optibrium

Comprehensive comparison of automatically generated QSAR models of target potency

ACS Spring National Meeting, March 29th 2012

Edmund Champness, Matthew Segall, Joelle Gola, Delphine Lariviere, Chris Leeding, Iskander Yusof, James Chisholm

© 2012 Optibrium Ltd.
Optibrium™, StarDrop™, Auto-Modeller™ and Glowing Molecule™ are trademarks of Optibrium Ltd.

Overview

- Objective
- Data sets
- Automatic QSAR model generation
- Results
  - Summary of model results
  - Comparison of algorithms
  - Comparison of data types
- Conclusion

© 2012 Optibrium Ltd.

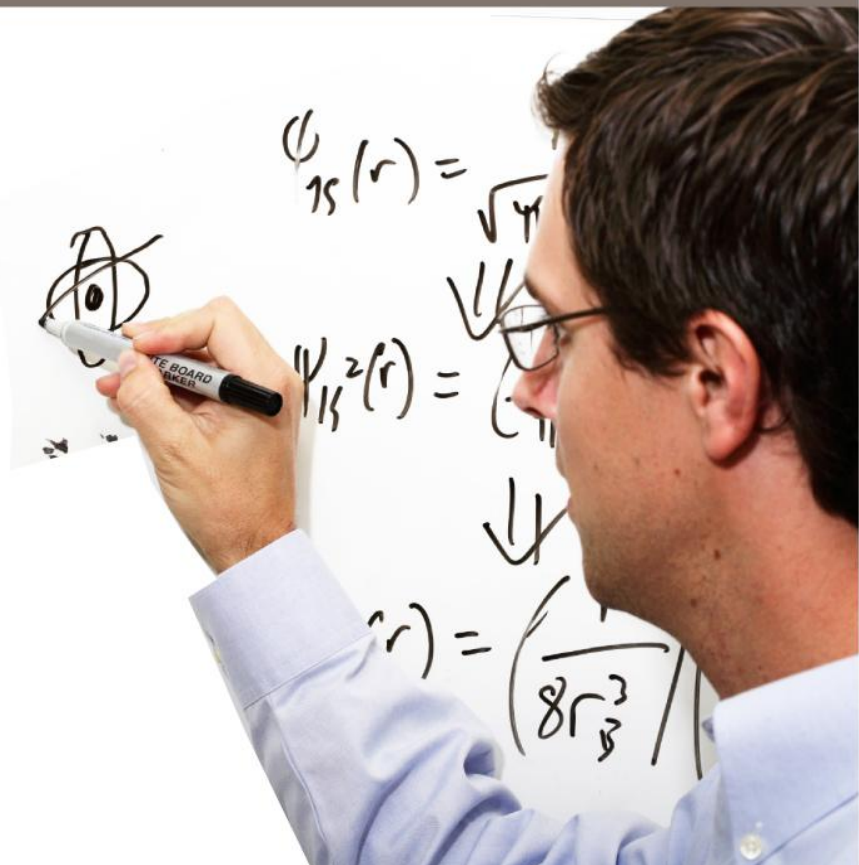
Objective

- Large sets of biological data now available in the public domain
- Assess potential for large-scale QSAR modelling of target potency
- Consistently compare different algorithms and data for 2D QSAR modelling

© 2012 Optibrium Ltd.

Data sets

optibrium



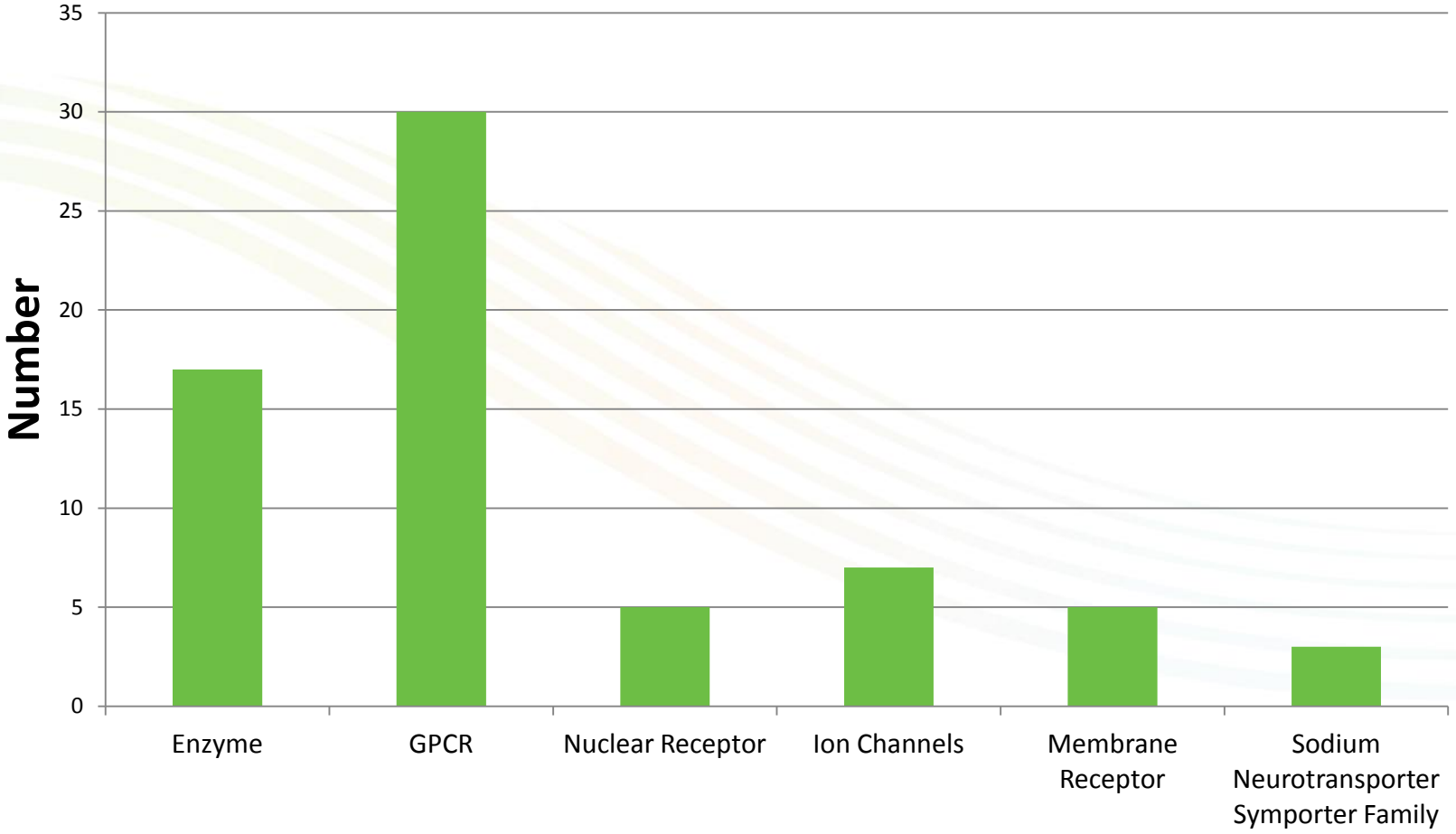
Data sets

- Data obtained from Chembl database
  - https://www.ebi.ac.uk/chembl/
- Data obtained for 67 targets selected from top human drug targets*
- Total of 104 data sets
  - 19 with IC_{50} values only
  - 12 with K_i values only
  - 36 with both IC_{50} and K_i values
- Data set sizes ranged from 29 to 1716 compounds

* Overington *et al.*, Nat. Rev. Drug Discov. 2006, **5**, 993-996

© 2012 Optibrium Ltd.

