Gaussian Processes: A method for automatic QSAR and ADME modelling

Olga Obrezanova, Joelle M.R. Gola, Matthew D. Segall

22 August 2007

BioFocus DPI
A Galápagos Company

Overview

- Gaussian Processes
  - A powerful computational modelling technique
- Application - predictive ADME and QSAR modelling (ADME – absorption, distribution, metabolism and excretion)
  - New techniques for finding method parameters
  - Examples and comparison with other methods
- Automatic modelling process

Background

- Machine learning method based on Bayesian approach. Not widely used in QSAR and ADME field yet.
- Advantages:
  - Does not require a priori determination of model parameters.
  - Nonlinear relationship modelling.
  - Built-in tool to prevent overtraining, no need for cross-validation.
  - Inherent ability to select important descriptors.
  - Provides uncertainty estimate for each prediction.
- Sufficiently robust to enable automatic model generation

Gaussian Processes: Key idea

- $D=\{Y, X\}$ – given data. We want to find function f:

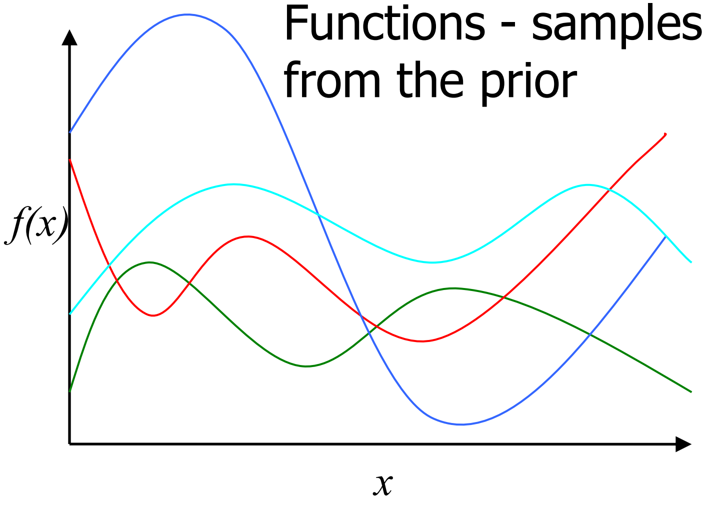
$$Y=f(X)+\text{noise}.$$

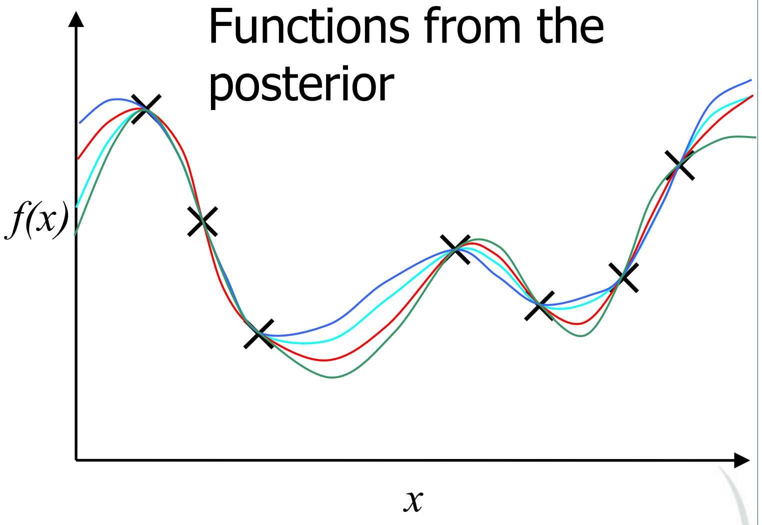
- Bayesian rule

$$\underset{\text{posterior}}{\underset{\nearrow}{P(f\,|\,D)}} \propto P(D\,|\,f)\,\underset{\text{prior}}{\underset{\nwarrow}{P(f)}}$$

- Prediction is a mean of posterior distribution.

- Gaussian Process defines a distribution over functions.


Functions - samples from the prior


Functions from the posterior

Gaussian Processes: Practical steps

- Structure of functions determined by covariance (kernel) function:

$$\mathrm{cov}(f(\boldsymbol{x}), f(\boldsymbol{x}')) = C(\boldsymbol{x}, \boldsymbol{x}')$$

- Distribution of functions (property values) is multivariate Gaussian with zero mean and covariance matrix

$$\boldsymbol{K} = \boldsymbol{C} + \theta_3 \boldsymbol{I}$$

  - Hyperparameter θ_3 is a variance of noise present in the observed values.

