# Assessment of the complementarity of machine learning methods in QSAR modeling using AZOrange

Jonna Stålring, Pedro Almeida and Scott Boyer
Safety Assessment, AstraZeneca R&D, Mölndal, Sweden

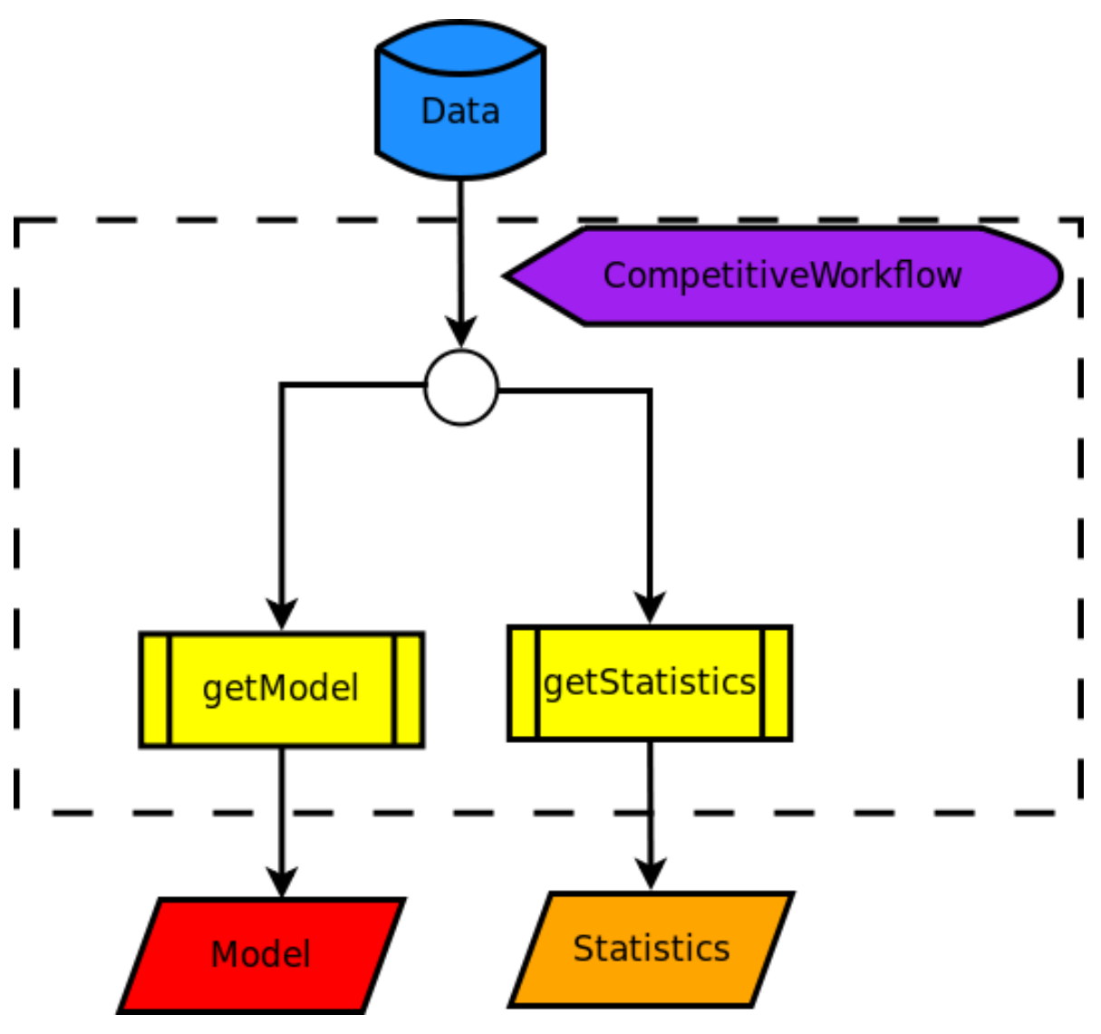
## Introduction

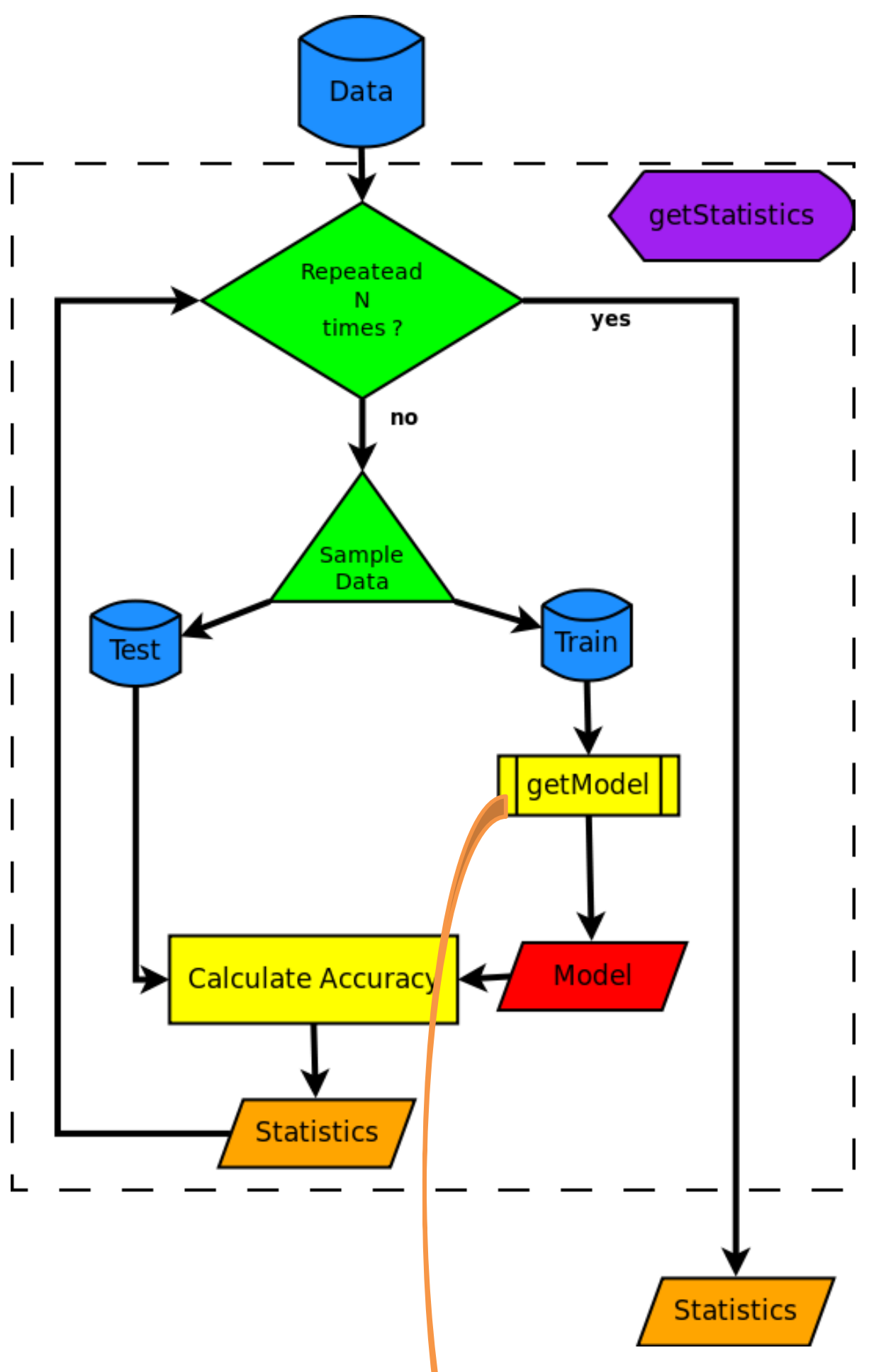
A multitude of machine learning algorithms are suitable for QSAR modeling and they rely upon various conceptual foundations. In general, no method can be identified as superior and an efficient data specific choice of modeling algorithm has the potential of increasing accuracy beyond what is possible with a single algorithm.

The AZOrange machine learning platform interfaces several algorithms for QSAR modeling. A process has been developed for automated assessment of the accuracy of all the AZOrange algorithms and consecutive automated model building and selection. To avoid overfitting and overestimation of the generalization accuracy, potentially resulting from any elaborate selection process, extensive re-sampling of external test sets is used. The accuracy is complemented by the accuracy variance between the folds in the assessment of model quality. The process includes automated construction of a consensus model, weighting the constituting models by their global accuracy.

## Method

The top level process accepts a data set and returns the selected AZOrange model together with extensively re-sampled validation statistics where no selection of algorithm or model hyper-parameters have been performed based on the validation sets.



