Predicting Properties of Small Molecules with Kernel-Based Machine Learning Methods

berlin

Klaus-Robert Müller, Matthias Rupp, Katja Hansen, Timon Schroeter, Gisbert Schneider et al.

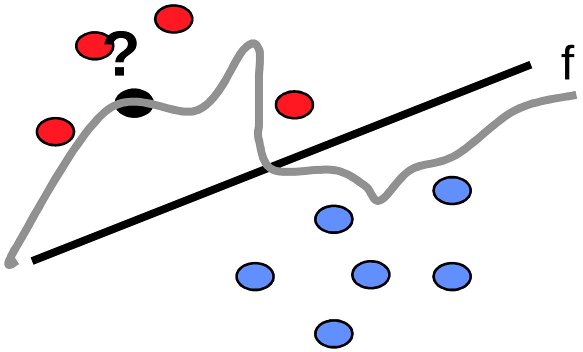
ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Boehringer
Ingelheim

idalab intelligent data analysis

Bayer HealthCare
Bayer Schering Pharma

Machine Learning in a nutshell



Typical scenario: learning from data

- given data set **X** and labels **Y** (generated by some joint probabilty distribution p(x,y))

- **LEARN/INFER** underlying **unknown** mapping

$$Y = f(X)$$

Example: distinguish toxic and non-toxic compounds, metabolically stable compounds ...

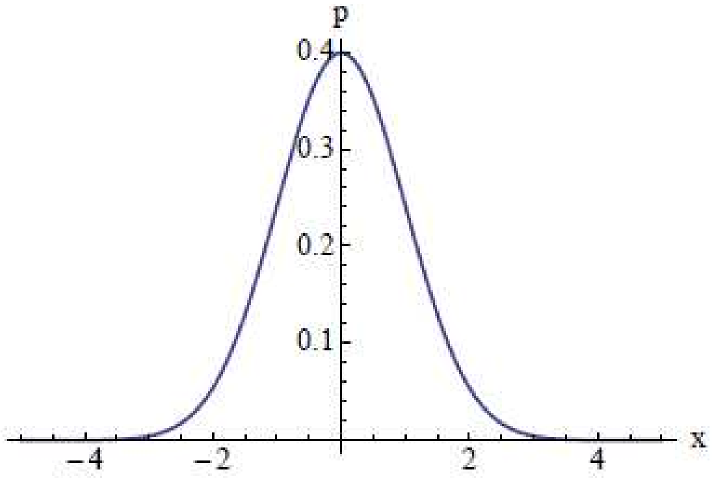
BUT: how to do this optimally with good performance on **unseen** data?

Gaussian Processes

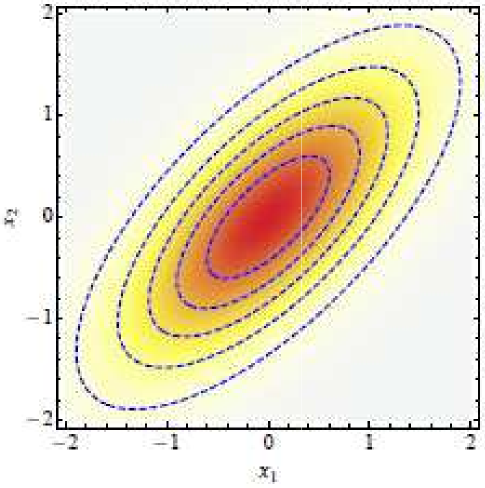
Formal: A *Gaussian process* is a collection of random variables, any finite number of which have a joint Gaussian distribution.

Informal: A generalization of normally distributed random variables to functions.

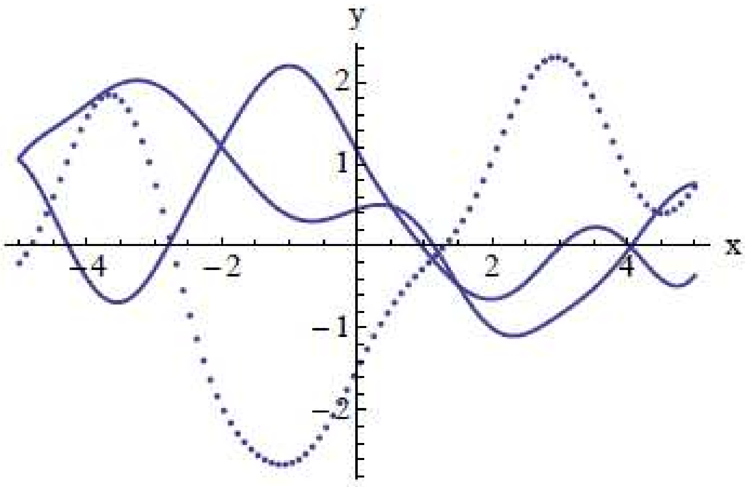
$$\mathcal{N}(\mu, \sigma)$$
$$\mu \in \mathbb{R}$$
$$\sigma \in \mathbb{R}$$



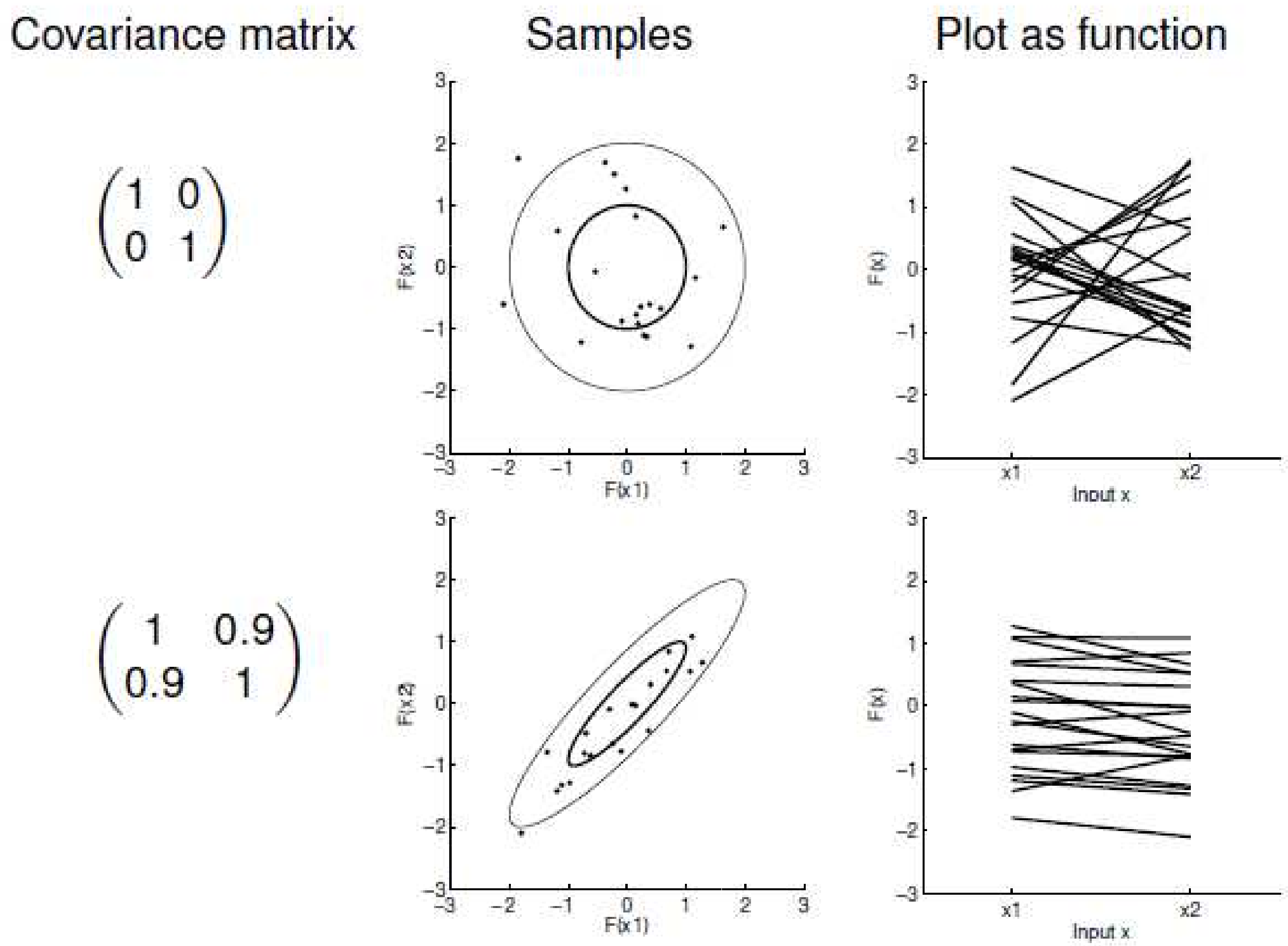
$$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$\boldsymbol{\mu} \in \mathbb{R}^d$$
$$\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$$



$$\mathcal{GP}(\mu, k)$$
$$\mu : \mathcal{X} \mapsto \mathbb{R}$$
$$k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$$



Gaussian Process in 2-dim



Covariance matrix

Samples

Plot as function

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

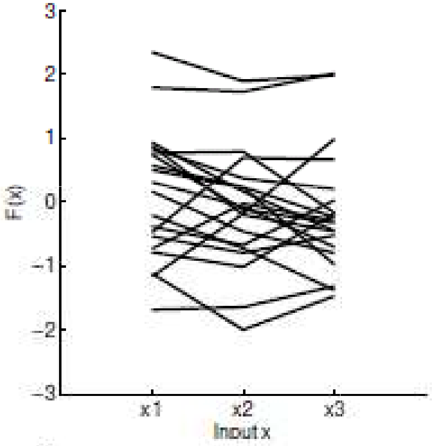
$$\begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$$

Gaussian Process in 3-dim

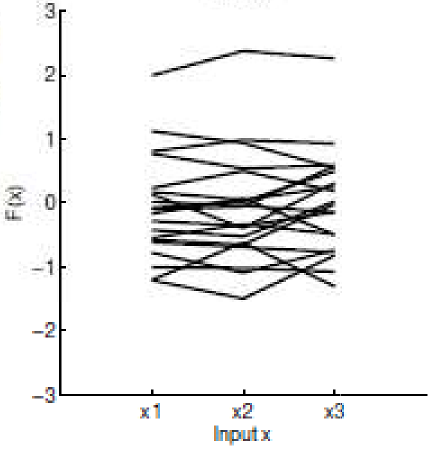
Covariance matrix

Plot as function

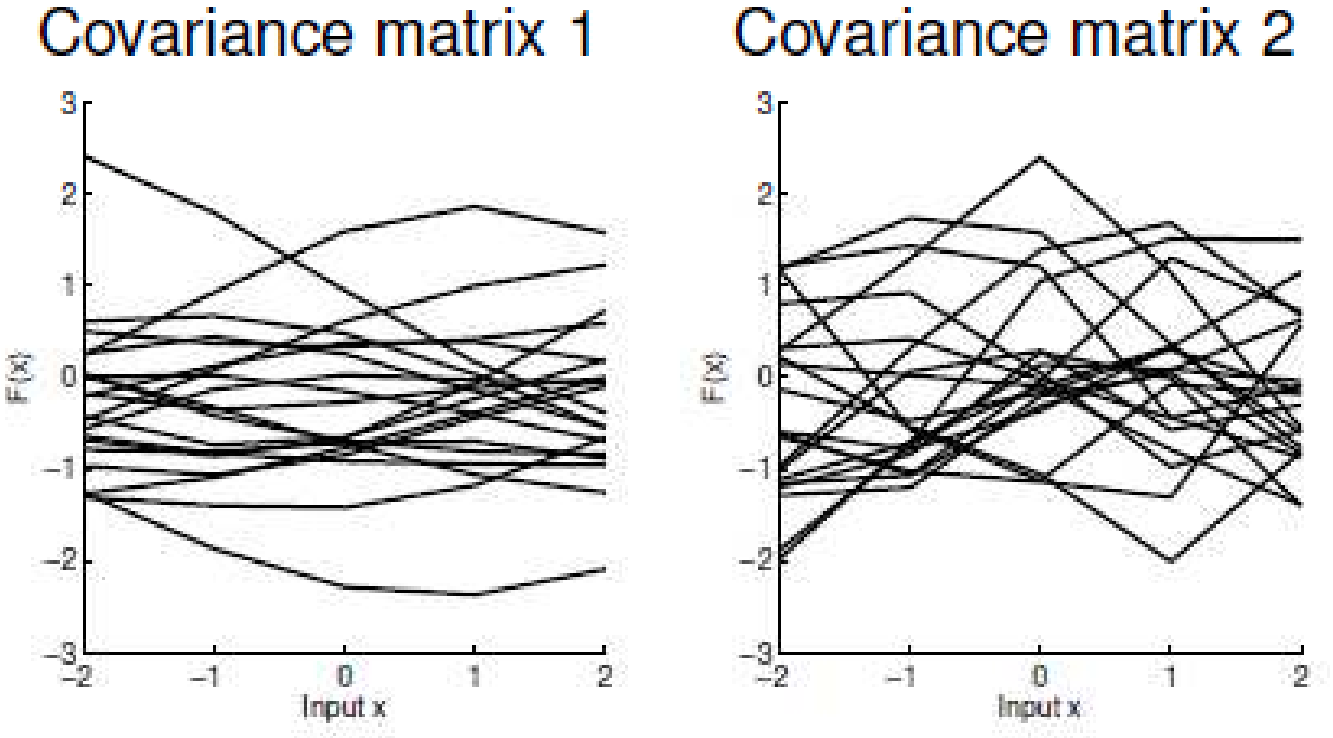
$$\begin{pmatrix} 1 & 0.8 & 0.6 \\ 0.8 & 1 & 0.8 \\ 0.6 & 0.8 & 1 \end{pmatrix}$$