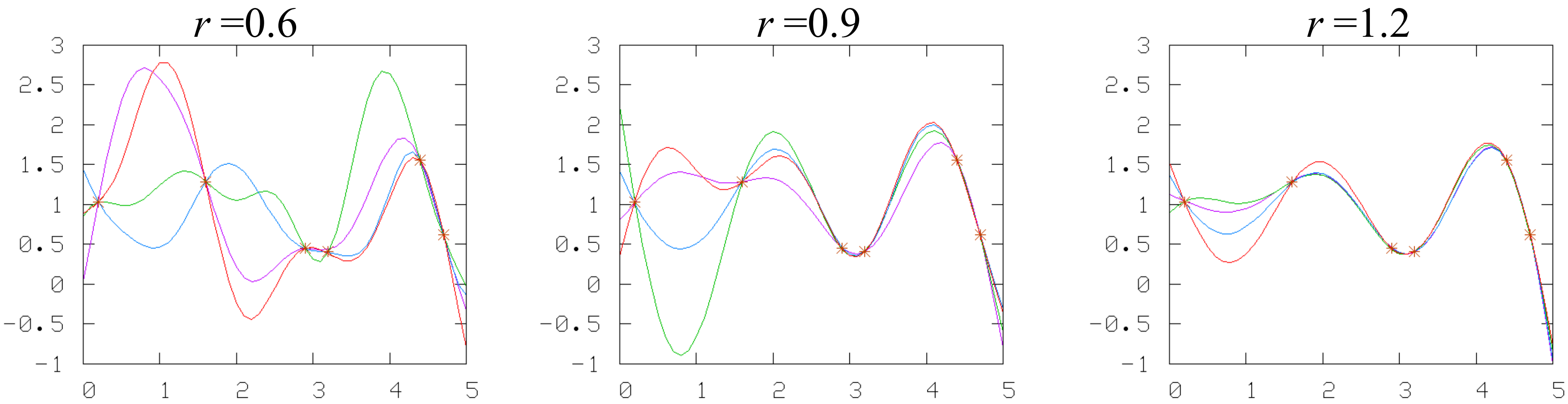
Gaussian Processes: Hyperparameters

- ARD covariance function

$$C(\boldsymbol{x},\boldsymbol{x}') = \theta_1 \exp\left[-\frac{1}{2}\sum_i (x_i - x'_i)^2 / r_i^2\right] + \theta_2$$

automatic relevance determination

- Control fit and smoothness via hyperparameters
  - ➢ θ_3 is a variance of noise in the observed values. Too small value leads to overfitting.
  - ➢ $\{r_i\}$ are length scale parameters.



Gaussian Processes: How to find hyperparameters?

- Use Bayesian inference in hyperparameters space.
  - Posterior for hyperparameters

$$P(\boldsymbol{\theta} \mid D) \propto P(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta})\, P(\boldsymbol{\theta})$$

  - Full integration over all hyperparameters
  - Or choose most probable value $\boldsymbol{\theta}$ that optimizes the marginal log-likelihood

$$\log P(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\theta}) = -\underbrace{\frac{1}{2}\log(\det \boldsymbol{K})}_{\text{Complexity penalty}} - \underbrace{\frac{1}{2}\boldsymbol{Y}^t \boldsymbol{K}^{-1}\boldsymbol{Y}}_{\text{fit}} - \frac{N}{2}\log 2\pi$$

- No need for cross-validation or validation set! Also prevents overtraining.

Gaussian Processes: Make predictions

- Want to make prediction y^* at unseen (test) point $\boldsymbol{x}^*$.

- Predictive distribution is Gaussian with mean and variance:

$$\left\langle y^* \right\rangle = \boldsymbol{k}^t \boldsymbol{K}^{-1} \boldsymbol{Y} \qquad \sigma^{*2} = C(\boldsymbol{x}^*, \boldsymbol{x}^*) - \boldsymbol{k}^t \boldsymbol{K}^{-1} \boldsymbol{k}$$

prediction

Confidence in prediction

  - ➢ $\boldsymbol{k}$ describes covariance of training and new points, $k_n = C(\boldsymbol{x}^*, \boldsymbol{x}^{(n)})$.

- For test set points need to add noise variance to GP variance.

ADME and QSAR modelling:

Techniques for determining hyperparameters

Finding hyperparameters

- Optimize the marginal log-likelihood

$$\log P(\boldsymbol{Y}\,|\,\boldsymbol{X},\boldsymbol{\theta}) = -\frac{1}{2}\log(\det \boldsymbol{K}) - \frac{1}{2}\boldsymbol{Y}^{t}\boldsymbol{K}^{-1}\boldsymbol{Y} - \frac{N}{2}\log 2\pi$$

- Conjugate gradient methods
  - ➢ Computationally demanding. Inversion of matrix NxN at each step, N is a number of compounds in the training set. Comp. cost O(N^3).
  - ➢ The function has multiple maxima. Search can get trapped in a local maximum.

- Need to find simplified approaches.

Techniques for finding hyperparameters

- "Fixed" values.
  - $$r_i = 4\sqrt{M}\,\sigma(\boldsymbol{x}_i), \quad \theta_2 = \sqrt{N}\sigma_{\boldsymbol{Y}} \,,$$

  M is a number of descriptors. Search for $\theta_1,\ \theta_3$.

- Forward variable selection provides feature selection.

- Optimization by conjugate gradient methods (only length scales).
  - Length scales show which descriptors are most relevant.

- Nested sampling.
  - Search in the full hyperparameter space.
  - Search does not get trapped in local maxima.

computational demand

Nested sampling

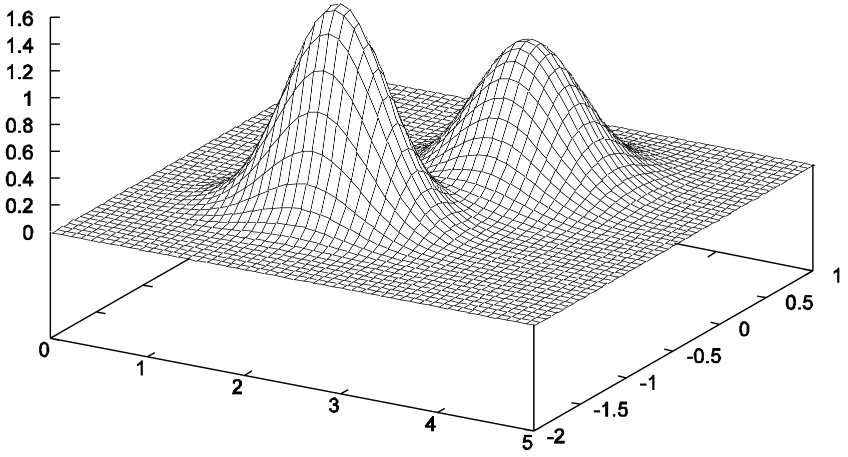
- Method by John Skilling to estimate evidence and generate posterior samples.
  (http://www.inference.phy.cam.ac.uk/bayesys/Valencia.pdf)

