

# 3D QSAR Methods: Phase and Catalyst Compared

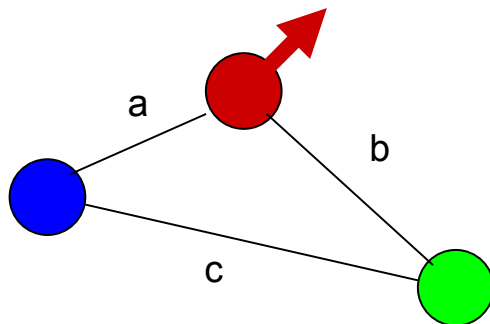
David Evans

Lilly Research Laboratories

UK-QSAR 18<sup>th</sup> October 2006

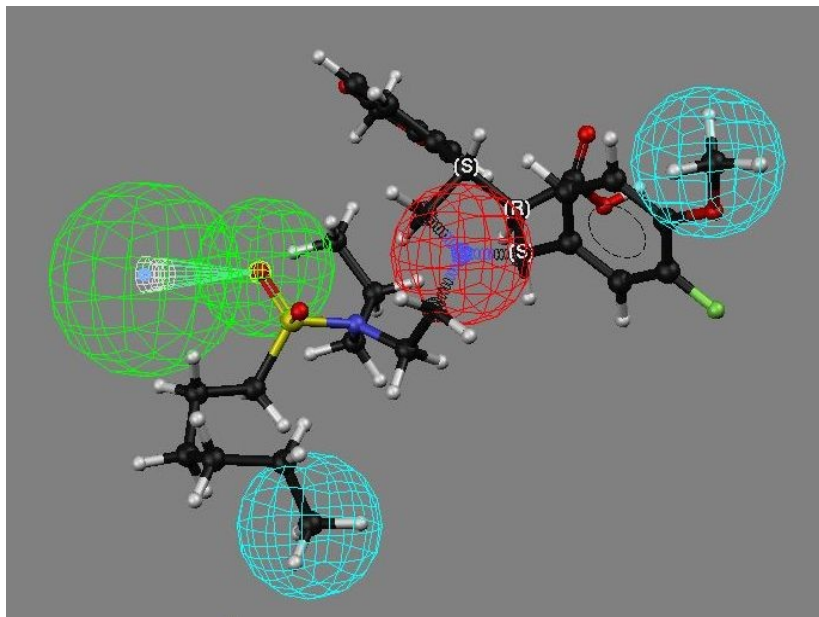
# Introduction

- **HypoGen** from Catalyst (Accelrys) and **Phase** (Schrodinger).
- Both aim to quantitatively predict activity using a training set of compounds with known activity.
- Both search for common 3D pharmacophores amongst active compounds in the training set — known as *hypotheses*.

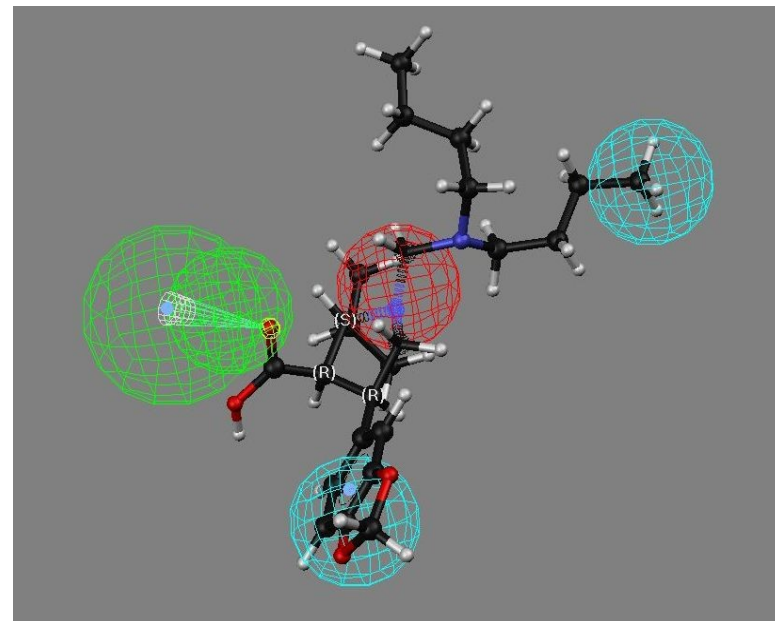


# Estimating Activity?

- **HypoGen** uses the distance between the pharmacophore features (*site points*) in the hypotheses and matching feature in the compounds.



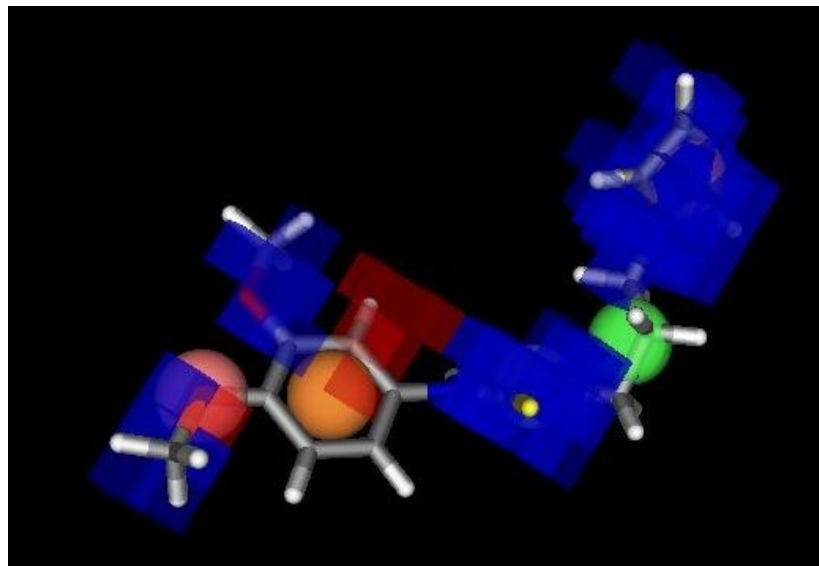
Active compound – features centred on site points