Distribution of Target Classes



© 2012 Optibrium Ltd.

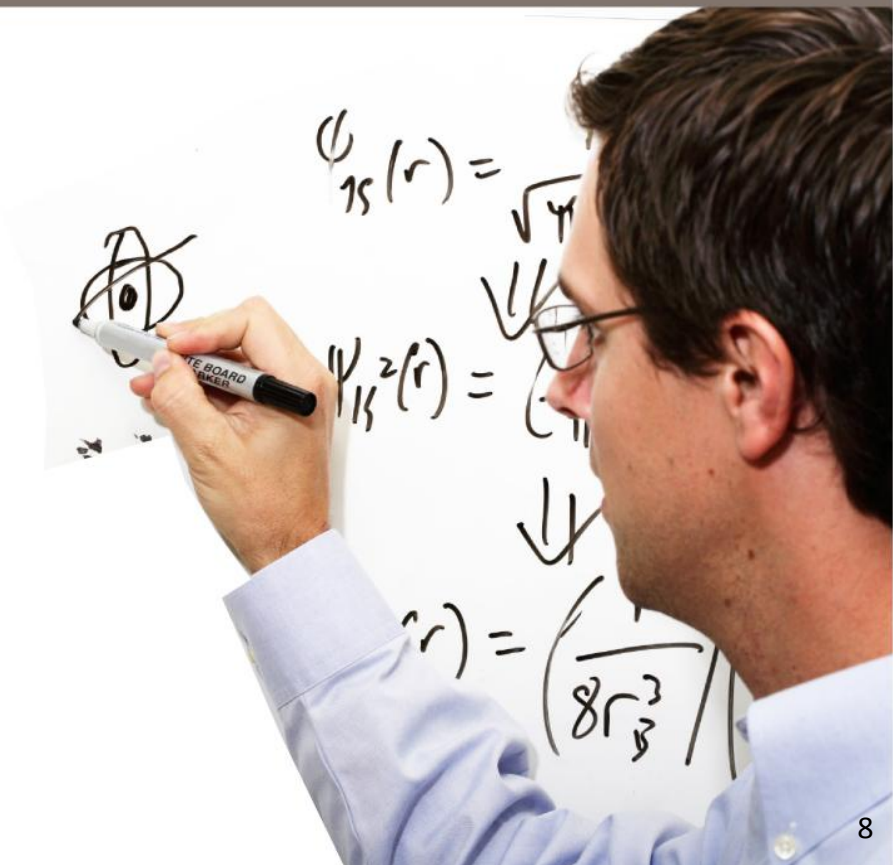
Cleaning the data

- Raw data from Chembl required further cleaning
- Rejected data with confidence <4
- Normalised units
  - Log(K_i) or log(IC_{50}) in nM
- Where multiple values present for the same compound
  - Calculate mean and standard error in mean
  - Reject compound in standard error in mean >0.5 log units
- Generate ‘parent’ structure
  - Remove salts
  - Not able to check structure for each compound (too many data points)

© 2012 Optibrium Ltd.

Automatic QSAR Model Generation

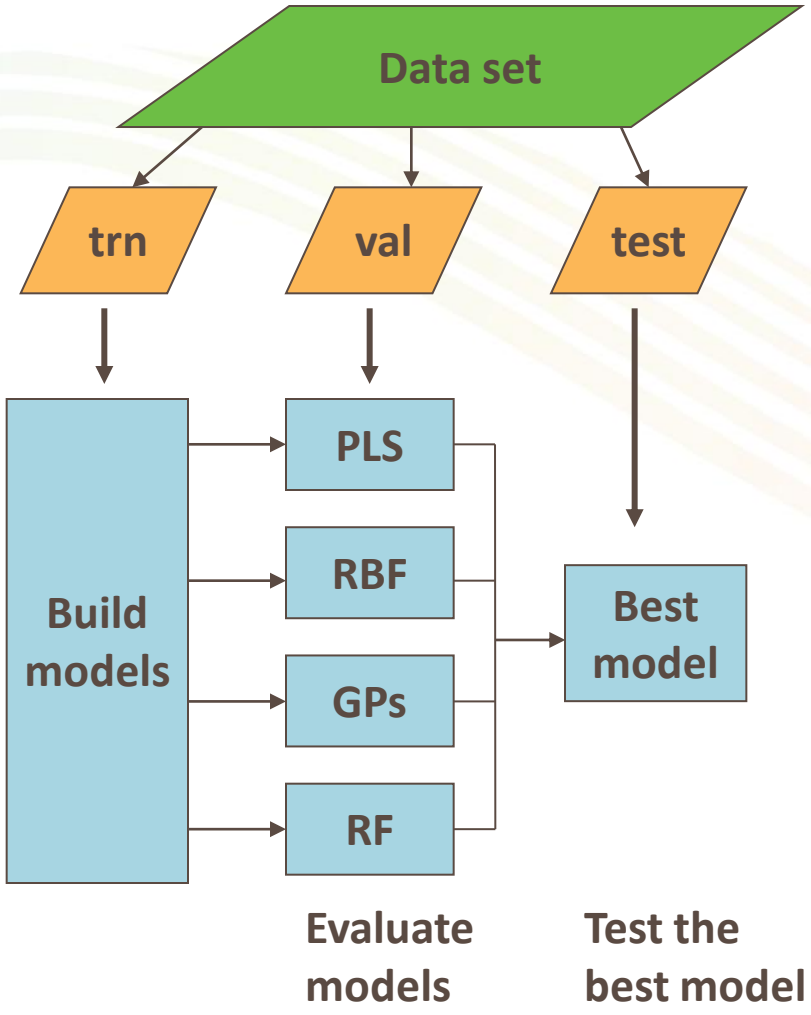
optibrium



Automatic Model Generation*
StarDrop Auto-Modeller

StarDrop 5
Software that guides you to successful drug discovery

* Obrezanova *et al.* JCAMD 2008, **22**, 431-440



- Split data set
- Calculate descriptors
- Multiple modelling techniques
- Select the best model by performance on the validation set
- Test with an **independent** set

© 2012 Optibrium Ltd.

Data Set Split

- Data set clustered using dbclus algorithm*
  - 2D path-based fingerprints
  - Tanimoto coefficient of 0.7
- Cluster centroids and singletons used as basis for training set
  - Ensure coverage of chemical diversity
- Remaining compounds distributed between subsets
  - Training 70%
  - Validation 15%
  - Test 15%

* Butina, J. Chem. Inf. Comput. Sci. 1999, **39**, 747-750

© 2012 Optibrium Ltd.

Descriptors

- Whole molecule properties
  - logP, TPSA, MW, V_x, flexibility...
- 321 SMARTS descriptors
  - Counts of substructures
- Descriptor preselection
  - Remove count descriptors with <4% occurrence
  - Remove descriptors with standard deviation <0.0005
  - Remove one of pairs with correlation R^2>0.95

© 2012 Optibrium Ltd.

