A comparison of scaling methods to obtain calibrated probabilities of activity for ligand-target predictions

Lewis H. Mervin[1], Avid M. Afzal[2], Ola Engkvist[3] and Andreas Bender[4§]

[1] Hit Discovery, Discovery Sciences, R&D, AstraZeneca, Cambridge, UK

[2] Data Sciences & Quantitative Biology, Discovery Sciences, R&D, AstraZeneca, Cambridge, UK

[3] Hit Discovery, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Sweden

[4] Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, UK

[§]Corresponding author(s)

Email addresses:

AB: ab454@cam.ac.uk

Abstract

In the context of bioactivity prediction, the question of how to calibrate a score produced by a machine learning method into either a relative or an absolute probability of binding to a protein target is not yet satisfactorily addressed. In this study, we compared the performance of three such methods, namely Platt Scaling (PS), Isotonic Regression (IR) and Venn-ABERS Predictors (VA) in calibrating prediction scores obtained from ligand-target prediction comprising the Naïve Bayes (NB), Support Vector Machines (SVMs) and Random Forest (RF) algorithms. Calibration quality was assessed on bioactivity data available at AstraZeneca for 40 million data points (compound-target pairs) across 2,112 targets and performance was assessed using Stratified Shuffle Split (SSS) and Leave 20% of Scaffolds Out (L20SO) validation. VA achieved the best calibration performances across all machine learning algorithms and cross validation methods tested, with the lowest (best) Brier score loss (representing the mean squared difference between an assigned probability and true label, weighted by fraction of predictions in that category). VA achieved the lowest Brier score loss for the RF with 0.023±0.013 and 0.028±0.019 during the SSS and L20SO test set, respectively, indicating high quality calibration (the probability of the true label lies within ~0.15 of the ideal values, across the probability bins evaluated) compared to the base estimator Brier score loss of 0.024±0.014 and 0.03±0.022. In comparison, the PS and IR methods can actually degrade the assigned probability estimates, particularly for the RF with scores of 0.048±0.025 and 0.048±0.026 for SSS, respectively, and 0.047±0.027 and 0.048±0.028 during L20SO. VA was also able to successfully calibrate the probability estimates for even small calibration sets, performing with a mean Brier score loss of 0.008±0.006 for targets with between 50 and 100 active training instances (i.e. estimates lie within ~0.02 of ideal values). VA is able to generate multi-probability values (lower and upper probability boundary intervals), which were shown to produce large discordance for test set molecules that are neither very similar nor very dissimilar to the active training set, and which hence were difficult to predict, suggesting that multi-probability discordance can be used as an estimate for target prediction uncertainty. Overall, we were in this work able to show that VA scaling of target prediction models is able to improve probability estimates in all testing instances, which is currently being applied for in-house target prediction models.

Introduction

In silico target prediction aims to annotate orphan compounds with their relative probabilities of binding protein targets of interest, usually at one or more activity thresholds[1-4]. Such methods are ideally designed to direct biological resources for subsequent experimental confirmation[5-7], given they are still only computational predictions and usually do not predict quantitative activity. Algorithms which utilize negative bioactivity space are capable of also predicting the relative likelihoods for input compounds to be inactive, which has been shown to statistically improve the quality of predictions during internal and external validations[8, 9].

Alternative approaches, which also *quantitatively* model ligand-protein activity (and which are hence more similar to multi-target QSAR methods) have recently been published[10-12], which, for a subset of targets and chemical space covered by the data, are able to approach assay reproducibility limits. Likelihood of activity and quantitative activity are two distinct axes of modelling bioactivity of compounds though, and generally a trade-off between both aspects exists (i.e., quantitative activity prediction, chemical coverage, and model performance cannot be optimized independently of each other at the same time). Hence, both quantitative activity predictions and target predictions, which aim to anticipate the likelihood of activity at a given threshold, have their value and distinct strengths and weaknesses.

An important consideration for *in silico* target prediction methods providing a likelihood of activity is not only their statistical quality (such as performance on an independent test set), but also the accuracy of the algorithmic output to assign a probability estimate that reflects the ground probability for each individual prediction[13, 14]. For example, a raw probability score of 0.90 may represent a high estimate that a compound is active;

however only 70% of the compounds with that corresponding score may be active in reality. Thus, the ground truth likelihood of a compound obtaining this score is lower than the relative output from the algorithm, and hence the model is poorly calibrated[15, 16]. In comparison, predictions from a perfectly calibrated (binary) single-label classifier would classify the samples such that, among the samples to which it gave a probability value of 0.90 (to belong to the active class), indeed 90% actually belong to the active class. While theoretically some algorithms aim to provide realistic probabilities, we will now discuss why this is often not the case in practice.

Different reasons can lead to 'raw' output probabilities of classification algorithms to deviate from true probabilities. On one hand, they might not be functioning even for a *single* model (poor performance); but also beyond that, chemogenomic data cannot be considered identically, independently distributed (i.i.d)[17], since compound-target associations are based on very different data distributions (i.e. different target classes, number of compounds, with higher or lower diversity)[18]. For example, published models using in-house data comprised a very different number of active compounds across the proteins modelled, with a median of 752 and (~6.5 times larger) standard deviation of 4,954 compounds per protein target model[19]. Another study modelled 15 different protein families, with a range of 17 to 615 targets per family, and (comparatively large) standard deviation of 174 targets across families[20]. Enzymes and kinases dominate protein family distributions (~30% and ~34% of training data set size[21]), which also negatively influences the probability estimates for the minority family classes (models with fewer data points lead to fewer positive predictions in most approaches). Model behavior is also influenced by the origin of inactive training sets (putative or experimental annotations)[22, 23], proportion of targets corresponding to protein complexes (~6% of ChEMBL data[24]) and degree of imbalance toward the inactive (majority) class[25-27], i.e. a median active:inactive compound ratio of 96 and (comparatively large) ~16 standard deviation[9]. The chemical space of training data is

also biased, with often 10−30 compound exemplars per scaffold[28], but also in some cases very highly populated scaffolds and high numbers of singleton scaffolds being present[29]. Bioactivity data for orthologue targets are also biased to distinct regions of the chemical space, outlining that probabilities generated for targets in related organisms often comprise separate distributions[30]. Hence, overall, bioactivity data is heavily biased, both with respect to the number, diversity, and particular distribution of data points in a given class, which often does not allow models to arrive at realistic probability estimates. Calibration is therefore an important step to account for the aforementioned biases, to enable the cross-model comparisons (i.e. poly-pharmacological assessment) of probability estimates across the proteome.

Algorithmic behavior is a further factor which influences both the output probability range and distribution of the 'raw' probability estimates generated. For example, Support Vector Machines (SVMs) provide no *direct* support for probability estimates associated with every output prediction, and consequently require additional work to convert the decision function into interpretable probability estimates[31, 32]. Naïve Bayes generates posterior probabilities populating extreme regions of the probability scale (very high or low values) due to repeated multiplications over conditional feature probabilities[33, 34]. Conversely, Random Forests bias predictions toward the midpoint when the predicted fraction of classes across the underlying trees are employed as probability estimates, and extreme values can only be achieved when an exceptionally high proportion of trees predicts either label[35, 36]. Deep neural networks are also more often poorly calibrated compared to a decade ago, due to overfitting from increased application of depth, width, weight decay, and application of Batch Normalization techniques, which manifest in probabilistic error rather than classification error[37, 38].

The posterior probabilities from the aforementioned methods are hence often poor estimates of the actual likelihood of a positive bioactivity prediction if used directly in

this way[39, 40]. Despite this, the assessment of calibration performance receives little attention in the field[41]. The current study will hence explore the application of existing probability calibration methods to the area of *in silico* target prediction, which aim to better calibrate the raw estimates of the actual probability of a positive bioactivity prediction[42]. This topic should not be confused with applicability domain (AD) estimation, which instead aims to identify when the assumptions imposed by a model are fulfilled, presenting sufficient evidence to make a prediction at all[43-45]. For example, approaches such as Conformal Prediction (CP) generate prediction intervals that are guaranteed to be valid in accordance to a user-set confidence level (a confidence level of 0.8 means that the conformal predictor will commit, at most, 20% errors)[46, 47]. This however needs to be contrasted with probability calibration or probability '*scaling*' (which is the topic of the current work), which aims to address the question of obtaining accurate likelihoods of predictions, based on previous observations given in a data set.

In this study, we explore three different scaling methods as an approach to improve algorithm output, namely a parametric approach based on Platt's Sigmoid Scaling[32], and two non-parametric approaches, namely Isotonic Regression Scaling[48] and Venn-ABERS Predictors[49]. The advantages, disadvantages and previous applications of these methods are summarized in **Table 1** and will be discussed in more detail in the following.

Platt (Sigmoid) Scaling (PS) is a method employed to calibrate the probabilistic output of a base estimator to better reflect a confidence in a prediction, and it was initially developed to generate interpretable outputs from SVMs,[32] but it is also applicable to the output of other machine learning methods. This procedure uses a cross validation split of the dataset and, for each split, employs the base estimator trained on the training samples to generate calibration predictions for the test set, as shown in **Figure 1**. A sigmoidal curve is then fitted to the distribution of the resulting base predictor

probabilities and the corresponding ground truth likelihood (i.e. the actual observed fraction of actives at a given base estimator probability bin). For each new input, query probabilities are then obtained and averaged across all folds to give an output probability estimate[50]. PS has been previously employed (and reported to provide valid estimates) for cytotoxicity prediction[51], and to enable the successful comparison of probability estimates for acute toxicity prediction that were generated from a range of different machine learning algorithms[52].

Isotonic Regression Scaling (IR) uses the same cross-validation approach and interpolation to calibration as PS; however it employs a non-parametric approach based on IR when fitting the curve to the calibration data[48]. Overall, IR is preferable for non-sigmoid calibration curves and in situations where large amounts of data are available[53], whilst PS has been shown to perform better in cases where there is limited calibration data (smaller numbers of ligand-target data points), where IR tends to overfit[16, 54]. Datasets with limited number of data points are relatively frequent in target prediction, as outlined above, and hence our initial expectation was that this trend would also be observed in the current study.

The final type of calibration method explored in this study are Venn–ABERS (VA) predictors, which are based on the idea of IR, though they apply this technique within the framework of Venn prediction[49], which are a non-parametric approach related to the Conformal Prediction (CP) framework[48, 55, 56]. While CP methods produce valid *region predictions* and predicted labels (i.e. a region that contains the true target prediction with a pre-defined probability), Venn predictors produce valid *probabilistic predictions*[57]. In this scenario, the IR is for every data point to two series of bioactivity labels, assuming either of the activity labels (i.e. that a compound is either active or inactive) of the new prediction object. This results in the production of *two* prediction values, $p0$ and $p1$, respectively (i.e. that a compound is truly active or inactive at a given threshold,

respectively) with the theoretical advantage of validity guarantees from Venn prediction[58]. VA has been successfully combined with conformal prediction to improve *p*-value interpretability[57], prediction of metabolic (site-of-metabolism) transformations[59], protein target prediction[58, 60], iterative screening[61] and cardio-vascular risk assessment[62].

Since *in silico* target prediction approaches assume the data are independent and identically distributed (i.i.d)[63, 64], and given the previously discussed advantages and disadvantages of the methods explored, we expect that VA will generally produce superior calibration results. PS, based on both theoretical considerations and previous studies, is only expected to perform better than IR for smaller target classes with limited calibration data.

In this study, we evaluated which method provides practically meaningful likelihoods of activity by exploring the impact of PS, IR and VA scaling on protein target prediction with Bernoulli Naïve Bayes (BNB), SVM and RF algorithms by combining the bioactivity data available in AstraZeneca ChemConnect[65] with additional inactive compounds from PubChem[66, 67]. Performance was assessed using Stratified Shuffle Split (SSS) and Leave 20% of Scaffolds Out (L20SO) validation, which emulates a scenario when protein target prediction models are tasked with extrapolating bioactivity predictions to novel chemical space, which is what we would assume to be a realistic scenario in the drug discovery context.

Methods

Sources of bioactivity data

Active compound set from AstraZeneca Chemistry Connect

The AstraZeneca Chemistry Connect[65] repository, comprising both in-house data and public repositories such as ChEMBL[68], was filtered for activity values (IC_{50}/EC_{50}/K_i/K_d) better than or equal to 10 μM from 'binding' or 'functional' human protein assays, as defined by Entrez Gene ID metadata. The 10 μM cut-off for activity specified here is in accordance with previously validated target elucidation methods[9, 21], and assigns both marginally and highly active compounds to targets. Compounds were filtered for targets with greater or equal to 50 active compounds to ensure proteins encompassing sufficient chemical space are retained for training. The resulting dataset includes 3,381,388 distinct compounds for 8,485,161 bioactivities spanning 2,112 targets. The targets modelled comprise a variety of target classifications; the three most populated target classes include 449 kinases, 222 GPCRs and 192 ion channels (see **Supplementary Table 1**).

Inactive compound set from AstraZeneca HTS screens

HTS bioactivity data from 420 in-house target-based screens (using Entrez Gene ID) spanning 400 targets was employed as a resource of inactive bioactivity data from AstraZeneca sources. These screens were filtered for activity values (K_i/K_d) worse than 10 μM. A compound was defined inactive if it was measured at least twice as many times as inactive versus as active in cases of conflicting annotations. Inactive data has coverage for a wide variety of targets, including 88 different GPCRs, 77 kinases and 31 proteases (see **Supplementary Table 2** for details). The resulting compound-target pairs resulted in a data set of 189,965,064 inactive data points, comprising 2,827,651 distinct compounds for 400 targets from in-house sources.

