

# Using Venn-Abers Predictors to assess Cardio-Vascular Risk

**Ernst Ahlberg**

ERNST.AHLBERG@ASTRAZENECA.COM

*Predictive Compound ADME & Safety, Drug Safety & Metabolism, AstraZeneca IMED Biotech Unit, Mölndal, Sweden*

**Ruben Buendia**

RUBEN.BUENDIA@HB.SE

*Dept. of Information Technology, University of Borås, 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*

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

## Abstract

This study investigates a method for predicting compound risk based on in vitro assay data and estimated  $C_{max}$ , the maximum concentration of a drug in the body. The method makes use of Venn-Abers predictors and Support Vector Machines to compute compound risk with respect to a biological target. The method has been applied to in vitro ion-channel data generated to assess cardiac risk and introduces a more intuitive way to reflect cardiac risk.

**Keywords:** Drug discovery, Venn-Abers Prediction, Cardio-Vascular Risk, Decision support

## 1. Introduction

Drug discovery is a costly and time consuming process that involves several steps and needs to take a multitude of factors into account (Bunnage, 2011; Gautam and Pan, 2016). First a biological pathway that can effect the disease of interest needs to be identified. Then a drug target, usually a protein or enzyme, needs to be located within that pathway. The drug target has to be susceptible to intervention by a compound in such a way that the disease state is affected and the patient is stabilized or better still cured of the disease. Focus is then shifted to find a compound that interacts with the target and gives the desired physiological effect. To achieve that, the compound needs to be highly active against the drug target but it also needs to be specific to that target and not be active against targets that can cause a safety risk, off targets. In compound optimization all these constraints need to be taken into account together with the information about how the human body can handle the compound through absorption, distribution, metabolism and excretion (ADME) (Ballard et al., 2012). Safety liabilities can stem from all organs in the body and their respective biological processes. In this work we focus on cardiovascular, CV, safety liabilities.

Cardiovascular safety liabilities are a major cause of drug attrition in all stages of drug discovery and development. In early stages of drug discovery compounds can be screened

to reveal potential CV risk using in vitro assays of selected ion channels that regulate heart function. At AstraZeneca, the following ion channels are screened routinely hERG, NaV1.5, CaV1.2, Kv4.3 and Kv7.1. Inhibition of these ion channels is an early indicator of CV risk, thus the assays can be used to rank compounds. Quantitative Structure-Activity Relationship (QSAR) models have been used to predict the outcome of these assays for specific compounds. In practice screening data has been preferred since such predictions generally lack a good measure of the risk for the individual compounds.

In 2005 Vovk, Gammerman and Shafer introduced the concept of Venn predictors ([Vovk et al., 2005](#)). Venn predictors offer a way to assign a calibrated probability to predictions. Venn-Abers prediction is a special case of Venn prediction that can be applied on top of a machine learning model under standard assumptions regarding data generation([Vovk and Petej, 2014](#)). In essence this means that the predicted probabilities reflect the long-term relative frequencies.

Assessing potential safety risk for a compound is commonly done in terms of a safety margin in terms of  $n$ -folds above expected  $C_{max}$ , the maximum concentration of a drug in the body after administration of the maximum recommended therapeutic dose. To mimic that, we propose to compute the probability that the  $C_{IC50}$ , ie the concentration at which 50% inhibition is observed, of an assay is below  $n$ -folds of  $C_{max}$  and thus create an individualized prediction for the compound rather than a general threshold. In this work we show how Venn-Abers predictors can be applied to a panel of assays, how it can be used together with  $C_{max}$  to assess cardiac risk of potential drugs and be an effective way to deliver enhanced decision making to projects.