Modelling Methods

- Partial Least Squares (PLS)*
  - Linear model
  - Number of components chosen by cross validation in training set
- Radial Basis Functions (RBF, GA-RBF)
  - Interpolation of training set in descriptor space.

$$y(X) = \sum_{i=1}^{N} w_i \Phi(| X - S_i |)$$

  - Genetic algorithm used to select descriptors in GA-RBF
  - Risk of overtraining for small data sets

* Wold *et al.* The Encyclopedia of Computational Chemistry, Schleyer *et al.* eds, 1999, 1-16

© 2012 Optibrium Ltd.

Modelling Methods II

- Gaussian Processes (GP) *
  - Bayesian approach – infer posterior distribution of functions that fit data
  - **Prediction** = mean of distribution, **uncertainty** = standard deviation
  - Training == learning hyperparameters
    - 6 methods: Fixed, 2DSearch, FVS, RFVS, Opt, Nest
- Random Forests (RF)†
  - Ensemble of random trees (100 trees)
  - **Prediction** = average of output of trees, **uncertainty** = standard deviation

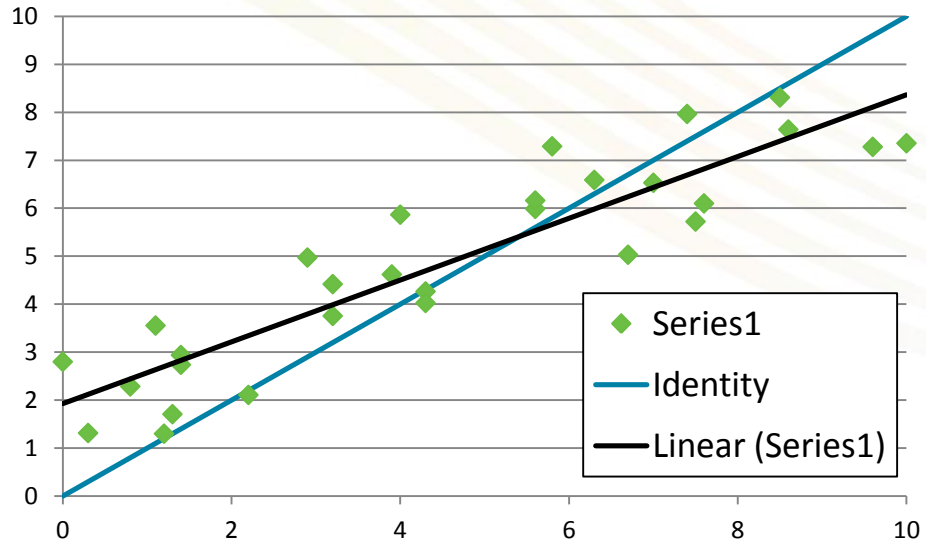
* Obrezanova *et al.* JCIM 2007, **47**,1847-57

† Breiman, L. Machine Learning 2001, **45**, 5-32

© 2012 Optibrium Ltd.

Evaluation of 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

© 2012 Optibrium Ltd.

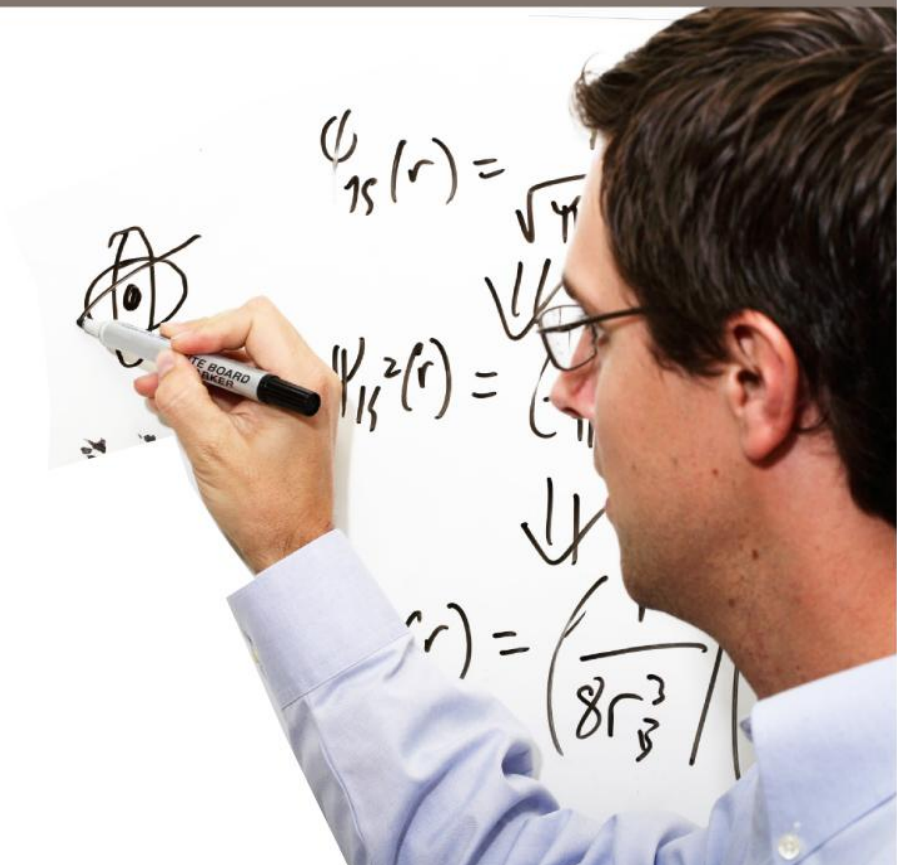
Assessing Predictive Ability
Domain of Applicability

- The diversity of the training set defines the **chemical space** of the model
- The position of a compound relative to chemical space is reflected in the reported confidence in the prediction

© 2012 Optibrium Ltd.

