

# Gaussian Processes: A method for automatic QSAR modelling of ADME properties?

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

11 October 2007

Autumn 2007 UK-QSAR meeting

**BioFocusDPI**  
A Galápagos Company

*Copyright © 2007 Galapagos NV*



# 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  $f$ :

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

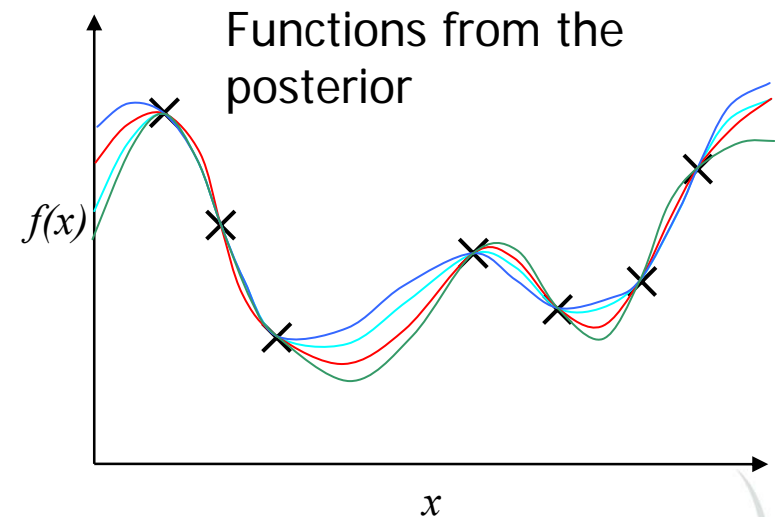
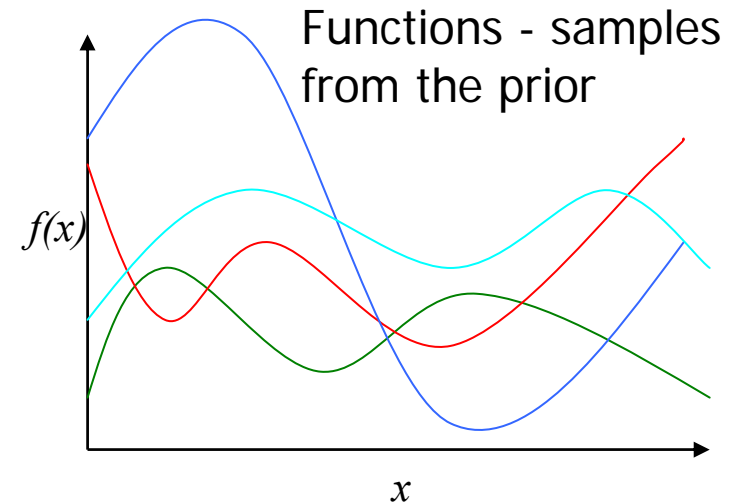
- Bayesian rule

$$P(f | D) \propto P(D | f) P(f)$$

posterior

prior

- Prediction is a mean of posterior distribution.
- Gaussian Process defines a distribution over functions.





# Gaussian Processes: Practical steps

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

$$\text{cov}(f(\mathbf{x}), f(\mathbf{x}')) = C(\mathbf{x}, \mathbf{x}')$$

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

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

- Hyperparameter  $\theta_3$  is a variance of noise present in the observed values.