## 2. Method

To obtain compound specific probabilities of cardiovascular risk it is necessary to take the required compound exposure or compound concentration into account. Compound safety is assessed as fold difference, or ratio, between the maximum concentration of the drug in the body,  $C_{max}$ , and the  $C_{IC50}$  of the off target interaction, thus experimental margin  $m_E = \frac{C_{IC50}}{C_{max}}$ . To predict a test compound its  $C_{max}$  value is used together with a desired margin, expressed in folds,  $f$ , of  $C_{max}$ , thus calculated margin  $m_C = f * C_{max}$ ,  $f \in [10, 30, 100]$ . Using the desired  $m_C$  gives a concentration, on the same scale as the  $C_{IC50}$  data from the in vitro assay. By using the compound specific concentration as cutoff for the  $C_{IC50}$  data results in a binary classifier that allows us to predict if the compound of interest is likely to have a  $C_{IC50}$  above or below the cutoff,  $m_C$ . To assess the target risk, the aim is to predict the probability that the off target  $C_{IC50}$  is below  $m_C$ . Thus, it is not sufficient to obtain a predicted label, a precise probability of the likelihood is also required.

In this paper we used Venn-Abers ([Vovk and Petej, 2014](#)) predictors to calculate that risk. Venn-Abers predictors are a special case of Venn predictors, ([Vovk et al., 2003](#)), implemented on top of a scoring classifier. Venn-Abers predictors inherits the properties of Venn predictors, i.e. that the multiprobabilistic predictions are perfectly calibrated ([Vovk and Petej, 2014](#)) which means that the probabilities are matched by observed frequencies. To apply Venn-Abers predictors, the only assumption needed is that the data is drawn independently from each other from an identical distribution (*i.i.d.*)

Many machine learning algorithms for classification are scoring classifiers, thus returning a prediction score  $s(x)$  where the actual label prediction is obtained by comparing the prediction score to a threshold. Support Vector Machines (SVM's), used here, are scoring classifiers, returning the distance to the hyper plane as a prediction score  $s(x)$ .

If scores can be calibrated by applying a monotonically increasing function  $g$  to  $s(x)$  then  $g(s(x))$  can be used as a valid probability, i.e. the predictors get the probabilities right, at least on average. Such a calibrator can be obtained by isotonic regression ([Ayer et al., 1955](#)). By obtaining the scores from the machine learning model for the calibration set and for the test example, isotonic regression can be applied twice for a binary problem, once for each label. By assuming that the test example belongs to a certain class and using isotonic regression to obtain the corresponding probability, the Venn-Abers method output a probability for each of the two labels, i.e.  $p_0$  and  $p_1$ . However, label specific probabilities are not desirable to make an over all assessment and thus log loss is used to obtain a precise probability that minimizes regret

$$p = \frac{p_1}{1 - p_0 + p_1}. \quad (1)$$

## 2.1. Data

In vitro data to assess cardiac liabilities is generated in a staged approach at AstraZeneca and uses the hERG channel as a first pass filter. If a compound has a  $C_{IC50}$  lower than  $10\mu\text{M}$  in the hERG in vitro test then follow up measurements are performed in the in vitro assays for NaV1.5, Kv4.3 and Kv7.1. As a result the data set for hERG is about five times larger compared to NaV1.5. For this study, targets have been limited to hERG and NaV1.5. The standard experimental concentration range in Molar (M) of the assays are between  $10\text{nM}$  and  $33\mu\text{M}$  with possible extension to  $100\mu\text{M}$ . The defined response from the experiment is a calculated  $C_{IC50}$  value.

Human  $C_{max}$  data compiled from internal sources at AstraZeneca together with experimental data in at least the hERG assay, generated a list of 67 compounds. This data set was used as test set and all compounds in the test set were removed from the training data prior to model building.

## 2.2. Training Procedure

For each compound and each  $m_C$  the proposed procedure the response is binary,  $[0, 1]$  and calculated from the  $C_{IC50}$  data by comparing to  $m_C$ . The training examples with  $C_{IC50} \leq m_C$  are assigned label 1, and training examples with  $C_{IC50} > m_C$  are assigned label 0. For each such data set the available training data was separated into a proper training set (75% of the data) used for building the machine learning model and a Venn-Abers calibration set (25% of the data).