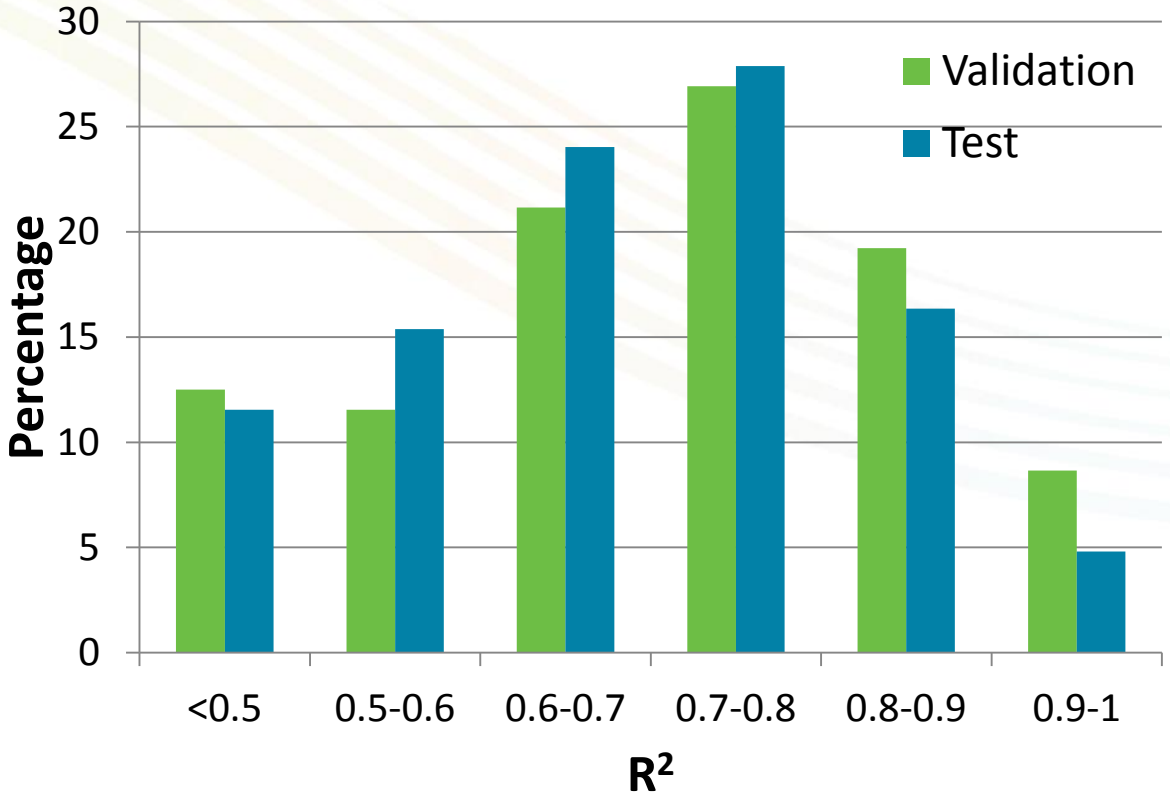
Results

optibrium



Summary of Validation Results
Coefficient of Determination

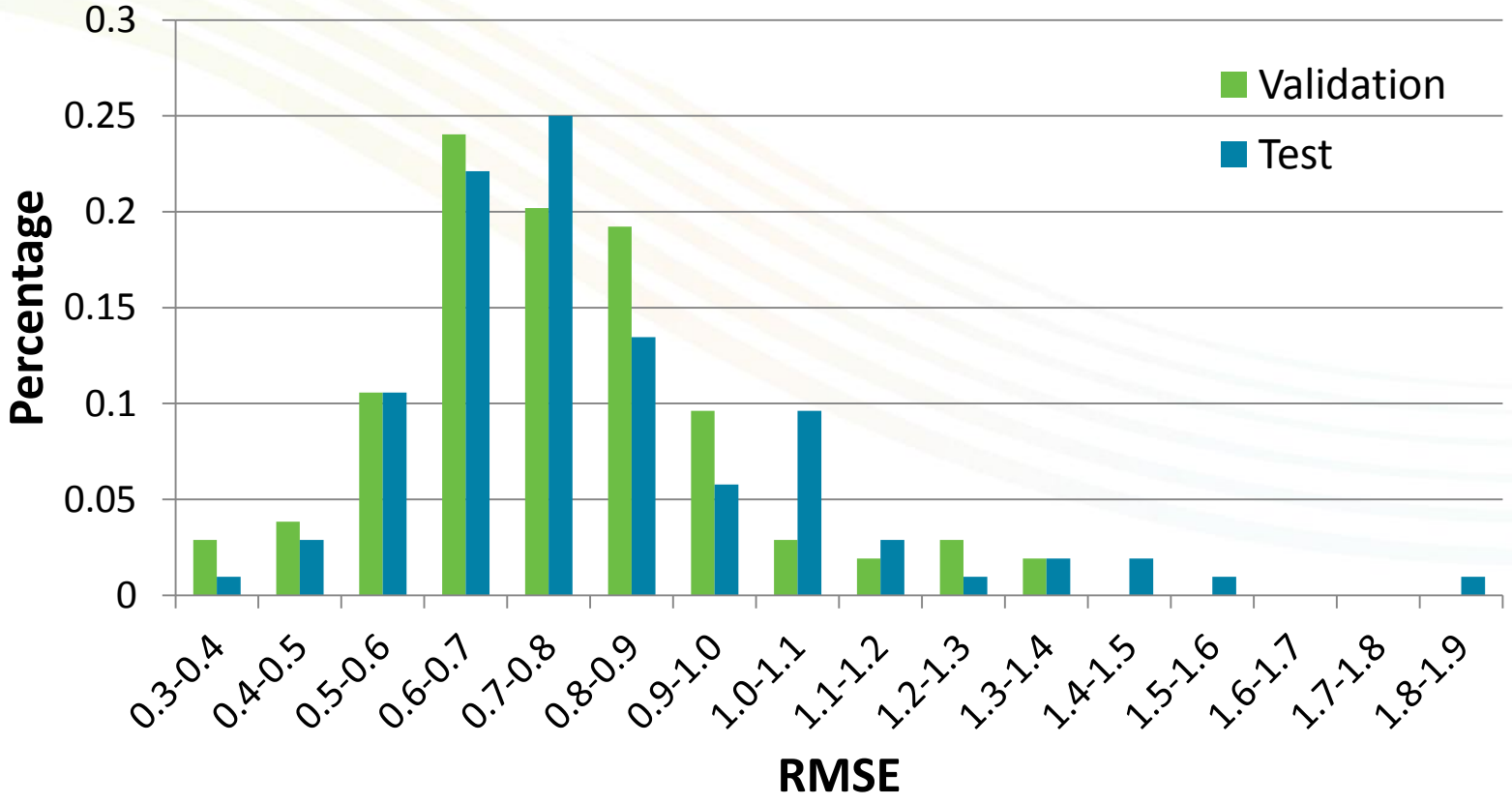
- 55% of data sets resulted in models with R^2>0.7 on validation set (50% on test set)
  - 76% with R^2 > 0.6



© 2012 Optibrium Ltd.

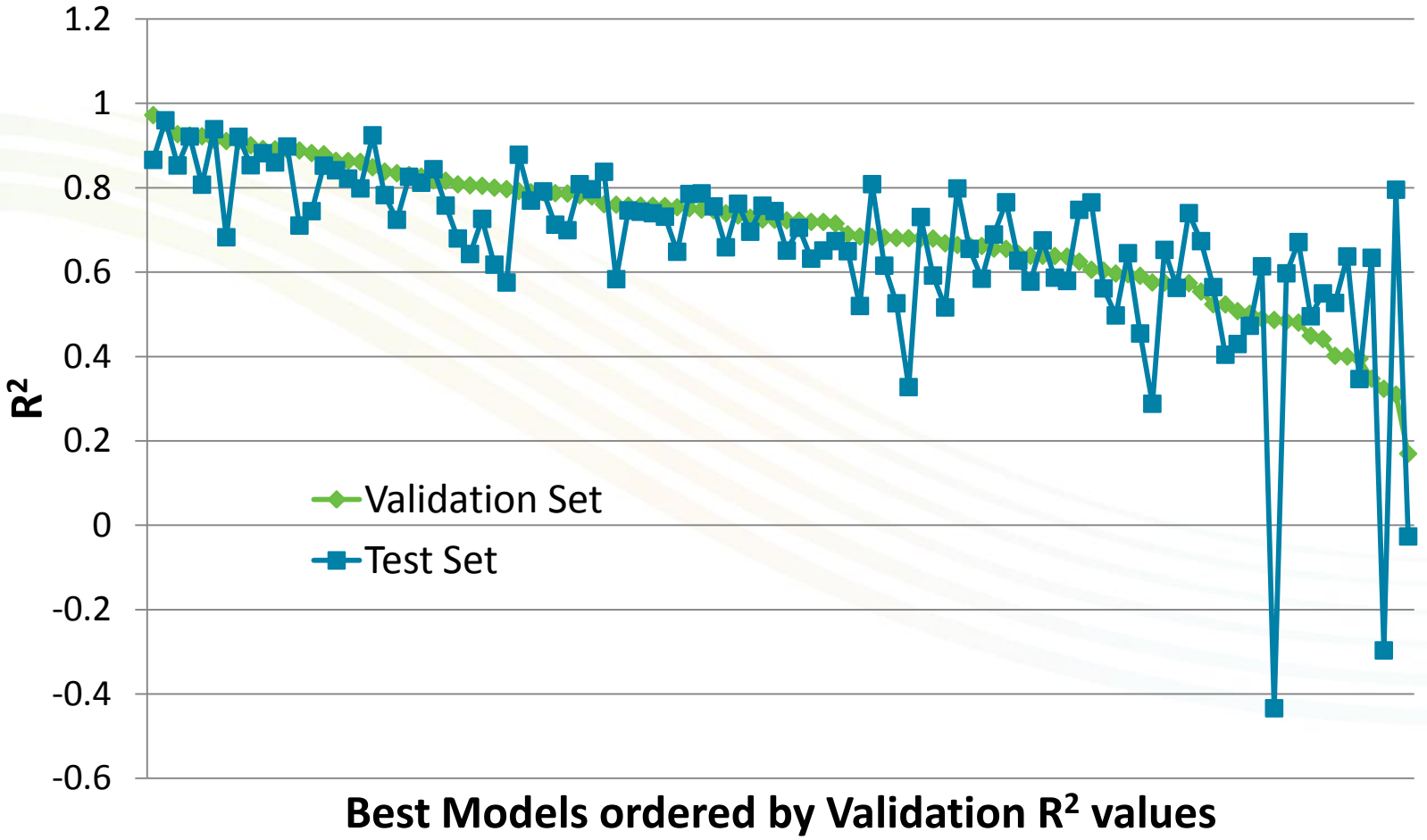
Summary of Validation Results
Root Mean Square Error

