optibrium

Gaussian Processes: We demand rigorously defined areas of uncertainty and doubt

ACS Spring National Meeting. COMP, March 16th 2016

Matthew Segall, Peter Hunt, Ed Champness

matt.segall@optibrium.com

“We demand rigidly defined areas of doubt and uncertainty”

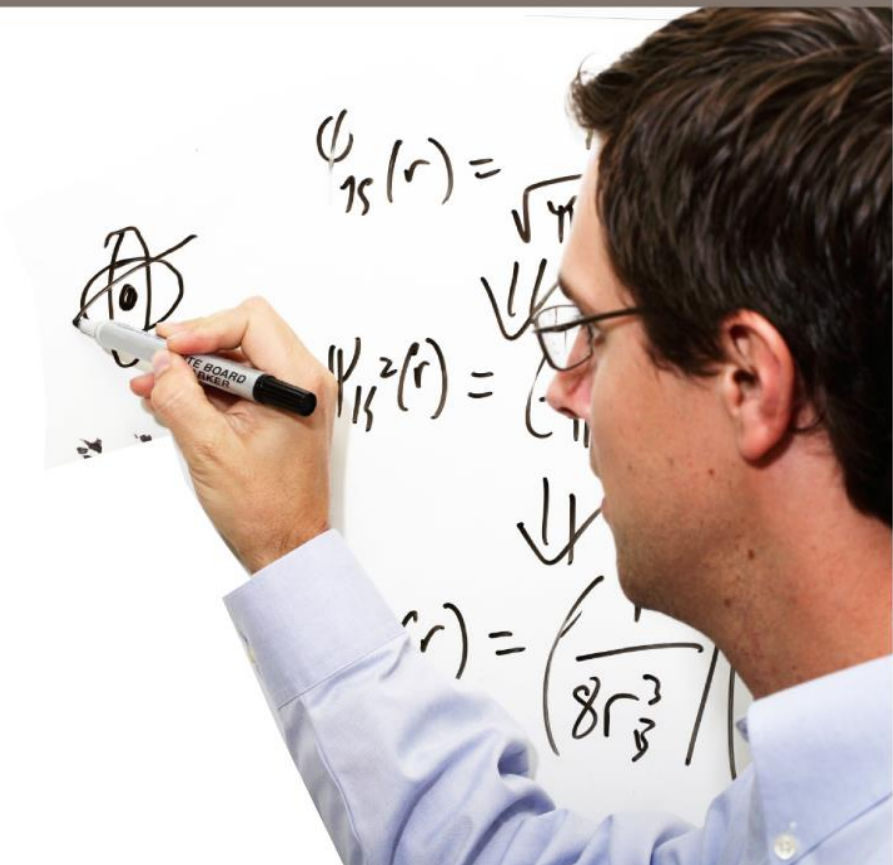
Douglas Adam’s, The Hitchhikers’ Guide to the Galaxy

Overview

- Uncertainties in model predictions
  - Domain of applicability
- Introduction to Gaussian Processes (GPs)
- Simple illustration of GPs
  - Uncertainty related to ‘domain of applicability’
  - Dealing with data variability as a source of uncertainty
  - Handling ‘missing’ or sparse data
- Automatic relevance determination
  - Identifying most important descriptors
- Practical example of GPs applied to QSAR
- Conclusions

Uncertainty in Model Predictions

optibrium



Quantitative Structure-Activity Relationships

$$y = f(x_1, x_2, x_3, \ldots) \pm \varepsilon$$

Statistical uncertainty

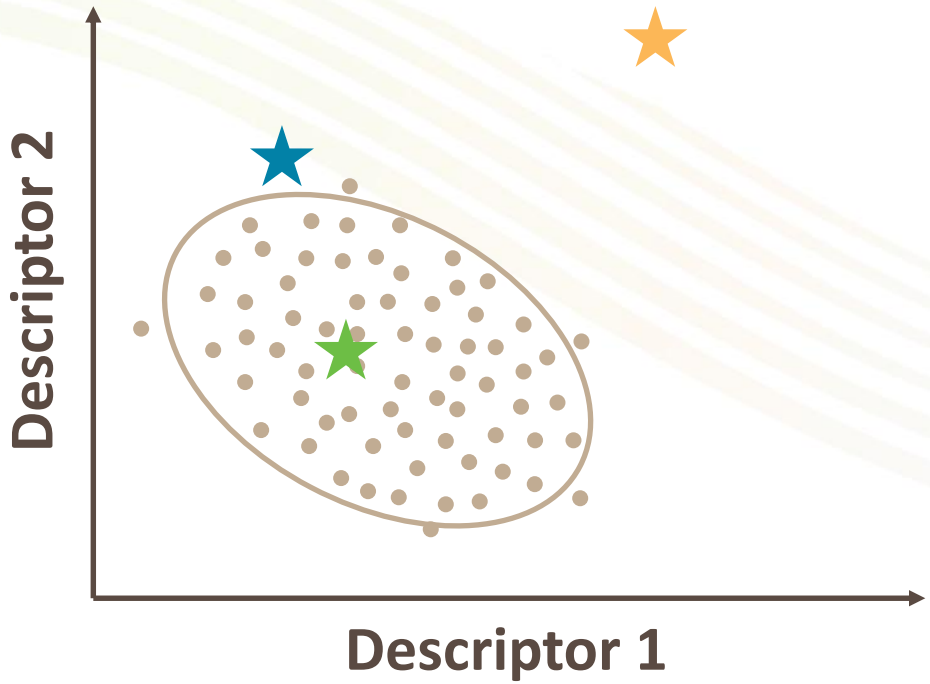
- Data
  - Quality data is essential
  - Public data needs very careful curation* (and may not be good enough)
- Descriptors, e.g.
  - Whole molecule properties, e.g. logP, MW, PSA…
  - Structural descriptors, SMARTS, fingerprints…
- Statistical fitting or machine learning method, e.g.
  - Multiple linear regression, partial least squares
  - Artificial neural networks, support vector machines, random forest, **Gaussian processes**…

*Waldman et al. JCAMD (2015) DOI: 10.1007/s10822-015-9865-0

Sources of Uncertainty in Model Predictions

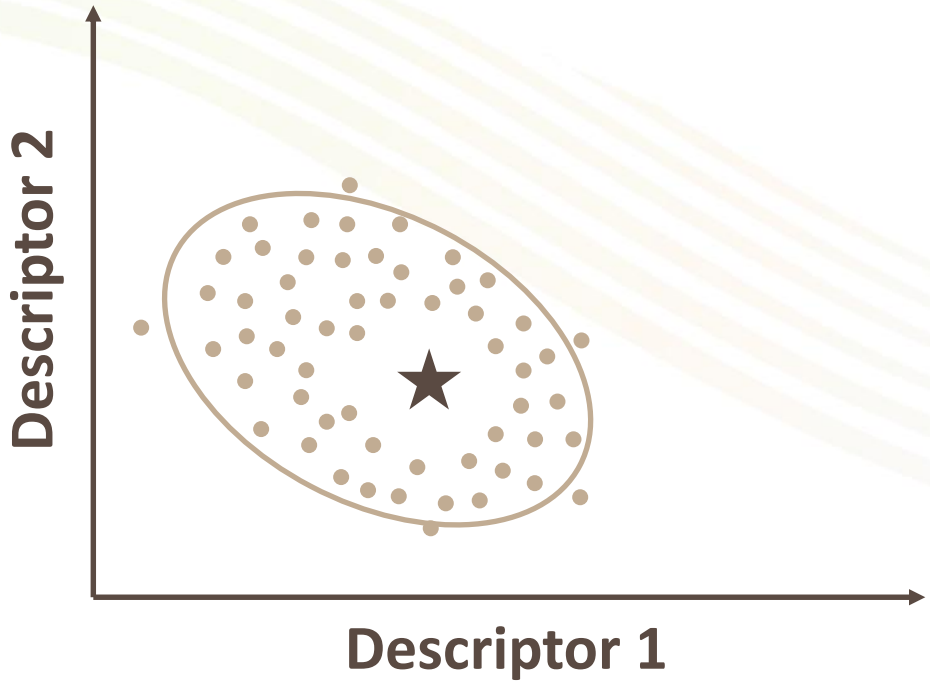
- Experimental noise in training data
- Descriptors may not capture all sources of variation
  - Modelled property may not be ‘smooth’ in descriptor space, limiting ability to interpolate
- New compound may be ‘different’ from those used to train the model
  - ‘Domain of applicability’
  - Models often have a limited ability to extrapolate beyond the descriptor space represented by the training set

Assessing Predictive Ability
Domain of Applicability



- The diversity of the training set defines the **domain of applicability** of the model
- The position of a new compound relative to the domain of applicability should be reflected in the reported confidence in the prediction
- Can we do better than ‘in’ or ‘out’ indication?

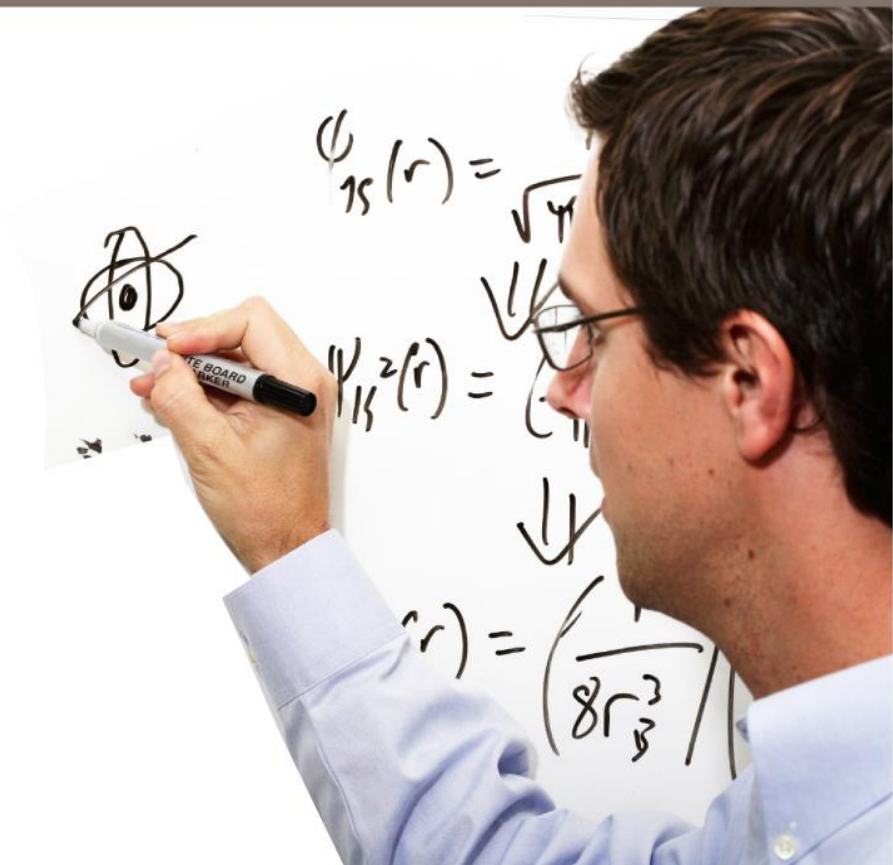
Assessing Predictive Ability
Dealing with ‘gaps’ in coverage



- Distribution of the training set may not be uniform and there may be ‘gaps’ or sparsely sampled regions
- Is this compounds ‘in’ or ‘out’ of the domain of applicability?