In this work a Support Vector Machine model system was used with a linear kernel and the cost parameter set to 0.05 for all models, based on previous experience. The signature descriptors ([Faulon and Churchwell, 2003](#); [Faulon et al., 2003](#)) were used as objects or feature vector and the binary response from the assay data as response. The signature descriptor describes a compound by a set of strings and corresponding counts where the strings represent subgraphs of the compound, centered at an vertex (atom) and expanded

to include all neighbor vertexes  $h$  edges (bonds) away from the centre vertex. Signature descriptors are calculated for all atoms in a compound. The signature descriptors used in this work were generated by CDK [Steinbeck et al. \(2003\)](#) and the signatures were limited to  $h \in [1, 3]$  thus one to three bonds away from the centre atom. The signature vectors are very sparse, with an information density below 1%. The distance to the hyperplane was used as scoring function for the Venn-Abbers method. A single probability prediction was obtained with the method that minimizes the maximum regret under log loss.

Using this procedure, a series of QSAR models were built to predict cardiac risk for individual compounds at  $m_C$  for each compound respectively. Compounds were considered safe if  $m_C < C_{IC50}$ . This way, a test compound  $x$  will be predicted to have an  $C_{IC50}$  over  $m_C$  for  $x$ , if similar compounds have an  $C_{IC50}$  over  $m_C$  for  $x$ .

The available training data for hERG was 70375 compounds which generated a total of 125352 signatures. For NaV1.5 the numbers were 24351 which generated 130469 signatures.

### 3. Results

#### 3.1. Example cases for risk assessment

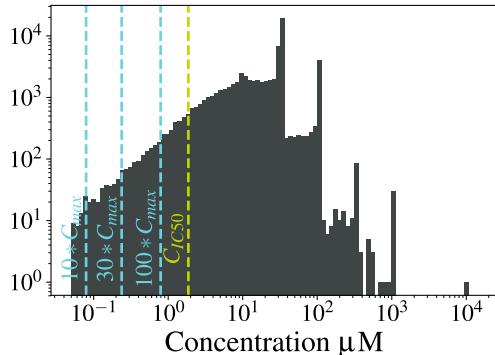
To demonstrate the benefit of using Venn-Abbers probabilities to assess cardiovascular risk three example compounds have been retrieved from the test data. The three compounds, one of which has low, one moderate and the last high risk exemplify how the data can be used and presented to projects and is showcased using hERG data.

Figure 1 shows a histogram of the training data based on  $C_{IC50}$  values for each test compound. On the histogram, the test compound  $C_{IC50}$  is marked as a dashed green line and the three  $m_C$  based on 10, 30 and 100 fold of test compound  $C_{max}$  are shown as blue lines. This plot shows how the  $C_{IC50}$  data is separated into binary data sets and Figure 2 shows the respective distribution of actives, label 1, and in-actives, label 0, in each data set respectively. For the low risk compound the training data sets are very unbalanced. As can be seen in Figure 1, the moderate and the high risk compound have the same  $C_{max}$ , consequently, the respective subfigures in Figure 2 are identical.

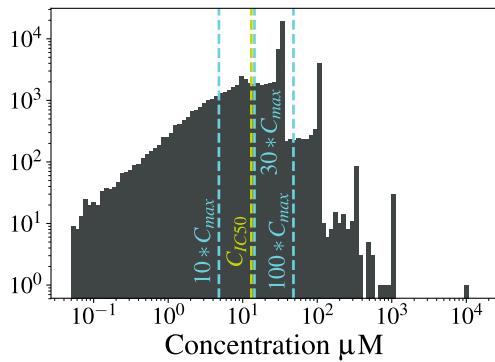
The resulting predictions are visualized in Figure 3. Both sub figures show predicted probabilities of risk on the y axis and each compound is distinguished by color. Sub figure 3(a) shows the results based on  $f$  on the x axis and shows  $m_E$  for each compound as dashed lines where the lines and the dots for the predicted probabilities share the same color for each compound. Sub figure 3(b) shows the predicted probabilities against concentration, with the  $C_{IC50}$  values for the respective compounds as dashed lines. The two sub figures shows a nice separation between the three cases and that it is possible to detect the change for the moderate risk compound as the required  $m_C$  increase. The figures also show the difference in risk between the low and high risk compounds that have very similar  $C_{IC50}$  values but where the  $C_{max}$  makes the difference resulting in a sufficient margin for the low risk compound and no margin for the high risk compound.