- Average RMSE on validation set = 0.76 log units (factor of 5.8)
- Average RMSE on test set = 0.8 log units (factor of 6.3)



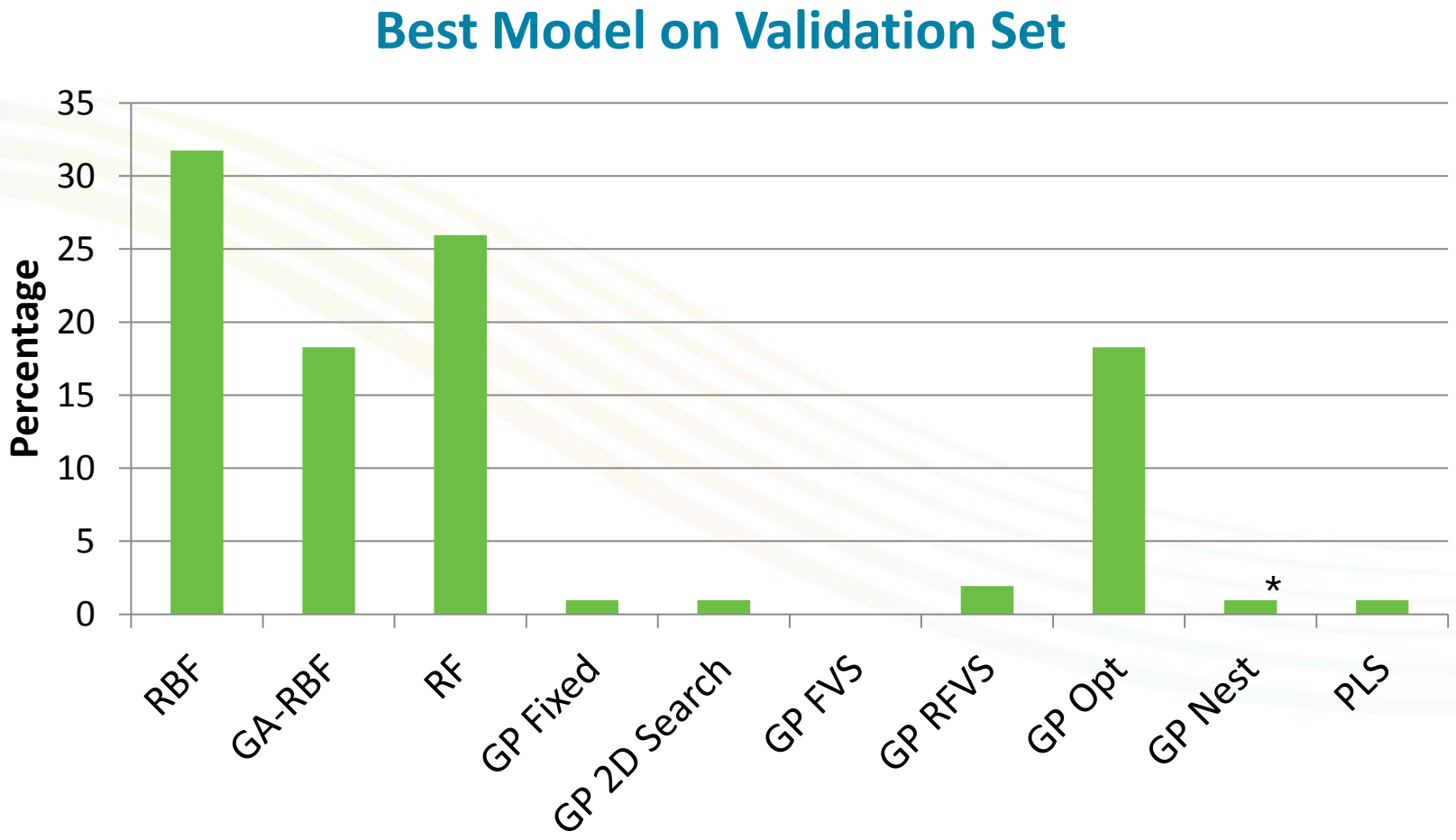
© 2012 Optibrium Ltd.

Comparison of Validation & Test Results



© 2012 Optibrium Ltd.

