Estimating the Predictivity of Free-Wilson Models

Background to Free-Wilson analysis

The datasets used in this analysis

Predictivity analysis

- Estimating predictivity
- Results

Conclusions

Free-Wilson

- A very old approach
- What medicinal chemists instinctively try to achieve
- Assigns a contributing value to a substituent's activity

Journal of Medicinal Chemistry

VOLUME 7, NUMBER 4 JULY 6, 1964

A Mathematical Contribution to Structure–Activity Studies

SPENCER M. FREE, JR., AND JAMES W. WILSON

Research and Development Division, Smith Kline and French Laboratories, Philadelphia, Pennsylvania

Received February 4, 1964

A mathematical technique is suggested as a means of describing structure–activity relationships of a series of chemical analogs. The data requirements included specific side chain arrangements and performance characteristics of all analogs tested. Two examples illustrate the use of the additive mathematical model where the performance characteristics are measures of biological activity. The results rank the structural changes per position by estimating the amount of biological response attributed to each change. The estimates are both positive and negative. Several uses for the mathematical solution are suggested.

410 citations in SciFinder
Compare with CoMFA paper[1]: 2,673 citations

1 Cramer, Richard D., III; Patterson, David E.; Bunce, Jeffrey D. *Journal of the American Chemical Society* (1988) 110(18) 5959

Free-Wilson Analysis

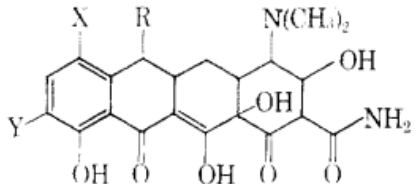


TABLE I

BIOLOGICAL ACTIVITY OF TEN TETRACYCLINES

Compound	Compound identification: R: H	R: CH_3	X: NO_2	X: Cl	X: Br	Y: NO_2	Y: NH_2	Y: CH_3CONH	Biological activity
III	1		1			1			60
IV	1			1		1			21
V	1				1	1			15
VI	1			1			1		525
VII	1				1		1		320
VIII	1		1				1		275
IX		1	1				1		160
X		1	1					1	15
XI		1			1		1		140
XII		1			1			1	75

Following the rules set out (under "Models"), one writes a series of 10 equations in 6 unknowns. (There are really 9 unknowns which reduce to 6 because the contributions at each position sum to zero.) The results are presented in Table II.

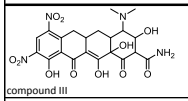
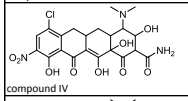
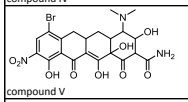
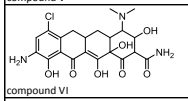
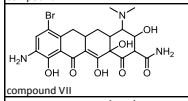
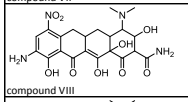
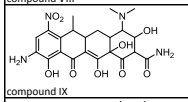
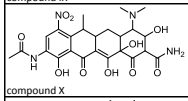
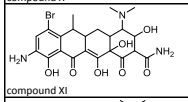
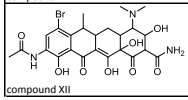
TABLE II

CONTRIBUTION OF STRUCTURAL CHANGES[a]

R		X		Y	
Side chain positions					
a[H]	75	b[Cl]	84	c[NH_2]	123
a[CH_3)	−112	b[Br]	−16	c[CH_3CONH—]	18
		b[NO_2]	−26	c[NO_2]	−218

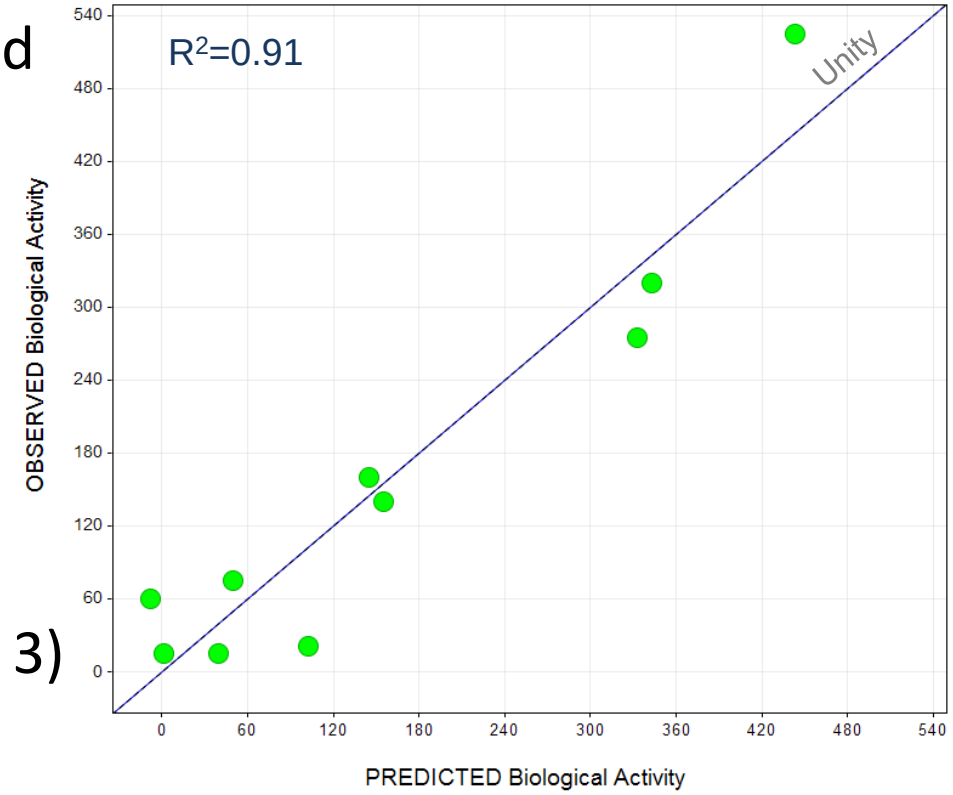
[a] The solution includes these restrictions: $6a[H] + 4a[CH_3] = 0$; $2b[Cl] + 4b[Br] + 4b[NO_2] = 0$; and $5c[NH_3] + 2c[CH_3CONH—] + 3c[NO_2] = 0$. The over-all average was 161.

Reformulate as a table for multi-linear regression

	Y-Block	X-Block		
Structure	Biological Activity	R SMILES	X SMILES	Y SMILES
 compound III	60	*[H]	[*][N+](=O)[O-]	[*][N+](=O)[O-]
 compound IV	21	*[H]	[*]Cl	[*][N+](=O)[O-]
 compound V	15	*[H]	[*]Br	[*][N+](=O)[O-]
 compound VI	525	*[H]	[*]Cl	[*]N
 compound VII	320	*[H]	[*]Br	[*]N
 compound VIII	275	*[H]	[*][N+](=O)[O-]	[*]N
 compound IX	160	*C	[*][N+](=O)[O-]	[*]N
 compound X	15	*C	[*][N+](=O)[O-]	[*]NC(=O)C
 compound XI	140	*C	[*]Br	[*]N
 compound XII	75	*C	[*]Br	[*]NC(=O)C

Model Predicted vs Observed

- Generates an (over-fitted) multi-linear regression model
- 8 coefficients are optimised to generate the model, one for each substituent
  - Variation in R = 2 (H, Me)
  - Variation in X = 3 (NO_2, Cl, Br)
  - Variation in Y = 3 (NO_2, NH_2, NHCOMe)
- 8 degrees of freedom (not 3)
  - 10 observations
- How to validate and assess predictivity?



