Venn-Abers Predictors for Improved Compound Iterative Screening in Drug Discovery

Ruben Buendia RUBEN.BUENDIA@ASTRAZENECA.COM
Dept. of Information Technology, University of Borås, Sweden

Ola Engkvist OLA.ENGKVIST@ASTRAZENECA.COM
Discovery Sciences, AstraZeneca IMED Biotech Unit, Mölndal, Sweden

Lars Carlsson LARS.A.CARLSSON@ASTRAZENECA.COM
Quantitative Biology, Discovery Sciences, AstraZeneca IMED Biotech Unit, Mölndal, Sweden
Department of Computer Science, Royal Holloway, University of London, Egham Hill, Egham, Surrey, United Kingdom

Thierry Kogej THIERRY.KOGEJ@ASTRAZENECA.COM
Discovery Sciences, AstraZeneca IMED Biotech Unit, Mölndal, Sweden

Ernst Ahlberg ERNST.AHLBERG@ASTRAZENECA.COM
Predictive Compound ADME & Safety, Drug Safety & Metabolism, AstraZeneca IMED Biotech Unit, Mölndal, Sweden

Editors: Alex Gammerman, Vladimir Vovk, Zhiyuan Luo, Evgueni Smirnov and Ralf Peeters

Abstract

Iterative screening, where selected hits from a given round of screening are used to enrich a compound activity prediction model for the next iteration, enables more efficient screening campaigns. The portion of the compound library that should be screened in each iteration is often arbitrarily decided. This is because no accurate information between screening size and the number of hits to be retrieved exists. In this article, a novel method based on Venn-Abers predictors was used to determine the optimal number of compounds to be screened in order to get a desired number of hits. We found that Venn-Abers predictors provide accurate information to support a reliable and flexible decision about the portion size of the compound library that should be screened in each iteration. In addition, the method exhibited great ability in producing an enriched subset in terms of hits and their diversity.

Keywords: Drug discovery, Iterative screening, Probabilistic prediction, Venn-Abers

1. Introduction

In early stages of drug discovery, chemists, pharmacologists, and clinicians used to work closely together in interdisciplinary project teams. Thus, knowledge about medicinal chemistry from different perspectives was put together to design potentially active and suitable compounds for further testing (Drews, 2000). In the 1980s, high-throughput screening (HTS) was developed; and since then it became increasingly accepted as a consequence of important improvements in: combinatorial chemistry, data processing, robotics and sensors.

HTS enables rapid screening of large collections of compounds to find relationships with drug targets (Macarron, 2006; Mayr and Fuerst, 2008), in order to get some active compounds or hits, i.e. typically a compound is considered to be a hit when its activity against a desired drug target exceeds a threshold. A drug target might be a protein or an enzyme which function can be altered through the intervention of a drug in a way to favourably influence disease. The most promising hits will be starting points for further investigation. This stage is called lead generation, whose is followed by lead optimization. A higher number of diverse hits increase the chances of finding promising leads.

HTS has become one of the main techniques in early stages of drug discovery. However, HTS has important limitations, in particular HTS campaigns are expensive and time consuming. This is because the great majority of all screened compounds are inactive. Efforts have been focus on the identification of smaller screening sets which are likely to contain a higher fraction of diverse hits among screened compounds (Phatak et al., 2009). Despite these efforts and computational advances, HTS campaigns remain an expensive endeavour. Iterative screening might be considered the state of the art approach to make HTS more efficient. This consist of screening libraries in an iterative fashion, screening a subset of the compound library and using the results to decide which subset of compounds to screen for the next cycle. The decision is taken by a machine learning model. Thus, this approach is an iterative process of model building and subsequent prediction of new compounds. Iterative screening was proven valuable to enhance hits discovery and diversity in (Paricharak et al., 2016).

In this work, machine learning methods for compound activity prediction complemented with probabilistic prediction were used. This is of great utility for triaging compounds, and it enables screening of a set of potentially active compounds which is significantly enriched in terms of hits compared to a purely random sample of the compound collection. In drug discovery this approach is referred to as Quantitative Structure-Activity Relationship (QSAR) modelling, where the chemical structure of compounds is used as the predictor variable, and whether the compound is a hit or not as the target variable.

Any scoring classifier allows ranking compounds to optimize experimental testing. However, in an iterative screening scenario, the choice of the portion size of the compound library to be screened in each iteration cannot be approached by traditional machine learning methods. To exemplify this, the selection of a small portion of the highest ranked compounds would yield high enrichment but few hits identified. On the contrary a large portion would yield more hits identified because the portion would be larger; but lower enrichment because compounds lower in the ranking were selected, and thus the average probability of selected compounds being a hit would be lower.

We hypothesize that probabilistic prediction has the potential to offer accurate information to support a reliable and flexible solution to this choice. In order to test this hypothesis, Fast Venn-Abers predictors were used in this work (Vovk et al., 2015; Toccaceli et al., 2016).

In this paper, a QSAR model complemented with Venn-Abers predictors to predict the probabilities of a large number of compounds being active against three different targets was implemented. The Compounds were ranked according to these probabilities. Thereafter, three subsets for each target were selected according to different probability thresholds. For each subset, the cumulative sum of the probabilities provided by Venn-Abers predictors was compared to the amount of hits selected. The closeness of these numbers indicates the ability of the method to determine the size of the portion of the compound library that needs to be screened in order to retrieve a desired number of hits. This is, in an iterative screening scenario, the ability to provide suitable information to determine the size of the subset to screen in the next iteration. Finally, enrichment offered by the model in terms of hits and their diversity, was evaluated against random selection.

2. The Choice of Fast Venn-Abers Predictors

During an iterative screening campaign, the optimal number of compounds to be screened at the next iteration is usually not known. To tackle this problem, conformal prediction (Vovk et al., 2005) has been suggested in Svensson et al. (2017). This method provides a framework for generating confidence predictors with a fixed error rate (Vovk et al., 2005). This is achieved by comparing predictions and actual values for compounds present in a so-called calibration set. In a simple fashion, in the case of two classes classification, the labels (e.g. active, inactive) are assigned to the compound being predicted in a way it results to four different classes: active, inactive, both labels simultaneously or none of the labels. This can limit the use of conformal prediction as the two latter classifications (e.g. both or none of the labels) present little practical use.