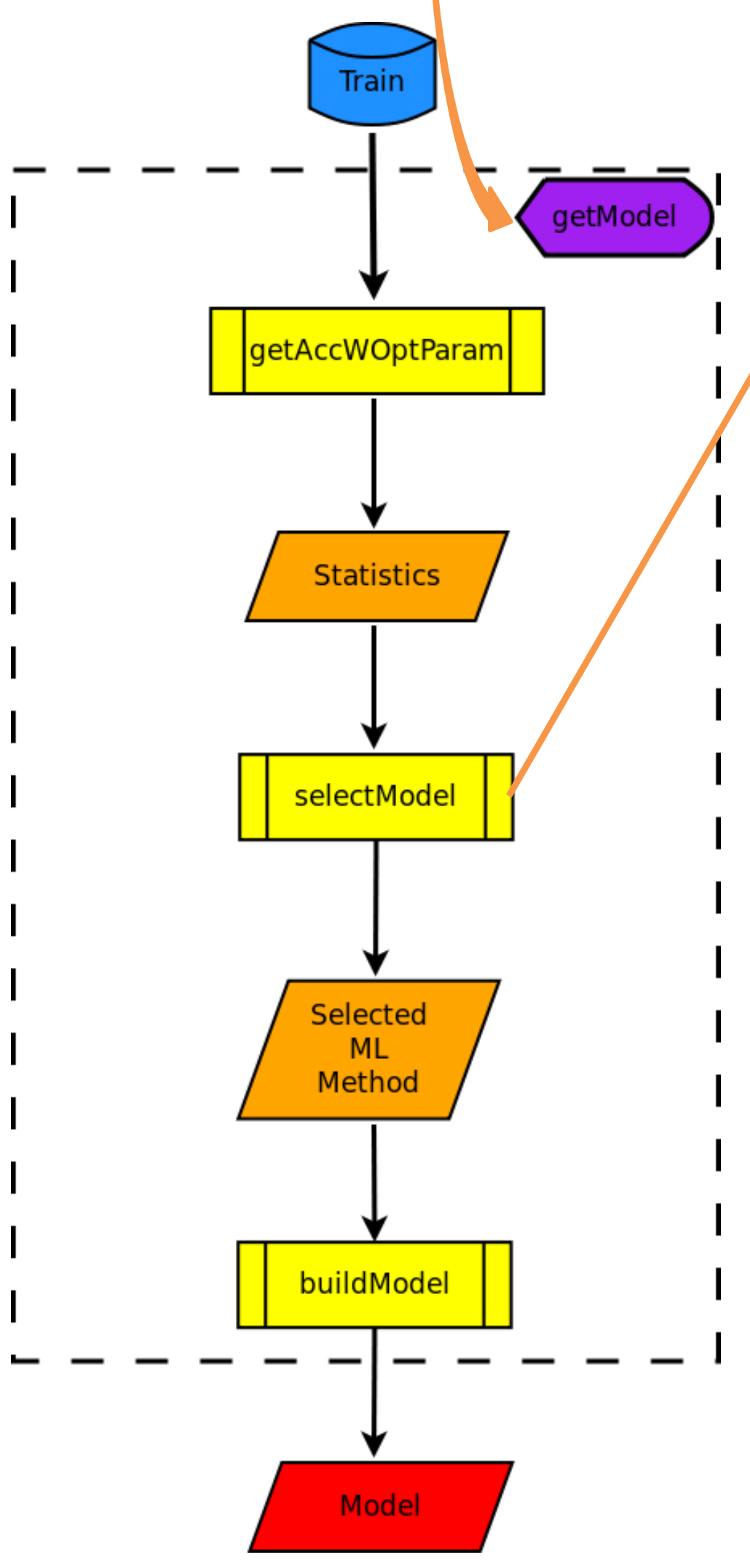


- The **getStatistics** process assesses the generalization accuracy expected while using the getModel method to automatically build the most accurate QSAR model for a given data set.

- The statistics contains accumulated results over all folds, as well as results from the individual folds.

- The **getModel** workflow automatically selects an ML algorithm, based on accuracy and stability, the most appropriate for a specific data set.

**Selection criteria:**
**1) Select the method with the highest Q2/CA amongst the stable methods**
**2) If no method is considered stable, select the method with the greatest Q2/CA**



- The accuracy is assessed in an outer loop within which no model hyper-parameter selection is performed, as displayed in the **getUnbiasedAccuracy** diagram.

- A model is considered stable if the variation in accuracy (Q2 or CA) over the folds is small (0.1/0.2 - data size and response type dependent).

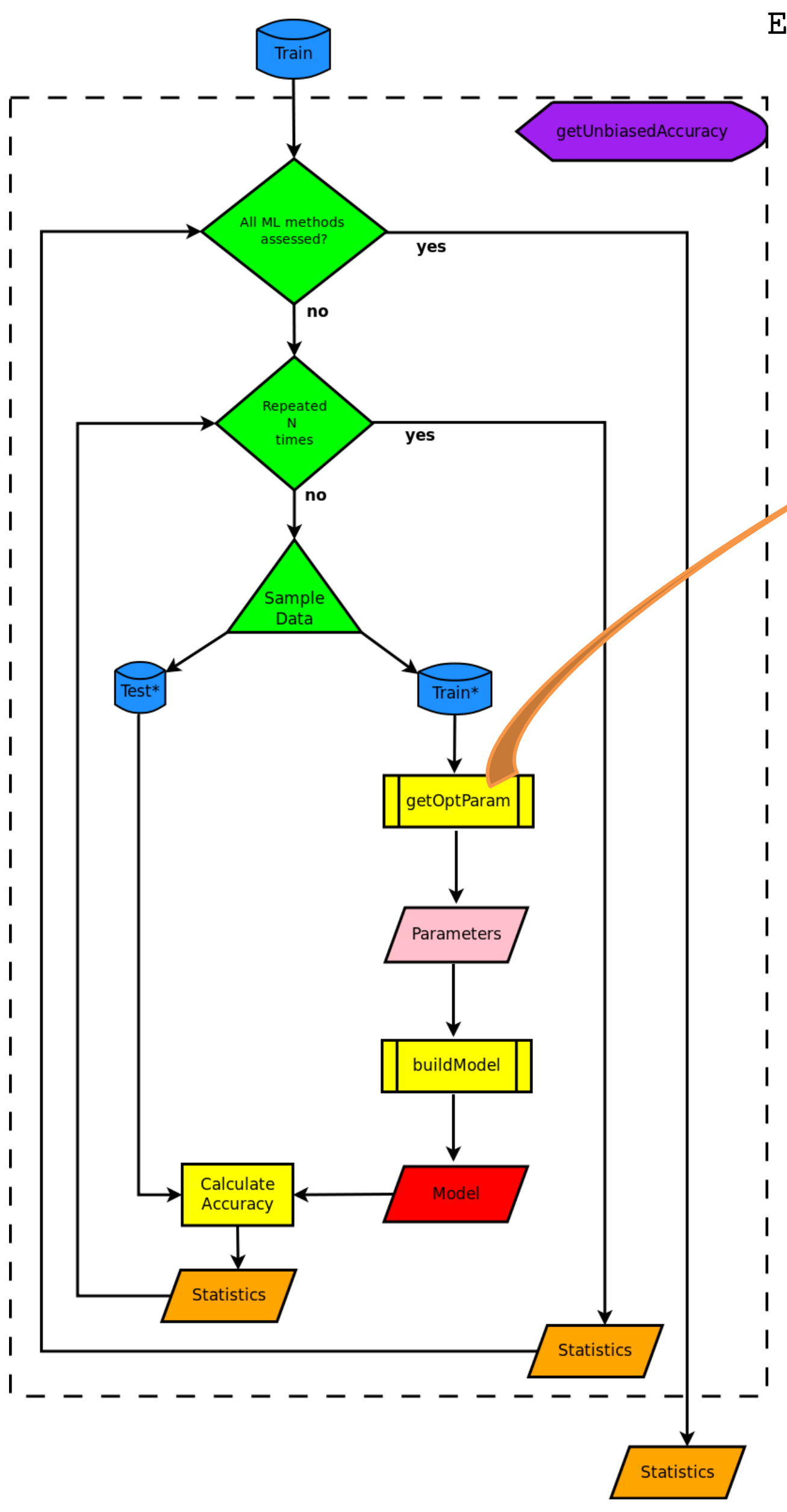
- A consensus learner is created using all stable methods:

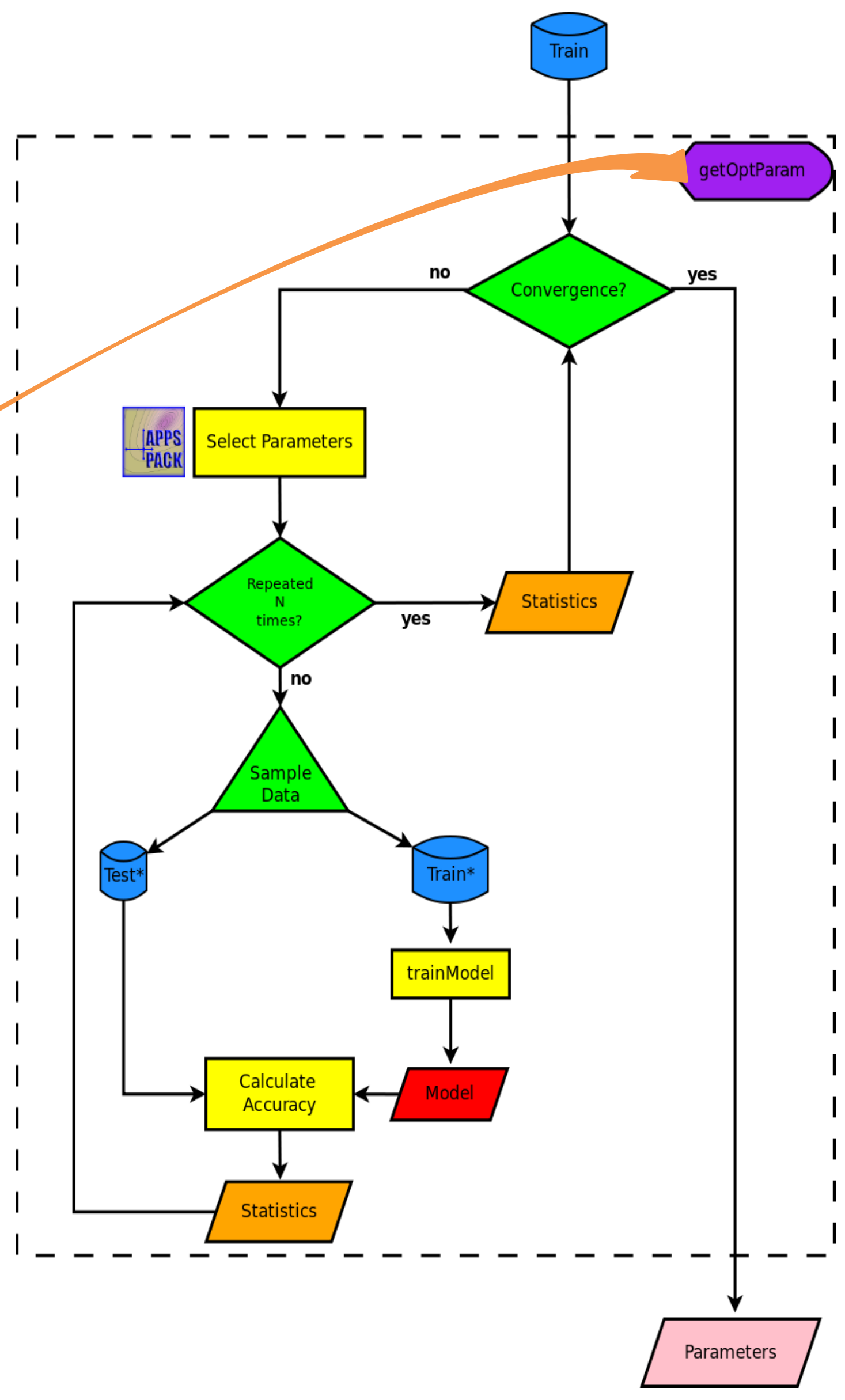
- Consensus Prediction Regression (all methods stable)

$$\mathrm{Pred}_{\mathrm{Cons}} = \frac{1}{\mathrm{Q2}_{\mathrm{sum}}}(\mathrm{Q2}_{\mathrm{ANN}} * \mathrm{Pred}_{\mathrm{ANN}} + \mathrm{Q2}_{\mathrm{SVM}} * \mathrm{Pred}_{\mathrm{SVM}} + \mathrm{Q2}_{\mathrm{RF}} * \mathrm{Pred}_{\mathrm{RF}} + \mathrm{Q2}_{\mathrm{PLS}} * \mathrm{Pred}_{\mathrm{PLS}})$$

- Consensus Decision Binary Classification

```
Avg(CA)POS = Average CA of all learners predicting POS
If  Avg(CA)POS >= Avg(CA)NEG; Predict POS
Else; Predict NEG
```

