Automatic Model Generation

Joelle Gola, Olga Obrezanova

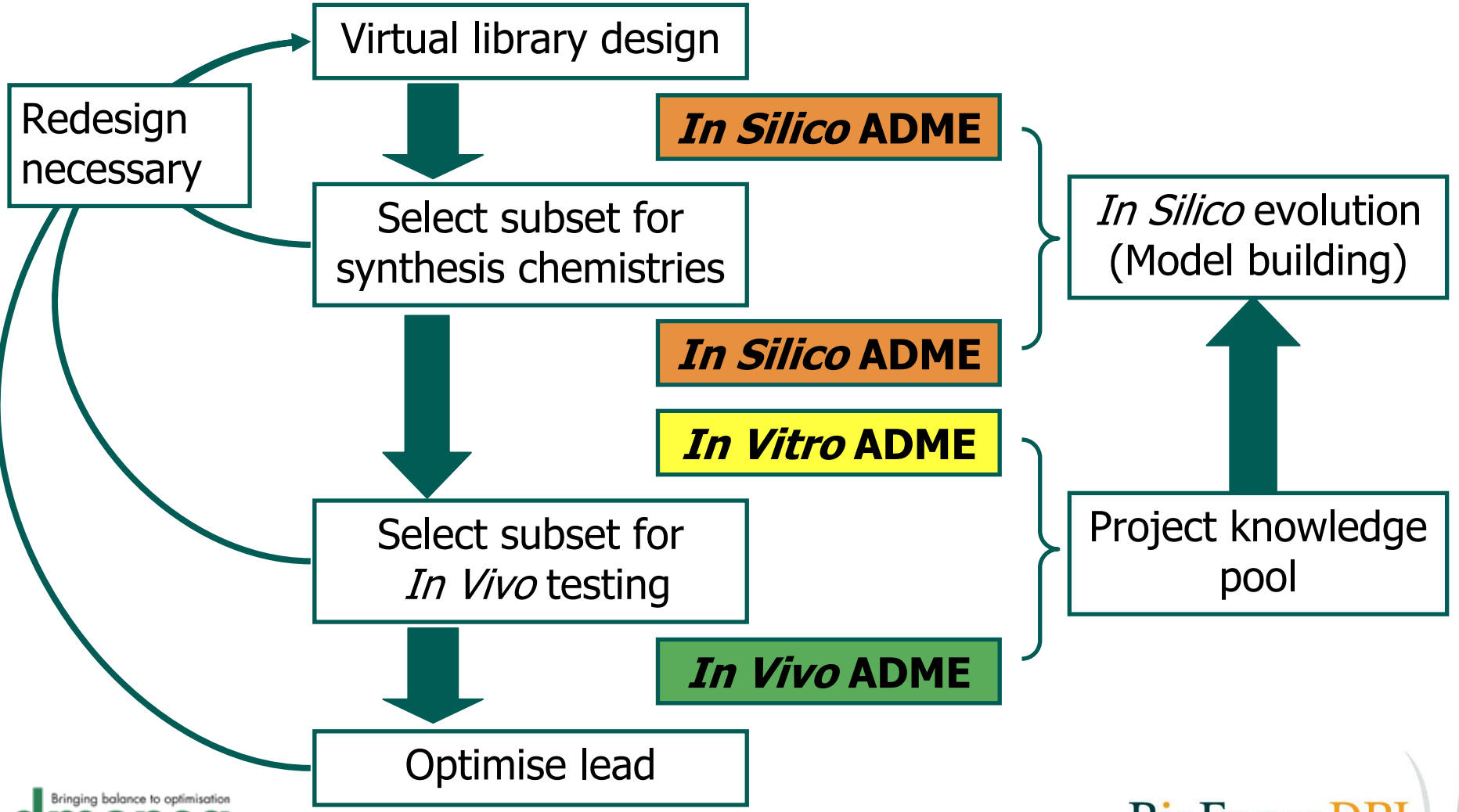
19 June 2007

BioFocus DPI
A Galápagos Company

Overview

- Background
- Process
- Techniques
- Illustrative examples
- Next steps

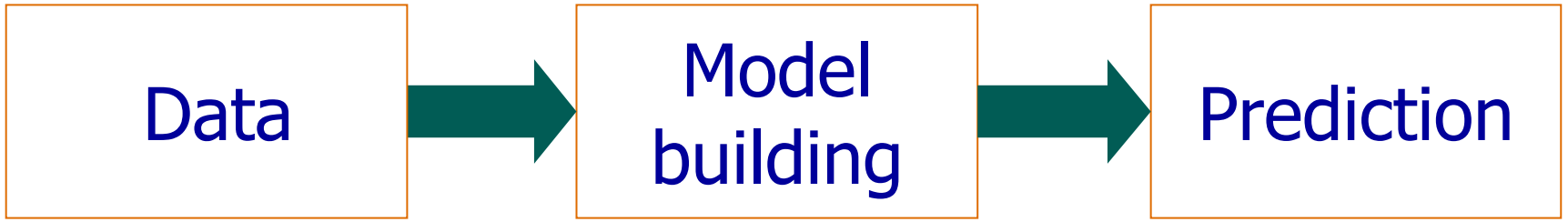
Application of ADME Predictions



Admensa Bringing balance to optimisation

BioFocus DPI A Galápagos Company

General Modeling Process



Quality
Volume

Descriptors
Set split
Techniques
Validation
Model selection

Uncertainty
Chemical space
Interpretation

Objectives

- Simplify and speed up the model building process
- Understand and make use of the model output
  - Compound in, prediction out….
- Flexible approach
  - Expert vs. non expert
- Facilitate integration of the model with the decision making process
  - Chemical space analysis
  - Probabilistic scoring