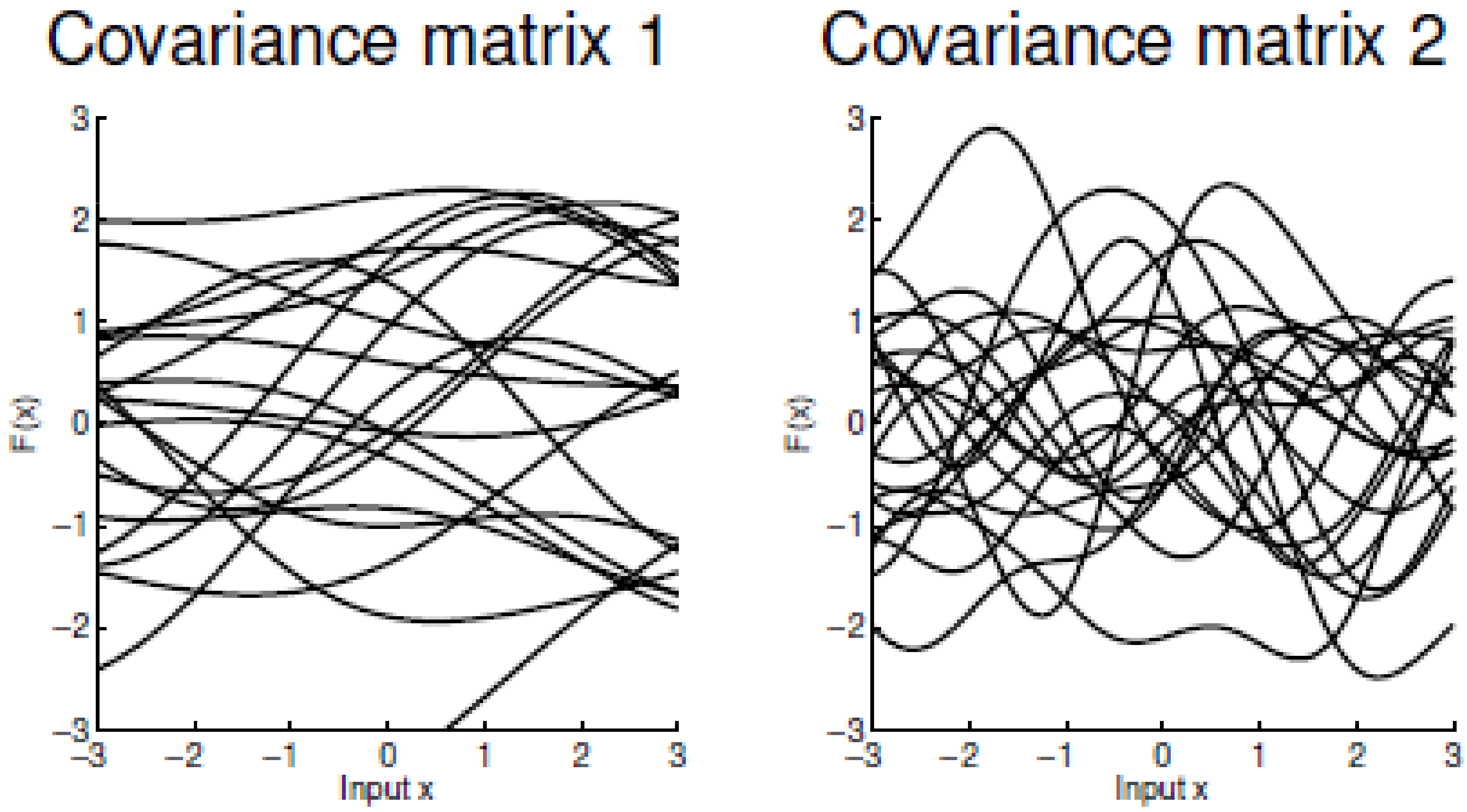
$$\begin{pmatrix} 1 & 0.95 & 0.9 \\ 0.95 & 1 & 0.95 \\ 0.9 & 0.95 & 1 \end{pmatrix}$$



Gaussian Process in 5-dim



Gaussian Process in 100-dim



And here is the GP

Specify prior over functions by specifying a covariance matrix K:

- Function on N points, $x_1, \dots, x_N$
- Covariance function k (“kernel function”)

$$k(x, x') = \operatorname{cov}\,[f(x), f(x')]$$

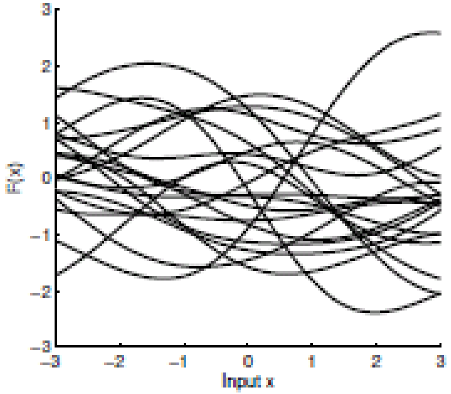
- Functional values $f(x_1), \dots, f(x_N)$ follow an N-variate Gaussian:

$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_N) \end{pmatrix} \sim \mathcal{N}(0, K)$$

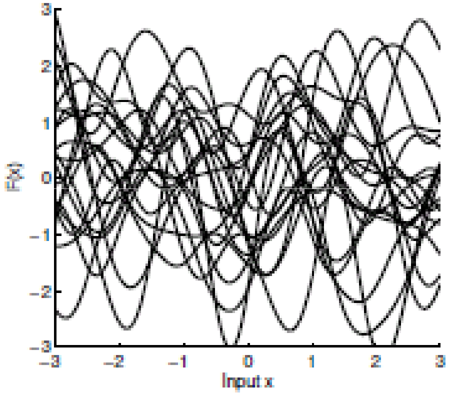
with $K_{i,j} = k(x_i, x_j)$

Covariance Functions

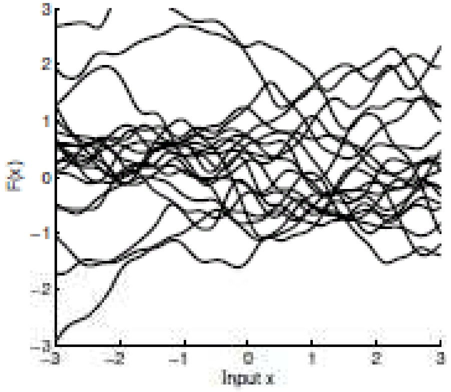
$$k(x,x') = \exp(-0.25(x-x')^2)$$



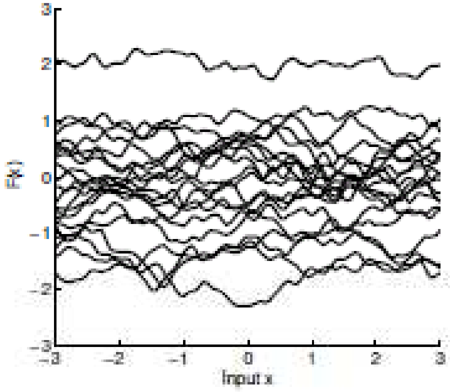
$$k(x,x') = \exp(-4(x-x')^2)$$



$$k(x,x') = (1+(x-x')^2)^{-0.1}$$



$$k(x,x') = (1+(x-x')^2)^{-0.01}$$



Gaussian Process Models

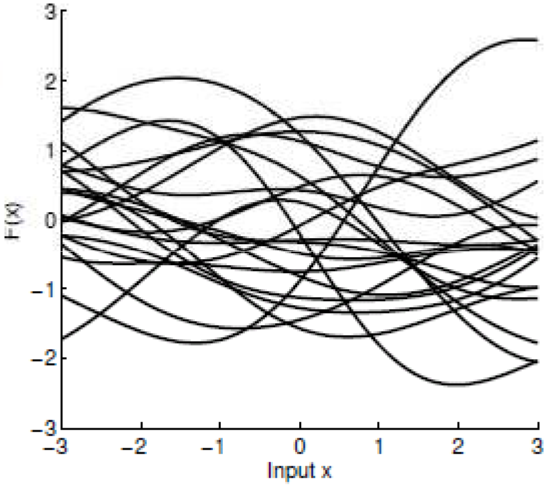
- Functional values $f(\mathbf{x}_1),\ldots,f(\mathbf{x}_n)$ for any finite set of n points form a n-variate Gaussian distribution.

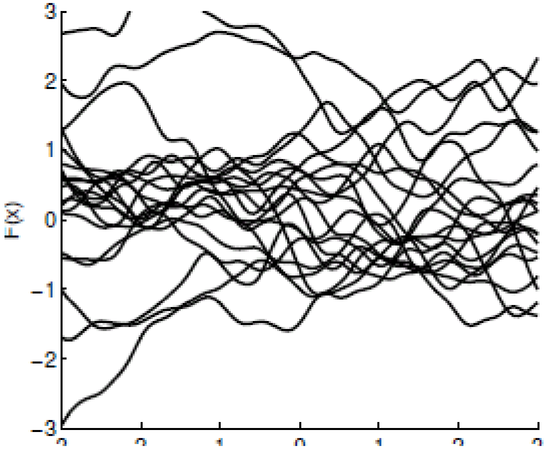
- Specified in terms of a covariance function (kernel function) k

$$k(x,x') = \mathrm{cov}\,[f(x),f(x')]$$

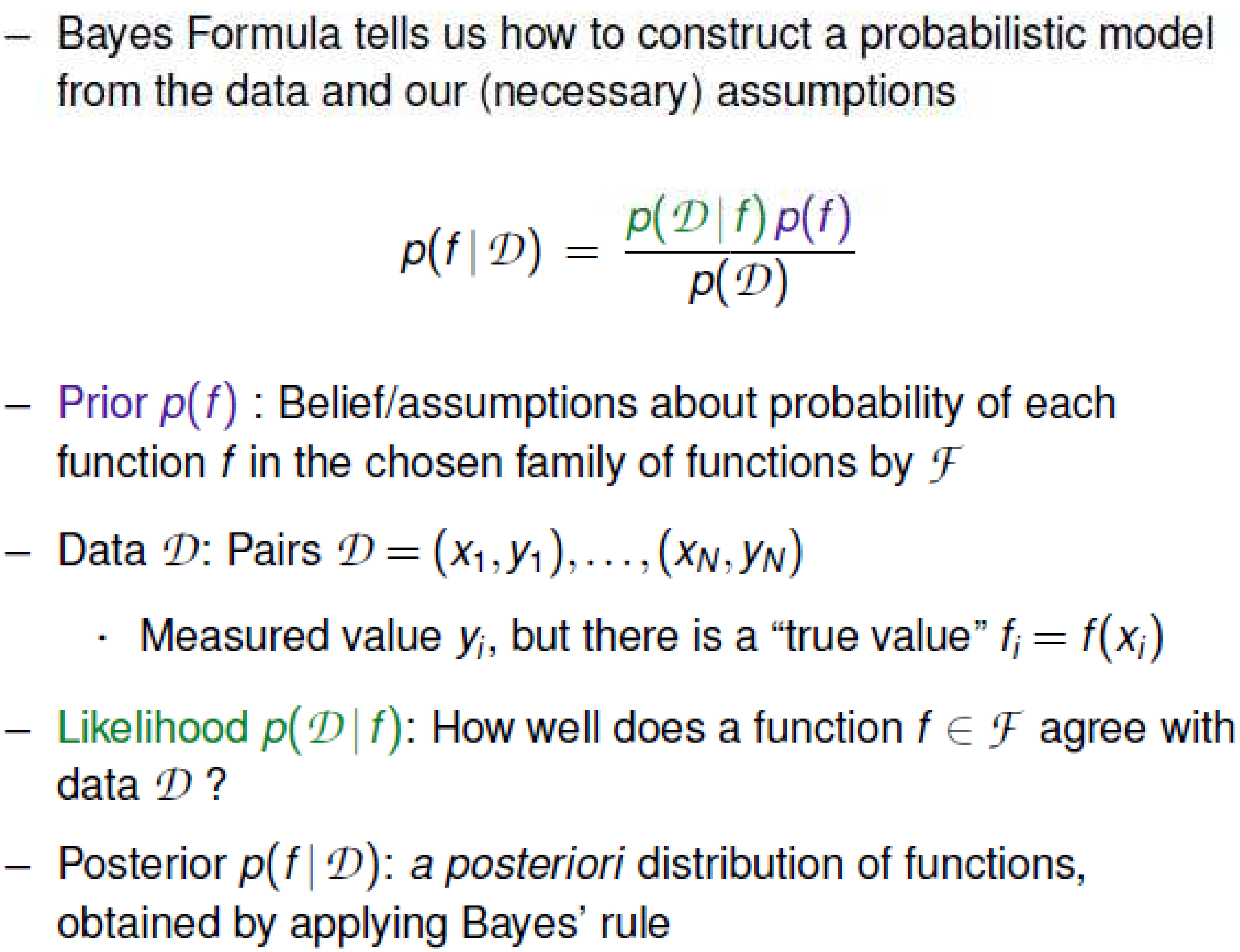
- Examples:

$$k(x,x') = \exp(-w(x-x')^2) \qquad \text{RBF}$$
$$k(x,x') = (1+w(x-x')^2)^{-\nu} \qquad \text{rational quadratic}$$