Introduction to Gaussian Processes (GPs)

optibrium



Modelling Techniques: Gaussian Processes

- A machine learning method based on Bayesian approach
- 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

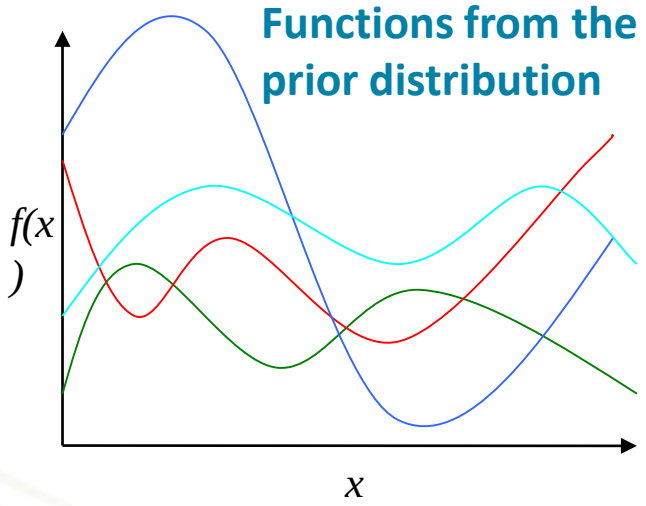
Modelling Techniques: Gaussian Processes

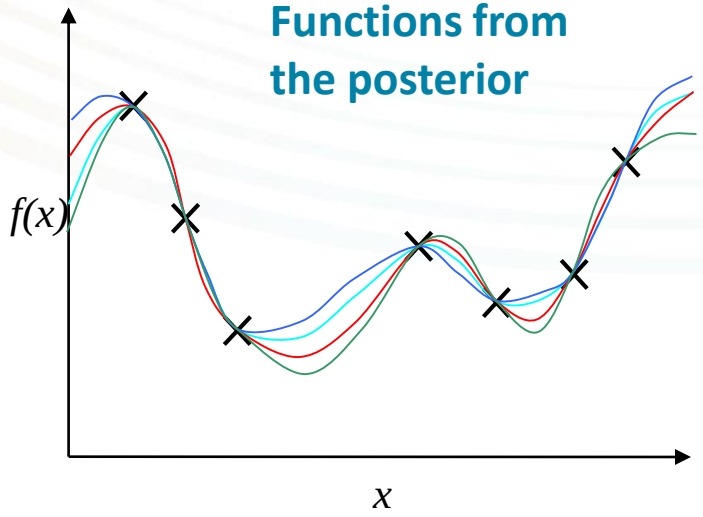
- Define **prior distribution** over functions (controlled by hyperparameters, covariance function – ARD function)

- **Posterior distribution**: retain functions which fit experimental data

- **Prediction** is the mean of posterior distribution.

- Standard deviation of the distribution provides estimate of the **uncertainty in prediction**



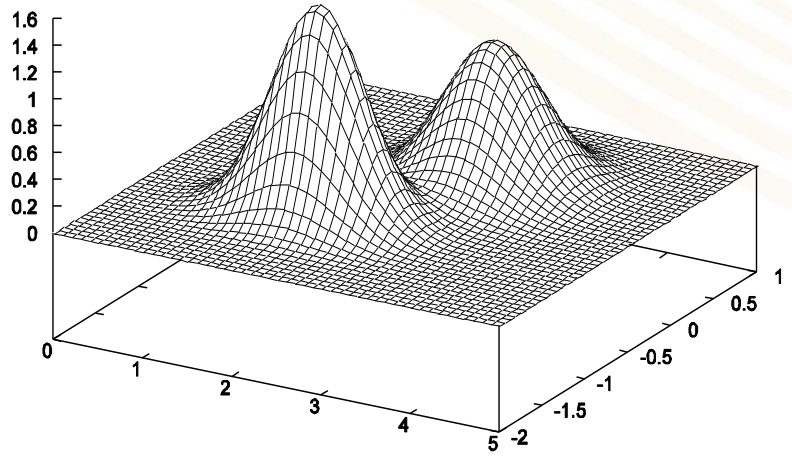


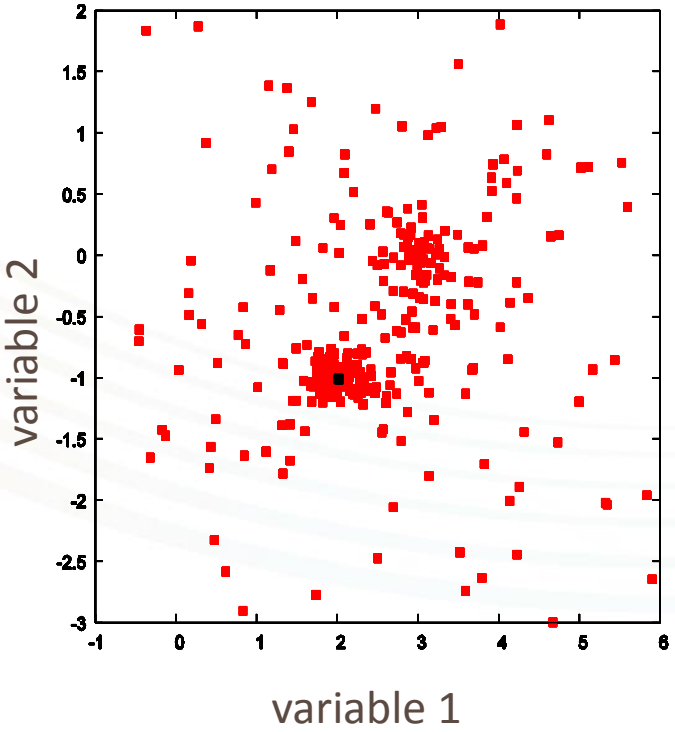
Gaussian Processes: Hyperparameters

- Learning the Gaussian Process ~ finding hyperparameters
  - Optimize the marginal log-likelihood (prevents overtraining)
  - Fits parameter corresponding to estimate of noise in input data (assuming normally distributed)
- Techniques for finding hyperparameters
  - "Fixed" values for length scales. Search for noise parameter
  - Forward variable selection provides feature selection
  - Optimisation by conjugate gradient methods
    - Length scales show which descriptors are most relevant
  - Nested sampling
    - Search in the full hyperparameter space
    - Search does not get trapped in local maxima

Gaussian Processes: Nested Sampling

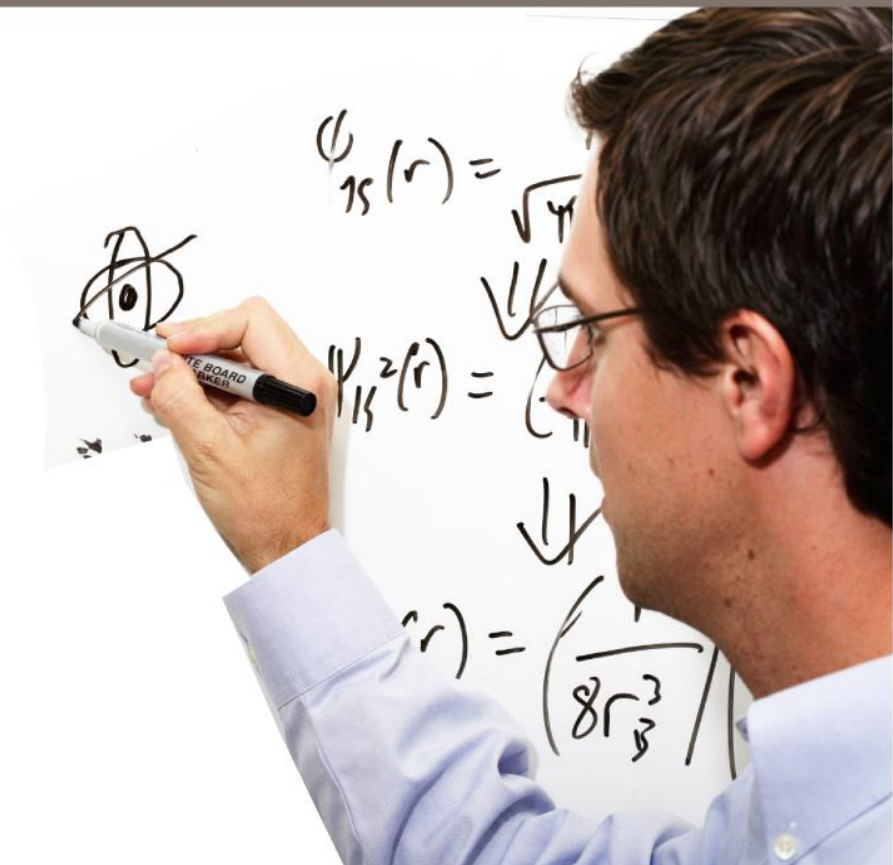
- Illustration for 2 variables
- Find maximum of likelihood:





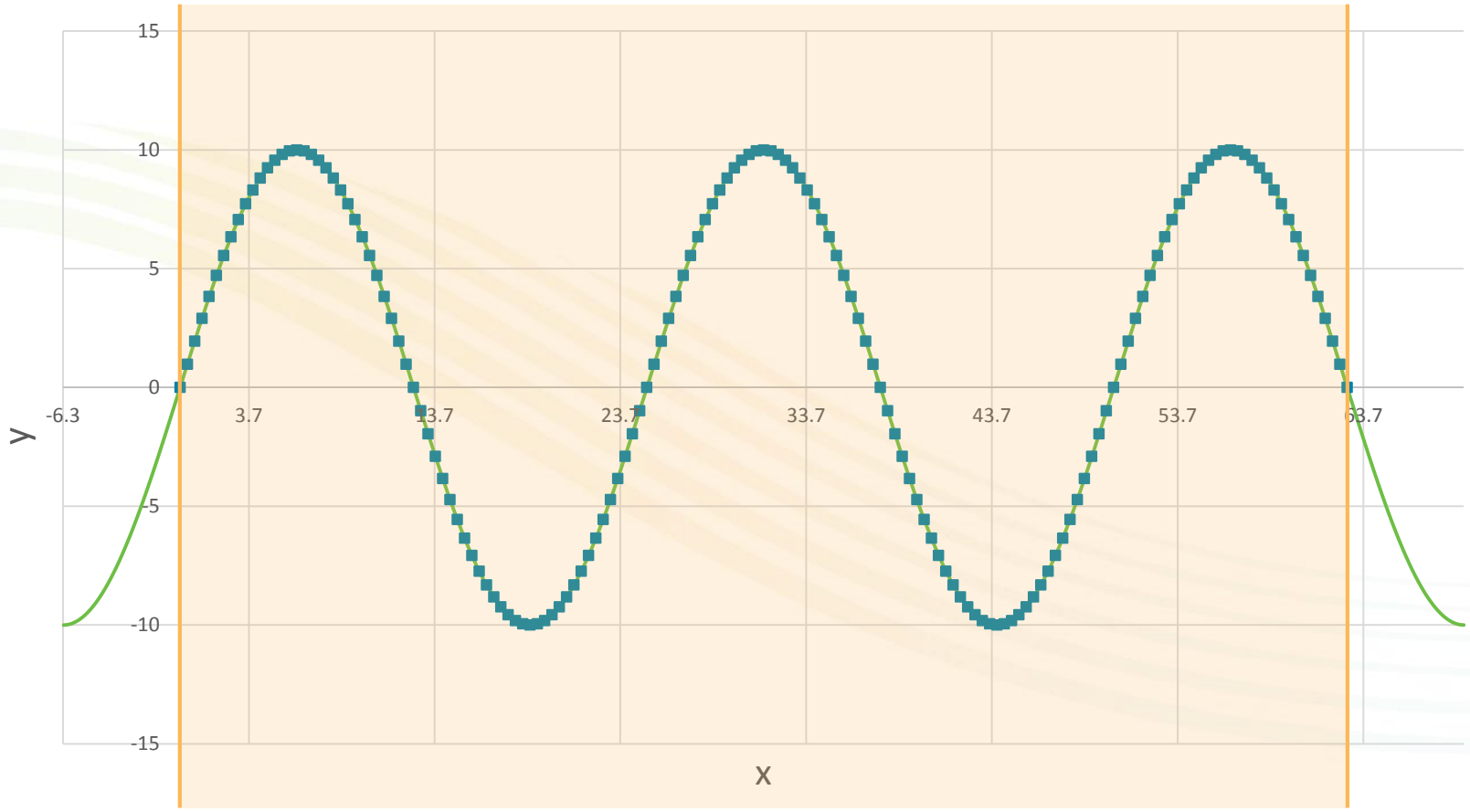
Simple Illustration of GPs

optibrium



‘Toy’ Example

Training set from sin function in 1 dimension

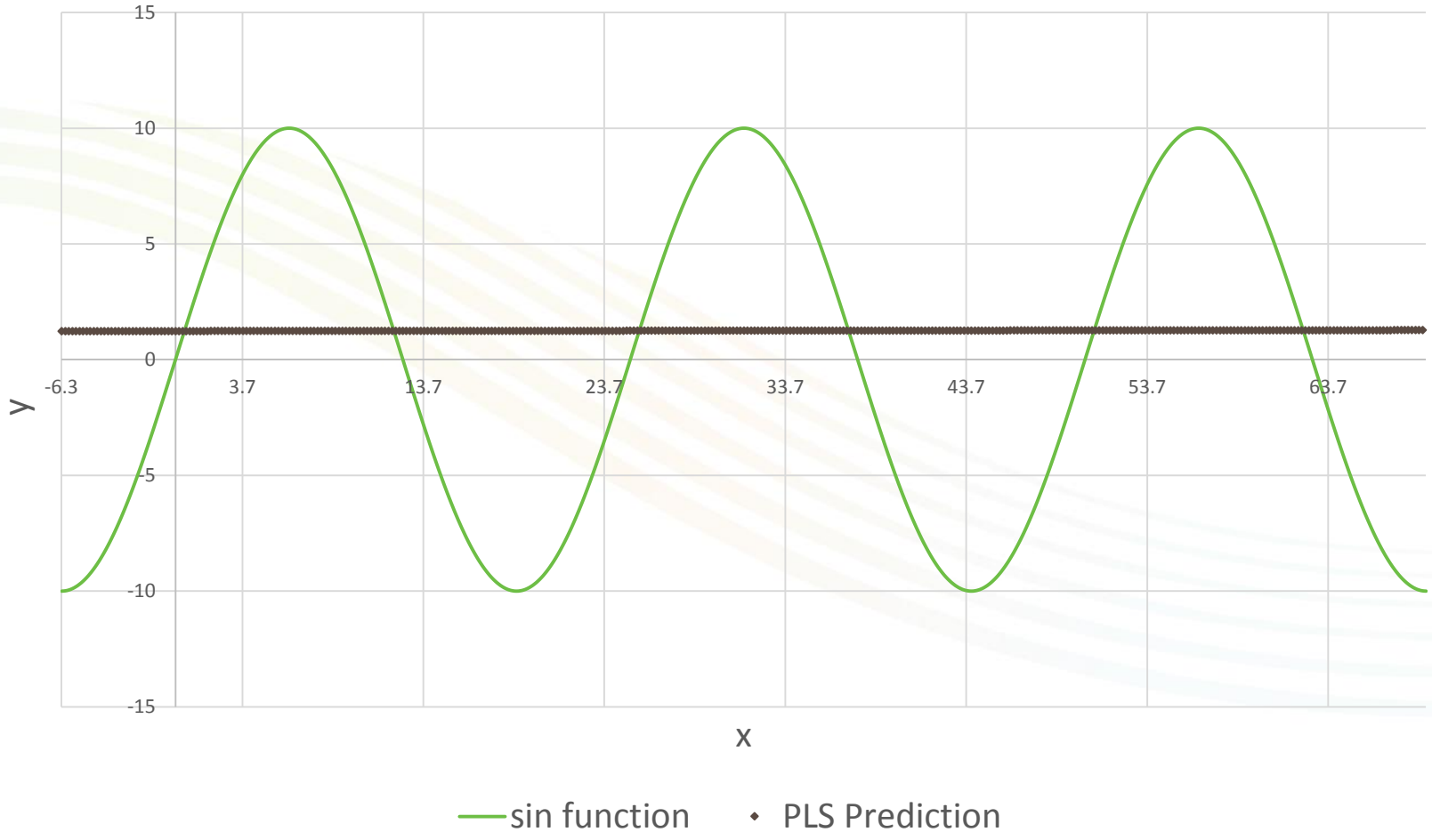


Lower boundary of training set

Upper boundary of training set

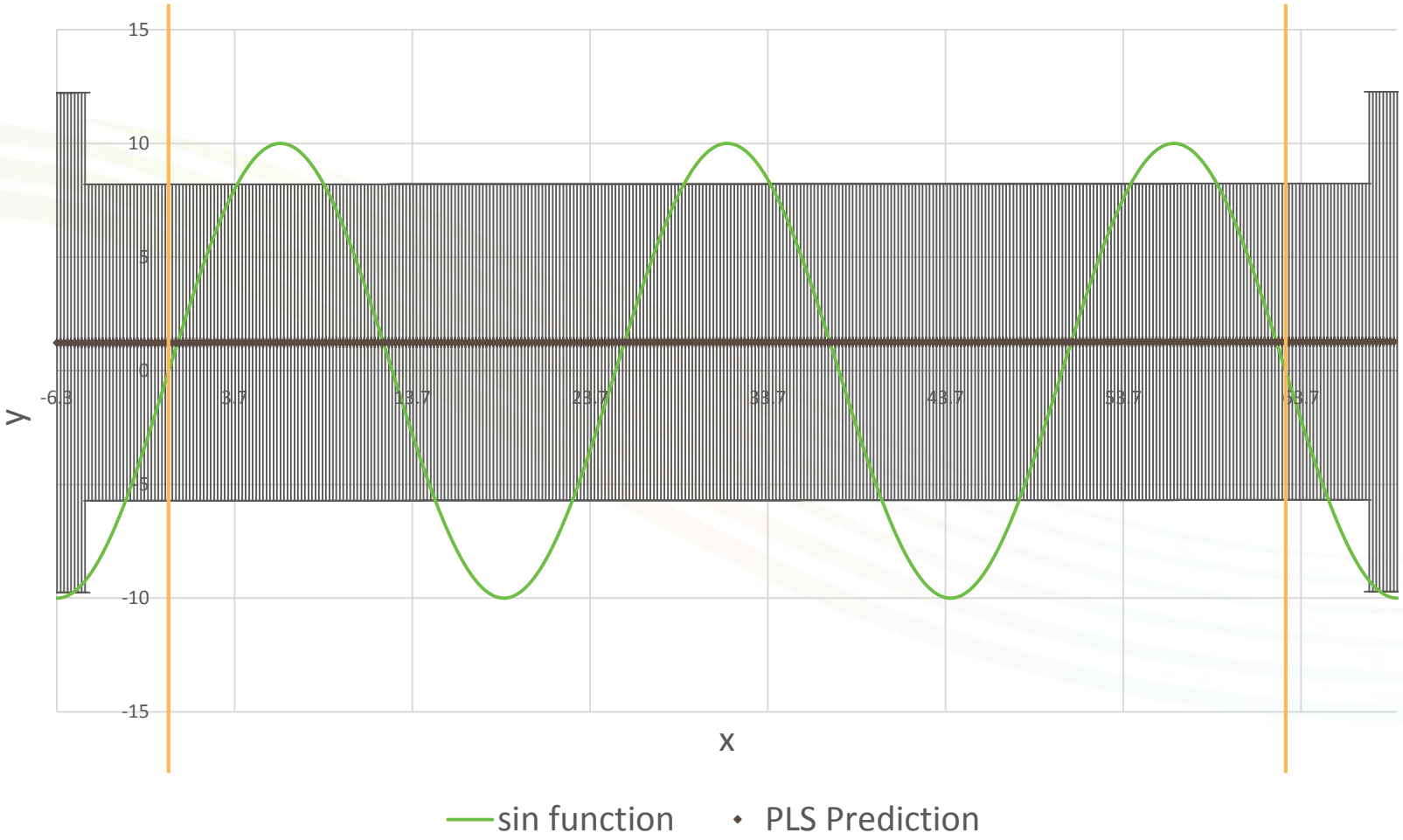
Partial Least Squares

Linear model not appropriate



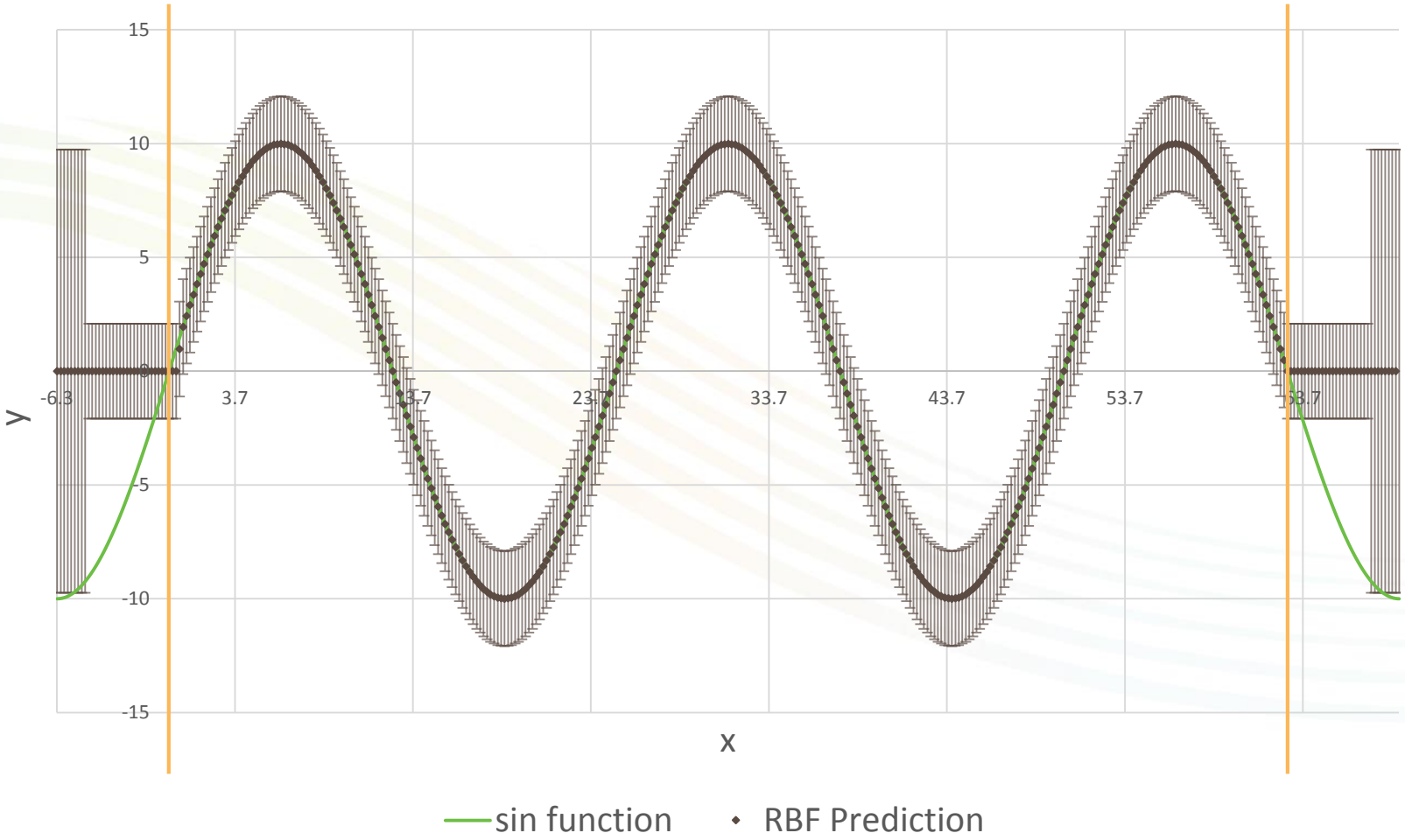
Partial Least Squares

