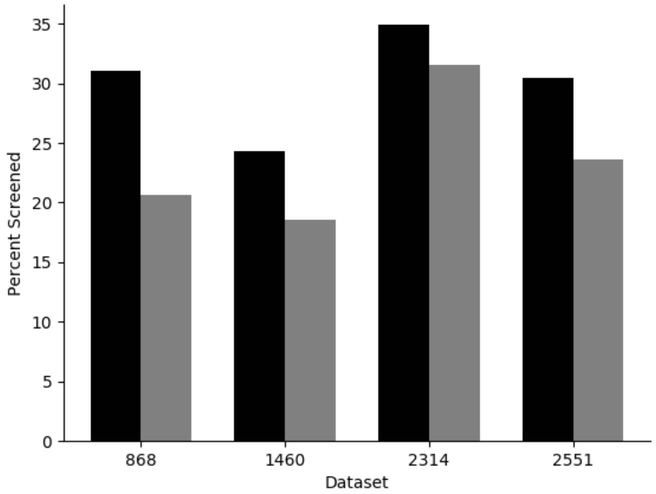

Figure 11: Average percentage screened compounds (black bars: fingerprints and grey bars: physicochemical descriptors)

4. Conclusions

This study investigated the usage of a combined strategy, with the aim of maximizing gain for screening campaigns, employing a gain-cost function in combination with CP in order to correctly identify a suitable significance level for the training set by internal validation that, subsequently, also represents a good level for screening an active compound enriched subset from the remainder of the library in question. Even though the internal validation did not always identify the correct optimal level for the enriched subset screening the decease, in terms of gain, was minimal with average decrease in gain of only 1.0 % and 2.1 %, respectively, for physicochemical descriptor-based models and fingerprint-based models.

Acknowledgments

The authors acknowledge Dr. Anna-Lena Gustavsson, Chemical Biology Consortium Sweden, CBCS, Karolinska Institutet, SciLifeLab, Stockholm, for fruitful discussions on the design of the gain-cost function.
The research at Swetox (UN) was supported by Stockholm County Council, Knut & Alice Wallenberg Foundation, and Swedish Research Council FORMAS. AMA gratefully acknowledges AstraZeneca for support (grant RG82877 RUL).

References

J. Bajorath. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.*, 1:882–894, 2002.

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

M. Eklund, U. Norinder, S. Boyer, and L. Carlsson. The application of conformal prediction to the drug discovery process. *Ann. Math. Artif. Intell.*, 74:117–132, 2013.

R. Macarron. Critical review of the role of hts in drug discovery. *Drug Discov. Today*, 11: 277–279, 2006.

R. Macarron, M. N. Banks, D. Bojanic, D. J. Burns, D. A. Cirovic, T. Garyantes, D. V. S. Green, R. P. Hertzberg, W. P. Janzen, J. W. Paslay, U. Schopfer, and G. Sitta Sittampalam. Impact of high-throughput screening in biomedical research. *Nat. Rev. Drug Discov.*, 10:188–195, 2011.

U. Norinder and S. Boyer. Conformal prediction classification of a large data set of environmental chemicals from toxcast and tox21 estrogen receptor assays. *Chem. Res. Toxicol.*, 29:1003–1010, 2016.

U. Norinder, L. Carlsson, S. Boyer, and M. Eklund. Introducing conformal prediction in predictive modeling. a transparent and flexible alternative to applicability domain determination. *J. Chem. Inf. Model.*, 54:1596–1603, 2014.

S. Paricharak, A. P. Ijzerman, A. Bender, and F. Nigsch. Analysis of iterative screening with stepwise compound selection based on novartis in-house hts data. *ACS Chem. Biol.*, 11:1255–1264, 2016a.

S. Paricharak, O. Mendez-Lucio, A. C. Ravindranath, A. Bender, A. P. IJzerman, and G. J. P. van Westen. Data-driven approaches used for compound library design, hit trage and bioactivity modeling in high-throughput screening. *Brief Bioinform.*, 2016b. doi: 10.1093/bib/bbw105. in press.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, Brucher M, M. Perrot, and . Duchesnay. Scikit-learn: Machine learning in python. *J. Chem. Inf. Comput. Sci.*, 34:1000–1008, 1994.

J. Sadowski, J. Gasteiger, and G. Klebe. Comparison of automatic three-dimensional model builders using 639 x-ray structures. *J. Chem. Inf. Comput. Sci.*, 34:1000–1008, 1994.

F. Svensson, U. Norinder, and A. Bender. Improving screening efficiency through iterative screening using docking and conformal prediction. *J. Chem. Inf. Model.*, 57, 2017a. doi: 10.1021/acs.jcim.6b00532. in press.

F. Svensson, U. Norinder, and A. Bender. Modelling compound cytotoxicity using conformal prediction and pubchem hts data. *Toxicol. Res.*, 6:73–80, 2017b.

A. Tropsha. *QSAR Modeling and QSAR Based Virtual Screening , Complexity and Challenges of Modern*, pages 7071–7088. Springer New York, New York, NY, 2009.

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World.* Springer, New York, NY, 2005.