# Using Conformal Prediction to Prioritize Compound Synthesis in Drug Discovery

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

**Susanne Winiwarter** SUSANNE.WINIWARTER@ASTRAZENECA.COM
*Predictive Compound ADME & Safety, Drug Safety & Metabolism, AstraZeneca IMED Biotech Unit, Mölndal, Sweden*

**Henrik Boström** HENRIK.BOSTROM@DSV.SU.SE
*Dept. of Computer and Systems Sciences, Stockholm University, Sweden*

**Henrik Linusson** HENRIK.LINUSSON@HB.SE
*Dept. of Information Technology, University of Borås, Sweden*

**Tuve Löfström** TUVE.LOFSTROM@HB.SE
*Dept. of Information Technology, University of Borås, Sweden*
*Dept. of Computer Science and Informatics, Jönköping University, Sweden*

**Ulf Norinder** ULF.NORINDER@SWETOX.SE
*Swetox, Karolinska Institutet, Unit of Toxicology Sciences, Sweden*
*Dept. of Computer and Systems Sciences, Stockholm University, Sweden*

**Ulf Johansson** ULF.JOHANSSON@JU.SE
*Dept. of Information Technology, University of Borås, Sweden*
*Dept. of Computer Science and Informatics, Jönköping University, Sweden*

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

**Oscar Hammar** OSCAR.HAMMAR@ASTRAZENECA.COM
*Quantitative Biology, Discovery Sciences, AstraZeneca IMED Biotech Unit, Mölndal, Sweden*

**Claus Bendtsen** CLAUS.BENDTSEN@ASTRAZENECA.COM
*Quantitative Biology, Discovery Sciences, AstraZeneca IMED Biotech Unit, Cambridge, UK*

**Lars Carlsson** LARS.A.CARLSSON@ASTRAZENECA.COM
*Quantitative Biology, Discovery Sciences, AstraZeneca IMED Biotech Unit, Mölndal, Sweden*

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

## Abstract

The choice of how much money and resources to spend to understand certain problems is of high interest in many areas. This work illustrates how computational models can be more tightly coupled with experiments to generate decision data at lower cost without reducing the quality of the decision. Several different strategies are explored to illustrate the trade off between lowering costs and quality in decisions.

AUC is used as a performance metric and the number of objects that can be learnt from is constrained. Some of the strategies described reach AUC values over 0.9 and outperforms

strategies that are more random. The strategies that use conformal predictor p-values show varying results, although some are top performing.

The application studied is taken from the drug discovery process. In the early stages of this process compounds, that potentially could become marketed drugs, are being routinely tested in experimental assays to understand the distribution and interactions in humans.
**Keywords:** Drug discovery, Conformal Prediction, ADME properties, Decision support

## 1. Introduction

Drug discovery is a costly and time consuming process that involves several steps (Bunnage, 2011; Gautam and Pan, 2016). First a biological pathway that is likely to affect a certain disease needs to be identified. Secondly, a specific drug target has to be discovered, a particular protein or enzyme within the pathway which function can be altered through intervention with a chemical entity in a way to favourably influence and, ultimately, cure the disease. One can identify a chemical entity that actually influences this target, i.e., a first compound that shows potency. However, such a hit is normally very far from the final drug candidate, since other compound properties need to be considered as well: The compound needs to be able to reach the place of action, i.e., it needs to be absorbed and correctly distributed within the body. It also needs to stay long enough to have an effect, i.e., it should not be metabolized or otherwise eliminated too fast. Finally, it must not have any adverse toxic effects. During lead identification and lead optimisation phases these additional properties are optimised together with the potency in an iterative process, the design-make-test-analyze (DMTA) cycle (Plowright et al., 2012). New chemicals that fit a certain design hypothesis will be made, submitted to certain, primary, tests and analysed. Results will be used in the design of the next set of compounds to be made. Obviously, the analysis will also define whether a compound is of interest to be submitted to additional, secondary assays, or, finally, be selected as drug candidate. The definition of primary and secondary assays may vary, dependent on company, drug discovery project or project stage. At AstraZeneca the primary assays used for all compounds are the respective potency assay and a set of assays to give essential information about absorption, distribution, metabolism and excretion (ADME) (Ballard et al., 2012). These early ADME assays include lipophilicity, solubility and metabolic stability screens and give a first insight into whether a compound has good or bad ADME properties. All data typically also feeds into machine learning algorithms that can be used to predict the property in question for new compounds, (Davis and Wood, 2013) and thus can help in compound design as well as when selecting which compounds to proceed further. The process described usually depends on the synthesis and experimental testing of hundreds of compounds each month, with only about 20% being of interest for further investigations. Thus, to confidently identify poor compounds at the design stage is of great value. Previous work has shown that conformal prediction can be useful in drug discovery using traditional machine learning based evaluation criteria (Eklund et al., 2015; Ahlberg et al., 2015; Norinder et al., 2014). In this work we want to evaluate several automated strategies to decide which compounds to make, based on conformal predictions of ADME properties. At the same time we need to ensure that the possibility to identify the optimal compound will be maintained.