#### 3.2. hERG

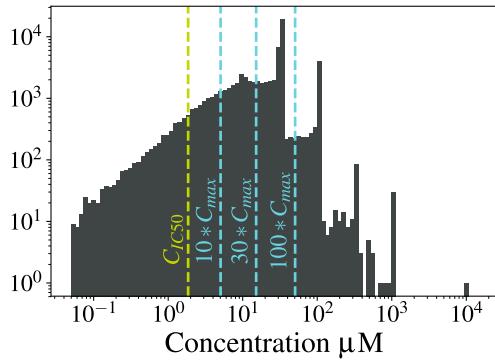
This section contains summary plots for the test data on the hERG target. Figure 4 is a representation of predicted risk for each  $f$ , showing that there is a good separation between



(a) low risk compound



(b) moderate risk compound



(c) high risk compound

Figure 1: Distribution of training data (IC50) and overlaid experimental  $C_{IC50}$  and  $m_C$  for the three example compounds

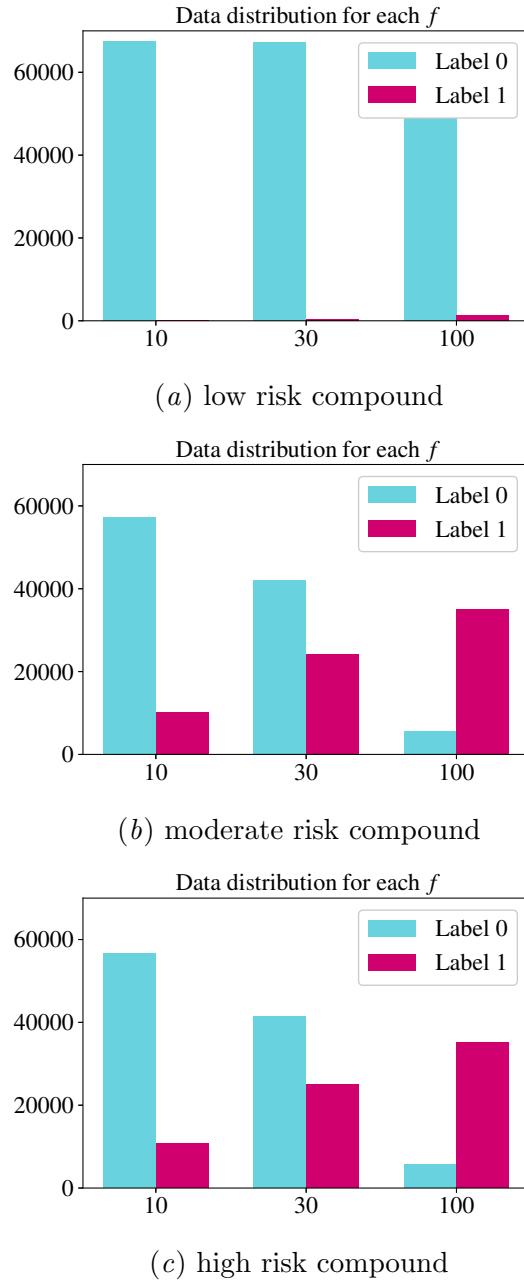


Figure 2: Distribution of binary training data for the three example compounds based on the three  $m_C$  margins

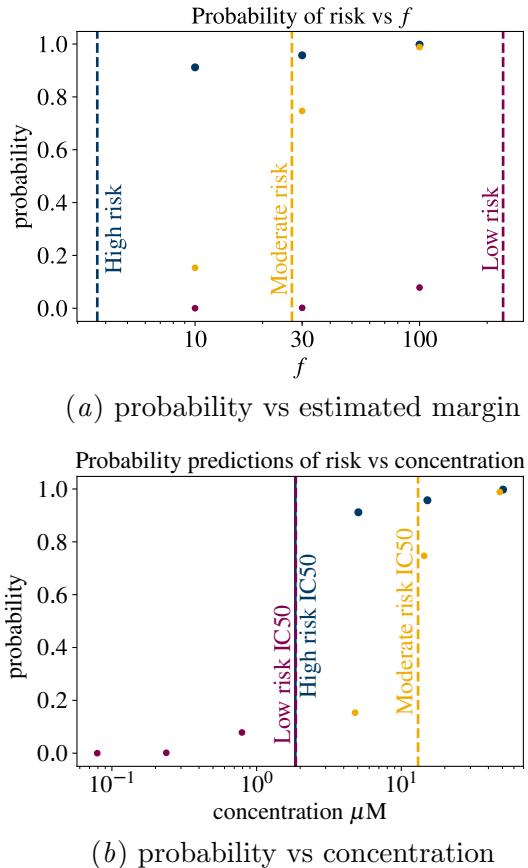


Figure 3: Predicted probabilities for the three example compounds. The dots represent the probabilities at the three margins ( $10, 30$  and  $100 * C_{max}$ ) for the three test compounds. The low, moderate and high risk compounds are colored purple, yellow and blue respectively.