Inactive compound set from PubChem

In order to also compile experimental inactive data points for proteins not covered in the internal AZ databases, the PubChem BioAssay[69] database was also used for additional experimentally confirmed inactive data points, as in previous work[9], *via* the EUtils and PubChem Power User Gateway (PUG) REST APIs[69]. This process involved the '*ESearch'* and '*ELink'* EUtils procedures to obtain a comprehensive list of all Entrez Gene ID's (GIDs) and Protein ID's (PIDs) associated to a given GID. These GIDs and PIDs were used in '*ELink'* to identify binding and functional assays held in the NCBI BioAssay database. A subsequent '*ELink'* step was used to link from these assays to Compound IDs (CIDs) with a compound-target '*activity_outcome'* annotation that has been declared as 'inactive' by the contributors of the screening data. Finally, inactive CIDs were mapped to SMILES using the PubChem PUG REST service. The active set of target-compound pairs were retained when conflicting inactive PubChem bioactivities arose, since active data has been calculated from dose-response curves and hence deemed to be more reliable. This workflow resulted in 419,121,152 inactive data points for 768,014 distinct compounds, spanning 2,116 targets (see **Supplementary Table 2** for details)

Sphere exclusion of putative negative bioactivity data and undersampling

The in-house and PubChem inactive data sets were combined, yielding 598,923,798 inactive data points spanning 2,161 targets. A Sphere Exclusion (SE) algorithm was applied to 1,500 targets with insufficient numbers of inactive data points (to achieve a ratio of 7:1 inactives to actives per target, or at least a minimum of 5,000 inactive data points) for both public and proprietary data, as in previous work[30]. In this procedure, compounds were randomly sampled from PubChem with a Tanimoto coefficient (Tc)

fingerprint (as outlined in the “Compound pre-processing and RDKit fingerprint generation” section) similarity to actives lower than 0.4 to obtain the desired number of compounds which could reasonably be assumed to be inactive against a given target. 16,188,048 additional putative inactive compounds were sampled in this manner (see **Supplementary Table 2** for details). Although model performance is not directly comparable if some models use experimental inactive compounds and some use putative inactive compounds from sphere exclusion (which artificially samples points further away from the actives and thereby inflates model performance), this sampling step is essential from the practical side (since data points for both active and inactive class are needed) and in order to have a large applicability domain in chemical space (there is benefit in having a reasonably large set of (putatively) inactive compounds available for training[26]). Conversely, 1,003 target models over the 5,000 absolute compound threshold required *under*-sampling of the inactive train set to achieve a 7:1 maximum ratio of inactive to active molecules, as in previous work[30]. In this procedure compounds were randomly removed from the inactive set to achieve the desired ratio. The putative inactive compounds were combined with the sub-sampled inactive bioactivity datasets, producing a final dataset of 38,902,310 inactive labelled compound annotations.

Supplementary Table 1 summarizes the sources and number of data points contributed by each data source and the large variance between the amount of bioactivity data available per family, with a median of 326,895 active compounds *per family* and standard deviation of 743,122. The table also outlines that different target families require putative sampling to obtain the 7:1 ratio to different extents (i.e. Nuclear Hormone Receptors (NHRs) had 822,412 inactive compounds added *via* sphere exclusion, versus Oxidoreducatases which do not require any additional sampling). Target classes also comprise different ratios of Murcko scaffolds to the size of the compound set (and hence different chemical diversity), such as GPCRs with a

compound:scaffold ratio of 3.3, versus Phosphatases with a ratio of 2.0. Overall this analysis outlines how the bioactivity data used for this work is highly imbalanced and biased, and shares the problems described in the introduction section.

Compound pre-processing and RDKit fingerprint generation

Compound structures were standardized using an in-house script[70] to remove salts, normalize charges, tautomerize (to the most favorable form) and to remove duplicates. RDKit[71] (Version 2019.03.4) was employed to remove structures without carbon, and to retain only compounds with atomic numbers between 21–32, 36–52, and greater than 53, and with a molecular weight between 100 and 1000 Da, to retain a small organic molecule-like chemical space. RDKit was used to generate 2,048-bit circular RDKit fingerprints[72], with the radius set to 2.

Outer and Inner Cross Validation Strategies

Five-fold Stratified Shuffle Split (SSS) cross validation was employed as a first outer split (see **Figure 1**) using the function '*StratifiedShuffleSplit*' in Scikit-Learn[73] with '*n_splits*' set to 5 and the '*train_size*' and '***test_size'*** set to 0.2 and 0.8, respectively. Leave 20% of Scaffolds Out (L20SO) cross validation was also employed (see **Figure 1**) as a more challenging outer split strategy to explore scaling method performance when the i.i.d assumption is not valid. L20SO was performed using the *'GroupShuffleSplit'* function with '*test_size'* of 0.2 and '***n_splits'*** set to 3, whilst suppling the *'groups'* function of the '*split*' method with the Murcko skeletons of training molecules using '*GetScaffoldForMol'* followed by '*MakeScaffoldGeneric*' in RDKit.

Both splits above served as an *outer* fold to benchmark the scaling methods. **Figure 1** outlines how PS, IR and VA are then applied to the base classifiers for each SSS or L20SO split. That is, the outer train split (used for benchmarking) is subsequently split

three-fold to produce an *inner* split (for scaling). The default probabilities generated by the unscaled classifiers are recorded for comparison against the absolute or scaled probabilities. The calibration (reliability) curves were generated in Scikit-Learn using the class '*calibration.calibration_curve*' with the number of bins set to 10. The Brier score loss, a metric to assess how well predictions are calibrated, was generated in Scikit-Learn using the function '*metrics.brier_score_loss'.* This score measures the ability of the model to distinguish between the classes across threshold bins through the mean squared difference between the predicted probability assigned to the classification items, and the actual outcome[74, 75]. Therefore, the lower the Brier score loss for a set of bioactivity predictions, the better the predictions are calibrated, where the best possible score of 0.0 represents that probability estimates are perfectly accurate and the lowest possible score of 1.0 outlines that the estimates are wholly inaccurate.

Platt Scaling and Isotonic Regression Scaling using Scikit-learn

Platt Scaling (PS) and Isotonic Regression Scaling (IR) were performed using the Scikit-Learn class '*CalibratedClassifierCV*' with the '*StratifiedShuffleSplit*' function used to split the (inner) folds, with ('*n_folds'*) set to 3 whilst supplying the '*sigmoid*' and '*isotonic'* method parameters, respectively. This function performs a three-fold cross validated calibration methodology on the training data and is kept constant in every run. The sigmoid or logistic function is fit to each fold, based on the generated probabilities from the base classifier trained on the train split and the true positive compound predictions from the test split. The protocol averages the interpolation for a given input compound based on the interpolated true positive rate between the sigmoid curves amongst all folds.

Venn-ABERS Predictors

Venn-ABERS Predictors (VA) were trained using the Scikit-Learn class '*StratifiedShuffleSplit*' (as above), with the number of (inner) folds ('*n_folds*') set to 3. For each split, VA Scaling was performed as described in Vovk *et al.,*[55] and implemented at https://github.com/ptocca/VennABERS using the training and test split of compounds and a base classifier. The output of this algorithm are two probabilities (*p0* and *p1),* which can be interpreted as upper and lower boundary of probability estimates for an individual active or inactive prediction. While there is a value in having multi-probability predictions for individual classes, it is required for compatibility to compare a single point probability prediction for the purpose of comparing VA with PS or IR. We constructed a single point probability prediction (P_i), for compound i, from a multi-probability prediction that minimizes the 'regret' under a given loss function, as described in Toccaceli *et al.*[58]:

$$P_i = \frac{p1}{1 - po + p1}$$

Equation 1

Finally, the single point probabilities were averaged across all splits to produce a final VA probability, which was then comparable with the PS and IR methods.

Target prediction methodology

Bernoulli Naïve Bayes

The Bernoulli Naïve Bayes (BNB) algorithms were trained using Scikit-Learn[73] with the *'alpha'* values of 1.0 (selected from hyper-parameter optimization). The Bernoulli algorithm explicitly penalizes the non-occurrence of a feature indicative of protein target activity (i.e. negative evidence within the fingerprints when a 0-bit is interpreted as the absence of a fingerprint feature in a molecule). This base classifier was trained using the binary matrix of the active and inactive compound-target fingerprints on a per-target

basis. In this procedure, each model is trained for a single target using the active and inactive compounds annotated for that target. The output is a raw posterior probability from the *'predict_proba'* function.

Support Vector Machine

Linear Support Vector Machines (SVMs) were trained in Scikit-Learn with the kernel set to linear (due to size and number of models and hence time to train) and with a *'C'* penalty parameter of 1.0 and '*class_weight*' set to '*balanced*' (obtained from hyper-parameter optimization). The raw output from the '*decision_function'* in Scikit-Learn was normalised between 0 and 1 using the MinMax scaling algorithm in *Equation 2*, where $P(C_1, \ldots., C_n)$ is the probability vector output from the SVM for each compound C, and P_i is a single point probability per-compound input.

$$P'(C_1, \ldots, C_n) = \sum_{i=1}^{n} \frac{P_i - \min(P(C_1, \ldots, C_n))}{\max(P(C_1, \ldots, C_n)) - \min(P(C_1, \ldots, C_n)}$$

Equation 2

Random Forest

The Random Forest (RF) classifier was deployed using 100 for the number of Trees in Scikit-Learn, with the number of features and maximum depth set to *'auto'* and the *'class_weight'* set to '*balanced'*. This base classifier was trained whilst providing the '*fit'* method the '*sample_weights*' of the ratio of active *versus* inactive compounds. The raw probability output from "*predict_proba*" is defined as the mean predicted fraction of class samples in a leaf across the Trees[73].

Results & Discussion

Calibration results from five-fold cross validation

We first investigated the desired behavior of each scaling methodology and contrasted this with the behavior of a perfectly calibrated classifier, when positive prediction scores generated by an algorithm perfectly encapsulate the number of positive instances obtaining that score. This is done by calculating the fraction of true active data points (ground likelihood) retrieved as a function of probability estimate, also known as a calibration (reliability) plot.

Results from the calibration plot for Stratified Shuffle Split (SSS) are shown in **Figure 2** with the overall Brier score loss for each line outlined in **Table 2**. Our findings show that VA produces the calibration points closest to perfect calibration (dashed line), and hence performs with the lowest (best) overall Brier score loss of 0.050, 0.043 and 0.033 for the BNB, SVM and RF, respectively. In context, a RF Brier score of 0.033 represents the average mean *squared* error of the true label probabilities across probability estimates (i.e. $(0.825\text{-}1)^2$ in this case). This means the scaled probability outputs are on average within 0.175 of the ideal values across the estimates generated and probability bins evaluated, which can be contrasted with the BNB estimates that are within 0.245 of the ideal values.

The predictions from VA can be considered *underconfident* for the SVM and RF algorithms within higher probability bins above 0.5, since the fraction of active compounds is higher than the forecast probabilities (i.e. perfect probabilities would be higher). In contrast, the PS and IR methods represent *overconfident* predictors, where the highest fraction of actives are disproportionately distributed within the larger probability estimate bins. The latter two methods hence obtain higher (worst) overall Brier loss scores, where PS obtained values of 0.072, 0.056 and 0.046 for the BNB,

SVM and RF methods, whilst IR obtained values of 0.067, 0.053 and 0.047, respectively. Hence, we can conclude that VA produces the best calibrated estimates for compound activity, despite the fact that this method still produces under-confident predictions in some cases.

We further explored the *per-target* Brier score loss for the scaling approaches across the 2,112 protein target models, since due to different distributions and numbers of data points per model we expected to see differences in behavior between them. Results from this analysis are shown in **Table 3** and **Figure 3**, and show that VA performs with the lowest mean Brier score loss (shown in bold) of 0.029±0.017, 0.025±0.012 and 0.023±0.013 (i.e. average of Brier score loss for 2,112 targets ± the standard deviation), across the range of BNB, SVM and RF models, respectively. There is hence little difference between VA performance across algorithms, since RF Brier score loss shows that scaled probability outputs are within 0.17 on average of the ideal probability estimates whilst BNB estimates are also within 0.15 of the ideal values.

Overall VA improved the baseline BNB, SVM and RF Brier scores by a margin of 0.01±0.02, 0.108±0.024 and 0.001±0.001 (± standard deviation across 2,112 targets) respectively, which in context means that output probability estimates are now 0.100±0.145, 0.320±0.155 and 0.032±0.032 closer to ideal values across the probability scale bins. In comparison, the PS and IR perform with inferior Brier score loss, with *degraded* performance compared to the base estimates of BNB and RF, by a Brier score loss margin of 0.018 and 0.013, and 0.024 and 0.024, respectively. Hence, we can conclude that VA produces the superior probability estimates on a *per-target* basis, compared to all other methods tested in this study.