Bayes Theorem



- Bayes Formula tells us how to construct a probabilistic model from the data and our (necessary) assumptions

$$p(f \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid f)\, p(f)}{p(\mathcal{D})}$$

- Prior $p(f)$: Belief/assumptions about probability of each function f in the chosen family of functions by $\mathcal{F}$
- Data $\mathcal{D}$: Pairs $\mathcal{D} = (x_1, y_1), \ldots, (x_N, y_N)$
  - Measured value y_i, but there is a “true value” $f_i = f(x_i)$
- Likelihood $p(\mathcal{D} \mid f)$: How well does a function $f \in \mathcal{F}$ agree with data $\mathcal{D}$?
- Posterior $p(f \mid \mathcal{D})$: *a posteriori* distribution of functions, obtained by applying Bayes’ rule

GP Training

GP regression with training data
$\mathcal{D} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$

1. Assume a covariance function $k_\theta(x, x')$ with parameters θ. E.g. rational quadratic:

$$k(x, x') = \frac{1}{(1 + w\|x - x'\|^2)^{-\nu}} \qquad (1)$$

2. Marginal likelihood for given θ and σ^2

$$L_\theta = -\frac{1}{2}\log\det(K_\theta + \sigma^2 I) - \frac{1}{2}\mathbf{y}^\top(K_\theta + \sigma^2 I)^{-1}\mathbf{y} - \frac{N}{2}\log 2\pi$$

3. Use a numeric optimizer to maximize marginal likelihood, obtain final covariance function k_θ
4. Compute kernel matrix K, $K_{ij} = k_\theta(\mathbf{x}_i, \mathbf{x}_j)$
5. Solve linear system $(K + \sigma^2\mathbf{1})\alpha = \mathbf{y}$

Prediction with GPs

- Prediction is a *probability distribution* (Gaussian):

$$p(f(x^*) \mid \mathcal{D}) = \mathcal{N}(\bar{f}^*, \bar{s}^*)$$

- Predictive mean

$$\bar{f}^* = \sum_{i=1}^{N} \alpha_i k(x^*, x_i)$$
$$\alpha = (K + \sigma^2 I)^{-1} \mathbf{y}$$

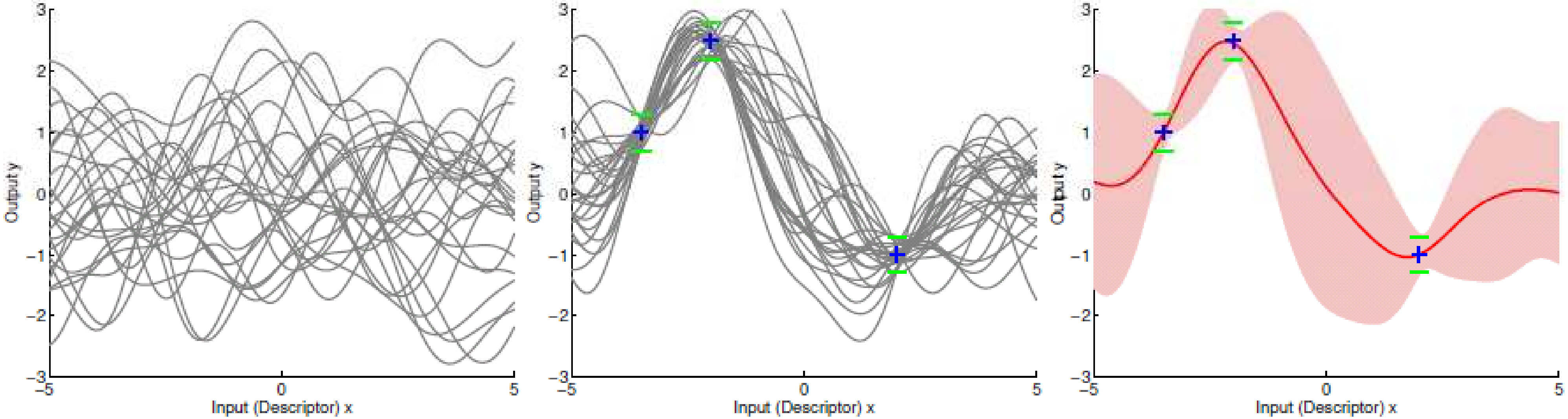
- Predictive standard deviation $\bar{s}^*$:

$$\bar{s}^* = \sqrt{k(x^*, x^*) - \mathbf{v}^\top (K + \sigma^2 I)^{-1} \mathbf{v}}$$

- Computationally not too demanding, fast

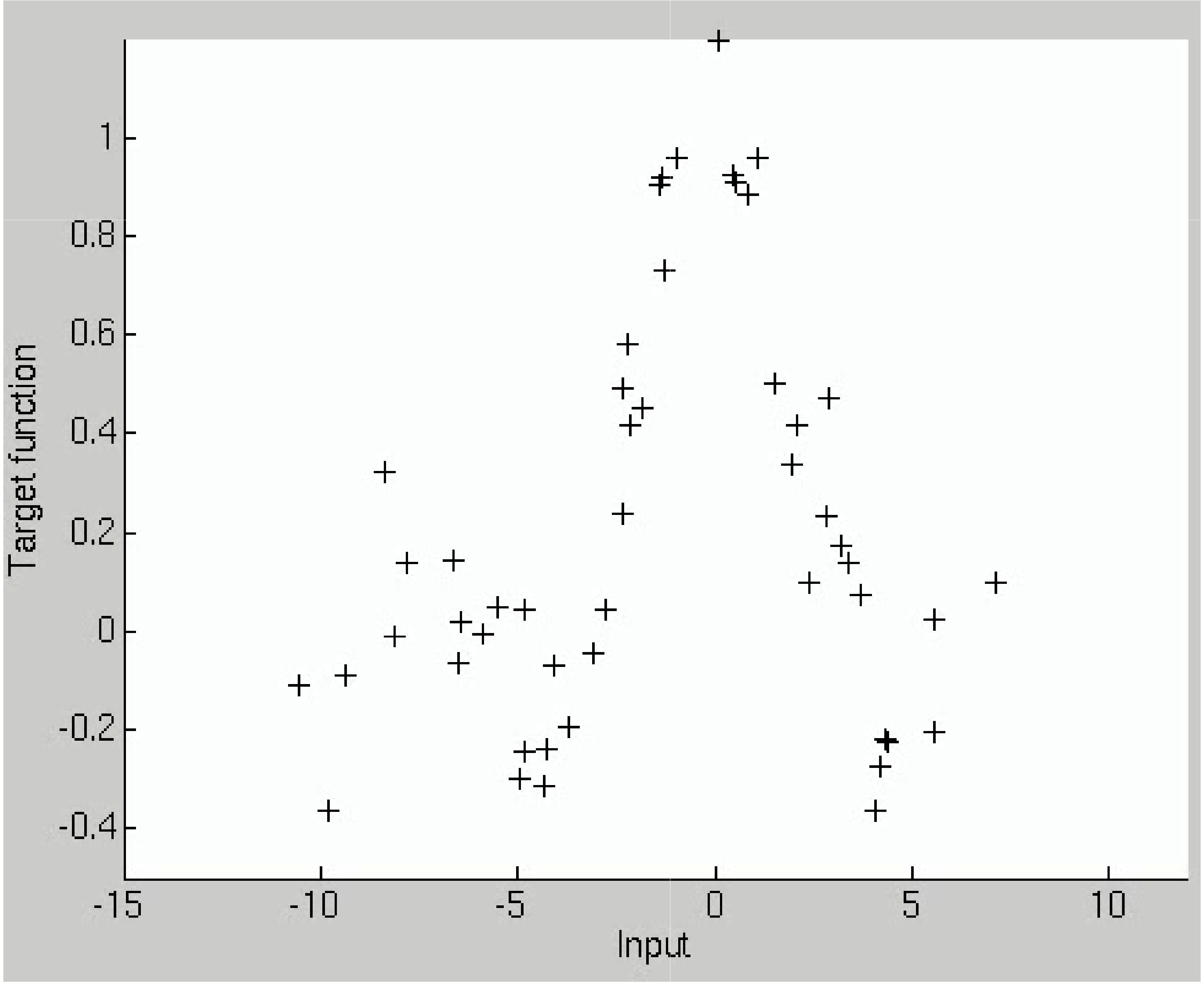
Notation: vector $\mathbf{y} = (y_1, \ldots, y_N)$, matrix K with $K_{i,j} = k(x_i, x_j)$, unit matrix I, vector $\mathbf{v}$ with $v_i = k(x^*, x_i)$

GP Learning – a cartoon



- Specify a huge number of possible functions
- Eliminate those that don’t agree with the data
- Average over what remains: Prediction is a probability distribution

GP the Movie



Application: predict chemical endpoints from descriptors

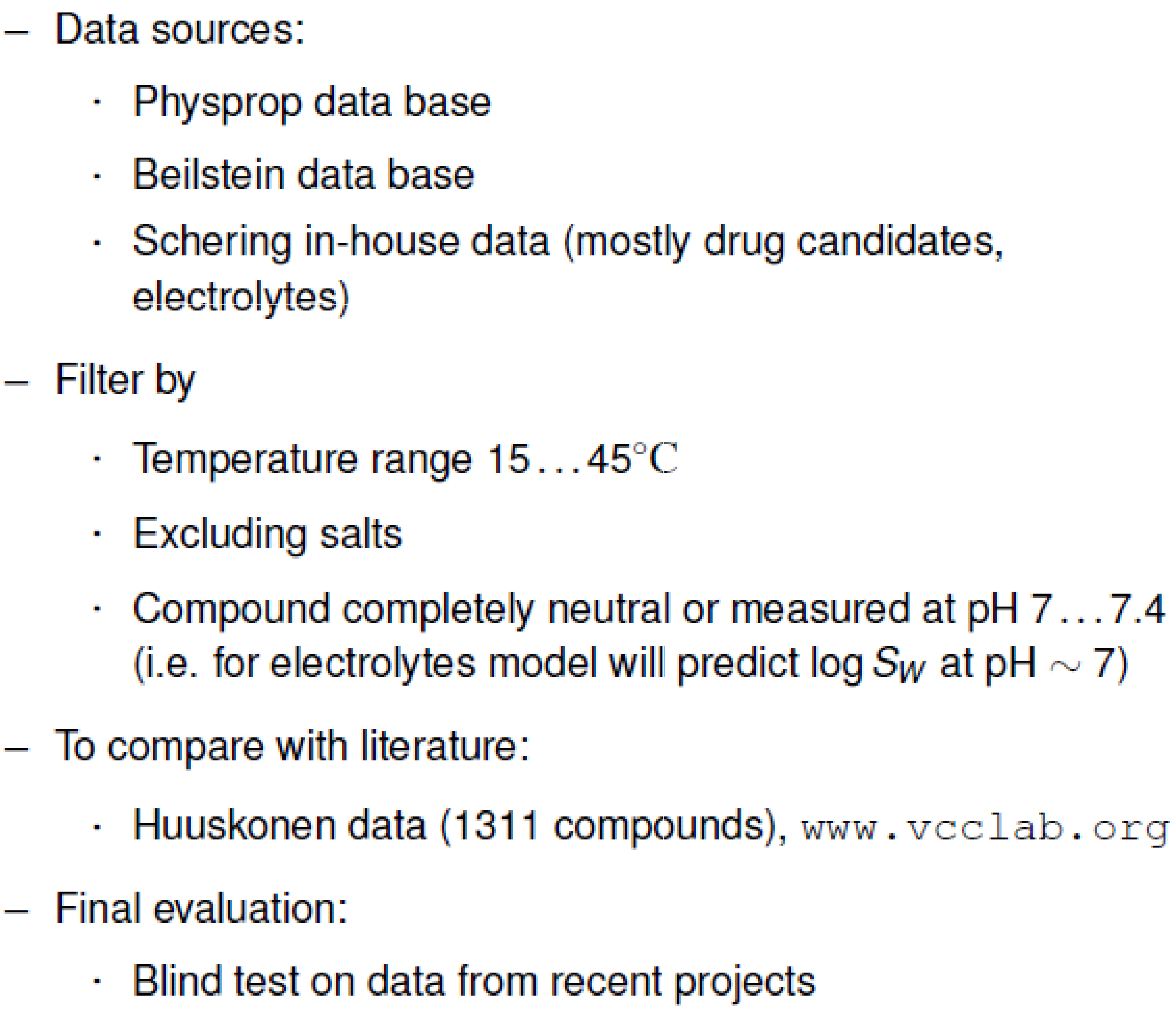
Develop customized tools to predict

- Water solubility, logP and logD
- Metabolic stability
- CYP P450 inhibition

that. . .