Linear model not appropriate



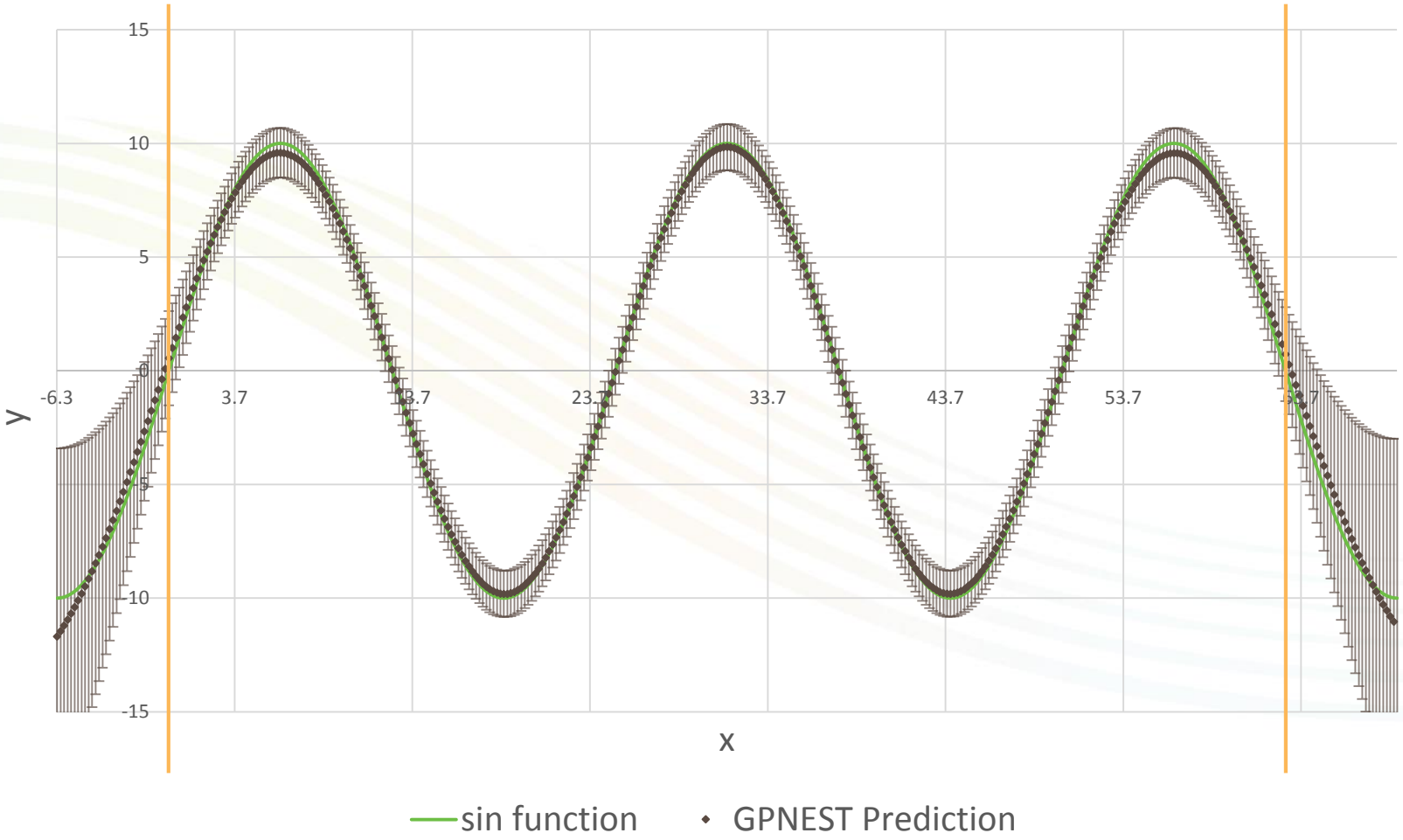
Domain of applicability based on Hotelling's T2 test with 95% confidence limit
Error bars based on RMSE error inside and outside of domain of applicability

Radial Basis Function Model



Domain of applicability based on Hotelling's T2 test with 95% confidence limit
Error bars based on RMSE error inside and outside of domain of applicability