To better understand why Brier score loss may be degraded by scaling we next analyzed how within-class changes in probabilities are assigned by the scaling methods

applied. To this end, we further explored the mean, median and standard deviation of the per-compound probability estimates assigned to all compounds across all protein target models after calibration in **Table 4**. If probabilities are inflated for *inactive* compound predictions, this would manifest in a high proportion of false positive compounds within higher probability estimate bins. The table numerically corroborates the trend of PS and IR overconfidence, since inactive compound probability estimates (which should be as low as possible) are increased from the BNB default of 0.057±0.204 (i.e. mean of per-compound probability estimates across all targets ± the standard deviation) to 0.175±0.175 and 0.147±0.192, for PS and IR respectively. RF also exhibits inflation for the probability estimates of inactive compounds, with an increase from the base estimator of 0.066±0.106 to 0.117±0.165 and 0.110±0.174 for PS and IR, respectively. Furthermore, the distribution of change as a function of base predictor probability estimate, shown in **Supplementary Figure 1** (Bland-Altman plot included in **Supplementary Figure 2**), outlines how the PS and IR probabilities are particularly inflated for inactive compound predictions in the lower half of the default probability scale (below 0.5). Taken together, these findings outline that both the PS and IR approaches assign enlarged probability estimates to inactive compounds, which hence manifests in a higher false positive rate.

VA alters the probabilities assigned to compounds to a lesser (more conservative) extent, for compounds with initially low 'raw' probability estimates compared to both PS and IR. In comparison, the mean probability estimates for the inactive compounds are actually often decreased (which is favorable behavior in this context), since the calibrated scores of 0.046±0104 are lower than the mean RF base estimates of 0.066±0.106. More specifically, 66.7% of the inactive probability estimates are decreased (see **Supplementary Table 3** for the ∆probability analysis). This trend is further illustrated by the distribution of markers around the VA curves across all machine learning methods in **Supplementary Figure 1**, and (in comparison to PS and

IR) do not often inflate probabilities within in the first half of the probability ranges as shown in the Bland-Altman plot (**Supplementary Figure 2**). Hence VA predictors assign lower probability estimates to inactive compounds, which manifests in a lower false positive rate.

We can hence conclude that VA provides the relatively best Brier score loss across the Stratified Shuffle Split (SSS) tests performed, where probability estimates are better calibrated, and scores more accurately reflect the actual distribution of compounds with those scores across the breadth of models evaluated. We hence expect that VA will also produce superior results when applied to an external distribution of compounds with scaffolds distinct from the training set, whilst PS and IR may confound performance due to the larger probability estimates assigned to input compounds. This is the analysis we have performed in the next step.

Calibration results from leave 20% of scaffolds out validation (L20SO)

The calibration plot obtained from the fraction of active data points retrieved as a function of probability estimate for the Leave 20% of Scaffolds Out (L20SO) validation which is hence a more difficult classification task, resulting in lower positive prediction scores for true positive compounds and a violation of the underlying i.i.d assumption, since the chemical space of test compounds are further from training compounds.

Results from the analysis are shown in **Figure 4**, with the corresponding overall Brier score loss for each method outlined in **Table 2**. The second column of the table sows Brier score loss is higher (worse) compared to the equivalent scores obtained during SSS benchmarking, with scores ranging from 0.074 (for the base BNB classifier) to 0.039 (for the VA method and RF classifier), which is expected when applying models to a distribution of chemical scaffolds distinct from training data.

In a similar manner to SSS, L20SO validation **Figure 4** highlights that VA generates a calibration curve closest to optimal calibration, with the lowest Brier score loss of 0.054, 0.048 and 0.039, for the BNB, SVM and RF, respectively. The output generated by the PS and IR methods is again distinct from VA, with inferior Brier score loss of 0.075, 0.058 and 0.048 and 0.070, 0.055 and 0.049, respectively (for methods in the same order as before). In a similar manner to SSS benchmarking, both PS and IR have calibration lines above the perfect calibration, characteristic of overconfident algorithms. Hence, we conclude that VA produces the best probability estimates of the methods benchmarked here, also when applied to an external dataset of chemistry.

We next investigated the *per-target* performance of the different methods, the results of which are shown in **Table 3**. It can be seen that that VA performs with the relatively best mean Brier score loss on a per-target basis, with the lowest scores of 0.033±0.023 (mean of the Brier score loss across the 2,112 models ± standard deviation), 0.032±0.02 and 0.028±0.019 for the BNB, SVM and RF, respectively. Only 27% of inactive VA predictions showed an increase (∆probability analysis in **Supplementary Table 3**), with mean inactive probability estimates of 0.043±0.101. The base estimate as a function of calibrated probability in **Supplementary Figure 3** and the absolute change in probability estimates after calibration, shown **Supplementary Figure 4**, further outline this trend. Similar to SSS, the figures show the conservative nature of the VA predictors to assign lower probability estimates within the lower range of the probability scale (between 0.0 and 0.5) with a higher proportion of markers below zero.

Results from the per-compound probability estimates, shown in **Table 4**, outline that the PS and IR methods inflate initially low inactive RF compound probability estimates (across all compounds and targets) during L20SO validation, consistent with the SSS result. Conversely, the RF algorithm obtained mean probability estimates of

0.109±0.161 for PS and 0.103±0.171 for IR, compared to the default algorithm estimates of 0.062±0.102. Hence, taken together we have shown that PS and IR methods actually degrade the quality of the probability estimates compared the base estimator, due to the inflation of the probability estimates generated.

We can hence conclude that VA predictors perform with smaller Brier score loss, indicating the best calibration both during SSS and during L20SO, and that VA scaling also performs with optimal Brier score loss for extrapolation to novel chemical scaffolds (violation of the i.i.d assumption). Therefore, VA scaling should be employed when extrapolating predictions to novel chemical space (i.e. new scaffolds and chemical features), since the probability estimates are closer to the perfect calibration and do not exhibit the same trend to assign over-confident probability estimates, compared to PS and IR.

Effect of model size on the performance of the scaling algorithms

Since calibration set size has been shown to influence the performance of the various scaling algorithms (through overfitting or a lack of sufficiently distributed calibration points)[16], we next explored the effect of target training set size as a function of the per-target Brier score loss for the Random Forest (RF) models (the most relatively optimal algorithm from SSS and L2SO validation based on Brier score loss). Target calibration set sizes were split into four bins; active training sizes between 50 and 100, between 100 and 500, between 500 and 1000, and larger than 1000, which are henceforth referred to as bins one, two, three and four, respectively.

Results from this analysis are shown in **Figure 5**, with the mean, standard deviation and median scores provided in **Supplementary Table 4**. Overall, VA achieves the lowest Brier score loss performance throughout all binned class sizes, with scores of 0.008±0.006 (mean of model scores within the bin ± the standard deviation),

0.020±0.015, 0.033±0.028 and 0.037±0.025, for bins one through four, respectively (illustrating that probability estimates are 0.090, 0.144, 0.182 and 0.193 closer to ideal probability estimates). This supports the view that there is hence a positive impact on the predictivity of VA scaling at any training set size. We conclude in particular that VA is applicable for even the smallest datasets, containing 50 to 100 calibration instances.

Converse to the findings in previous literature (which were performed on different datasets[48, 58]) which outlined PS outperforms IR for small calibration sets, our results indicate PS and IR exhibit relatively *similar* performance for models within the smaller training set sizes. For example, the PS and IR approaches achieved a mean Brier score loss 0.045±0.049 and 0.047±0.050 for bin one and 0.051±0.049 and 0.049±0.048 for bin two, respectively. Taken together, these findings support the view that IR overfitting on small amounts of calibration data is no more detrimental to performance than enforcing a PS sigmoidal form to calibration data on small data sets. Overall, PS and IR actually degrade Brier score loss compared to the RF base estimator (with scores of 0.009±0.006, 0.021±0.014 and 0.036±0.026, respectively within the first three bins), and hence we conclude (on the datasets employed in this work, which due to size and diversity we would however assume to be representative of large-scale target prediction tasks) that PS or IR should not be applied to small bioactivity training datasets when using the BNB and RF algorithms.

The choice of scaling method does not produce significantly different Brier score loss performance for the default RF, PS and IR methods for the largest calibration set in bin four (1000+ active compounds), with a similar overall distribution obtained across the three scaling methods. Here, the base estimator, PS, IR and VA methods obtain scores of 0.041±0.030, 0.046±,0.027, 0.045±0.026 and 0.037±0.025, respectively. Although VA hence performs still with the lowest Brier score loss of the methods tested, we conclude that the underlying probability estimates from the base algorithm are overall

already more reliable for larger training set sizes, leading to overall less impact of scaling or the particular method chosen (though VA scaling still performs better by a small margin).

We have shown here that VA generates the most accurate predictions for both smaller and larger sized models (to somewhat different extents), and thus they are better suited to calibrating for the actual probability of activity for either the active or inactive class, regardless of calibration set size.

Uses for the multi-point probabilities from the Venn-ABERS predictors

We next examined how to best use the multi-point probabilities $p0$ and $p1$ produced from the VA predictors during L20SO validation for the RF, beyond combining both values into a single probability value using *Equation 1*. The rationale is that probabilities $p0$ and $p1$ provide information for class membership of the inactive and active class, respectively, which we could hence use to provide an estimate for the expected uncertainty of the resulting model. Our results, depicted in **Figure 6** and available in **Supplementary Table 5,** show the relationship between the $p0$, $p1$ and single point probabilities compared with the similarity to nearest neighbor compounds (based on circular RDKit (ECFP_4) Tanimoto coefficient (Tc) fingerprint similarity between test and training set). The Tc bins 0.0-0.25, 0.25-0.5, 0.5-0.75 and 0.75-1.0 are henceforth referred to as bins one, two, three and four within this section, respectively. Our results show a large discordance (margin) between the $p0$ and $p1$ values toward the center of the Tc similarity scale, with a mean discordance of 0.014±0.031 and 0.011±0.029 (mean compound prediction discordance within that bin across all L20SO predictions ± the standard deviation) for Tc bins two and three, respectively. This region encompasses testing compounds which are neither similar nor dissimilar to the training set, and which hence produce unconfident RF predictions since the RF classifier would assign a probability estimate of 0.5 when neither activity label is certain (i.e. 50% of the

Trees in the Random Forest predict the compound as active and inactive, as outlined in **Table 1**). Conversely, there are smaller margins between the *p0* and *p1* values for the low and high similarity bin one and four with mean discordances of 0.011±0.029 and 0.005±0.016, respectively. This is hence an indicator for more confident test cases, when a compound is either likely to be active due to similarity to an active compound, and *vice-versa*, when an input compound is highly likely to be inactive due to dissimilarity to actives in the training set. Taken together, these findings demonstrate that scaling-derived *p0* and *p1* values are linked to conceptual chemical meaning, i.e. chemical structure similarity to the input data of the models generated. Therefore, we propose that this probability discordance renders itself to predicting the uncertainty of subsequent model predictions.

Summary

This work explores the application of three different scaling techniques, namely Platt's (Sigmoid) Scaling (PS), Isotonic Regression Scaling (IR) and Venn-ABERS Predictors (VA) for scaling prediction scores obtained from ligand-target prediction using the Naïve Bayes (NB), Support Vector Machines (SVMs) and Random Forest (RF) algorithms. Data available at AstraZeneca and PubChem were combined and calibration performance assessed using Stratified Shuffle Split (SSS) and a Leave 20% of Scaffolds Out (L20SO) methodology. Out of the three methods tested, we found that the VA scaling method provided the best probability estimates of during both SSS and L20SO validation, obtaining an overall (aggregated across all compounds and targets) Brier score loss of 0.050, 0.043 and 0.033 for the BNB, SVM and RF during SSS, respectively, and 0.054, 0.048 and 0.039 during L20SO, illustrating that the output probabilities are better reflected in the actual distribution of active compounds. VA were also shown generated better estimates on a *per-target* basis, for both smaller and larger sized models, and thus they are better suited to calibrating for the actual probability for compound activity regardless of the amount of calibration data available. In comparison, the calibration analysis for PS and IR showed class membership estimates were over-confident and further from perfect calibration (and even in many cases degrade calibration performance compared to the base classifier for the BNB and RF). A final analysis into the scaling-derived $p0$ and $p1$ values from VA highlighted how the margin between these values could be used to provide an estimate for the anticipated prediction [un-]certainty of the model for a particular data point. Overall, this analysis hence provides a step towards scaling model output in *in silico* target prediction to arrive at more reliable class probability estimates.

Acknowledgements

LHM thanks the Biotechnology and Biological Sciences Research Council (BBSRC) [BB/K011804/1]; and AstraZeneca, grant number RG75821. All authors thank Paolo Toccaceli for helpful discussions and the Venn-ABERS implementation in Python. All authors thank Marianna Trapotsi for proofreading the manuscript.