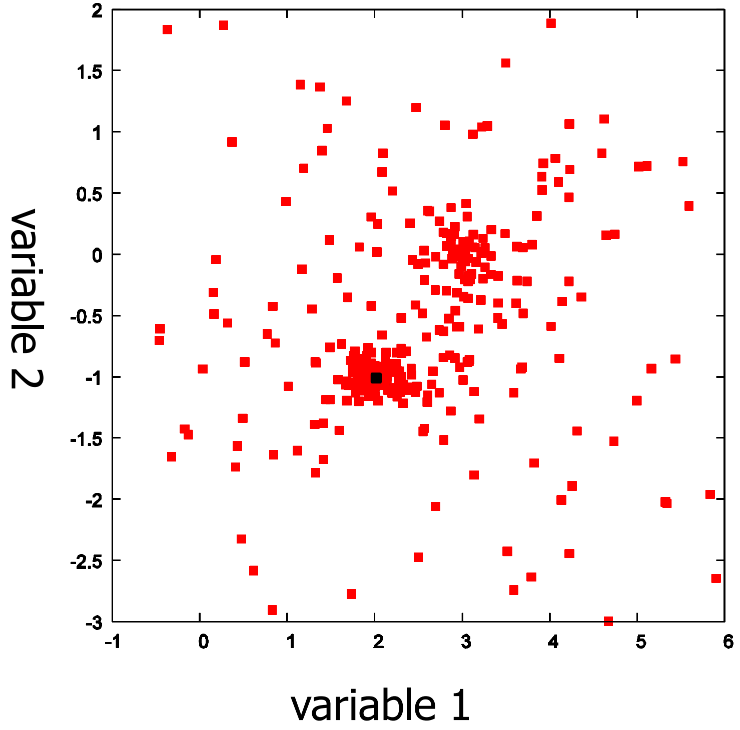
- We want to find most probable hyperparameter values, i.e that give the maximum of the likelihood.

- Key idea:
  - Sample uniformly from wide prior space of all hyperparameters.
  - Iteratively replace samples with low likelihood by new samples with high likelihood.
  - At the end of the process we have points corresponding to high likelihood values.

Nested sampling: Example

- 2 variables.
- Find maximum of likelihood:





ADME and QSAR modelling:

Examples and comparison

Benzodiazepine set

- F. Burden, JCICS 2001, 41, 830-835.
- 245 ligands for the benzodiazepine receptor (in vitro binding affinities as pIC$_{50}$).
- 59 descriptors:
  - Randic and Kier-Hall indices (E-Dragon: www.vcclab.org),
  - counts of atoms, rings and functional groups.
- Test set - 15%.
  - Burden's set split is not known to us.
  - Used set split based on uniform sample of Y values.

Benzodiazepine set: Results

Method	Desc	r^2_{corr}(trn)	r^2_{corr}(test)
PLS	38(3)	0.32	0.53
GP-Basic	38	0.52	0.53
GP-FVS	15	0.52	0.54
GP-Opt	9	0.62	0.51
GP-Nest	38	0.68	0.65
ASNN+kNN	36	0.73	0.64
BRANN	39	0.75	0.71
GPmodel	39	0.76	0.66
GPlinear	39	0.78	0.71

GP-Nest
on test set:
RMSE=0.46
R^2=0.63
r^2_{corr}=0.65

← VCCLAB (www.vcclab.org)

} Burden results

Training set - 208 compounds, test set - 37 compounds.

hERG inhibition set

- Inhibition of human ether-a-go-go related gene by medication.
- 137 compounds with patch-clamp pIC_{50} values.
- 166 descriptors:
  - 2D SMARTS based + logP, PSA, charge, etc.
- Test set - 20%.
  - Set split based on clustering analysis (Tanimoto level = 0.7).

hERG inhibition: Results

Method	Desc	R^2 (trn)	R^2 (test)
PLS	166(2)	0.63	0.74
GP-Basic	166	0.79	0.76
GP-FVS	17	0.76	0.80
GP-Opt	26	0.82	0.81
GP-Nest	166	0.81	0.77
ASNN+kNN	166	0.94	0.77

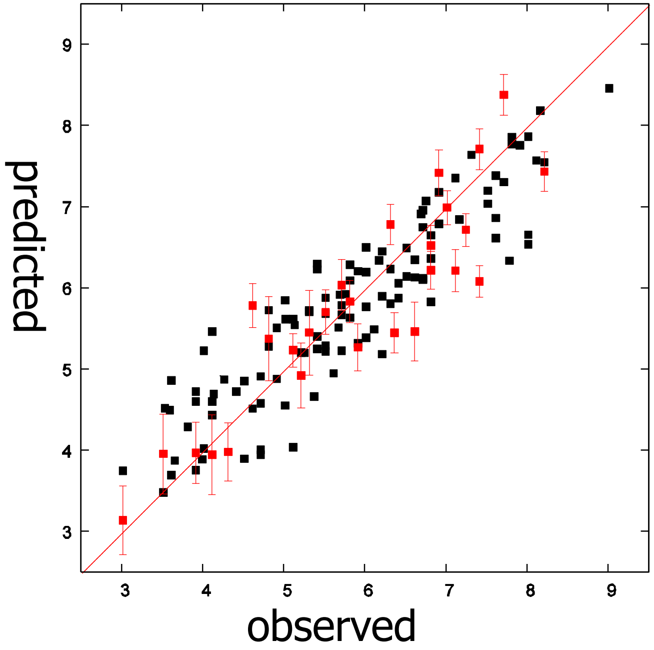
GP-Opt
on test set:
RMSE=0.6
R^2=0.81
r^2_{corr}=0.81