The efficiency of a conformal predictor is then defined as the number of single label predictions which can vary depending on the confidence level applied. Thus, in order to optimize efficiency, multiple confident levels have to be tested which increase the computational time and the process complexity. In addition, for an iterative screening purpose, a confidence level that maximizes efficiency might not lead to a dataset of a suitable size for the next screening iteration. A confidence level that optimizes a gain-cost function for the training set has been suggested, but this assumes that the results remain consistent to the test set (Svensson et al., 2017).

Probabilistic prediction provides the probability distribution of each label for two given training and test sets. If these probabilities are true, it means that their cumulative sums for a certain number of compounds to screen would correspond to the number of hits retrieved. For the probabilities to be true, the predictors need to get the probabilities right, at least on average, for each value of the prediction. In such case, we can state that the probabilistic predictor is valid. Unfortunately, it is not possible to make point probabilistic predictions.

Venn predictors (Vovk et al., 2003) circumvent this problem by returning multiple probabilities, i.e. same number of probabilities as possible labels, one of which is the valid one. Besides, validity is restricted to calibration, i.e. probabilities are matched by observed frequencies. The only assumption made by Venn predictors is that the data is drawn independently from each other from an identical distribution (i.i.d.). Venn-Abers predictors are a natural class of Venn predictors that can be computed from any scoring classifier, e.g. support vector machine (SVM).

The Venn-Abers method enables automatic transformation of a scoring classifier into a Venn-Abers predictor (Vovk and Petej, 2014). The Venn-Abers method is a modification of the Zadrozny and Elkan's method (Zadrozny and Elkan, 2002) which is based on isotonic regression (Ayer et al., 1955) while the former overcomes the tendency to overfitting. This is because Venn-Abers predictors inherit the properties of Venn predictors, and therefore are perfectly calibrated, i.e. probabilities are matched by observed frequencies.

However, Venn-Abers predictors present two limitations. Firstly, in a transductive form, it requires the underlying classifier model and the isotonic regression being calculated twice for each test sample. This can lead to high computing cost which can be reduced by employing an inductive approach (see below). Secondly, as Venn-Abers predictors apply to binary problems, they returns two probabilities (p_0 and p_1) for each prediction which can be difficult to use in some cases.

In the inductive approach, the overall data set is split into a proper training and calibration sets. The training set is used to build the underlying machine learning classifier model, e.g. SVM, which produces scores. The calibration data are employed to "train" the isotonic calibrator which transforms the scores into estimated probabilities. The SVM scores are computed only once, while the isotonic regression is recalculated two times for each test object. Inductive Venn-Abers predictors (IVAPs) are automatically valid and perfectly calibrated (Vovk et al., 2015). While IVAPs reduce the computational load compared to the transductive Venn-Abers approach, it is still resource demanding for large datasets (e.g. big pharma HTS collection). As recalculating the isotonic regression for every label values and test object is the main limiting factor, it is therefore possible to exploit the fact that only one data point is added to an otherwise fixed calibration set. This way, a new algorithm that produces exactly the same results in a much reduced computational effort was proposed in (Vovk et al., 2015). This algorithm make the computational cost of Venn-Abers affordable even on large data set sizes. The algorithm is called Fast Venn-Abers (Toccaceli et al., 2016), and it is the one used in this work.

Point probabilities can be derived from the multiprobabilities output by Venn-Abers predictors. Different options to calculate the point probabilities can be used. One possibility is to combine the two probabilities p_0 and p_1 to minimize the regret, which is, according a chosen loss function, the error for choosing the wrong probability. In the case of a log loss, the probability p can be calculated following the Equation (1). Subsequently, the limitation of dealing with multiprobabilistic predictions is overcome as a single value p is suggested. In addition, the difference between p_1 and p_0 provide a valuable measurement of uncertainty. However, the point probabilities no longer benefit the calibration property of the multiprobabilistic prediction. Nevertheless, experimental evidence suggests that the point predictions still exhibit high accuracy (Vovk et al., 2015). It is important to note that the multiprobability prediction as represented by p_0 and p_1 returned by an IVAP always satisfy $p_0 < p_1$. For practical reasons, p_0 and p_1 can be interpreted as the lower and upper probability margins. In practice, (p_0 and p_1) are often close to each other for large training and calibration sets (Vovk et al., 2015).

$$p = \frac{p_1}{1 - p_0 + p_1} \tag{1}$$

An alternative to Venn-Abers predictors is Platt scaling (Platt, 2000). Platt scaling is a similar method but requires distribution assumptions as it fits a sigmoid as calibrator. The advantage of Venn-Abers predictors lay in the uncertainty information contained in the difference between p_0 and p_1, as well as in the fact that they are theoretically valid. Furthermore, Fast Venn-Abers predictors have been reported to outperform Platt scaling along several datasets (Vovk et al., 2015; Toccaceli et al., 2016). Their improvement over Platt scaling is larger the more the functional dependency of probabilities to scores departs from a sigmoid (Toccaceli et al., 2016).

2.1. Practical application of IVAPs

Here, a brief description of how to apply inductive Venn-Abers predictors is provided. Firstly the training set needs to be separated into proper training and calibration. Then, a machine learning model is used to calculate prediction scores for the calibration set and a test sample. These scores, including the test score, are sorted in lexicographical order. Using the sorted prediction scores and the real labels for the calibration data, two isotonic regressors are computed, one where the test sample assumes positive label and another one where it assumes negative label. The intersection of the test score with the positive regression line is p_1, i.e. the maximum probability of the test sample having the positive label. Analogously, its intersection with the negative regression line is p_0, i.e. the minimum probability of the test sample having the positive label.

3. Methods

3.1. Data

AstraZeneca in-house HTS data was utilized in this work. More specifically, three large HTS assays were selected. The activity of compounds, of a large compound library, were experimentally tested against three different potential drug targets. The three assays are described in Table 1. They were chosen to represent typical cases with high, moderate and low proportion of hits, denoted *HHRT* (High Hit Rate Target), *MHRT* (Moderate Hit Rate Target), and *LHRT* (Low Hit Rate Target) respectively.