Figures

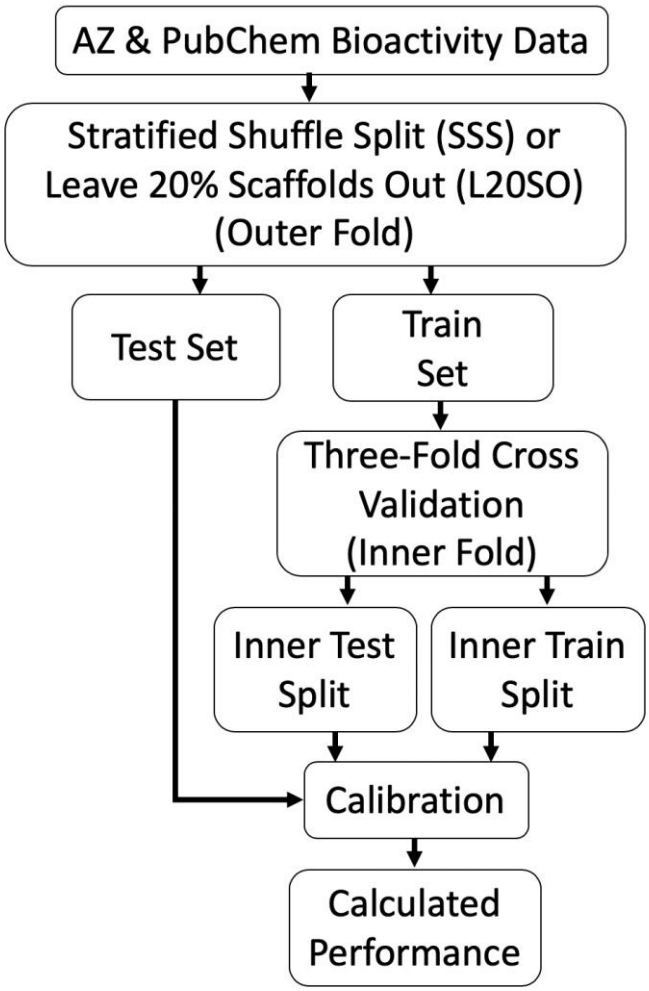


Figure 1. Flow chart of the methods to calibrate models and determine model performance. Stratified Shuffle Split (SSS) and Leave 20% of Scaffolds Out (L20SO) validation were used to split the training data into outer training and test splits. The outer train set was split using three-fold stratified cross validation for calibration. The inner train and test split were used to train and calibrate each model. Models were finally benchmarked using the test set from SSS or L20SO, using the calibrated model, with calculation of the Brier score loss.

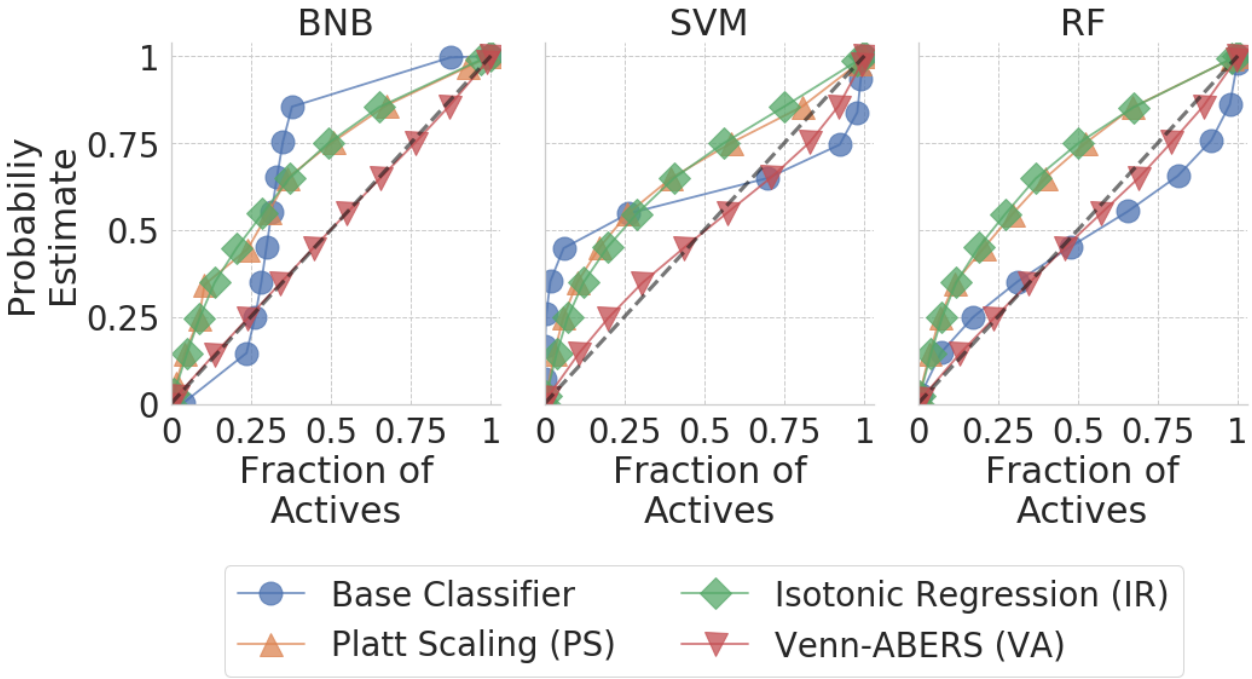


Figure 2. Calibration curves for the Stratified Shuffle Split (SSS) cross validation for Bernoulli Naïve Bayes (BNB), Support Vector Machines (SVMs) and Random Forests (RF). Venn-ABERS (VA) (red) achieves a reliability curve closest to a perfectly calibrated model across all classifiers tested, where the calibrated probabilities are better reflected in the true distribution of active compounds. In comparison, Platt scaling (PS) (yellow) and Isotonic Regression (IR) (green) methods produce overconfident calibration lines (above perfect calibration).

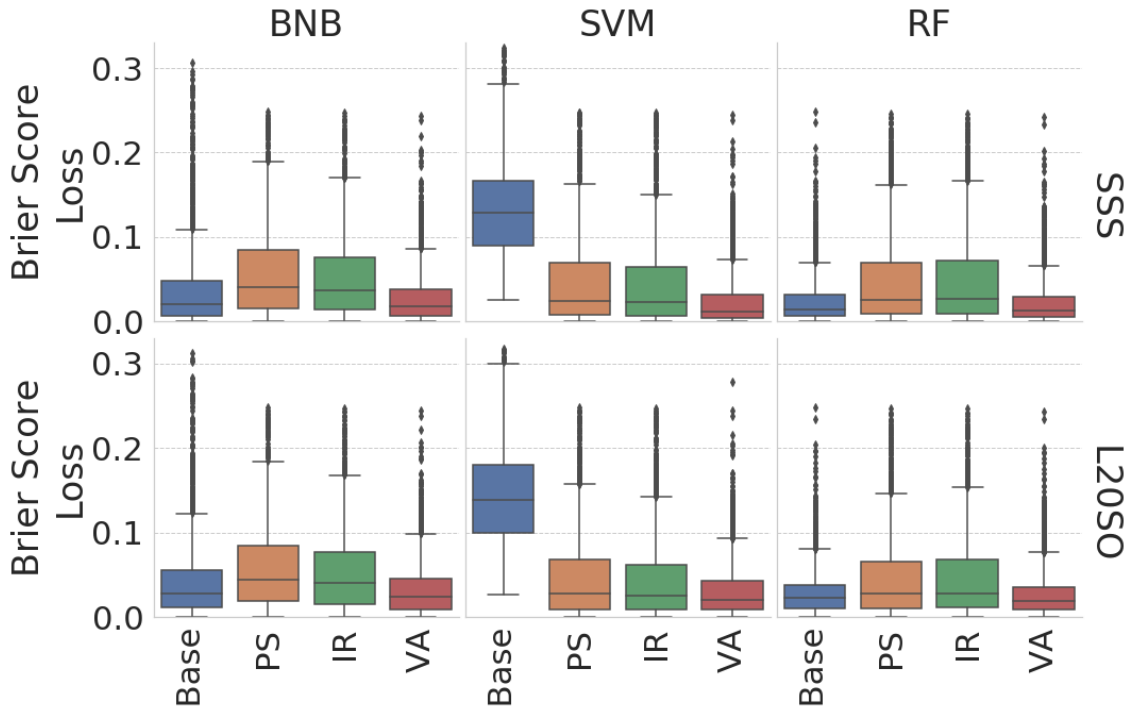


Figure 3. Per-target (*n=2,112*) Brier score loss of the Platt Scaling (PS), Isotonic Regression (IR) and Venn-ABERS (VA) scaling methods across different machine learning methods during Stratified Shuffle Split (SSS) and Leave 20% of Scaffold out (L20SO) cross validation. The Bernoulli Naïve Bayes (BNB), Linear Support Vector Machine (SVM) and Random Forest (RF) algorithms show different behavior in response to calibration. Our results show VA produces the most optimal per-target Brier score loss (lower scores are better) across the BNB, SVM and RF. VA achieved higher respective scores during L20SO, due to more challenging tasks of extrapolating to novel chemical scaffolds. PS and IR often even degrade Brier Score Loss for SSS and L20SO when compared to the base scores generated for the BNB and RF methods.

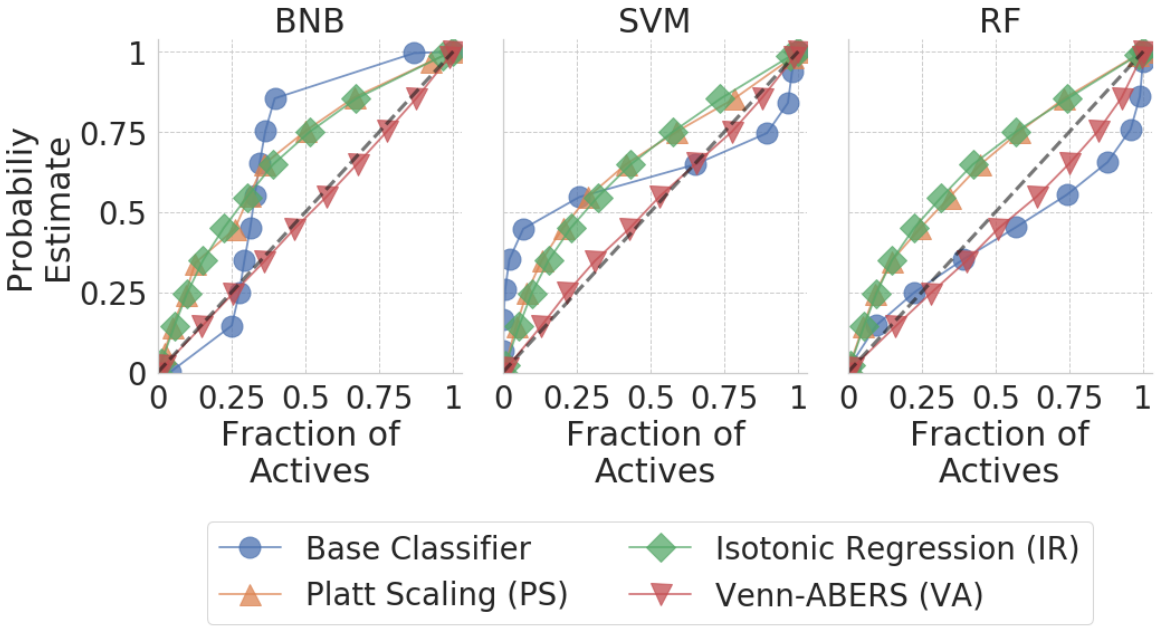


Figure 4. Calibration curve for Leave 20% Scaffolds Out (L20SO) validation for Bernoulli Naïve Bayes (BNB), Support Vector Machines (SVMs) and Random Forests (RF). It can be seen that all curves are further from the prefect calibration scenario, due to the more difficult classification task of extrapolating to novel chemical space. VA produces the relatively most optimal calibration curves of the methods tested. PS and IR result in overconfident calibration curves (above the diagonal), consistent with our previous Stratified Shuffle Split (SSS) results.

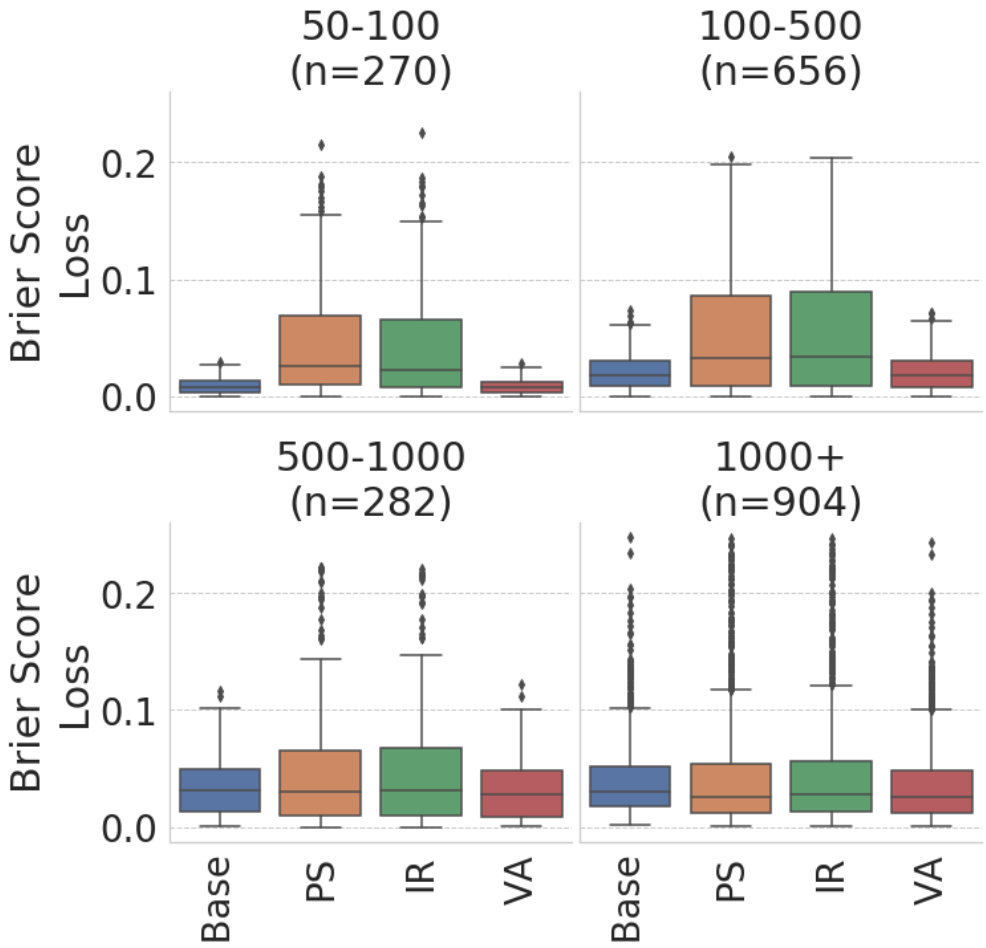


Figure 5. Brier Score Loss as a function of training set size and scaling methods using the Random Forest (RF) algorithm (which showed overall best performance based on preceding investigations). VA maintains performance across all target training set bin sizes, even in case of lower numbers of training points. In comparison, PS and IR exhibit degraded performance for smaller bins containing 50-100 and 100-200 data points in the training set. IR performs with similar performance to PS for the 50-100 data point bin, and hence IR overfitting on small amounts of calibration data is no more confounding than the sigmoidal assumption of calibration data form imposed by the PS approach for target prediction (which is different from previous studies, which were however performed on different data sets).

