Towards Evolutionary-based Automated Machine Learning for Small Molecule Pharmacokinetic Prediction

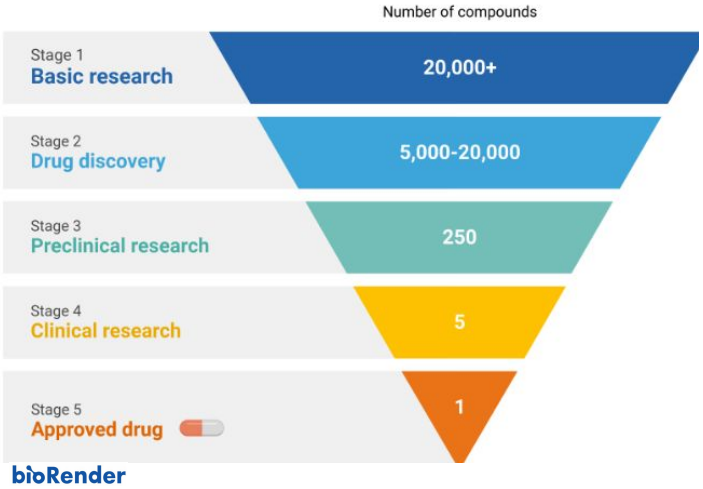
Alex G. C. de Sá
David B. Ascher

Background - Small Molecule Research in Drug Discovery

- Small molecules are low molecular weight (900–1,000 Daltons) organic compounds.
  - caffeine, roundup, aspirin, and paracetamol.

- Small molecule research plays a significant role in drug development and discovery:
  - **90% of currently marketed drugs**.

- Drug discovery is a costly, time-consuming and uncertain endeavour:
  - 12-15 years, on average.
  - Exceeds $ 2.5 billion.
  - 80-90% of projects are discontinued before getting tested in humans.

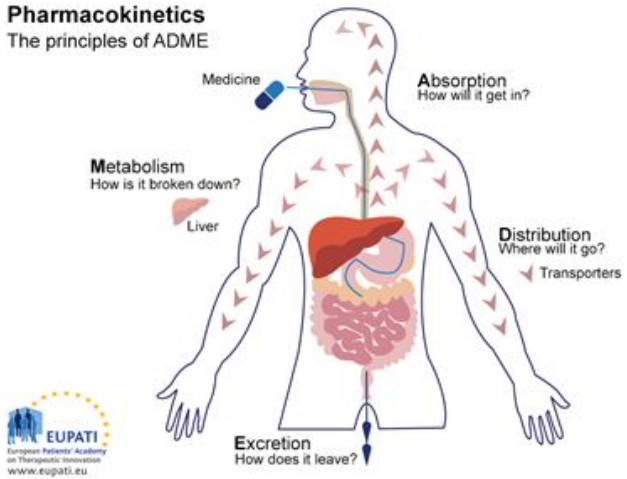


Background - Small Molecule Research in Drug Discovery

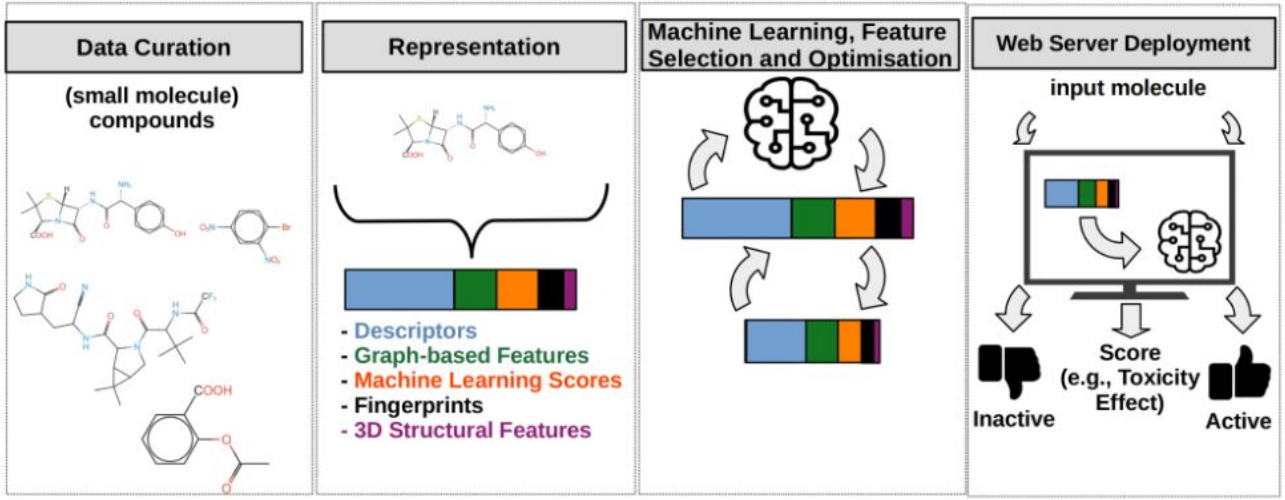
- Effectiveness *versus* compound's "fitness":
  - Although a compound can be considered good for a particular case, it might have issues on properties in later stages of its development.
  - **Pharmacokinetics: A**bsorption, **D**istribution, **M**etabolism and **E**xcretion (**ADME**) of the molecule in the (human) organism.