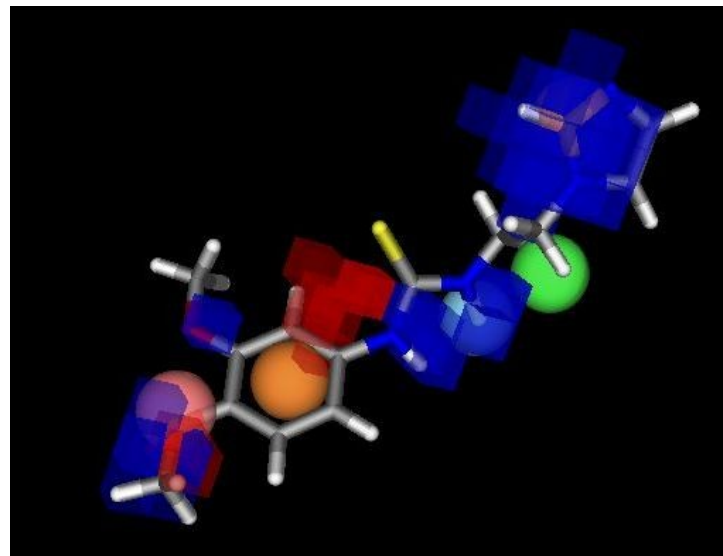
Less active compound – features cannot be centred

# Estimating Activity?

- **Phase** scores compounds using a grid-based 3D QSAR model built from the compounds aligned on the hypotheses.



Active compound – more blue regions



Less Active compound – less blue

# HypoGen Methodology

Select Training Set



Generate Conformers



Find possible Hypotheses from Actives



Reject some Hypotheses using Inactives



Refine and Score Hypotheses

Minimum of 16 compounds, spread of 4 orders of magnitude in activity, includes representatives of different structural classes.

CHARMM-like force field, 'poling' method to give maximum sampling of conformational space.

'Constructive': Active compounds are those within a threshold of the maximum activity (9-fold by default).

Finds all hypotheses which match a minimum (default 4) number of site points in all these actives.

'Subtractive': Inactives are those with activity 3.5 orders of magnitude below the most active compound.

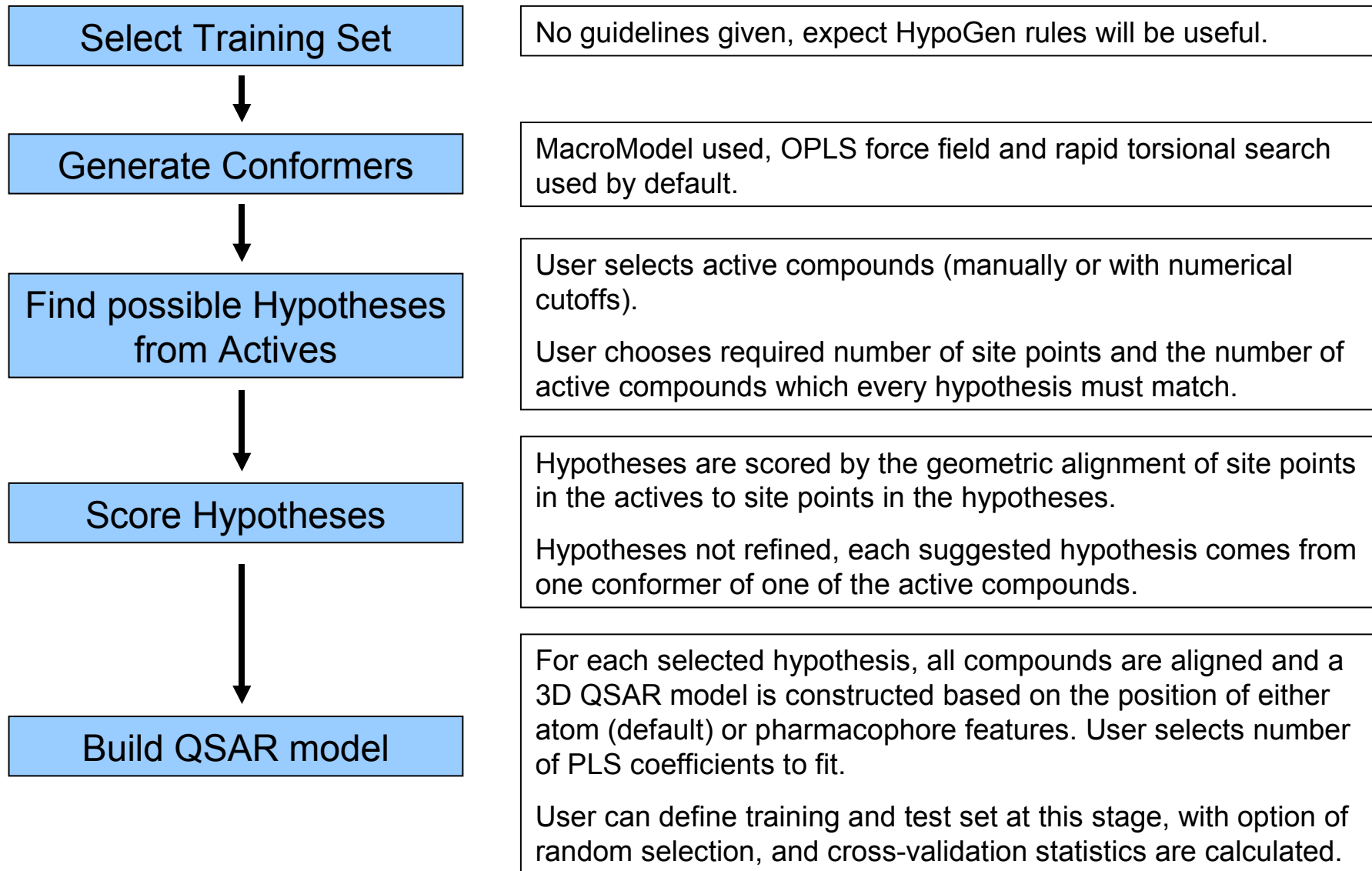
Any hypothesis which matches more than half of these inactives is rejected.