Comments about Free-Wilson

Assumes additivity

- Not always valid, but identifies opportunities to exploit
- When additivity is valid, should also be within the domain of applicability

Medicinal chemists naturally identify with Free-Wilson, and usually have a good idea what the best groups are for a particular endpoint

- More difficult with multiple endpoints and thousands of possibilities and Free-Wilson can identify best combinations for multi-objective optimisation

Synthetic accessibility is often solved

- Constituents of new compounds have already been synthesised

Squeezes the most out of the existing data

- Predictions restricted to existing core and R-groups

Data overview

- Project handovers
- Particular groups might not be represented in certain assays; can select for screening

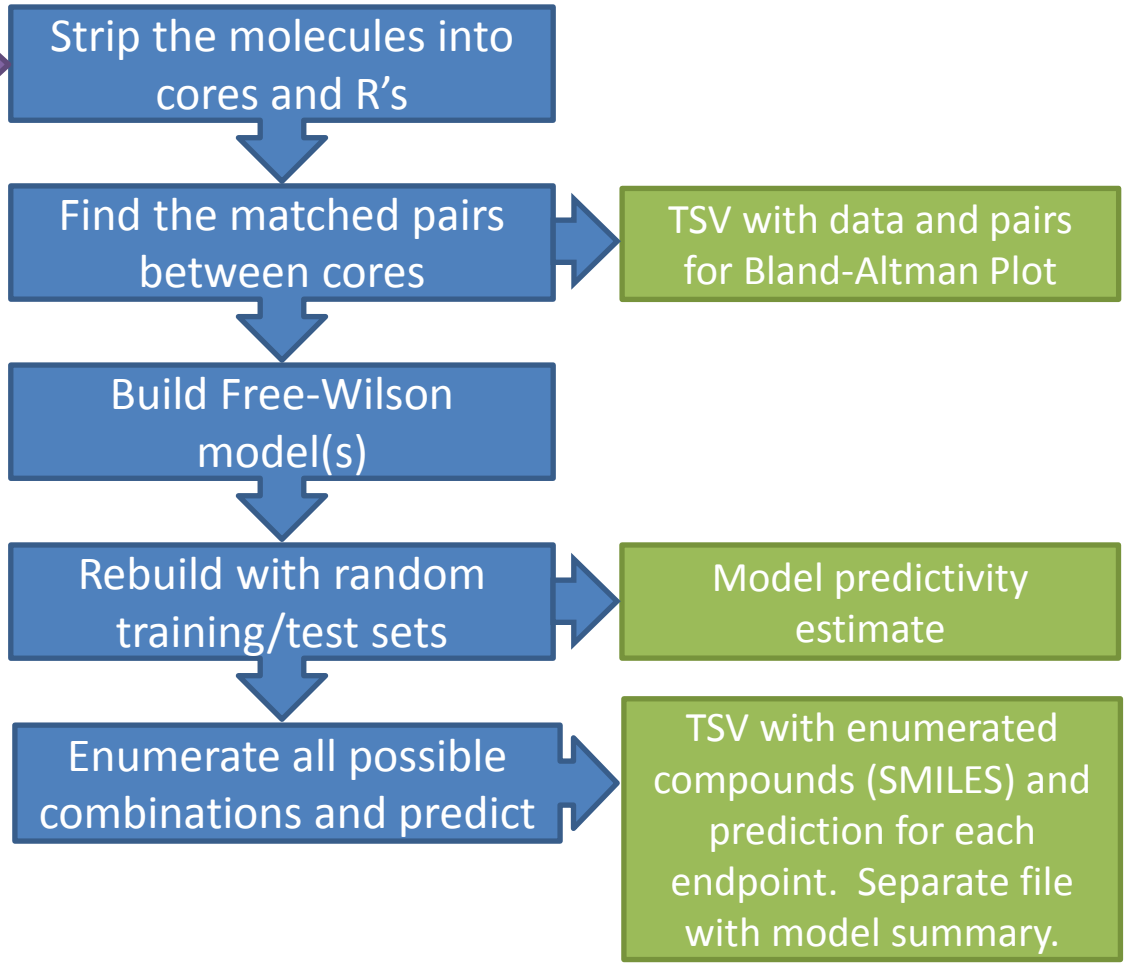
Overview of code ("FWenum")

1. SDF of compounds and data
2. SDF of core and R definitions
3. Data tag(s) to build model
4. Data tag for molecule name

```
> FWenum –i compoundsWithData.sdf \
         -r coreAndRDefinitions.sdf \
         -d "SDFTag for data identifier" \
        [-d "SDFTag another identifier" \]
         -n "SDFTag for name of compounds"
```

Requires:

- OEChem toolkit
- python 2.7
- R v3.1+
- rpy2



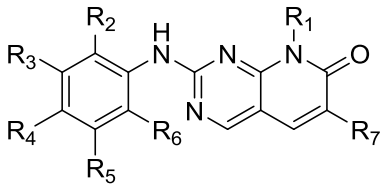
Datasets

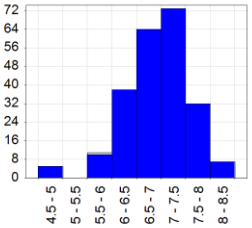
- Two sources of data
  - 11 datasets from Chen paper[2]
  - ChEMBL (v21) EGFR1 and EGFR2 data set
    - ‘Epidermal growth factor receptor erbB1’ (target id CHEMBL203)
      - 484 assays with 5 or more data points
    - ‘Receptor protein-tyrosine kinase erbB-2’ (target id CHEMBL1824)
      - 141 assays with 5 or more data points
    - Aggregate all µM and nM data (taking geometric mean if multiple assay values for each compound
      - 1,084 compounds with measurements in both EGFR1 and EGFR2
    - Good example set of data
      - Similar size to late lead optimisation
      - Flawed as a genuine EGFR dataset

2 Chen, H.; Carlsson, L.; Eriksson, M.; Varkonyi, P.; Norinder, U.; Nilsson, I.
“Beyond the Scope of Free-Wilson Analysis: Building Interpretable QSAR Models with Machine Learning Algorithms”
Journal of Chemical Information and Modeling (2013) 53(6) 1324-1336

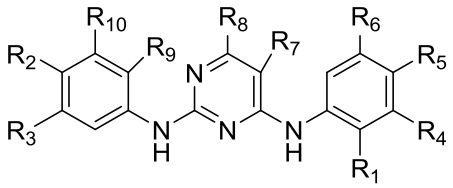
Chen Dataset: Cores and Potency Bins

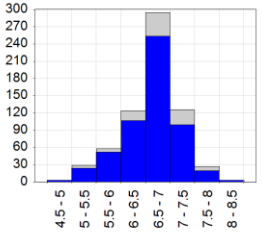
CDK5 (229)





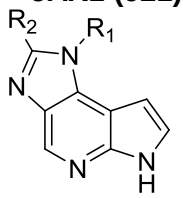
IL4 (563)

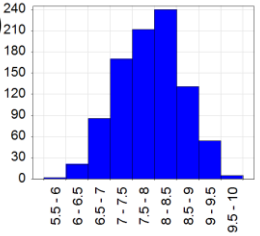




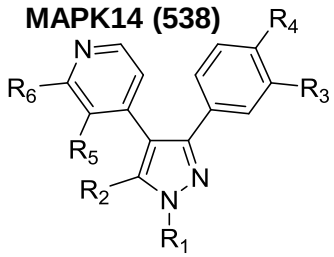
Potency distribution

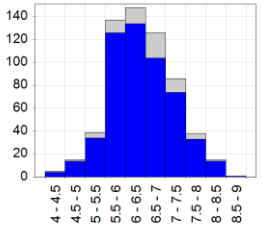
JAK1 (921)



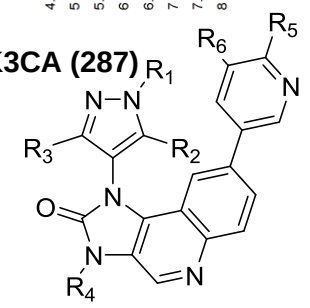


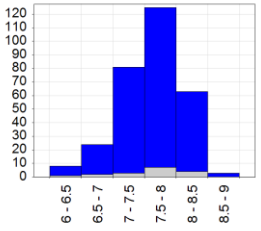
MAPK14 (538)



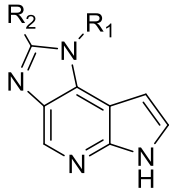


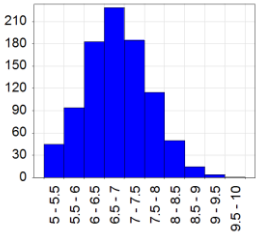
PIK3CA (287)



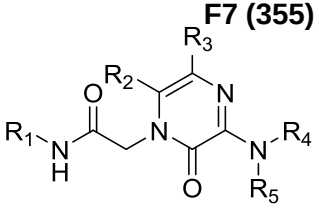


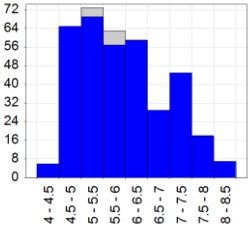
TYK2 (921)