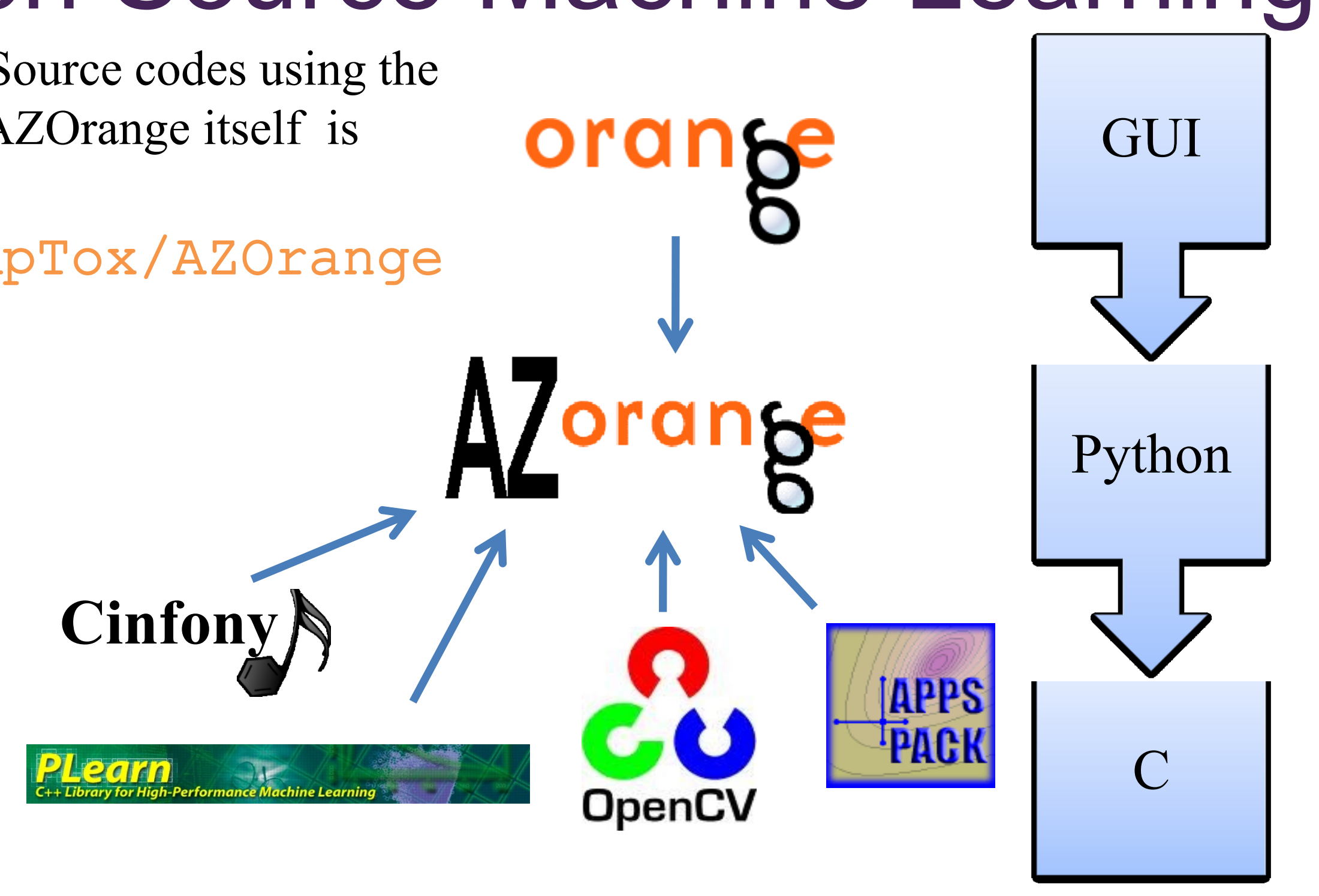




## AZOrange - Open Source Machine Learning

- AZOrange integrates several Open Source codes using the Orange platform as a framework and AZOrange itself is available as an Open Source package: https://github.com/AZCompTox/AZOrange

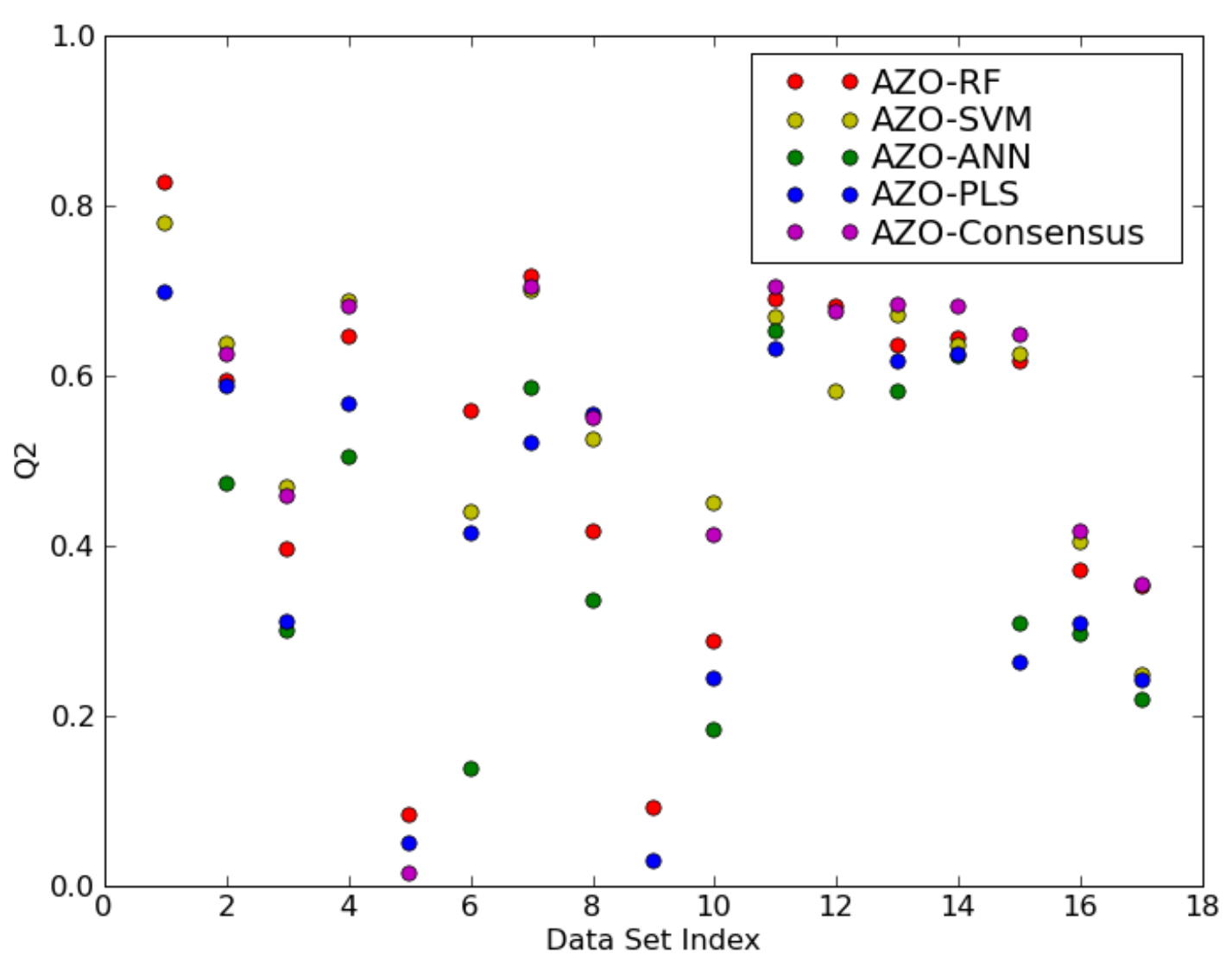
- The AZOrange methods are customized for automated model development. In particular, model hyper-parameters are automatically selected using the pattern search algorithm in APPSPACK.