- are accurate on in-house data
- provide individual error bars for each prediction
- check the domain of applicability
- are easily retrainable
- are fast (library design)

Data available: solubilty (physico-chemical property)

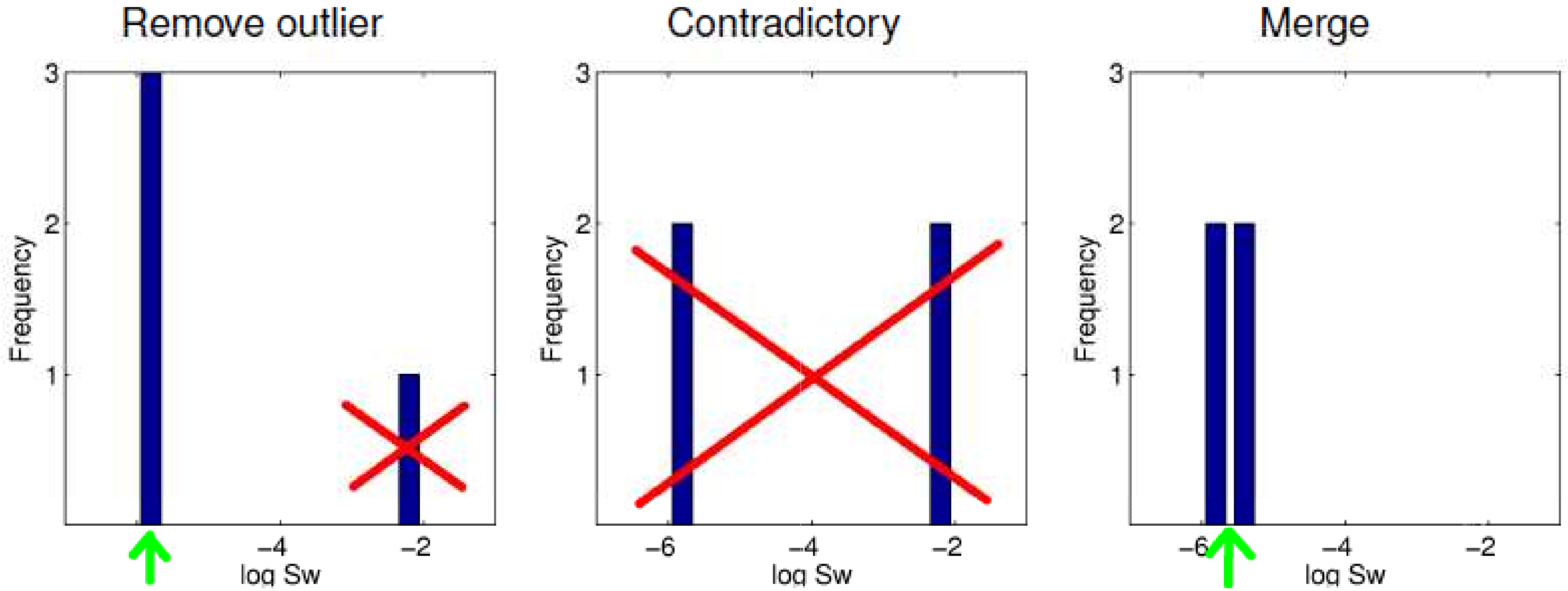


- Data sources:
  - Physprop data base
  - Beilstein data base
  - Schering in-house data (mostly drug candidates, electrolytes)
- Filter by
  - Temperature range 15 . . . 45°C
  - Excluding salts
  - Compound completely neutral or measured at pH 7 . . . 7.4 (i.e. for electrolytes model will predict log S_W at pH $\sim$ 7)
- To compare with literature:
  - Huuskonen data (1311 compounds), `www.vcclab.org`
- Final evaluation:
  - Blind test on data from recent projects

[Schwaighofer et al. JCIM 2007, Schroeter et al, ChemMedChem 2007]

Issues: Multiple Measurements

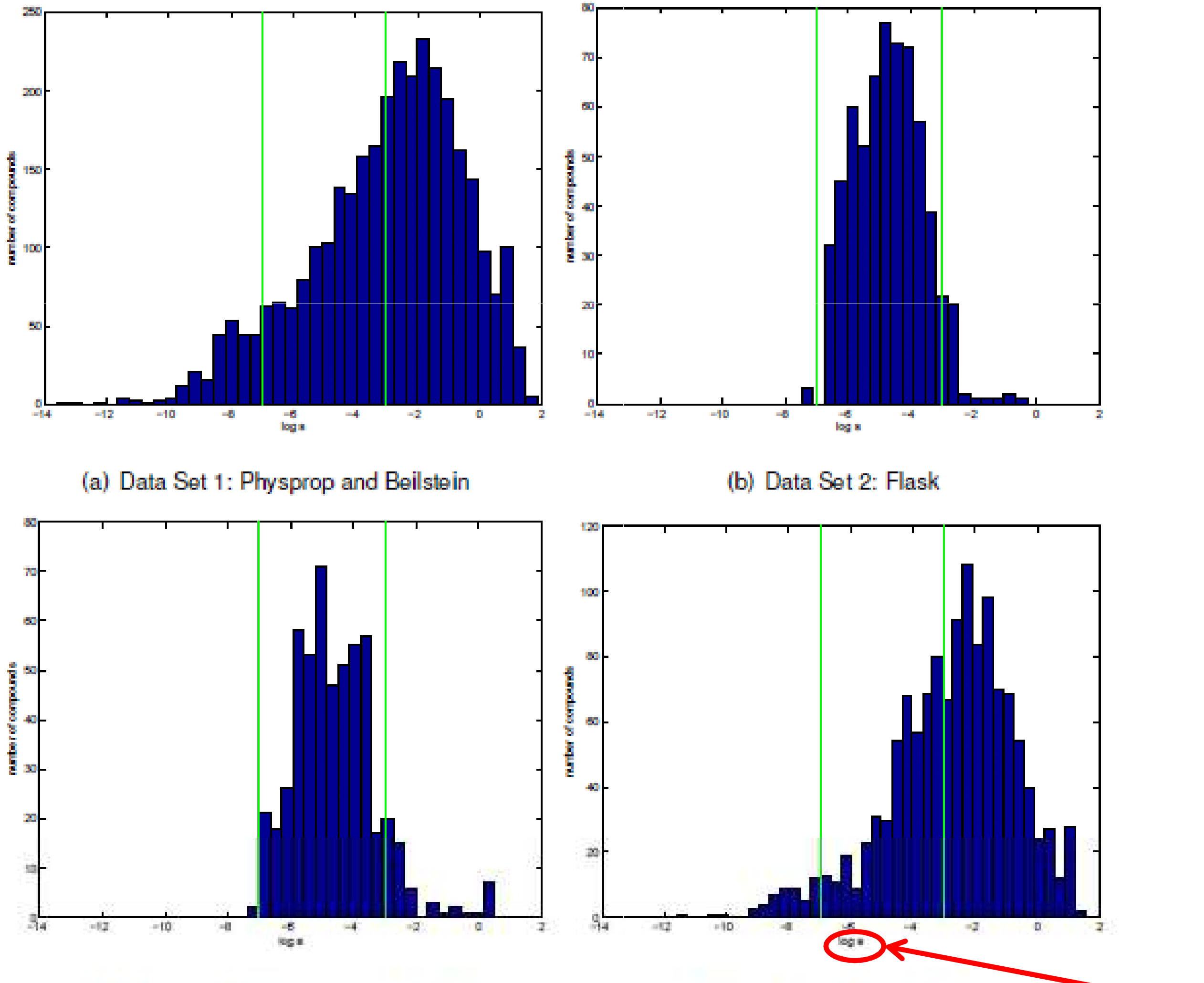
# Measurements M	$M = 1$	$M = 2$	$3 \leq M \leq 10$	$M > 10$
# Compounds	2857	858	320	23



GP models *learned* plausible noise levels

- $\sigma_1 = 0.46$ for compounds with single measurements
- $\sigma_2 = 0.15$, $\sigma_3 = 0.026$ for compounds with multiple measurements

Fitness for Purpose



(a) Data Set 1: Physprop and Beilstein

(b) Data Set 2: Flask

(c) Data Set 3: Flask external validation

(d) Data Set 4: Huuskonen

Log s

Descriptors

Full set of 1664 Dragon descriptors (Todeschini et al) includes, among others

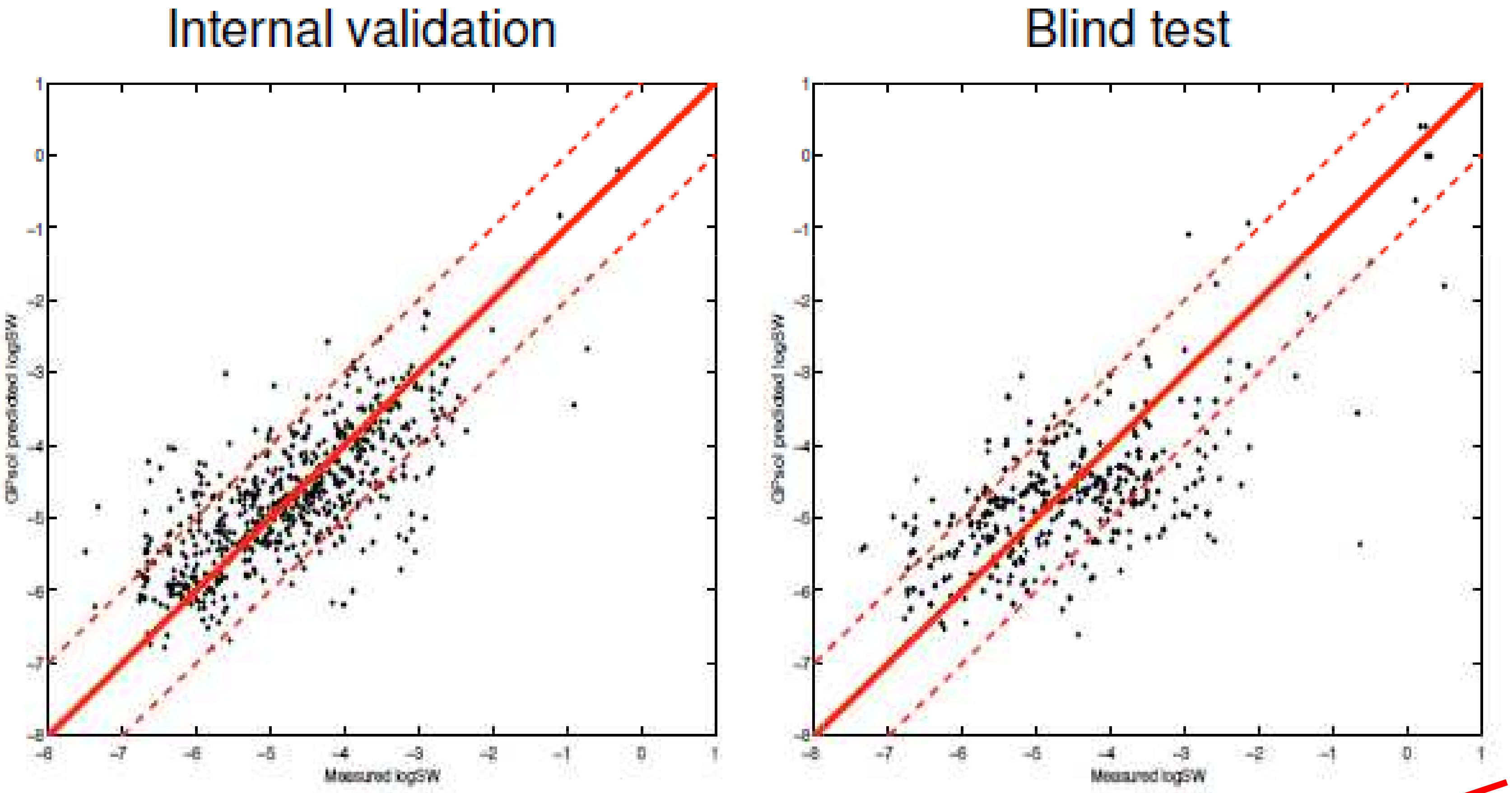
- constitutional & topological descriptors
- walk & path counts
- eigenvalue-based indices
- counts of functional groups & atom-centered fragments

Descriptors with highest weight include

- Number of hydroxy-, carboxylic acid and keto groups
- LogD at ph 7
- Total polar surface area
- Number of nitrogen & oxygen atoms

ML model can tell which descriptors are important

Results Solubility Schering in House (at pH 7)