This paper focus on the DMTA process and automated strategies for the selection of compounds for synthesis and testing using machine learning and conformal prediction. The

aim is to use cost as a more practical measure from a process point of view. The possibility to reduce cost comes in two stages, one with just reducing testing of compounds in assays where the activity of the compound can be predicted with sufficient confidence. Such a decision has the potential of saving between 5 and 50$ per compound and assay depending on the complexity of the assay etc. The second and preferred approach would be to intervene prior to synthesis and thus remove the cost of making and testing the compound all together. The cost for synthesizing a compound tends to exceed $1000, thus eliminating compounds from synthesis potentially has big impacts on cost. If one would e.g. reduce the compounds synthesized by 20% then savings in excess of $1M would be expected when considering only five thousand compounds overall.

The aim of this study is to assess different strategies to select which compounds to synthesize and ultimately test in ADME assays based on different types of conformal predictors. To evaluate possible reduction in synthesis, a desirability function has been created that uses the results from the ADME assays and assigns a label `Good` or `Bad` to each compound. All data have been ordered by the experimental test date, to mimic the actual data generating process, and evaluations are made using teaching schedules (Vovk et al., 2005) in batches of 500 compounds, which corresponds to a periodic updating schedule resembling the DMTA process. Preset values of reduction of number of compounds to synthesize have been set to [20, 40, 60]% and all methods are compared using area under ROC curve (AUC) for each level of reduction.

## 2. Method

Each strategy will start with a set of initial training examples, $T_T$, and select to learn from subsequent objects by applying a choice function to the objects in the prediction set, $T_P$. The selected objects are then added to $T_T$ with an upper bound constraint on the final size of $T_T$. In this manner the training set will gradually grow as choices are made on which compounds to test.

The updating procedure can be explained using a data set $T_D$ with examples ordered by testing date, the number of examples in $D$ ($n_D$), the number of examples in the initial training set $n_T$, the number of examples to predict in each iteration $n_P$, a choice function $F_c$ and $n_m$ the maximal number of examples that can be measured. Given these parameters the updating procedure presented in Algorithm 1, generates a training set $T_T$ and starts an iterative process of model building and subsequent prediction of new examples. The predicted examples are evaluated using the choice function. If the prediction of an example meets the criteria of the choice function, then the prediction is taken as sufficient. If the example fails the criteria of the choice function, then the example is learned and added to the training set.

### 2.1. Data

Data from four high throughput ADME assays routinely run for newly synthesized compounds, (Ballard et al., 2012) are considered in the present study: lipophilicity, solubility and metabolic stability in both human liver microsomes and rat hepatocytes. Lipophilicity is an important compound property influencing both potency and pharmacokinetic properties of a compound. Higher lipophilicity is associated, for example, with higher cell membrane

**Algorithm 1:** Updating Procedure

**Input:** $T_D, n_T, n_P, n_D, p$
**Output:**
$T_T \leftarrow \{z_0, \ldots, z_{n_T}\}, z \in T_D$;
**while** $n_T < n_D$ **do**
  $T_P \leftarrow \{z_{n_T+1}, \ldots, z_{n_T+n_P}\}, z \in T_D$;
  *build CP model on* $T_T$;
  *predict on* $T_P$;
  $T_T \leftarrow T_T \cup \{z_i, i \in [n_T + 1, \ldots, n_T + n_P]; If(F_c(z_i))\}$;
  $n_T \leftarrow n_T + n_P$;
**end**