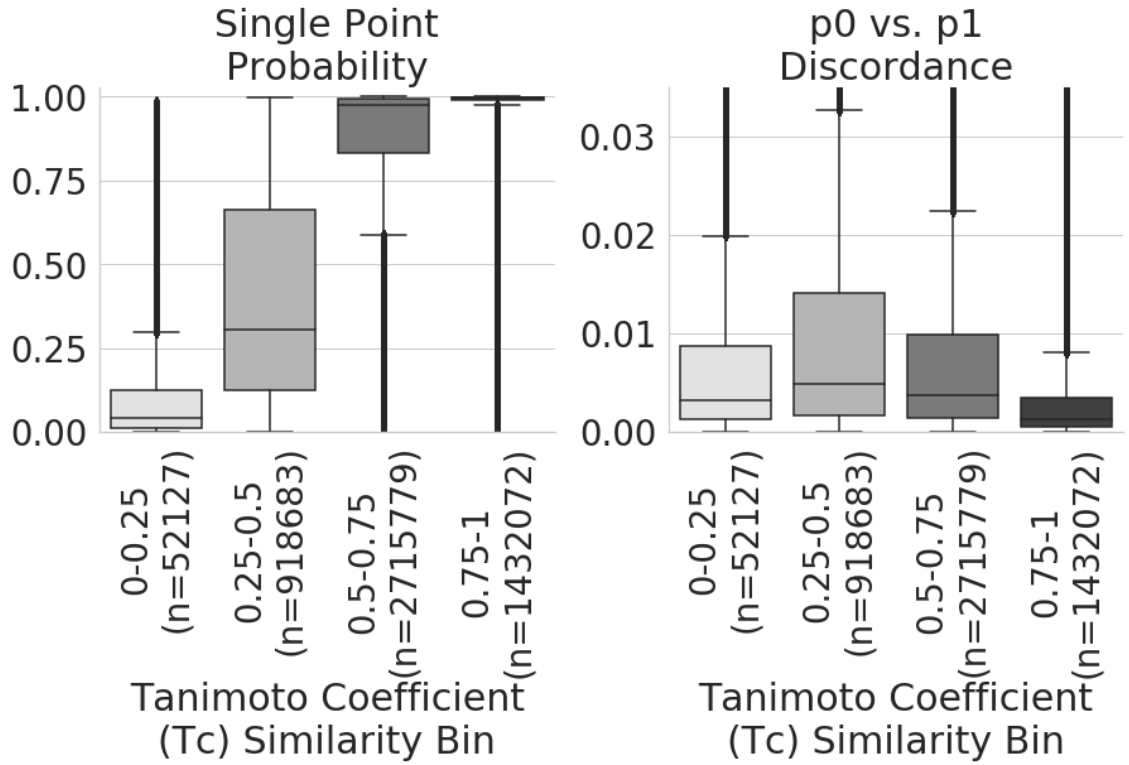


Figure 6. Analysis of *p0*, *p1* and single point probabilities obtained from Venn-ABERS scaling across nearest neighbor train-test Tanimoto Coefficient (Tc) similarity bins. During L20SO validation we observe that the discordance (margin) between the $p0$ and $p1$ derived values are higher for intermediate (central) similarities for the similarity bins 0.25 to 0.5 and 0.5 to 0.75, resulting in less confident predictions. Conversely, the very dissimilar and similar bins 0.0-0.25 and 0.75-1.0 have lower $p0$ and $p1$ discordance. Hence, we propose that $p0$ and $p1$ values could be used to provide an indication of prediction (un-)certainty.

Tables

Table 1. Classifiers and scaling methods used in this work, with their advantages, disadvantages and exemplary previous applications.

Algorithm/ Scaling	Core Methodology for class probability	Advantages	Disadvantages	Example previous applications
Bernoulli Naïve Bayes (BNB)	• Class conditional posterior probability	(default)	• Likelihood output populates extreme regions of the probability scale near 0 and 1	9, 21, 25, 39, 52, 76
Linear Support Vector Machine (SVM)	• Signed distance of prediction to the hyperplane	(default)	• Decision function scale ranges from negative to positive values, with many close to the mid-point due to the margin property of the hinge loss	8, 26, 31, 63, 77, 78
Random Forest (RF)	• Mean predicted fraction of active class samples in a Leaf across all the Trees in the Forest	(default)	• Difficulty making predictions near 0 and 1, since the underlying base models variance biases predictions that should be near 0 or 1 away from these values[14]	10, 11, 19, 25, 39, 52
Platt Scaling (PS)	• Parametric approach using a sigmoidal curve	• Applicable for small calibration sets • well suited to sigmoidal form	• Parametric form assumes a sigmoid sigmoidal distribution • Assumed symmetry, which is not true for highly unbalanced bioactivity data	• Initially developed for SVMs[32] toxicity prediction to ensemble predictions[3, 52]
Isotonic Regression (IR)	• Non-parametric approach using isotonic curve	• Non-parametric method makes no assumption on curve form	• Requires large numbers of calibration points • Tendency to overfit	• Target prediction[58]
Venn-ABERS (VA)	• Non-parametric approach using multi-probabilistic Venn Predictors based on IR	• Validity guarantees provided by Venn predictors • Susceptibility for IR to overfit is reduced	• *p0* or *p1* probability intervals must be combined into a single prediction for comparison between the scaling methods	• Combined with conformal prediction to improve p-value interpretability[57]. • Metabolic transformation prediction[59] and target prediction[58, 60, 79]

Table 2. Overall Brier score loss performance during Stratified Shuffle Split (SSS) and Leave 20% of Scaffolds Out (L20SO) validation. Results from the Brier score loss performance highlights that Venn-ABERS (VA) provides superior probability estimates (lower values in bold) compared to both the Platt Scaling (PS) and Isotonic Regression (IR) across all the algorithms tested, during both SSS and L20SO validation. Overall the effect of PS and IR scaling is minor overall except for the SVM. PS and IR methods perform with inferior performance than the baseline for the RF and BNB, which is due to the increased probabilities for inactive compounds, resulting in a larger number of false positive predictions.

Learner	Scaling	Overall Brier score loss	
		SSS (*n=58,777,503*)	L20SO (*n=23,533,709*)
BNB	Base	0.069	0.074
	PS	0.072	0.075
	IR	0.067	0.070
	VA	**0.050**	**0.054**
SVM	Base	0.156	0.158
	PS	0.056	0.058
	IR	0.053	0.055
	VA	**0.043**	**0.048**
RF	Base	0.035	0.042
	PS	0.046	0.048
	IR	0.047	0.049
	VA	**0.033**	**0.039**

Table 3. Per-target Brier score loss performance during Stratified Shuffle Split (SSS) and Leave 20% of Scaffolds Out (L20SO) validation across all targets modelled in this work. Results from the Brier score loss performance highlights that Venn-ABERS (VA) provides superior probability estimates (highlighted in bold, lower values) compared to both the Platt Scaling (PS) and Isotonic Regression (IR) across all the algorithms tested, during both SSS and L20SO validation. The PS and IR methods perform with inferior performance, which is due to the increased probabilities for inactive compounds, resulting in false positive predictions.

Learner	Scaling	SSS per-target Brier score loss (*n=2,112*)			L20SO per-target Brier score loss (*n=2,112*)		
		Mean	Standard Deviation	Median	Mean	Standard Deviation	Median
BNB	Base	0.039	0.052	0.02	0.044	0.053	0.028
	PS	0.057	0.053	0.04	0.058	0.051	0.044
	IR	0.052	0.049	0.037	0.053	0.048	0.04
	VA	**0.029**	0.032	**0.017**	**0.033**	0.033	**0.023**
SVM	Base	0.133	0.056	0.129	0.144	0.058	0.139
	PS	0.046	0.052	0.025	0.046	0.05	0.028
	IR	0.042	0.049	0.022	0.043	0.047	0.026
	VA	**0.025**	0.032	**0.012**	**0.032**	0.034	**0.02**
RF	Base	0.024	0.029	0.014	0.03	0.029	0.022
	PS	0.048	0.053	0.025	0.047	0.05	0.027
	IR	0.048	0.053	0.026	0.048	0.05	0.028
	VA	**0.023**	0.028	**0.013**	**0.028**	0.029	**0.019**

Table 4. Analysis of the probability estimates assigned to compounds across all protein target models. Venn-ABERS (VA) assigns the relatively most optimal probability estimates to compounds, with lower probability estimates in the mean and median columns for inactive compounds, whilst maintaining comparatively high estimates for the active compounds, when compared to the base score, as well as Platt scaling (PS) and Isotonic Regression (IR) methods, during both Stratified Shuffle Split (SSS) and Leave 20% Compounds Out (L20SO) validation. Overall, VA is more conservative in its predictions, whilst PS and IR may generate overconfident predictors with higher false positive rates. Standard deviations are large due to the deviation between probability estimates assigned to compounds across the range of target models. Large variances in scores are observed due to the different number and diversity of data points in every bioactivity class, as well as a different ratio of active to inactive data points.

Scaling	Classifier	Label	SSS (*n=58,777,503*)			L20SO (*n=23,533,709*)		
			Mean	Std. Dev	Median	Mean	Std. Dev	Median
Base	BNB	Inactive	0.057	0.204	0.000	0.057	0.205	0.000
	SVM		0.388	0.122	0.384	0.389	0.127	0.384
	RF		0.066	0.106	0.020	0.062	0.102	0.020
	BNB	Active	0.826	0.354	1.000	0.803	0.372	1.000
	SVM		0.669	0.120	0.678	0.671	0.130	0.681
	RF		0.811	0.266	0.950	0.746	0.278	0.870
PS	BNB	Inactive	0.175	0.175	0.119	0.171	0.175	0.116
	SVM		0.131	0.188	0.038	0.121	0.187	0.028
	RF		0.117	0.165	0.046	0.109	0.161	0.039
	BNB	Active	0.815	0.281	0.957	0.796	0.299	0.955
	SVM		0.860	0.224	0.975	0.841	0.246	0.973
	RF		0.887	0.232	0.997	0.858	0.258	0.995
IR	BNB	Inactive	0.147	0.192	0.063	0.145	0.192	0.061
	SVM		0.122	0.182	0.038	0.113	0.180	0.030
	RF		0.110	0.174	0.028	0.103	0.171	0.024
	BNB	Active	0.835	0.267	0.993	0.815	0.280	0.986
	SVM		0.866	0.233	0.989	0.843	0.258	0.987

	RF		0.889	0.229	0.999	0.860	0.251	0.995
VA	BNB	Inactive	0.067	0.018	0.124	0.066	0.017	0.124
	SVM		0.066	0.015	0.125	0.065	0.011	0.133
	RF		0.046	0.009	0.104	0.043	0.007	0.101
	BNB	Active	0.766	0.970	0.324	0.741	0.953	0.335
	SVM		0.791	0.950	0.285	0.778	0.959	0.305
	RF		0.840	0.993	0.284	0.798	0.978	0.307

Supplementary Table

Supplementary Table 1. Description of the bioactivity training data across different protein target families. There are biases toward the Kinase, GPCR and Ion Channel protein families, due to the wealth of bioactivity data for these target classes. Annotations are based on in-house methods.