- Cost of development:
  - ADME studies are expensive (*In vivo* or *in vitro*).

- Computational models to assist in this scenario:
  - Virtual (pre-)screening of the compounds.
  - **Machine learning approach.**



Background - Typical Methodological Predictive Pipeline



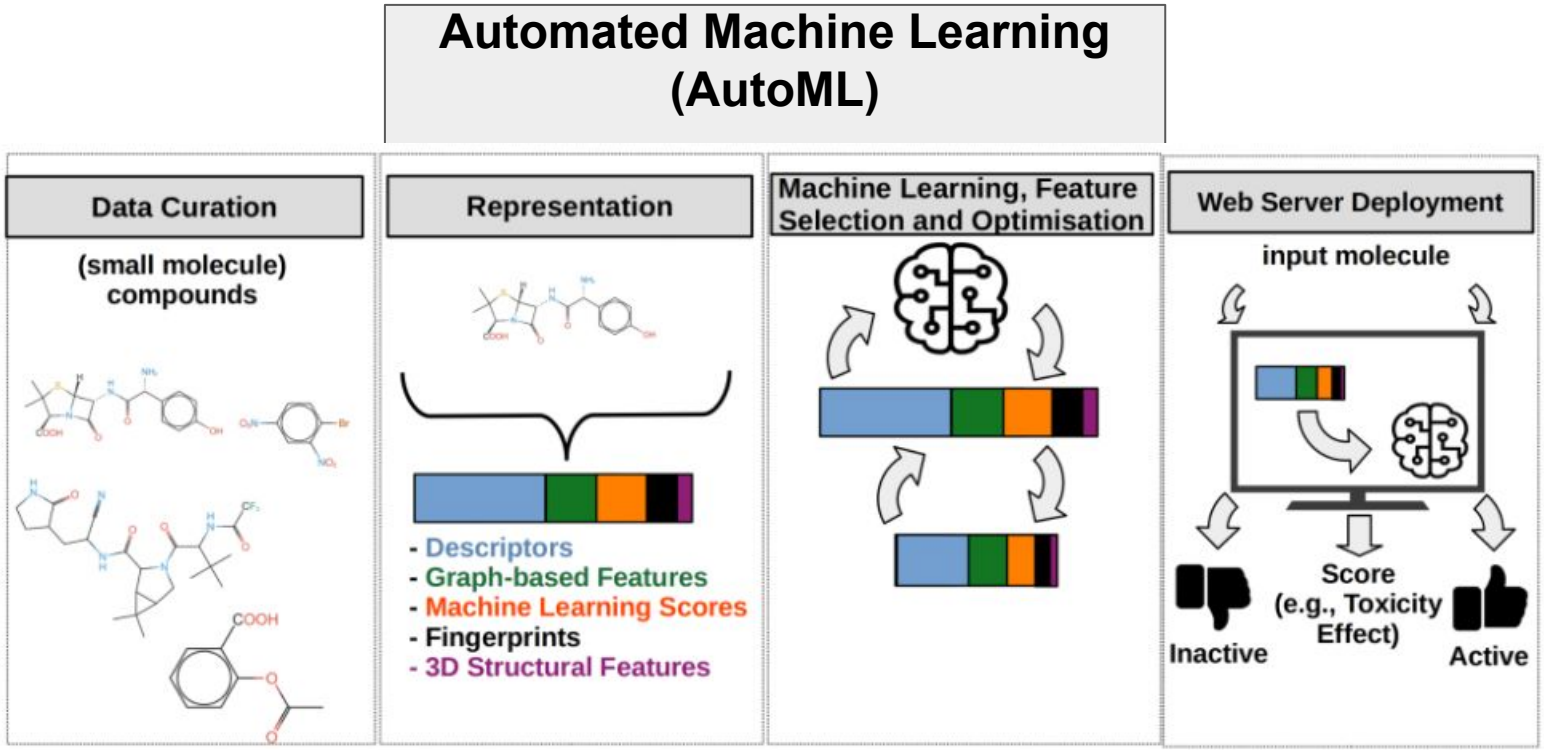
Pros:

- Good predictive performance on small molecule pharmacokinetics properties.
- Web-based tools:
  - Easy-to-use.
  - Fast at performing predictions.

Cons:

- Bias on decisions.
- Static models.
- Not personalised to the researcher/company's data.

Background - Proposed Pipeline



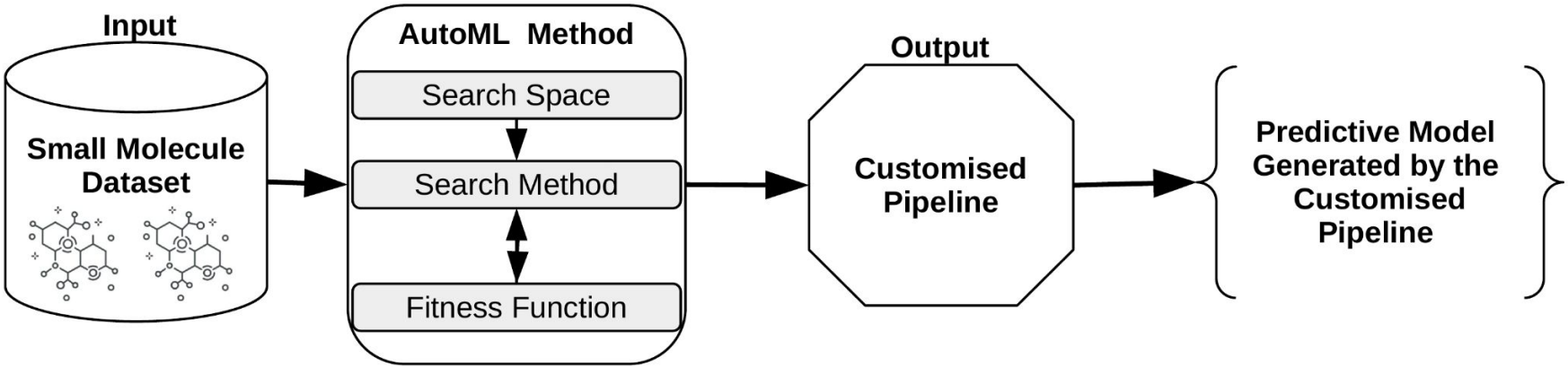
Related Work on AutoML for Pharmacokinetic (PK) Prediction

- Traditional Evolutionary-based AutoML Methods:
  - Tree-Based Pipeline Optimization Tool (TPOT).
  - RECIPE (REsilient ClassifIcation Pipeline Evolution).
  - Automated MEKA (Auto-MEKA).

- Non-evolutionary-based AutoML for Small Molecule Property Prediction:
  - **AutoQSAR:** exhaustive searches based on accuracy to rank the ML pipelines.
  - **ZairaChem:** uses 5 AutoML methods independently, aiming to specify different ML aspects to the task.
  - **Uni-QSAR:** employs stacking to ensemble ML models and predict molecular properties
  - **Qptuna:** applies Bayesian optimisation for searching and optimising ML pipelines in the context of molecule property prediction.

AutoML for Pharmacokinetic (PK) Prediction

First evolutionary-based AutoML method in the context of PK Prediction



Search Space

- Main ML building blocks for PK prediction:
  - Representation (5) - 31 combinations
  - Feature Scaling (5)
  - Feature Selection (5)
  - ML Modelling (6)
  - Hyper-parameters

- Context-Free Grammar (CFG) to defined this search space:
  - 25 (non-redundant) production rules
  - 24 non-terminals
  - 317 terminals

<Start> ::= <feature_definition> [<feature_scaling>] [<feature_selection>] <ML_algorithms>

<feature_definition> ::= General_Descriptors | Advanced_Descriptors | Graph-based_Signatures | Toxicophores | Fragments | General_Descriptors Advanced_Descriptors | General_Descriptors Graph-based_Signatures | ... | General_Descriptors Advanced_Descriptors Graph-based_Signatures Toxicophores Fragments

<feature_scaling> ::= <Normalizer> | <MinMaxScaler> | <MaxAbsScaler> | <RobustScaler> | <StandardScaler>

<Normalizer> ::= Normalizer <norm>
<norm> ::= l1 | l2 | max

...

