Adaptive sampling for alchemical free energy calculations

Hannah Bruce Macdonald - John Chodera

Memorial Sloan Kettering Cancer Center, NYC

Email: Hannah.brucemacdonald@choderalab.org
Twitter: @hannahbruce

Memorial Sloan Kettering Cancer Center

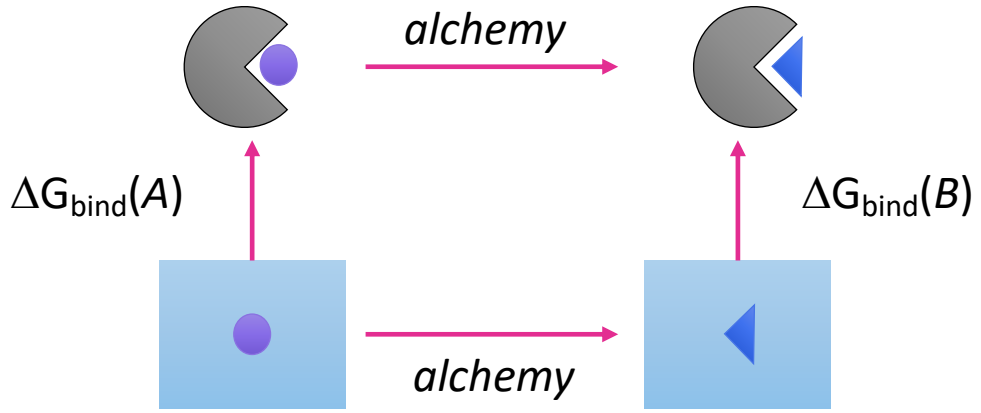
Outline

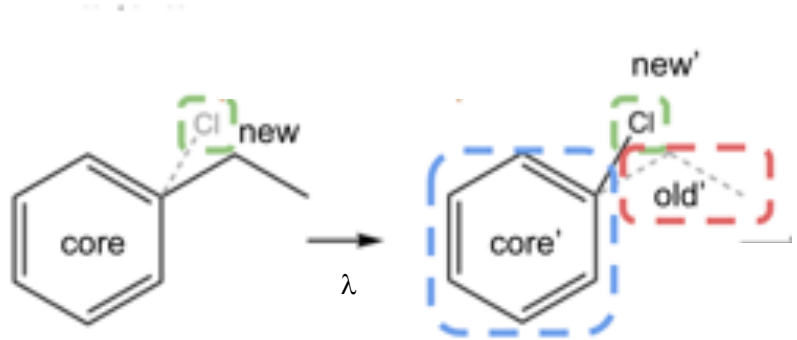
- Introduction to free energy calculations
- Part A - Adaptive sampling for given rewards (Bayesian bandits)
- Part B - Intelligent experiment design

For a given amount of resources, how can we best use that to ask questions?

Free energy calculations

- Alchemically perturb one ligand into another
- Doing this in two phases allows for free energy cycles to be formed

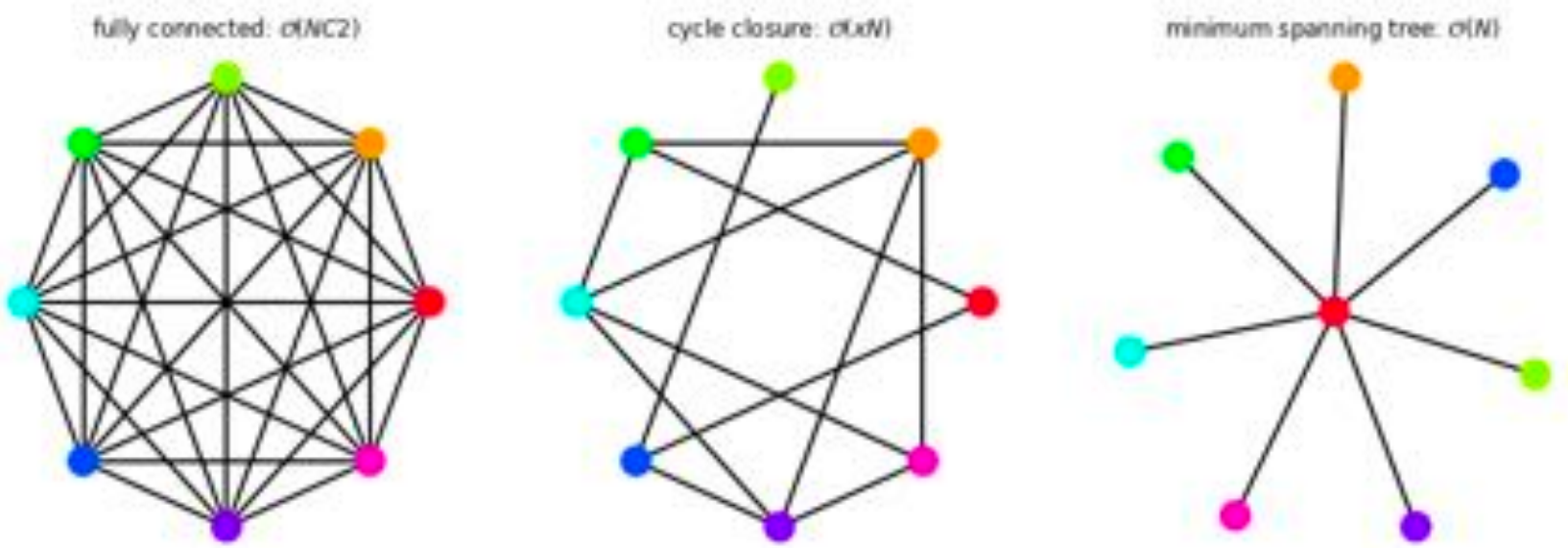




Motivation: Free energy calculations are computationally expensive – how can we run them most efficiently?

Pairwise comparisons of ligands

- Relative free energy calculations do pairwise comparisons of ligands within a set.

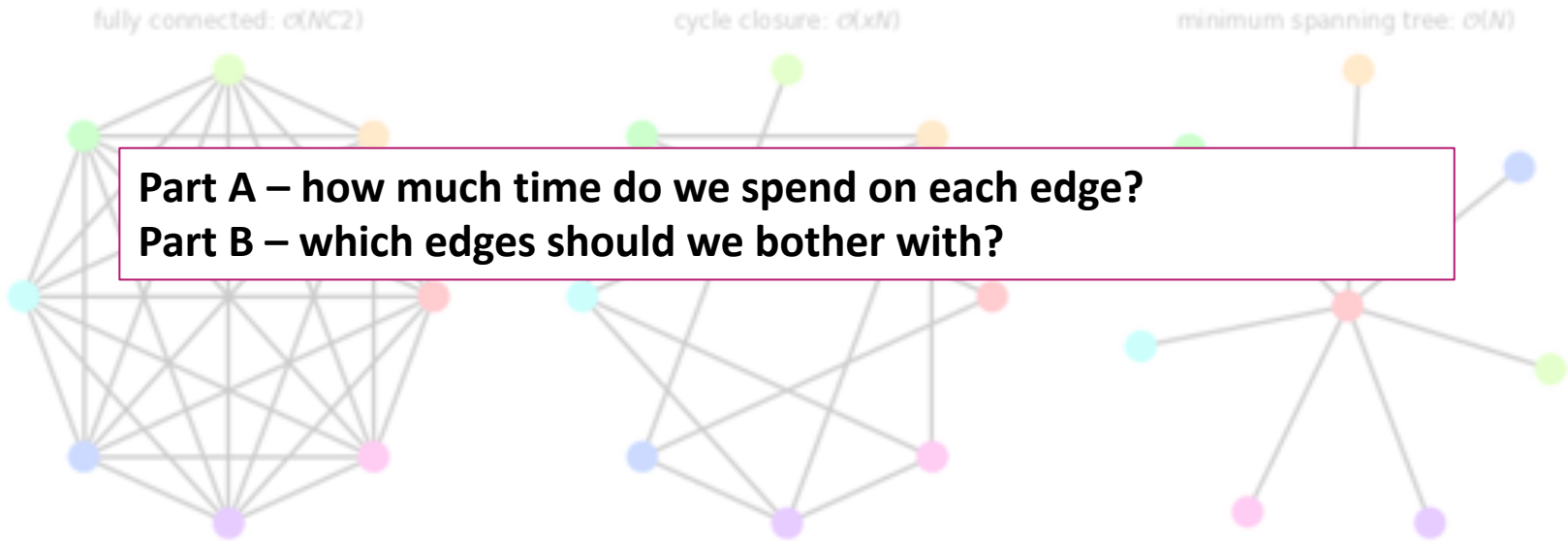


Computationally expensive

More error prone

Pairwise comparisons of ligands

- Relative free energy calculations do pairwise comparisons of ligands within a set.