← VCCLAB (www.vcclab.org)

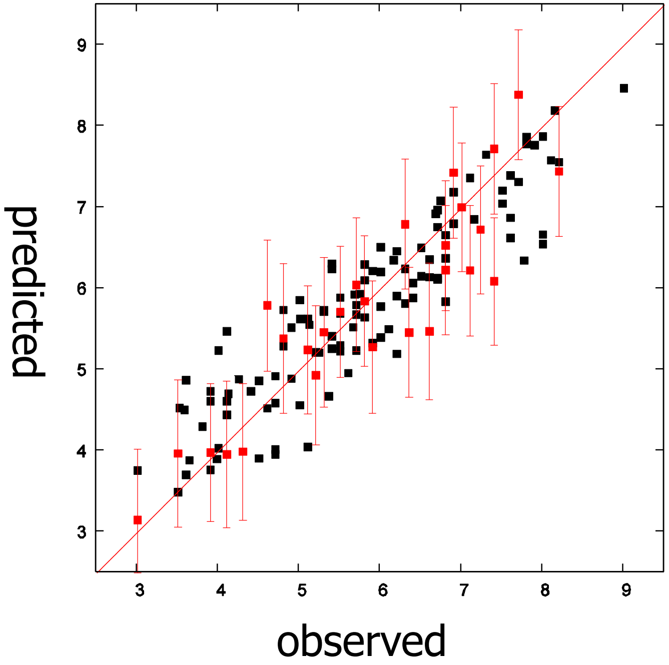
Training set - 110 compounds,
test set - 27 compounds.

hERG inhibition model

Predicted pIC_{50} values versus observed with error bars.
Training set in black. Test set in red.



- Original GP error bars, do not include experimental noise variance
- Applicability of the model



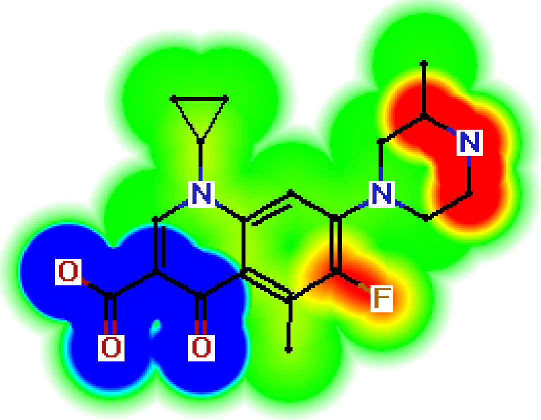
- Error bars include noise variance
- Confidence in prediction

ADMEnsa
Bringing balance to optimization

BioFocus DPI
A Galápagos Company

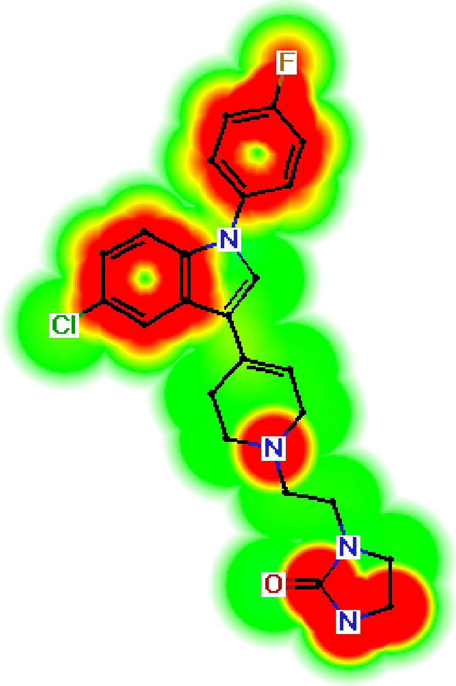
hERG inhibition model: Descriptors

- Important features:
  - Lipophilicity
  - Negative charge
  - Positively charged nitrogen at pH 7.4
  - Aromaticity index
  - HB donor – acceptor pairs separated by 6 bonds
  - Ketone
  - Amide