Classification	#Target models	#Actives	#Inactives	#Sphere Exclusion (Putative) Inactives	#Murcko Skeleton Scaffolds	Ratio compounds: scaffolds
Kinase	449	2,401,875	6,358,405	1,410,230	3,513,304	2.9
Other	380	889,769	2,986,157	1,212,943	2,063,307	2.5
GPCR	222	2,139,718	4,314,921	1,483,299	2,395,925	3.3
Ion Channel	192	667,823	1,322,885	1,431,875	1,271,155	2.7
Transporter	164	336,593	690,023	694,637	704,208	2.4
Hydrolases	150	326,895	1,078,919	539,106	774,500	2.5
Protease	144	642,495	1,580,399	668,626	1,016,426	2.8
Oxidoreductases	132	434,312	900,535	873,930	822,412	2.7
Transferases	118	216,032	781,959	391,271	576,936	2.4
Lyases	39	64,842	107,988	273,027	187,327	2.4
NHR	34	266,440	851,785	-	351,058	3.2
Phosphatase	31	22,096	172,054	29,746	110,765	2.0
Ligases	22	18,422	113,739	49,836	86,343	2.1
Isomerases	20	22,810	118,656	41,909	81,915	2.2
Lipase	15	35,039	83,246	92,334	83,110	2.5

Supplementary Table 2. Sources of bioactivity data and the number of data points extracted. Intermediate steps of inactive data extraction are not shown in bold. AZ comprises both in-house and public data, so little as little data bias is introduced as possible, whilst maintaining the largest training set possible.

Data source	Label	Number of data points	Number of targets	Requires subsampling?
AZ ChemConnect (ChEMBL version 26 and in-house data)	**Active**	**8,505,197**	**2,112**	**No**
AZ HTS Screens	Inactive	189,965,064	400	Yes
PubChem	Inactive	419,121,152	2,116	Yes
Sphere Exclusion (SE)	Inactive	16,188,048	1,003	Yes
Final Inactive Dataset (after sub- sampling)	**Inactive**	**38,902,310**	**2,112**	**No**

Supplementary Table 3. Effect of scaling compared to the base algorithm scores. VA assigns conservative probability estimates outlined by a higher percentage of probability estimates in the "Decreased" and "No Change" columns, compared to PS and IR, which over-inflate probabilities for the RF during SSS and L20SO (with a higher proportion of "Increased" probability estimates compared to the base model). Absolute numbers of data points are shown in brackets.

Label	Learner	Scaling	SSS (*n=58,777,503*)			L20SO (*n=23,533,709*)		
			Decreased	No Change	Increased	Decreased	No Change	Increased
Inactives	BNB	PS	16.79% (7,728,517)	73.99% (34,066,972)	9.22% (4,244,912)	16.67% (3,075,300)	74.03% (13,655,582)	9.29% (1,714,359)
		IR	47.61% (21,919,701)	13.15% (6,053,449)	39.24% (18,067,251)	48.03% (8,859,512)	12.27% (2,262,334)	39.7% (7,323,395)
		VA	51.37% (23,649,466)	7.45% (3432144)	41.18% (18,958,791)	50.77% (9,364,792)	8.62% (1,590,161)	40.61% (7,490,288)
	SVM	PS	54.25% (24,976,775)	0.24% (112,227)	45.51% (20,951,399)	54.04% (9,968,680)	0.24% (44,773)	45.71% (8,431,788)
		IR	53.72% (24,733,488)	0.26% (120,662)	46.02% (21,186,251)	53.53% (9,873,319)	0.25% (46,973)	46.22% (8,524,949)
		VA	51.23% (23,587,023)	0.26% (120,284)	48.51% (22,333,094)	51.46% (9,491,693)	0.26% (47103)	48.29% (8,906,445)
	RF	PS	60.89% (28,033,370)	5.38% (2,476,549)	33.73% (15,530,482)	60.28% (11,118,819)	7.17% (1,322,432)	32.55% (6,003,990)
		IR	62.43% (28,744,333)	7.94% (3,657,744)	29.62% (13,638,324)	61.81% (11,401,407)	9.14% (1,686,298)	29.05% (5,357,536)
		VA	66.71% (30,712,592)	6.57% (3,024,712)	26.72% (12,303,097)	65.93% (12,160,286)	8.4% (1,549,487)	25.67% (4,735,468)
Actives	BNB	PS	37.31% (4,752,550)	53.79% (6,851,595)	8.89% (1,132,957)	40.74% (2,072,847)	49.68% (2,528,132)	9.58% (487,489)
		IR	56.34% (7,176,532)	23.36% (2,975,597)	20.29% (2,584,973)	59.06% (3,005,413)	19.14% (974,125)	21.79% (1,108,930)
		VA	54.65% (6,960,823)	17.23% (2,194,104)	28.12% (3,582,175)	57.96% (2,949,237)	14.39% (732334)	27.65% (1,406,897)
	SVM	PS	57.98% (7,385,027)	0.28% (35,677)	41.74% (5,316,398)	59.91% (3,048,400)	0.34% (17,330)	39.75% (2,022,738)
		IR	58.55% (7,457,258)	0.29% (37,538)	41.16% (5,242,306)	60.68% (3,087,636)	0.33% (17,001)	38.99% (1,983,831)
		VA	60.14% (7,660,134)	0.24% (31,085)	39.62% (5,045,883)	61.94% (3,151,933)	0.31% (15,643)	37.75% (1,920,892)
	RF	PS	56.96% (7,255,559)	8.59% (1,093,497)	34.45% (4,388,046)	56.77% (2,888,779)	2.92% (148,811)	40.3% (2,050,878)
		IR	51.02% (6,498,004)	8.88% (1,130,922)	40.1% (5,108,176)	53.98% (2,746,966)	2.76% (140,315)	43.26% (2,201,187)
		VA	66.48% (8,467,948)	6.42% (817,212)	27.1% (3,451,942)	64.46% (3,279,820)	1.73% (88,000)	33.81% (1,720,648)

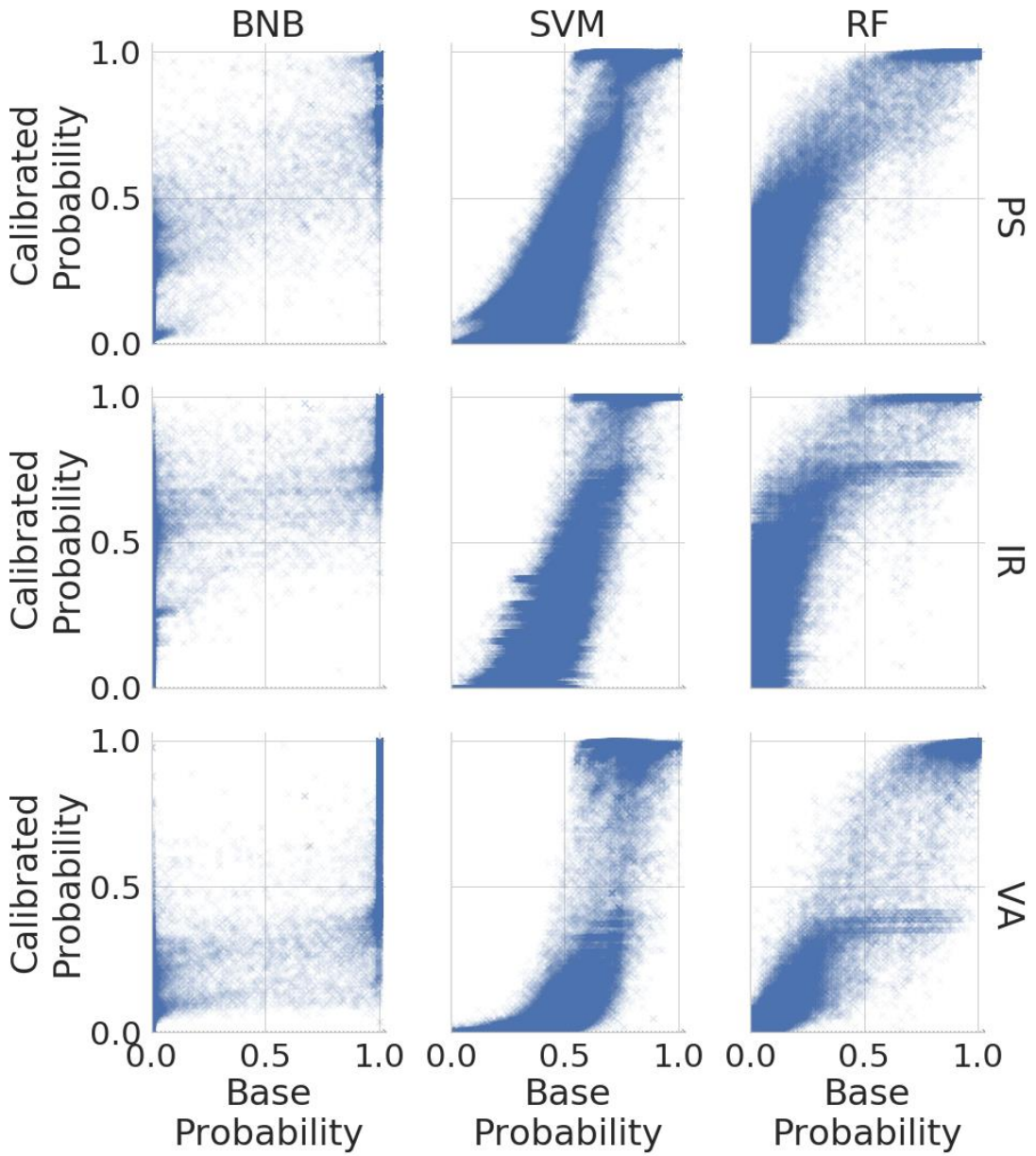
Supplementary Table 4. Effect of calibration set size on RF Brier score loss performance. VA performs with the best Brier score loss across all bins, outlining that this method is able to provide reliable probability estimates for even small number of calibration instances. In comparison, PS and IR produce worse probability estimates compared to the base estimator for the small bins 50-100 and 100-500, hence these methods should not be used for targets with small number of calibration instances.

Bins	Scaling	Mean	Standard Deviation	Median
50-100 (n=270)	Base	0.009	0.006	0.008
	PS	0.047	0.050	0.026
	IR	0.045	0.049	0.023
	VA	**0.008**	0.006	**0.008**
100-500 (n=656)	Base	0.021	0.014	0.018
	PS	0.049	0.048	0.033
	IR	0.051	0.049	0.034
	VA	**0.020**	0.015	**0.018**
500-1000 (n=282)	Base	0.036	0.026	0.032
	PS	0.047	0.051	0.030
	IR	0.048	0.051	0.032
	VA	**0.033**	0.027	**0.028**
1000+ (n=904)	Base	0.041	0.035	0.030
	PS	0.045	0.051	0.026
	IR	0.046	0.052	0.027
	VA	**0.037**	0.036	**0.025**

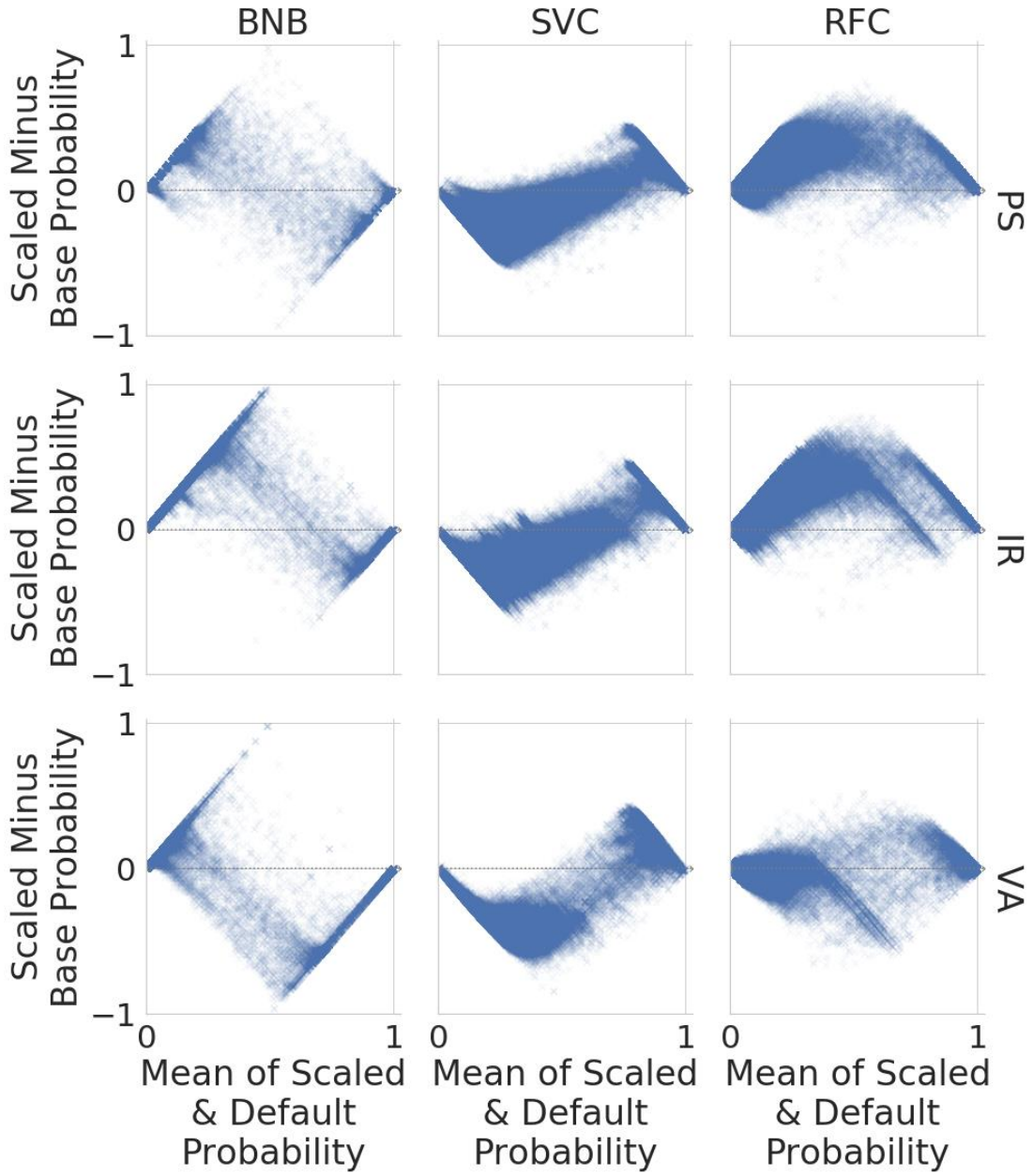
Supplementary Table 5. Distribution of single point probability, p0 and p1 values across train-test set Tanimoto Coefficient (Tc) similarity bins. We show here that there is larger p0 vs. p1 discordance for difficult testing instances (similarity bins 0.25-0.5 and 0.5-0.75), where input chemistry is neither very similar or very dissimilar to the nearest neighbor in the training set, and hence the interval between p0 and p1 could be used in the future to estimate the confidence of a prediction.