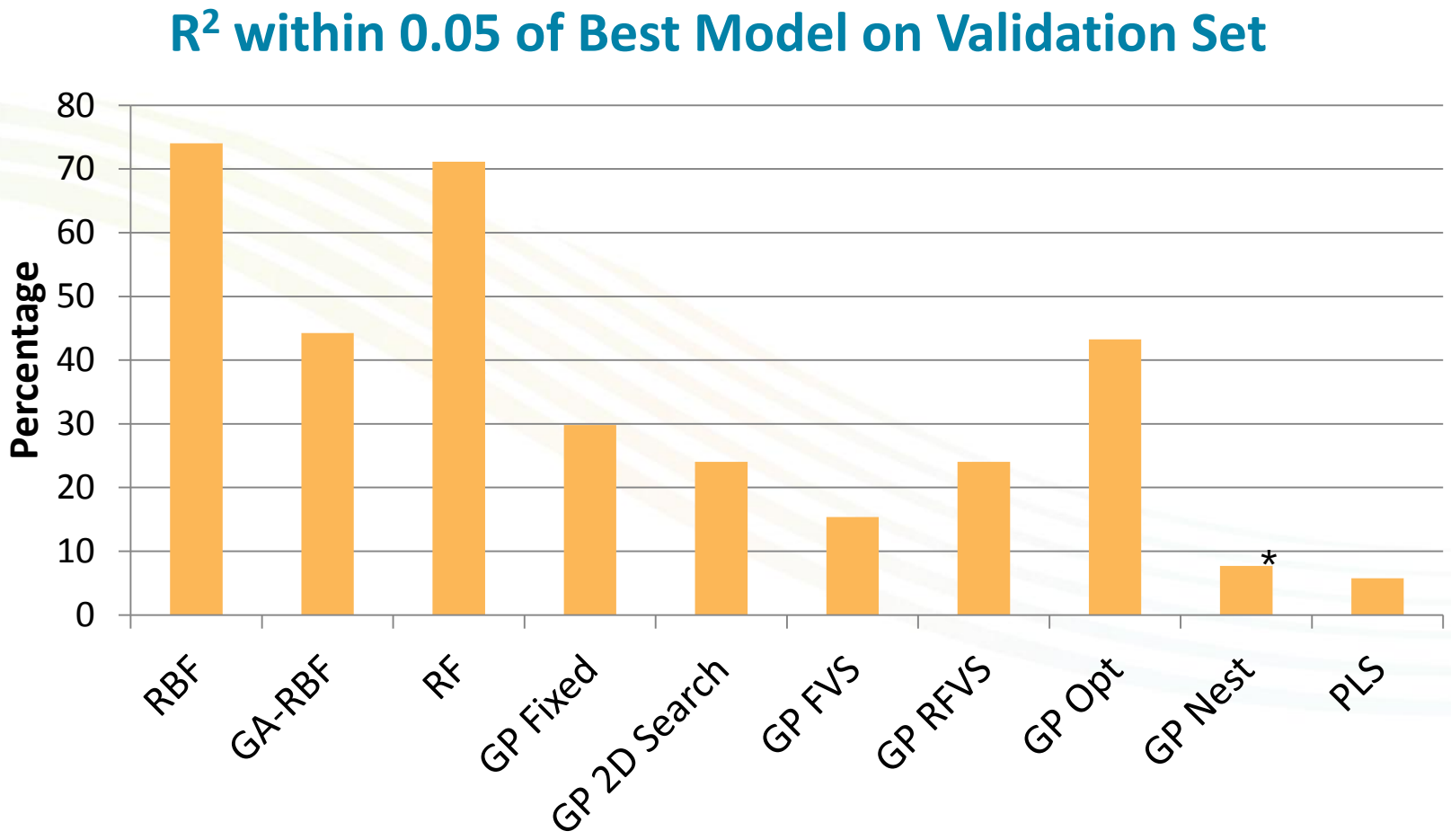
Comparison of Modelling Methods



* GP Nest only applied to 47 smallest data sets.

© 2012 Optibrium Ltd.

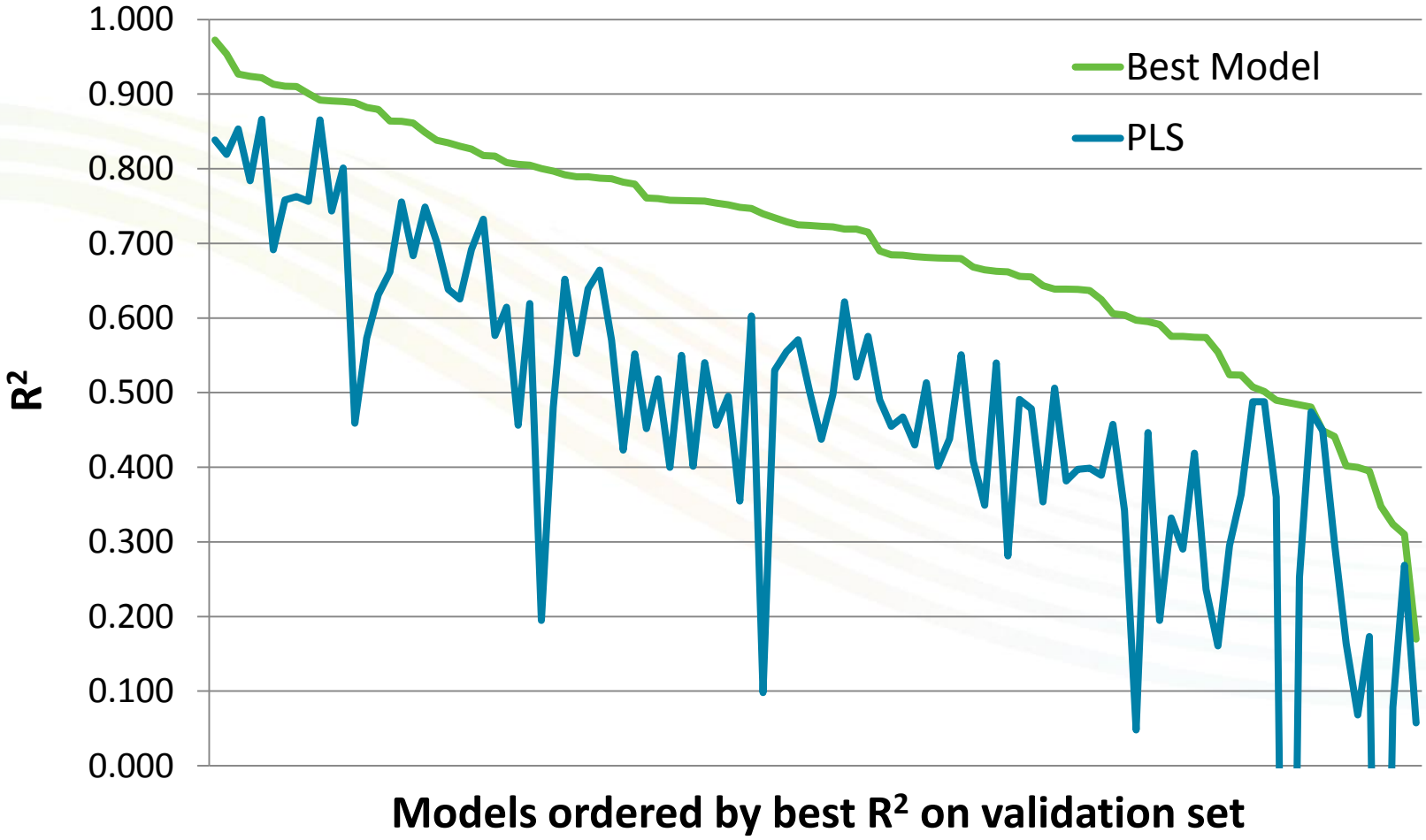
Comparison of Modelling Methods



* GP Nest only applied to 47 smallest data sets.

© 2012 Optibrium Ltd.