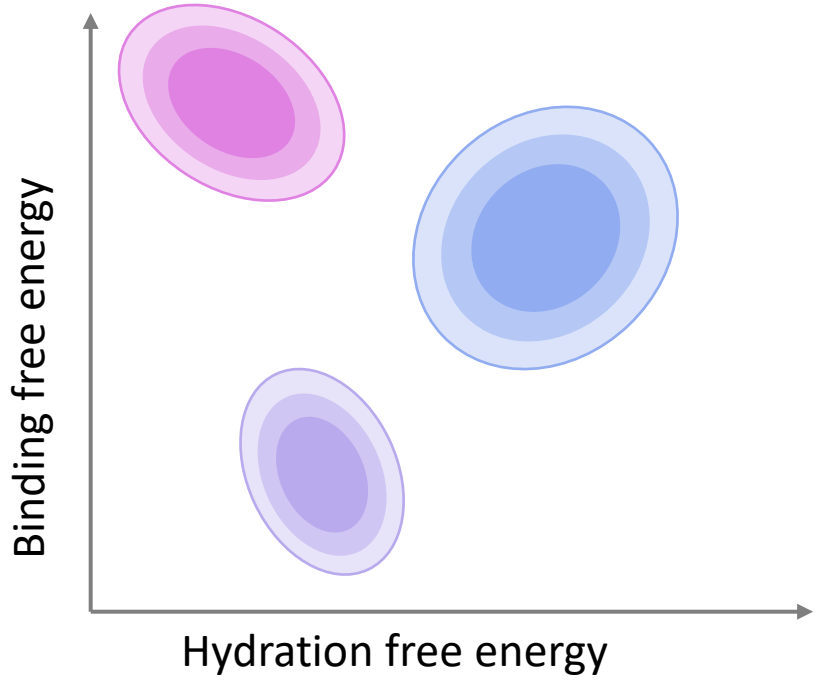


Perses

- github.com/choderalab/perses
- Open-source relative free energy software, developed in the Chodera lab
- Single-topology type calculations (*dual-topology coming soon*)
- Uses *openmm* as MD engine

Part A - Adaptive sampling

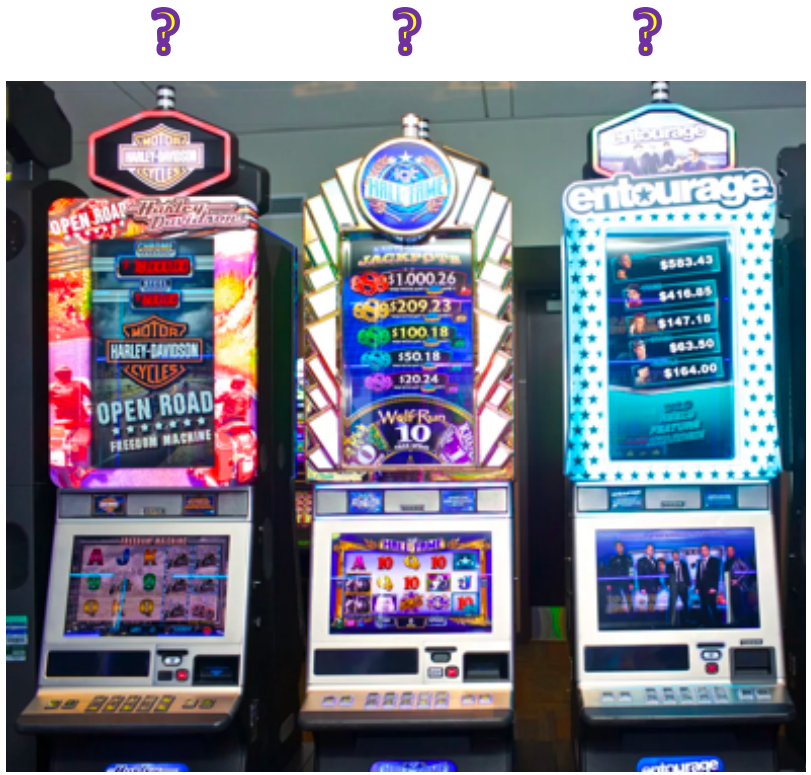
- When considering a set of molecules and
  - Sampling high affinity ligands?
  - Uncertain ligands?
  - Multiple properties?



Bayesian bandits

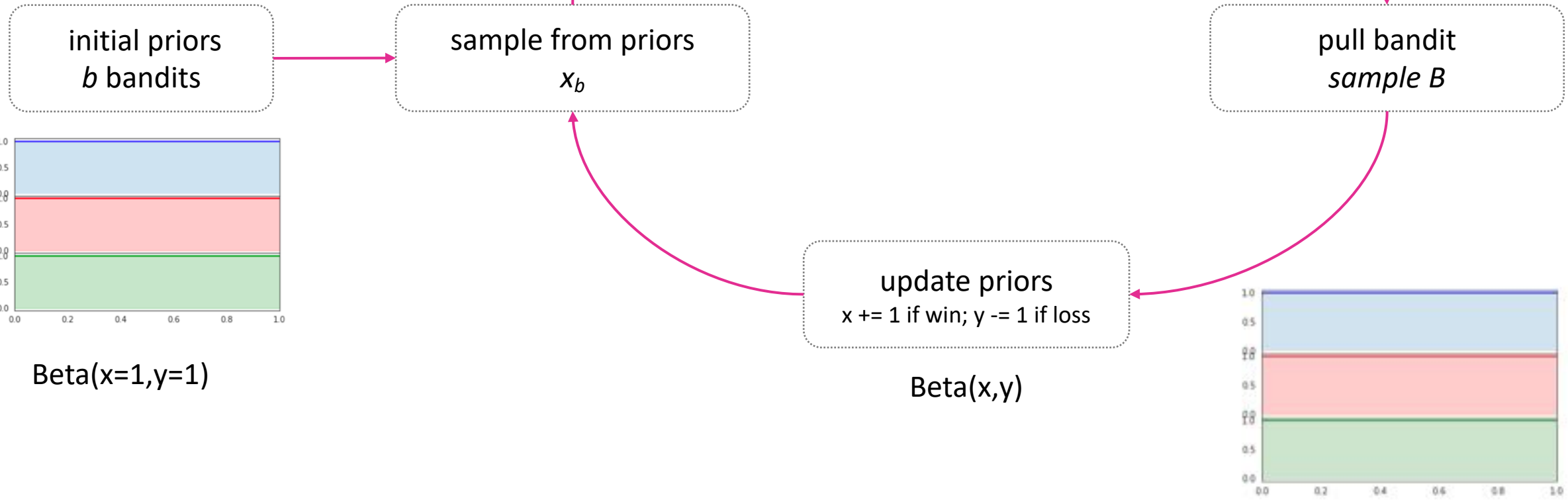
- Or multi-armed bandits
- Decision making based on what we understand of the system (so far)
- As we sample more, our understanding improves

- Applications:
  - *Gambling*
  - *A/B testing*
  - *Drug trials*



Bayesian bandits

- Beta-Bernoulli bandit
- Thompson sampling



Bayesian bandits

- Beta-Bernoulli bandit
- Thompson sampling

select bandit
$B = \text{argmax}(x_b)$

initial priors
b bandits

sample from priors
x_b

pull bandit
sample B

Exploration-exploitation tradeoff

update priors
x += 1 if win; y -= 1 if loss

Beta(x=1,y=1)

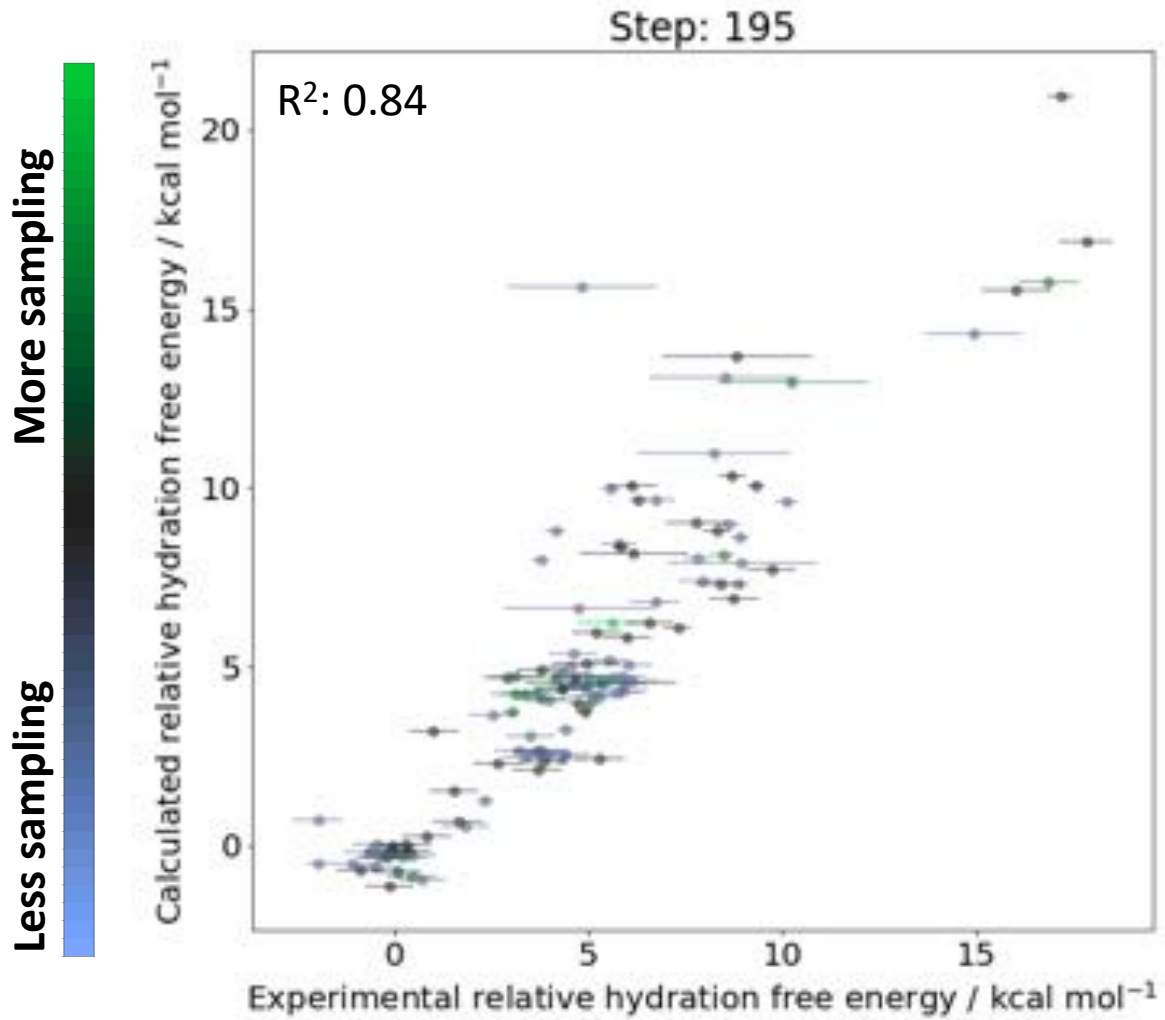
Beta(x,y)