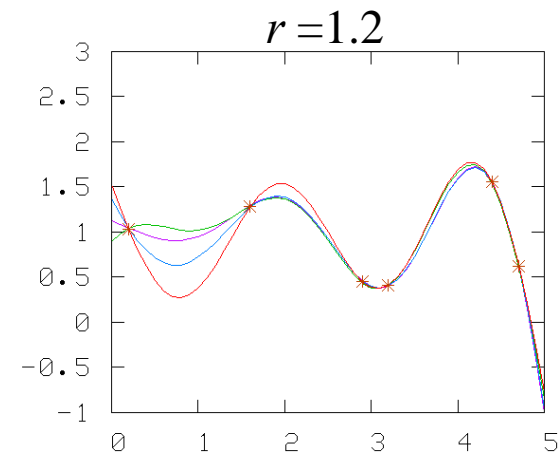
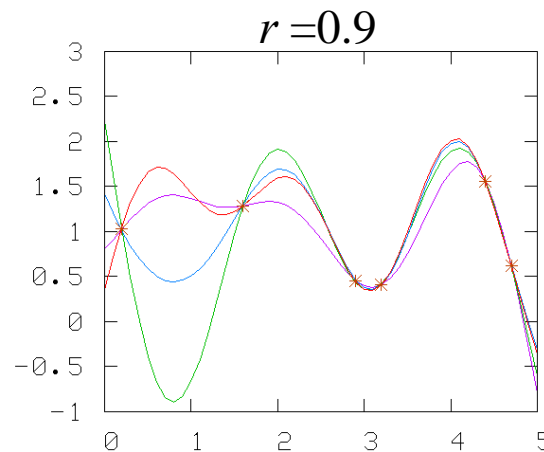
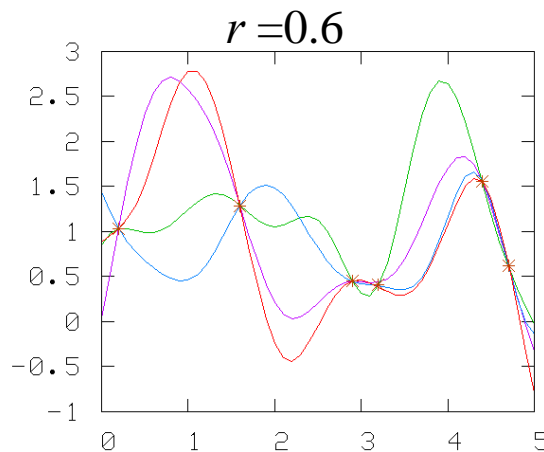
# Gaussian Processes: Hyperparameters

- ARD covariance function

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

automatic relevance  
determination

- Control fit and smoothness via hyperparameters
  - $\theta_3$  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

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

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

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

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



# Gaussian Processes: Make predictions

- Want to make prediction  $y^*$  at unseen (test) point  $\mathbf{x}^*$ .
- Predictive distribution is Gaussian with mean and variance:

$$\langle y^* \rangle = \mathbf{k}^t \mathbf{K}^{-1} \mathbf{Y}$$

prediction

$$\sigma^{*2} = C(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}^t \mathbf{K}^{-1} \mathbf{k}$$

Confidence in prediction

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

- 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

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

- 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.