'Optimization': All remaining hypotheses are scored.

$Cost = Error + Feature Weight + Complexity$

Hypotheses are refined by adjusting the site point positions to minimize the cost, using simulated annealing.

# Phase Methodology



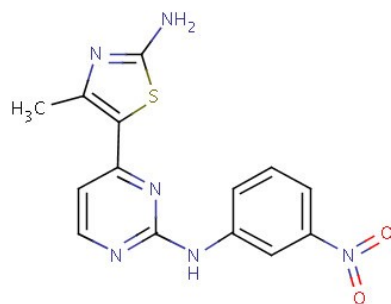
# Data Sets

- Eight public data sets.  $K_i$  or IC50 values.

[www.bindingDb.org](http://www.bindingDb.org)

[www.cheminformatics.org](http://www.cheminformatics.org)

- Chosen without reference to previous 3D QSAR work (with exception of ETA set).



CDK2

52 compounds

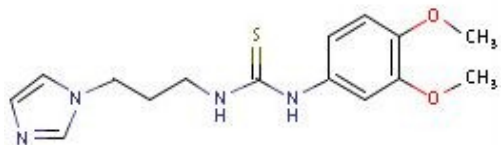
MPS 0.62

- Aim to have similar sized test set and training set
- Assess based on predictions on test set.

Mean Pairwise Similarity (MPS):

The average Tanimoto Coefficient of MACCS fingerprints across all pairs of compounds in the set (MOE).