Bayesian bandits

- What would be a pharmaceutically relevant reward?
  - Increasing the sampling of highly soluble ligands
  - Increasing the sampling of uncertain results
  - Increasing the sampling of favourable binders

Bayesian bandits

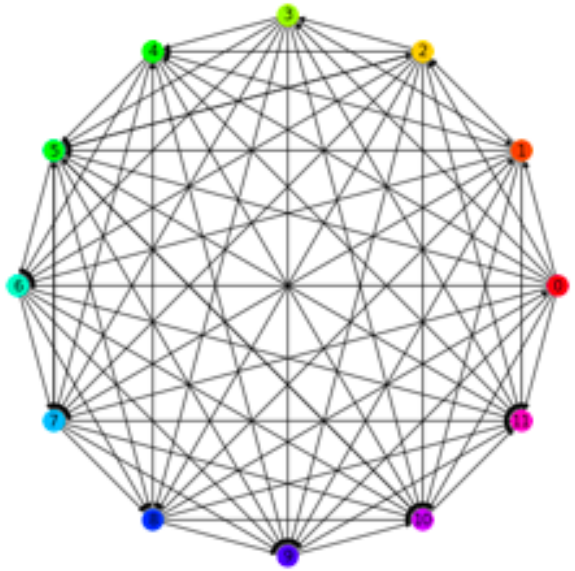
- Hydration free energies of 196 benzene derivatives from the **freesolv**[1] dataset
- Toy data – sampling from experimental results

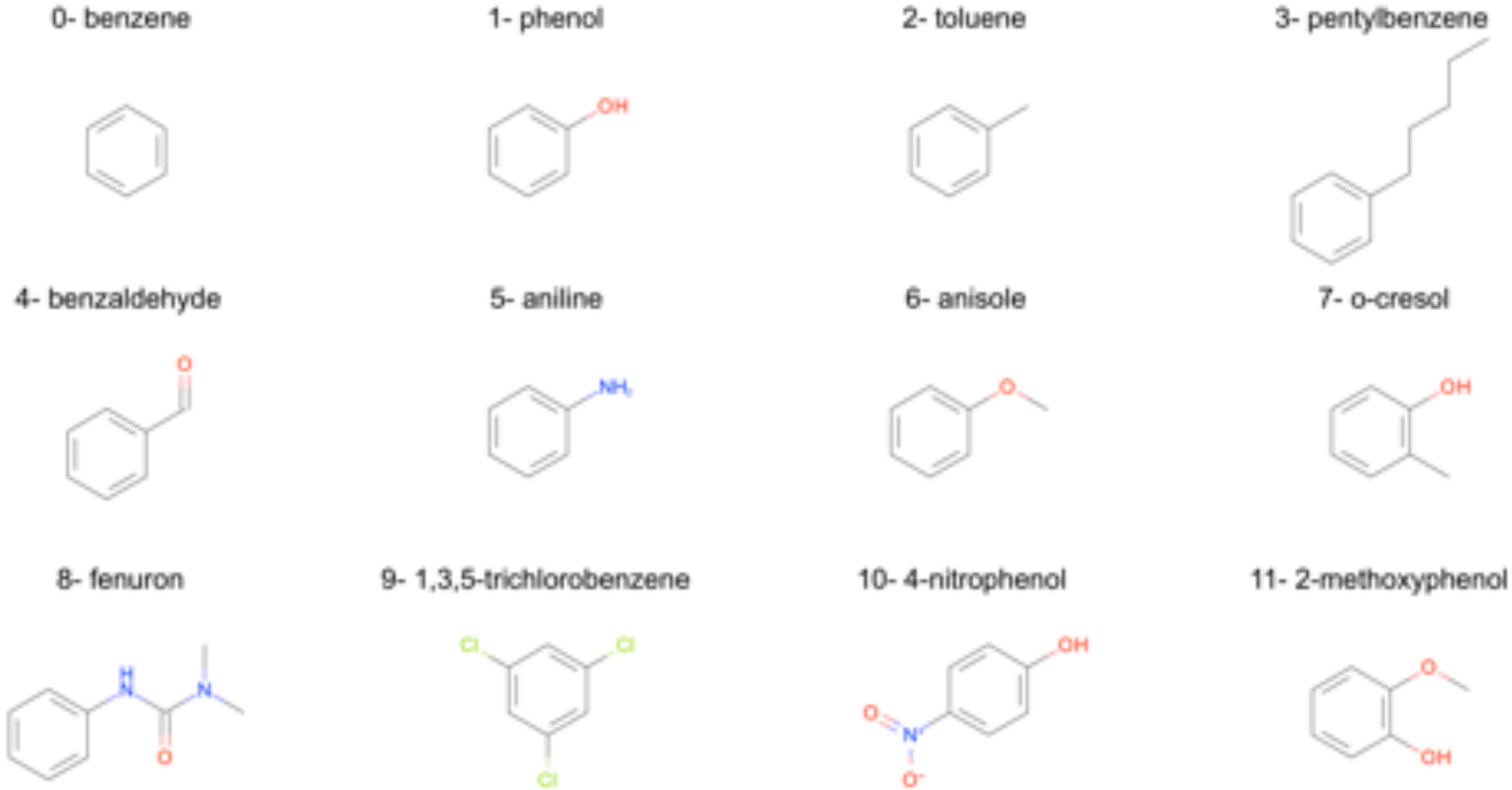


1) Mobley, David L., and J. Peter Guthrie. "FreeSolv: a database of experimental and calculated hydration free energies, with input files." *Journal of computer-aided molecular design* 28.7 (2014): 711-720.

Bayesian bandits

- Hydration free energies of 12 benzene derivatives from the **freesolv**[1] dataset



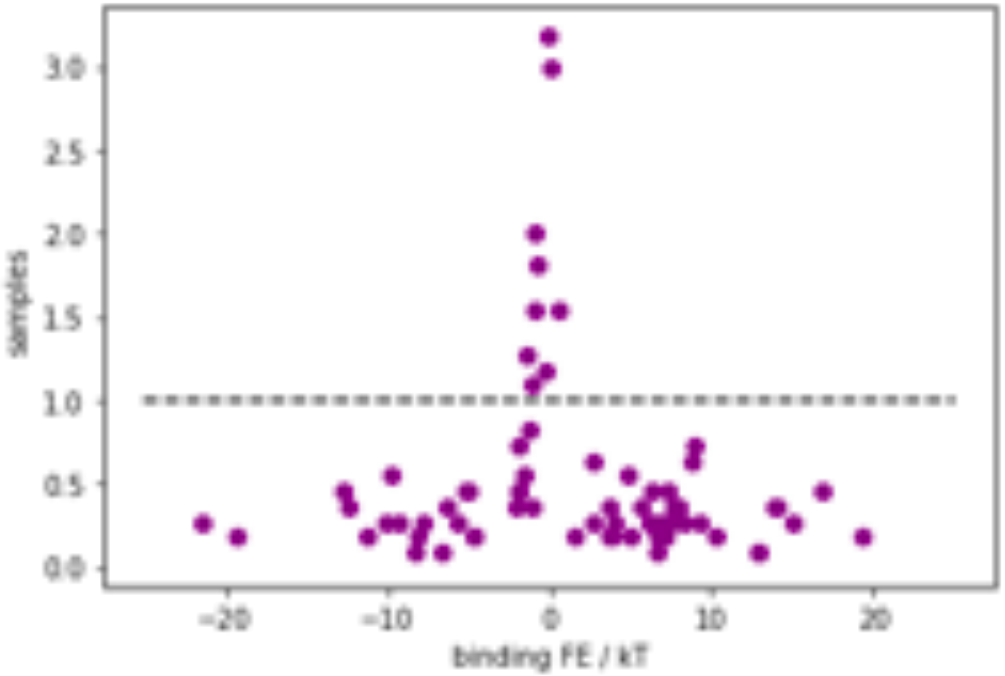


1) Mobley, David L., and J. Peter Guthrie. "FreeSolv: a database of experimental and calculated hydration free energies, with input files." *Journal of computer-aided molecular design* 28.7 (2014): 711-720.