Process

Structure + Activity

↓

Descriptor Calculation

↓

Feature Selection

↓

Set Split

↓

Modeling Techniques

↓

Model Selection

↓

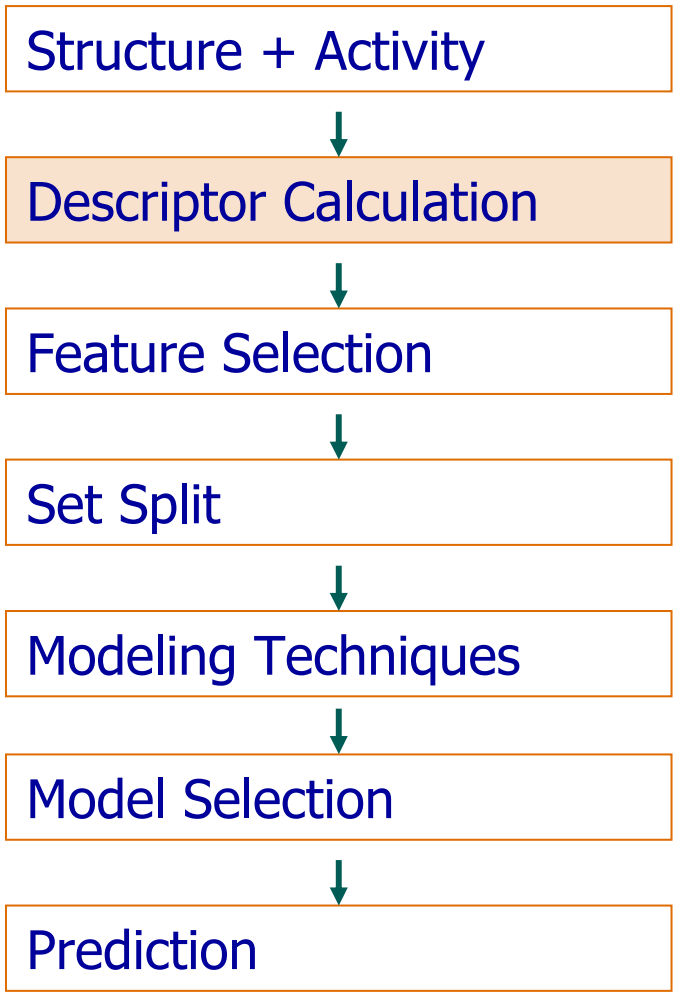
Prediction

- Checking input data

Admensa Bringing balance to optimisation

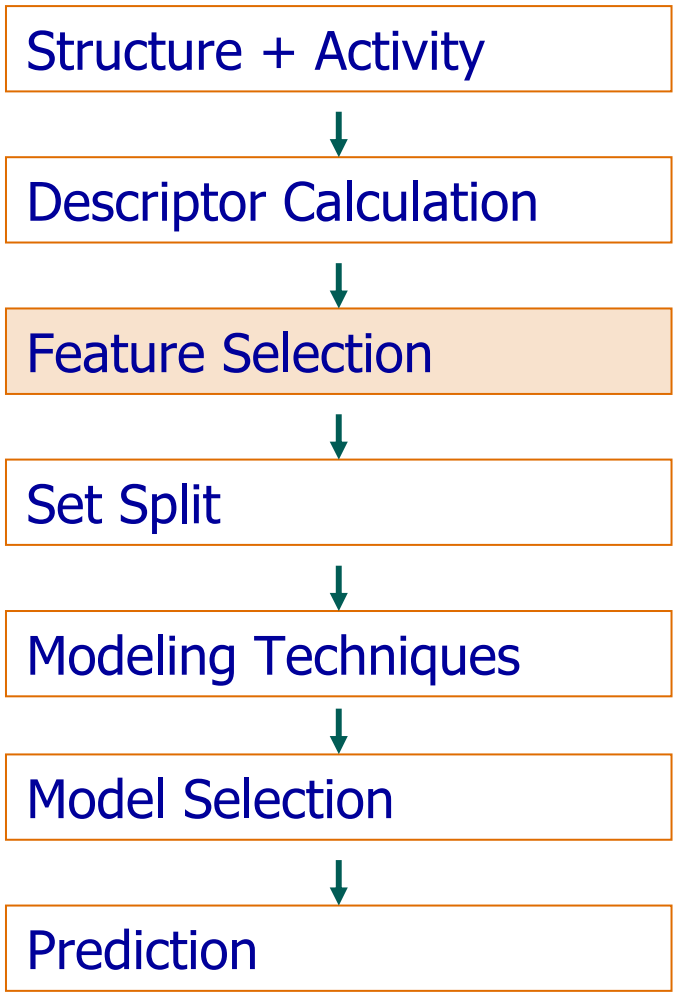
BioFocus DPI A Galápagos Company

Process



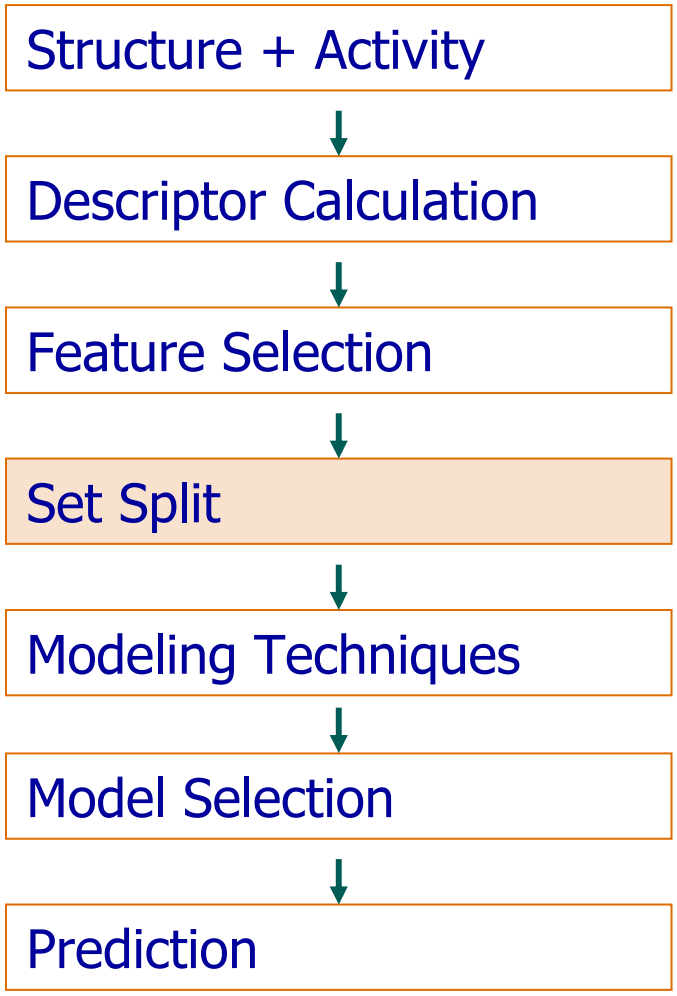
- Whole molecule properties
- ~330 2D SMARTS based descriptors
- Input own descriptors (additional columns)
- User defined 2D SMARTS based descriptors

Process



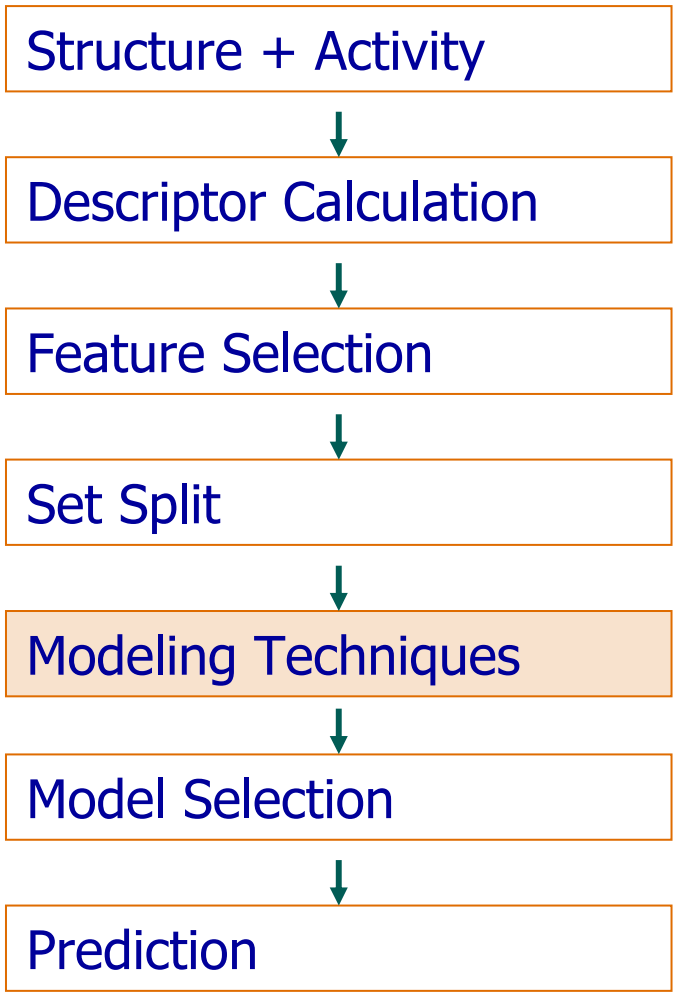
- Exclude descriptors with:
  - Less than 4% occurrences
  - Standard deviation: 0.0005
- Pair-wise correlated descriptors:
  - Correlation >=95%