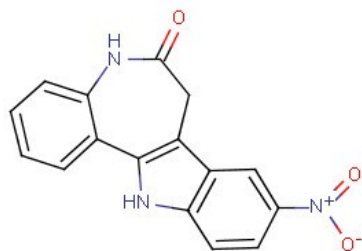
# Data Sets



Glutaminyl cyclase

63 compounds

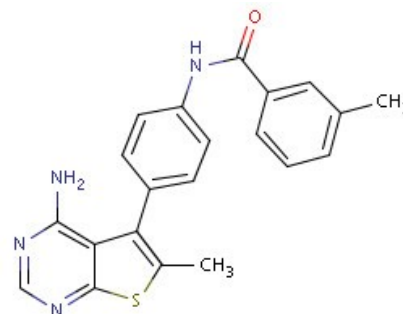
MPS 0.47



CDK1

79 compounds

MPS 0.64



VEGFR2

63 compounds

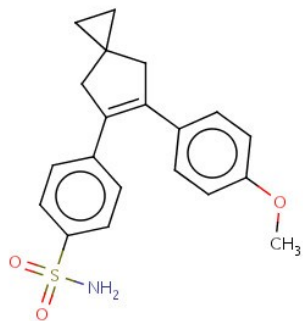
MPS 0.77



ETA

61 compounds

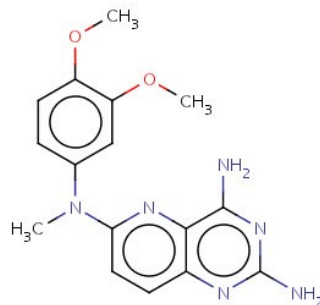
MPS 0.77



COX2

81 compounds

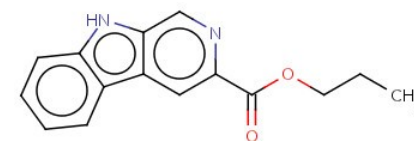
MPS 0.62



DHFR

73 compounds

MPS 0.69



BZR

73 compounds

MPS 0.47

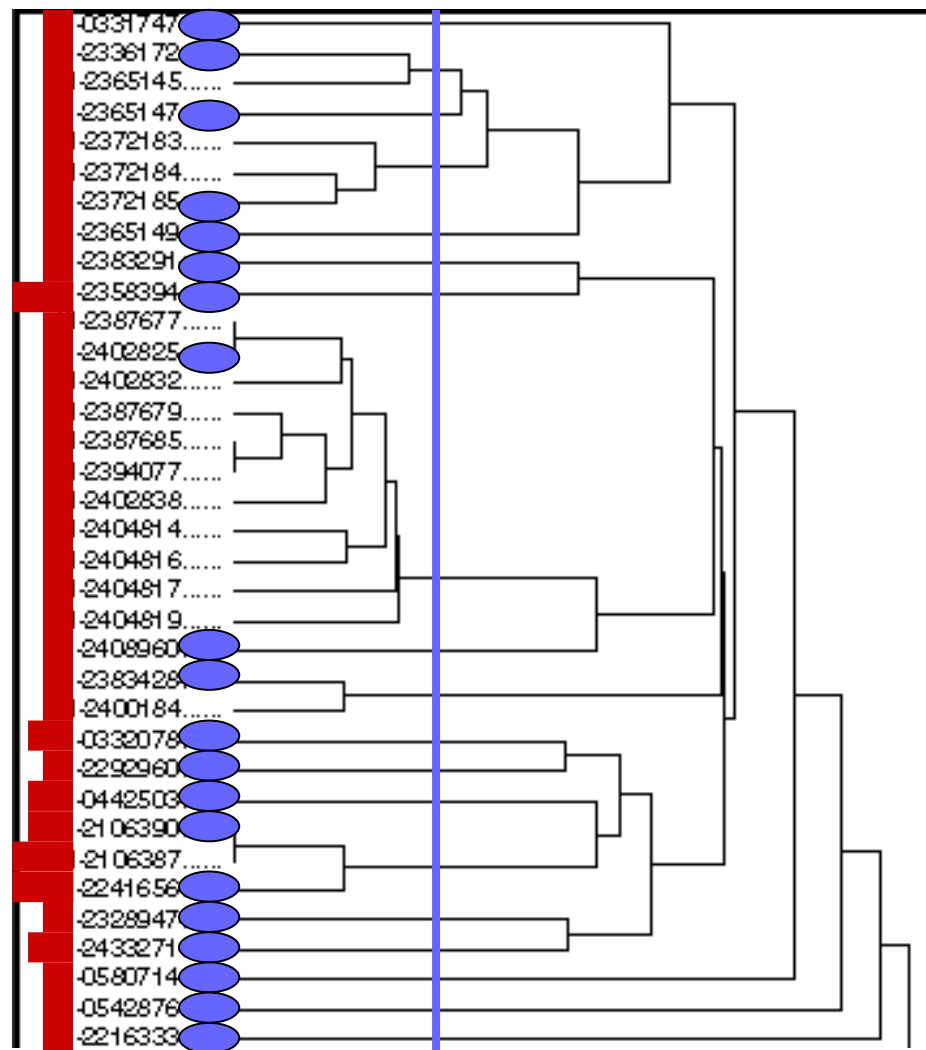
# Preparing Training Sets

Catalyst 'rules' require:

- 4 orders of magnitude spread of activity
- No similar compounds which are also similar in activity.

Lilly 'CatScan' program makes training sets and runs automated Catalyst jobs.

5. Cluster with fingerprint similarity.
6. Include more than one compound per cluster only if they differ significantly in activity.



# Generating Models

- Each program: 5 jobs for each data set. Command-line scripts.
- Phase:
  1. Default
  2. Pharmacophore QSAR grid
  3. Cutoff for defining active (1-3 pK units)
  4. 3 + Hypothesis scoring to include activity of reference ligand
  5. 3 + Lower tolerance of matching molecules to site points
- HypoGen:
  1. Default
  2. Minimum spacing of site points.
  3. Weight and tolerances of scoring different features vary
  4. 2 + 3
  5. Add excluded volume as extra site point.

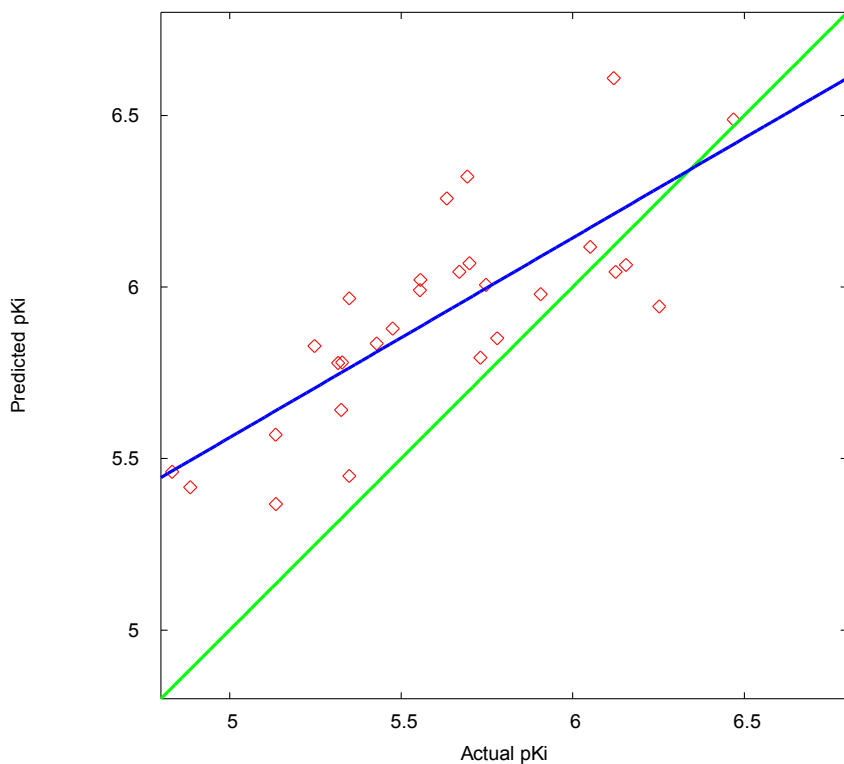
# Results on Test Sets

- 5 x 10 top scoring hypotheses → Predict on Test Set
- Results are presented for the highest  $R^2_{\text{test}}$ .

System	Phase						HypoGen					
	1	2	3	4	5	Best	1	2	3	4	5	Best
CDK2	-0.34	-0.27	0.06		-0.20	0.06	-0.31	-0.28	-0.09	-0.26	-0.48	-0.09
GlutCyc	0.11	0.04	0.13	0.20	0.03	0.20	-1.10	-0.93	-0.40	-0.52	-0.92	-0.40
CDK1	0.37		0.46	0.46	0.48	0.48	-0.02	-0.05	-0.29	0.32	0.00	0.32
VEGFR2	0.18	-0.69	0.17	0.17	0.20	0.20	0.10	0.05	0.27	0.10	0.09	0.27
ETA	0.13	-0.43	0.16	0.28	0.28	0.28	-0.63	-0.21	-0.19	0.09	-0.52	0.09
COX2	0.18	0.13	0.18	0.23	0.18	0.23	-0.06	0.05	-0.03	0.13	-0.03	0.13
DHFR	0.34	0.26	0.27	0.27	0.26	0.34	0.24	0.22	0.20	0.22	0.21	0.24
BZR	0.26	0.11	0.48	0.46	0.43	0.48	0.11	0.15	0.08	0.03	0.14	0.15
Mean	0.15	-0.12	0.24	0.30	0.21	0.28	-0.21	-0.13	-0.05	0.01	-0.19	0.08