the classes in the predicted data. Figure 5 shows predicted probabilities against  $m_E$  for each  $f$ . The respective  $f$  is highlighted using a dashed line. This shows that the separation is very good for cases that are far from the decision point and that it is still quite good for cases closer to the actual cutoff. With a model like this it could also be of interest to rerun some compounds in the experimental assay that are close to the cutoff and that are likely to be incorrectly predicted. Figure 6 shows the cumulative distributions for the Venn-Abers probabilities and for the log loss probability  $p$ . The obtained probabilities are tight for all test cases indicating low uncertainty. In comparison, the dotted line shows the experimental results, tested, for the same examples.

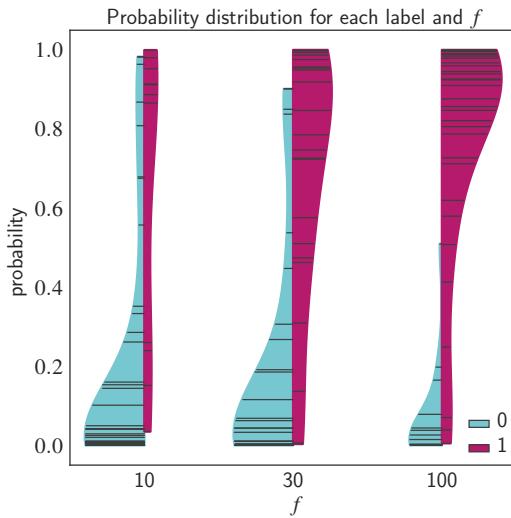


Figure 4: Distribution of predicted probabilities,  $p$ , for each  $f$  grouped by experimental labels 0,1

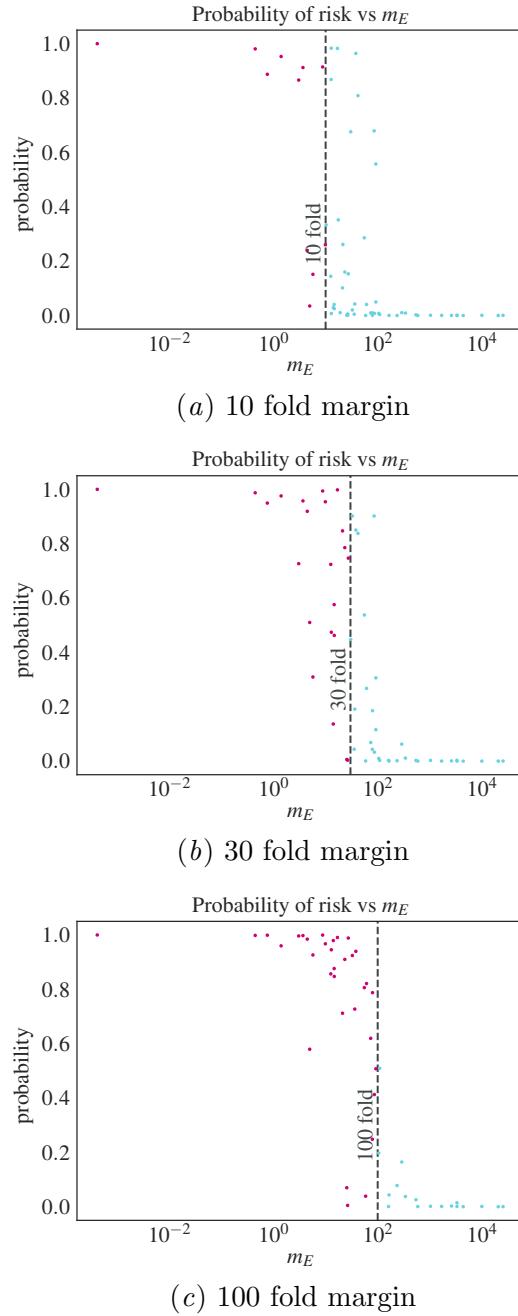


Figure 5: Log-loss weighted Venn-Abers probabilities vs  $m_E$  with  $f$  highlighted by a dashed line