Process



- Require 3 sets:
  - Training: Building models
  - Validation: Selecting best model
  - Test: Checking best model
- Clustering based method:
  - Structural fingerprint
  - Tanimoto level: 0-1
  - % compounds in the training set
- 3 user defined sets

Process

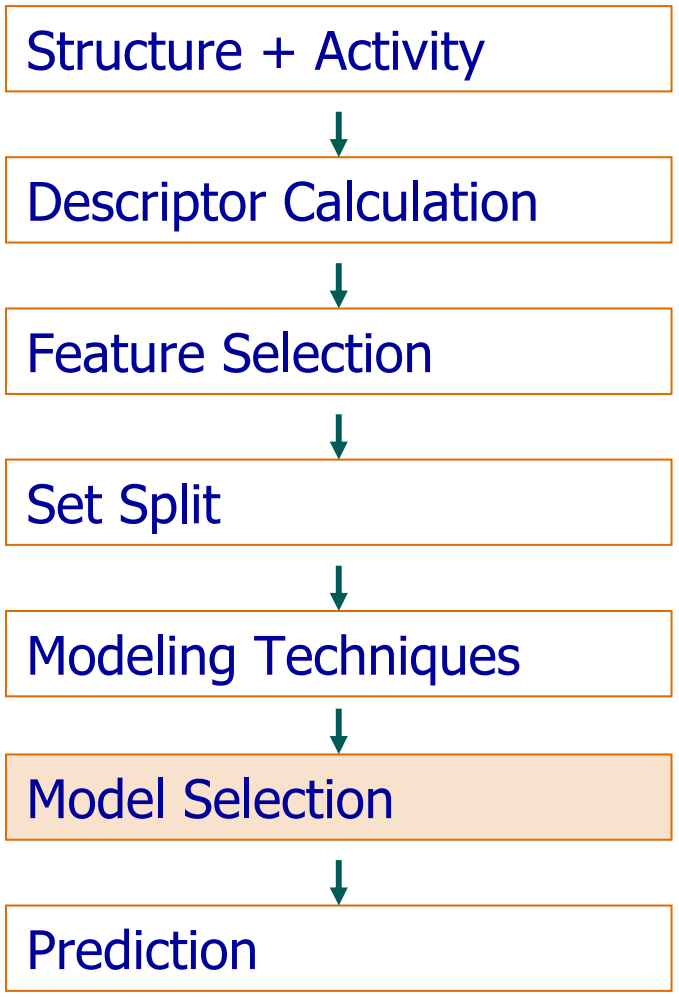


- Classification models:
  - Decision Tree
- Continuous models:
  - Partial Least Squares
  - Gaussian Processes
  - Radial Basis Functions
- Automatically run all the appropriate models
- User selects preferred techniques

Admensa Bringing balance to optimisation

BioFocus DPI A Galápagos Company

Process

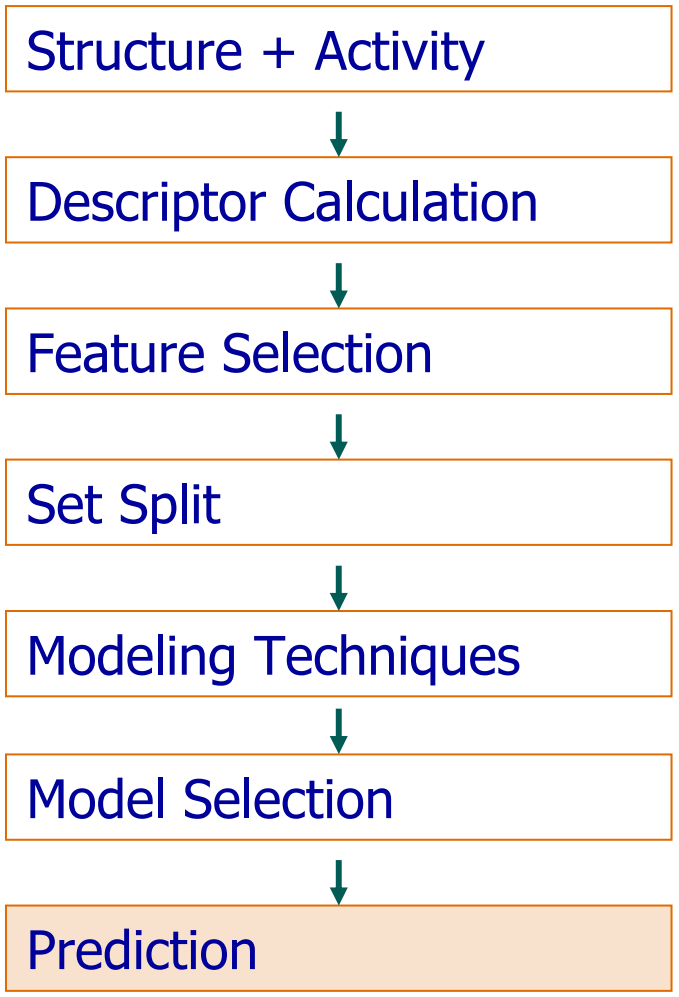


- Compare statistical results on validation set:
  - Set of rules to automatically select the best model
  - User choice

Admensa Bringing balance to optimisation

BioFocus DPI A Galápagos Company

Process



- Further validation of the chosen model using test set
- Uncertainty in prediction
- Glowing Molecule

Admensa Bringing balance to optimisation

BioFocus DPI A Galápagos Company

Techniques: Classification Models

- Recursive partitioning approach to building classification models
- Based on C4.5 approach developed by Ross Quinlan

Rules:

- Generate up to 20 different models by automatically varying method settings:
  - Simple decision trees with various stopping conditions
  - Pruned decision trees
  - Rule sets built from decision trees