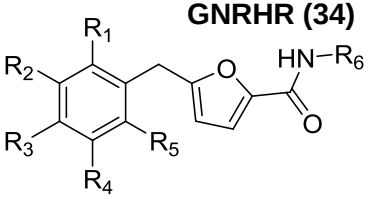


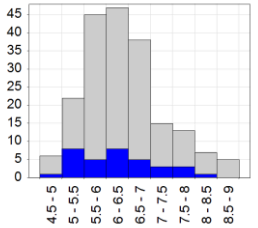
F7 (355)



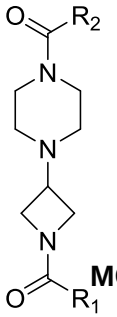


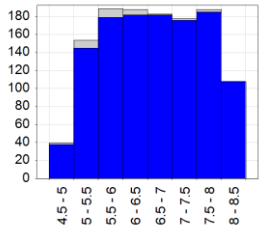
GNRHR (34)



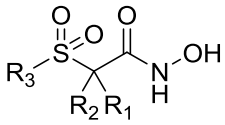


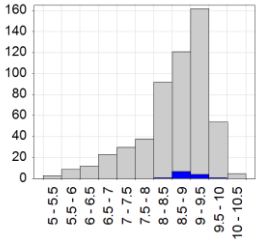
MGLL (1195)



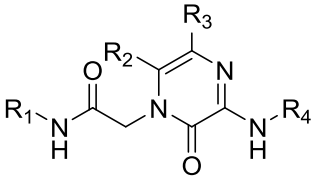


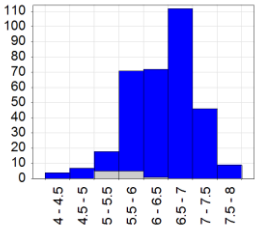
MMP2 (13)





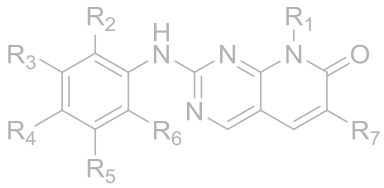
PRSS2 (328)

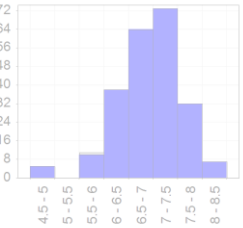




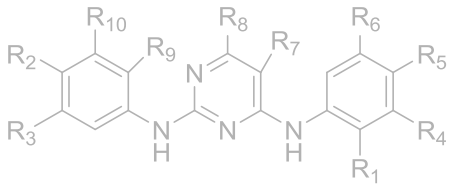
Chen Dataset: Cores and Potency Bins

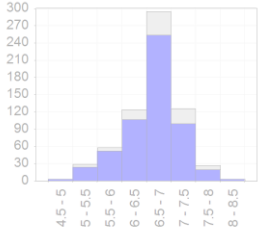
CDK5 (229)



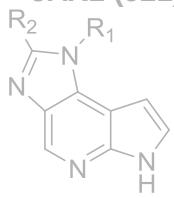


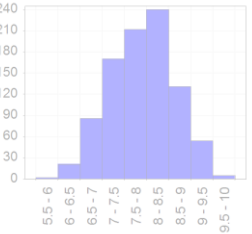
IL4 (563)



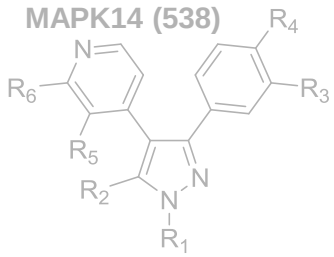


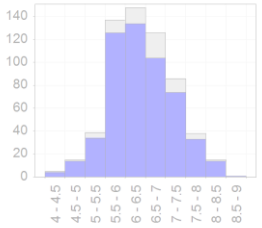
JAK1 (921)



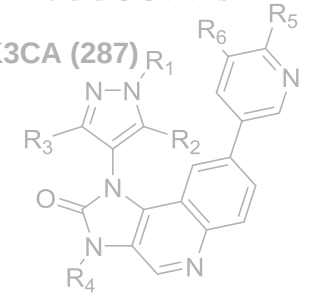


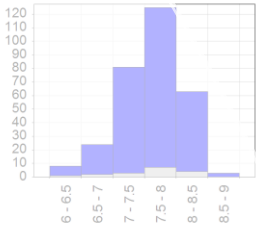
MAPK14 (538)



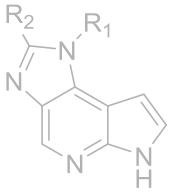


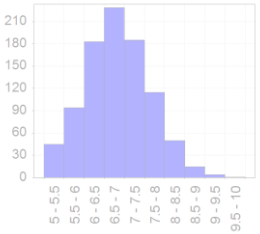
PIK3CA (287)



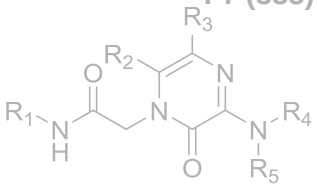


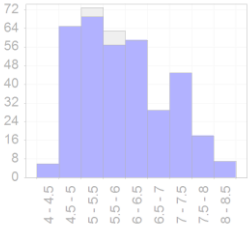
TYK2 (921)



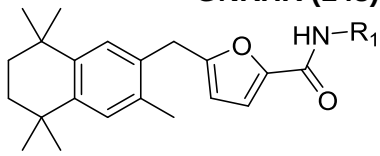


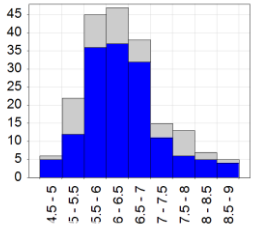
F7 (355)

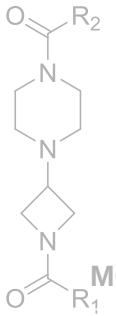




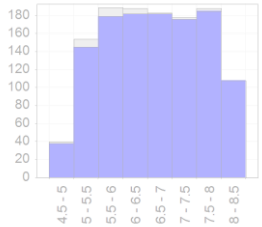
GNRHR (148)



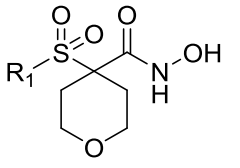


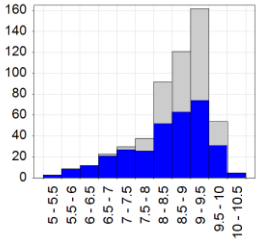


MGLL (1195)

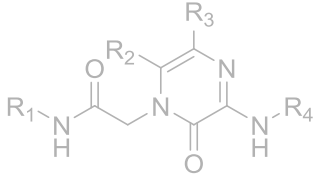


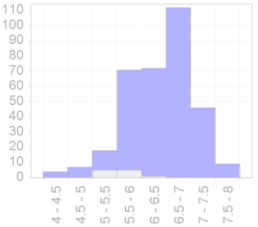
MMP2 (323)



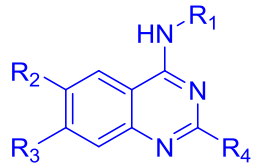


PRSS2 (328)