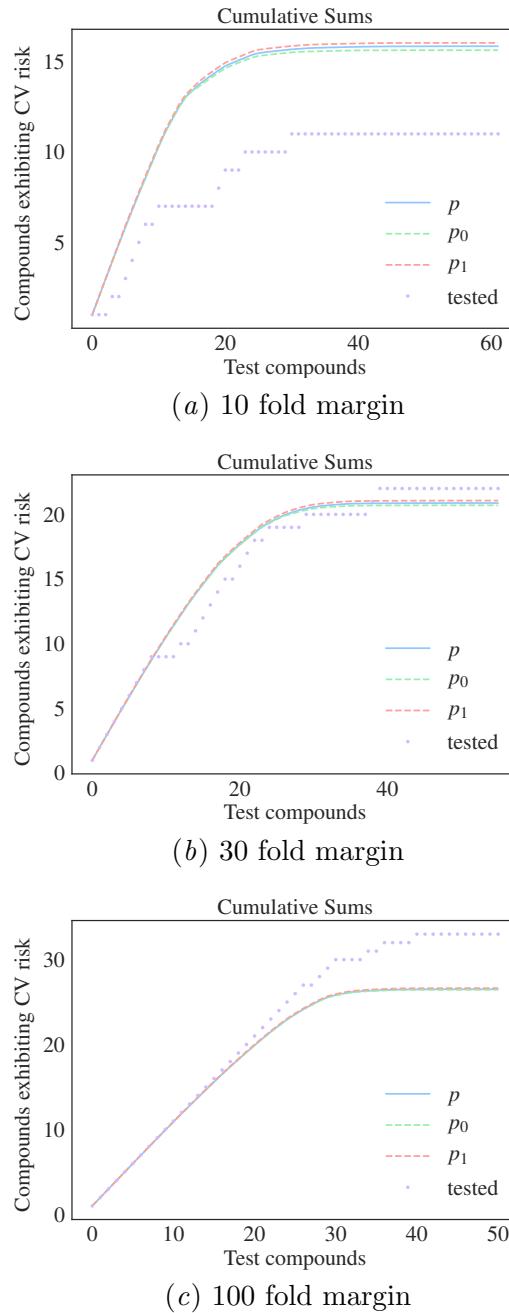


Figure 6: Cumulative  $p$ ,  $p_0$  and  $p_1$  together with the activity distribution for the experimentally tested compounds. For the hERG endpoint

### 3.3. NaV1.5

Summary plots for the test data on the NaV1.5 target. Figure 7 is a representation of predicted risk for each  $m_C$ , showing that there is a separation between the classes in the predicted data. Figure 8 shows predicted probabilities against  $m_E$  for each  $f$ . The respective  $f$  is highlighted using a dashed line. This shows that the separation for NaV1.5 is less clear than for hERG. In comparison, the dotted line shows the experimental results, tested, for the same examples. Figure 9 shows the cumulative distributions for the Venn-Abers probabilities and for the log loss weighted probability  $p$ .

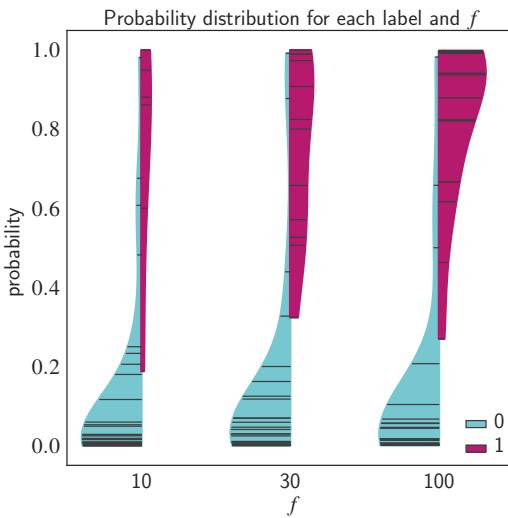


Figure 7: Distribution of predicted probabilities,  $p$ , for each margin grouped by experimental labels 0,1