Internal validation

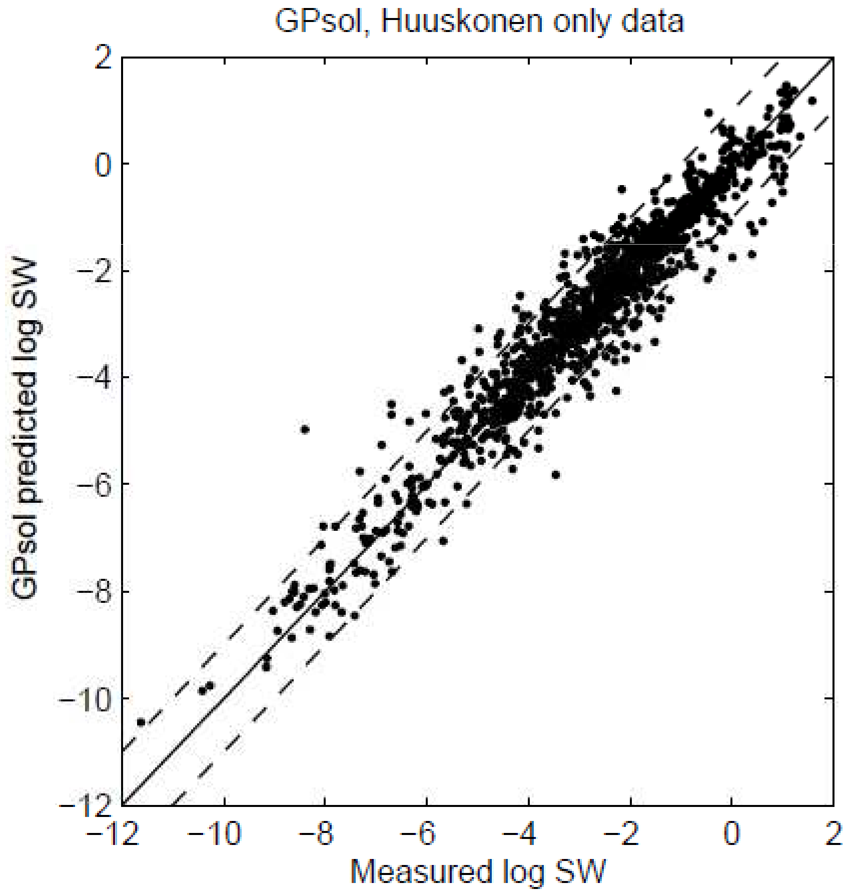
Blind test

5-fold CV

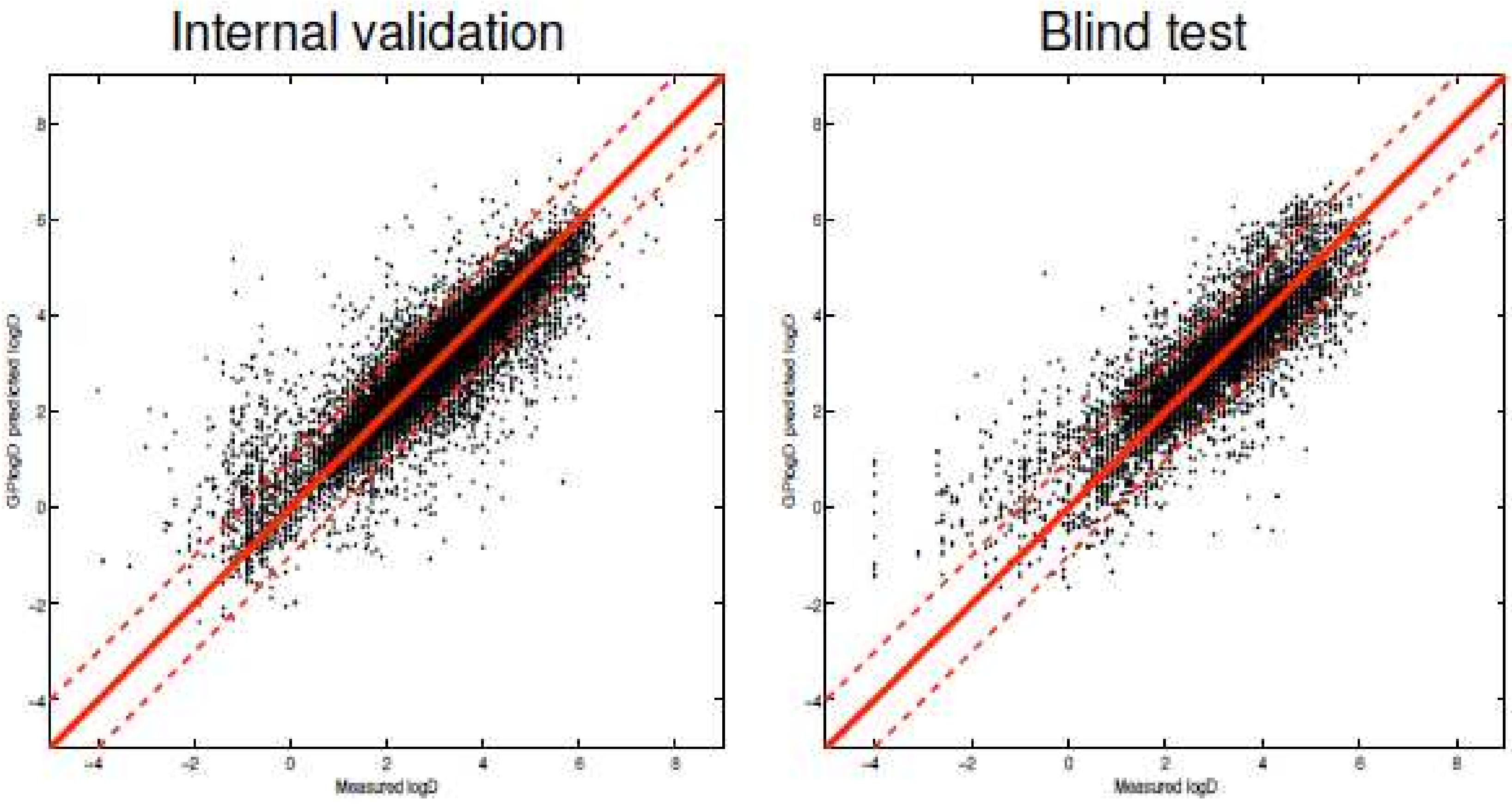
	MAE	r^2	% ±1
Internal validation (∼ 4,000 compounds)	0.57	0.56	84%
Blind test (∼ 500 compounds)	0.73	0.51	75%
Best commercial tool (out of 6) on blind test data	0.99	0.43	58%

Results Solubility Huuskonen

		r^2	rmse
Huuskonen	2000	0.88	0.71
Tetko	2001	0.85	0.81
		0.90	0.66
Liu	2001		0.87
Ran	2001		0.76
Bruneau	2001		0.82
Engkvist	2002	0.95	
Yan	2003	0.82	
		0.92	
Yan	2003	0.89	
		0.94	
Lind	2003	0.89	0.68
Yan	2004	0.94	
Hou	2004	0.90	
Fröhlich	2004	0.90	
Clark	2005	0.84	
Rapp	2005	0.92	
		0.91	
This study	**2006**	**0.93**	**0.57**



Results LogD (at pH 7)


Internal validation

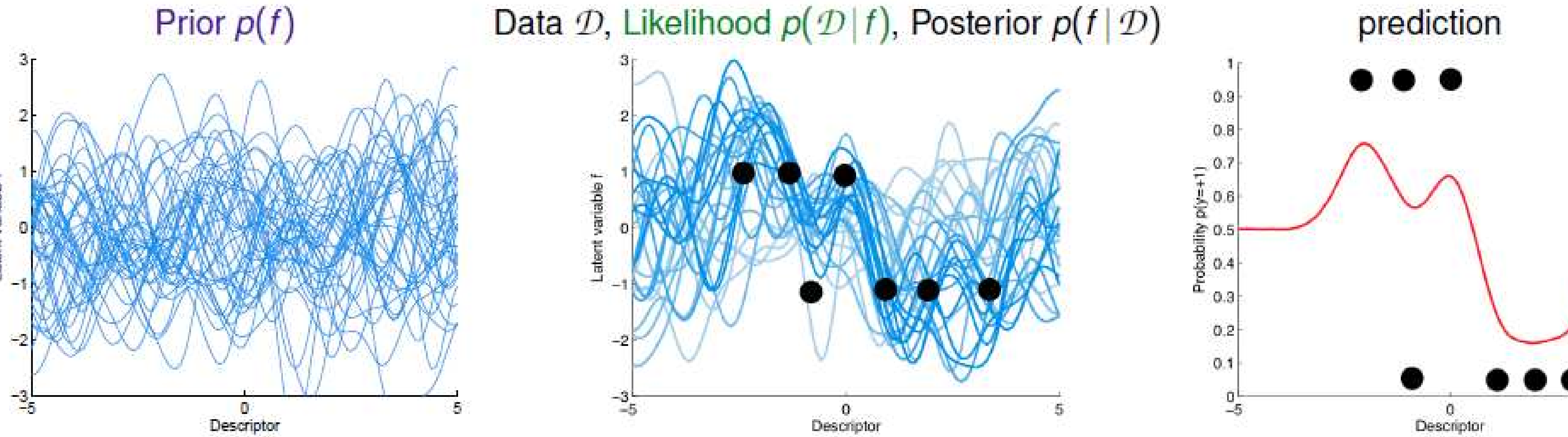
Blind test

	MAE	r^2	% ±1
Internal validation (∼ 22,000 compounds)	0.45	0.79	89%
Blind test (∼ 7,000 compounds)	0.60	0.71	81%
Best commercial tool (out of 3) on blind test data	1.40	0.27	44%

Blind Test

[Schroeter et al 2008]

GPs for Classification

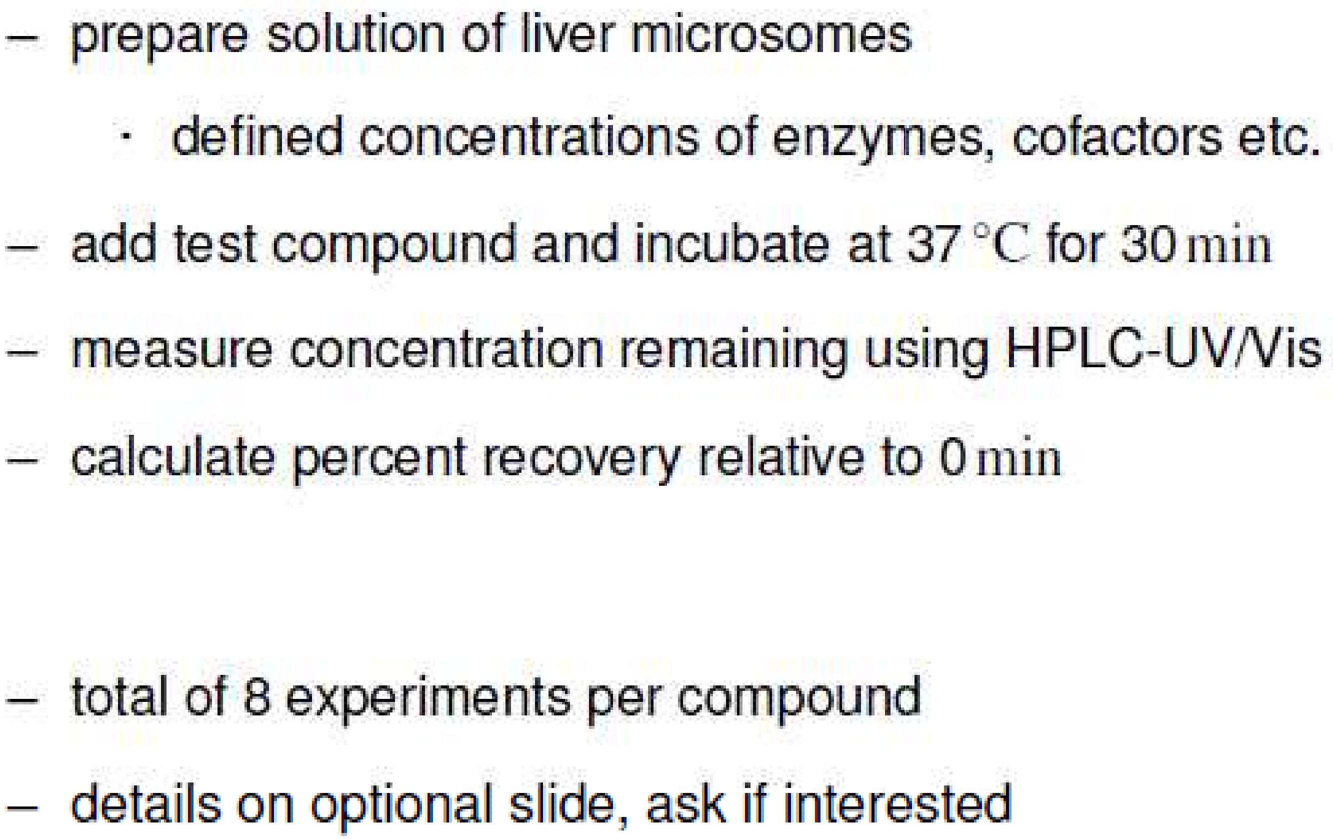


Prior $p(f)$

Data $\mathcal{D}$, Likelihood $p(\mathcal{D}\,|\,f)$, Posterior $p(f\,|\,\mathcal{D})$

prediction

Measuring Metabolic Stability (bio-chemical property)