- $r_i = 4\sqrt{M} \sigma(\mathbf{x}_i), \quad \theta_2 = \sqrt{N} \sigma_Y,$

- $M$  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





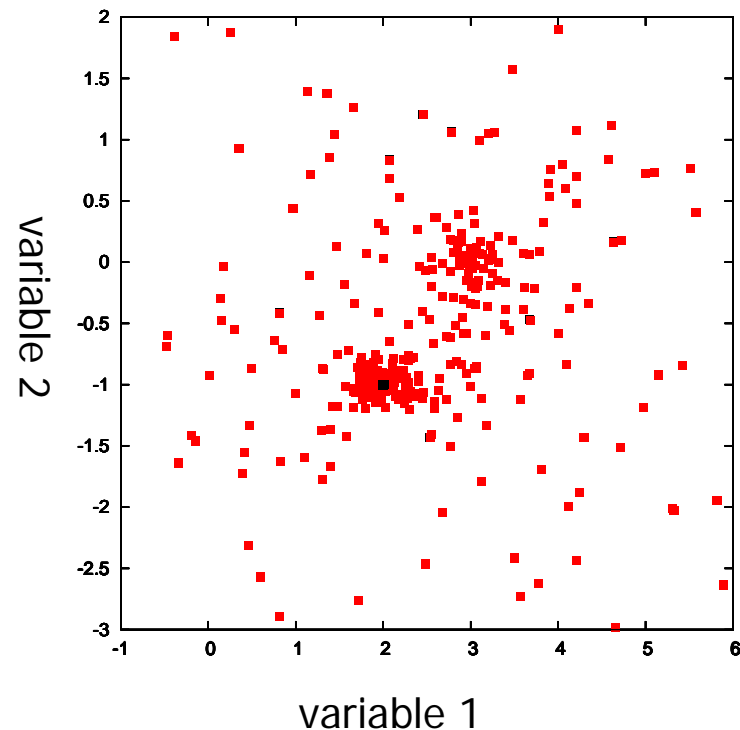
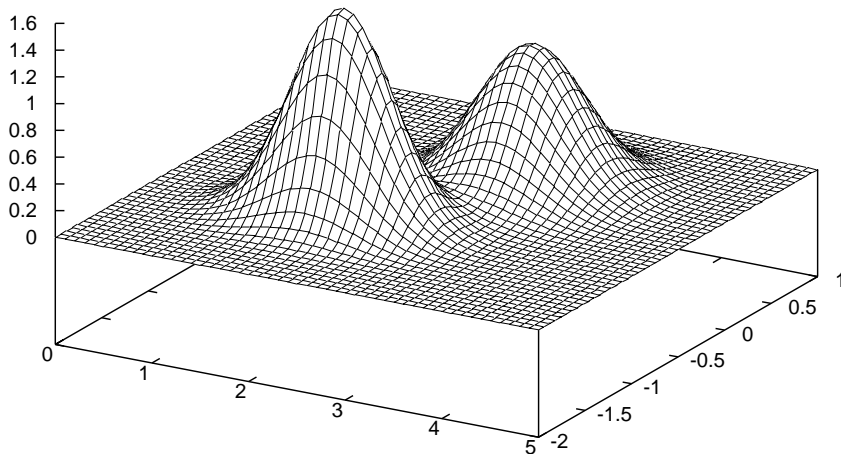
# Nested sampling

- 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](http://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	$r^2_{\text{corr}}$ (trn)	$r^2_{\text{corr}}$ (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
<b>GP-Nest</b>	<b>38</b>	<b>0.68</b>	<b>0.65</b>
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  
 $R^2=0.63$   
 $r^2_{\text{corr}}=0.65$

← VCCLAB ([www.vcclab.org](http://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 <sup>2</sup> (trn)	R <sup>2</sup> (test)
PLS	166(2)	0.63	0.74
GP-Basic	166	0.79	0.76
GP-FVS	17	0.76	0.80
<b>GP-Opt</b>	<b>26</b>	<b>0.82</b>	<b>0.81</b>
GP-Nest	166	0.81	0.77
ASNN+kNN	166	0.94	0.77