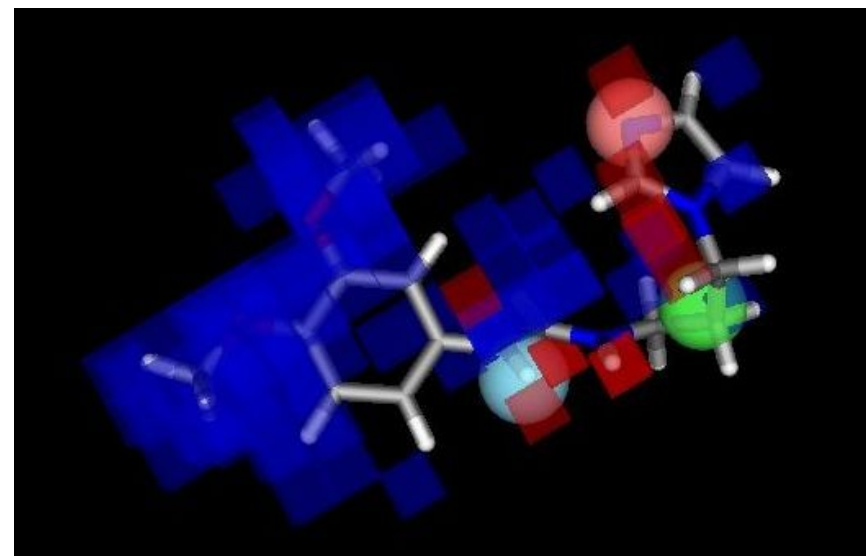
# Example Correlations

## Glutaminyl Cyclase Phase '1' hypothesis



$$R^2 = 0.02$$

$$\text{Pearson } r = 0.78$$



3D QSAR model projected on most active compound

# Results on Test Sets

- Pearson  $r$ -values
- Measure correlation, not exact match. Amber 99.9% chance  $r > 0$ .

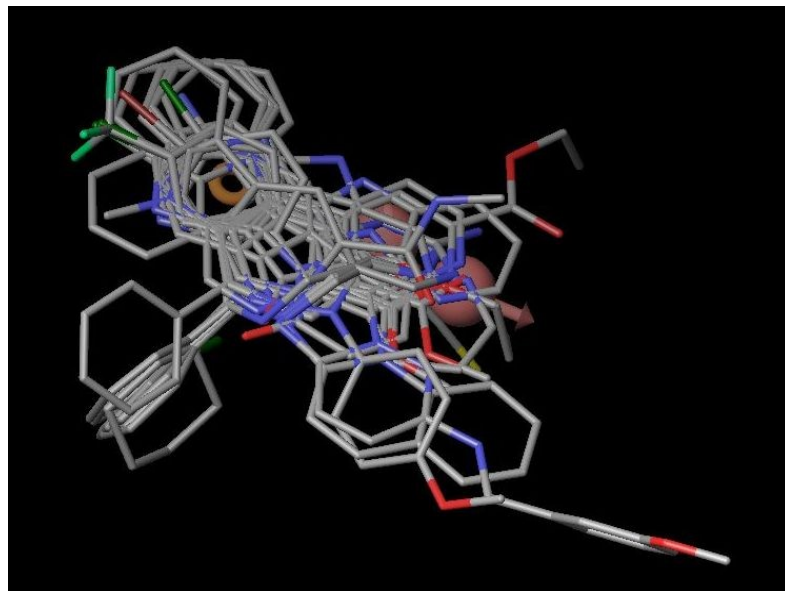
System	Phase						HypoGen					
	1	2	3	4	5	Best	1	2	3	4	5	Best
CDK2	0.32	0.62	0.39		0.37	0.62	0.46	0.53	0.51	0.58	0.46	0.58
GlutCyc	0.78	0.52	0.60	0.73	0.55	0.78	0.74	0.76	0.77	0.74	0.70	0.77
CDK1	0.65		0.78	0.78	0.78	0.78	0.36	0.39	0.35	0.64	0.39	0.64
VEGFR2	0.67	0.41	0.62	0.62	0.63	0.67	0.61	0.61	0.63	0.61	0.60	0.63
ETA	0.51	0.36	0.49	0.66	0.63	0.66	0.77	0.75	0.79	0.78	0.78	0.79
COX2	0.45	0.41	0.45	0.49	0.45	0.49	0.17	0.44	0.17	0.41	0.22	0.44
DHFR	0.58	0.52	0.52	0.52	0.51	0.58	0.54	0.57	0.55	0.57	0.53	0.57
BZR	0.52	0.39	0.71	0.68	0.67	0.71	0.45	0.43	0.42	0.34	0.47	0.47
Mean	0.56	0.40	0.57	0.56	0.57	0.66	0.51	0.56	0.52	0.58	0.52	0.61

# Comments

- Important to **test all hypotheses** to find good model, not just top scoring.
- Experimentation with more extensive conformational sampling in both programs did not yield significantly better results.
- Phase gives statistics excluding compounds which do not match the hypothesis
  - Feature as initial step to identifying different classes/ binding modes
  - Need to write own code for fair comparison
- Catalyst prediction much slower than Phase (with Best fit), harder to test as many hypotheses.

# Transplant Conformers

Assess whether poor conformer generation or poor scoring was responsible for the poor models.



Input to Catalyst  
as training set  
conformer library



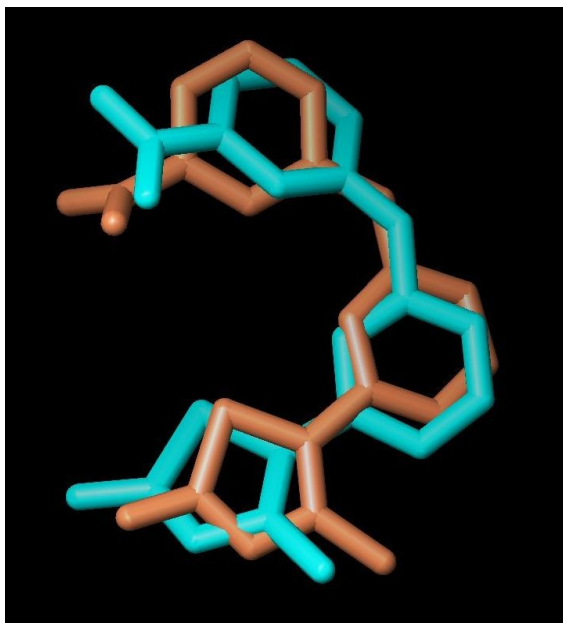
Good Catalyst Model ?  
NO  
Pearson  $r = 0.43$