hERG pIC_{50} obs. = 4.3

predicted= 3.99 ±0.84

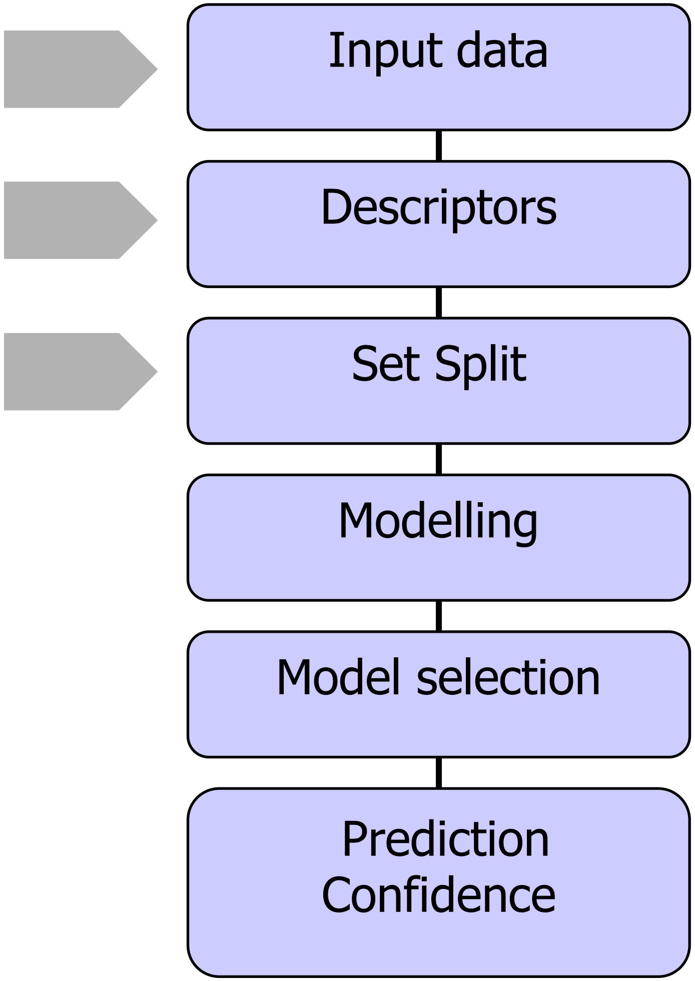


hERG pIC_{50} obs. = 8

predicted= 7.88 ±0.8

Automatic modelling process

Automatic Model Generation Process


Input data

Descriptors

Set Split

Modelling

Model selection

Prediction Confidence

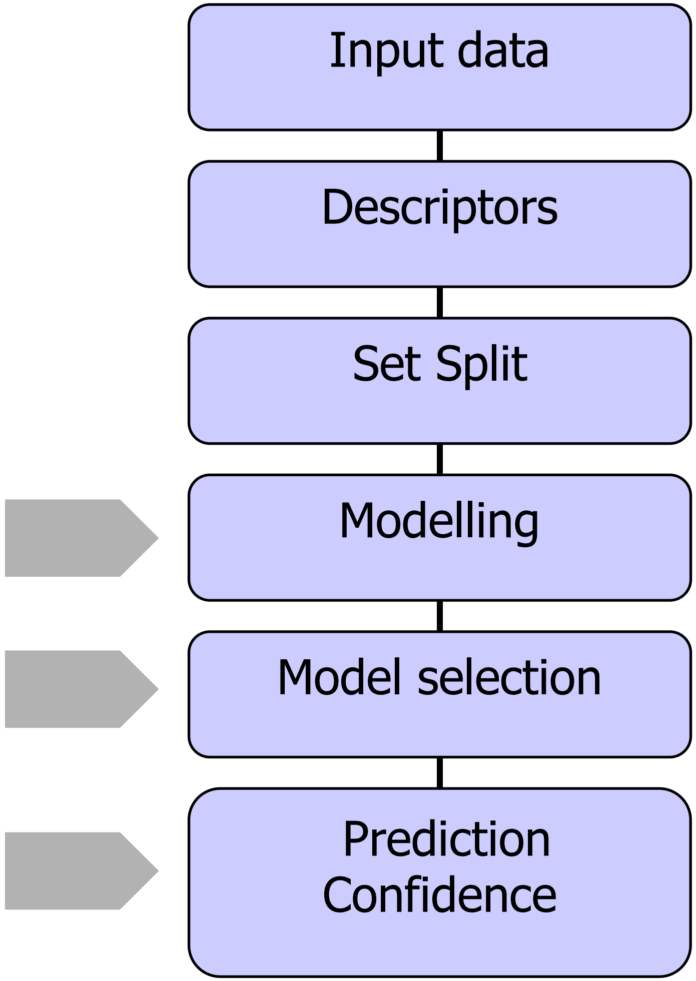
- User provides structures and property values.

- 2D SMARTS based descriptors and logP, flexibility, charge, PSA, etc. A user can import own descriptors.

- Split into 3 sets:
  - training (building a model),
  - validation (model selection),
  - test (independent).

- Clustering by structural similarity or Y – based. Or user's own split.

Automatic Model Generation Process



- Modelling continuous data:
  - PLS
  - Gaussian Processes (5 techniques)
  - Radial Basis Functions + GA

  categorical data:
  - Decision trees (C4.5)