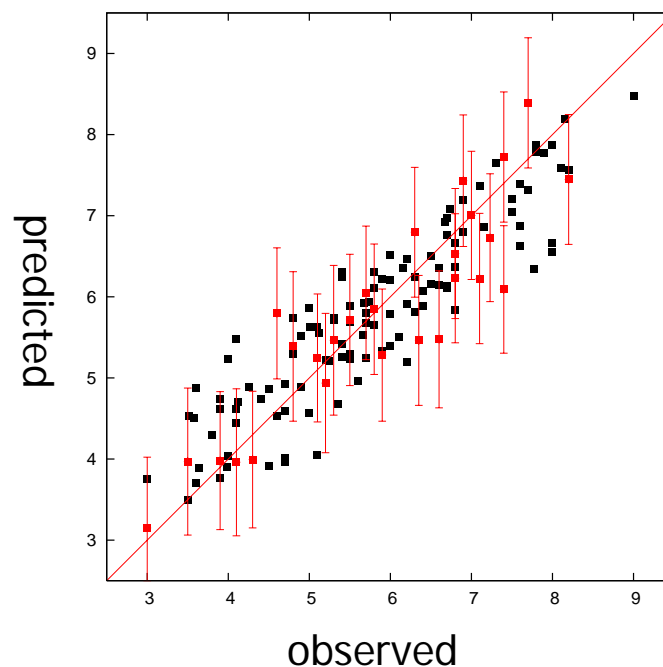
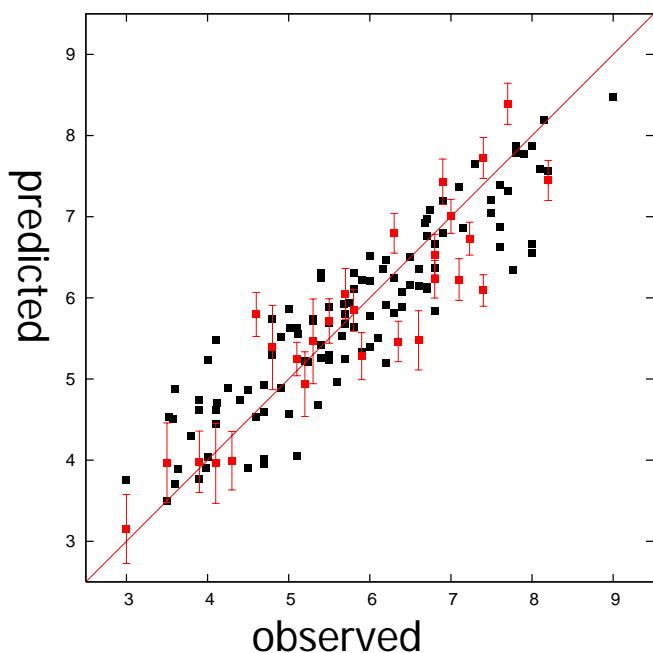
GP-Opt  
on test set:  
RMSE=0.6  
R<sup>2</sup>=0.81  
r<sup>2</sup><sub>corr</sub>=0.81

← VCCLAB ([www.vcclab.org](http://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.

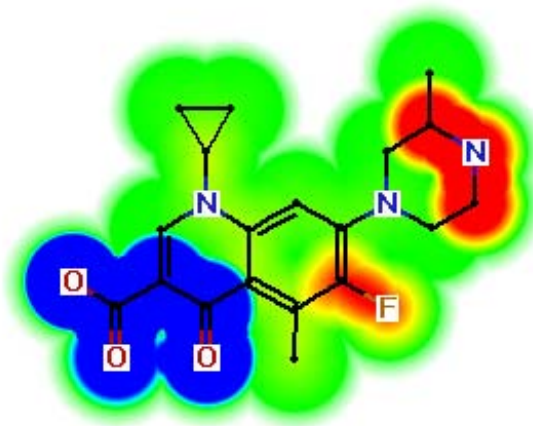


- Original GP error bars, do not include experimental noise variance
- **Applicability of the model**

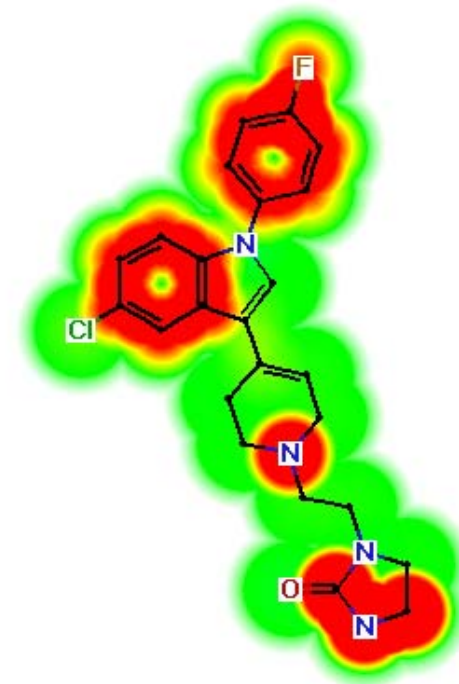
- Error bars include noise variance
- **Confidence in prediction**