Phase BZR training set  
compounds aligned with best  
hypothesis

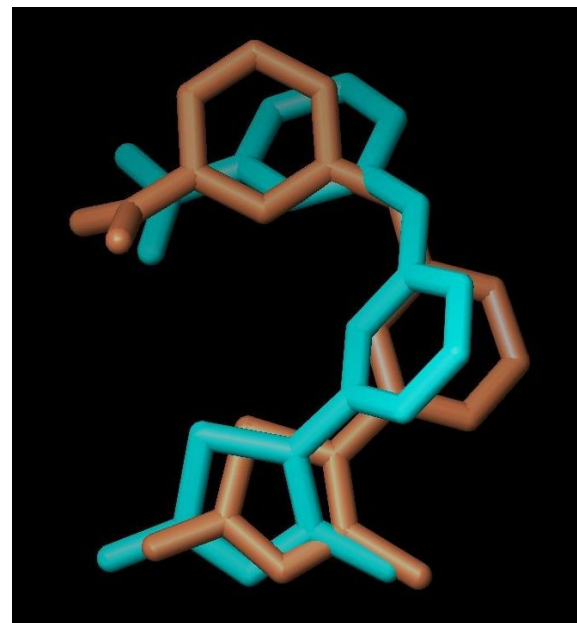
- Indicates that scoring is the problem
- Programs unable to model activity even if compounds are input in 'correct' overlays?

# CDK2: Structure-Based Comparison

- CDK2 data set: PDB structure (2C5P) of the most active compound in complex with receptor.
- Does the conformer generation in the two programs find the crystallised conformation?



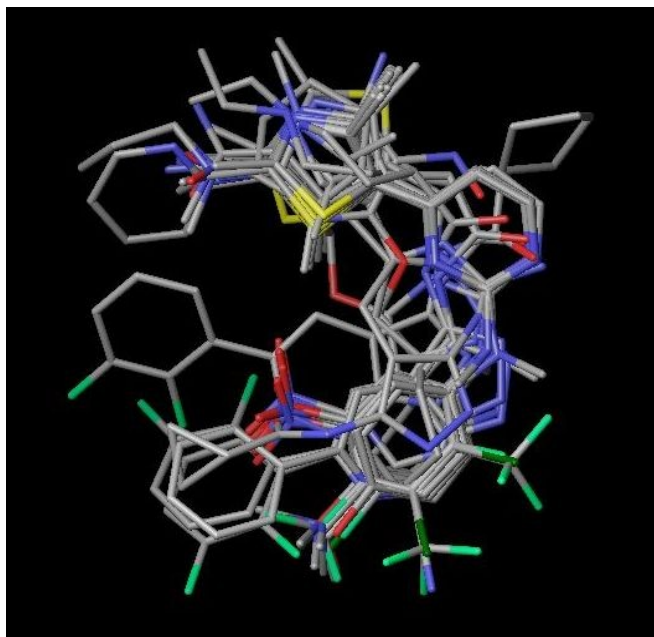
Phase best alignment RMS = 2.4 Å



Catalyst best alignment RMS = 3.1 Å

# CDK2: Alignments to X-ray structure

- Generate alignments of all compounds to the crystal conformation with ROCS (OpenEye). Shape + feature matching.
- Input these to Phase and Catalyst.



ROCS alignment

Input to Phase



Best  $r_{\text{test}} = 0.51$

Input to HypoGen

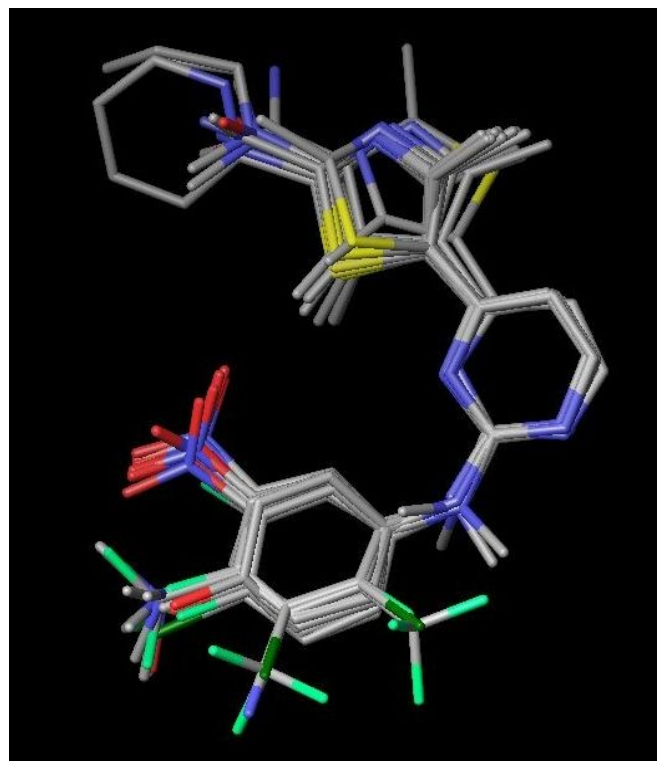


Best  $r_{\text{test}} = 0.56$

Overlay still not convincing –  
different series of compounds not  
overlaid correctly?