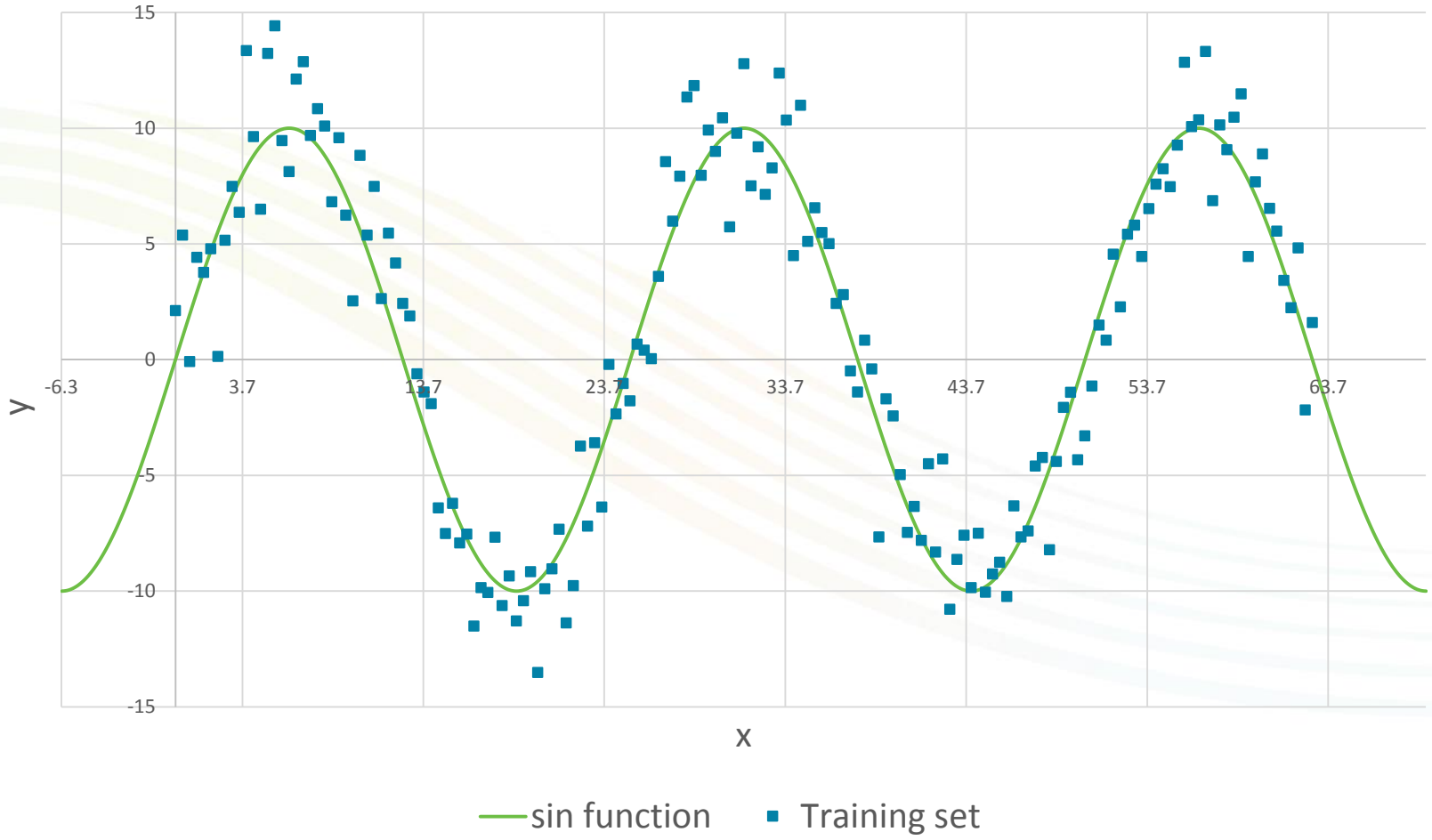
Gaussian Processes (Nested Sampling)



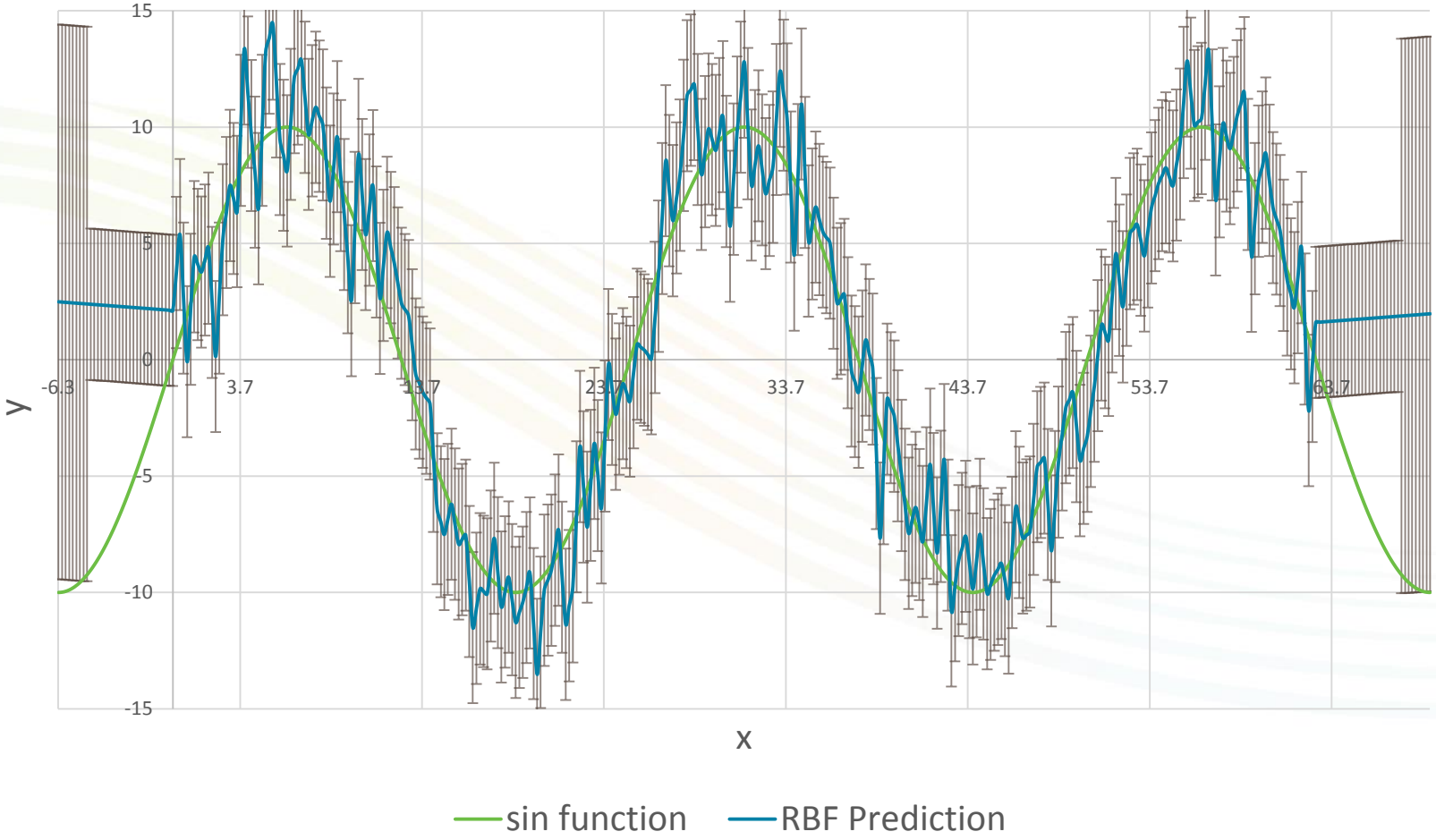
Error bars estimated by standard deviation of Gaussian process for each predicted value