Bayesian bandits

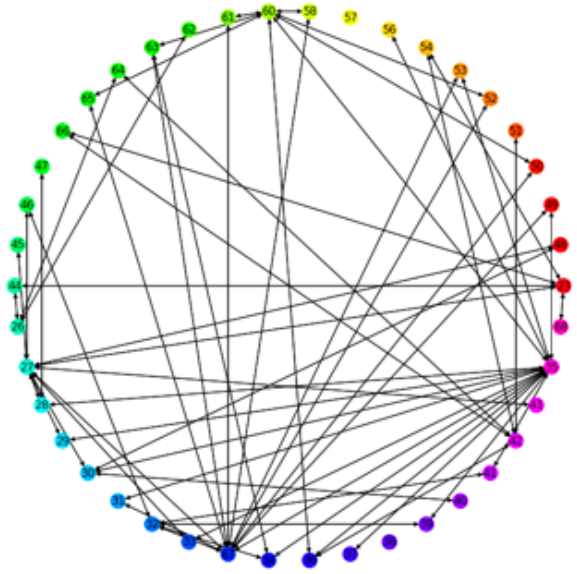
- Increasing sampling of ‘inconclusive’ relative free energies

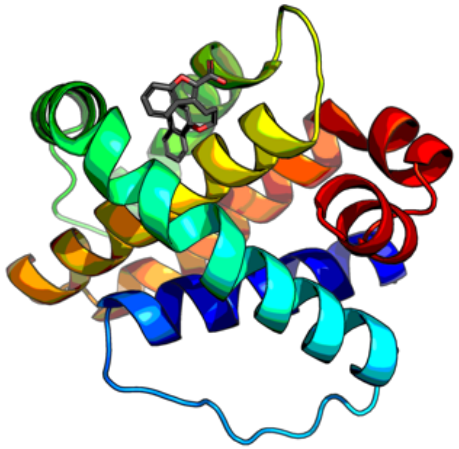
More sampling of relative free energy calculations close to zero

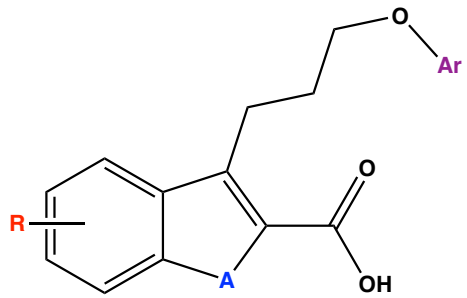


Bayesian bandits

- Schrodinger dataset
- 42 ligands

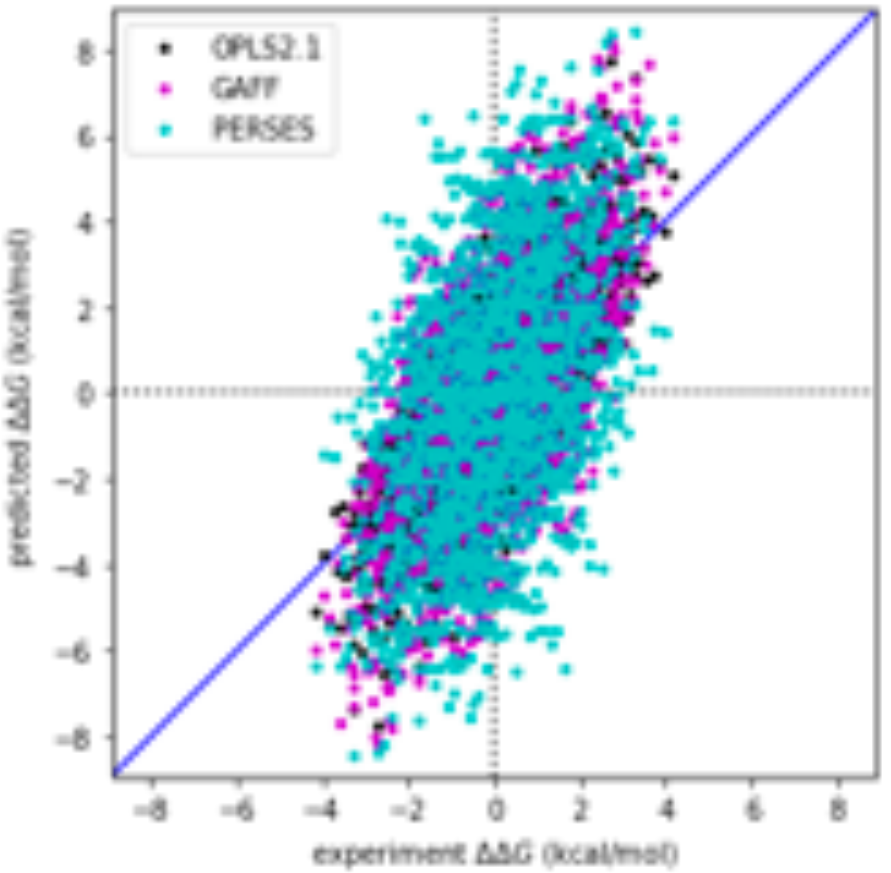




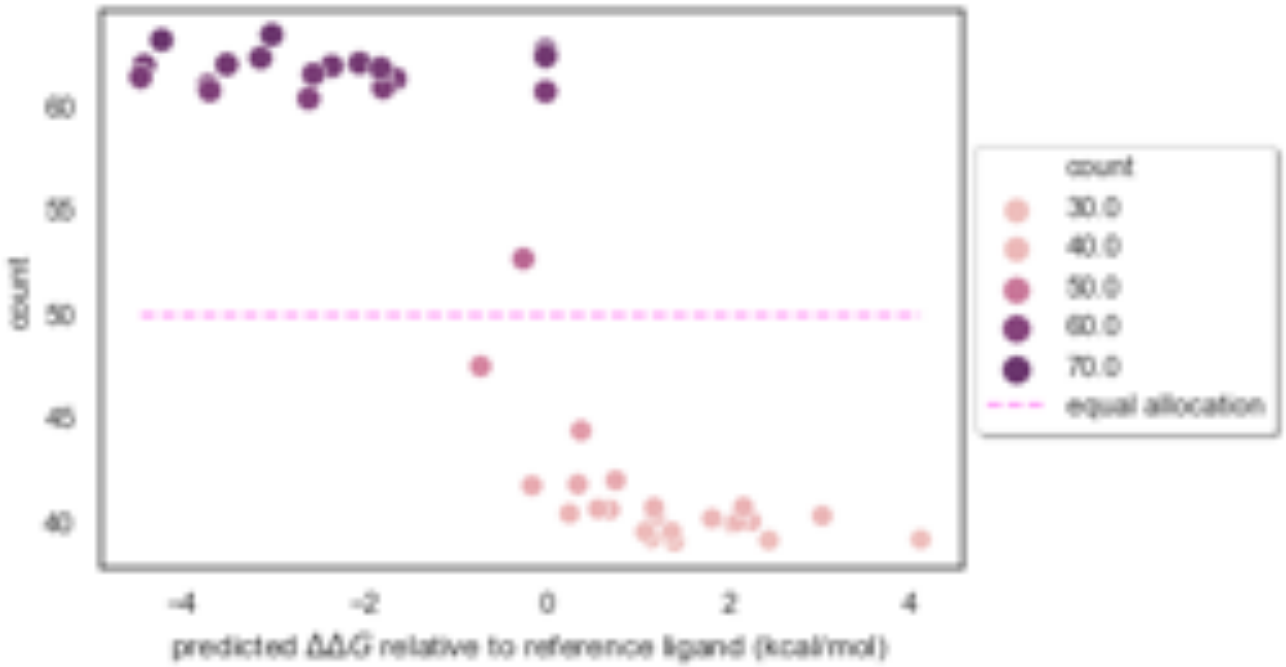


mcl1 all-pairs ΔΔG (N = 1722)

RMSE:	OPLS2.1	1.49	[95%:	1.44,	1.54] kcal/mol
RMSE:	GAFF	2.11	[95%:	2.05,	2.18] kcal/mol
RMSE:	PERSES	2.70	[95%:	2.62,	2.78] kcal/mol
MUE :	OPLS2.1	1.20	[95%:	1.16,	1.24] kcal/mol
MUE :	GAFF	1.68	[95%:	1.62,	1.74] kcal/mol
MUE :	PERSES	2.20	[95%:	2.12,	2.28] kcal/mol
R2 :	OPLS2.1	0.02	[95%:	-0.07,	0.10] kcal/mol
R2 :	GAFF	-0.97	[95%:	-1.15,	-0.80] kcal/mol
R2 :	PERSES	-2.20	[95%:	-2.50,	-1.92] kcal/mol
rho :	OPLS2.1	0.77	[95%:	0.75,	0.79] kcal/mol
rho :	GAFF	0.65	[95%:	0.62,	0.68] kcal/mol
rho :	PERSES	0.49	[95%:	0.46,	0.52] kcal/mol



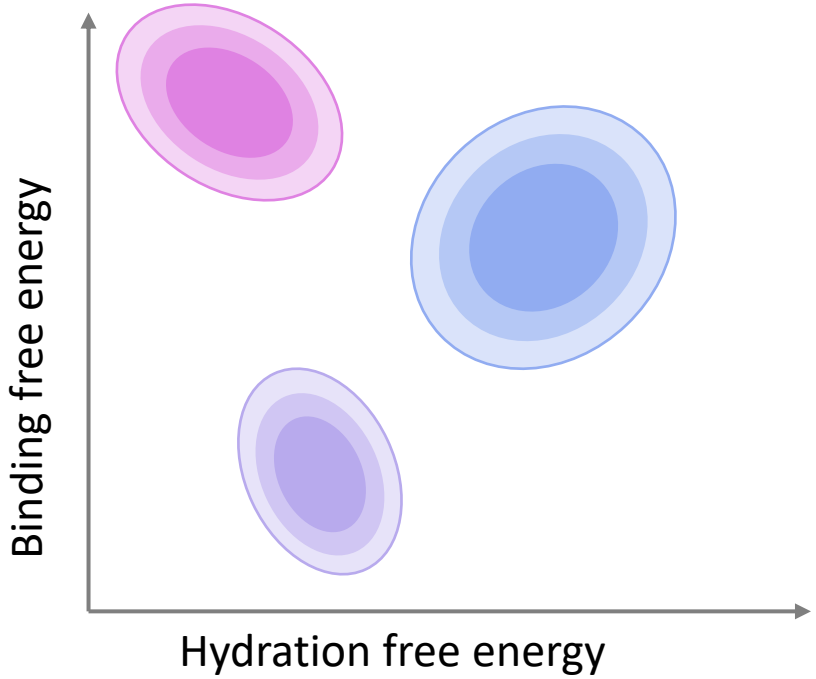
Bayesian bandits



Increased sampling of molecules with higher affinity to the reference.