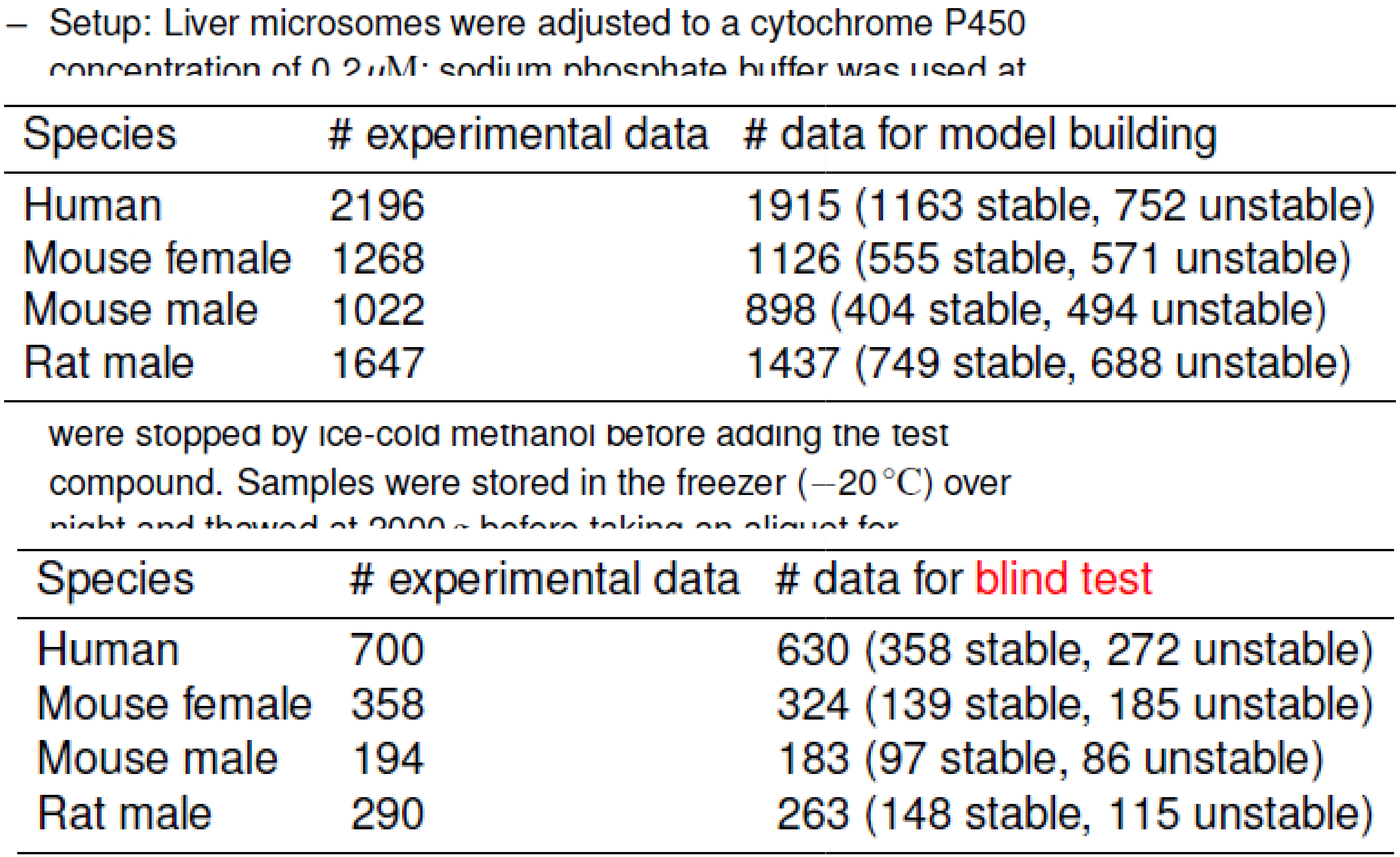


- prepare solution of liver microsomes
  - defined concentrations of enzymes, cofactors etc.
- add test compound and incubate at 37 °C for 30 min
- measure concentration remaining using HPLC-UV/Vis
- calculate percent recovery relative to 0 min

- total of 8 experiments per compound
- details on optional slide, ask if interested

[Schwaighofer et al, J Comp Mol Des 2008]

Measuring Metabolic Stability: Detailed set-up



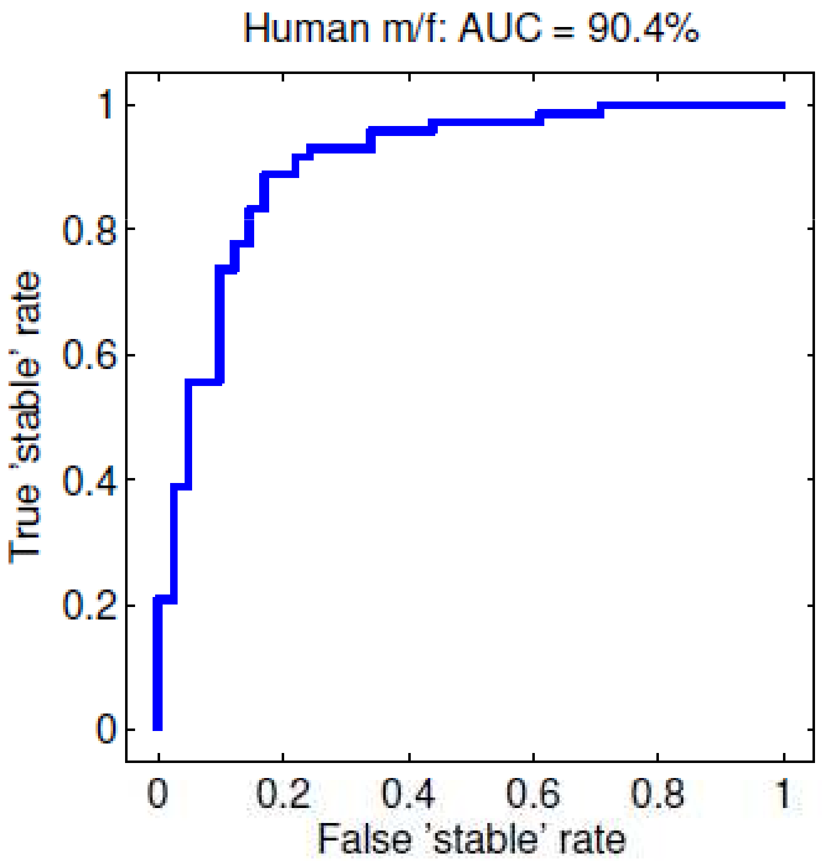
– Setup: Liver microsomes were adjusted to a cytochrome P450 concentration of 0.2 μM; sodium phosphate buffer was used at

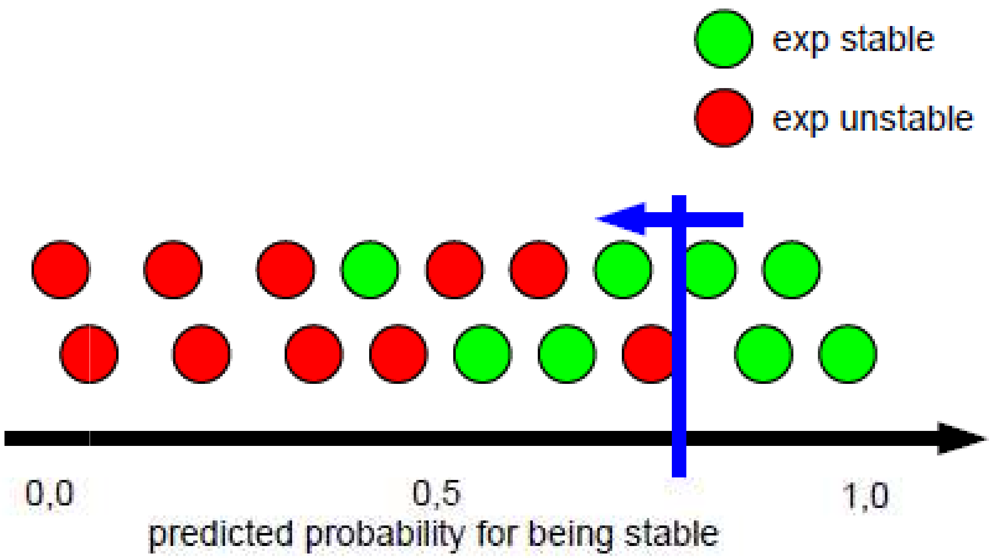
Species	# experimental data	# data for model building
Human	2196	1915 (1163 stable, 752 unstable)
Mouse female	1268	1126 (555 stable, 571 unstable)
Mouse male	1022	898 (404 stable, 494 unstable)
Rat male	1647	1437 (749 stable, 688 unstable)

were stopped by ice-cold methanol before adding the test compound. Samples were stored in the freezer (–20 °C) over night and thawed at 2000 g before taking an aliquot for

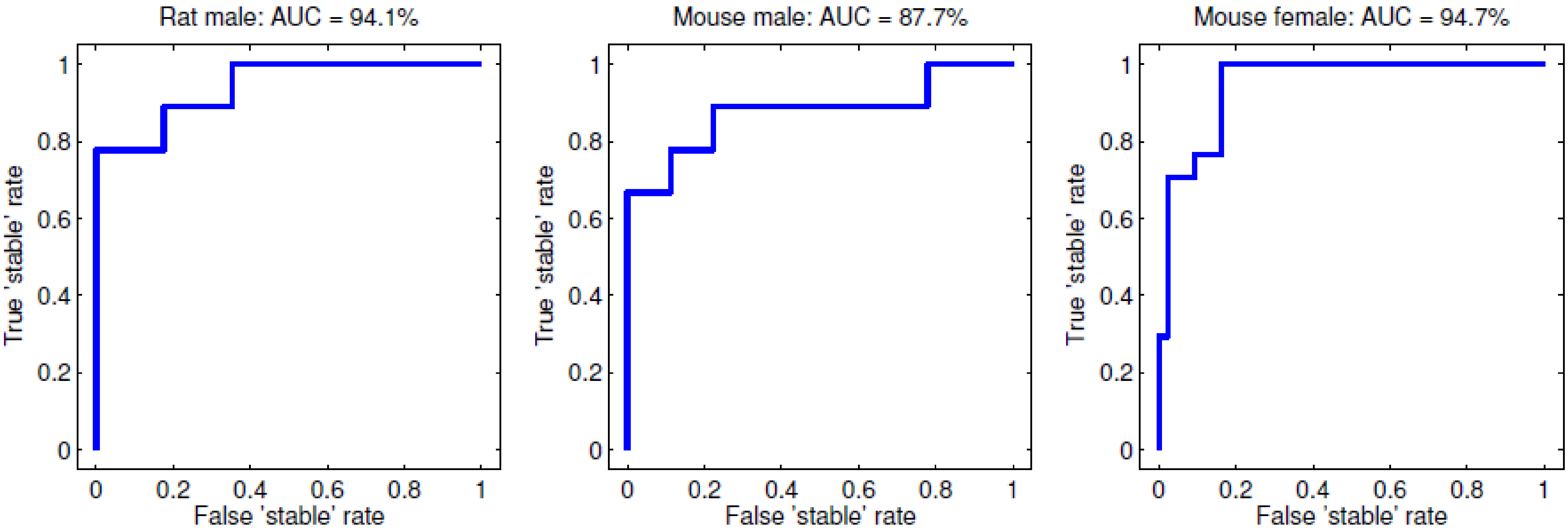
Species	# experimental data	# data for blind test
Human	700	630 (358 stable, 272 unstable)
Mouse female	358	324 (139 stable, 185 unstable)
Mouse male	194	183 (97 stable, 86 unstable)
Rat male	290	263 (148 stable, 115 unstable)

Quantifying Performance



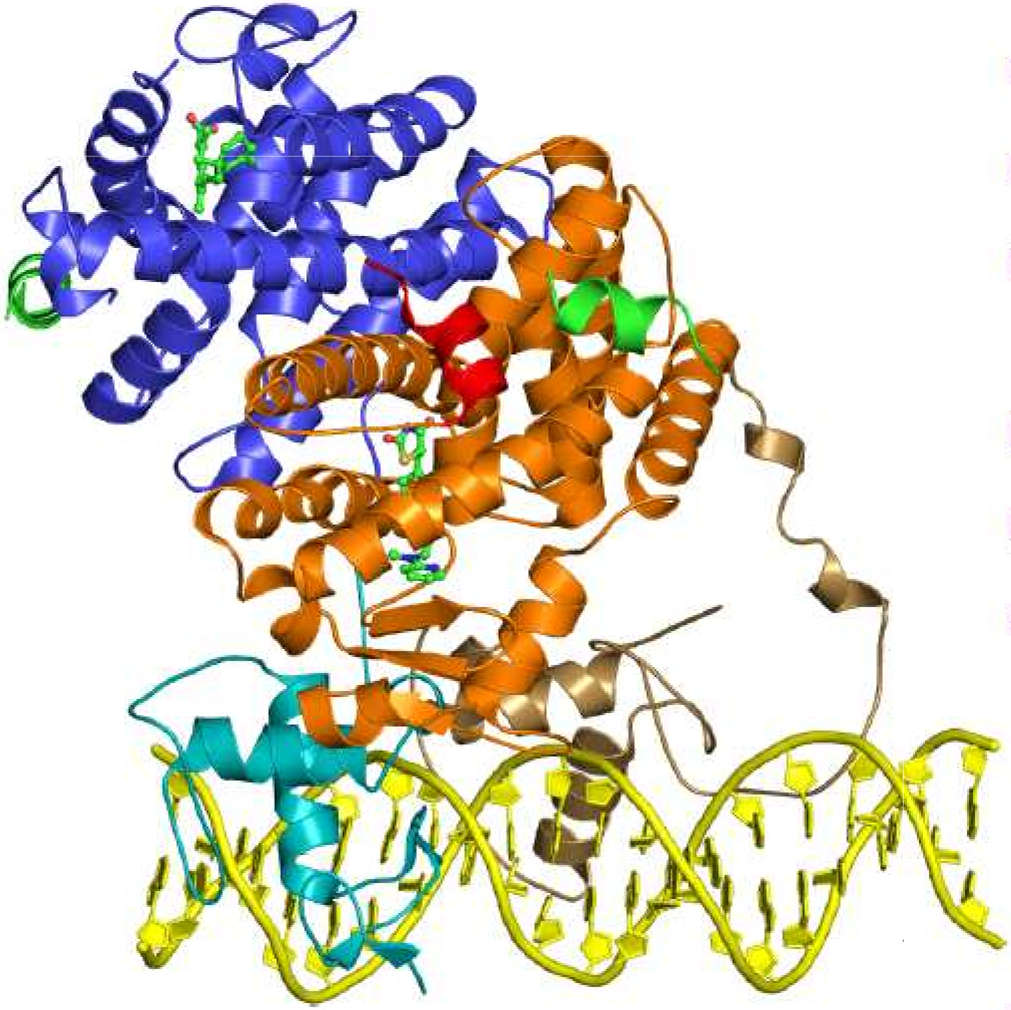


Model Performance



Predicting biological properties

PPARγ = <u>P</u>eroxisome <u>P</u>roliferator-<u>A</u>ctivated <u>R</u>eceptor γ



- Nuclear receptor
- 3 isoforms: α, β/δ, γ
- Related to type 2 diabetes and dyslipidemia
- Heterodimerization with RXR
- Large binding pocket (1.5 nm^3)
- Native ligands: fatty acids, lipid metabolites
- Objective: find new agonists

Chandra et al: *Nature* 456 (7220): 350, 2008.