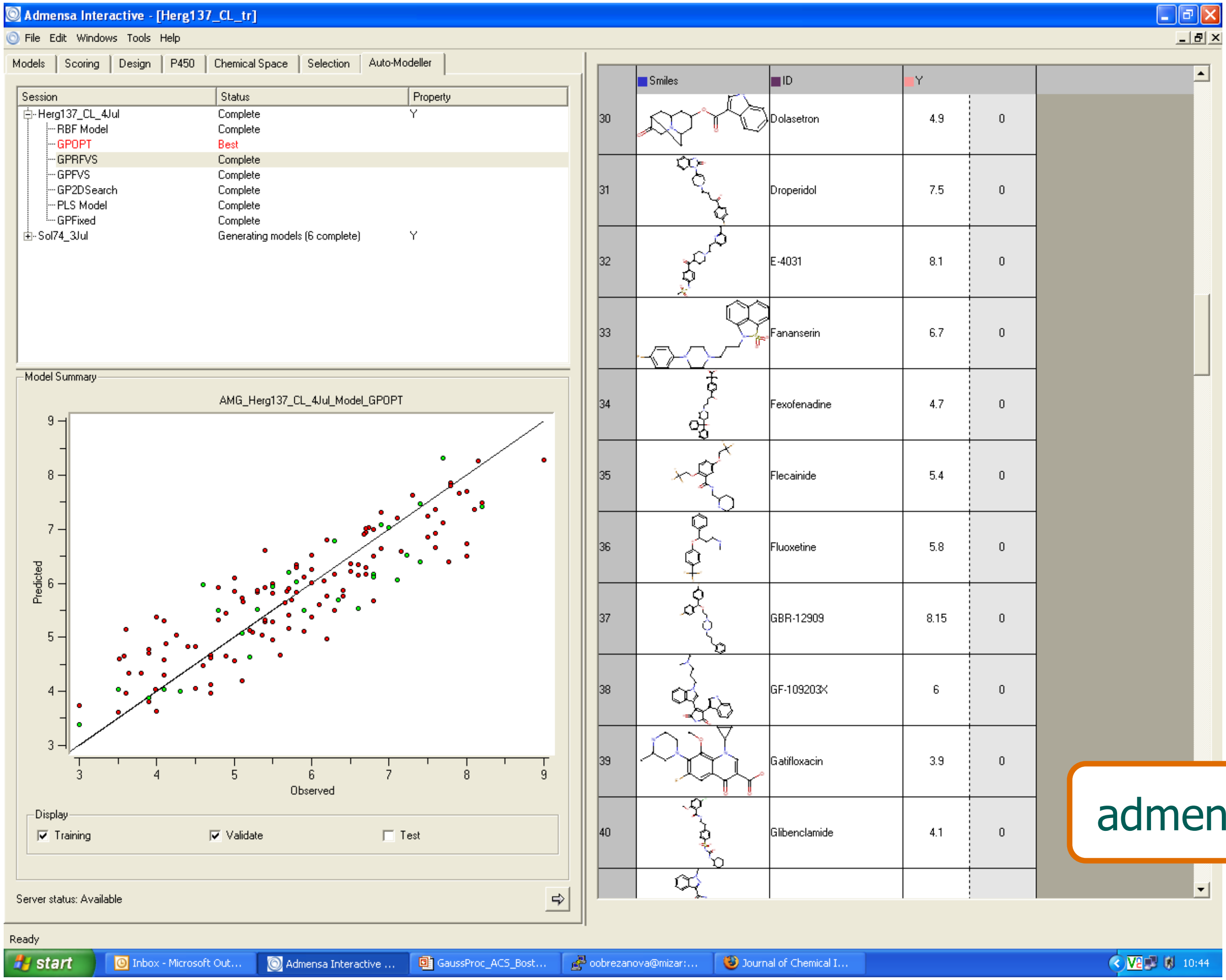
- Best model selection is based on performance of validation set.

- Estimation of uncertainty for each prediction.

ADMEnsa
Bringing balance to optimization

BioFocus DPI
A Galápagos Company

ADMEnsa Interactive. Auto-Modeler.



Session	Status	Property
Herg137_CL_4Jul	Complete	Y
RBF Model	Complete	
GPOPT	Best	
GPRFVS	Complete	
GPFVS	Complete	
GP2DSearch	Complete	
PLS Model	Complete	
GPFixed	Complete	
Sol74_3Jul	Generating models (6 complete)	Y

☑ Training ☑ Validate □ Test

Server status: Available

Ready

	Smiles	ID	Y	
30		Dolasetron	4.9	0
31		Droperidol	7.5	0
32		E-4031	8.1	0
33		Fananserin	6.7	0
34		Fexofenadine	4.7	0
35		Flecainide	5.4	0
36		Fluoxetine	5.8	0
37		GBR-12909	8.15	0
38		GF-109203X	6	0
39		Gatifloxacin	3.9	0
40		Glibenclamide	4.1	0

admensa-support@glpg.com

ADMEnsa
Bringing balance to optimization

BioFocus DPI
A Galápagos Company

Conclusions

- Gaussian Processes is a powerful nonlinear modelling technique:
  - No *a priori* determination of model parameters.
  - Built-in tool to prevent overtraining, no need for cross-validation.
  - Works well for a big pool of descriptors.
  - Identifies relevant descriptors.
  - Uncertainty with each prediction.

- Application to building QSAR and ADME models. New techniques for determining model parameters.

- Automatic model generation process accessible through an intuitive desktop environment.

References

- The Gaussian Processes Website. www.gaussianprocess.org
- D. MacKay. Information Theory, Inference, and Learning Algorithms. Cambridge University Press, 2003.
- C. Rasmussen, C. Williams. Gaussian Processes for Machine Learning. The MIT Press, 2006.
- Obrezanova et al. *J. Chem. Inf. Model.* E-publication ahead of print, 28 June, 2007.

Acknowledgements:

- Gábor Csányi (Cavendish Laboratory, University of Cambridge)
- Joelle Gola
- Matthew Segall
- Ed Champness
- Chris Leeding
- Andre Kramer

Spare slides

Comparison: hERG inhibition set

Method	Desc	R^2 (trn)	R^2 (test)	Time (min)
PLS	166(2)	0.63	0.74	0.2
RBF-GA	21	1	0.77	
GP-Basic	166	0.79	0.76	2.3
GP-FVS	17	0.76	0.80	19
GP-Opt	26	0.82	0.81	13
GP-Nest	166	0.81	0.77	170
ASNN	166	0.94	0.69	188
ASNN+kNN	166	0.94	0.77	