	Nearest Neighbour Train-Test ECFP_4 Tanimoto Coefficient (Tc) Similarity Bin	Number data points	Mean	Standard Deviation	Median
Single Point Probability	0.0-0.25	52,127	0.104	0.155	0.04
	0.25-0.5	918,683	0.398	0.316	0.307
	0.5-0.75	2,715,779	0.854	0.237	0.976
	0.75-1.0	1,432,072	0.975	0.085	0.997
p0 vs. p1 Discordance	0.0-0.25	52,127	0.011	0.029	0.003
	0.25-0.5	918,683	0.014	0.031	0.005
	0.5-0.75	2,715,779	0.011	0.027	0.004
	0.75-1.0	1,432,072	0.005	0.016	0.001

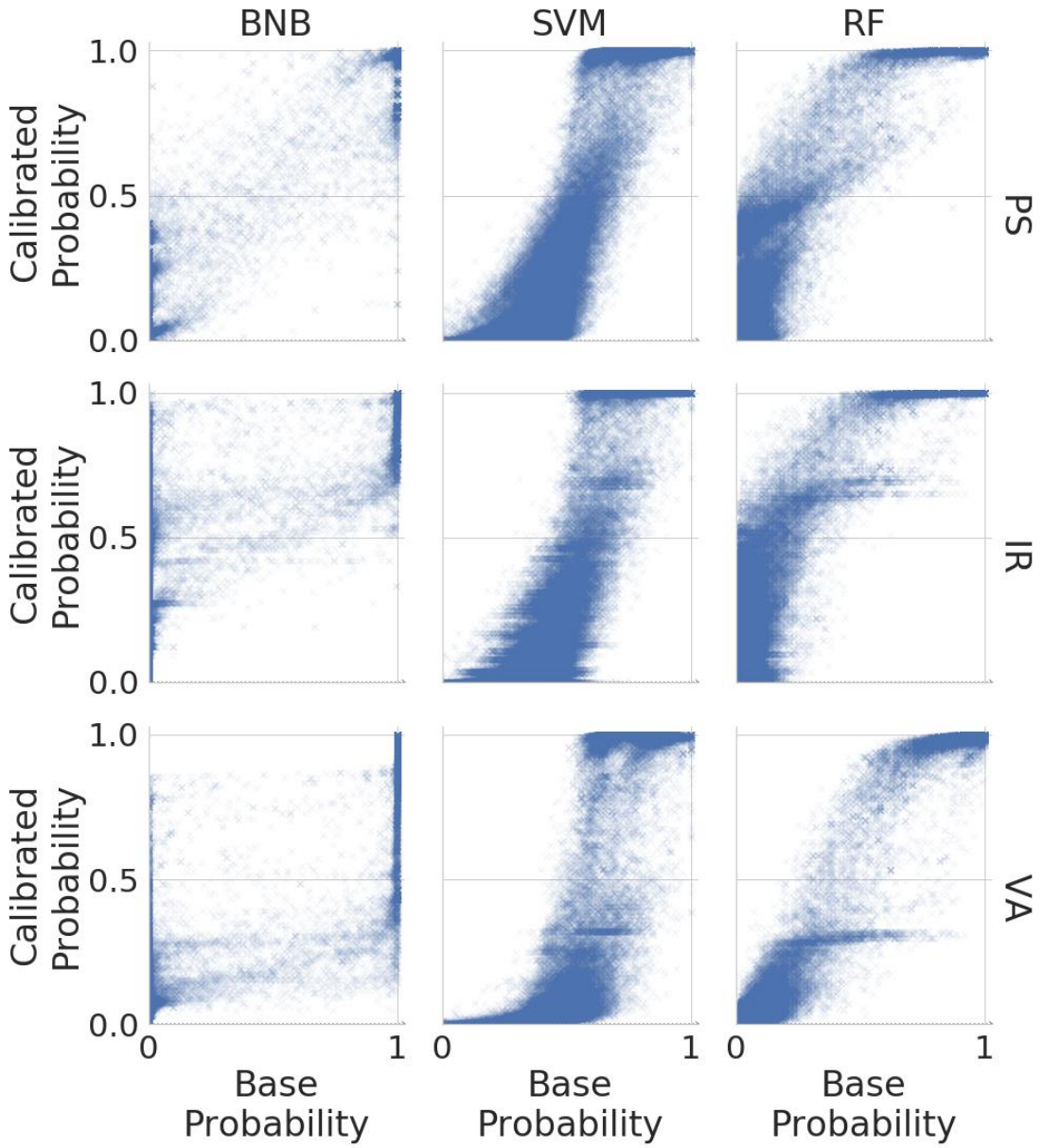
Supplementary Figures



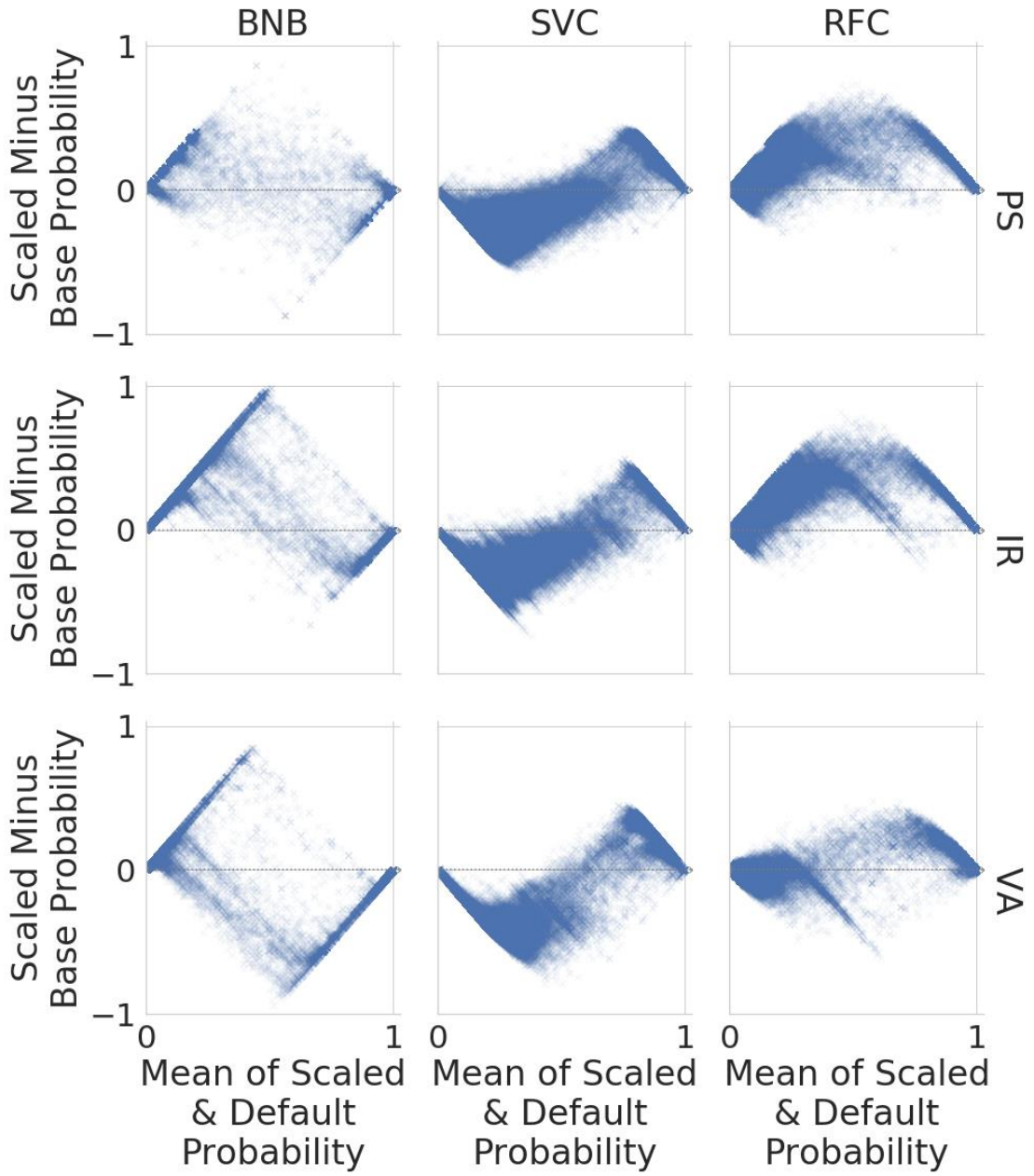
Supplementary Figure 1. Effect of scaling during SSS validation versus base probabilities (*n=58,777,503*). Platt (Sigmoid) and Isotonic Scaling inflate the probabilities generated by the models, as shown by the spread markers towards the top right of these plots. In comparison, Venn-ABERS markers show a tendency to form conservative probabilities indicated by the density of markers shifted toward the right of the bottom of the curve.



Supplementary Figure 2. SSS validation Bland-Altman plot of the scores generated using Platt (sigmoid) (PS), Isotonic Regression (IR), Venn-ABERS (VA) Scaling (*n=58,777,503*). PS and IR produce markers above the zero line, illustrating the tendency for PS and IR to produce inflated probabilities. VA produces conservative estimates with markers below the *zero line* for already low base probability estimates.



Supplementary Figure 3. Effect of scaling during L20SO validation (*n*=23,533,709). Platt (Sigmoid) Scaling (PS) and Isotonic Regression (IR) methods inflate the probabilities generated by the models, as shown by the shifted markers towards the upper right of these plots. In comparison, Venn-ABERS markers form a tighter curve with a tendency to form conservative probabilities, particularly indicated by the shifted density of markers toward the bottom right of the curve.



Supplementary Figure 4. L20SO validation Bland-Altman plot of the scores generated using Platt (sigmoid) (PS), Isotonic Regression (IR), Venn-ABERS (VA) Scaling *(n*=23,533,709). PS and IR produce a higher number of markers above the zero line then compared to VA, illustrating the tendency for PS and IR to produce inflated probabilities. VA shows markers below the *zero y-axis* around, when the base algorithm assigns already low predictions.

References

1. Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L., Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases. *J Chem Inf Model* **2006**, 46, 1124-1133.

2. Mayr, A.; Klambauer, G.; Unterthiner, T.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Clevert, D. A.; Hochreiter, S., Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem Sci* **2018**, 9, 5441-5451.

3. Unterthiner, T.; Mayr, A.; Klambauer, G.; Steijaert, M.; Wegner, J. K.; Ceulemans, H.; Hochreiter, S. Deep learning for drug target prediction. In: *Conference Neural Information Processing Systems Foundation (NIPS 2014)*, **2014**.

4. Lenselink, E. B.; Ten Dijke, N.; Bongers, B.; Papadatos, G.; van Vlijmen, H. W. T.; Kowalczyk, W.; AP, I. J.; van Westen, G. J. P., Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *J Cheminform* **2017**, 9, 45.

5. Drakakis, G.; Wafford, K. A.; Brewerton, S. C.; Bodkin, M. J.; Evans, D. A.; Bender, A., Polypharmacological *In Silico* Bioactivity Profiling and Experimental Validation Uncovers Sedative-Hypnotic Effects of Approved and Experimental Drugs in Rat. *ACS Chem Biol* **2017**, 12, 1593-1602.

6. Wen, M.; Zhang, Z.; Niu, S.; Sha, H.; Yang, R.; Yun, Y.; Lu, H., Deep-Learning-Based Drug-Target Interaction Prediction. *J Proteome Res* **2017**, 16, 1401-1409.

7. Lavecchia, A.; Cerchia, C., In silico methods to address polypharmacology: current status, applications and future perspectives. *Drug Discov Today* **2016**, 21, 288-298.

8. Heikamp, K.; Bajorath, J., Comparison of confirmed inactive and randomly selected compounds as negative training examples in support vector machine-based virtual screening. *J Chem Inf Model* **2013**, 53, 1595-1601.

9. Mervin, L. H.; Afzal, A. M.; Drakakis, G.; Lewis, R.; Engkvist, O.; Bender, A., Target prediction utilising negative bioactivity data covering large chemical space. *J Cheminform* **2015**, 7, 51.

10. Martin, E. J.; Polyakov, V. R.; Zhu, X. W.; Tian, L.; Mukherjee, P.; Liu, X., All-Assay-Max2 pQSAR: Activity Predictions as Accurate as Four-Concentration IC50s for 8558 Novartis Assays. *J Chem Inf Model* **2019**, 59, 4450-4459.