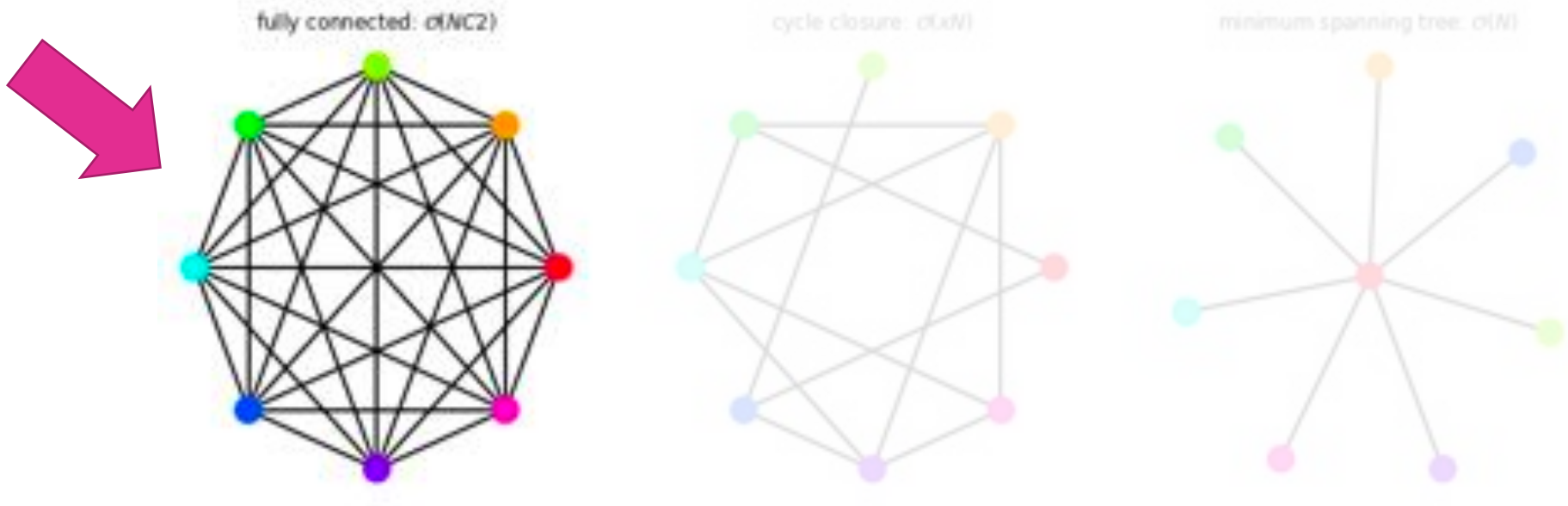
Summary – Part A

- Bayesian bandits can direct simulations towards features of interest:
  - High affinity ligands (solubility or binding)
  - Uncertain ligands
- Could reduce the computational time required to answer questions
- Multi-objective design



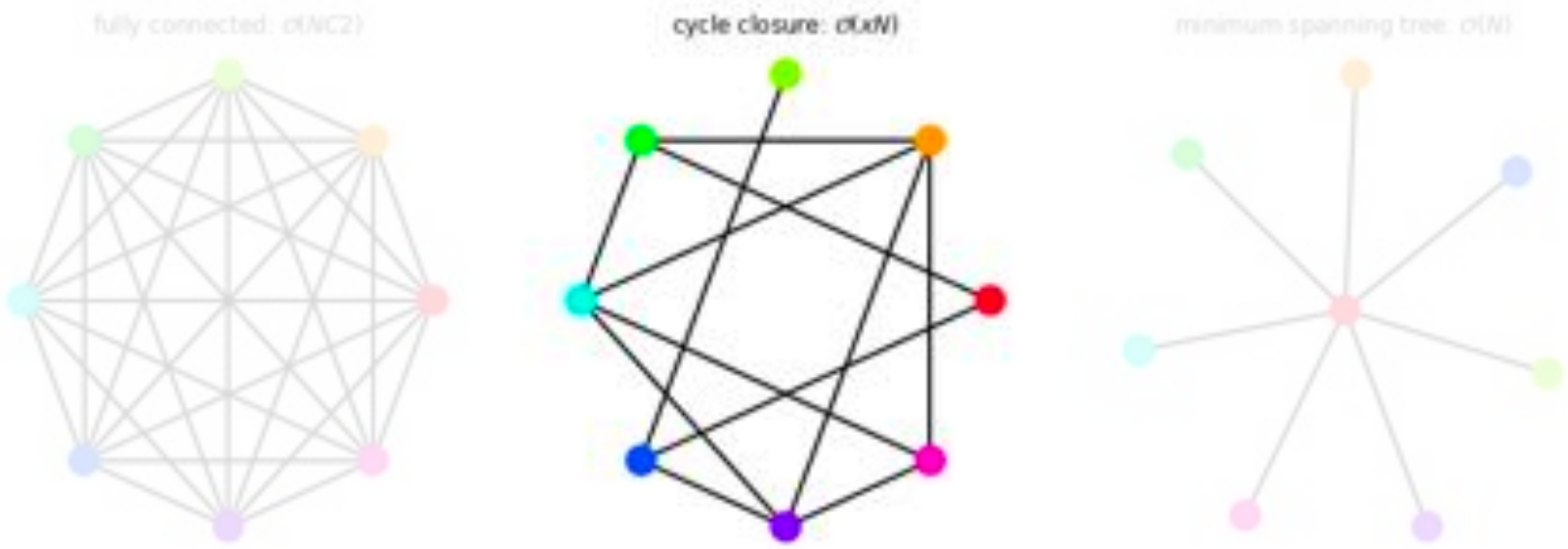
Part B – optimal map design

- 66 x 5 ns = 330 ns of simulation for hydration free energies of 12 molecules
- High effort
- Scales terribly



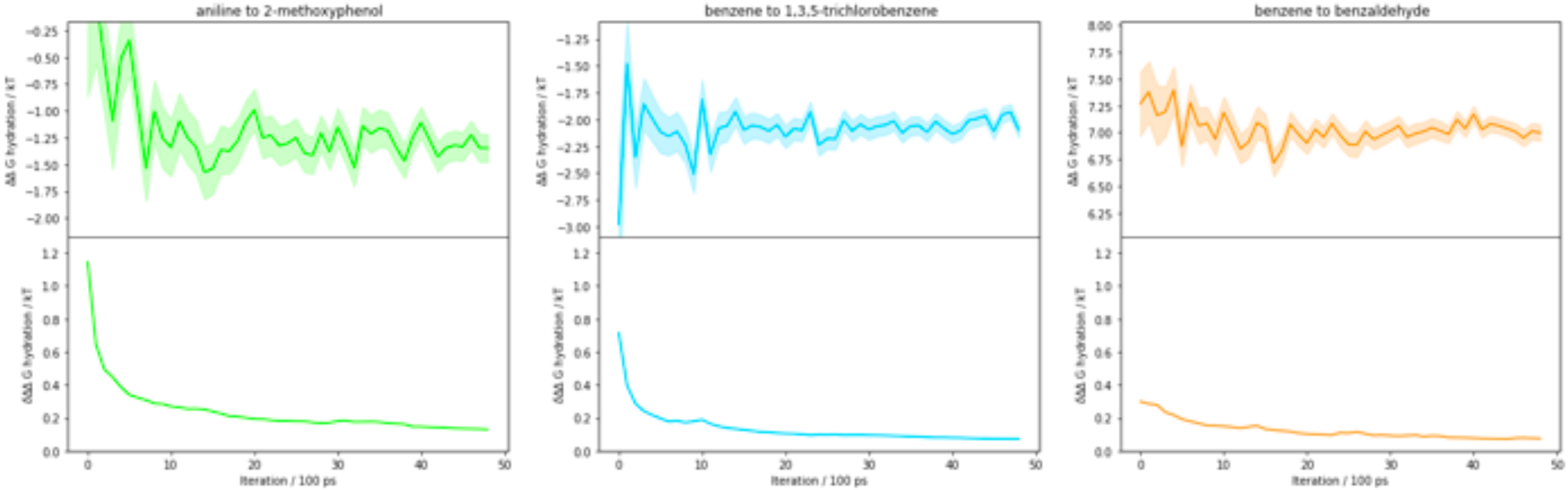
Results

- How can we best move to something lower effort?
- Which, and how many ‘edges’ should we use?
- The ‘best’ edges have the smallest variance, or highest efficiency