## 4. Concluding remarks

This study shows that Venn-Abers in combination with SVM can be used effectively to predict cardiac risk based on hERG and NaV1.5 data and can be an effective way to triage compounds. Today,  $C_{max}$  is used together with a predicted  $C_{IC50}$  to calculate the predicted margin which requires a regression model for predicting  $C_{IC50}$ . By framing the question as a binary problem not only can we assign valid probabilities to the risk but it also allows us to use all generated data, including data outside assay experimental limits. The result is a more intuitive way to assess risk as that is the predicted endpoint, rather than predicting a  $C_{IC50}$  which is subsequently converted into a margin which in turn describes the risk. To fully assess the potential of the method larger test sets are needed as well as predictions of compound  $C_{max}$ . In early projects  $C_{max}$  is not known but can be estimated. In future work, we will use these probabilities and produce a measure of over all risk across multiple targets incorporating mechanistic knowledge to further enhance the value to projects. In

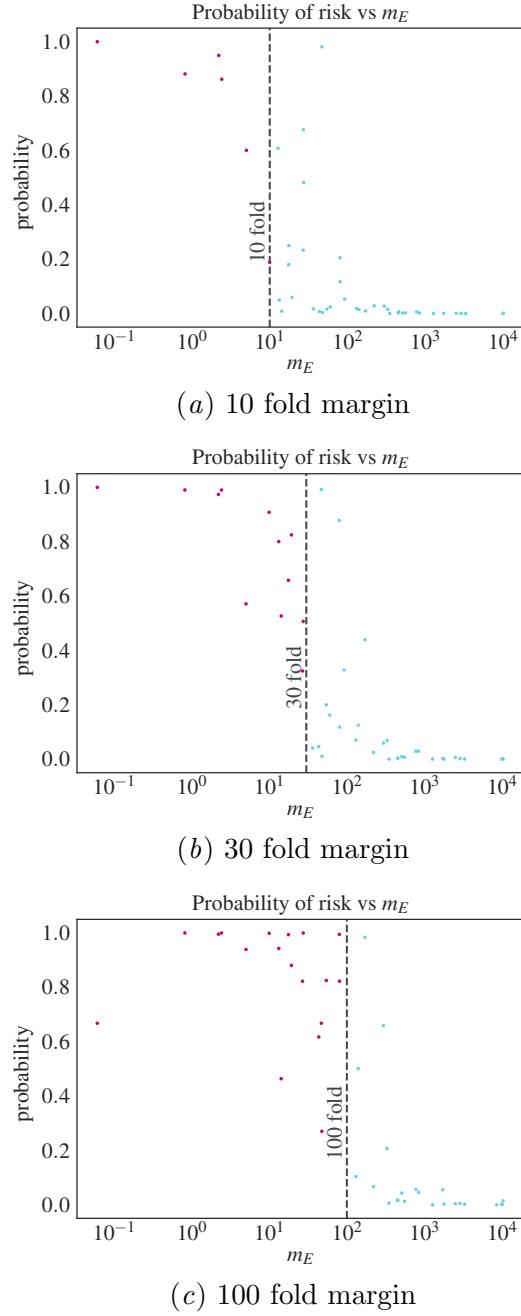


Figure 8: Log-loss weighted Venn-Abers probabilities vs  $m_E$  with  $f$  highlighted by a dashed line