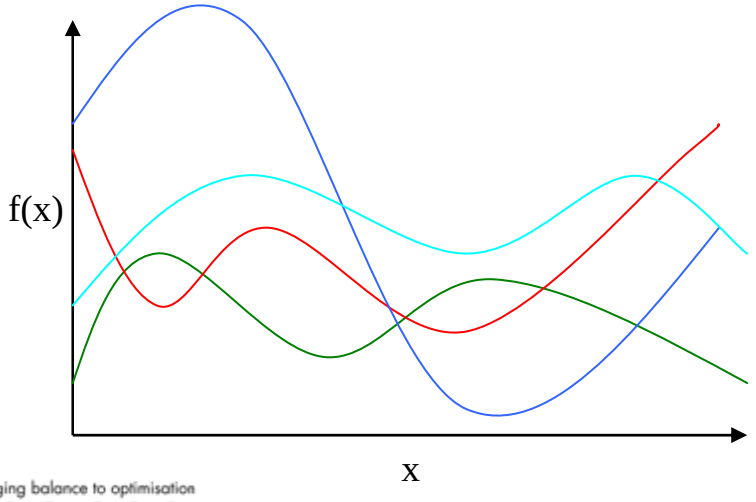
Techniques: Continuous Models

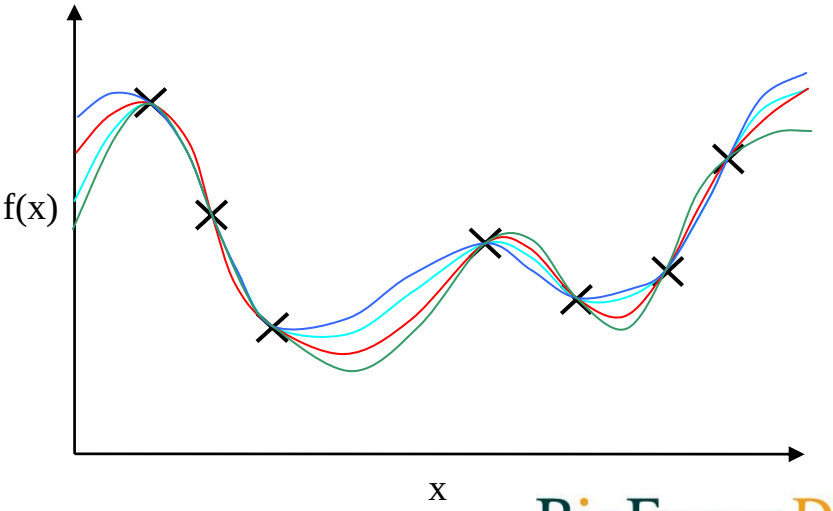
- Suite of modeling techniques include:
  - Partial Least Squares
  - Gaussian Processes
  - Radial Basis Functions

Techniques: Gaussian Processes

- Powerful machine learning technique based on a Bayesian statistical approach
  - Bayes Theorem: $$P(f \mid Y, X) \propto P(Y \mid f, X) P(f)$$

    Posterior distribution — Prior distribution
  - Gaussian process defines the distribution over functions





Admensa Bringing balance to optimisation

BioFocus DPI A Galápagos Company

Techniques: Gaussian Processes

- Automatic determination of the model parameters:
  - Inherent ability to select relevant descriptors
- Implemented 5 techniques*:

Techniques	Descriptor Selection
GP-Fixed	-
GP-2DSearch	-
GP-FVS	√
GP-RFVS	√
GP-OPT	√

Increasing computational demand

* Obrezanova et al., J.Chem.Inf.Model., accepted

Techniques: Radial Basis Functions

- For given Y and x, RBF is used to find function f(x)
  - f(x) = Y + noise
- f(x) chosen as:

$$f(x) = \sum_{i=1}^{N} a_i \left\| x - x^{(i)} \right\|$$

- RBF requires f(x) to pass through all training points:

$$y_j = \sum_{i=1}^{N} a_i \left\| x^{(j)} - x^{(i)} \right\|, \quad j = 1 \ldots N$$

weights a_i can be found from this linear system of equations

Techniques: Radial Basis Functions

- Good method for small or large data sets
  - But sensitive to noise created by excessive descriptors

Solution:

- Coupled with a genetic algorithm, GA-RBF
- Automatically select RBF or GA-RBF:

# Compounds per descriptor	GA-RBF	RBF
< 5	√	-
≥ 5	-	√

Techniques: Performance Measures

- Classification models

$$Accuracy = \frac{Correctly\ \ predicted\ \ compounds}{Total\ \ number\ \ of\ \ compounds}$$

$$Kappa = \frac{(Observed\ \ agreement - Chance\ \ agreement)}{(Total - Observed\ \ agreement)}$$

- Continuous models

$$R^2 = 1 - \frac{\sum_i (y_i^{pred} - y_i^{obs})^2}{\sum_i (y_i^{obs} - \overline{y^{obs}})^2}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i^{pred} - y_i^{obs})^2}$$

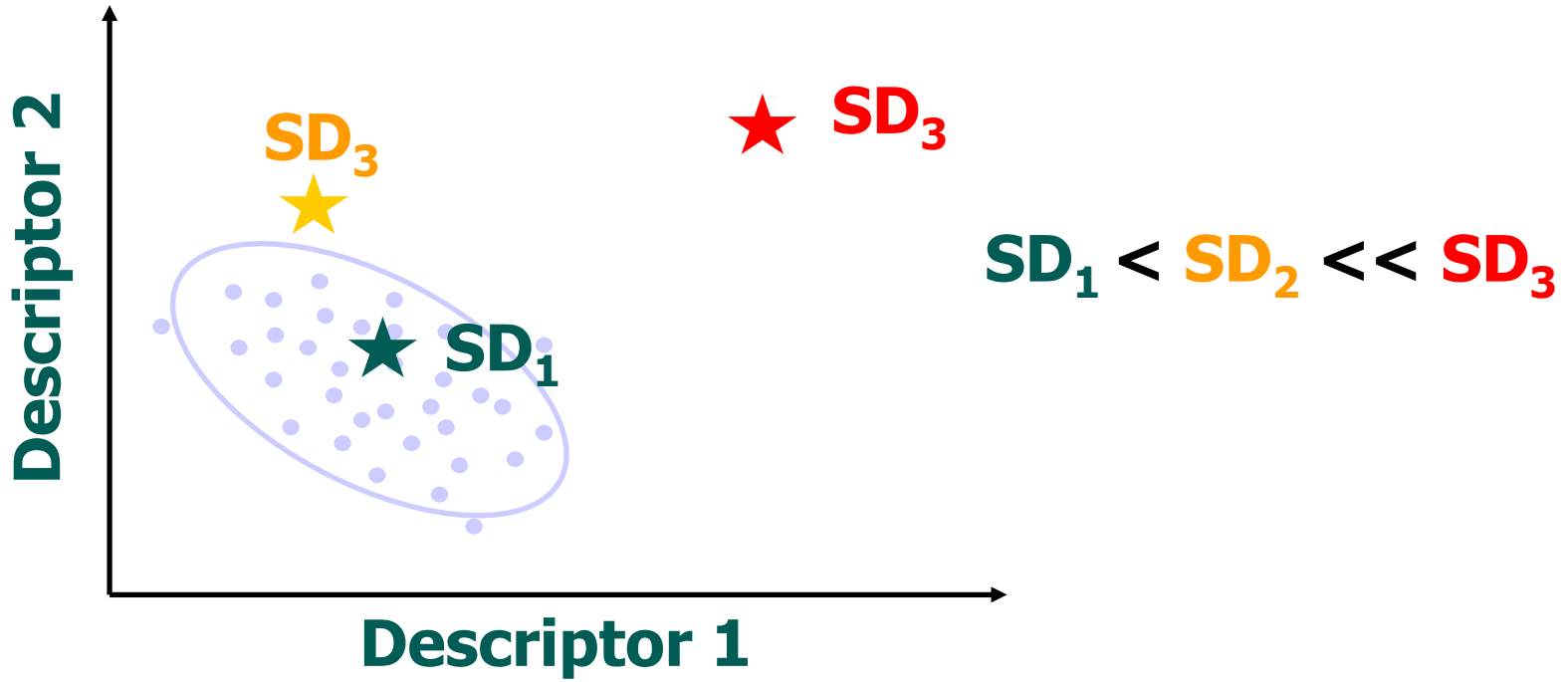
Techniques: Choosing Best Model

- Selection of the best model is done by its performance on the validation set.