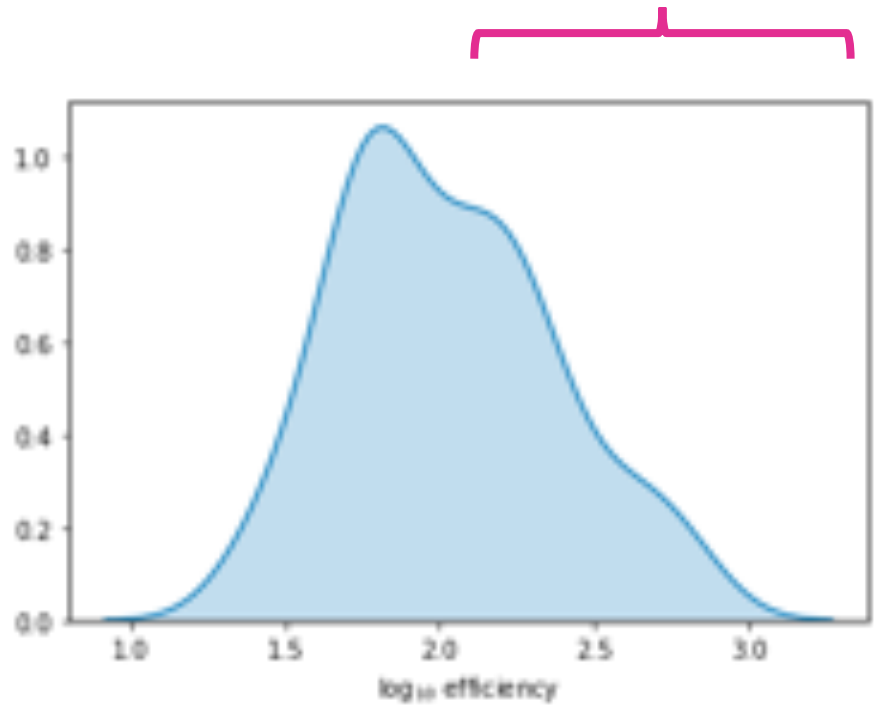
How best to simplify the graph?

- Not all edges are equal



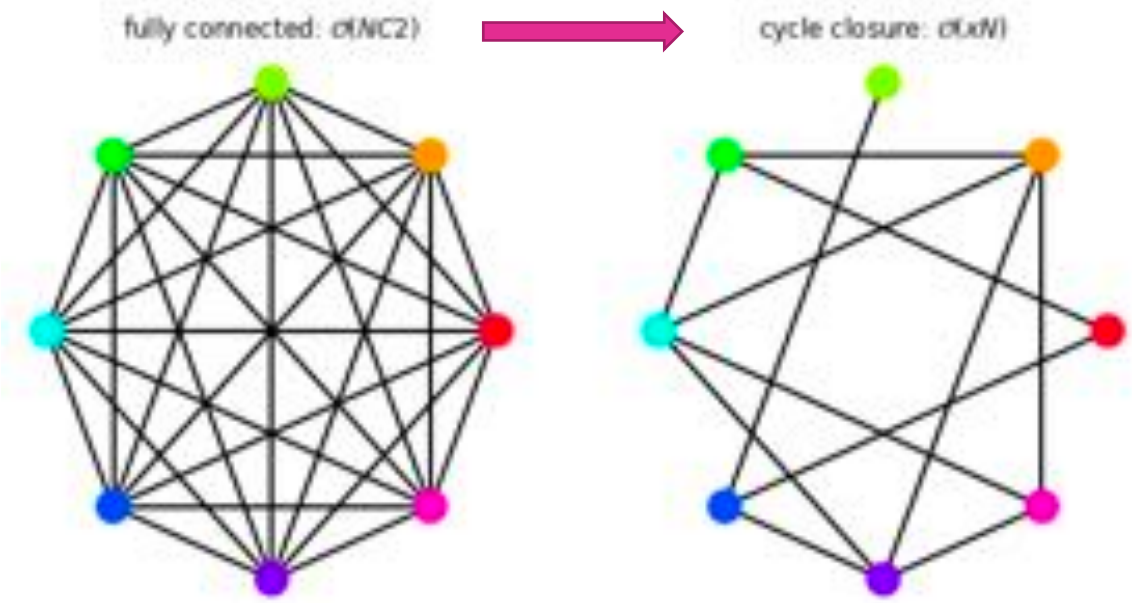
Results – efficiency

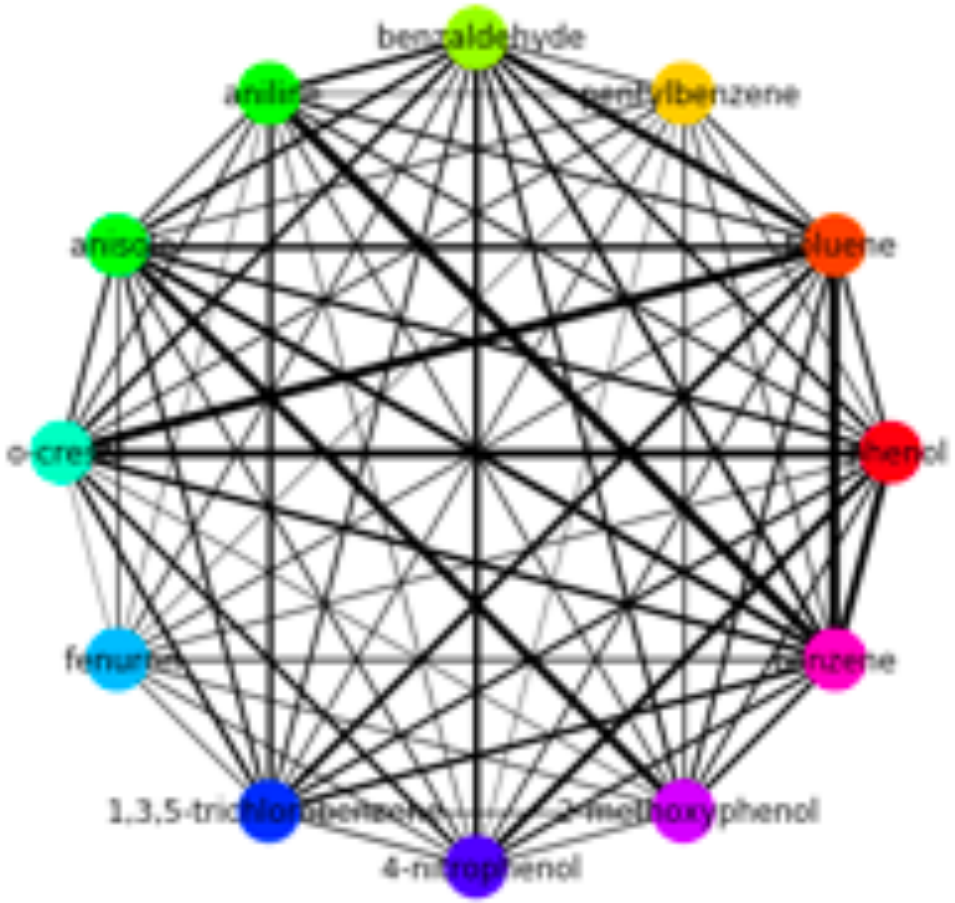
- Efficiency is the inverse of the variance
- $e_i = \sigma_i^{-2}$
- Doubling the efficiency halves the required simulation time



Results – efficiency

- How can we minimize the graph to fewer edges most effectively?





Thicker line = more efficient = better

Optimal graphs

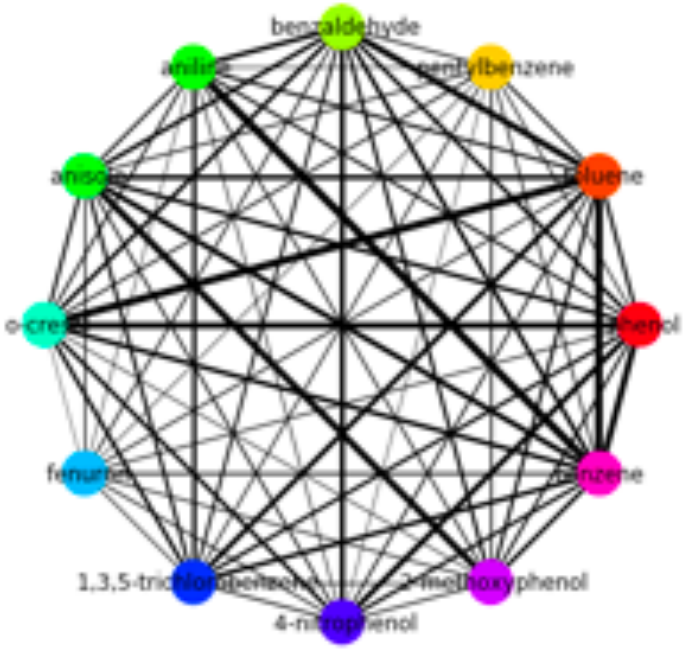
- Two preprints addressing this:
- "*Optimal measurement network of pairwise differences*"[1] **DiffNet**
- "*Optimal Designs of Pairwise Calculation: an Application to Free Energy Perturbation in Minimizing Prediction Variability*"[2]

- Choices in optimal -
  - A-optimal: minimizes variances relative to a single vertex
  - D-optimal: minimizes variances for all edges

1) Xu, Huafeng. "Optimal measurement network of pairwise differences." arXiv preprint arXiv:1906.08599 (2019).
2) Yang, Qingyi, et al. "Optimal Designs of Pairwise Calculation: an Application to Free Energy Perturbation in Minimizing Prediction Variability." ChemrXiv preprint arXiv:7965140.v2 (2019).