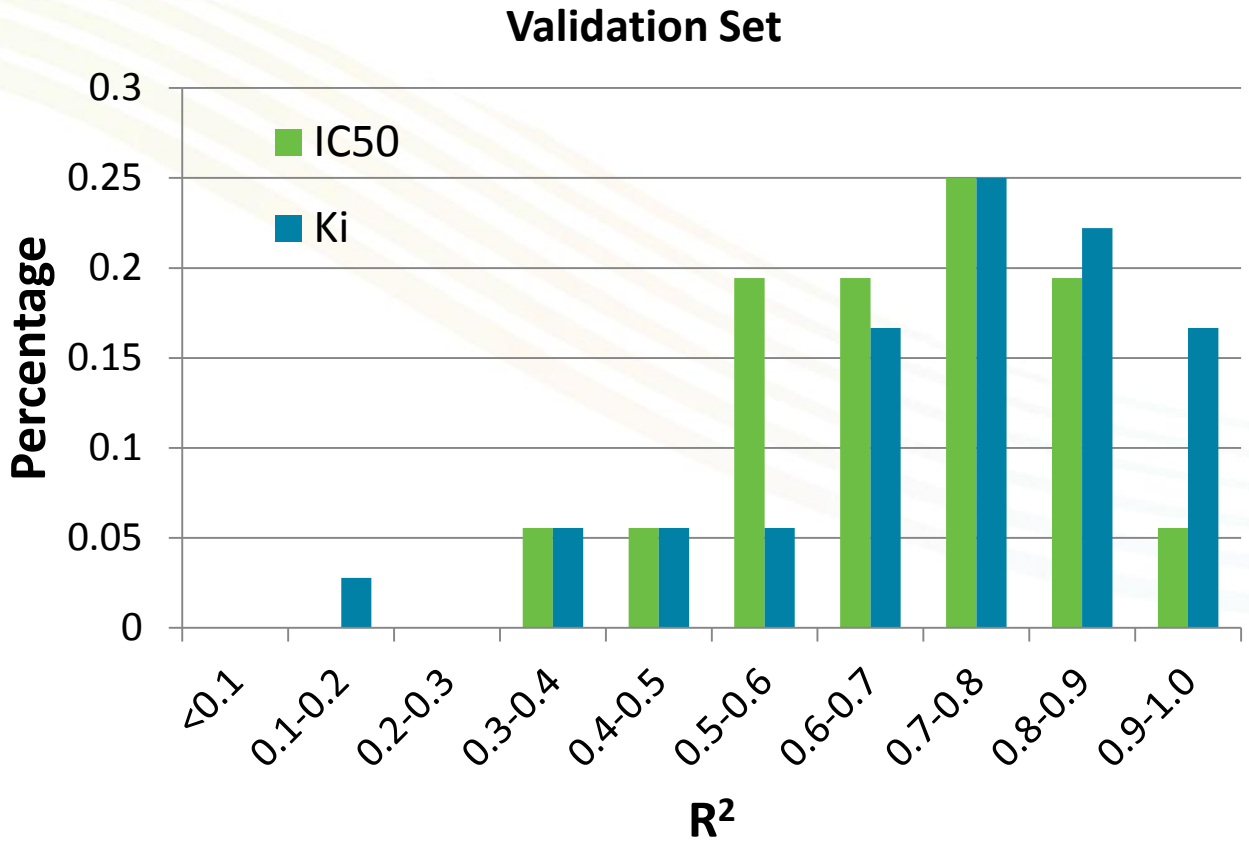
Linear vs Non-linear Models



© 2012 Optibrium Ltd.

Comparison of Data Types
IC_{50} vs K_i

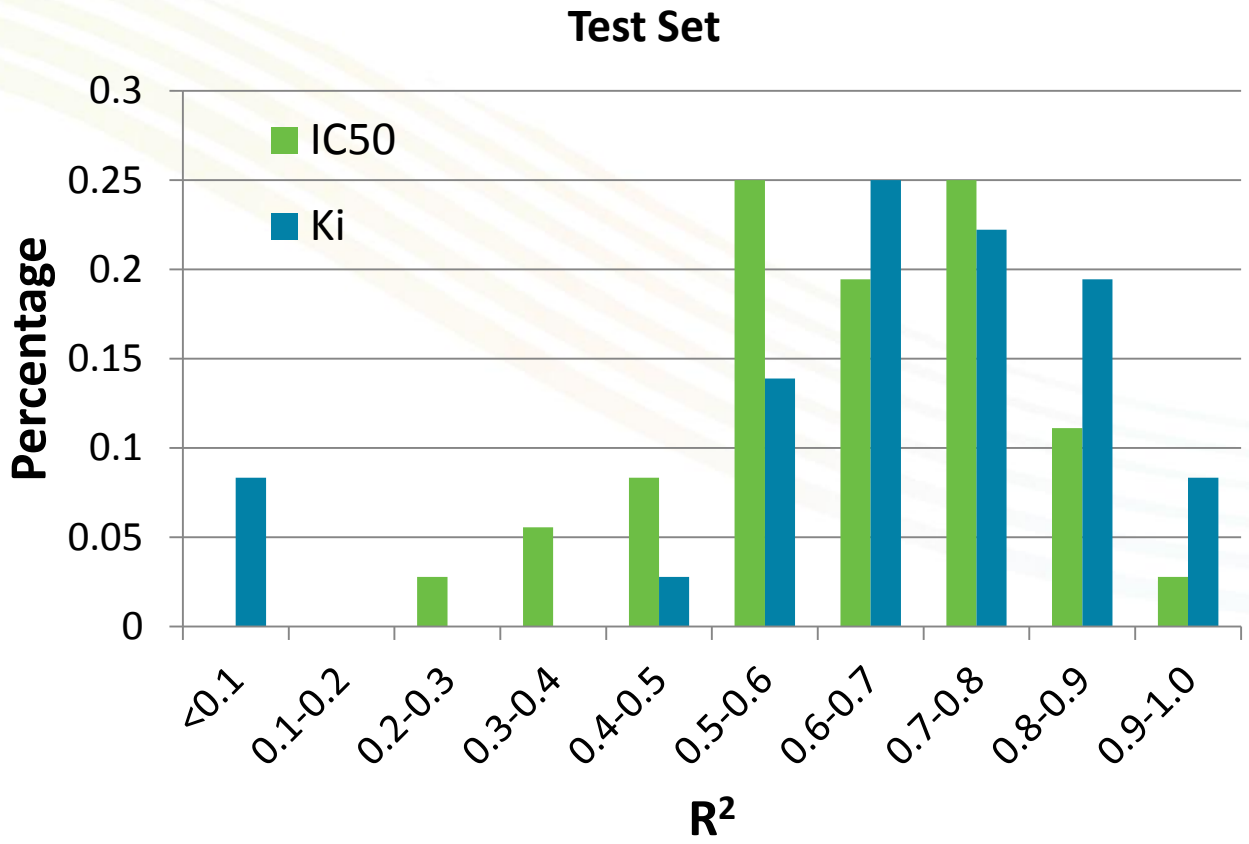
- Mean validation set R^2: IC_{50} = 0.69, K_i = 0.72



© 2012 Optibrium Ltd.