<StandardScaler> ::= StddScaler <with_mean> <with_std>
<with_mean> ::= True | False
<with_std> ::= True | False

<feature_selection> ::= <Variance_Threshold> | <Select_Percentile> | <SelectFPR> | <SelectFWE> | <SelectFDR>

<Variance_Threshold> ::= VarianceThreshold <threshold>
<threshold> ::= 0.0 | 0.05 | 0.10 | 0.15 | ... | 0.85 | 0.90 | 0.95 | 1.0

...

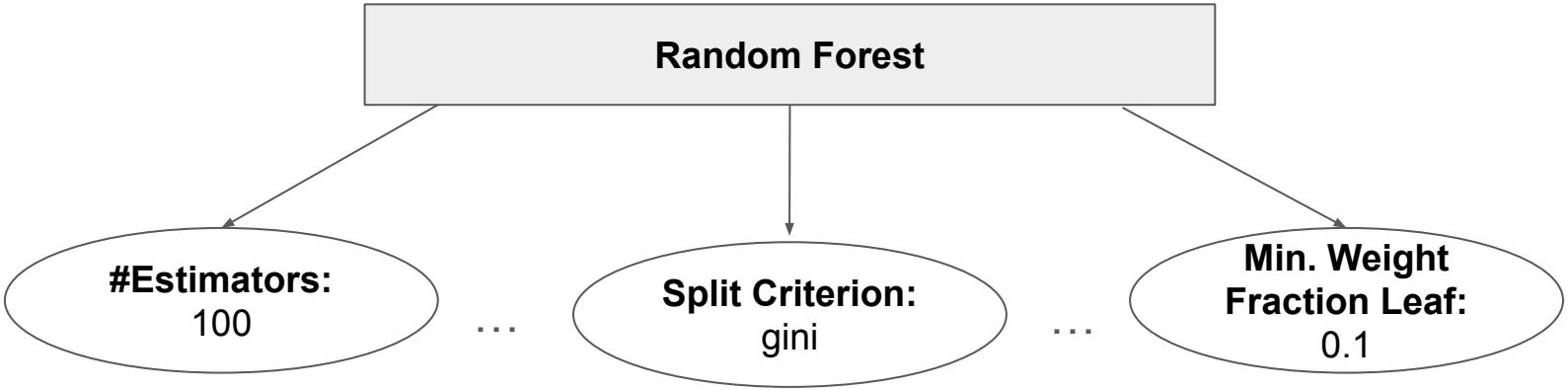
<ML_algorithms> ::= <AdaBoost> | <DecisionTree> | <ExtraTree> | <RandomForest> | <ExtraTrees> | <XGBoost>

<AdaBoost> ::= AdaBoost <algorithm> <n_estimators> <learning_rate>
<algorithm> ::= SAMME.R | SAMME
<n_estimators> ::= 5 | 10 | 15 | 20 | ...| 300 | 500 | 550 | ... | 950 | 1000 | 1500 | 2000 | 2500 | 3000
<learning_rate> ::= 0.01 | 0.02 | 0.03 | ... | 2.0

...