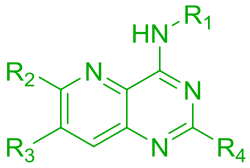




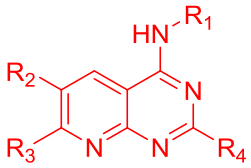
EGFR Dataset: Cores and Potency Bins



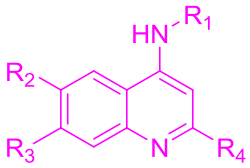
quinazolin-4-amine



pyrido[3,2-d]pyrimidin-4-amine

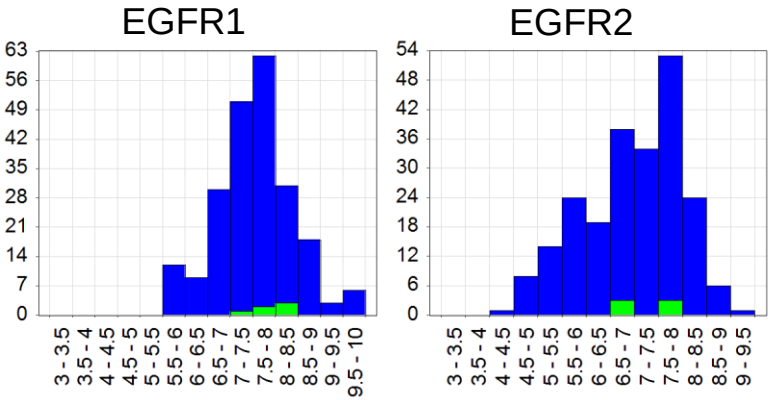


pyrido[2,3-d]pyrimidin-4-amine



quinolin-4-amine

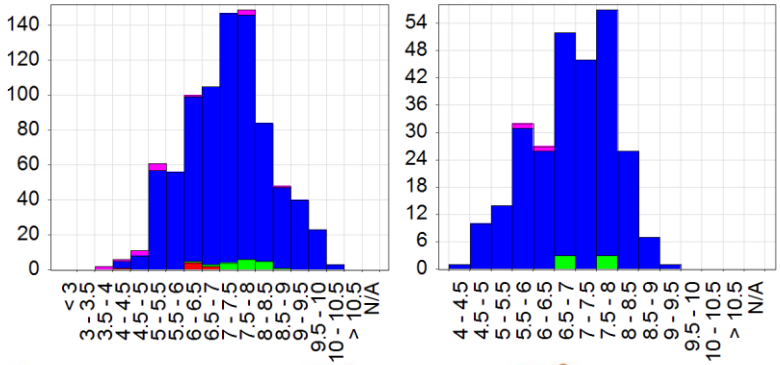
Combined datasets (each compound has an EGFR1 and EGFR2 measurement)



quinazolin-4-amine: EGFR1 (216), EGFR2 (216)

pyrido[3,2-d]pyrimidin-4-amine: EGFR1 (6), EGFR2 (6)

Individual datasets (each compound does not have both an EGFR1 and EGFR2 measurement)



quinazolin-4-amine: EGFR1 (795), EGFR2 (256)

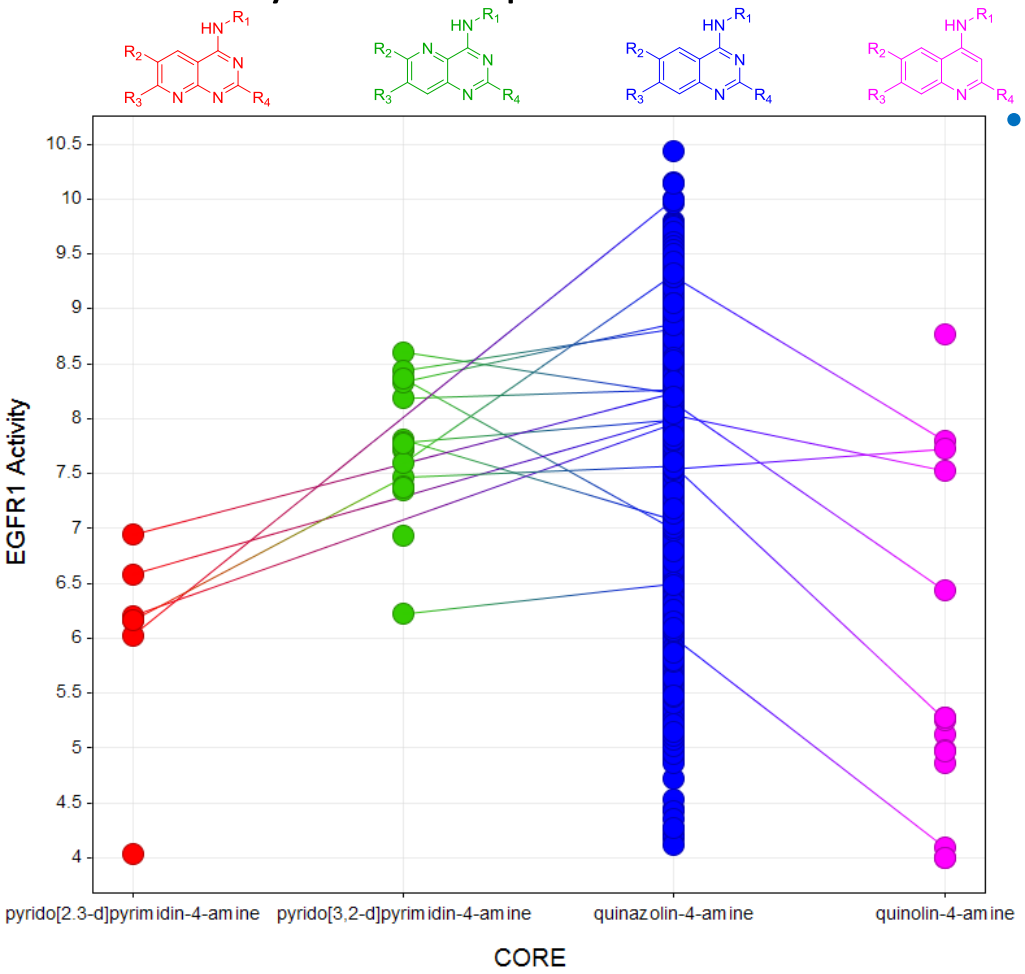
pyrido[3,2-d]pyrimidin-4-amine: EGFR1 (18), EGFR2 (6)

pyrido[2,3-d]pyrimidin-4-amine: EGFR1 (7),

quinolin-4-amine: EGFR1 (15), EGFR2 (2)