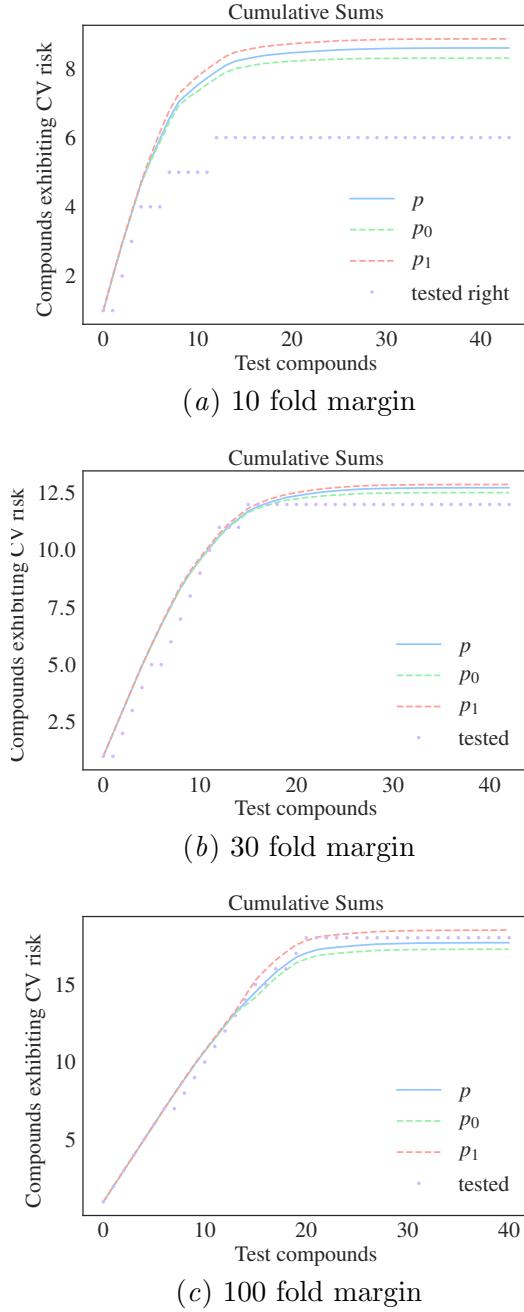


Figure 9: Cumulative  $p$ ,  $p_0$  and  $p_1$  together with the activity distribution for the experimentally tested compounds. For the NaV1.5 endpoint

this study human  $C_{max}$  data has been used to define the concentrations required to obtain specific margins. In early drug discovery human  $C_{max}$  data is generally not available. For future use and effective application in discovery projects  $C_{max}$  needs to be estimated from in vitro and in vivo data. For early screens, data on primary target interaction can be used together with data on ADME properties. The method has a huge potential as it simplifies the decision process and that the result can be more reliable since more data is available for modeling compared to the regression case where data needs to have defined  $C_{IC50}$  values.

## Acknowledgments

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

We are also grateful for the support given by the ExCAPE project funded by the European Unions Horizon 2020 Research and Innovation programme under Grant Agreement no. 671555.

## 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.

Peter Ballard, Patrick Brassil, Khanh H. Bui, Hugues Dolgos, Carl Petersson, Anders Tunek, and Peter J. H. Webborn. The right compound in the right assay at the right time: an integrated discovery dmpk strategy. *Drug Metabolism Reviews*, 44:224–252, 2012. doi: 10.3109/03602532.2012.691099.

M. E. Bunnage. Getting pharmaceutical r&d back on target. *Nat Chem Biol*, 7(6):335–9, 2011. ISSN 1552-4469 (Electronic) 1552-4450 (Linking). doi: 10.1038/nchembio.581.

Jean-Loup Faulon and Carla J Churchwell. Signature Molecular Descriptor. 2. Enumerating Molecules from Their Extended Valence Sequences. *J. Chem. Inf. Comput. Sci.*, 43:721–734, 2003.

Jean-Loup Faulon, Donald P Jr Visco, and Ramdas S Pophale. Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.*, 43:707–720, 2003.

A. Gautam and X. Pan. The changing model of big pharma: impact of key trends. *Drug Discov Today*, 21(3):379–84, 2016. doi: 10.1016/j.drudis.2015.10.002.

Christoph Steinbeck, Yongquan Han, Stefan Kuhn, Oliver Horlacher, Edgar Luttmann, and Egon Willighagen. The chemistry development kit (cdk) an open-source java library for chemo- and bioinformatics. *J. Chem. Inf. Comput. Sci.*, 43(2):493–500, 2003. doi: 10.1021/ci025584y. PMID: 12653513.

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-Verlag New York, Inc., Secaucus, NJ, USA, 2005. ISBN 0387001522.