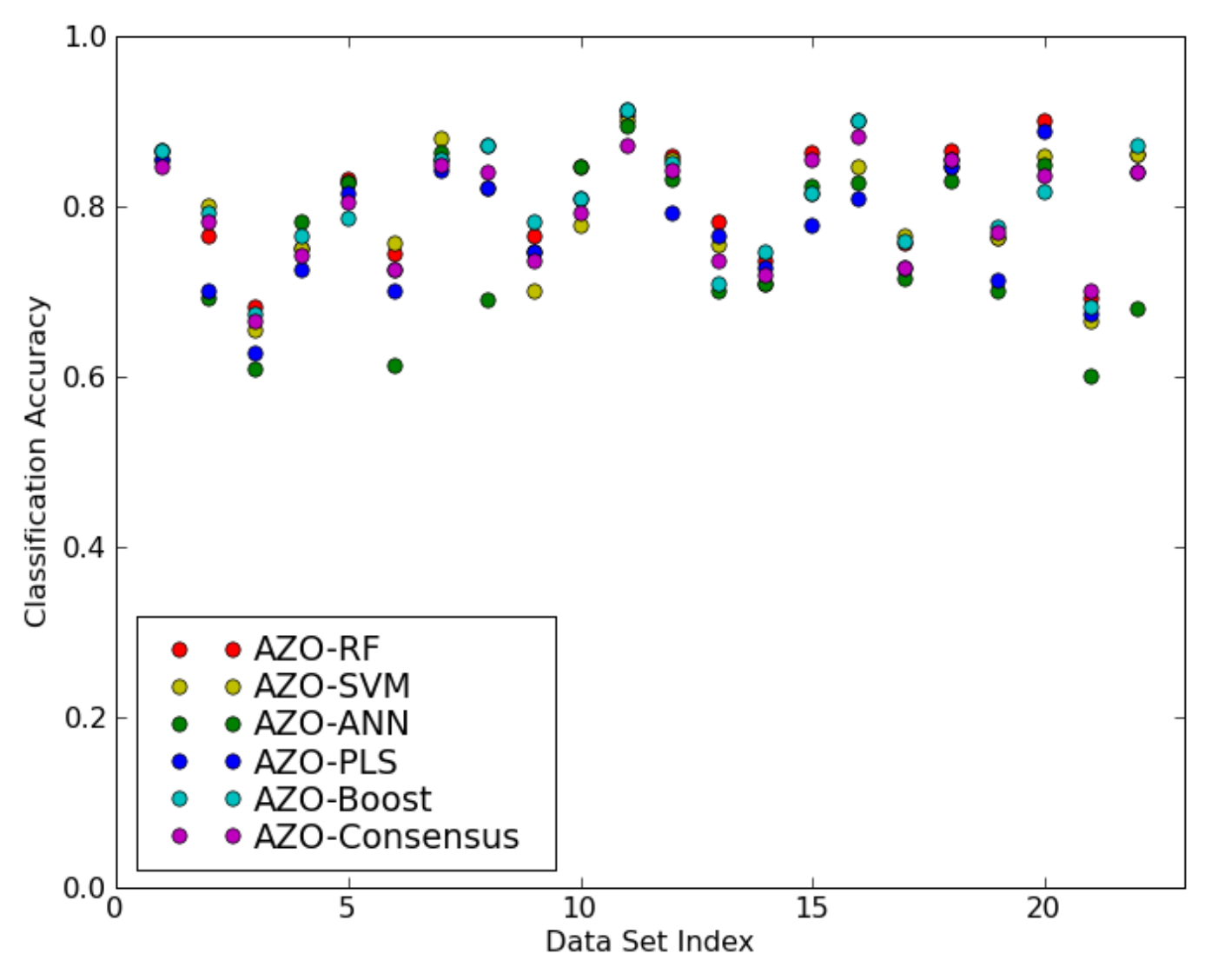
## Data sets

| Origin | Endpoints | Size | Response dimension |
|---|---|---|---|
| **PubChem** | | | |
| 10 Bioactivities | AhR, NF-kB, Thyroid, hERG, KCNQ2, PPAR, ER, M1, STAT3, STAT1 | 1500 – 10 000 | Regression and Classification (balanced) |
| **Congeneric Series** | | | |
| 12 Targets | ACE, ACHE, AMPH1, BZR, COX2, DHFR, EDC, HIVPR, HIVRT, HPTP, THERM, THR | 100 - 400 | Regression and Classification |
| **EPA** | Long Term Carcinogenicity | 1069 | Classification |
| **ChemInformatics web page** | ER, Melting point, Solubility | 1000 - 4000 | Regression and Classification |
| **FDA** | Max Recommended Daily Dose | 1200 | Regression |