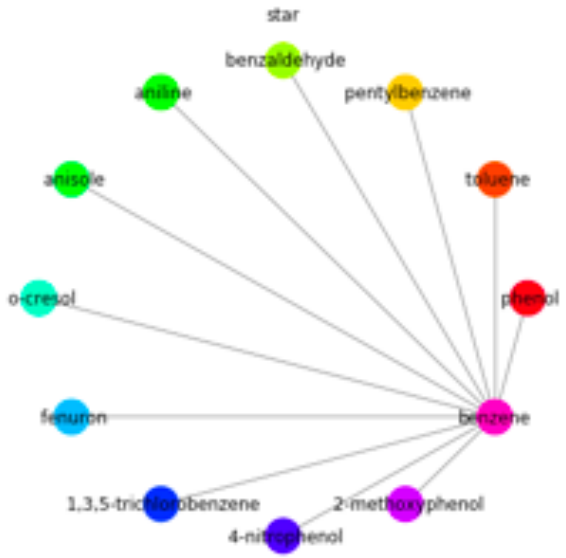
DiffNet

- N by N matrix of the statistical fluctuations of the simulations **C**
  - Where statistical fluctuation is $s_i = \sqrt{n_i \sigma_i^2}$

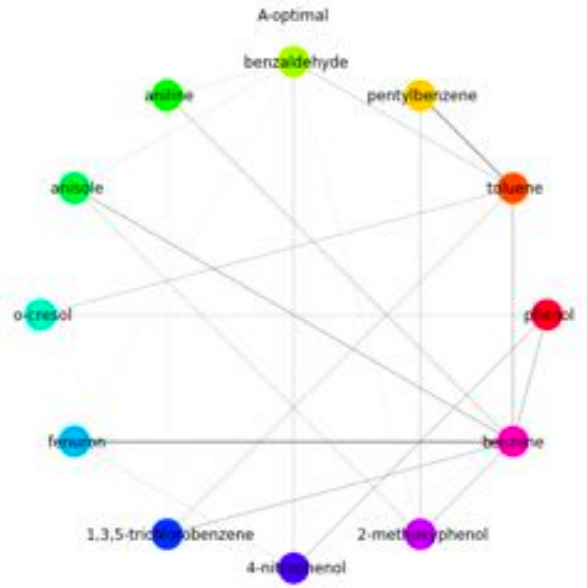


- Choices in optimal -
  - A-optimal: minimizes variances relative to a single vertex
    - Minimize trace(**C**)
  - D-optimal: minimizes variances for all edges
    - Minimize ln det(**C**)

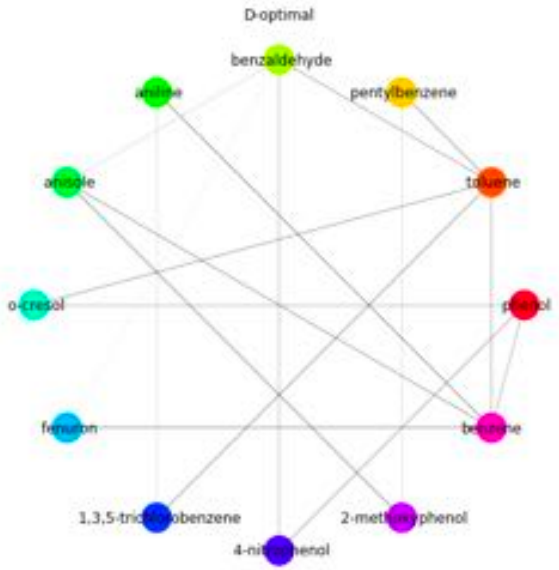
Results – DiffNet



D: -52.28
A: 0.13



D: -57.51
A: 0.07

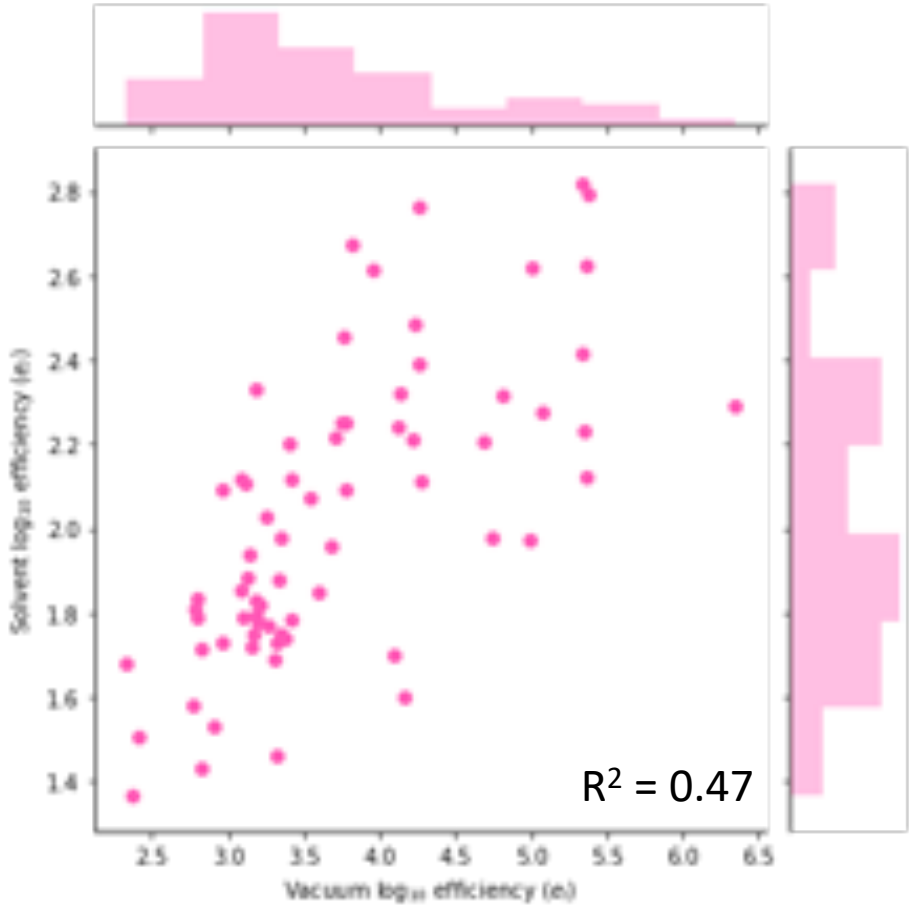


D: -59.19
A: 0.07

- These results are generated from the results… could we do this prospectively?
- Would need an estimate for the variance

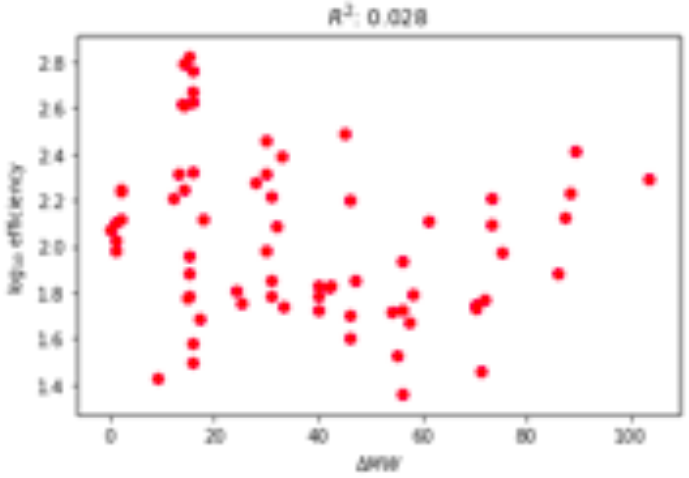
Estimating the variance *a priori*

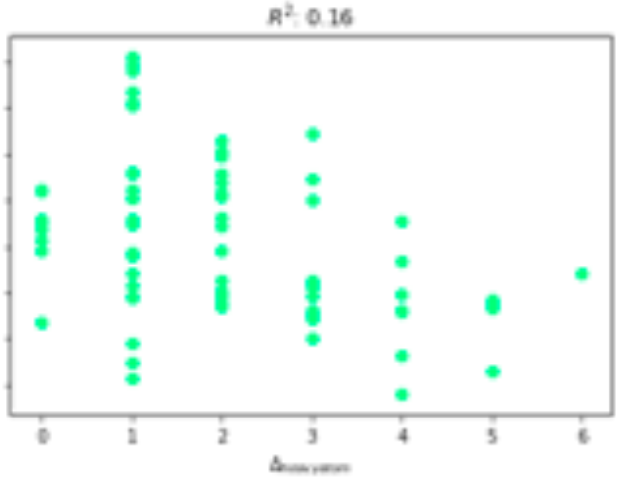
- Results from cheaper simulations?

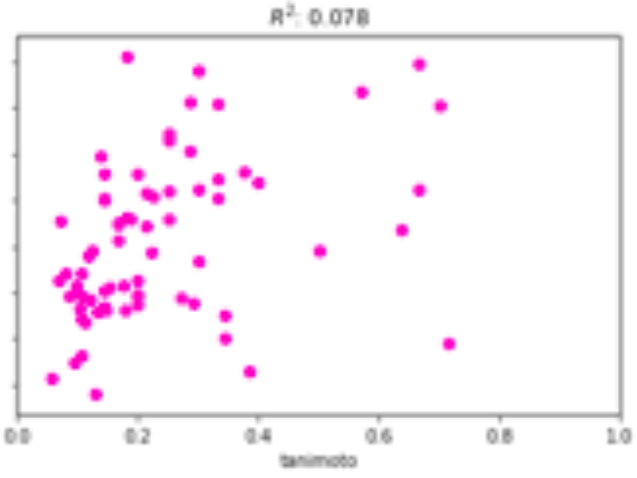


Estimating the variance *a priori*

- Chemical predictors?