permeability, lower water solubility, lower metabolic stability, higher unspecific binding to both plasma proteins and target proteins, higher volume of distribution and higher risk for binding to other targets besides the intended. logD values between 1 and 3 are usually considered as favourable. Solubility is of interest to estimate compound absorption but also to understand whether low solubility may impair other assay results. Lower Solubility, below 100 $\mu$M, will decrease a compounds chance to become a successful drug. Metabolic stability is an essential property to estimate a compounds pharmacokinetic (PK) properties and consequently the necessary therapeutic dose. At AstraZeneca both human liver microsomes and rat hepatocytes are routinely used as part of this assessment. The former determine the ability of human cytochrome P450 enzymes to break down a compound, whereas rat hepatocytes indicate if additional types of metabolism such as glucuronidations can occur. Intrinsic clearance values below 10$\mu$l/min/mg protein for microsomes and 10$\mu$l/min/million cells for hepatocytes are here accepted as sufficient. All assays follow standard procedures: shake-flask methods are used for lipophilicity (measured as n-octanol water distribution coefficient at pH 7.4, log D) (Wenlock et al., 2011) and solubility (Wan and Holmn, 2009). Metabolic stability in human liver microsomes and rat hepatocytes are determined as intrinsic clearance (CLint) from the compound disappearance rate in an incubation over 30 minutes and 2 hours, respectively (Sohlenius-Sternbeck et al., 2010; Temesi et al., 2010). Test compounds are supplied in form of a 10 mM solution in dimethylsulfoxid (DMSO) for all assays. In order to reduce the influence of the solvent for the solubility determination, DMSO is evaporated at assay start. The assays show good reproducibility over time, with the solubility assay being the most variable (Winiwarter et al., 2015). The data set consists of compounds measured over a period of five years and contains 53515 compounds in total. Provided that 3 out of 4 targets were in their desired range a compound would be categorized as `Good` else it would be categorized as `Bad`. The distribution between the `Good` and `Bad` classes are 29% and 71% respectively. This dataset is not publicly available which limits us to describe its exact nature. However, the compounds are described by structural chemical descriptors, so called signature fingerprints. The initial training set is 39084 compounds and the test set is 14431 compounds.

### 2.2. Strategies

In the following, several strategies are described that will be applied according to the updating procedure outlined in Algorithm 1. Each strategy will start with the set of initial training examples, $T_T$, and select to learn from subsequent objects by applying a choice function to the objects in the prediction set, $T_P$ with a lower bound constraint on the final size of $T_T$.

For all strategies, covered in this work, that produce a pair of p-values, these values are combined into a final single valued score based on the credibility and the confidence (Vovk et al., 2005) of the object. This score is

$$s = \begin{cases} -p_0(1-p_1)/2 + 0.5 & \text{if } p_0 > p_1, \\ p_1(1-p_0)/2 + 0.5 & \text{if } p_0 \leq p_1. \end{cases} \quad (1)$$

Here the indices 0 and 1 represent the labels `Bad` and `Good`, respectively. Selection of compounds was made using p-values and the score, $s$, was used for the evaluation.

#### 2.2.1. Random choice, Mondrian Cross-Conformal Prediction

This strategy is added as a reference strategy to be compared to the two following strategies. The underlying machine-learning algorithm is the $C$-SVC and a radial basis function kernel as defined in Scikit-learn version 0.17 (Pedregosa et al., 2011). Here, we are using a Mondrian, label-conditional cross conformal predictor. The division of the training examples is done by a stratified five fold sampling using the `cross_validation.StratifiedKFold` function in Scikit-learn, using one fold as calibration examples and the remaining examples as proper training examples. The predicted p-values from each of the five inductive conformal predictors are averaged for each prediction object, within each class label respectively. The nonconformity measures are defined by the decision function values of the underlying $C$-SVC model. The initial training examples are used to find the optimal values for the cost coefficient of the $C$-SVC, $C$, and the width of the radial basis function, $\gamma$, by applying a grid search over the values $C \in \{10^c : c = 3i/9 : i \in \{0, 1, \ldots, 9\}\}$ and $\gamma \in \{10^g : g = -3i/9 - 5 : i \in \{0, 1, \ldots, 9\}\}$ and selecting the pair that leads to the highest accuracy when using `GridSearchCV` in Scikit-learn and five folds. For this strategy, the choice function randomly selects a fraction of the predicted objects for each updating step to be learnt and subsequently added to the training set for the next iteration of the updating procedure. For example, when the constraint of learning from 80% of the objects is applied, the choice function will randomly select 20% of the objects and these objects will not be learnt at any stage in the following updating steps. This strategy is referred to as **random**.