# 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 =  $3.99 \pm 0.84$



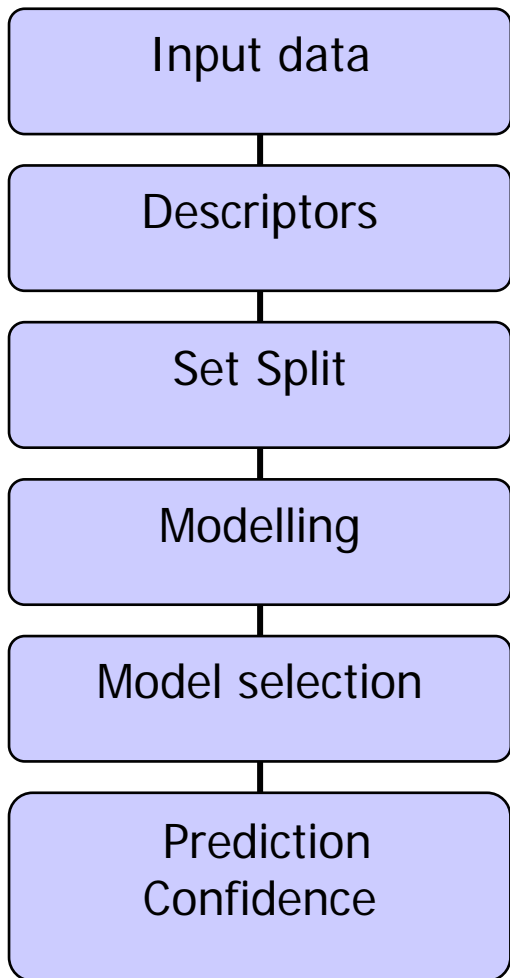
hERG  $pIC_{50}$  obs. = 8  
predicted =  $7.88 \pm 0.8$



# Automatic modelling process



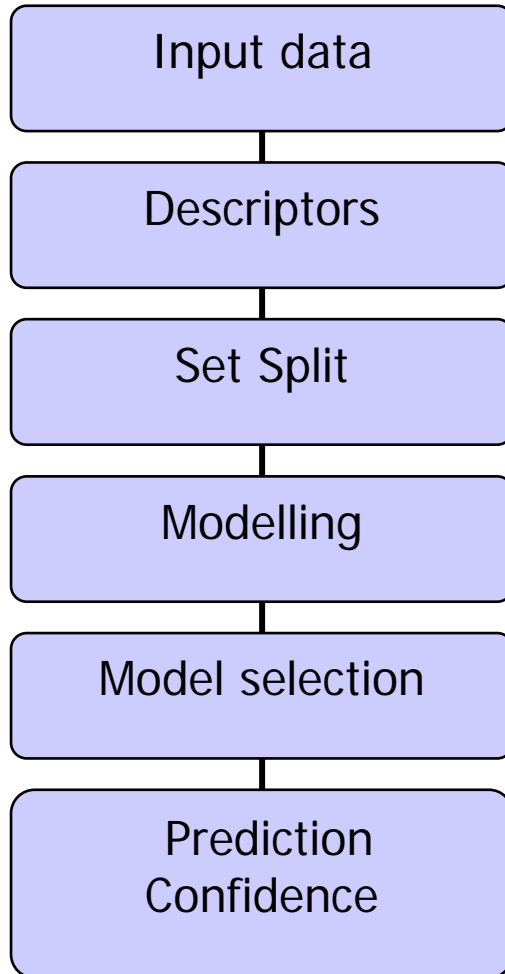
# Automatic Model Generation Process



- 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.

# ADMEEnsa Interactive. Auto-Modeler.

The screenshot displays the Admensa Interactive software interface. The top menu bar includes File, Edit, Windows, Tools, and Help. Below the menu, there are tabs for Models, Scoring, Design, P450, Chemical Space, Selection, and Auto-Modeler. The main window is divided into several sections:

- Session List:** A table showing the status of various modeling sessions. The 'Best' model is highlighted in red.
- Model Summary:** A scatter plot titled 'AMG\_Herg137\_CL\_4Jul\_Model\_GPOPT' showing Predicted vs. Observed values. The plot includes a diagonal line representing a perfect fit and data points in red and green. Below the plot are checkboxes for 'Training', 'Validate', and 'Test'.
- Compound List:** A table listing 11 compounds with their SMILES, IDs, and predicted values (Y).