Table 1: Characteristics of the Three Datasets Used

Target	Active	Inactive	% of Active
HHRT	47946	1643931	2.83%
MHRT	16410	1964086	0.83%
LHRT	6626	1954785	0.34%

The compounds were described by signature descriptors (Faulon et al., 2003) derived from the chemical structure of the compounds. Each signature corresponds to the number of occurrences of a particular substructure in the compound. The resulting dataset can be viewed as a sparse matrix of attributes were each signature is a column, and each row is a compound. The number of signatures is in the order of hundreds of thousands, however the density of the matrix, understood as features of compounds with non-zero value, is always under 0.1%.

3.2. Machine learning and Fast Venn-Abers Predictors

For each model, the dataset was split into an initial proper training set of 100 000 instances, a calibration set of 50 000 instances, and a test set containing the rest of the instances. The proper training and calibration sets correspond, approximately, to a 5% and a 2.5% of the full dataset respectively. Compounds for all three sets were picked randomly from each full HTS essay; thus proper training, calibration and testing were IID.

All models consisted of a support vector machine (Vapnik., 1995), and were implemented using LibLinear (Fan et al., 2008). The penalty parameter, i.e. cost, was automatically optimized by a new functionality in the LibLinear algorithm version 2.20. Venn-Abers predictors were calculated using the python/numpy function *VennABERS.py*. This function was developed by Paolo Toccacely in the frame of the EU project ExCAPE (Toccaceli et al., 2016); and it is an implementation of the Fast Venn-Abers Predictor described in (Vovk et al., 2015).

For each essay, after optimizing the cost using the proper training data, an SVM is trained as well using the proper training data. The resulting model was used to output scores for each compound of the calibration and test sets. The distance to the hyperplane was used as scoring function. Thereafter, the calibration set and the test scores were used as input to the *VennABERS.py* function which output calibrated probabilities of each compound in the test set for being a hit. Finally, since Venn-Abers produce two probabilities per instance, the log loss function in Equation (1) was used to obtain single probabilistic predictions that facilitate the ranking of compounds.

3.3. Screening Summary

The screening workflow proposed here used an initial screening of 150 000 compounds for each HTS essay. The results from this initial screening were used to produce probabilistic predictions, i.e. p, of compounds being a hit, which were used to rank compounds. At this point, three different probability thresholds at 1%, 5% and 10% probability were used to select compounds. Because the datasets were randomly split for each model, and new models were built for each threshold, the number of hits in proper training, calibration, and test sets slightly varied for each threshold.

3.4. Performance evaluation measures

Performance of the models was evaluated using the area under the ROC curve (Bradley, 1997), AUC. Enrichment of selected sets of compounds were evaluated in terms of number of hits retrieved and their respective diversity. This way, number of hits retrieved and percentage of hits selected were quantified for the different thresholds and targets. In order to evaluate diversity, molecular framework (MF) (Bemis and Murcko, 1996), topological framework (Schuffenhauer et al., 2007) (TF), and clusters were considered. Clusters were generated by FLUSH (Blomberg et al., 2009) on the basis of the ECFP4 fingerprint (Rogers and Hahn, 2010). Only MF and TF shared by at least a hit were considered. In the same way only clusters containing hits were considered. Both, hits and diversity enrichment, were compared against random selection.

Further, to evaluate the validity of the method, the cumulative sums of labels and probabilities of predicted compounds were plotted for each target. Afterwards, the number of expected and retrieved hits for the different thresholds were calculated. The minimum number of expected hits was calculated as the cumulative sum of p_0 of screened compounds for each combination of thresholds and targets. Analogously, the maximum number of expected hits was calculated as the cumulative sum of p_1 of screened compounds for each case. The actual number of expected hits was calculated as the cumulative sum of p of screened compounds for each case. In addition, calibration validity plots have been provided for a visual representation of the validity of the method.

Whereas experiments could have been run just one time per target, then selecting different thresholds on the output, they were repeated for each threshold and again for ROC and calibration plots. The rationale was controlling the influence of the random selection of proper training and calibration sets. Otherwise, both SVM and the Venn-Abers method are deterministic, therefore repeated experiments on the same model would provide the same outcome.

4. Results

4.1. Performance of the Models

Figures 1, 2 and 3 show the ROC of the three different models corresponding to the three targets. A target with a higher proportion of hits represent a less skewed dataset, i.e. a lower ratio Active/Inactive. Thus, in principle, it would have higher performance. This way the AUC of HHRT was 0.92 representing very high performance. The AUC of MHRT was 0.75 and the AUC of LHRT was 0.74, thus the performance of the MHRT was just slightly higher than the one of the LHRT, on the specific draw of the data set used.

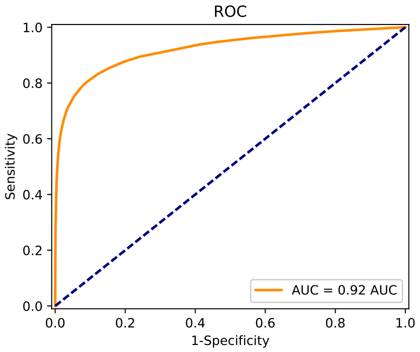

Figure 1: ROC curve of the model for the HHRT.

4.2. Hits Enrichment

Table 2 provides results representing hits enrichment. As expected lower thresholds and targets with higher hit rate select a higher number of compounds. Also as expected, higher thresholds offer higher accuracy and enrichment. Nevertheless, even at the lowest threshold (1%) of the HHRT, the enrichment is substantial when compared to random selection.

4.3. Diversity Enrichment

Table 3 shows the amount and percentage of MF, TF, and clusters contained in the test set that are represented in retrieved hits. Surely, lower thresholds selected more compounds and retrieved more hits; therefore lower thresholds identify a higher number and percentage

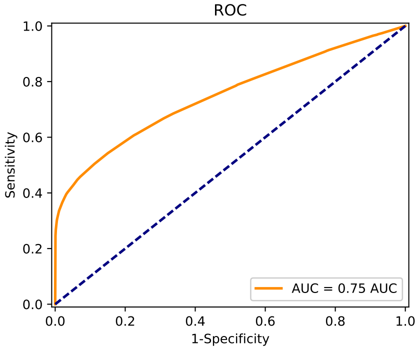

Figure 2: ROC curve of the model for the MHRT.