#### 2.2.2. Largest p-values, Mondrian Cross-Conformal Prediction

The p-values for the different labels in this strategy are calculated in the same way as in Section 2.2.1. At each updating step, the choice function selects the fraction, as specified by the constraint, of objects with the largest predicted p-value of any class label not to be learnt. This corresponds to the prescribed reduction in synthesis. This strategy is referred to as **both**.

#### 2.2.3. Largest p-values bad labels, Mondrian Cross-Conformal Prediction

Also for this strategy, the p-values are calculated in the same way as in Section 2.2.1. The difference for this strategy is in the choice function. Now we form a subset of objects for which the p-value of the `Bad` class label is greater than the p-value for the `Good` class label. The selection is taken to be the largest fraction of `Bad` class label p-values regardless of the p-values for the `Good` class. This strategy would only consider to leave out objects based on one of the class p-values and thus introducing a more conservative view for the other class where many high credible predictions would be learnt. This strategy is referred to as **bad**.

#### 2.2.4. Random forest, Largest p-values, Mondrian Cross-Conformal Prediction

The p-values for the different labels in this strategy are calculated in the same way as in Section 2.2.2 using the FEST[1] Random Forest implementation with 100 trees instead of a Support Vector Machine.

#### 2.2.5. Random forest, Mondrian out-of-bag calibration

The underlying model is a random forest with 500 trees using a depth-limit of 500 levels and 200 randomly selected features, which is used together with a Mondrian (class-conditional) approach to calculate p-values using out-of-bag instances for calibration; the other parameters are set according to the default of the used Julia implementation[2].

The original training set is used to construct a first model and different *selection strategies* are then employed for deciding what examples to synthesize and predict for each batch of new examples, where synthesized examples are merged with the original training set when training subsequent models.

The three considered selection strategies are:

**random** Randomly divide the new batch into examples to be synthesized and predicted.

**confidence** Select examples with the highest p-value (for any of the two classes) for prediction and the remaining (low-confidence) examples for experimental testing, to learn the true label.

**half-each** Select half of the examples for prediction based on the highest scores according to Eq. 1 and half of the examples according to the lowest scores.

#### 2.2.6. Random forest, Differences in confidence

A class-conditional conformal classifier is trained using a random forest classifier, using 500 trees and a max depth of 500. Training and calibration data consists of all previously synthesized compounds, from all previous batches. For each batch, obtain $p_{good}$ and $p_{bad}$ for all test objects. For each batch, we expect to synthesize $x\%$ of compounds.

1. http://lowrank.net/nikos/fest
2. http://github.com/henrikbostrom/RandomForest

**diff** Calculate $p_{diff} = p_{good} - p_{bad}$ for each test object. Reveal labels for all test objects where $p_{diff} < percentile(p_{diff}, x)$, and output predictions for the remaining test objects. Pro: test objects belonging to the *Good* class are either synthesized or obtain a prediction where the *Bad* class has a low confidence. Con: *Bad* objects with high confidence predictions for the *Good* class are not revealed; i.e., many *Good* predictions are *Bad* objects, and we do not learn to distinguish these by not revealing the labels.

**absdiff** Calculate $p_{absdiff} = |p_{good} - p_{bad}|$ for each test object. Reveal labels for all test objects where $p_{absdiff} < percentile(p_{absdiff}, x)$, and output predictions for the remaining test objects. Pro: most *Good* objects are either synthesized or obtain a prediction with a high confidence for the *Good* class and a low confidence for the *Bad* class. Con: same as **diff**; additionally, some *Good* objects obtain high confidence predictions for the *Bad* class (and low confidence for the *Good* class).

**rand** For each test object calculate

$$w_i = 1 - m_i + \beta, \tag{2}$$

where $m_i$ is either **diff** or **absdiff** for the test object, and $\beta$ is a parameter. Normalize the weights using

$$w_i = \frac{w_i}{\sum_{j=1}^{n} w_j}. \tag{3}$$

Randomly reveal the true labels for $x\%$ of test objects using $w_1, \ldots, w_n$ as a probability distribution function.