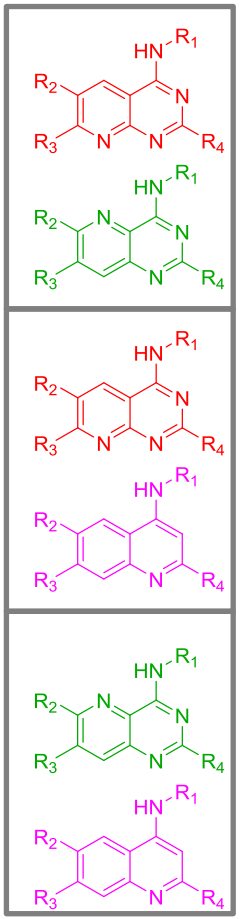
Matched Pairs

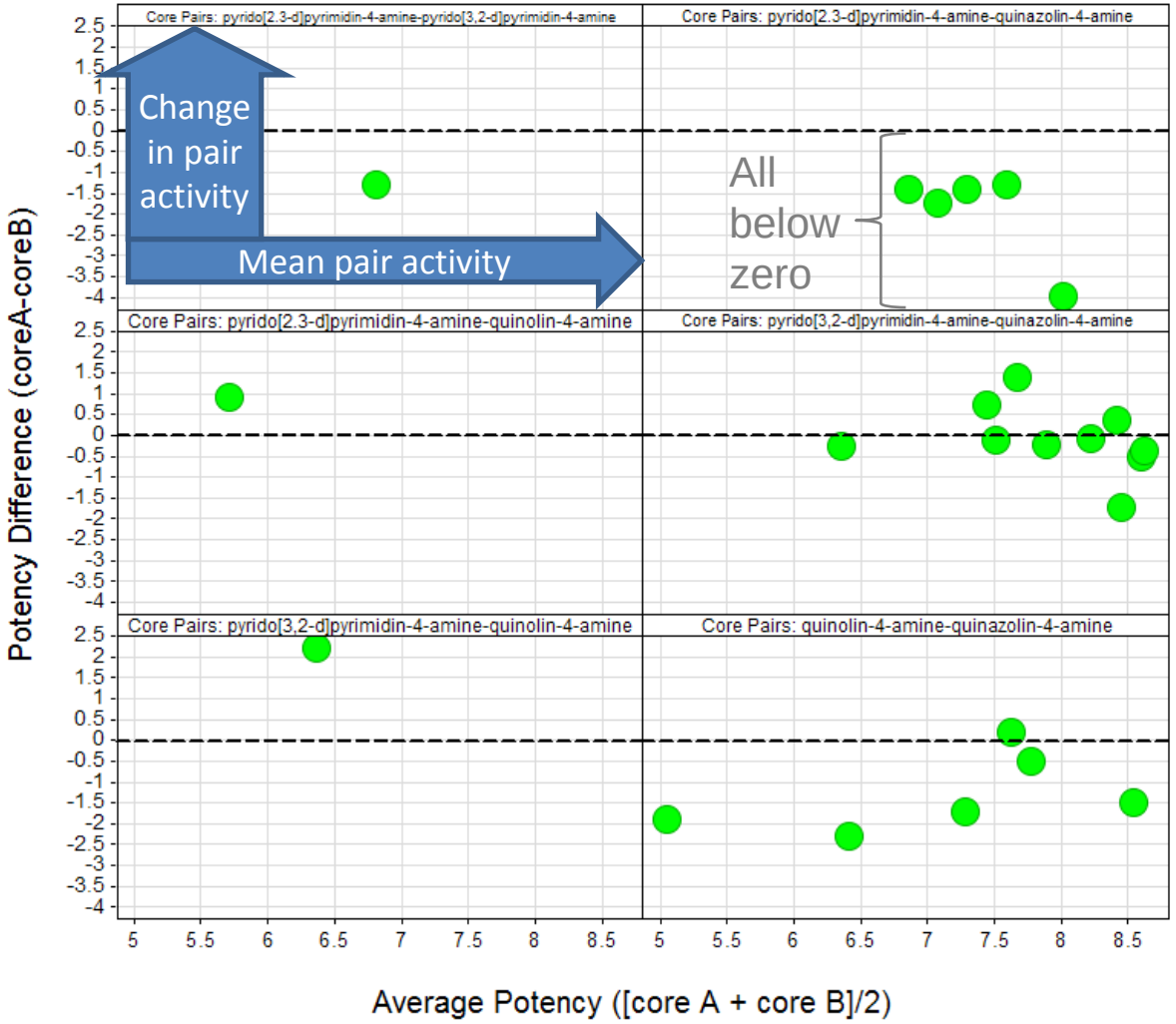
- Identify matched pairs between cores

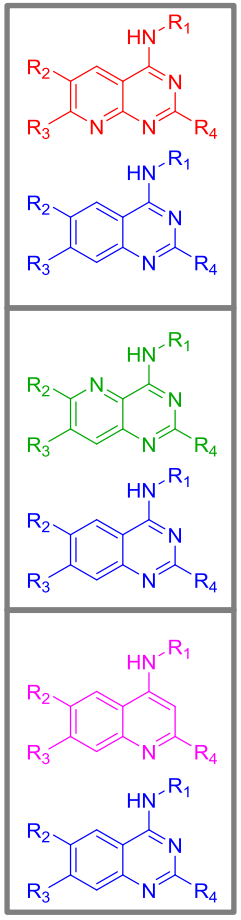


- Can identify which compounds might have been missed in the best core

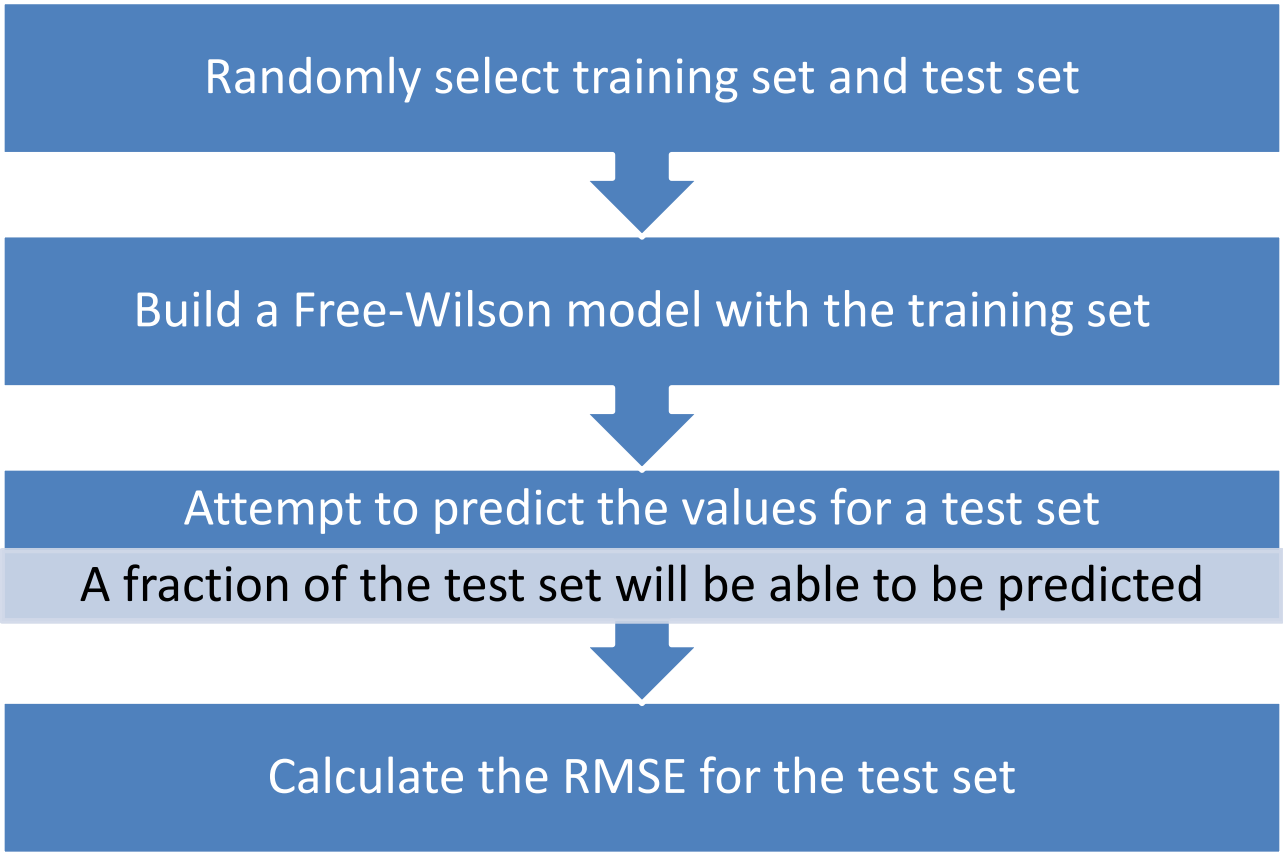
Bland-Altman Plot





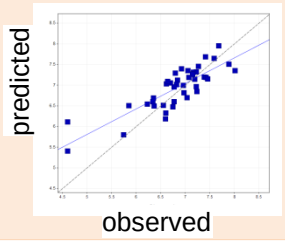
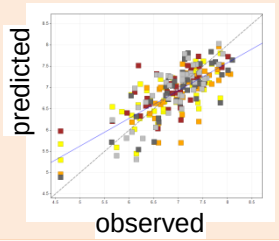
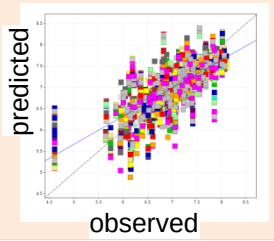
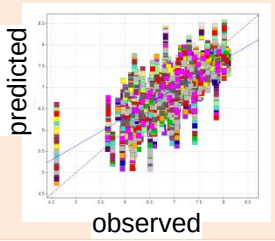
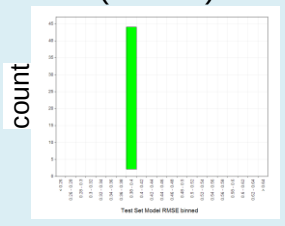
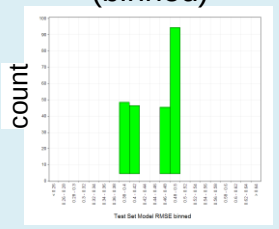
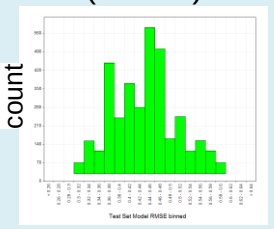
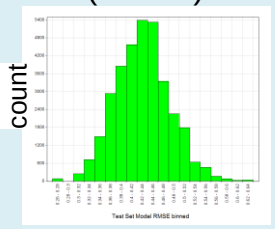