# CDK2: Alignments to X-ray structure

- Restrict to one series of compounds



ROCS alignment

Input to Phase



Best  $r_{\text{test}} = 0.46$

Input to Catalyst



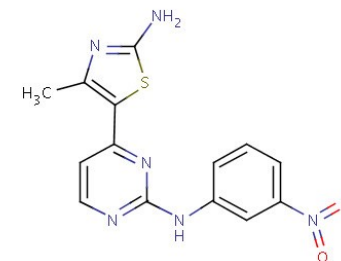
Best  $r_{\text{test}} = 0.56$

Input to CoMFA



$r_{\text{test}} = 0.42$

$r_{\text{LOO}}$  on entire  
data set = 0.35



# Conclusions

- Phase  $R^2$  and Pearson  $r$  values higher than Catalyst on average.
- Evaluated recommended strategies for improving models.
- Hypothesis: Is the problem with scoring & not conformations?
  - If so why?
    - Data sets just too small to build model?
    - Multiple binding modes for ligands and proteins.
    - $\Delta G$  depends on free ligand state as well as bound state.

# Acknowledgements

## *Lilly, Erl Wood*

- Mike Bodkin
- Dave Thorner
- Sandra Milutinovic
- Linda Hiron

## *Lilly, Indianapolis*

- Thompson Doman

## *Lilly, Alcobendas*

- Howard Broughton

## *Accelrys*

- Katalin Nadassy

## *Schrodinger*

- Jas Gata-Aura
- Stuart Murdock
- Steve Dixon

[www.bindingDb.org](http://www.bindingDb.org)

[www.cheminformatics.org](http://www.cheminformatics.org)

# Appendix: 50/50 Train/Test

- 5 x 10 top scoring hypotheses → Predict on Test Set
- Results are presented for the highest  $R^2_{\text{test}}$ .

System	Phase Original						Phase 50/50					
	1	2	3	4	5	Best	1	2	3	4	5	Best
CDK2	-0.34	-0.27	0.06		-0.20	0.06	0.08	0.17				0.17
GlutCyc	0.11	0.04	0.13	0.20	0.03	0.20	0.25	0.16	0.23	0.23	0.22	0.25
CDK1	0.37		0.46	0.46	0.48	0.48	0.53	0.23	0.60	0.60	0.61	0.61
VEGFR2	0.18	-0.69	0.17	0.17	0.20	0.20	0.39	0.09	0.40	0.40	0.30	0.40
ETA	0.13	-0.43	0.16	0.28	0.28	0.28	0.40	0.30	0.55		0.55	0.55
COX2	0.18	0.13	0.18	0.23	0.18	0.23	0.12	0.13	0.12	0.09	0.13	0.13
DHFR	0.34	0.26	0.27	0.27	0.26	0.34	0.44	0.18	-0.03			0.44
BZR	0.26	0.11	0.48	0.46	0.43	0.48	0.36	0.01	0.50	0.47	0.42	0.50
Mean	0.15	-0.12	0.24	0.30	0.21	0.28	0.32	0.16	0.34	0.36	0.37	0.38

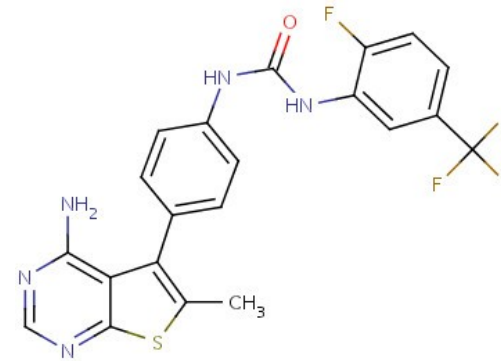
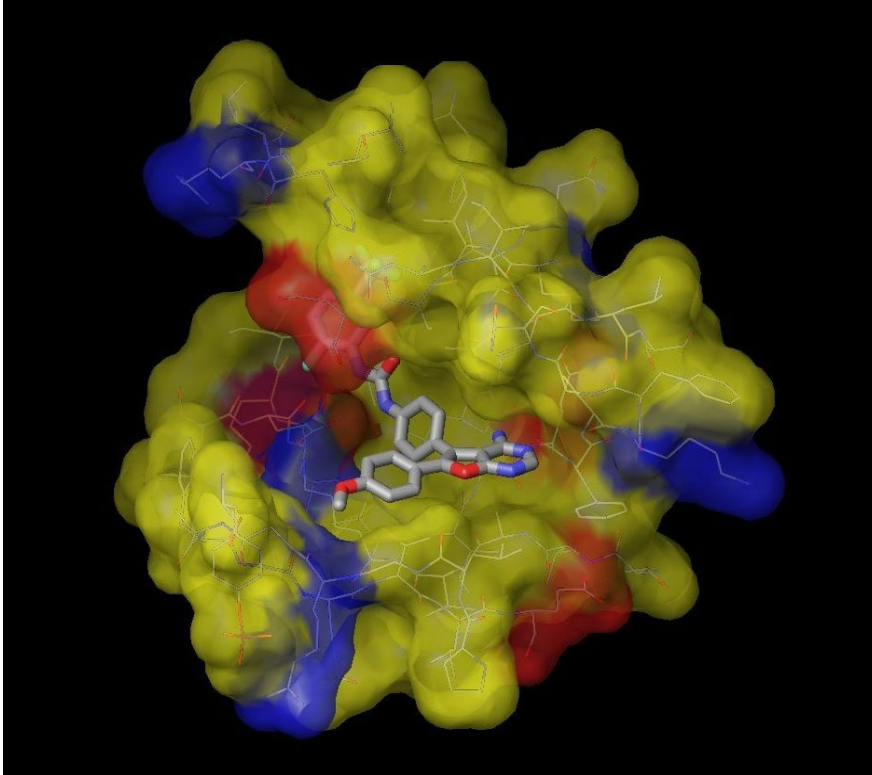
# Appendix: 50/50 Train/Test

- Pearson  $r$ -values
- Measure correlation, not exact match. Amber 99.9% chance  $r > 0$ .