Rules:

- Classification model
  - The higher Kappa statistic (0<Kappa< 1)
  - If identical Kappa statistics then choose model with best statistical results on training set
- Continuous model
  - Smaller RMSE

Techniques: Uncertainties in Prediction



- The position of compounds relative to the chemical space of the model is reflected in the reported standard deviation (SD)

Techniques: Interpretation

- Molecule in, prediction out…

"Why did *this* model predict *that* value for *my* molecule?"

"What can I do to my molecule to improve the prediction?"

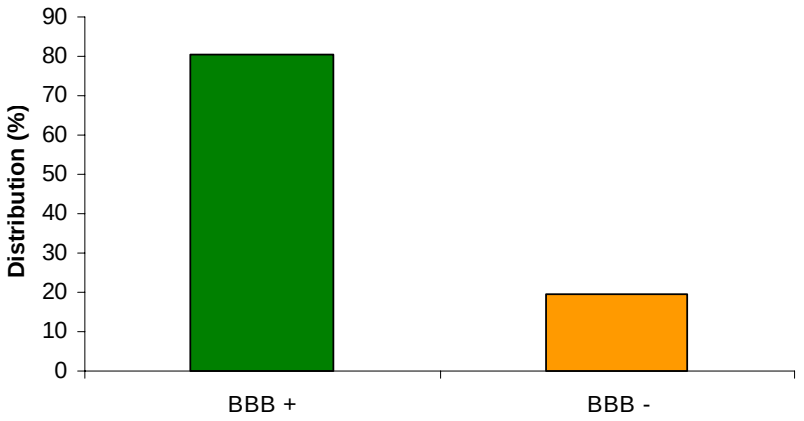
- Models already encode some of these answers!
  - Reveal non-linear as well as linear relationships between property being modelled and the selected descriptors
  - Proprietary algorithm: Glowing Molecule

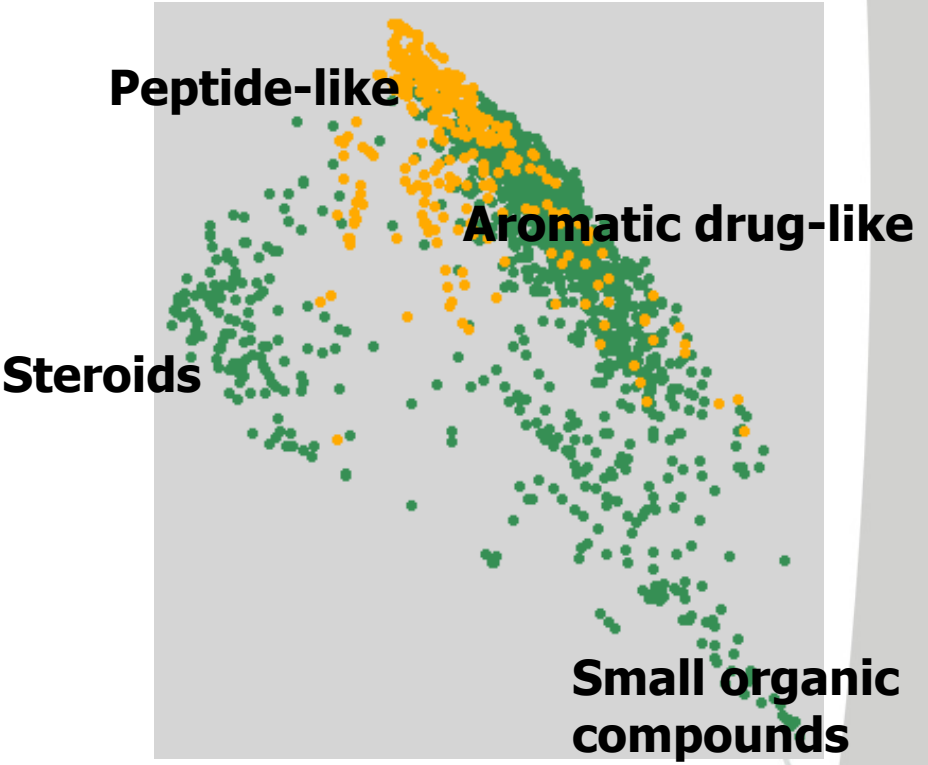
Illustrative Examples

- Example 1: BBB classification model
- Example 2: Continuous hERG model
- Example 3: Building a local model

Example 1: BBB Classification Model

- BBB± data set from a CNS library published by Zhao et al. Originally prepared by Adenot and Lahana
  - J. Chem. Inf. Model, 2007,47,170
  - 1593 compounds
  - Chemically Diverse Set





Admensa Bringing balance to optimisation

BioFocus DPI A Galápagos Company

Example 1: BBB Classification Model

	Training (1273 cpds-112 Descriptors)			Validation (158 cpds)		Test (160 cpds)	
Models	# Descriptors	Kappa	Accuracy	Kappa	Accuracy	Kappa	Accuracy
DT1	7	0.78	0.94	0.88	0.96		
DT2	5	0.77	0.94	0.88	0.96		
DT4	6	0.77	0.94	0.88	0.96		
DT7	1	0.63	0.91	0.70	0.90		
DT13	13	0.87	0.96	0.91	0.97		
DT14	15	0.89	0.97	0.90	0.96		
DT15	15	0.86	0.96	0.95	0.98		
DT16	20	0.84	0.96	0.91	0.97		
DT17	13	0.84	0.96	0.90	0.96		
DT19	28	0.88	0.97	0.92	0.97		
DT20	26	0.86	0.96	0.91	0.97		

Example 1: BBB Classification Model

	Training (1273 cpds-112 Descriptors)			Validation (158 cpds)		Test (160 cpds)	
Models	# Descriptors	Kappa	Accuracy	Kappa	Accuracy	Kappa	Accuracy
DT1	7	0.78	0.94	0.88	0.96		
DT2	5	0.77	0.94	0.88	0.96		
DT4	6	0.77	0.94	0.88	0.96		
DT7	1	0.63	0.91	0.70	0.90		
DT13	13	0.87	0.96	0.91	0.97		
DT14	15	0.89	0.97	0.90	0.96		
DT15	15	0.86	0.96	0.95	0.98		
DT16	20	0.84	0.96	0.91	0.97		
DT17	13	0.84	0.96	0.90	0.96		
DT19	28	0.88	0.97	0.92	0.97		
DT20	26	0.86	0.96	0.91	0.97		

Example 1: BBB Classification Model

	Training (1273 cpds-112 Descriptors)			Validation (158 cpds)		Test (160 cpds)	
Models	# Descriptors	Kappa	Accuracy	Kappa	Accuracy	Kappa	Accuracy
DT1	7	0.78	0.94	0.88	0.96		
DT2	5	0.77	0.94	0.88	0.96		
DT4	6	0.77	0.94	0.88	0.96		
DT7	1	0.63	0.91	0.70	0.90		
DT13	13	0.87	0.96	0.91	0.97		
DT14	15	0.89	0.97	0.90	0.96		
DT15	15	0.86	0.96	0.95	0.98	0.95	0.98
DT16	20	0.84	0.96	0.91	0.97		
DT17	13	0.84	0.96	0.90	0.96		
DT19	28	0.88	0.97	0.92	0.97		
DT20	26	0.86	0.96	0.91	0.97		