Assessing the Model Predictivity

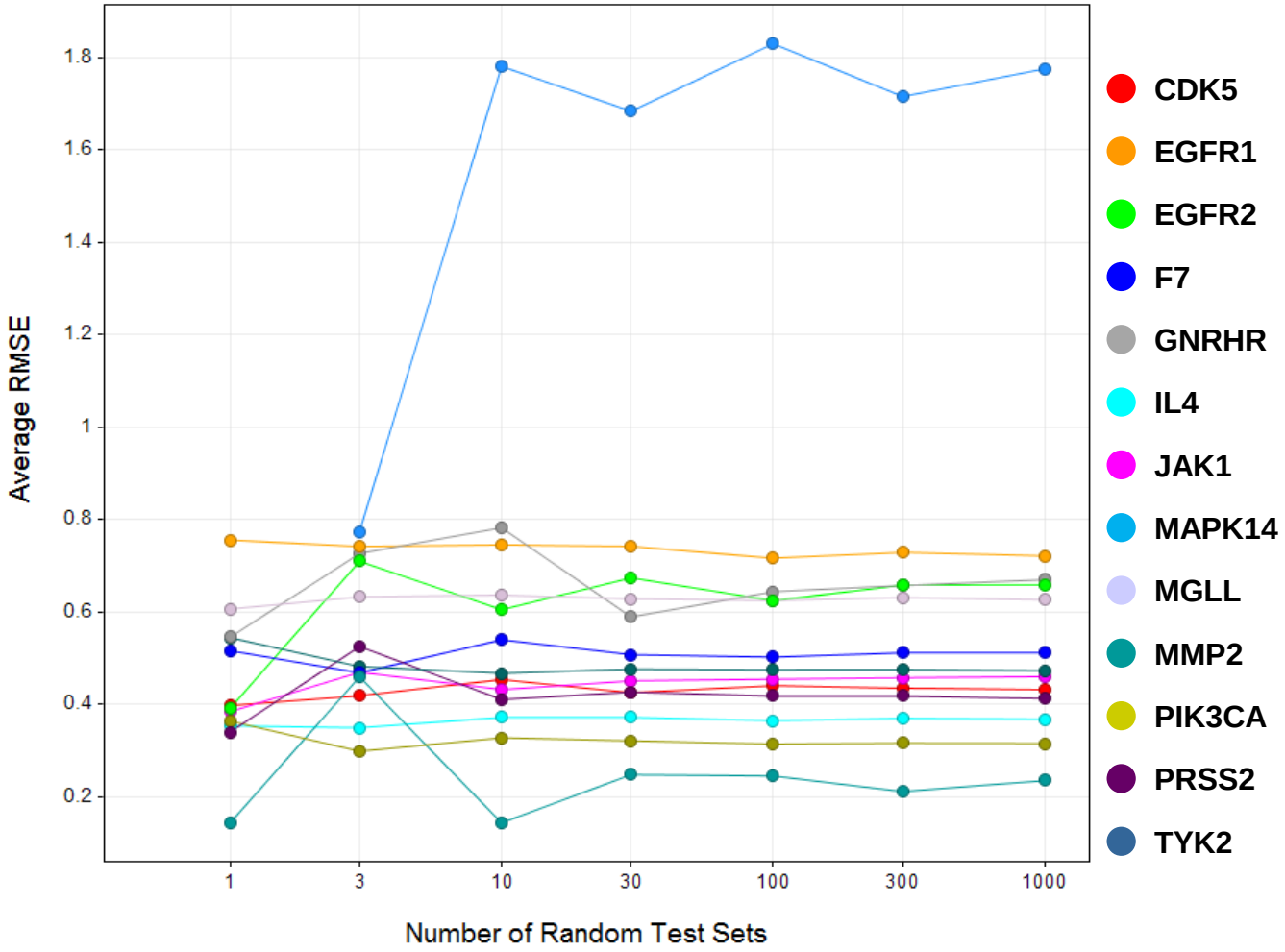


- How many times to repeat test set prediction for a predictivity estimate?
- What percentage training/test set split to use?

CDK5 Test Set RMSE's by Iteration

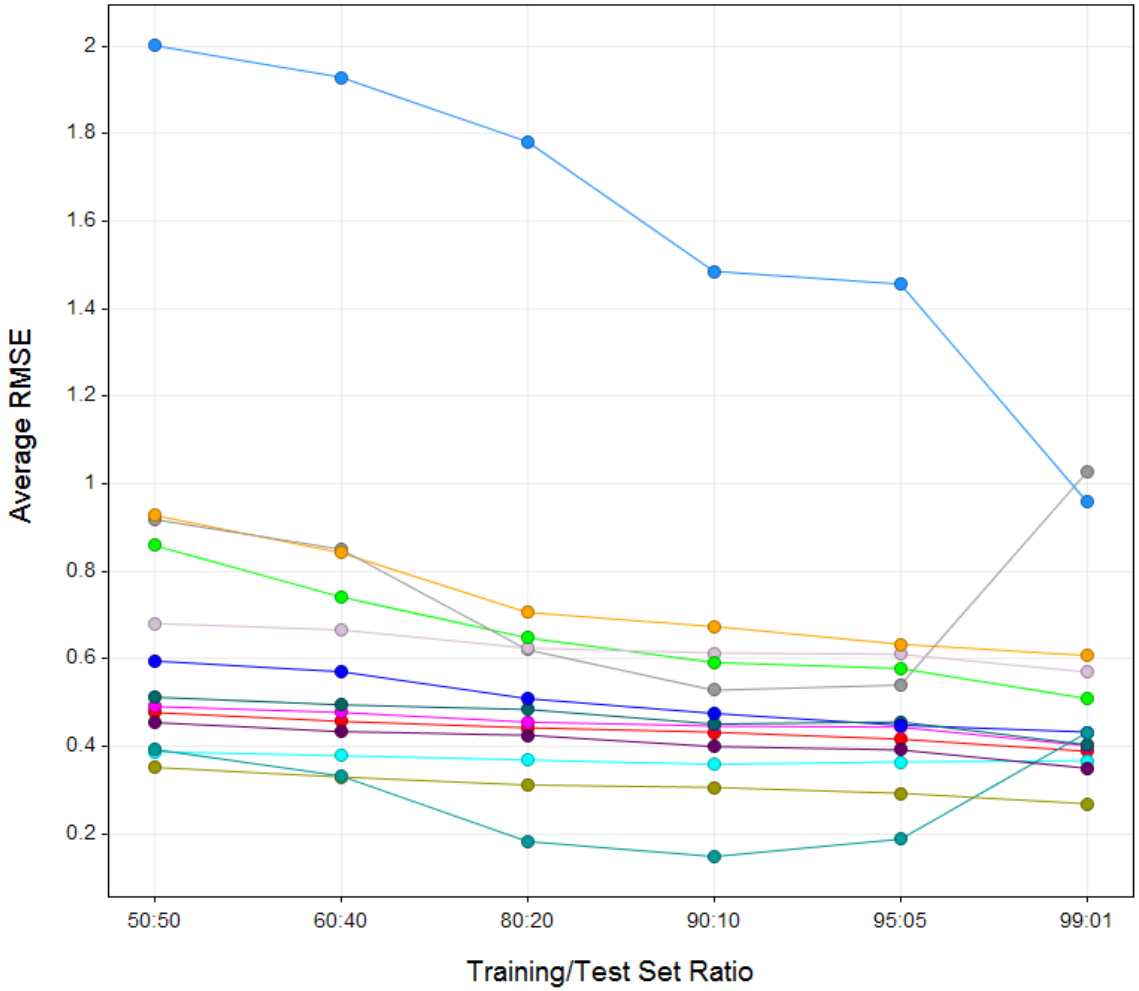
1 random 20% test set	**10 random** 20% test sets	**100 random** 20% test sets	**1000 random** 20% test sets
			
Test Set RMSE (binned) 	Test Set RMSE (binned) 	Test Set RMSE (binned) 	Test Set RMSE (binned) 
Average RMSE Average **0.397**	Average RMSE Average **0.453**	Average RMSE Average **0.440**	Average RMSE Average **0.431**
%age of Test Set Used **22.8%**	%age of Test Set Used **23.6%**	%age of Test Set Used **23.2%**	%age of Test Set Used **23.2%**