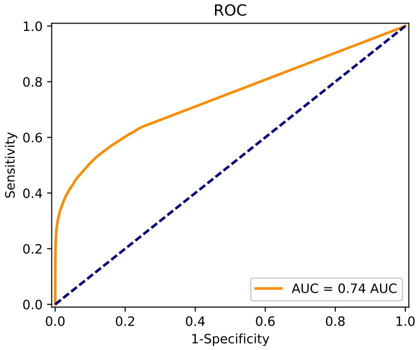

Figure 3: ROC curve of the model for the LHRT.

Table 2: Hits Enrichment Evaluation Results

Target (Compounds in test set)	Threshold	Hits in test set (%Hits)	Compounds Selected (% test set)	%Hits in the selection	Hits Retrieved (If Random)	%Hits Selected (If Random)
HHRT (1541705)	1%	43740 (2.8%)	505663 (32.8%)	7.9%	39969 (14346)	91.4% (32.8%)
	5%	43652 (2.8%)	107639 (7.0%)	30.5%	32851 (3047)	75.3% (7.0%)
	10%	43719 (2.8%)	76763 (5.0%)	39.6%	30434 (2177)	69.6% (5.0%)
MHRT (1830049)	1%	15185 (0.83%)	147804 (8.1%)	4.6%	6835 (1226)	45.0% (8.1%)
	5%	15210 (0.83%)	17301 (0.95%)	26.7%	4616 (144)	30.3% (0.95%)
	10%	15161 (0.83%)	8195 (0.45%)	49.5%	4057 (68)	26.8% (0.45%)
LHRT (1810970)	1%	6124 (0.37%)	45589 (2.5%)	4.8%	2175 (154)	32.8% (2.5%)
	5%	6127 (0.37%)	9870 (0.55%)	17.1%	1683 (33)	27.5% (0.55%)
	10%	6095 (0.34%)	4268 (0.24%)	33.6%	1432 (15)	23.5% (0.24%)

of MF, TF and clusters. In the same way, in targets with higher hit rate, more hits and consequently more MF, TF, and clusters were identified.

Comparing the percentage of compounds selected to the percentage of MF, TF, and clusters identified, diversity enrichment seems to be remarkable. This relative enrichment is higher the lower the number of compounds selected. To exemplify this, with a threshold of 10% probability in the LHRT, 0.24% of test compounds were selected; in those compounds, 16.2%, 16.8%, and 13.8% of all MF, TF and clusters respectively, contained in the test set, were represented. On the other extreme, with a threshold of 1% probability in the HHRT, 32.8% of the compounds were selected; in those compounds, 86.4%, 79.1%, and 83.7% of MF, TF and clusters respectively, were represented. Nevertheless, it is still to prove whether this diversity enrichment would hold after several iterations.

Random selection retrieve a much lower amount of hits that selection by probability ranking. For that reason, random selection identify a much lower amount of MF, TF, and clusters. However, more MF, TF, and clusters per hit were represented in randomly selected hits; what, although not desired, could be expected. The reason is that QSAR models predict compounds with structures similar to known hits as hits. Therefore, predicted hits have a higher probability to share MF, TF or cluster with known hits.

4.4. Venn-Abers Method Empirical Validation

Figures 4, 5 and 6 show the cumulative sums of labels and probabilities of ranked compounds against number of compounds to screen, for each target respectively. These figures are a suitable proof of validity since they provide an estimation of the number of hits that would be retrieved together with the retrospectively observed number of hits retrieved. In addition the cumulative sums of p_0 and p_1 show the upper and lower margins of expected hits for each number of compounds to screen. Finally, the curve provide information about where the relation between the number of compounds screened and the number of hits retrieved is maximized.

Expected and actual number of hits were close for all three targets. The difference between upper and lower margins of expected hits, i.e. cumulative sums of p_1 and p_0 respectively, was low for all three targets. This mean that average uncertainty was low. Nevertheless, HHRT exhibited significant higher precision and lower uncertainty than MHRT which exhibited significant higher precision and lower uncertainty than LHRT.

Figures 7, 8 and 9 visualize the accuracy of the calibration on the test set. We observe how the calibration is very accurate for the bins containing a high number of predicted compounds. Otherwise, calibration is less accurate for bins with very low number of compounds. This is logical because perfect calibration can only be expected on average, i.e. the average of output probabilities would match the average of predicted compounds being hits. Thus the calibration accuracy is subjected to the law of large numbers (Bernoulli, 1713).

Table 4 shows the hits retrieved, for each combination of thresholds and targets, against predictions output by the Venn-Abers predictors for the selected subsets of compounds. This way the cumulative sums of p and labels are shown. In addition the cumulative sums of p_0 and p_1 are shown because they represent the minimum and maximum hits retrieval that can be expected.

Table 3: Diversity Enrichment Evaluation Results

Target	Thold	MF in test set	MF Retrieved (Random)	%MF Selected (Random)	TF in test set	TF Retrieved (Random)	%TF Selected (Random)	Clusters in test set	Clusters Retrieved (Random)	%Clusters Selected (Random)
HHRT	1%	19171	16566 (7703)	86.4% (40.2%)	6278	4964 (3112)	79.1% (49.6%)	16107	13474 (7836)	83.7% (48.6%)
	5%	19128	12721 (2134)	66.5% (11.2%)	6272	3853 (1121)	61.4% (17.9%)	16083	9317 (2290)	57.9% (14.2%)
	10%	19127	11615 (1637)	60.7% (8.6%)	6271	3577 (900)	57.0% (14.4%)	16080	8115 (1789)	50.5% (11.1%)
MHRT	1%	9573	3190 (973)	33.3% (10.2%)	4150	1185 (630)	28.6% (15.2%)	9358	2784 (1013)	29.7% (10.8%)
	5%	9602	1862 (137)	19.4% (1.4%)	4149	714 (116)	17.2% (2.8%)	9399	1203 (140)	12.8% (1.5%)
	10%	9569	1542 (62)	16.1% (0.6%)	4136	590 (54)	14.3% (1.3%)	9369	859 (64)	9.2% (0.7%)
LHRT	1%	4218	1100 (143)	26.1% (3.4%)	1944	474 (110)	24.4% (5.7%)	4352	1142 (150)	26.2% (3.4%)
	5%	4187	833 (29)	19.9% (0.7%)	1947	384 (25)	19.7% (1.3%)	4327	770 (29)	17.8% (0.7%)
	10%	4182	676 (17)	16.2% (0.4%)	1945	326 (16)	16.8% (0.8%)	4319	597 (17)	13.8% (0.4%)