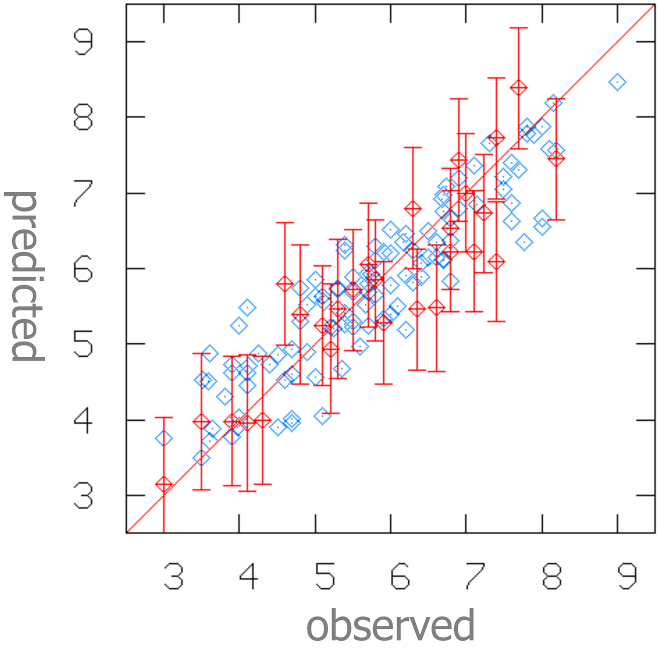
VCCLAB

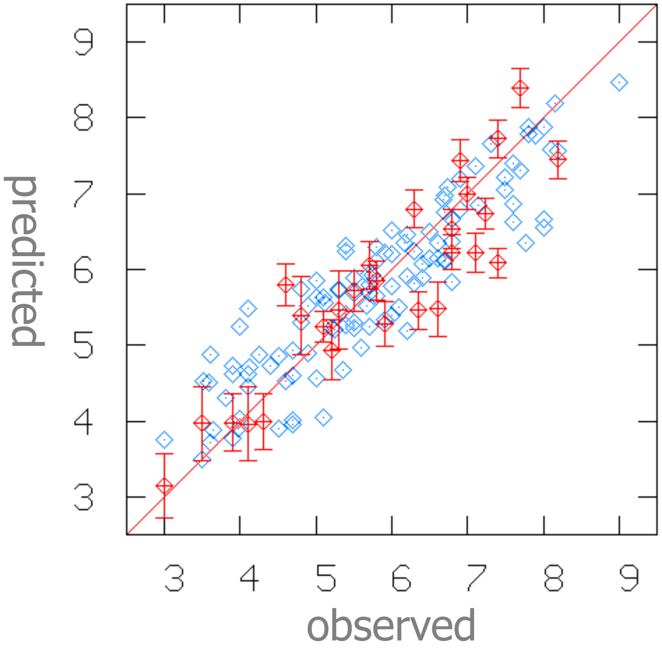
GP-Opt
on test set:
RMSE=0.6
R^2=0.81
R^2corr=0.8
1

Training set - 110 compounds,
test set - 27 compounds.

hERG inhibition model

Predicted pIC50 values versus observed with errorbars.
Training set in blue. Test set in red.





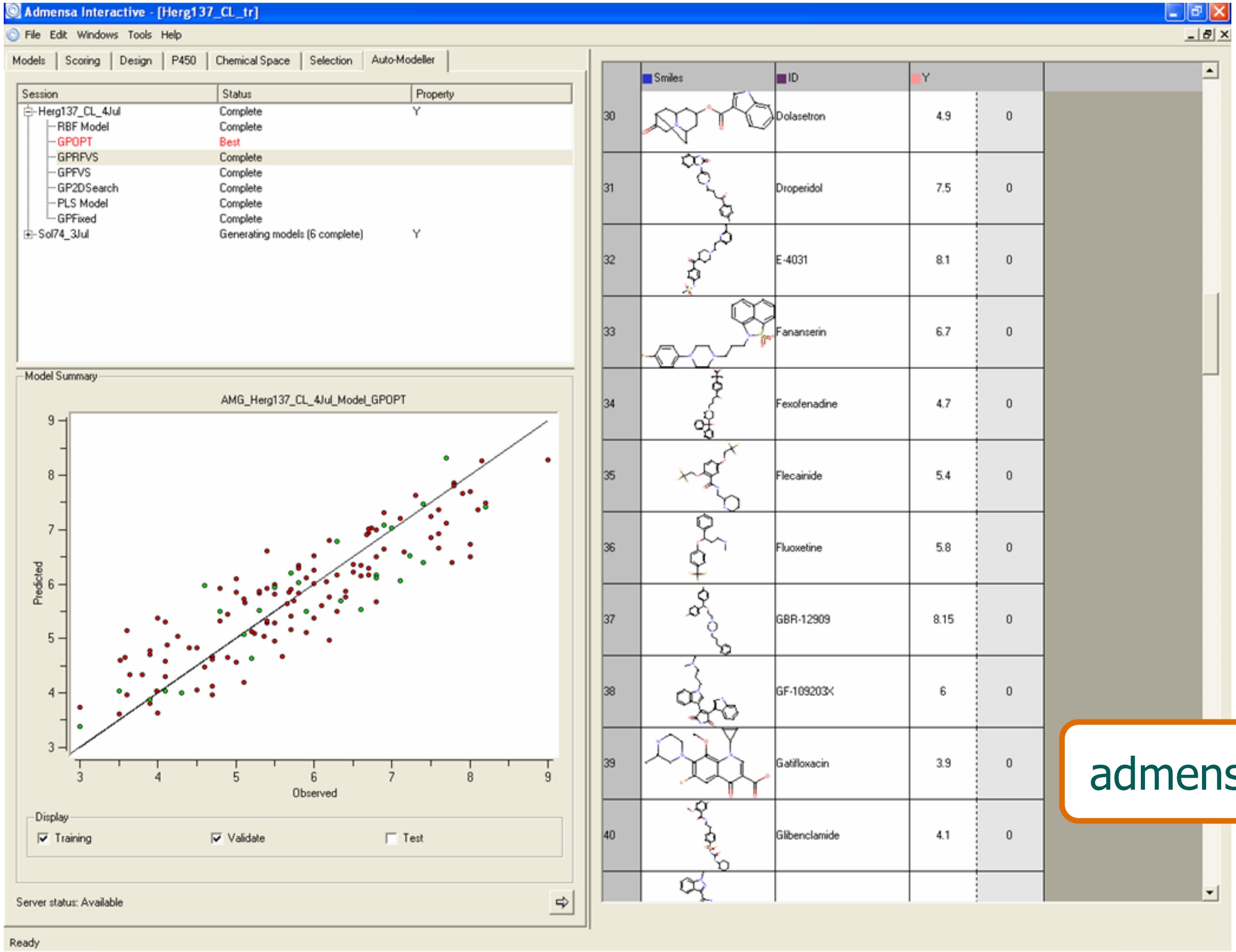
- Error bars include noise variance
- Confidence in prediction