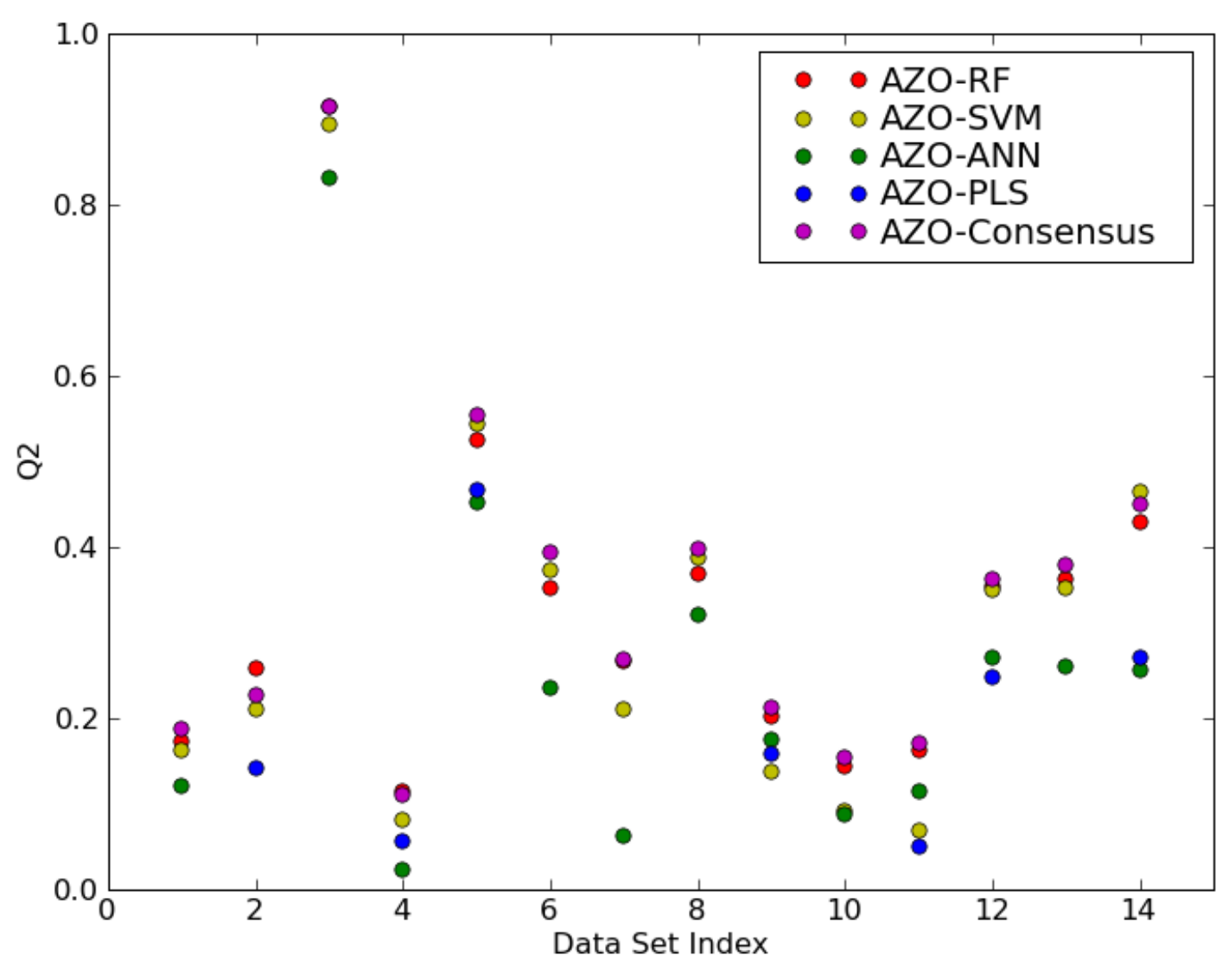
**Table 1. Accuracy of the individual model on the external test sets**