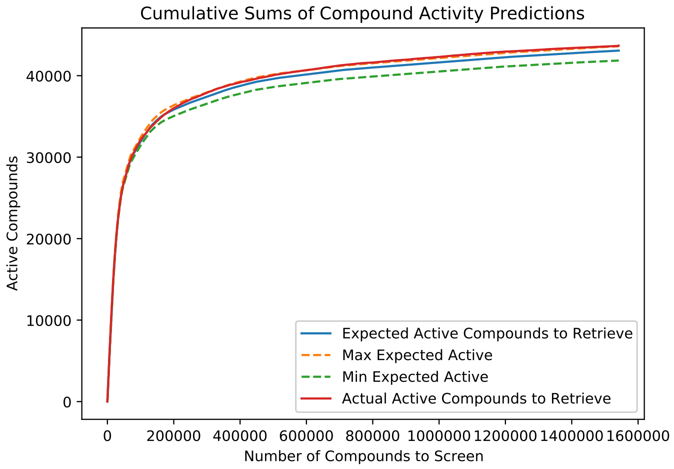

Figure 4: Cumulative sums of predicted probabilities and hits retrieved for the HHRT. The cumulative sums of p_0 and p_1 represent predictions of minimum and maximum hits to retrieve; the cumulative sum of p represent the prediction of hits to retrieve; the cumulative sum of labels represent the retrospectively observed hits retrieved.

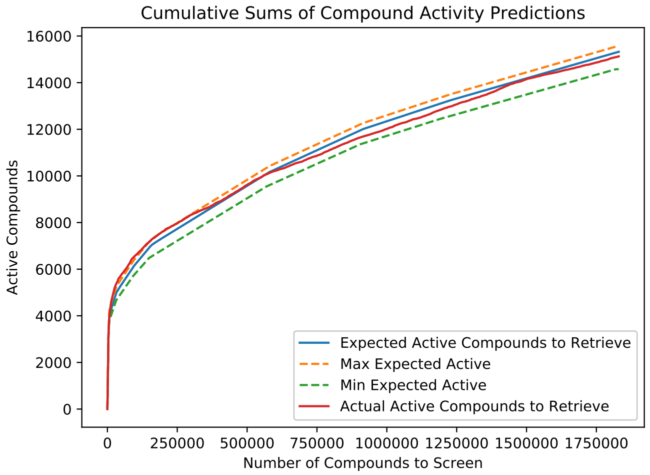

Figure 5: Cumulative sums of predicted probabilities and hits retrieved for the MHRT. The cumulative sums of p_0 and p_1 represent predictions of minimum and maximum hits to retrieve; the cumulative sum of p represent the prediction of hits to retrieve; the cumulative sum of labels represent the retrospectively observed hits retrieved.

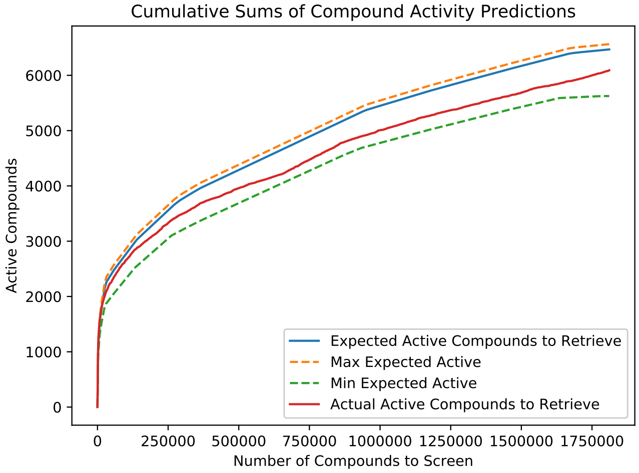

Figure 6: Cumulative sums of predicted probabilities and hits retrieved for the LHRT. The cumulative sums of p_0 and p_1 represent predictions of minimum and maximum hits to retrieve; the cumulative sum of p represent the prediction of hits to retrieve; the cumulative sum of labels represent the retrospectively observed hits retrieved.

Empirical results of Table 4 show remarkably well-calibrated predictions. This way the cumulative sum of labels is always very close to the cumulative sum of p; with a relative root square error of $3.1 \pm 3.9\%$. In addition the cumulative sum of labels is always in between the cumulative sums of p_0 and p_1; except in one case (HHRT at 10% threshold) where it is very slightly under the low margin. This might happen, despite perfect calibration, because probabilities must be matched by observed frequencies on average in the long run.

Table 4: Method Validation Evaluation Results

Target	Threshold	Hits Retrieved (cumsum of labels)	Hits expected (cumsum of p) [at least (cumsum of p_0) - at most (cumsum of p_1)]	Relative difference between expected and retrieved
HHRT	1%	39969	40232 [39345 - 40699]	0.66%
	5%	32851	32421 [31964 - 32866]	1.32%
	10%	30434	30969 [30511 - 31415]	1.74%
MHRT	1%	6835	6852 [6372 - 7104]	0.25%
	5%	4616	4833 [4612 - 5035]	4.59%
	10%	4057	4138 [3935 - 4361]	1.98%
LHRT	1%	2175	2477 [2112 - 2604]	12.98%
	5%	1683	1665 [1442 - 1756]	1.08%
	10%	1432	1483 [1296 - 1593]	3.50%