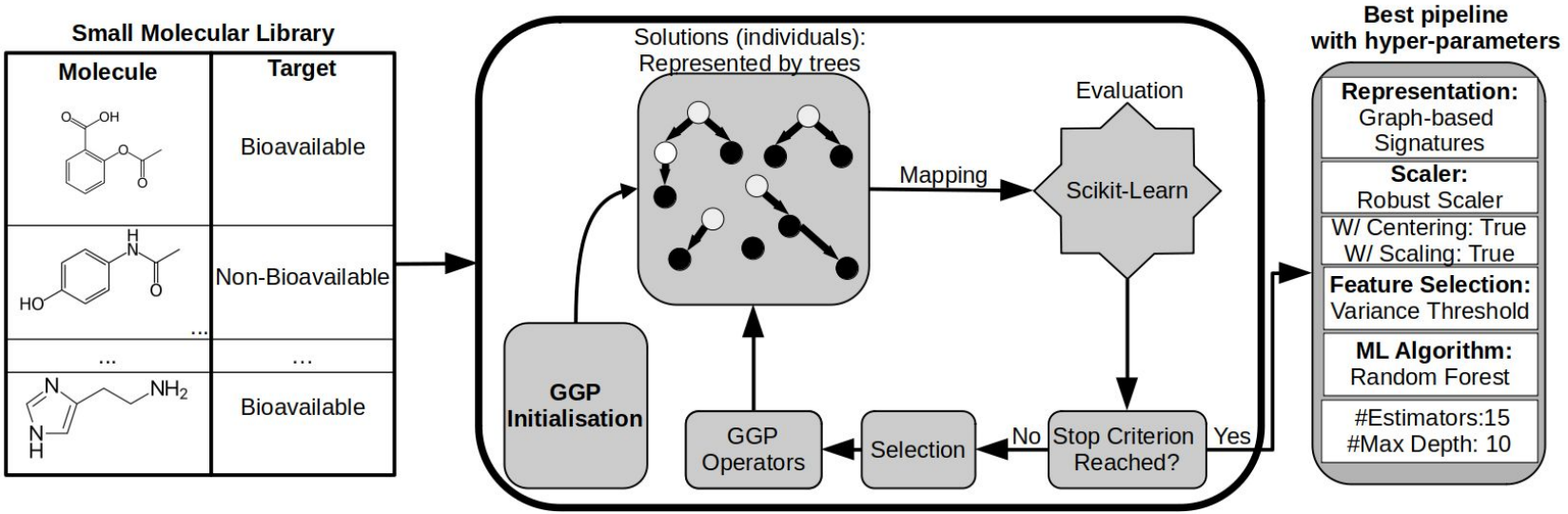
<XGBoost> ::= XGBoost <n_estimators> <max_depth> <max_leaves> <learning_rate>
<max_depth> ::= 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | None
<max_leaves> = 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10

Search Space Hardness



- Hierarchical search space with mixed types of variables (hyper-parameters).
  - Hyper-parameters can be categorical, integer or float, or being disabled by another hyper-parameter.

Search Method - Grammar-based Genetic Programming (GGP)



GGP Initialisation and Operators: Respect the grammar rules.

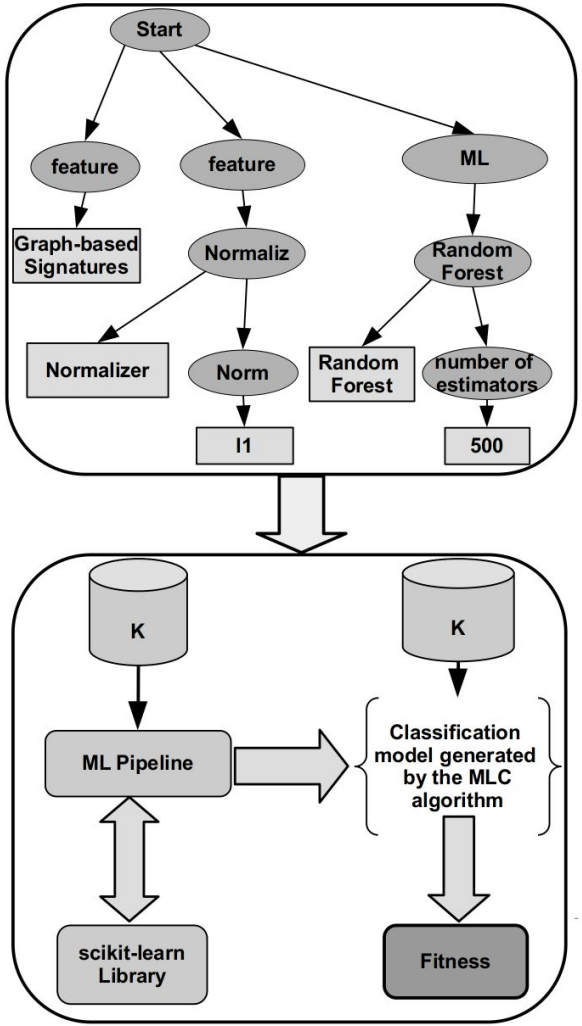
- Only valid individuals are generated.