#### 2.2.7. Gradient Boosted Tree Classifier, Largest p-values, Mondrian Conformal Prediction

A Mondrian class conditional conformal gradient boosted classifier is trained on all available and predicted synthesized compounds, i.e., from all previous batches, where the predicted labels are assumed to be correct. The base classifier "GradientBoostingClassifier" (`http://scikit-learn.org`), version 0.17 (Pedregosa et al., 2011) with all parameters at default was used for model building. The original number of attributes (105781) was filtered down to attributes that occured at least 30 times in the training set. After filtering 7988 attributes remained for modeling. At each updating step, the choice function selects the fraction, as specified by the constraint, of objects with the largest predicted p-value of the `Good` class label to be synthesized. The risk of hurting performance by systematically including low-confidence compounds is hopefully alleviated by including high-confidence predictions.

## 3. Results

In the following, the performance metric, AUC, has been calculated only on the examples that the different strategies did not learn from and using the s score defined in Section 2.2. The predictions on all objects that were not selected for learning have been compared to their true labels through the calculation of AUC using the `roc` function (Robin et al., 2011).

### 3.1. Mondrian Cross-Conformal Predictors

The results for the Mondrian cross-conformal predictors refers to the strategies presented in Sections 2.2.1,2.2.2 and 2.2.3. The results in Table1 show that the **both** strategy achieves the highest AUC for all constraints. It is also noticeable that the **bad** strategy performs worse than the **random** strategy.

Table 1: AUC for Mondrian cross-conformal predictors

| Predict fraction<br>Strategy | **0.20** | **0.40** | **0.60** |
|---|---|---|---|
| **random** | 0.7967 | 0.7963 | 0.7964 |
| **both** | **0.9041** | **0.8749** | **0.8424** |
| **bad** | 0.5961 | 0.6591 | 0.6875 |

In Table 2, the results for using random forests with Mondrian cross-validation-calibration for the three considered fractions of each batch that are to be predicted, i.e., 0.2, 0.4 and 0.6, are shown. The results are similar to those, for the same strategy, using SVM instead of RF.

Table 2: AUC for Mondrian cross-conformal predictors using FEST RF

| Predict fraction<br>Strategy | **0.20** | **0.40** | **0.60** |
|---|---|---|---|
| **both** | 0.9059 | 0.8529 | 0.7846 |

### 3.2. Mondrian out-of-bag calibration

In Table 3, the results for using random forests with Mondrian out-of-bag-calibration together with the selection strategies *random*, *confidence*, *half-each* and *absdiff* (as described in Sections 2.2.5 and 2.2.6, respectively) or the three considered fractions of each batch that are to be predicted, i.e., 0.2, 0.4 and 0.6.

Table 3: AUC for random forests with Mondrian out-of-bag calibration

| Predict fraction<br>Strategy | **0.20** | **0.40** | **0.60** |
|---|---|---|---|
| **random** | 0.7585 | 0.7527 | 0.7586 |
| **confidence** | 0.8550 | 0.8400 | 0.8004 |
| **half-each** | 0.8453 | 0.8222 | 0.7935 |
| **absdiff** | **0.8627** | **0.8430** | **0.8008** |

### 3.3. Differences in confidence

In table 4, the results for the strategies based on differences in confidence are presented, for the three considered fractions to be predicted. The **absdiff** strategy performs the best.

Table 4: AUC for gradient boosted trees

| Predict fraction<br>Strategy | **0.20** | **0.40** | **0.60** |
|---|---|---|---|
| **absdiff** | **0.9202** | **0.8727** | **0.8363** |
| **diff** | 0.6146 | 0.6529 | 0.6729 |

### 3.4. Gradient Boosted Tree Classifier

In Table 5 the results from using a gradient boosted tree classifier (GBT) for the three fractions are shown.
The performance of GBT, with respect to AUC, is not as good as some of the Mondrian cross-conformal or Random Forest results.

Table 5: AUC for gradient boosted trees

| Predict fraction | **0.20** | **0.40** | **0.60** |
|---|---|---|---|
| **GBT** | 0.6561 | 0.6376 | 0.6508 |

## 4. Concluding remarks

The possibility to reduce the number of synthesised compounds in early stages of drug discovery discovery has been evaluated using strategies for automated compound prioritization based on early ADME data. The results show that there is a clear possibility to reduce the number of synthesized compounds with minimal reduction in the quality of data that decisions are taken on. The gold standard in this work is the assay data, hence measuring all compounds would return an AUC of 1.0. In that respect, obtaining AUC values for the proposed methods above 0.9 is very promising, like in Tables 4 and Table 1. The results show that it is possible to reduce testing by 20% and significantly lower the development costs for projects. One of the drivers behind this work is cost and under a limited budget these methods can be used instead of assay data for early decisions. For later stages of candidate drug selection it is advised to generate the assay data if it is needed.