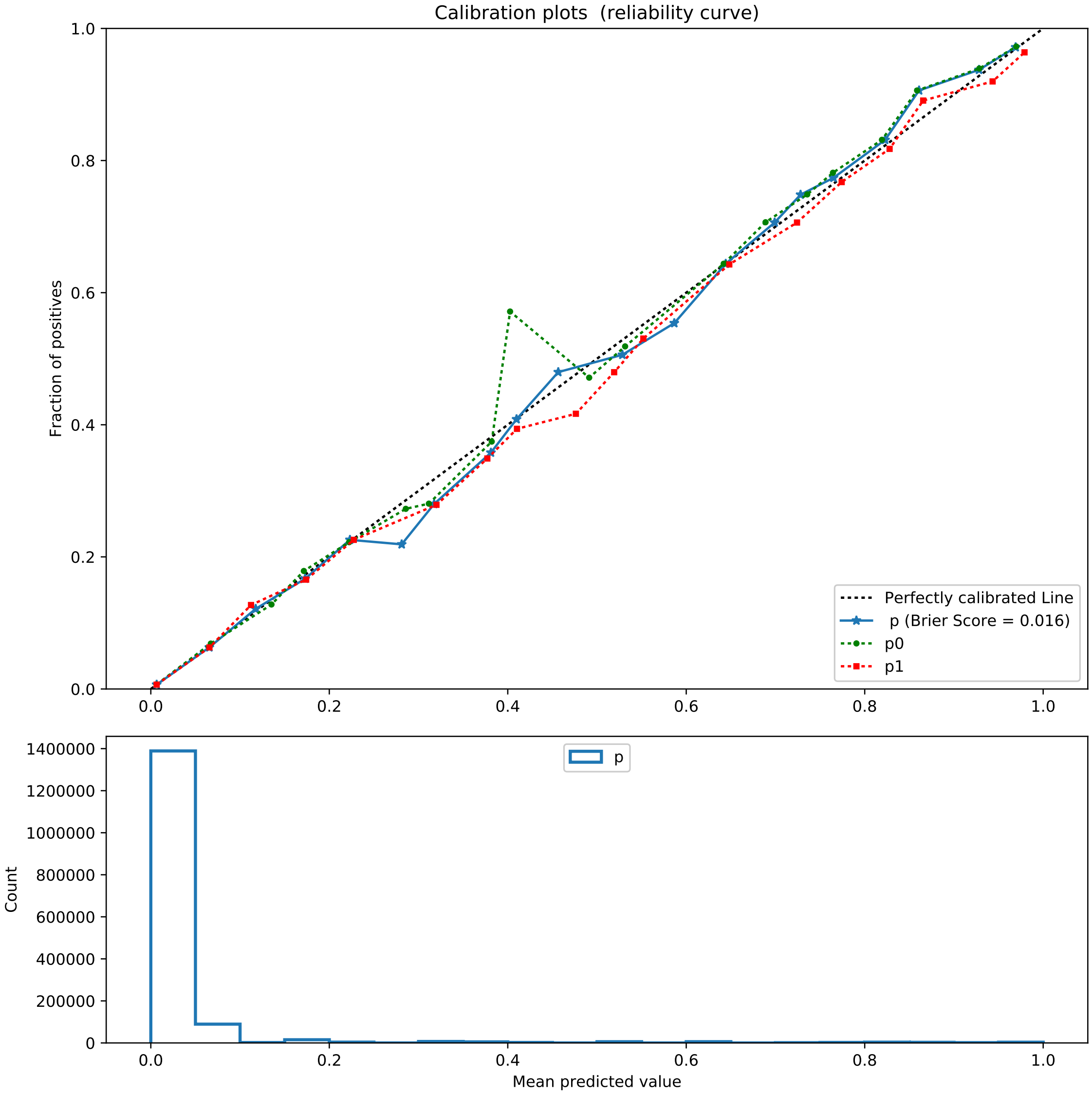

Figure 7: Calibration plots for HHRT, showing how well calibrated the predictions are compared to observed outcome. The plots have been generated by picking 20 bins linearly distributed between 0 and 1, forming the expected probability on the x axis. The fraction of positives is the retrospectively observed probability of being a hit.