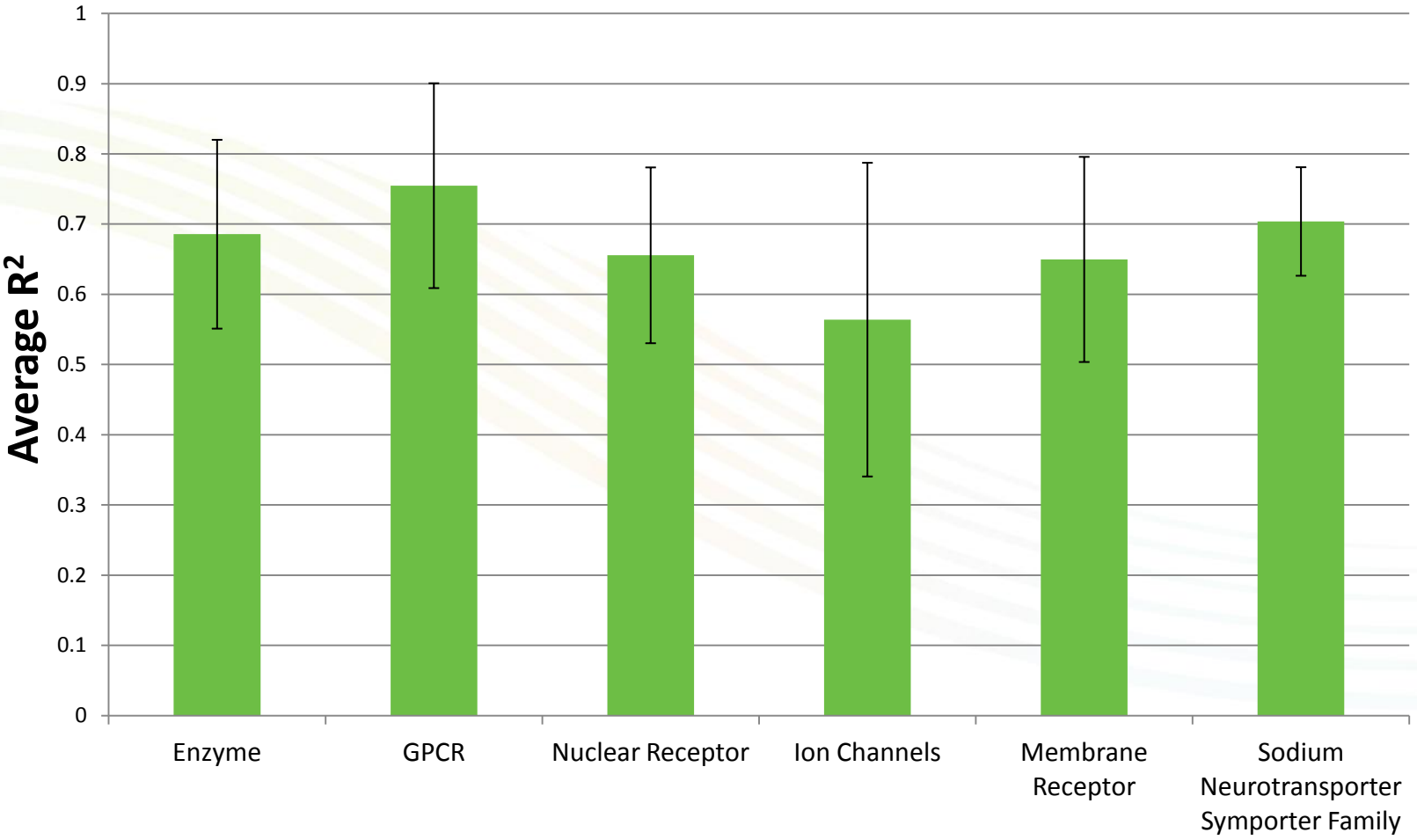
Comparison of Data Types
IC_{50} vs K_i

- Mean test set R^2: IC_{50} = 0.64, K_i = 0.65



© 2012 Optibrium Ltd.

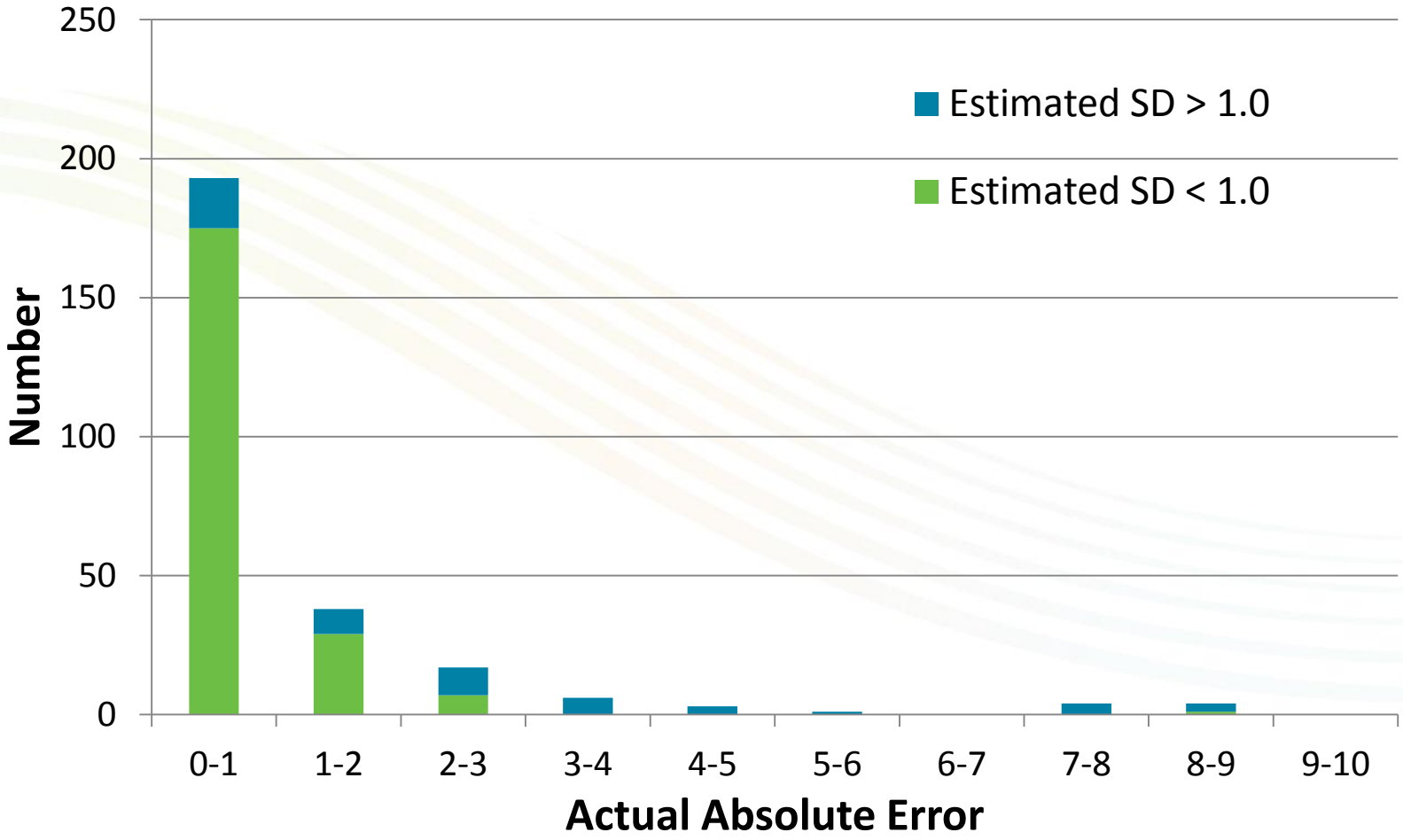
Comparison of Target Classes



© 2012 Optibrium Ltd.

Identification of Low Confidence Predictions

GP Opt Model of Alpha 1a Adrenergic Recept.



© 2012 Optibrium Ltd.

Conclusions

StarDrop™5
Software that guides you to successful drug discovery

- Built 104 2D QSAR models of target activity using StarDrop Auto-Modeller and public domain data sets
  - Consistent descriptors, set splits and validation
  - 9 methods based on 4 algorithms
- Achieved $R^2 > 0.7$ for 55% of data sets
- Linear models (PLS) consistently performed worse than non-linear
  - RBF and RF most often best
- Possible applications
  - Activity profiling – identify potential for off-target interactions
  - Compound repurposing
  - Virtual screening?
- Future work: Other descriptors including 3D
- All models and data sets will be available soon on the Optibrium on-line community
  - **www.optibrium.com/community**
  - If you would like to be notified when they are available please email me: **matt.segall@optibrium.com**

© 2012 Optibrium Ltd.