Virtual Screening: Optimization Criteria

“Target is binding affinity” — oversimplification

- False negatives and false positives may have different costs
  → need to reduce false positives (in our case)

PPARγ study:

- Learn binding affinity (pK_i) instead of receptor activation (EC_{50})
- Ignore other criteria during learning
- Do “cherry-picking” at the end
- Use fraction of inactives in top 20 as performance measure
- Use Gaussian process variance estimates

The Study

PPARγ study:

- Published data set ($n = 144$)
- Used leave-k-clusters-out cross-validation

PPARγ study:

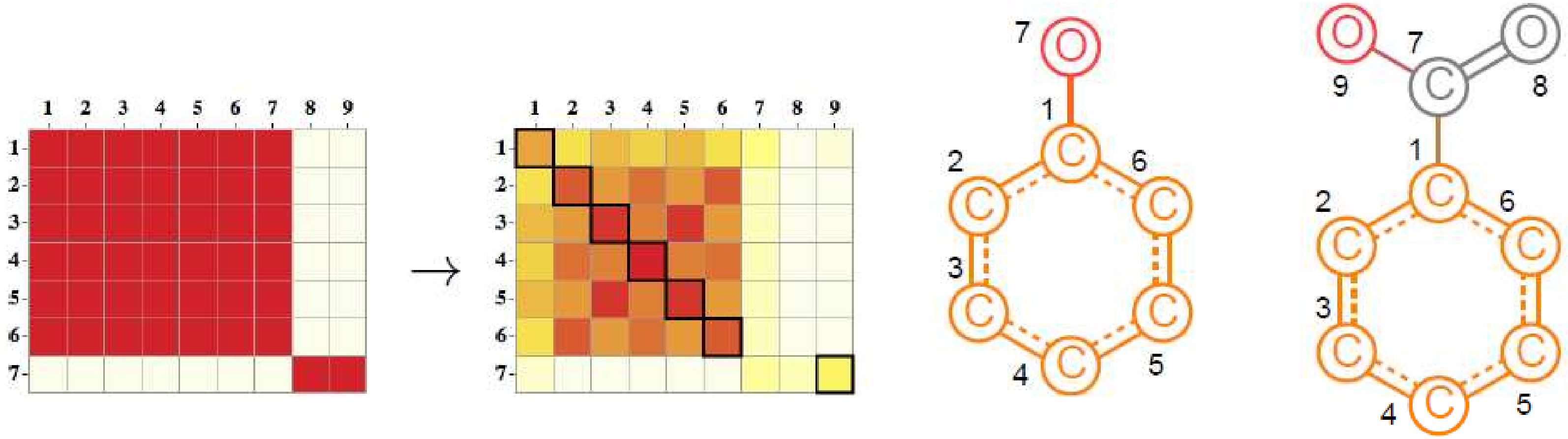
- CATS2D ($d = 210$), MOE 2D ($d = 184$) descriptors
- ISOAK graph kernel
- Multiple kernel learning

Choosing the Kernel

ISOAK = iterative similarity optimal assignment kernel

$$\mathbf{X}_{v,v'} = (1-\alpha)k_v(v,v') + \alpha \max_{\pi} \frac{1}{|v'|} \sum_{\{v,u\}\in E} \mathbf{X}_{u,\pi(u)} k_e(\{v,u\},\{v',\pi(u)\})$$

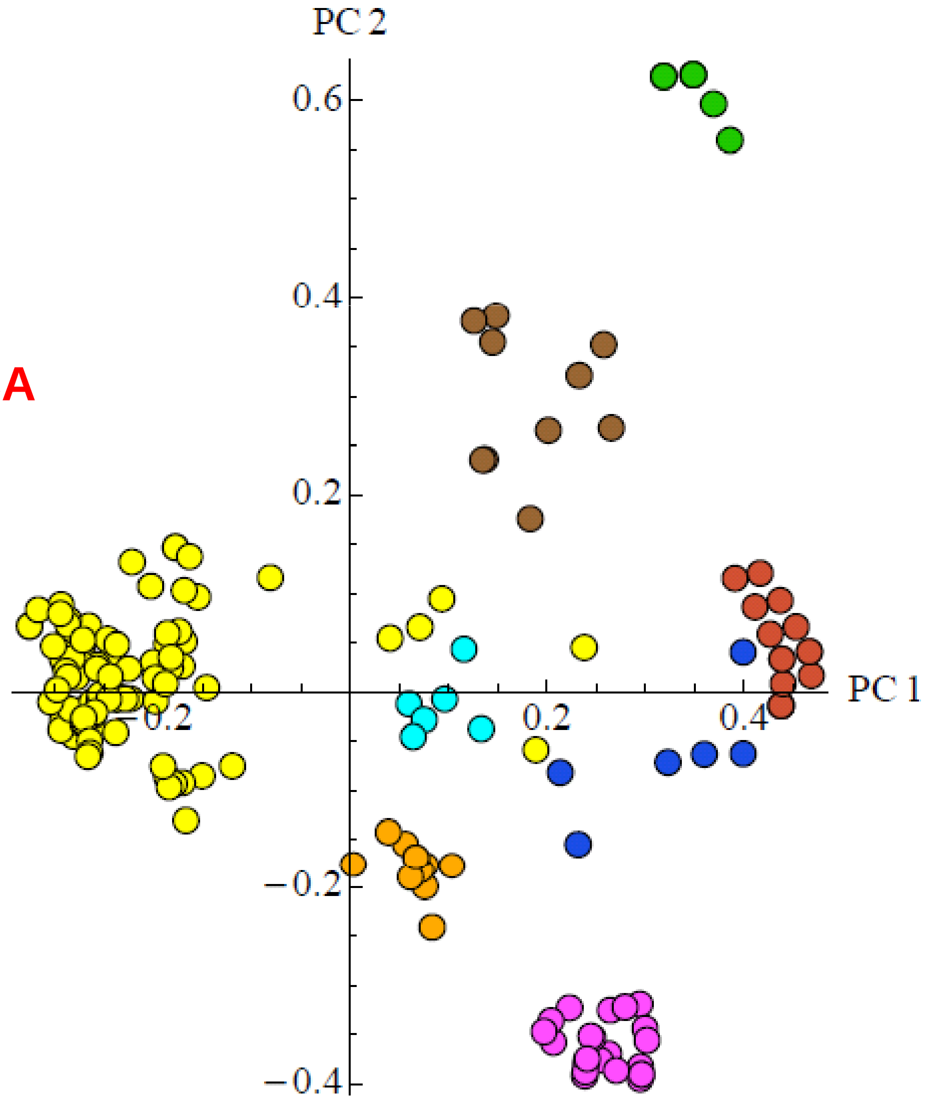
α controls recursiveness; π assigns neighbors of v to neighbors of v'



Rupp et al, J. Chem. Inf. Mol. Model. 47(6): 2280, 2007.

PPAR-Gamma Data Set

kPCA



Kernel principle component analysis with ISOAK graph kernel ($n = 176$)

● tyrosines, ● TZDs, ● indoles, ● oxadiazoles, ● fatty acids, ● tertiary amides, ● tyrosines N, ● TZD-fatty acid hybrids

Rücker et al.: *Bioorg. Med. Chem.* 14(15): 5178, 2006; Rupp, PhD thesis, 2009.

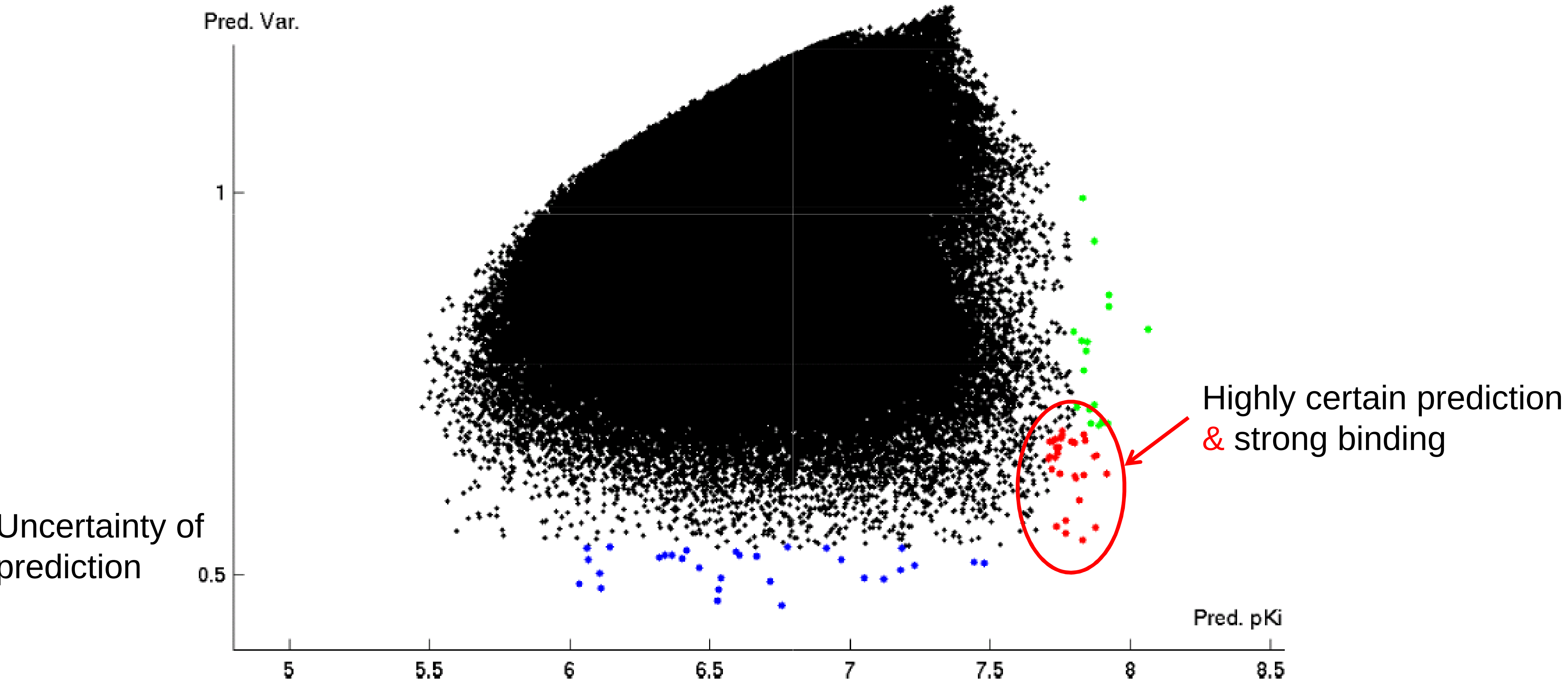
Results

- Top 30 of three best performing models
- 16 cherry-picked compounds with novel scaffolds

- PPARγ selective activator (EC$_{50}$ $9.3 \pm 0.3\,\mu$M), natural product related
- 3 dual PPARα/γ activators (μM range, two $\leq 10\mu$M)
- 4 selective PPARα activators (μM range, one $\leq 10\mu$M)

- 8 out of 16 compounds are active
- 4 out of 16 compounds with EC$_{50}$ $\leq 10\mu M$

Virtual Screening: cherry picking



Highly certain prediction & strong binding

Uncertainty of prediction

Higher *pKi* values indicate stronger binding

Detailed Results

- PPARγ affinity is a non-linear function of structure
- Compound weighting by activity did not improve predictions
- Separate kernels in MKL worsened MAE but improved FI_{20}

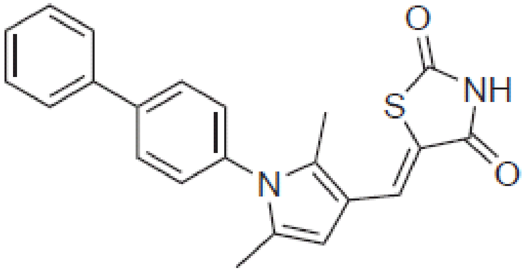
Fraction of inactives in top

	Cross-validation		y-scrambling	
Model	MAE	FI_{20}	MAE	FI_{20}
KRR/MOE 2D/linear	1.45 ± 0.04	0.78 ± 0.05	1.45 ± 0.04	0.78 ± 0.05
SVM/MOE 2D/RBF	0.69 ± 0.08	0.29 ± 0.14	1.10 ± 0.10	0.68 ± 0.24
GP/CATS2D/RBF+RQ	$\mathbf{0.66 \pm 0.09}$	0.27 ± 0.14	1.08 ± 0.02	0.57 ± 0.17
GP/all+ISOAK/MKL	0.70 ± 0.11	$\mathbf{0.21 \pm 0.09}$	1.11 ± 0.06	0.65 ± 0.12

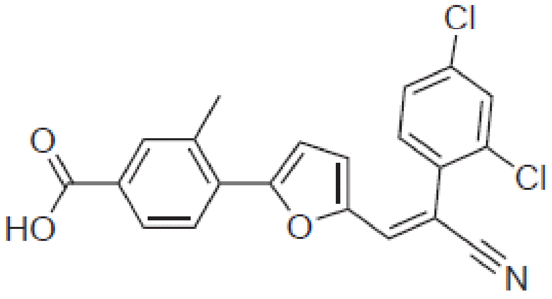
- 5 (best MAE model) + 10 (best FI_{20} model) = 15 compounds selected for assay tests