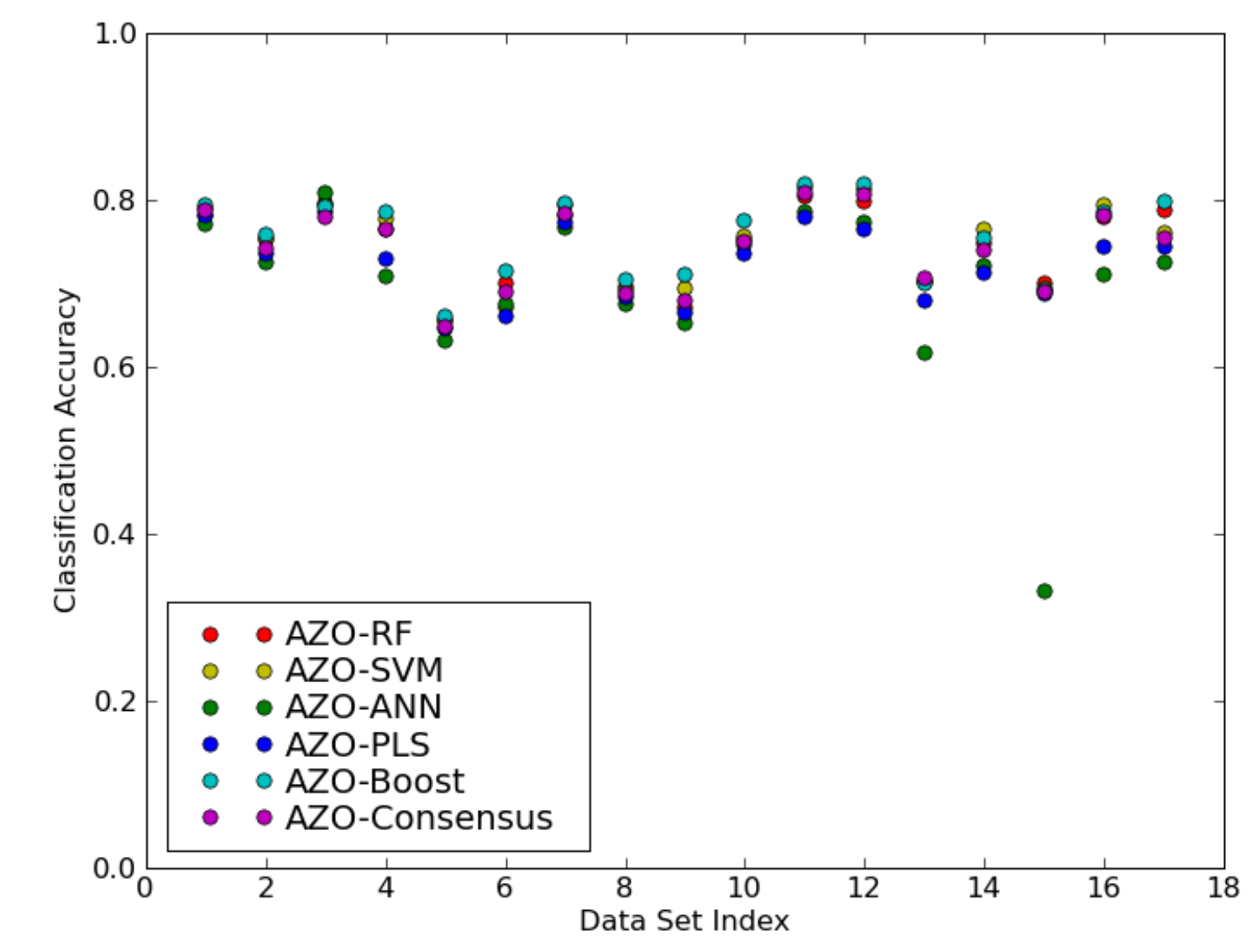
Small Congeneric Data Regression



Small Congeneric Data Classification



Large Global Data Regression



Large Global Data Classification

## Descriptors

- Chemical structure was represented using either the set of 177 physio-chemical descriptors of RDkit or by circular fingerprints as implemented in RDkit with radius 1.

## Results

- The results include the data sets for which at least one model was stable.

- The global regression sets almost exclusively favor the consensus learner.

- The boosted tree algorithm is selected for most of the global classification data sets.

- The greater the rank sum (Table2) the more accurate the learner.

- The Bonferroni-Dunn test assesses the significance of the differences in rank sum between the learners.

**Table 2. The rank sums and the number of times each method was selected**

| | | RF | SVM | ANN | PLS | Boost | Consensus | Sign Diff (95%) | Sign Diff (90%) |
|---|---|---|---|---|---|---|---|---|---|
| CONGENERIC | **Classification Rank Sum** | 4.98 | 3.52 | 2.45 | 2.64 | 4.32 | 3.09 | 1.45 | 1.31 |
| | **Times Selected** | 9 | 4 | 2 | 1 | 6 | 1 | | |
| | **Regression Rank Sum** | 3.82 | 3.65 | 1.41 | 2.35 | NA | 3.76 | 1.35 | 1.21 |
| | **Times Selected** | 6 | 4 | 0 | 1 | NA | 6 | | |
| GLOBAL | **Classification Rank Sum** | 4.06 | 4.65 | 1.65 | 1.70 | 5.41 | 3.53 | 1.65 | 1.49 |
| | **Times Selected** | 1 | 2 | 1 | 0 | 12 | 1 | | |
| | **Regression Rank Sum** | 3.93 | 3.14 | 1.86 | 1.34 | NA | 4.71 | 1.49 | 1.33 |
| | **Times Selected** | 2 | 1 | 0 | 0 | NA | 11 | | |

- The **getUnbiasedAccuracy** chart shows the assessment of the generalization accuracy expected when using getOptParam for automated model hyper-parameter selection.

- **getOptParam** performs numerical optimization of the model hyper-parameters using the pattern search algorithm of APPSPACK.

- The objective function used in the parameter selection is the generalization accuracy in a 5 fold cross validation.

- The model hyper-parameters to be optimized are empirically identified, as well as their ranges.

## Conclusions

- The process developed automatically makes a data sets specific selection of machine learning algorithm and performs extensive external validation.

- No algorithm can be identified as superior as many of the methods are frequently selected as the most accurate.

- For the global data sets the automatically constructed consensus model and the boosted tree algorithm are almost exclusively selected for the regression and classification data sets, respectively.

## Acknowledgements

We would like to acknowledge the Open Source community and in particular the Artificial Intelligence Laboratory at the University of Ljubljana, without which this project had not been possible. In addition, tribute should be given to Philip Werner at AstraZeneca who implemented the generalized consensus learner.

AstraZeneca