The results in this work are very encouraging and any future work in establishing better strategies with potential theoretical support would be very useful in the process of discovering potential drugs.

In future work, it will be important to more carefully define the problem statement and to review appropriate measures of success. The benefits of using AUC as performance metric, in contrast to e.g., lift, accuracy, precision or recall, are that it does not require any threshold to be set for determining class membership for the predicted instances and that it

is insensitive to differences between the underlying class distributions from which training and test instances are sampled. However, similar to the other standard performance metrics, the AUC does not incorporate misclassification costs, which in the considered domain typically are distributed unevenly, e.g., the cost of incorrectly predicting a bad compound as good is often much higher than for the opposite type of error. An evaluation of the investigated approaches with respect to performance metrics that do include misclassification costs, as well as the cost of obtaining labels through synthesis and testing, is hence one of the most important directions for extending this work.

## Acknowledgments

This work was supported by the Swedish Knowledge Foundation through the project Data Analytics for Research and Development (20150185).
The research at Swetox (UN) was supported by Stockholm County Council, Knut & Alice Wallenberg Foundation, and Swedish Research Council FORMAS.

## References

Ernst Ahlberg, Ola Spjuth, Catrin Hasselgren, and Lars Carlsson. Interpretation of conformal prediction classification models. In *International Symposium on Statistical Learning and Data Sciences*, pages 323–334. Springer International Publishing, 2015.

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.

A. M. Davis and D. J. Wood. Quantitative structure-activity relationship models that stand the test of time. *Mol Pharm*, 10(4):1183–90, 2013. doi: 10.1021/mp300466n.

Martin Eklund, Ulf Norinder, Scott Boyer, and Lars Carlsson. The application of conformal prediction to the drug discovery process. *Annals of Mathematics and Artificial Intelligence*, 74(1-2):117–132, 2015.

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.

Ulf Norinder, Lars Carlsson, Scott Boyer, and Martin Eklund. Introducing conformal prediction in predictive modeling. a transparent and flexible alternative to applicability domain determination. *Journal of chemical information and modeling*, 54(6):1596–1603, 2014.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Alleyn T. Plowright, Craig Johnstone, Jan Kihlberg, Jonas Pettersson, Graeme Robb, and Richard A. Thompson. Hypothesis driven drug design: improving quality and effectiveness of the design-make-test-analyse cycle. *Drug Discovery Today*, 17:56–62, 2012. doi: 10.1016/j.drudis.2011.09.012.

Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frdrique Lisacek, Jean-Charles Sanchez, and Markus Mller. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77, 2011.

Anna-Karin Sohlenius-Sternbeck, Lovisa Afzelius, Peteris Prusis, Jan Neelissen, Janett Hogstraate, J. Johansson, Eva Floby, A. Bengtsson, O. Gissberg, J. Sternbeck, and Carl Petersson. Evaluation of the human prediction of clearance from hepatocytes and microsome intrinsic clearance for 52 drug compounds. *Xenobiotica*, 40(9):637–649, 2010.

David G. Temesi, Scott Martin, Robin Smith, Christopher Jones, and Brian Middleton. High-throupghput metabolic stability studies in drug discovery by orthogonal acceleration time-of-flight (oatof) with analogue-to-digital signal capture (adc). *Rapid Communications in Mass Spectometry*, 24:1730–1736, 2010.

Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005. ISBN 0387001522.

Hong Wan and Anders G. Holmn. High throughput screening of physiochemical properties and in vitro adme profiling in drug discovery. *Combinatorial Chemistry & High Throughput Screening*, 12:315–329, 2009.

Mark C. Wenlock, Tim Potter, Patrick Barton, and Rupert P. Austin. A method for measuring the lipophilicity of compounds in mixtures of 10. *Journal of Biomolecular Screening*, 16:3, 2011.

S. Winiwarter, B. Middleton, B. Jones, P. Courtney, B. Lindmark, K. M. Page, A. Clark, and C. Landqvist. Time dependent analysis of assay comparability: a novel approach to understand intra- and inter-site variability over time. *J Comput Aided Mol Des*, 29(9): 795–807, 2015. doi: 10.1007/s10822-015-9836-5.