- Original GP error bars, do not include experimental noise variance
- Applicability of the model

ADMEnsa
Bringing balance to optimization

BioFocus DPI
A Galápagos Company

Admensa Interactive. Auto-Modeller.



Session	Status	Property
Herg137_CL_4Jul	Complete	Y
RBF Model	Complete	
GPOPT	Best	
GPRFVS	Complete	
GPFVS	Complete	
GP2DSearch	Complete	
PLS Model	Complete	
GPFixed	Complete	
Sol74_3Jul	Generating models (6 complete)	Y

☑ Training ☑ Validate □ Test

Server status: Available

Ready

	Smiles	ID	Y	
30		Dolasetron	4.9	0
31		Droperidol	7.5	0
32		E-4031	8.1	0
33		Fananserin	6.7	0
34		Fexofenadine	4.7	0
35		Flecainide	5.4	0
36		Fluoxetine	5.8	0
37		GBR-12909	8.15	0
38		GF-109203X	6	0
39		Gatifloxacin	3.9	0
40		Glibenclamide	4.1	0

admensa-support@glpg.com

ADMEnsa
Bringing balance to optimization

BioFocus DPI
A Galápagos Company

Gaussian Processes: Practical steps

- Structure of functions determined by covariance (kernel) function:

$$\mathrm{cov}(f(\boldsymbol{x}), f(\boldsymbol{x}')) = C(\boldsymbol{x}, \boldsymbol{x}')$$

- Distribution of functions is multivariate Gaussian with zero mean and covariance matrix

$$\boldsymbol{K} = \boldsymbol{C} + \theta_3 \boldsymbol{I}$$

- ARD covariance function (automatic relevance determination)

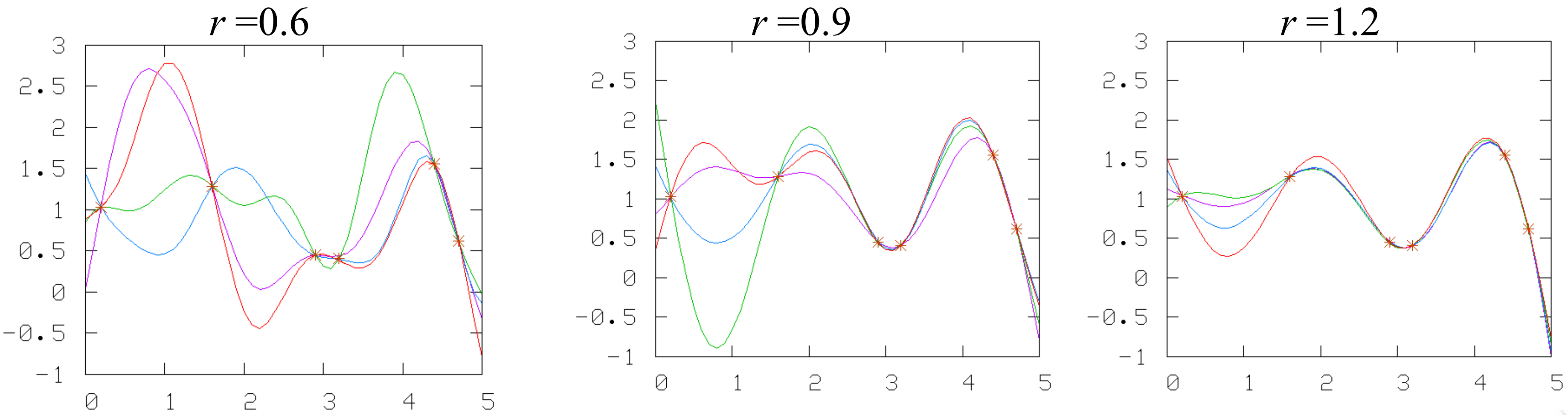
$$C(\boldsymbol{x}, \boldsymbol{x}') = \theta_1 \exp\left[-\frac{1}{2} \sum_i (x_i - x'_i)^2 / r_i^2 \right] + \theta_2$$

- Control fit and smoothness via hyperparameters.
  - θ_3 is a variance of noise present in the observed values.
  - $\{r_i\}$ are length scale parameters.

Gaussian Processes: Hyperparameters

- Noise variance θ_3 : too small value leads to overtraining.

- Length scale parameters $\{r_i\}$: large values mean that corresponding descriptor does not influence the property values very much. Automatic relevance determination.


r =0.6

r =0.9

r =1.2