All Test Set RMSE's by Iteration



- RMSE's stabilize after 100 iterations

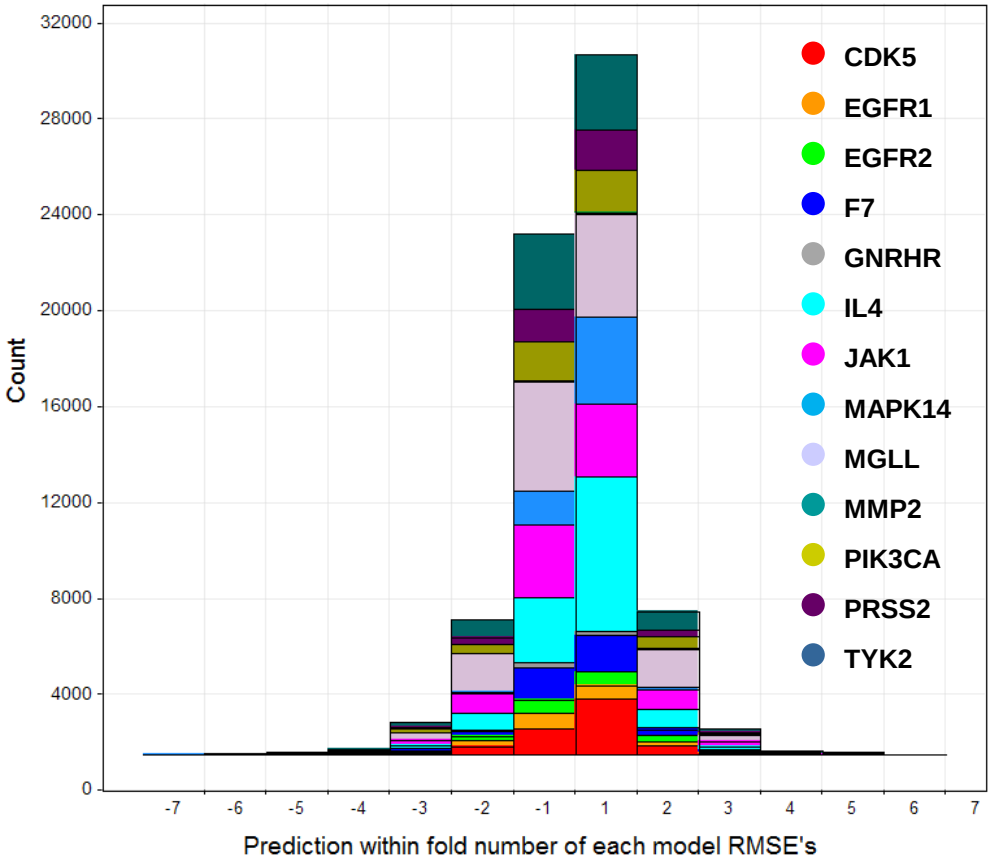
Test Set RMSE by Train/Test Ratio



- 100 iterations
- Test set RMSE's reduce as training set proportion increases, but stability decreases at 99:01
- 80:20 split ok. Possibly a slight over estimate of model RMSE

Estimating Predictivity

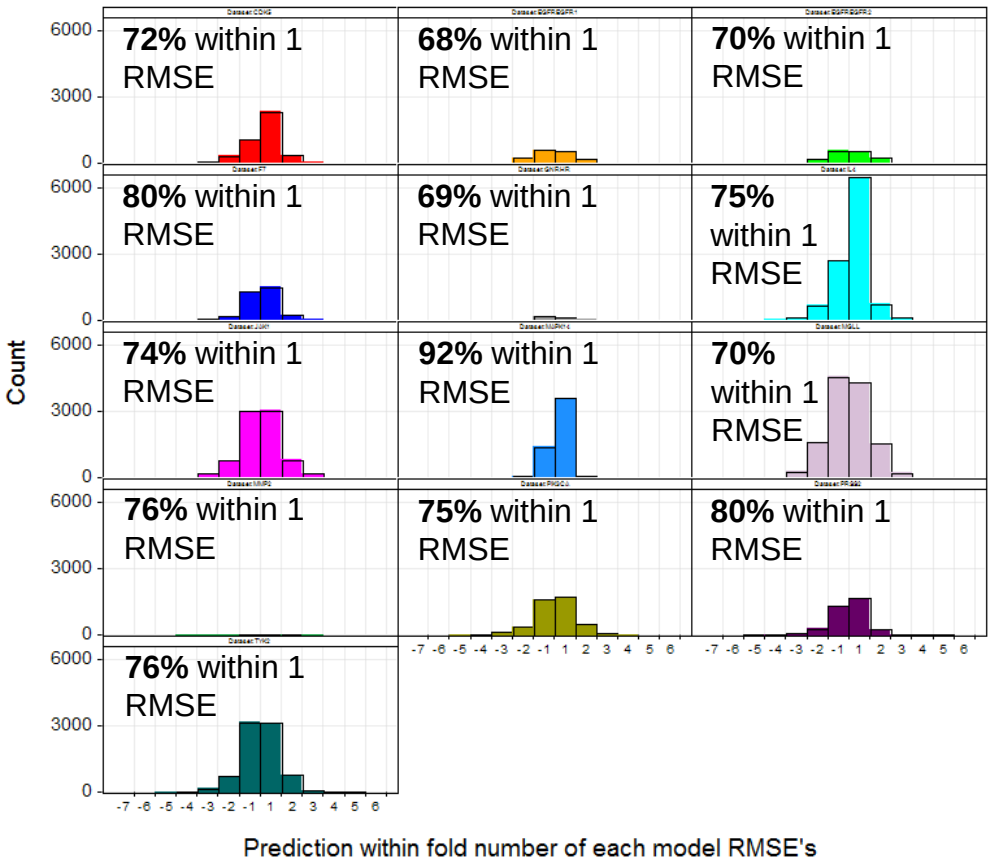
- Use average test set RMSE from 100 random 80:20 split predictions as a guide to overall model predictivity



- Categorise the predictions from the 100 random 20% test sets within a fold value of the average RMSE for each dataset model
  - $int\left(\frac{observed-predicted}{average\ model\ RMSE}\right) \pm 1$
- Shows the distribution of over predictions and under predictions is symmetrical
- 74.7% of predictions are within 1 model RMSE, 94.6% within 2 RMSE values

Estimating Predictivity

- Use average test set RMSE from 100 random 80:20 split predictions as a guide to overall model predictivity