Smiles	ID	Y
<chem>CC1=CC=C(C=C1)C2=CC=CC=C2C3=CC=CC=C3C4=CC=CC=C4C5=CC=CC=C5C6=CC=CC=C6C7=CC=CC=C7C8=CC=CC=C8C9=CC=CC=C9</chem>	Dolasetron	4.9
<chem>CC1=CC=C(C=C1)C2=CC=CC=C2C3=CC=CC=C3C4=CC=CC=C4C5=CC=CC=C5C6=CC=CC=C6C7=CC=CC=C7C8=CC=CC=C8C9=CC=CC=C9</chem>	Dropeidol	7.5
<chem>CC1=CC=C(C=C1)C2=CC=CC=C2C3=CC=CC=C3C4=CC=CC=C4C5=CC=CC=C5C6=CC=CC=C6C7=CC=CC=C7C8=CC=CC=C8C9=CC=CC=C9</chem>	E-4031	8.1
<chem>CC1=CC=C(C=C1)C2=CC=CC=C2C3=CC=CC=C3C4=CC=CC=C4C5=CC=CC=C5C6=CC=CC=C6C7=CC=CC=C7C8=CC=CC=C8C9=CC=CC=C9</chem>	Fananserin	6.7
<chem>CC1=CC=C(C=C1)C2=CC=CC=C2C3=CC=CC=C3C4=CC=CC=C4C5=CC=CC=C5C6=CC=CC=C6C7=CC=CC=C7C8=CC=CC=C8C9=CC=CC=C9</chem>	Fexofenadine	4.7
<chem>CC1=CC=C(C=C1)C2=CC=CC=C2C3=CC=CC=C3C4=CC=CC=C4C5=CC=CC=C5C6=CC=CC=C6C7=CC=CC=C7C8=CC=CC=C8C9=CC=CC=C9</chem>	Flecainide	5.4
<chem>CC1=CC=C(C=C1)C2=CC=CC=C2C3=CC=CC=C3C4=CC=CC=C4C5=CC=CC=C5C6=CC=CC=C6C7=CC=CC=C7C8=CC=CC=C8C9=CC=CC=C9</chem>	Fluoxetine	5.8
<chem>CC1=CC=C(C=C1)C2=CC=CC=C2C3=CC=CC=C3C4=CC=CC=C4C5=CC=CC=C5C6=CC=CC=C6C7=CC=CC=C7C8=CC=CC=C8C9=CC=CC=C9</chem>	GBR-12909	8.15
<chem>CC1=CC=C(C=C1)C2=CC=CC=C2C3=CC=CC=C3C4=CC=CC=C4C5=CC=CC=C5C6=CC=CC=C6C7=CC=CC=C7C8=CC=CC=C8C9=CC=CC=C9</chem>	GF-109203X	6
<chem>CC1=CC=C(C=C1)C2=CC=CC=C2C3=CC=CC=C3C4=CC=CC=C4C5=CC=CC=C5C6=CC=CC=C6C7=CC=CC=C7C8=CC=CC=C8C9=CC=CC=C9</chem>	Galifloxacin	3.9
<chem>CC1=CC=C(C=C1)C2=CC=CC=C2C3=CC=CC=C3C4=CC=CC=C4C5=CC=CC=C5C6=CC=CC=C6C7=CC=CC=C7C8=CC=CC=C8C9=CC=CC=C9</chem>	Glibenclamide	4.1

Server status: Available

Ready

[admensa-support@glpg.com](mailto:admensa-support@glpg.com)



# 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](http://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.* 47 (5), 2007, pp.1847-1857.



# Acknowledgements:

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