Training Set with Noise

Normally distributed error with standard deviation of 2

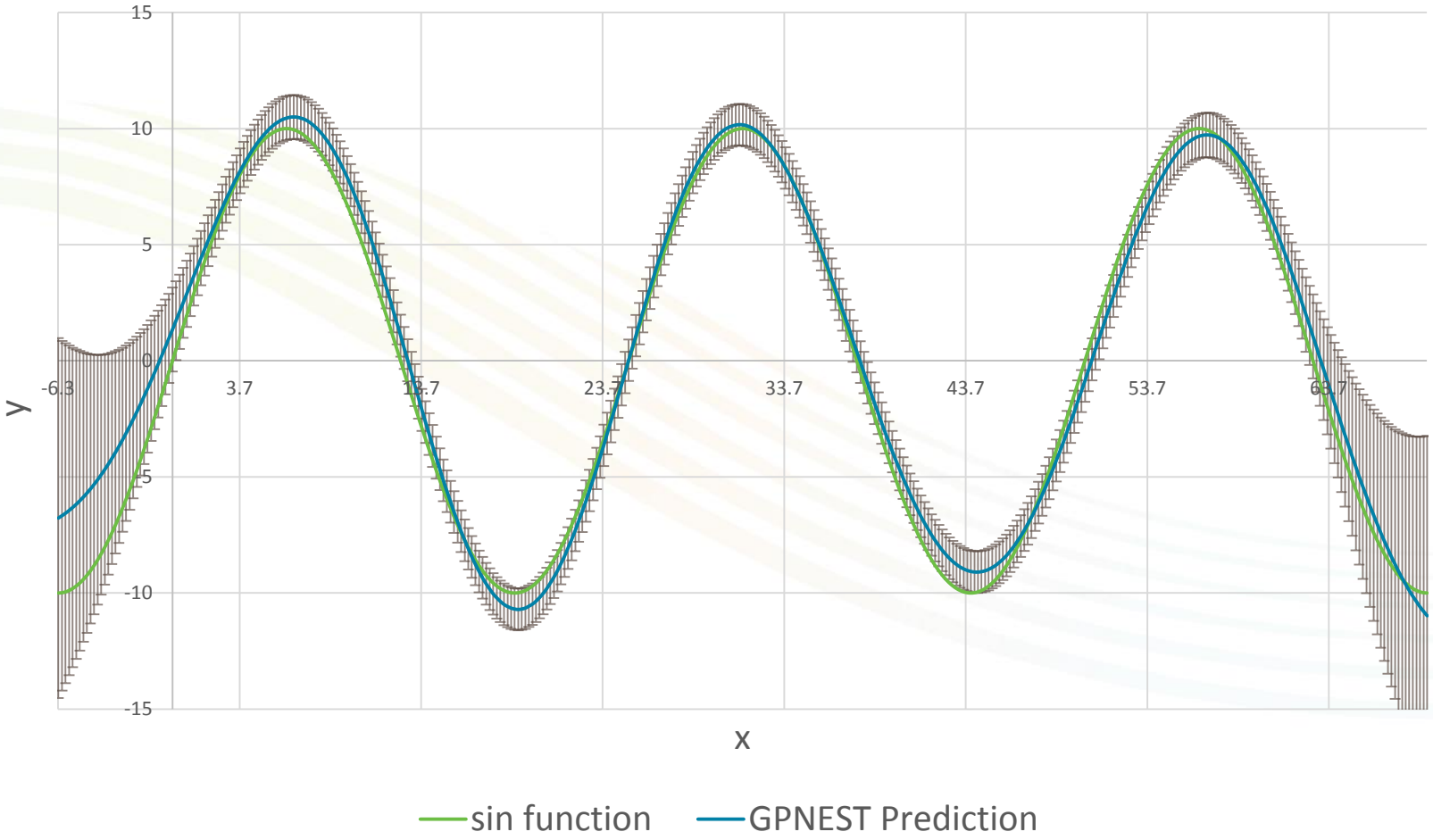


RBF Model of Noisy Data



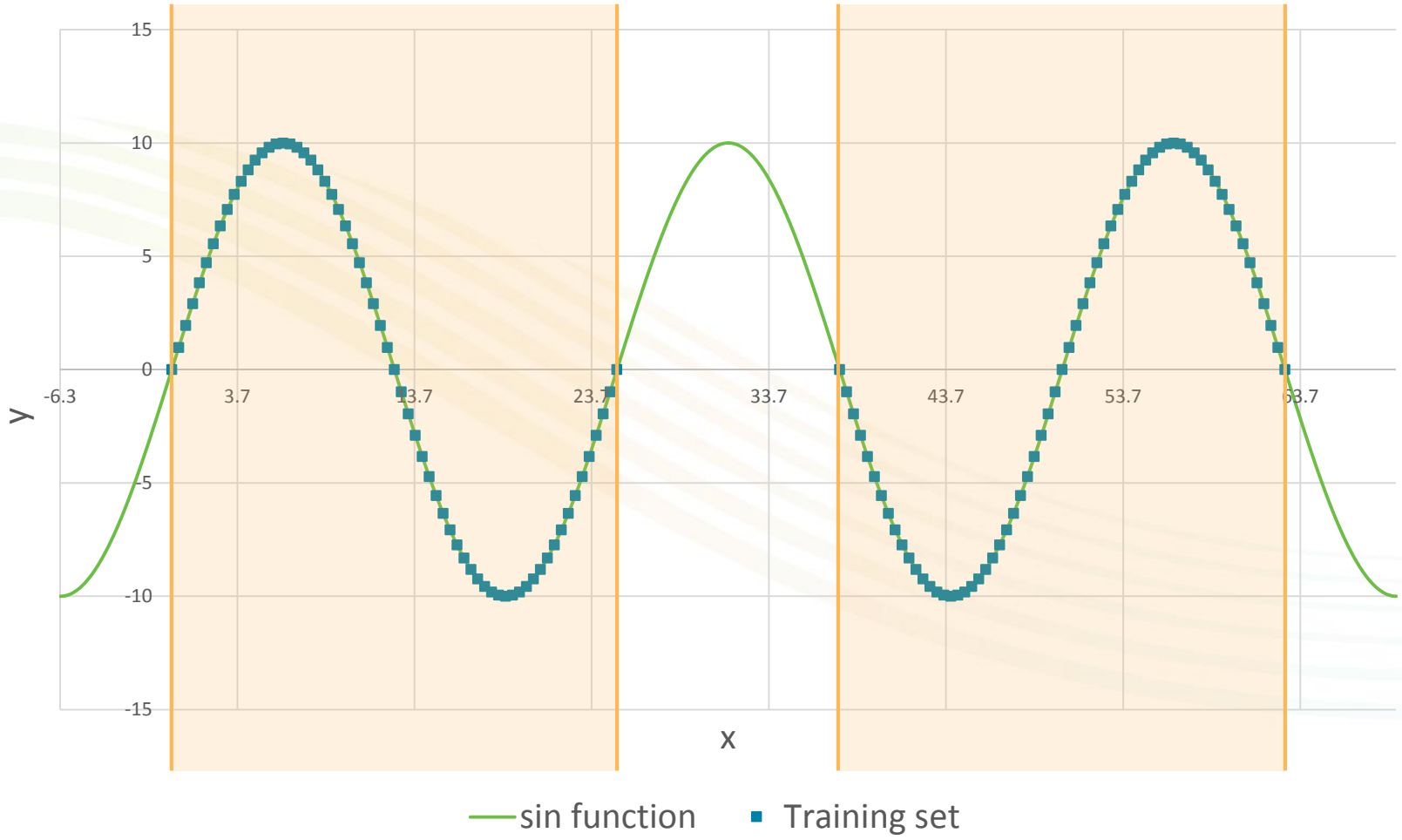
Greater error in prediction in domain of applicability, so error bars increase accordingly

GP Model of Noisy Data

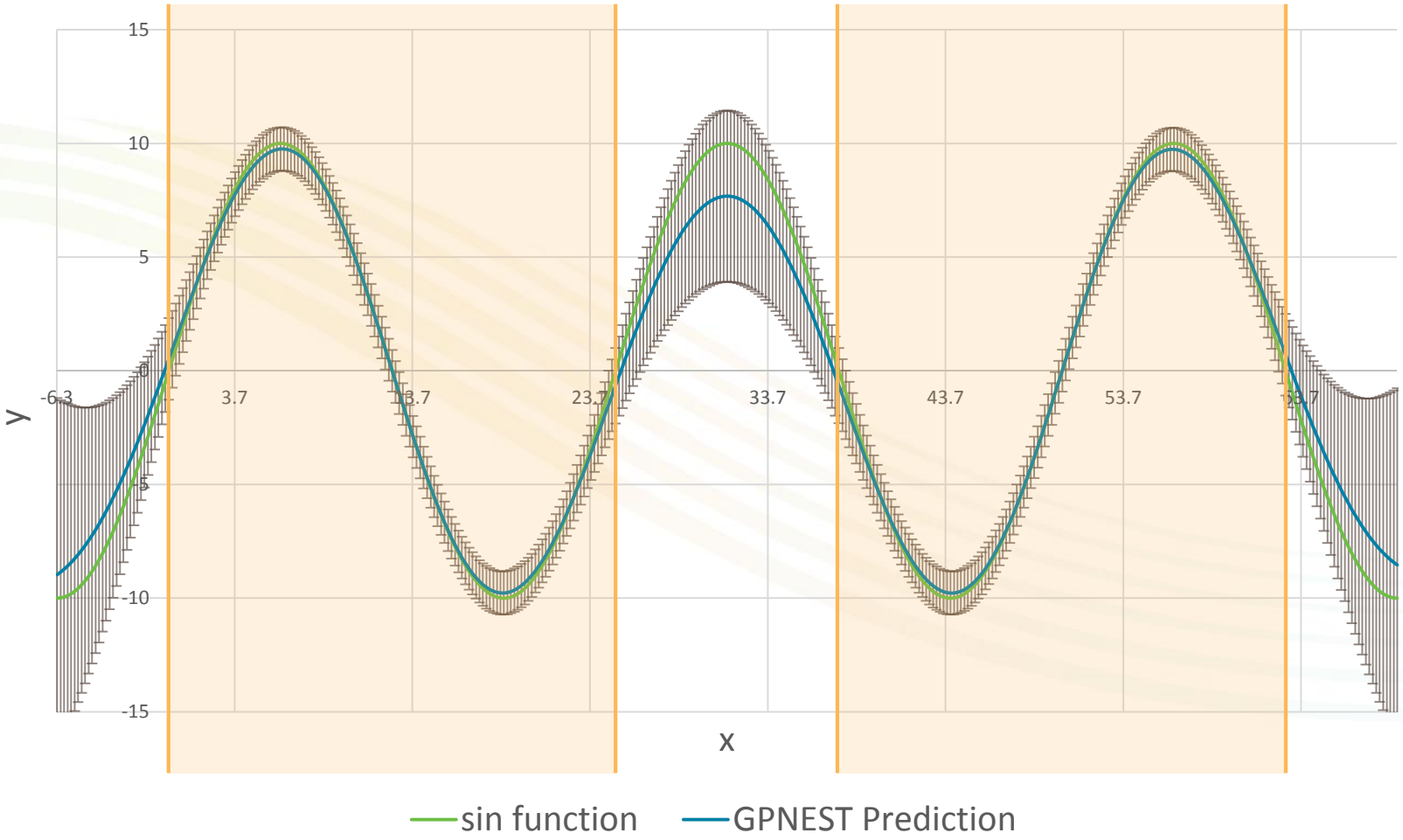


GP infers underlying functional form, fits model of noise and corrects error bars accordingly
More difficult to extrapolate, but this is accounted for outside domain of applicability

Training Set with Missing Data



GP Model Built with Missing Data

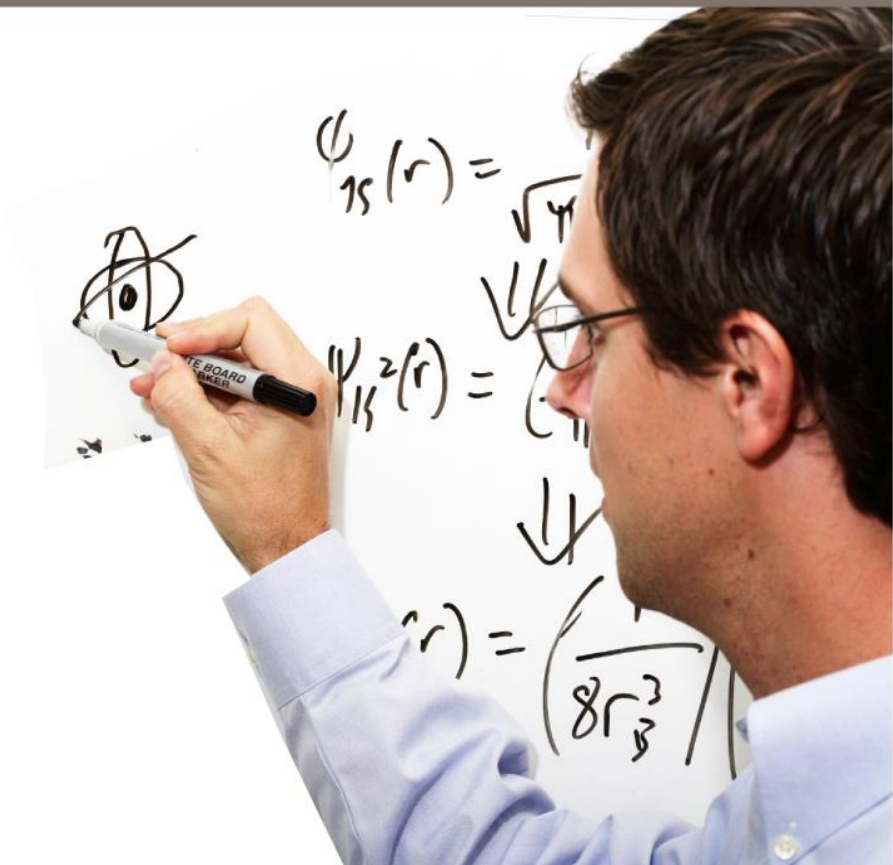


Where data is sparse or missing uncertainty in prediction is higher, as reflected by larger error bars