Results: prospective validation

- Cell-based reporter gene (luciferase) assay
- 8 out of 15 active, 4 in lower micro-molar range

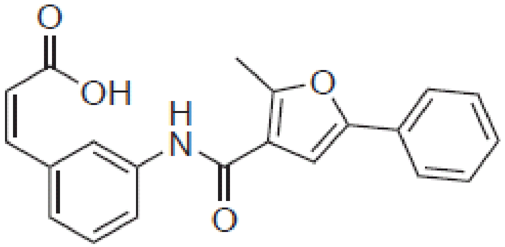


hPPARα $\mathrm{EC}_{50} = 1.25 \pm 0.37\,\mu\mathrm{M}$

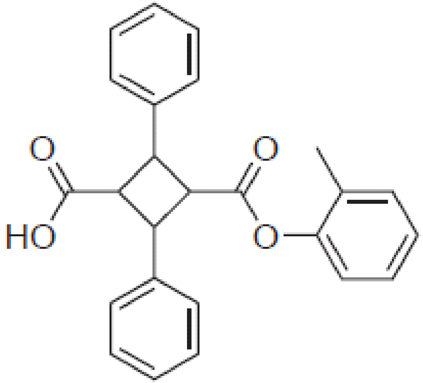


hPPARα $\mathrm{EC}_{50} = 12.98 \pm 4.21\,\mu\mathrm{M}$

hPPARγ $\mathrm{EC}_{50} = 3.75 \pm 0.2\,\mu\mathrm{M}$



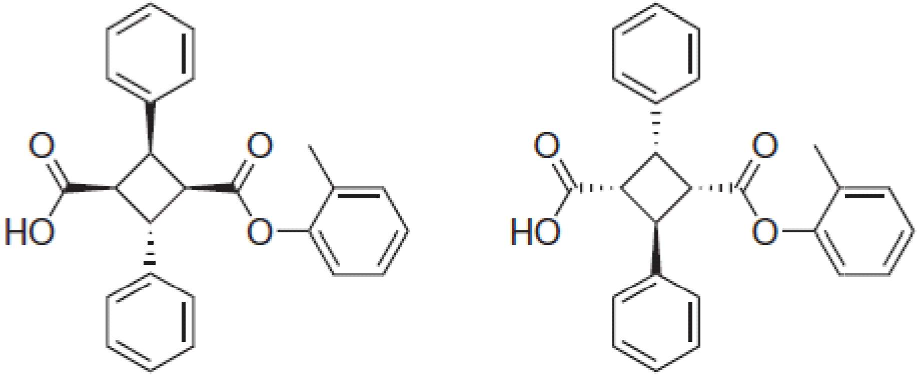
hPPARα $\mathrm{EC}_{50} = 13.48 \pm 8.53\,\mu\mathrm{M}$



hPPARγ $\mathrm{EC}_{50} = 10.03 \pm 0.2\,\mu\mathrm{M}$

Rau et al., Planta Med. 72(10): 881, 2006.

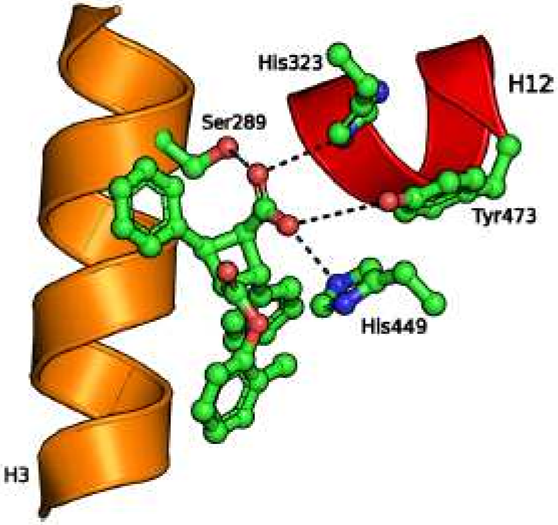
Best Hit: a natural product



stereochemistry

Cynodon dactylon

- Natural product
- Occurs in plant cell walls
- Photo-dimerization of trans-cinnamic acid

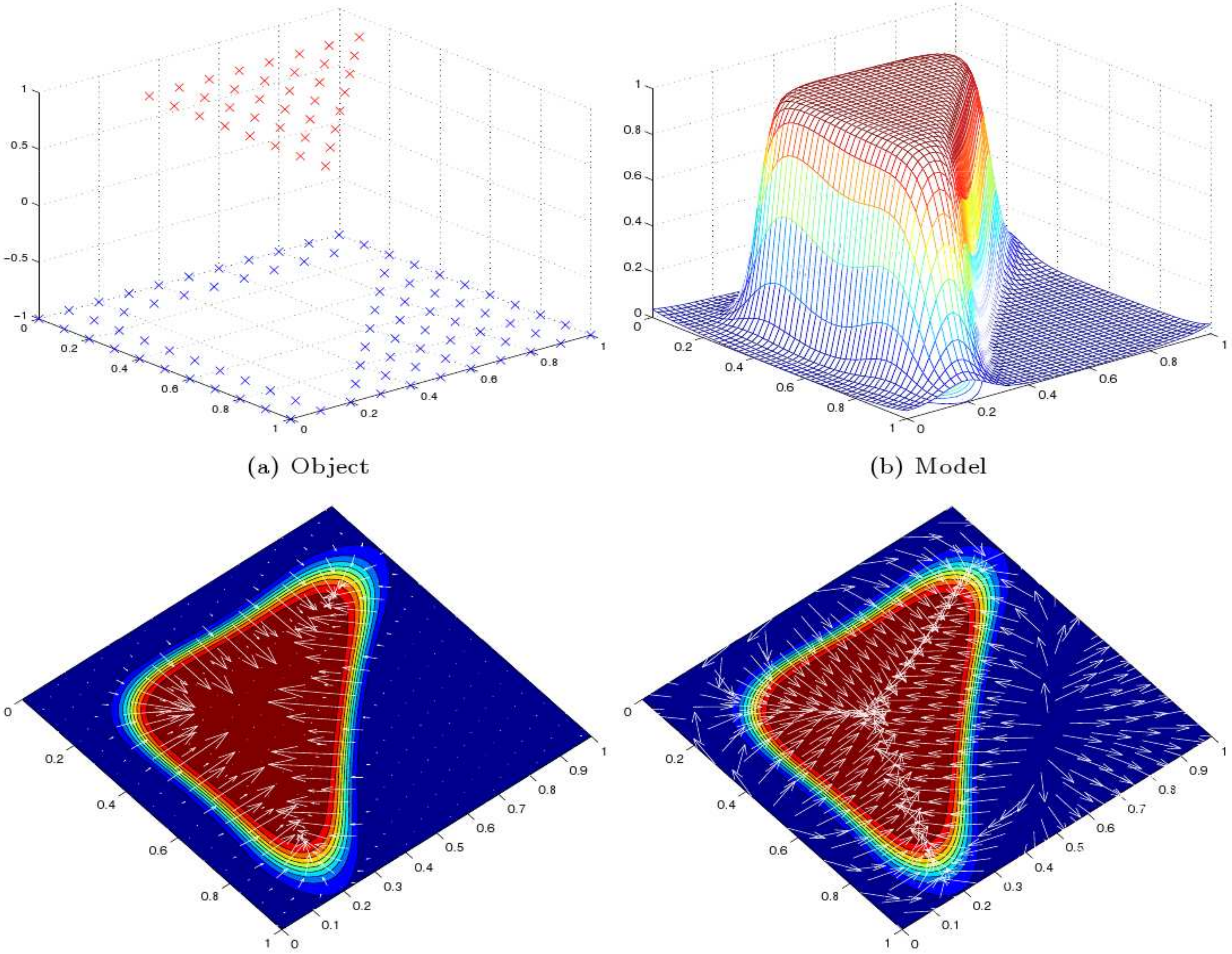


putative binding mode

[Rupp et al., ChemMedChem 2009, Steri et al., Bioorg Med Chem Lett 2010]

Misc Remarks

Explaining single Predictions



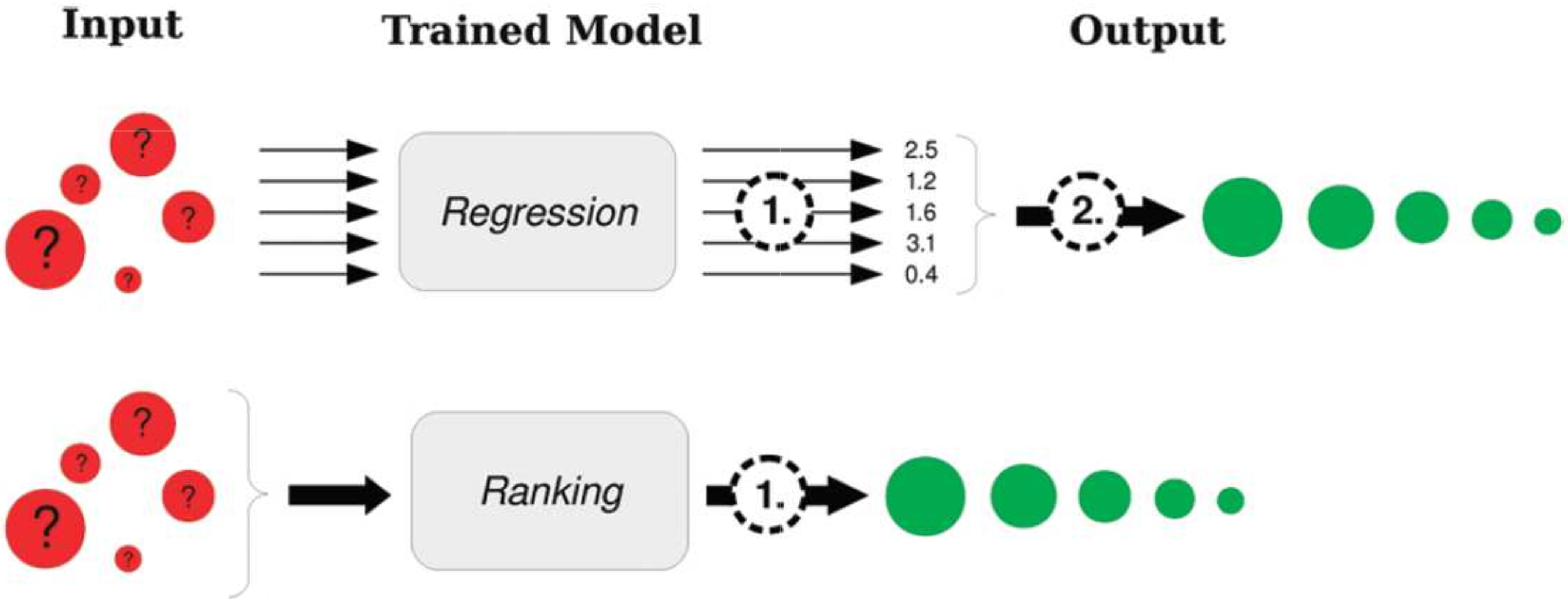
(a) Object

(b) Model

(c) Local explanation vectors

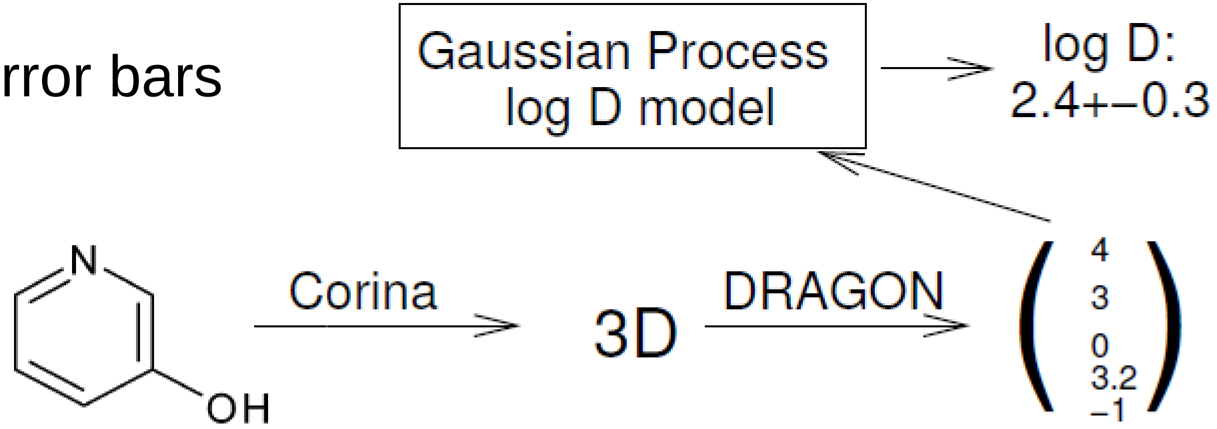
(d) Direction of explanation vectors

Ranking or Regression



Conclusion

- GPs and SVM have been applied in many practical applications
- CYP, hERG, metabolic stability, toxicity, log p, log d, solubility, mutagenicity
- ranking, explaining, error bars
- Kernel holds the key



Gaussian Process log D model → log D: 2.4+–0.3

Corina → 3D; DRAGON → $\begin{pmatrix} 4 \\ 3 \\ 0 \\ 3.2 \\ -1 \end{pmatrix}$

OH

- Machine Learning Methods are universal tools and useful