11. Merget, B.; Turk, S.; Eid, S.; Rippmann, F.; Fulle, S., Profiling Prediction of Kinase Inhibitors: Toward the Virtual Assay. *J Med Chem* **2017**, 60, 474-485.

12. Martin, E.; Mukherjee, P., Kinase-kernel models: accurate in silico screening of 4 million compounds across the entire human kinome. *J Chem Inf Model* **2012**, 52, 156-170.

13. Zadrozny, B.; Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In: *Eighth ACM SIGKDD international conference on Knowledge discovery and data mining* **2002**, 694-699.

14. Niculescu-Mizil, A.; Caruana, R. Predicting good probabilities with supervised learning. In: *22nd international conference on Machine learning* **2005**, 625-632.

15. Rüping, S. Robust probabilistic calibration. *Springer* **2006**, 743-750.

16. Flach, P. A., Classifier Calibration. *Encyclopedia of Machine Learning and Data Mining* **2016**, 1-8.

17. Jacobs, K. Independent Identically Distributed (IID) Random Variables. *Discrete Stochastics* **1992**, 65-101.

18. Humbeck, L.; Koch, O., What Can We Learn from Bioactivity Data? Chemoinformatics Tools and Applications in Chemical Biology Research. *ACS Chem Biol* **2017**, 12, 23-35.

19. Mervin, L. H.; Afzal, A. M.; Brive, L.; Engkvist, O.; Bender, A., Extending in Silico Protein Target Prediction Models to Include Functional Effects. *Front Pharmacol* **2018**, 9, 613.

20. Mervin, L. H.; Cao, Q.; Barrett, I. P.; Firth, M. A.; Murray, D.; McWilliams, L.; Haddrick, M.; Wigglesworth, M.; Engkvist, O.; Bender, A., Understanding Cytotoxicity and Cytostaticity in a High-Throughput Screening Collection. *ACS Chem Biol* **2016,** 11, 3007-3023.

21. Koutsoukas, A.; Lowe, R.; Kalantarmotamedi, Y.; Mussa, H. Y.; Klaffke, W.; Mitchell, J. B.; Glen, R. C.; Bender, A., *In silico* target predictions: defining a

benchmarking data set and comparison of performance of the multiclass Naïve Bayes and Parzen-Rosenblatt window. *J Chem Inf Model* **2013**, 53, 1957-1966.

22. Smusz, S.; Kurczab, R.; Bojarski, A. J., The influence of the inactives subset generation on the performance of machine learning methods. *J Cheminform* **2013**, 5, 17.

23. Gobbi, A.; Lee, M.-L., DISE: directed sphere exclusion. *Journal of chemical information and computer sciences* **2003**, 43, 317-323.

24. Bosc, N.; Atkinson, F.; Felix, E.; Gaulton, A.; Hersey, A.; Leach, A. R., Large scale comparison of QSAR and conformal prediction methods and their applications in drug discovery. *J Cheminform* **2019**, 11, 4.

25. Kurczab, R.; Smusz, S.; Bojarski, A. J., The influence of negative training set size on machine learning-based virtual screening. *J Cheminform* **2014**, 6, 32.

26. Rodriguez-Perez, R.; Vogt, M.; Bajorath, J., Influence of Varying Training Set Composition and Size on Support Vector Machine-Based Prediction of Active Compounds. *J Chem Inf Model* **2017**, 57, 710-716.

27. Sun, J.; Carlsson, L.; Ahlberg, E.; Norinder, U.; Engkvist, O.; Chen, H., Applying Mondrian Cross-Conformal Prediction To Estimate Prediction Confidence on Large Imbalanced Bioactivity Data Sets. *J Chem Inf Model* **2017**, 57, 1591-1598.

28. Krier, M.; Bret, G.; Rognan, D., Assessing the scaffold diversity of screening libraries. *J Chem Inf Model* **2006**, 46, 512-524.

29. Langdon, S. R.; Brown, N.; Blagg, J., Scaffold diversity of exemplified medicinal chemistry space. *J Chem Inf Model* **2011**, 51, 2174-2185.

30. Mervin, L. H.; Bulusu, K. C.; Kalash, L.; Afzal, A.; Svensson, F.; Firth, M. A.; Barrett, I.; Engkvist, O.; Bender, A., Orthologue chemical space and its influence on target prediction. *Bioinformatics* **2017**, 34, 72-79.

31. Plewczynski, D.; von Grotthuss, M.; Spieser, S. A.; Rychlewski, L.; Wyrwicz, L. S.; Ginalski, K.; Koch, U., Target specific compound identification using a support vector machine. *Comb Chem High Throughput Screen* **2007**, 10, 189-196.

32. Platt, J., Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* **1999**, 10, 61-74.

33. Murphy, K. P., Naive Bayes classifiers. *University of British Columbia* **2006**.

34. Zhang, H., The optimality of naive Bayes. *University of New Brunswick* **2004**.

35. Breiman, L., Random forests. *Machine learning* **2001**, 45, 5-32.

36. Dankowski, T.; Ziegler, A., Calibrating random forests for probability estimation. *Stat Med* **2016**, 35, 3949-3960.

37. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K. Q. On calibration of modern neural networks. In: *34th International Conference on Machine Learning* **2017**, 1321-1330.

38. Johansson, U.; Gabrielsson, P. Are Traditional Neural Networks Well-Calibrated? In: *International Joint Conference on Neural Networks (IJCNN)* **2019**, 1-8.

39. Mitchell, J. B., Machine learning methods in chemoinformatics. *Wiley Interdiscip Rev Comput Mol Sci* **2014**, 4, 468-481.

40. Drakakis, G.; Koutsoukas, A.; Brewerton, S. C.; Bodkin, M. J.; Evans, D. A.; Bender, A., Comparing Global and Local Likelihood Score Thresholds in Multiclass Laplacian-Modified Naïve Bayes Protein Target Prediction. *Comb Chem High Throughput Screen* **2015**.

41. Van Calster, B.; McLernon, D. J.; van Smeden, M.; Wynants, L.; Steyerberg, E. W., Calibration: the Achilles heel of predictive analytics. *BMC Med* **2019**, 17, 230.

42. Vaicenavicius, J.; Widmann, D.; Andersson, C.; Lindsten, F.; Roll, J.; Schön, T. B., Evaluating model calibration in classification. *arXiv preprint arXiv:1902.06977* **2019**.

43. Dimitrov, S.; Dimitrova, G.; Pavlov, T.; Dimitrova, N.; Patlewicz, G.; Niemela, J.; Mekenyan, O., A stepwise approach for defining the applicability domain of SAR and QSAR models. *J Chem Inf Model* **2005**, 45, 839-849.

44. Aniceto, N.; Freitas, A. A.; Bender, A.; Ghafourian, T., A novel applicability domain technique for mapping predictive reliability across the chemical space of a QSAR: reliability-density neighbourhood. *Journal of cheminformatics* **2016**, 8, 69.

45. Norinder, U.; Carlsson, L.; Boyer, S.; Eklund, M., Introducing conformal prediction in predictive modeling for regulatory purposes. A transparent and flexible alternative to applicability domain determination. *Regul Toxicol Pharmacol* **2015**, 71, 279-284.

46. Eklund, M.; Norinder, U.; Boyer, S.; Carlsson, L., The application of conformal prediction to the drug discovery process. *Annals of Mathematics and Artificial Intelligence* **2015**, 74, 117-132.

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

48. Vovk, V., Venn predictors and isotonic regression. *CoRR abs/1211.0025* **2012**.

49. Vovk, V.; Petej, I., Venn-abers predictors. *arXiv preprint arXiv:1211.0025* **2012**.

50. Franc, V.; Zien, A.; Schölkopf, B. Support vector machines as probabilistic models. In: *28th International Conference on Machine Learning (ICML-11)* **2011**, 665-672.

51. Svensson, F.; Norinder, U.; Bender, A., Modelling compound cytotoxicity using conformal prediction and PubChem HTS data. *Toxicology Research* **2017**, 6, 73-80.

52. Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S., DeepTox: Toxicity Prediction using Deep Learning. *Frontiers in Environmental Science* **2016**, 3.

53. Kull, M.; Silva Filho, T. M.; Flach, P., Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electronic Journal of Statistics* **2017**, 11, 5052-5080.

54. Vezhnevets, A.; Barinova, O., Avoiding boosting overfitting by removing confusing samples. *Machine Learning 2007*, 430-441.

55. Vovk, V.; Petej, I.; Fedorova, V. Large-scale probabilistic predictors with and without guarantees of validity. *Advances in Neural Information Processing Systems* **2015**, 892-900.

56. Zhou, C. Conformal and Venn Predictors for Multi-probabilistic Predictions and Their Applications. Department of Computer Science (Thesis). *Royal Holloway, University of London*, **2015**.

57. Vovk, V.; Petej, I.; Fedorova, V. From conformal to probabilistic prediction. In: *International Conference on Artificial Intelligence Applications and Innovations* 2014, 221-230.

58. Toccaceli, P.; Nouretdinov, I.; Luo, Z.; Vovk, V.; Carlsson, L.; Gammerman, A., ExCAPE WP1. Probabilistic prediction (http://www.clrc.rhul.ac.uk/projects/ExCAPE/Report_wp1_2-20160612.pdf, accessed on 19th Dec 2019)

59. Arvidsson, S.; Spjuth, O.; Carlsson, L.; Toccaceli, P. Prediction of Metabolic Transformations using Cross Venn-ABERS Predictors. *Conformal and Probabilistic Prediction and Applications* **2017**, 118-131.

60. Thomas, J., Exascale Compound Activity Prediction Engine (https://cordis.europa.eu/project/rcn/197536/en, accessed on 19th Dec 2019).

61. Buendia, R.; Kogej, T.; Engkvist, O.; Carlsson, L.; Linusson, H.; Johansson, U.; Toccaceli, P.; Ahlberg, E., Accurate Hit Estimation for Iterative Screening Using Venn-ABERS Predictors. *J Chem Inf Model* **2019**, 59, 1230-1237.

62. Ahlberg, E.; Buendia, R.; Carlsson, L. Using Venn-Abers predictors to assess cardio-vascular risk. *Conformal and Probabilistic Prediction and Applications* **2018**, 132-146.

63. Meng, F. R.; You, Z. H.; Chen, X.; Zhou, Y.; An, J. Y., Prediction of Drug-Target Interaction Networks from the Integration of Protein Sequences and Drug Chemical Structures. *Molecules* **2017**, 22.

64. Fakhraei, S.; Raschid, L.; Getoor, L., Drug-target interaction prediction for drug repourposing with probabilistic similarity logic. In: *12th International Workshop on Data Mining in Bioinformatics* **2013**, 10-17.

65. Muresan, S.; Petrov, P.; Southan, C.; Kjellberg, M. J.; Kogej, T.; Tyrchan, C.; Varkonyi, P.; Xie, P. H., Making every SAR point count: the development of Chemistry Connect for the large-scale integration of structure and bioactivity data. *Drug Discov Today* **2011**, 16, 1019-1030.

66. Wang, Y.; Xiao, J.; Suzek, T. O.; Zhang, J.; Wang, J.; Bryant, S. H., PubChem: a public information system for analyzing bioactivities of small molecules. *Nucleic Acids Res* **2009**, 37, 623-633.

67. Bolton, E. E.; Wang, Y.; Thiessen, P. A.; Bryant, S. H., PubChem: integrated platform of small molecules and biological activities. *Annual reports in computational chemistry* **2008**, 4, 217-241.

68. Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Krüger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P., The ChEMBL bioactivity database: an update. *Nucleic Acids Res* **2014**, 42, 1083-1090.

69. Coordinators, N. R., Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **2013**, 41, 8-20.

70. OEChem, T., version 2.0.0. OpenEye Scientific Software, Santa Fe, NM (https://docs.eyesopen.com/, accessed on 19th Dec 2019).

71. Landrum, G. RDKit: Open-source cheminformatics. (http://www.rdkit.org, accessed on 19th Dec 2019),

72. Morgan, H. L., The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation* **1965**, 5, 107-113.

73. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V., Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research* **2011**, 12, 2825-2830.

74. Brier, G. W., The statistical theory of turbulence and the problem of diffusion in the atmosphere. *Journal of Meteorology* **1950**, 7, 283-290.

75. Jolliffe, I. T.; Stephenson, D. B., *Forecast verification: a practitioner's guide in atmospheric science* **2003**.

76. Nigsch, F.; Bender, A.; Jenkins, J. L.; Mitchell, J. B., Ligand-target prediction using Winnow and naive Bayesian algorithms and the implications of overall performance statistics. *J Chem Inf Model* **2008**, 48, 2313-2325.

77. Wale, N.; Karypis, G., Target fishing for chemical compounds using target-ligand activity data and ranking based methods. *J Chem Inf Model* **2009**, 49, 2190-2201.

78. Jacob, L.; Hoffmann, B.; Stoven, V.; Vert, J. P., Virtual screening of GPCRs: an in silico chemogenomics approach. *BMC Bioinformatics* **2008**, 9, 363.

79. Nouretdinov, I. Improving Reliable Probabilistic Prediction by Using Additional Knowledge. *In: Conformal and Probabilistic Prediction and Applications* **2017**, 193-200.