Automatic Relevance Determination

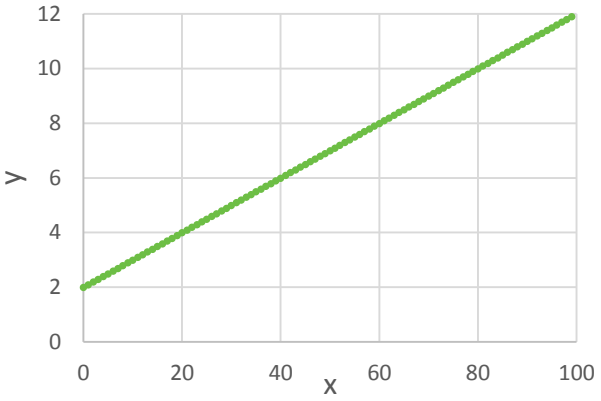
Identifying most important descriptors

optibrium



Experiment
Detecting relevant descriptors

- 100 training data points



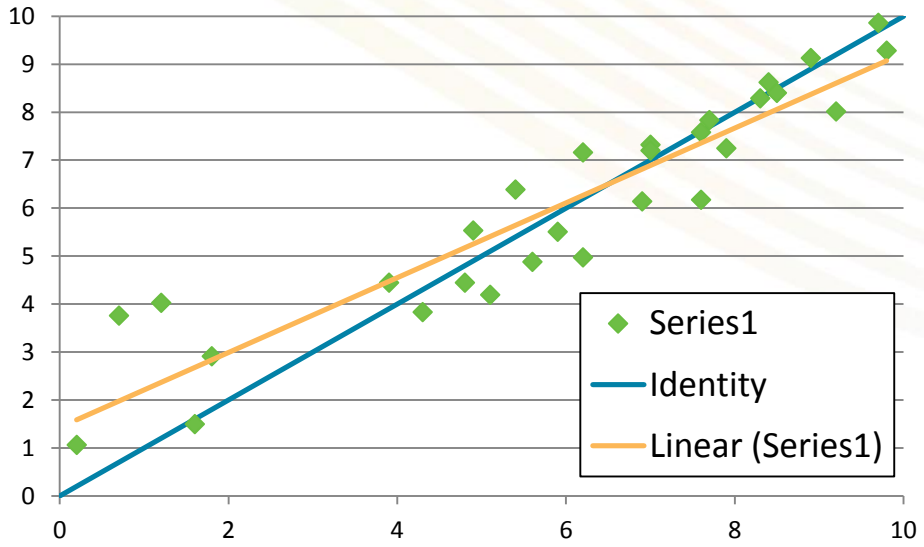
- One descriptor (x) with perfect linear correlation with property (y)
- Hide this descriptor in a data set containing N_{random} randomly generated descriptors

Identifier	y	x	Rnd 1	Rnd 2	Rnd 3	Rnd 4	…	Rnd N_{random}
Compound 1	0	2	1.252494	4.741985	2.14597	9.457343	3.958759	4.780421
Compound 2	1	2.1	8.64592	1.747653	6.76429	3.99527	7.626024	8.251326
Compound 3	2	2.2	7.064635	5.553097	0.355306	9.588649	2.791829	9.871042
Compound 4	3	2.3	4.783329	5.126768	3.525646	2.39005	0.392087	7.550868
…	4	2.4	7.077723	1.938555	2.028159	7.487378	9.672227	9.300353
Compound 100	99	11.9	0.588886	7.136536	9.538188	1.295742	3.522841	9.480185

- Can a method find the relevant descriptor?

Assessing Predictive Ability
Validation of Regression Models

- Coefficient of Determination – $R^2 = 1 - \frac{\sum_{i=1}^{N}(y^{obs} - y^{pred})^2}{\sum_{i=1}^{N}(y^{obs} - \overline{y^{obs}})^2}$
  - Measure of fit to identity line y=x
  - N.B. Not the same as square correlation coefficient r^2_{corr} which is measure of fit to best fit line – R^2 is a stricter test

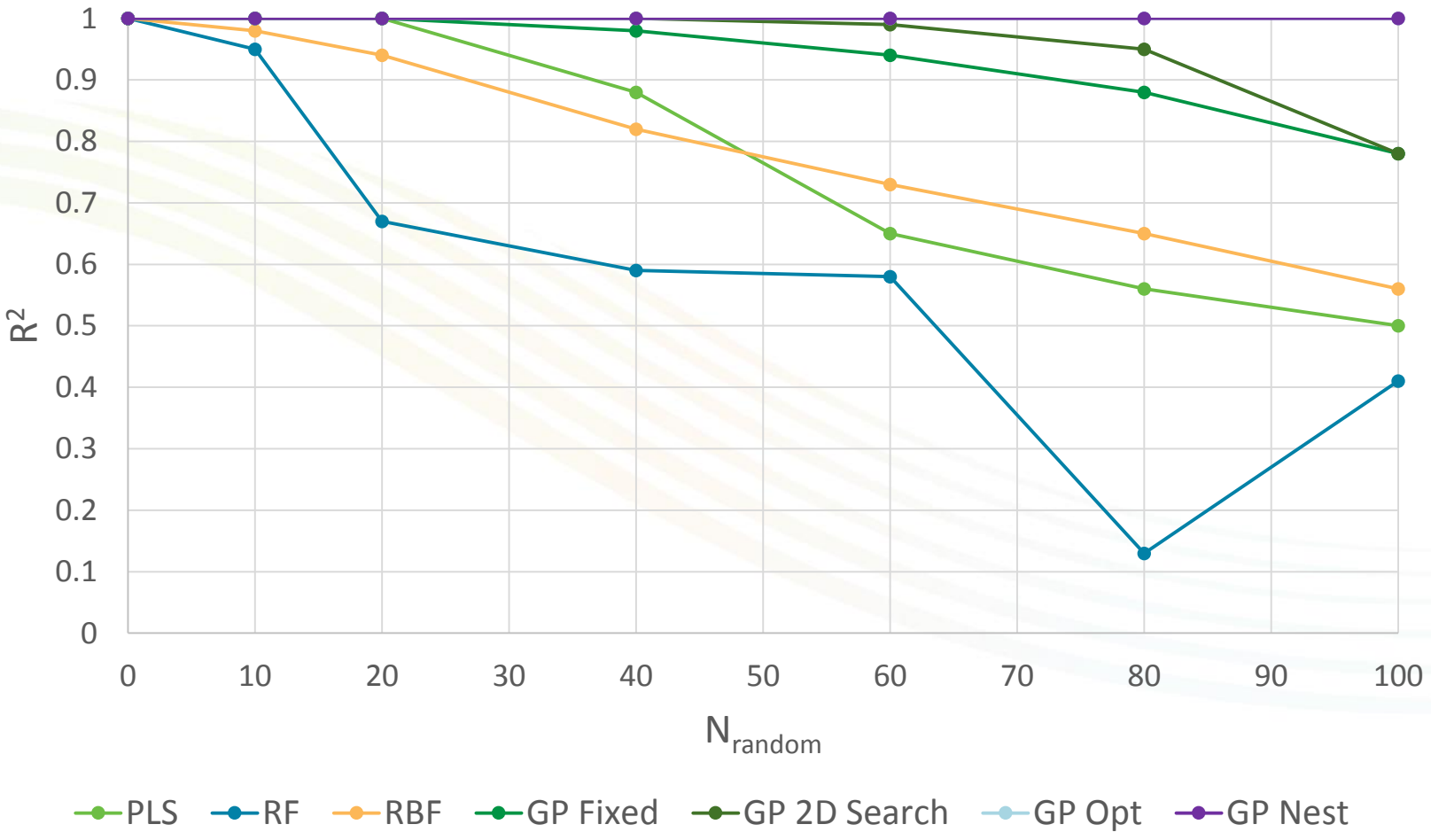


Best fit line
y= 0.71x + 2.3
r^2_{corr} = 0.86

R^2 = 0.74

- Root mean square error - RMSE

Results
Detecting relevant descriptors

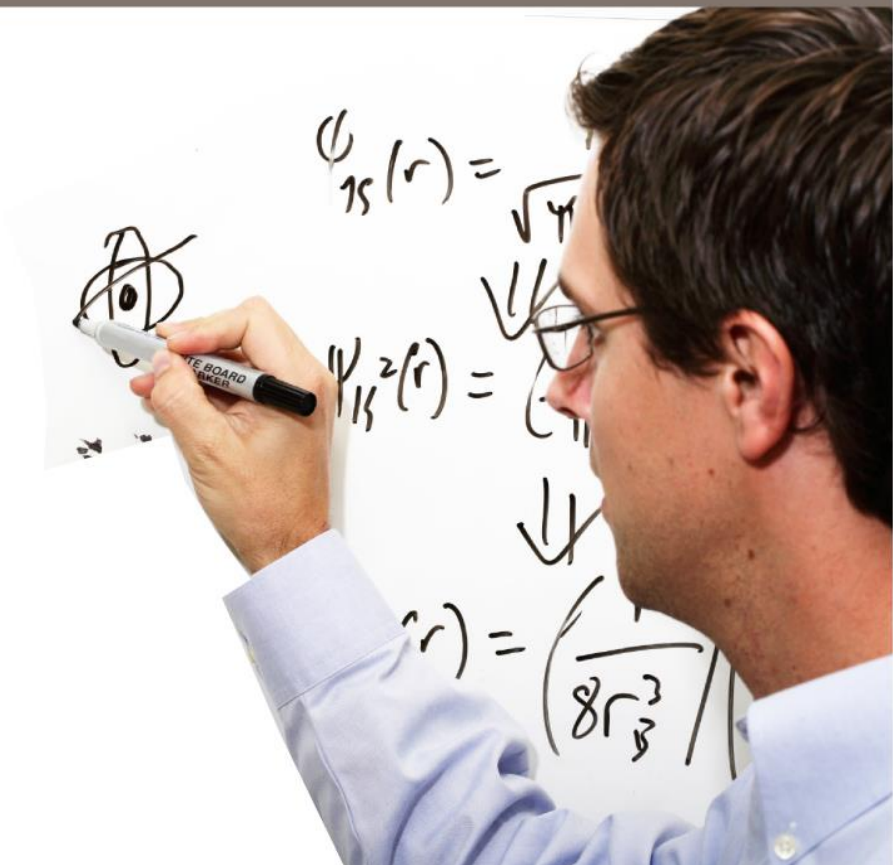


Number of Descriptors Use in Model

- Often quoted rule of thumb… at least 5 compounds in training set per descriptor
- This is relevant for simple models where the only complexity control is the number of descriptors in the model
- But, GP Nest model includes all descriptors
  - Influence of random descriptor on model is negligible
  - Posterior probability of complex models is low
- Including additional non-influential descriptors in the model can be valuable
  - Detect when new compound differs significantly from training set
  - Uncertainty in prediction will increase