Example 1: BBB Classification Model

Validation set		Predicted	
		BBB -	BBB +
Obs.	BBB -	38	3
	BBB +	0	117

Test set		Predicted	
		BBB -	BBB +
Obs.	BBB -	38	1
	BBB +	2	119

- DT15 is a rule based model with 10 rules

Rule 1:

If PSA > 108.8 and q137 > 26 and calc-logP <= 3.5 then

BBB - confidence = 0.98

Rule 2:

If thioEther = 0 and PSA =< 128.2 and ed70 = 0 and Negative Charge = 0 then

BBB+ confidence = 0.97

Admensa Bringing balance to optimisation

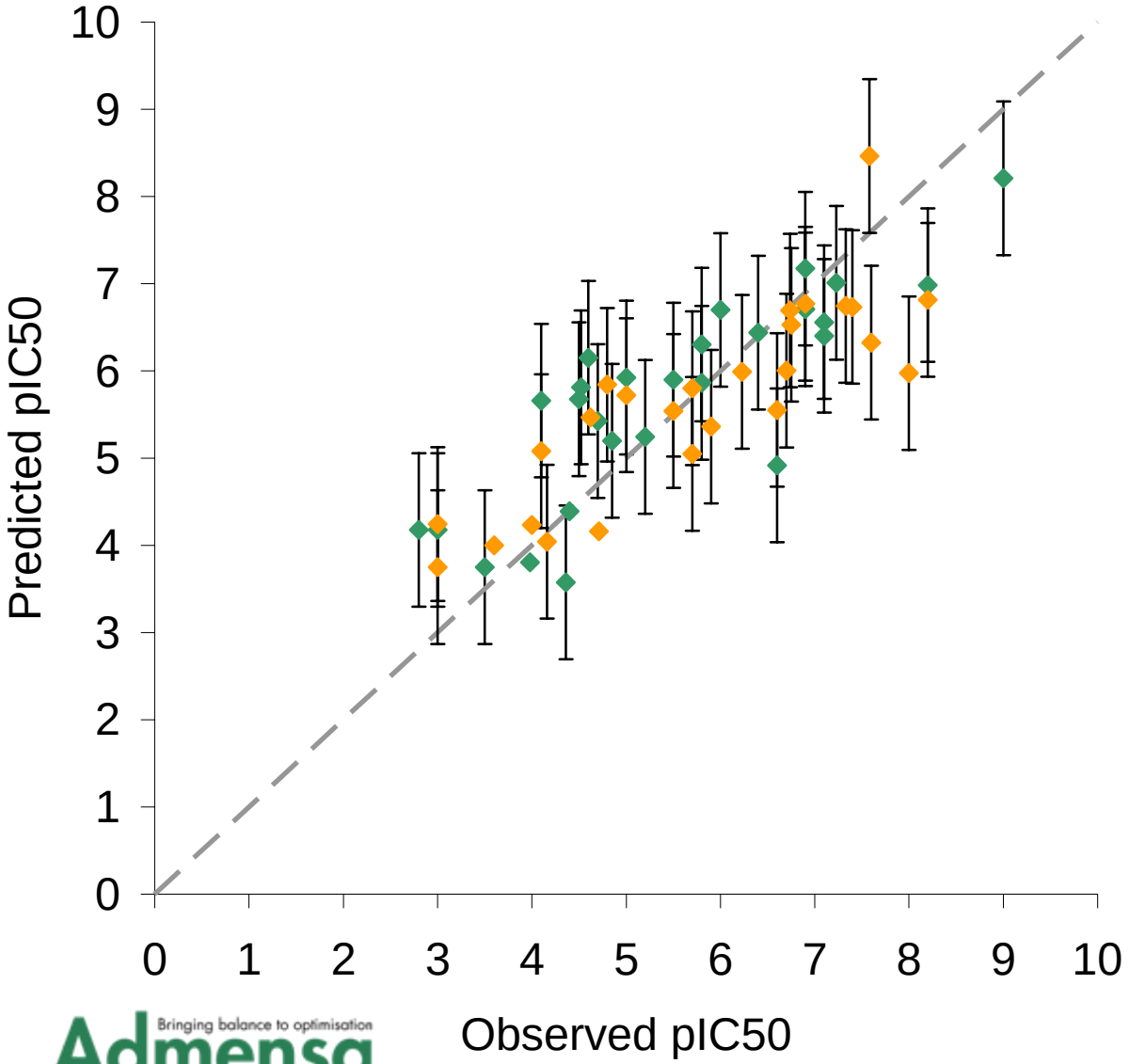
BioFocus DPI A Galápagos Company

Example 2: Continuous hERG Model

- 177 pIC_{50} values from the literature
  - Patch clamp measurements in mammalian cells – mainly HEK293

	Training (124 cpds-143 descriptors)			Validation (27 cpds)		Test (26 cpds)	
Models	# Descriptors	R^2	RMSE	R^2	RMSE	R^2	RMSE
PLS	143	0.69	0.770	0.61	0.953		
GPFixed	143	0.85	0.538	0.66	0.883		
GP2DSearch	143	0.83	0.576	0.65	0.892		
GPFVS	41	0.78	0.661	0.68	0.864	0.71	0.822
GPRFVS	19	0.81	0.611	0.65	0.897		
GPOPT	42	0.84	0.558	0.63	0.925		
GA-RBF	23	1	0	0.64	0.910		

Example 2: Continuous hERG Model



Validation set:
R^2 = 0.68

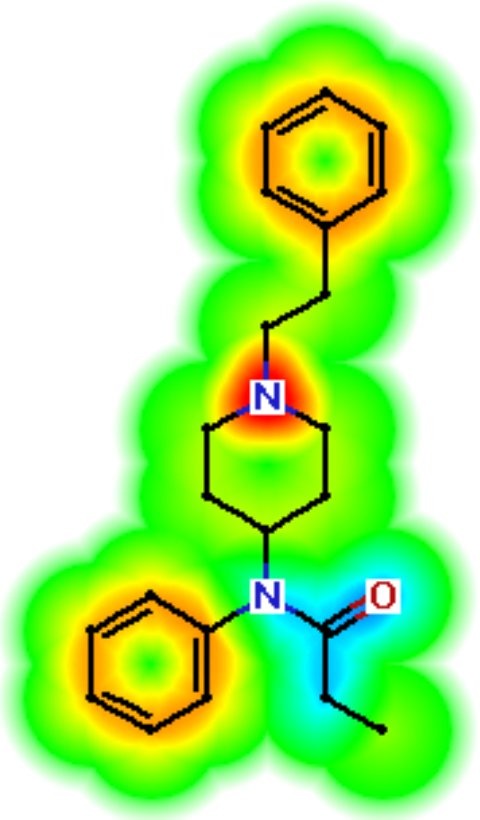
Test set:
R^2 = 0.71

Admensa Bringing balance to optimisation

BioFocus DPI
A Galápagos Company

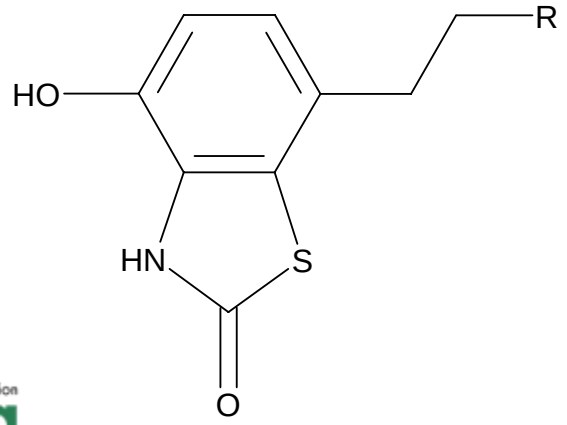
Example 2: Continuous hERG Model

- Important chemical features for hERG binding:
  - Lipophilicity
  - Negative charge
  - Positively charged nitrogen at pH 7.4
  - Aromaticity index
- Redesign using Glowing Molecule