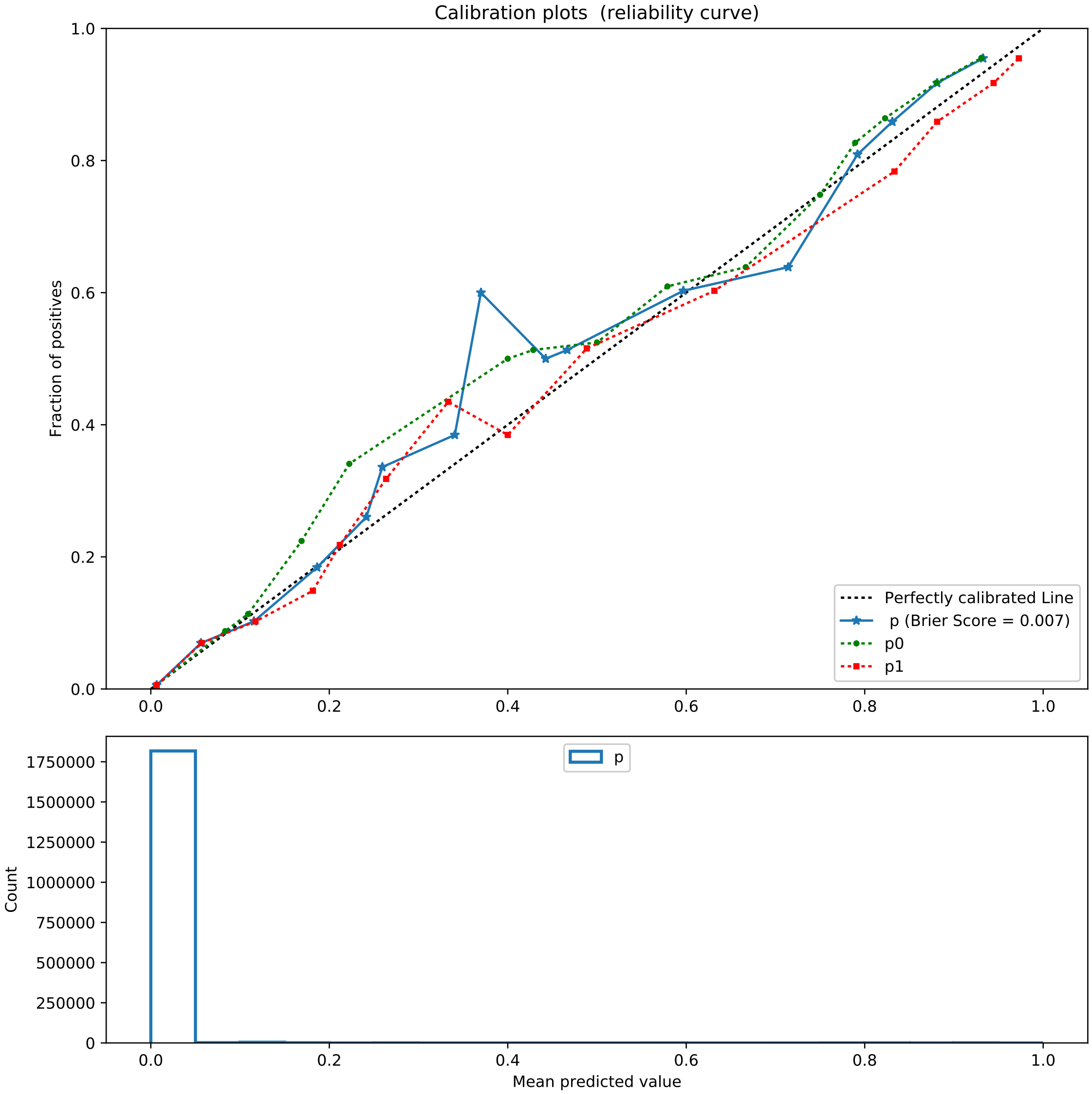

Figure 8: Calibration plots for MHRT, showing how well calibrated the predictions are compared to observed outcome. The plots have been generated by picking 20 bins linearly distributed between 0 and 1, forming the expected probability on the x axis. The fraction of positives is the retrospectively observed probability of being a hit.

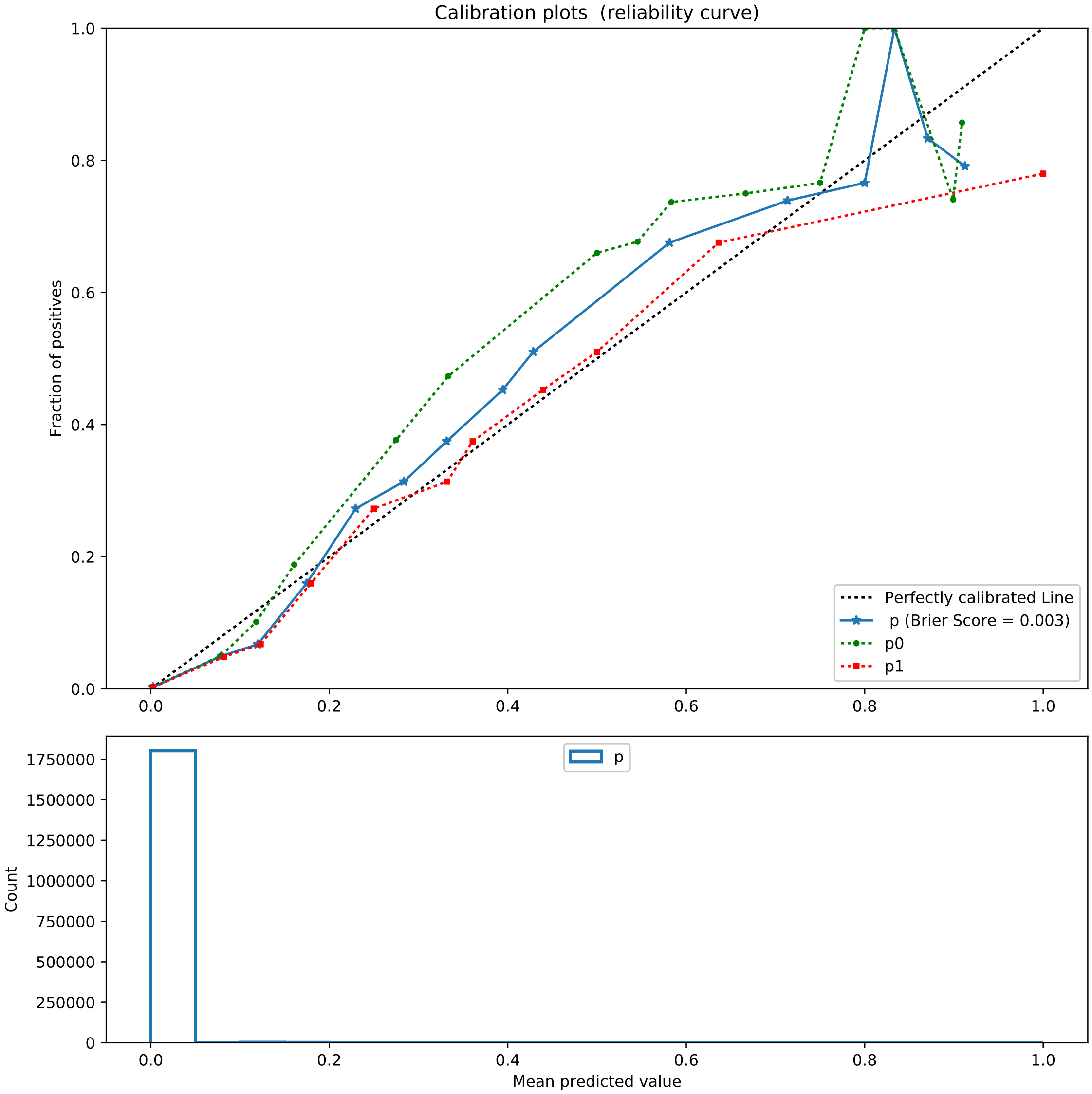

Figure 9: Calibration plots for LHRT, showing how well calibrated the predictions are compared to observed outcome. The plots have been generated by picking 20 bins linearly distributed between 0 and 1, forming the expected probability on the x axis. The fraction of positives is the retrospectively observed probability of being a hit.

5. Conclusions and Future Work

In this work a novel method to determine the number of compounds that need to be screened in order to get any desired number of hits was proposed. The method consisted of an SVM combined with Venn-Abers predictors. In an iterative screening scenario, this method offer the necessary information to decide the size of the portion of the compound library that should be screened in each iteration.

The method was validated retrospectively using three drug targets which are representative targets with low, moderate and high proportion of hits. Results showed that the proposed method results in an accurate and reliable estimation of the number hits that would be retrieved after screening a certain number of compounds.