Practical Application to QSAR Modelling

optibrium

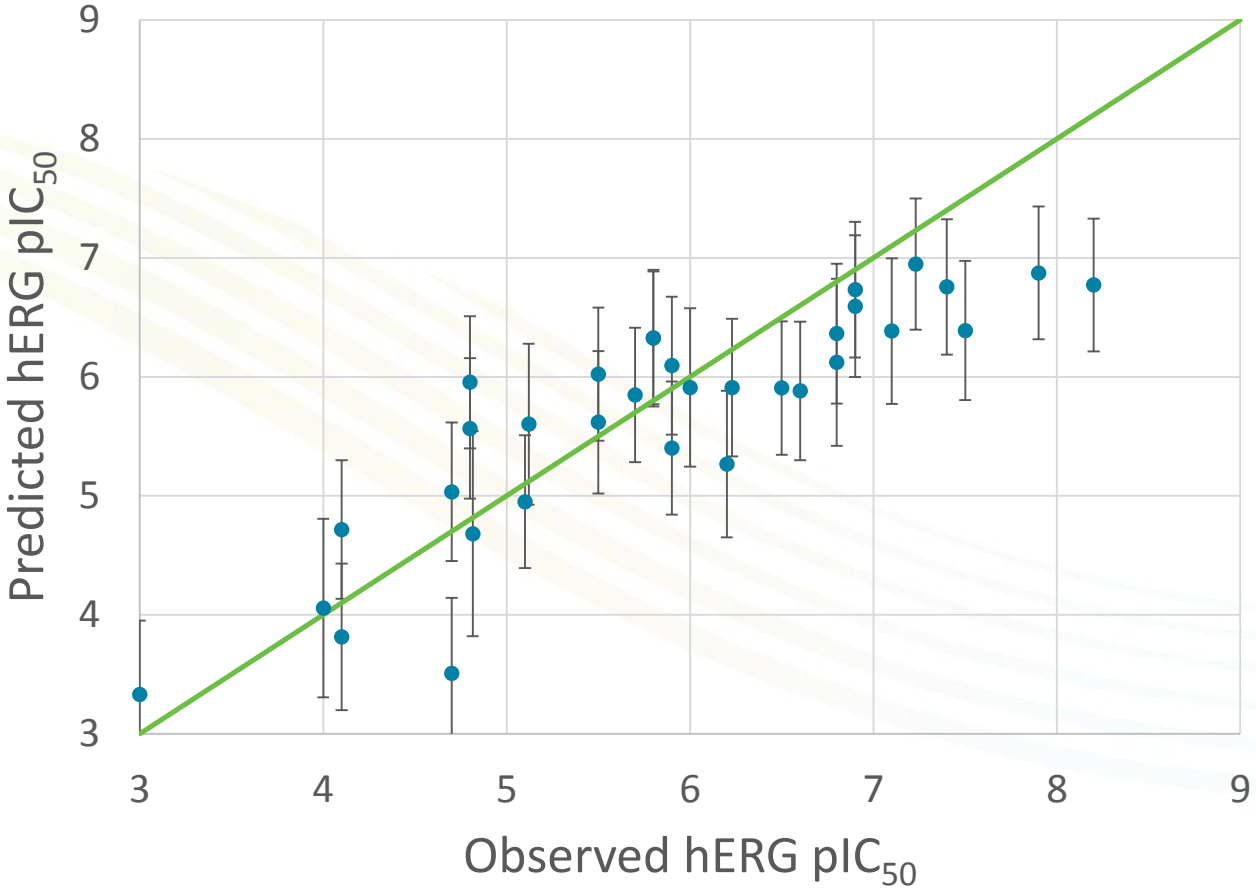


hERG pIC_{50}

- Diverse data set of 168 compounds
  - All manual patch clamp measurements in mammalian cells
- Divided into training (135) and external test (33) sets
- Descriptors including
  - Whole molecule properties: logP, V_x, TPSA, MW, flexibility…
  - 156 structural descriptors expressed as SMARTS

Method	R^2	RMSE
GP (Nested sampling)	0.72	0.64
RF	0.68	0.68
RBF	0.70	0.66

hERG pIC_{50} GP Nested Sampling Results



64% within 1 SD, 94% within 2 estimated SD (assuming 0.5 log units uncertainty in expt. data)

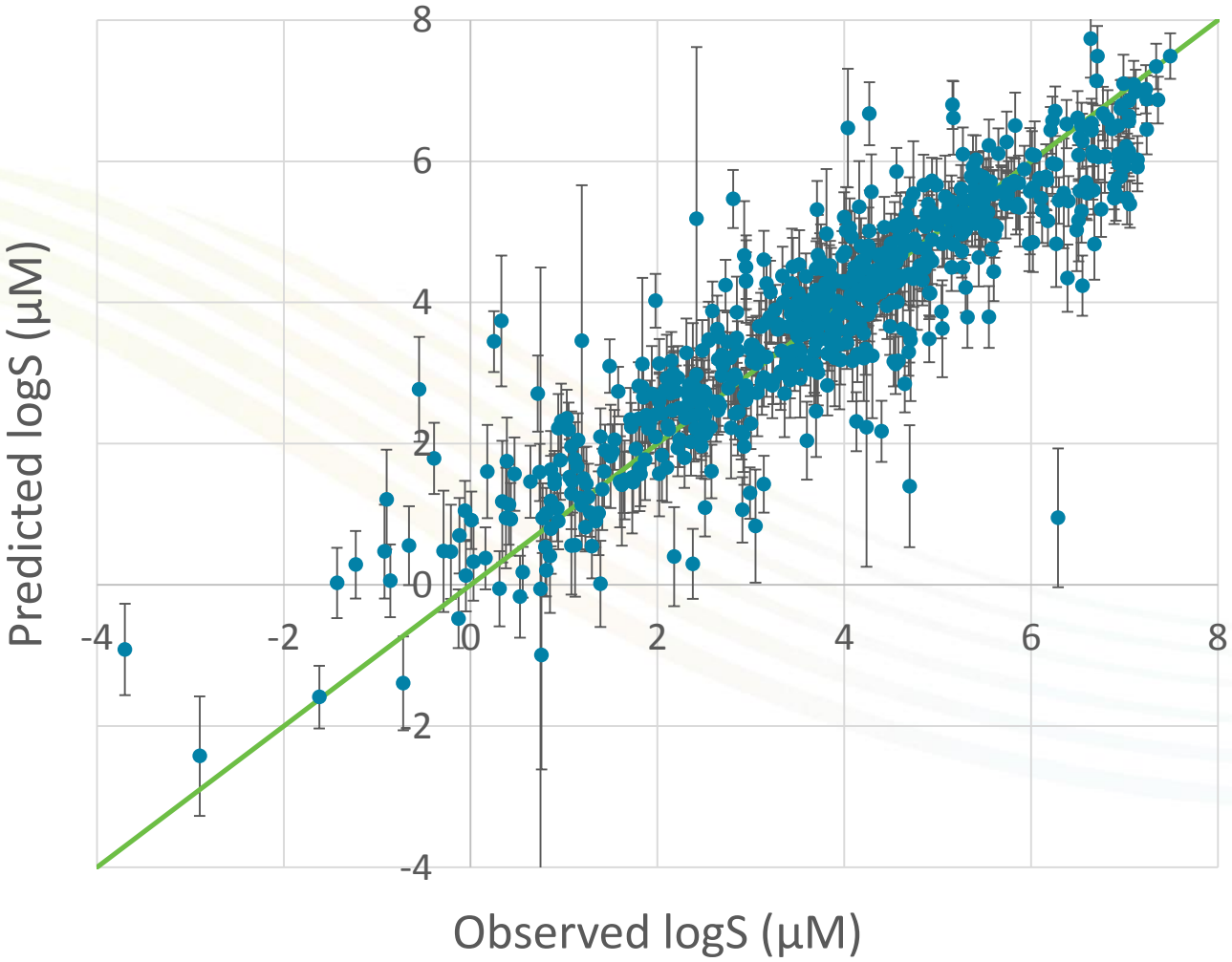
Aqueous Solubility (logS)

- Diverse data set of 3313 compounds
  - Log of intrinsic, thermodynamic, aqueous solubility in µM, measured between 20 and 30 °C
- Divided into training (2650) and external test (663) sets
- Descriptors including
  - Whole molecule properties: logP, V_x, TPSA, MW, flexibility…
  - 164 structural descriptors expressed as SMARTS

Method	R^2	RMSE
GP (Fixed)	0.81	0.80
RF	0.78	0.87
RBF	0.84	0.73
PLS	0.75	0.91

Obrezanova et al. J. Comp.-Aided Mol. Des. (2008) **22**(6-7) pp. 431-440

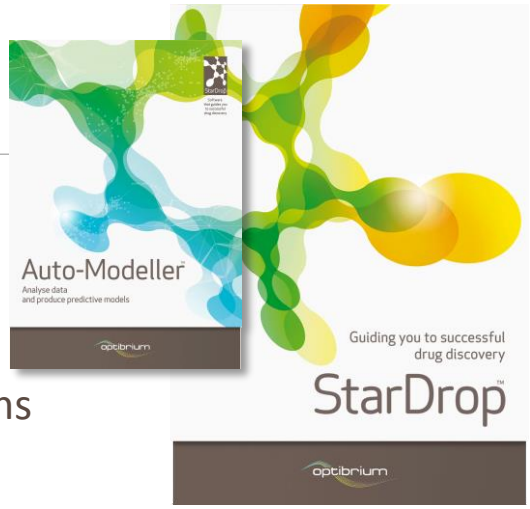
Aqueous Solubility (logS)



51% within 1 SD, 80% within 2 estimated SD (assuming 0.3 log units uncertainty in expt. data)

Obrezanova et al. J. Comp.-Aided Mol. Des. (2008) **22**(6-7) pp. 431-440

Conclusions



- Gaussian Processes
  - Bayesian non-linear modelling technique
  - Generates a probability distribution over possible functions
  - Explicitly calculates uncertainties for each prediction
  - Similar performance to methods such as random forests, radial basis functions…
  - Can also be used for classification
- Limitations
  - Most expensive optimisations methods are computationally expensive (e.g. nested sampling $O(N^4)$)
  - Can't deal with potential sources of variability (e.g. structural features) not captured by descriptors
- More information (www.optibrium.com/community)
  - Obrezanova *et al.* JCIM (2007) **47**(5) pp. 1847-57
  - Obrezanova *et al.* JCAMD (2008) **22**(6-7) pp. 431-440
  - Obrezanova *et al.* JCIM (2010) **50** (6), pp. 1053-1061

Acknowledgements

- Olga Obrezanova
- Iskander Yusof
- Chis Leeding
- All our other colleagues at Optibrium