- CDK5
- EGFR1
- EGFR2
- F7
- GNRHR
- IL4
- JAK1
- MAPK14
- MGLL
- MMP2
- PIK3CA
- PRSS2
- TYK2

- For all datasets, at least two thirds of all predictions are within 1 RMSE

Combining Coefficient Standard Errors

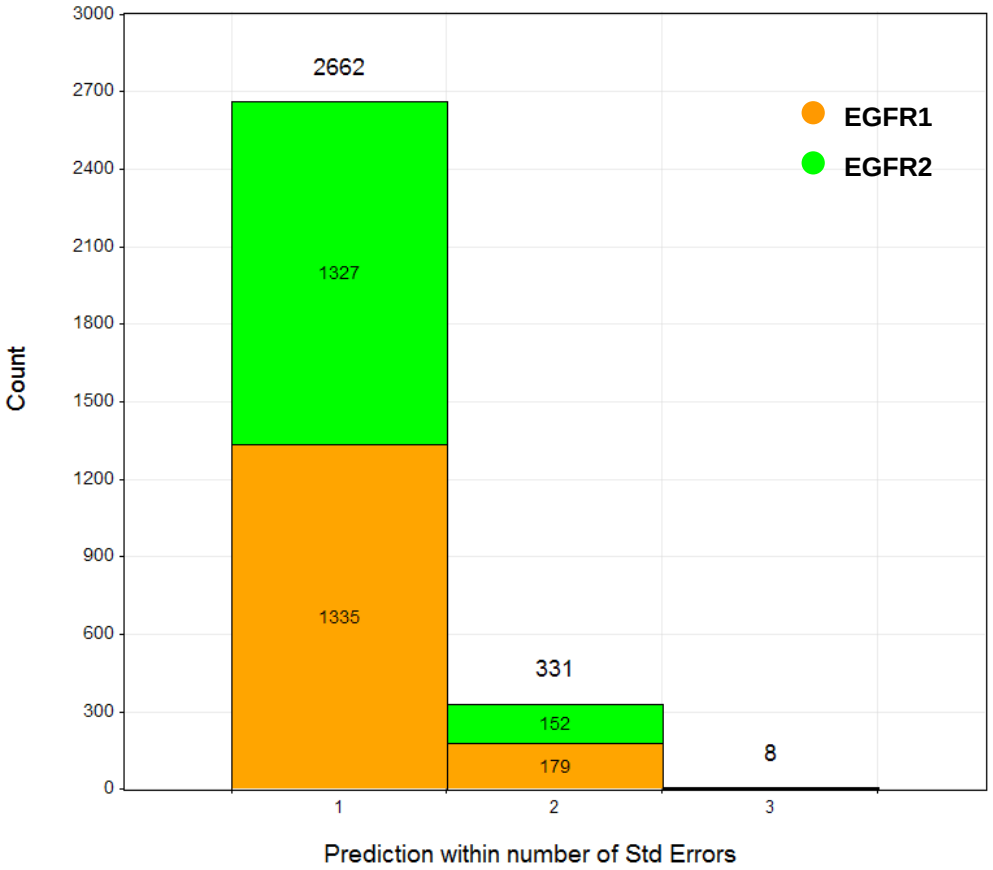
- R (the statistics software) generates an estimated standard error for each R-group coefficient value, representing the variance in the values associated for that R group

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.15E+00	1.0927592	7.46E+00	2.43E-09
xquinazolin-4-amine	9.73E-01	0.3063513	3.18E+00	2.72E-03
x1c1ccc2c(c1)nc(s2)c3ccc(c(c3)F)[R1]	-1.60E+00	0.5816389	-2.76E+00	8.49E-03
x1c1ccc2c(c1)nc(s2)c3ccc(cc3F)[R1]	-1.33E+00	0.5816389	-2.29E+00	2.70E-02
x1c1ccc2c(c1)nc(s2)c3ccc(c(c3)O)[R1]	-1.03E+00	0.4598259	-2.24E+00	3.03E-02

- Combine individual core and R-group standard errors for an estimate of the standard error for <u>each individual predicted</u> molecule
  - $\sigma_{individual\ compound} = \sqrt{\sigma_{core}^2 + \sigma_{R1}^2 + \sigma_{R2}^2 + \cdots + \sigma_{RN}^2}$

Estimating Predictivity II

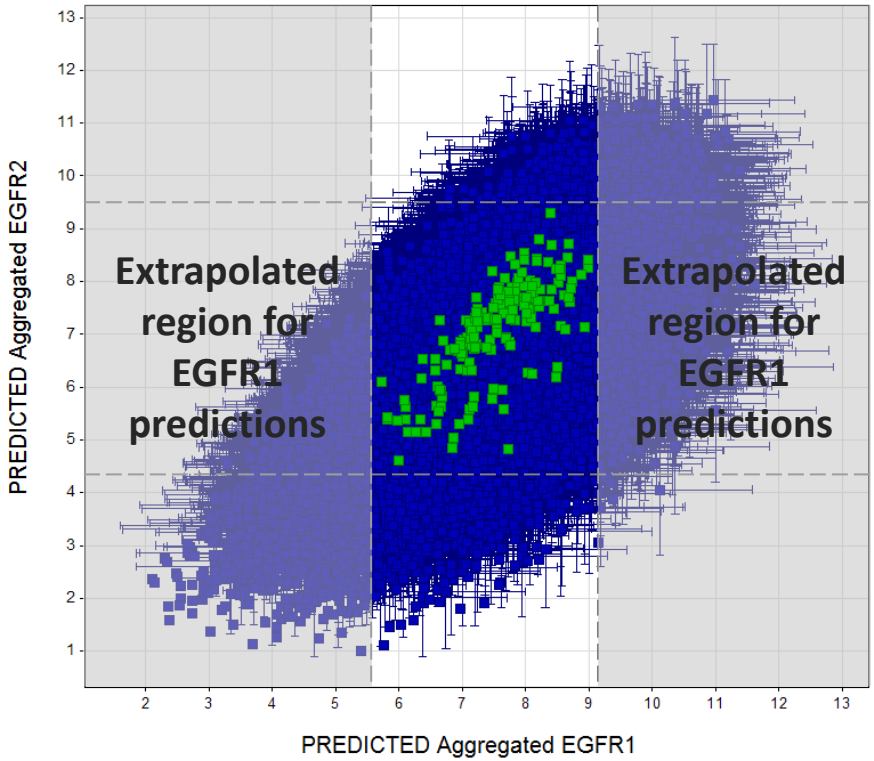
- Use combined standard errors as a guide to overall model predictivity

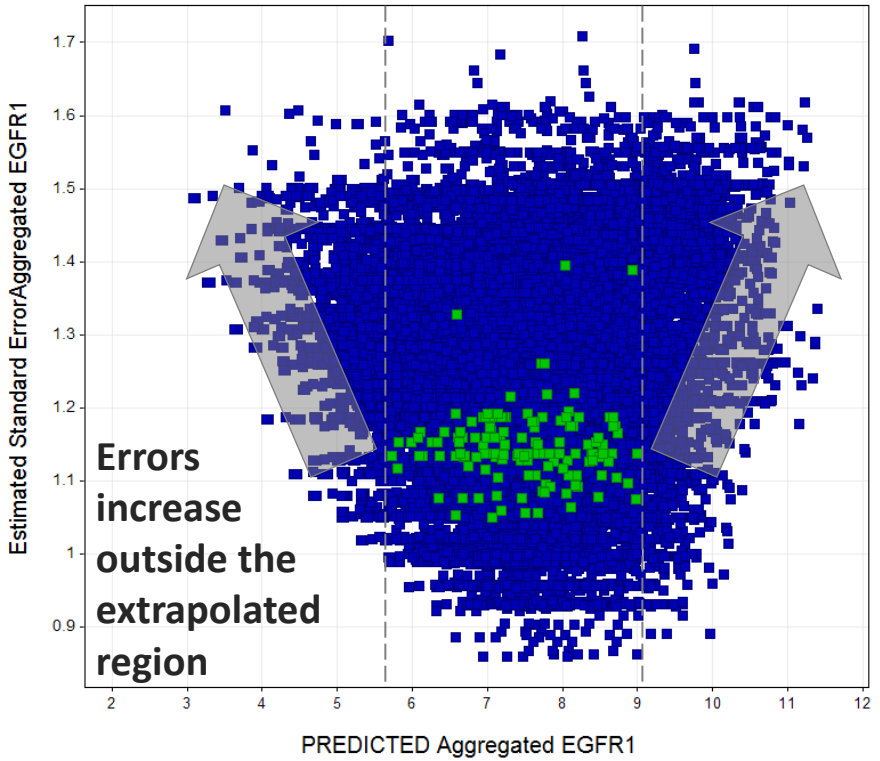


- 88.7% of predictions are within 1 combined standard error; 99.7% within 2 combined standard error values
- Estimates do not exist when only a single observation for any constituent core/R-group is used to make a prediction
- Generate individual error estimates for each predicted compounds

Estimating Predictivity II

- Add individual error estimates to predictions of enumerated products
  - Identify predictions based on single observations
- Frequently see extrapolated points
  - Individual errors increase as predictions extend into the extrapolated domain (don't get with overall model predictivity from RMSE)





Conclusion

- Assess the predictivity of a Free-Wilson model from the average RMSE of one hundred random 20% training sets
  - On average three quarters of potency predictions are within one RMSE value and 95% within two RMSE values
- For predictions derived from multiple core/R-groups, can assign combined and individual estimates of the error
  - Nearly 90% of predictions are within one combined standard error value

Acknowledgements

- Sygnature Discovery Computational Chemistry
  - Colin Sambrook Smith
  - Louise Birch
  - Kam Chohan
  - Amit Kumar Garg
  - Silvia Paoletta
  - Ting Qin
  - Bill Tatsis
- OpenEye Support
  - Jose Batista
  - James Haigh
  - Gunther Stahl