Example 3: Building a Local Model

- Austin et al. measured logD values at pH 7.4 for 78 compounds
  - J Med Chem. 2003 Jul 17;46(15):3210-20
- QSAR model highlighted the importance of lipophilicity and ionization in controlling beta(2) duration
- Wants to predict logD7.4 values for new compounds

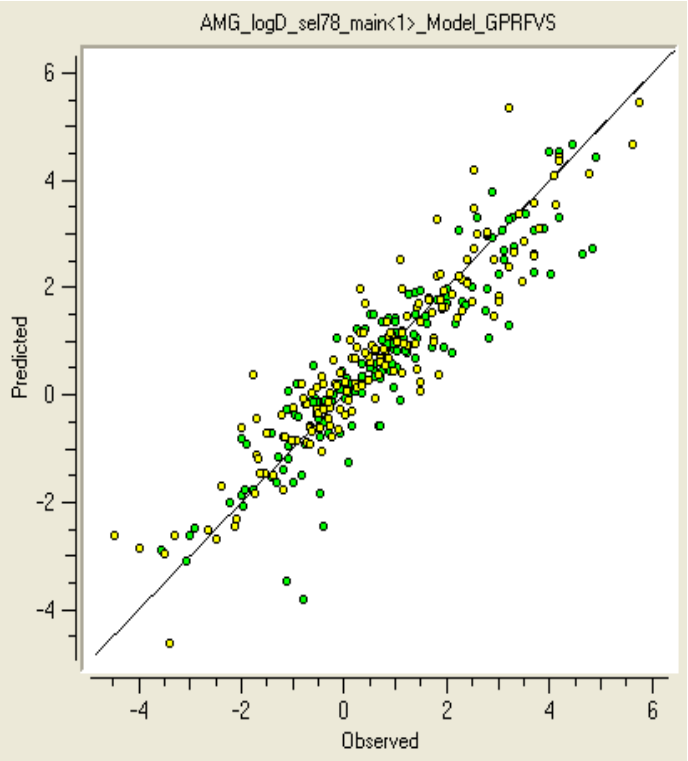


Admensa Bringing balance to optimisation

BioFocus DPI A Galápagos Company

Example 3: Global Model

- Current global model was built on 1044 compounds
- logD values measured at pH 7.4 were extracted from StARLITe™



Training Set:
GPRFVS Model, 52 descriptors
R^2 = 0.91 RMSE = 0.57

Validation set:
R^2 = 0.82
RMSE = 0.75

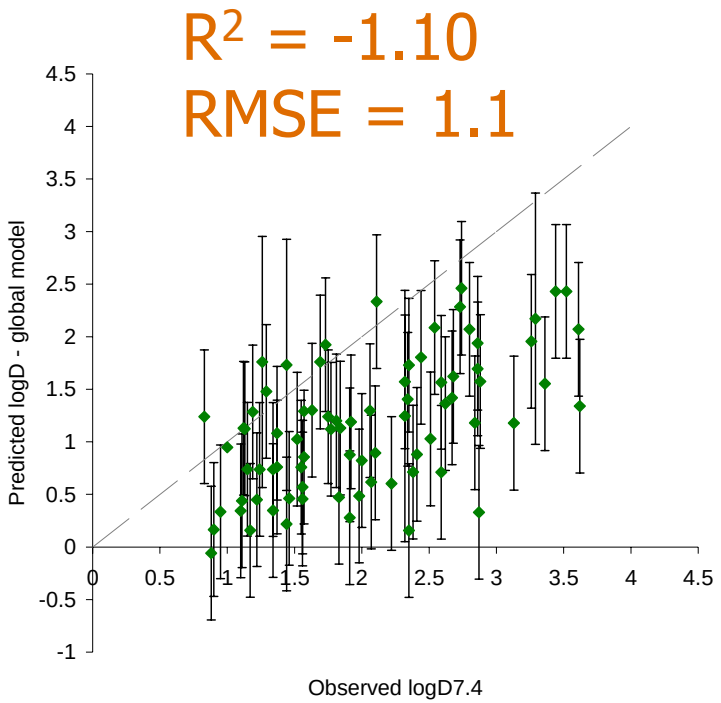
Test set:
R^2 = 0.86
RMSE = 0.68

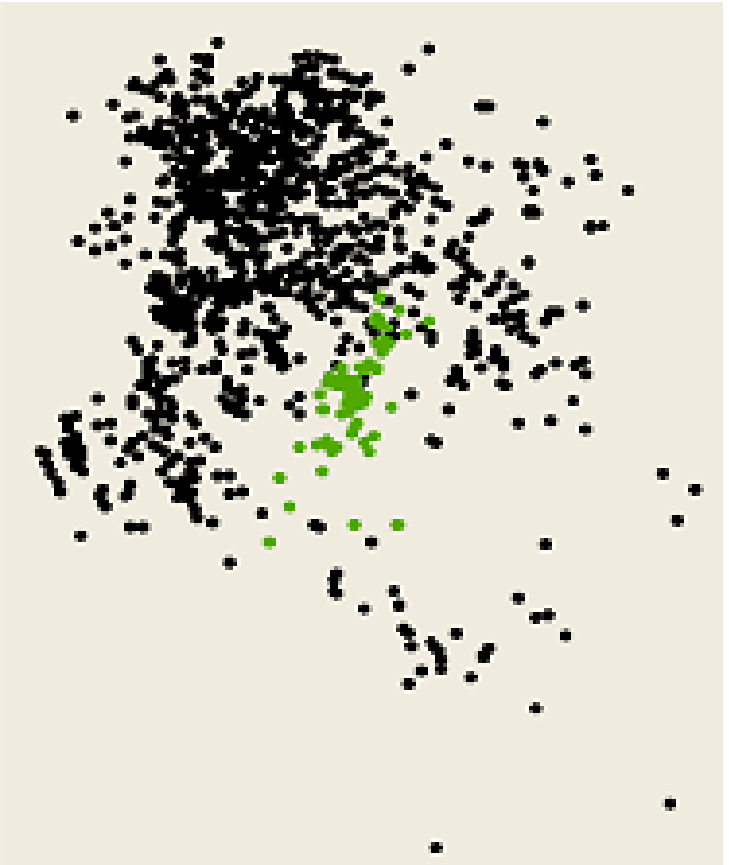
Admensa Bringing balance to optimisation

BioFocus DPI A Galápagos Company

Example 3: Applying Global Model

Results for the 78 Austin compounds:



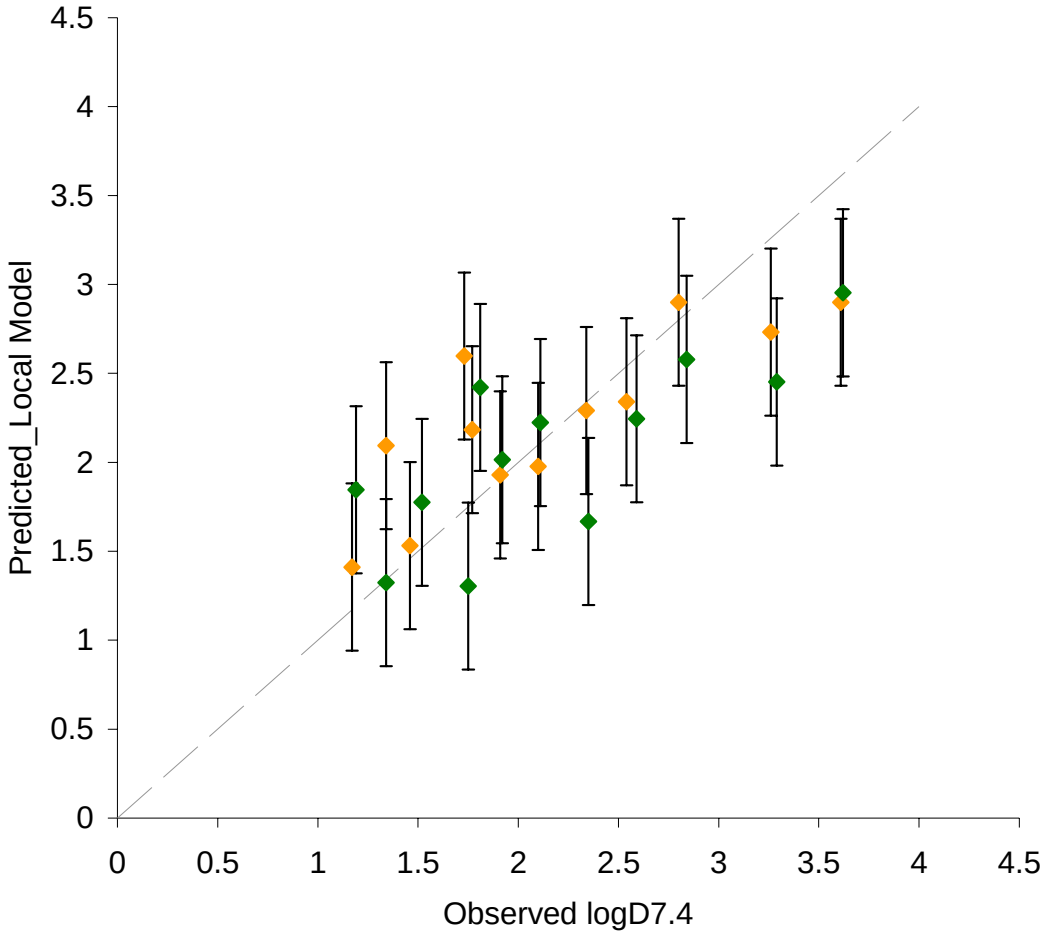


Example 3: Building a Local Model

Build a model on the 78 Austin compounds.

	Training (54 cpds – 72 descriptors)			Validation (12 cpds)		Test (12 cpds)	
Models	# Descriptors	R^2	RMSE	R^2	RMSE	R^2	RMSE
PLS	72	0.44	0.54	0.37	0.57		
GPFixed	72	0.70	0.39	0.32	0.60		
GP2DSearch	72	0.86	0.26	0.39	0.57		
GPFVS	22	0.82	0.31	0.55	0.49	0.62	0.45
GPRFVS	6	0.81	0.32	0.46	0.54		
GPOPT	18	0.87	0.26	0.47	0.53		
GA-RBF	7	1	0	-0.13	0.78		

Example 3: Building a Local Model



Validation set:
R^2 = 0.55
RMSE = 0.49

Test set:
R^2 = 0.62
RMSE = 0.45

Admensa Bringing balance to optimisation

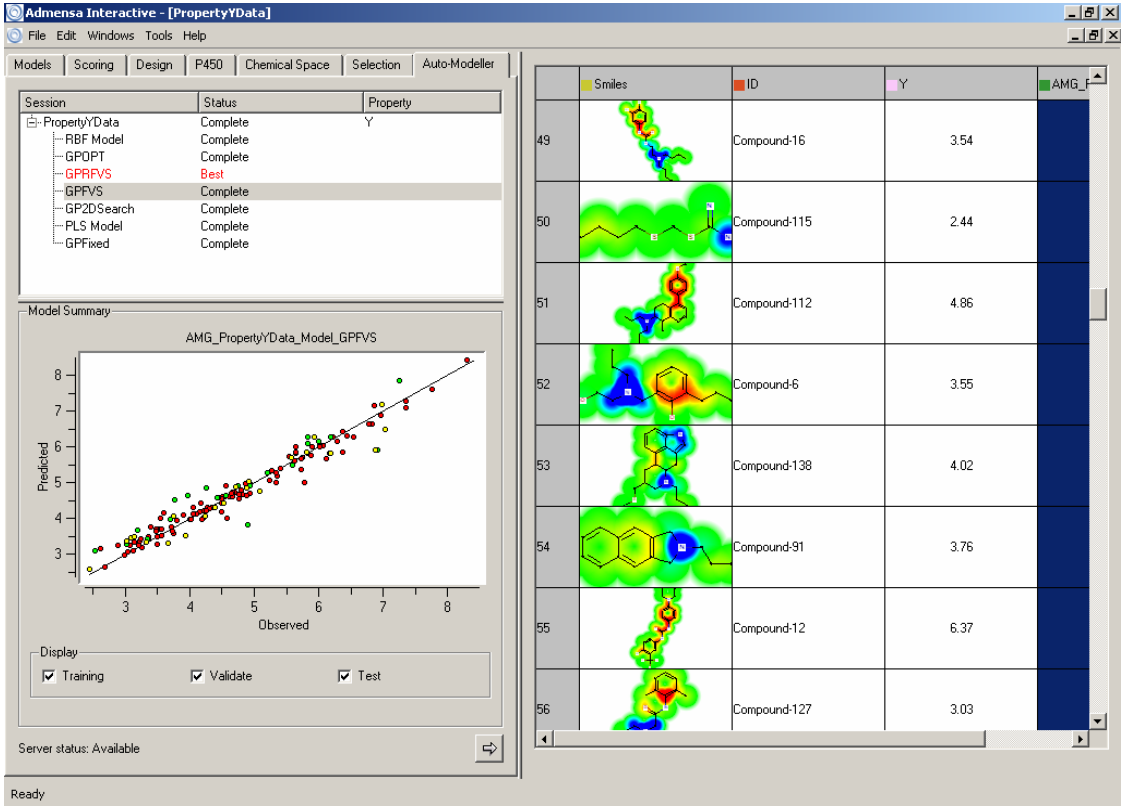
BioFocus DPI
A Galápagos Company

Next Steps

- New modeling techniques:
  - Gaussian Processes with Nested Sampling
  - Apply Gaussian Processes to categorical problems
- New descriptors:
  - Automatically designed descriptors to take into account of features not present in current set of descriptors
- Set split:
  - Look at other techniques

Acknowledgments

- Matthew Segall
- Edmund Champness
- Olga Obrezanova
- Christopher Leeding
- Andre Kramer



admensa-support@glpg.com

Admensa Bringing balance to optimisation

BioFocus DPI A Galápagos Company