Fitness Function

Assessing how good the pipeline is for PK property prediction.

- Scikit-Learn
- Average of K-fold Cross-Validation (K=5)
- Matthew's Correlation Coefficient (MCC)

$$MCC = \frac{((TP \times TN) - (FP \times FN))}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$



Computational Experiments

Classification Datasets

ID	Dataset	Abbreviation	PK Category	# Molecules
1	Caco-2 permeability	Caco-2	Absorption	663
2	P-glycoprotein I Inhibitor	PGP I Inhibitor	Absorption	1223
3	P-glycoprotein II Inhibitor	PGP II Inhibitor	Absorption	1023
4	P-glycoprotein I Substrate	PGP I Substrate	Absorption	1272
5	Skin Permeability	Skin Perm.	Absorption	404
6	Cytochrome P450 CYP2C9 Inhibitor	CYP2C0 Inhibitor	Metabolism	14,706
7	Cytochrome P450 CYP2C19 Inhibitor	CYP2C19 Inhibitor	Metabolism	14,572
8	Cytochrome P450 CYP2D6 Inhibitor	CYP2D6 Inhibitor	Metabolism	14,738
9	Cytochrome P450 CYP2D6 Substrate	CYP2D6 Substrate	Metabolism	666
10	Cytochrome P450 CYP3A4 Inhibitor	CYP3A4 Inhibitor	Metabolism	18,558
11	Cytochrome P450 CYP3A4 Substrate	CYP3A4 Substrate	Metabolism	669
12	Renal Organic Cation Transporter 2 Substrate	OCT2 Substrate	Excretion	904

Search Method's Parameters:

Parameter	Value
Population Size	100
Stopping Criterion	1 hour
Crossover Probability	0.90
Mutation Probability	0.10
Elitism Size	1
Data Resampling	Every 5 Generations
Individual's time budget	5 minutes

Comparison:

- **AutoML Method:** Avg 20 runs
- pkCSM and XGBoost
- Modified Friedman Test with Nemeyi's Post Hoc Test

Summary of AutoML Performance

Overall, good predictive performance from 5-fold cross-validation to blind testing, with a few exceptions:

- PGP II Substrate
- CYP2D6

Trying to understand a few cases where the performance on blind test went above CV:

- PGP II Inhibitor
- CYP3A4 Substrate

ID	Dataset	5-fold CV	Blind Test
1	Caco-2	0.589 (0.024)	0.570 (0.042)
2	PGP I Inhibitor	0.786 (0.023)	0.792 (0.044)
3	PGP II Inhibitor	0.617 (0.019)	0.754 (0.037)
4	PGP II Substrate	0.494 (0.037)	0.287 (0.107)
5	Skin Perm.	0.406 (0.025)	0.420 (0.110)
6	CYP2C9 Inhibitor	0.565 (0.026)	0.578 (0.028)
7	CYP2C19 Inhibitor	0.599 (0.028)	0.619 (0.019)
8	CYP2D6 Inhibitor	0.506 (0.034)	0.528 (0.024)
9	CYP2D6 Substrate	0.469 (0.018)	0.284 (0.088)
10	CYP3A4 Inhibitor	0.528 (0.024)	0.563 (0.024)
11	CYP3A4 Substrate	0.255 (0.013)	0.427 (0.042)
12	OCT2 Substrate	0.475 (0.019)	0.371 (0.062)

Comparison to pkCSM and XGBoost

The average results of the proposed AutoML method were close to pkCSM and XGBoost.

When using the **best selected pipelines** by the **AutoML method**, results are clearly better.

Dataset	Avg. AutoML Method	Best AutoML-Selected	pkCSM	XGBoost
Caco-2	0.57	**0.61**	**0.609**	0.579
PGP I Inhibitor	0.792	**0.837**	0.776	0.82
PGP II Inhibitor	0.754	**0.783**	0.716	0.696
PGP II Substrate	**0.287**	**0.289**	0.214	0.232
Skin Perm.	**0.42**	0.394	0.108	0.368
CYP2C9 Inhibitor	0.578	**0.615**	0.601	0.553
CYP2C19 Inhibitor	0.619	**0.647**	0.583	0.59
CYP2D6 Inhibitor	0.528	**0.556**	0.408	0.488
CYP2D6 Substrate	0.284	**0.334**	0.197	0.267
CYP3A4 Inhibitor	0.563	0.59	**0.623**	0.534
CYP3A4 Substrate	0.427	0.274	0.289	**0.44**
OCT2 Substrate	0.371	**0.427**	0.353	0.402
Avg. Value	0.516	**0.53**	0.456	0.497
Avg. Ranking	2.417	**1.417**	3.25	0.917