We can observe Figures 4, 5 and 6 to exemplify the way to use the proposed method. In a prospective scenario, we would need to screen an initial set of 150 000 compounds, selected randomly from the compound library, and assign them to proper training and calibration sets. After applying the method we would obtain Figures 4, 5 and 6 for, HHRT, MHRT and LHRT respectively; however without the line of actual hits to retrieve. Observing the curve provides information about where the relation between gain and cost is maximized, i.e. considering hits retrieved being the gain and number of compounds screened the cost. At that point, a decision on how many compounds should be screened needs to be taken. This decision might have different restraining factors. For this reason the flexibility of the method is a great advantage. We can imagine three different scenarios. For example, for the MHRT, screening a number of compounds that stop just before the inflection point of the curve is decided. This is a sensible decision to maximize enrichment in a reduced number of iterations. In this case around 100 000 compounds would be selected and around 6 000 hits retrieved. Otherwise, for the HHRT, it could be decided to screen a reduced number of compounds because this maximizes enrichment while still retrieving a substantial number of hits. Finally, for the LHRT, time constrains could produce the decision of screening a large number of compounds in order to generate a lead with only one iteration. This way, around 250 000 compounds are screened in order to retrieve over 3 000 hits. In all three cases, the difference between the number of hits expected over which the decision was taken and the actual number of hits retrieved would have been very small.

In addition to the proposal and validation of the probabilistic prediction method, the ability of the model to produce enriched subsets was evaluated in terms of hits and their diversity. The model exhibited very good ability to rank compounds. This way, in all subsets selected, i.e. combinations of targets and thresholds, the enrichment in terms of number and percentage of hits was remarkable as compared to random selection. This way, for the smallest subset which corresponded to a 0.24% of the test dataset, the enrichment of hits was over 100 times higher than random. For the largest subset which corresponded to a 32.8% of the test dataset, the enrichment of hits was near 3 times higher than random. Diversity enrichment was as well high. However, randomly selected hits were more diverse than hits selected by the proposed method. This indicates that an overall strategy for compound selection could benefit from exploration approaches. Otherwise, the present method is only targeting exploitation.

In conclusion, the method proposed in this article have the potential to make HTS for drug discovery more efficient. Further, the results obtained warrant a more comprehensive study using more essays against other targets. In addition, enrichment should be compared to current screening strategies. Finally, exploration strategies need to be as well considered.

Acknowledgments

This work was supported by the Swedish Knowledge Foundation through the project Data Analytics for Research and Development (20150185).

References

M. Ayer, H. D. Brunk, G. M. Ewing, W. T. Reid, and E. Silverman. An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics*, 26:641–647, 1955.

G. W. Bemis and M. A. Murcko. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.*, 39:2887–2893, 1996.

J. Bernoulli. *Ars conjectandi:*. Impensis Thurnisiorum, fratrum, 1713. URL https://books.google.se/books?id=kD4PAAAAQAAJ.

Niklas Blomberg, David A. Cosgrove, Peter W. Kenny, and Karin Kolmodin. Design of compound libraries for fragment screening. *Journal of Computer-Aided Molecular Design*, 23(8):513–525, Aug 2009. ISSN 1573-4951. doi: 10.1007/s10822-009-9264-5. URL https://doi.org/10.1007/s10822-009-9264-5.

Andrew P. Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *PATTERN RECOGNITION*, 30(7):1145–1159, 1997.

J. Drews. Drug discovery: a historical perspective. *Science*, 287(5460):1960–1964, 2000.

R.E. Fan, K.W. Chang, C.J. Hsieh, X.R. Wang, and C.J. Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

J.L. Faulon, J.D.P. Visco, and R.S. Pophale. The signature molecular descriptor. 1. using extended valence sequences in qsar and qspr studies. *Journal of Chemical Information and Computer Sciences*, 43(3):707–720, 2003.

R. Macarron. Critical review of the role of hts in drug discovery. *Drug Discovery Today*, 11 (7-8):277–279, 2006.

L.M. Mayr and P. Fuerst. The future of high-throughput screening. *J. Biomol. Screening*, 13:443–448, 2008.

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, 2016.

S.S. Phatak, C.C. Stephan, and C.N. Cavasotto. Highthroughput and in silico screenings in drug discovery. *Expert Opin. Drug Discovery*, 4:947–959, 2009.

J. C. Platt. Probabilities for sv machines. *Advances in Large Margin Classifiers*, pages 61–74, 2000. URL https://ci.nii.ac.jp/naid/10027760454/en/.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t. URL https://doi.org/10.1021/ci100050t. PMID: 20426451.

Ansgar Schuffenhauer, Peter Ertl, Silvio Roggo, Stefan Wetzel, Marcus A. Koch, and Herbert Waldmann. The Scaffold Tree - Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *Journal of Chemical Information and Modeling*, 47(1):47–58, January 2007. doi: 10.1021/ci600338x. URL http://dx.doi.org/10.1021/ci600338x.

F. Svensson, A. M. Afzal, U. Norinder, and A. Bender. Maximizing gain in high-throughput screening using conformal prediction. *J Cheminform*, 10:7, 2017.

P. Toccaceli, I. Nouretdinov, Z. Luo, V. Vovk, L. Carlsson, and A Gammerman. Excape project, wp1, probabilistic prediction. Technical report, Royal Holloway, University of London and AstraZeneca, 2016.

Vladimir Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag, New York, 1995.

V. Vovk and I. Petej. Venn-abers predictors. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014.

Vladimir Vovk, Glenn Shafer, and Ilia Nouretdinov. Self-calibrating probability forecasting. In *Advances in Neural Information Processing Systems 16 [Neural Information Processing Systems, NIPS 2003, December 8-13, 2003, Vancouver and Whistler, British Columbia, Canada]*, pages 1133–1140, 2003. URL http://papers.nips.cc/paper/2462-self-calibrating-probability-forecasting.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

Vladimir Vovk, Ivan Petej, and Valentina Fedorova. Large-scale probabilistic predictors with and without guarantees of validity. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 892–900, Cambridge, MA, USA, 2015. MIT Press. URL http://dl.acm.org/citation.cfm?id=2969239.2969339.

B. Zadrozny and C. Elkan. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 694–699, New York, 2002. ACM Press.