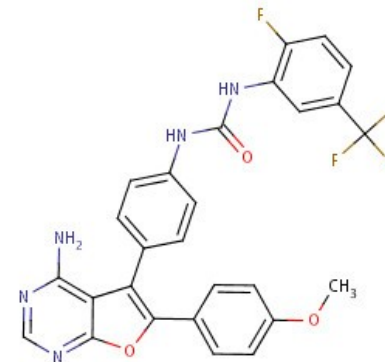
System	Phase Original						Phase 50/50					
	1	2	3	4	5	Best	1	2	3	4	5	Best
CDK2	0.32	0.62	0.39		0.37	0.62	0.35	0.55				
GlutCyc	0.78	0.52	0.60	0.73	0.55	0.78	0.59	0.55	0.53	0.53	0.55	0.59
CDK1	0.65		0.78	0.78	0.78	0.78	0.73	0.48	0.78	0.78	0.78	0.78
VEGFR2	0.67	0.41	0.62	0.62	0.63	0.67	0.65	0.32	0.69	0.69	0.57	0.69
ETA	0.51	0.36	0.49	0.66	0.63	0.66	0.66	0.61	0.74		0.74	0.74
COX2	0.45	0.41	0.45	0.49	0.45	0.49	0.47	0.41	0.45	0.42	0.46	0.47
DHFR	0.58	0.52	0.52	0.52	0.51	0.58	0.66	0.42				
BZR	0.52	0.39	0.71	0.68	0.67	0.71	0.61	0.44	0.71	0.69	0.66	0.71
Mean	0.56	0.40	0.57	0.56	0.57	0.66	0.59	0.47	0.56	0.52	0.54	0.65

# VEGFR2: Alignments to X-ray structure

- VEGFR2 data set has analogous ligand in the PDB (1YWN).



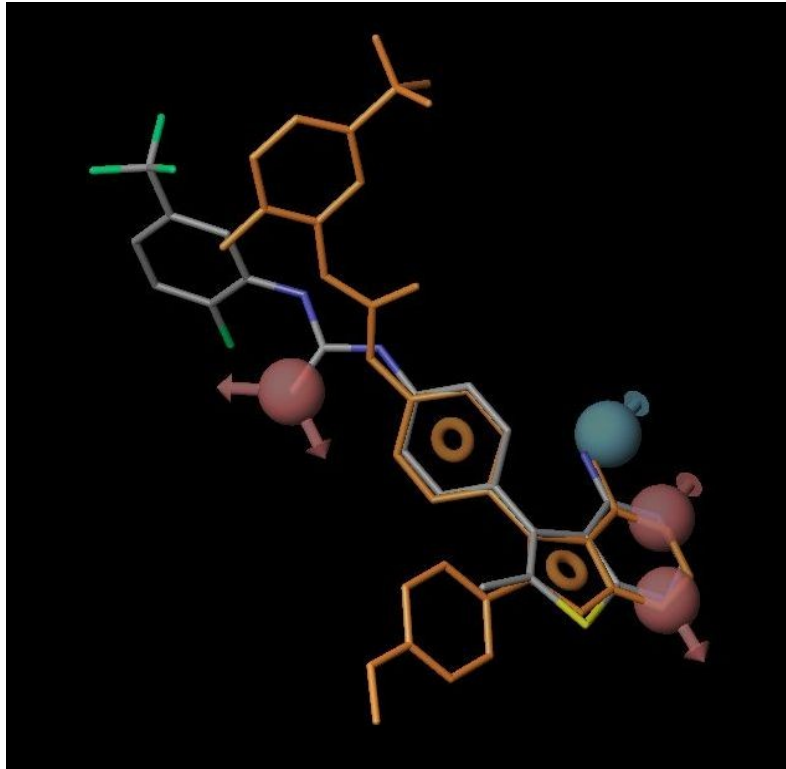
Data set ligand



PDB ligand

# VEGFR2: Alignments to X-ray structure

- Compare alignment with best Phase hypothesis ( $r_{\text{test}} = 0.67$ ) to crystal



Orange = crystal

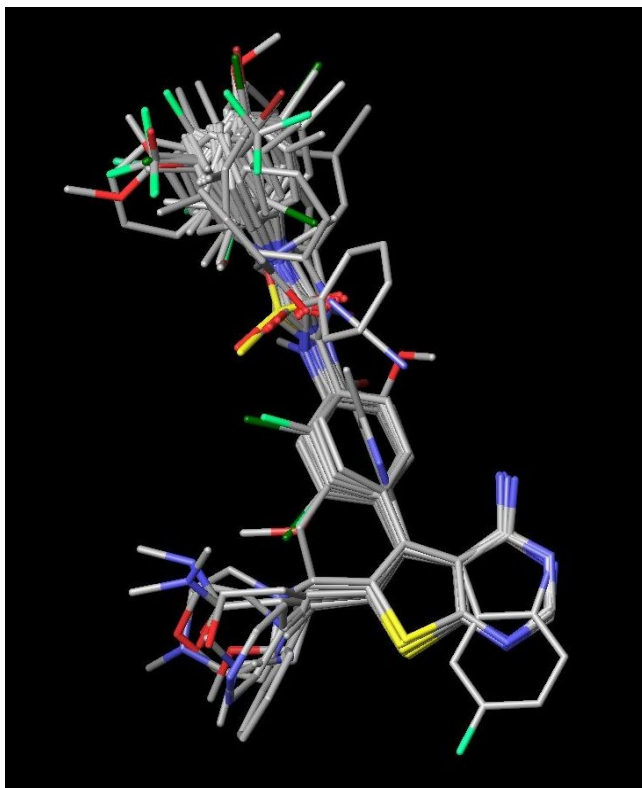
Urea oriented incorrectly

- Alignment good apart from orientation of urea.
- No reason to expect program to get this correct in absence of receptor structure.
- Urea interacts with Glu CO<sub>2</sub><sup>-</sup> in protein

# VEGFR2: Alignments to X-ray structure

Original  
Results:

	Phase	Catalyst
VEGFR2	0.67	0.52



ROCS alignment

Input to Phase



Best  $r_{\text{test}} = 0.55$

Input to HypoGen



Best  $r_{\text{test}} = 0.62$

Input to CoMFA



$r_{\text{test}} = 0.41$   
 $r_{\text{LOO}}$  on entire  
data set = 0.62