Understanding AutoML results for Pharmacokinetics

Representation:
Combination of General Descriptors, Advanced Descriptors and Graph-based Signatures (**13%**).
Combination of Advanced Descriptors and Graph-based Signatures (**9.6%**).
Combination of General Descriptors, Graph-based Signatures, Toxicophores and Fragments (**9.2%**).

No Scaling
Tree-based machine learning models.

Feature Selection
No Feature Selection **(17.9%),** Select FDR (**15.8%**), Select FPR (**13.8%**), and Select Percentile (**12.1%**).

ML algorithms
Gradient Boosting (**42.5%**), Random Forest (**20%**), Extremely Randomised Trees (**17.5%**), and XGBOOST (**15.4%**).

Summary

- New evolutionary-based AutoML method that is able to find pipelines that perform similarly or better that current approaches in the context of PK prediction.
  - Method available at:
    - **https://github.com/alexgcsa/ecada_2024**

- It can be a starting point when you have a new PK problem to solve.
  - **Aim:** End-to-end approach.

Future work

- Extend the evaluation - Deep-PK and admetSAR 3.0 datasets.
  - Include regression in the evaluation.
- Extend the search space.
- Improve the AutoML method in terms of search and evaluation.
  - Surrogate model to guide the search.
  - Evaluation considering scaffold and molecule similarity.

Towards Evolutionary-based Automated Machine Learning for Small Molecule Pharmacokinetic Prediction

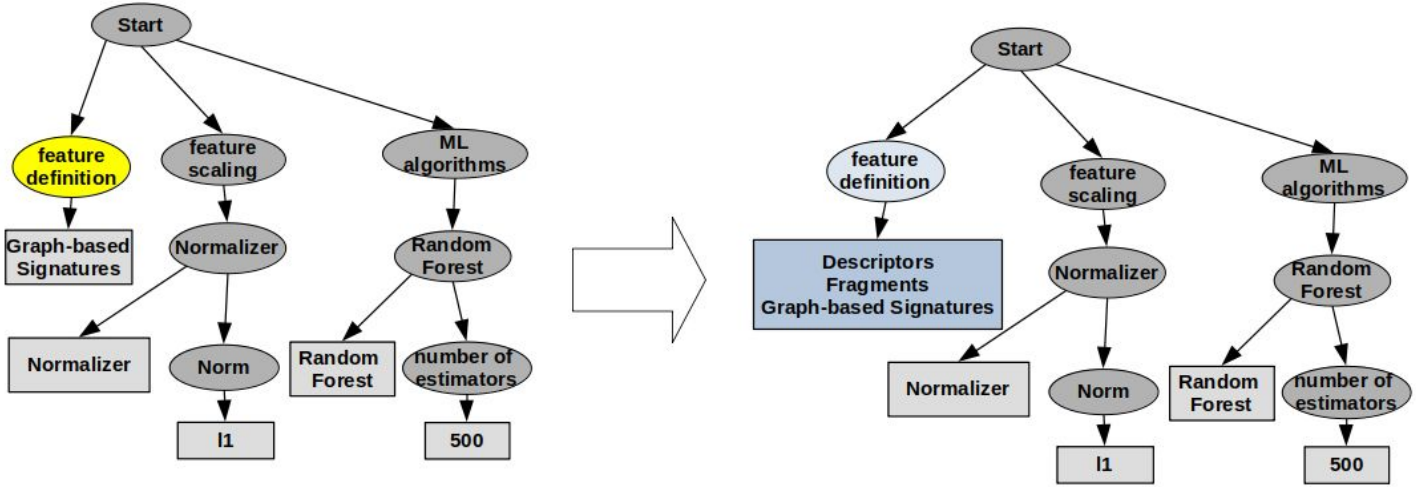
Alex G. C. de Sá
David B. Ascher

Supplementary Slides

Search Method - GGP initialisation and operators

- **Initialisation** and **GGP operators** take into account the grammar to create new individuals (ML pipelines).

- **Mutation:**



Search Method - GGP initialisation and operators

- **Crossover:**