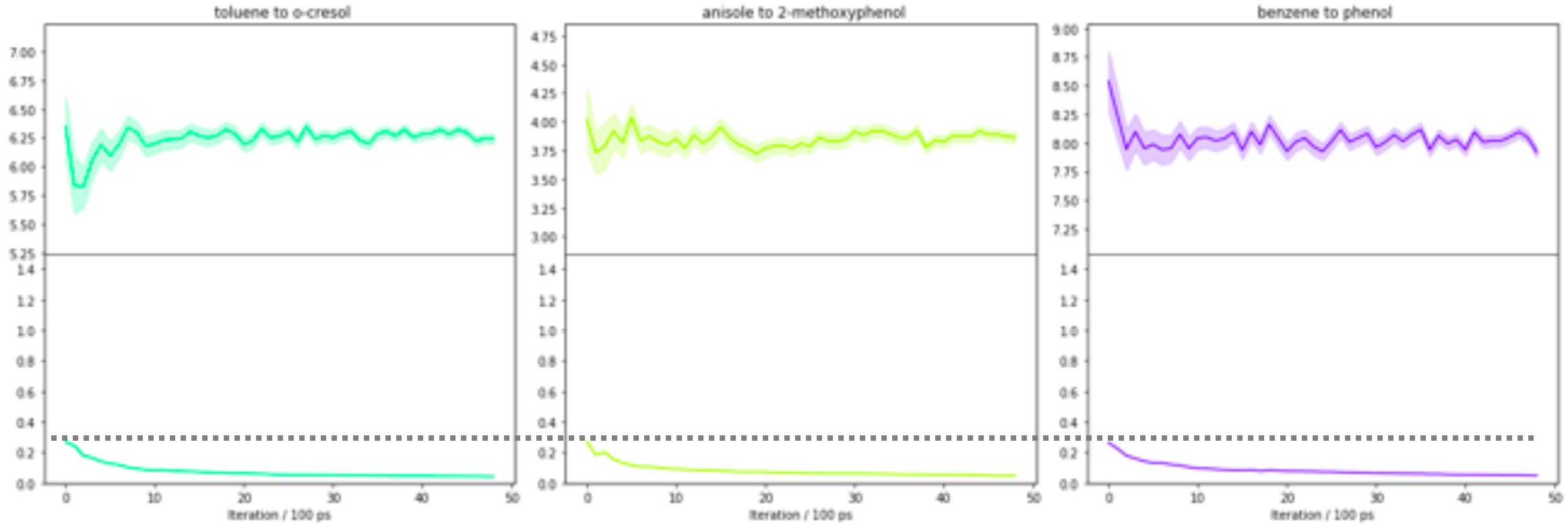




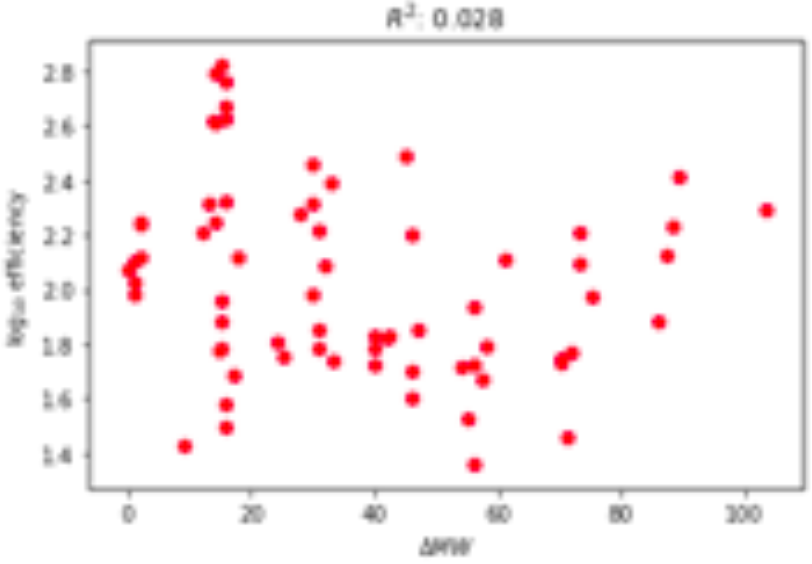


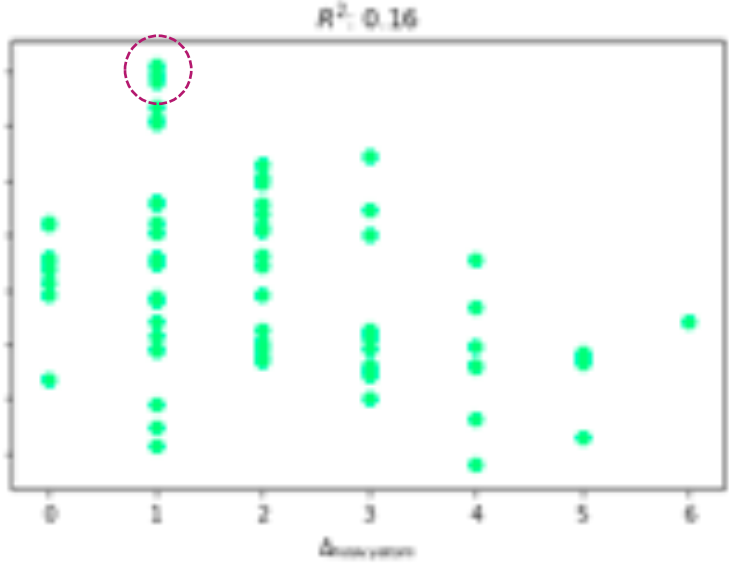
Estimating the variance *a priori*

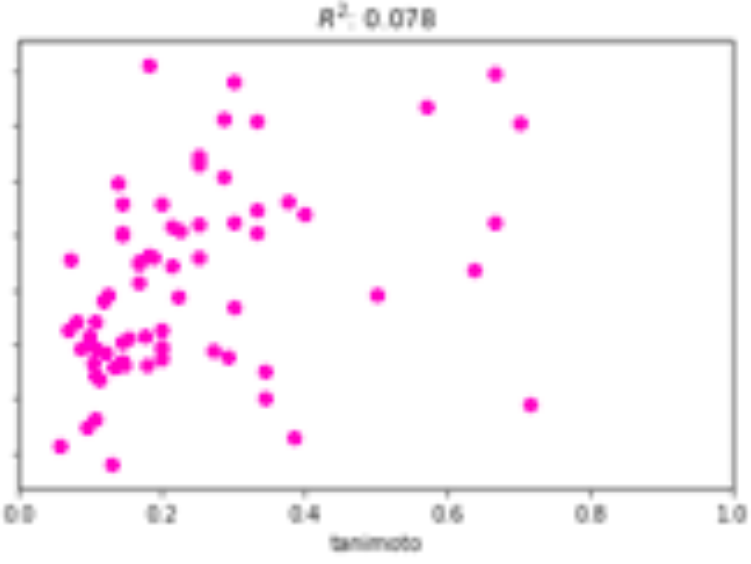
- Possibly this is something that could be learnt?



Results – similarity measures







NOTE: This is a very small dataset

Summary – Part B

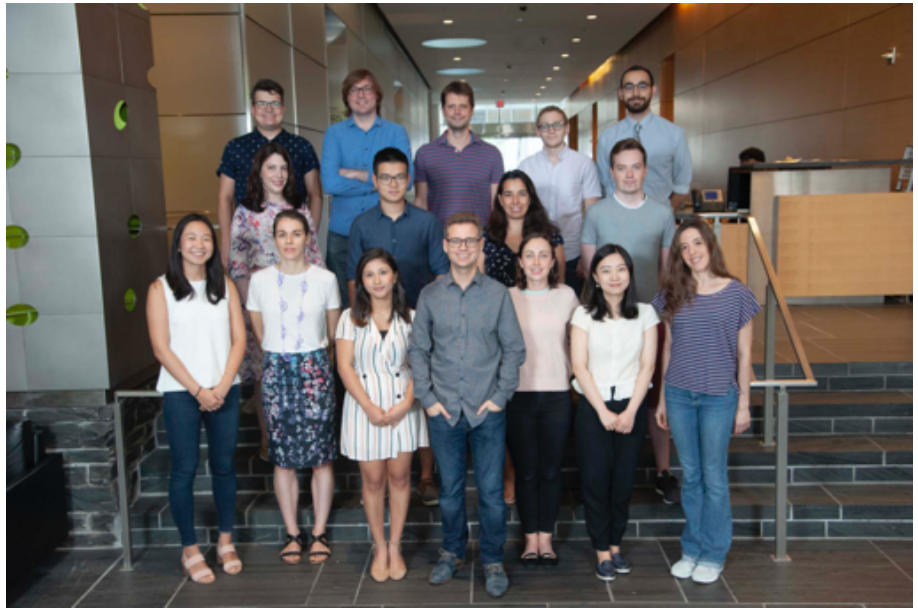
- By optimizing free energy calculations, we can minimize variance (error bars) for a given ‘amount’ of simulation
- Conversely, we could use less computer time to get error bars of a target *size*.
- Need a good method to predict the variance *a priori*
  - Vacuum variance?
  - Machine learning?
  - Updating on the fly?

Future Work

- Fully demonstrate this for a protein-ligand system
- Implement this *on-the-fly*
  - DASK for handling workflow
- Combining absolute and relative free energies optimally
  - (and other types of relative free energies)
- Optimize perturbations via protocols
- Improving the predictions of variance
- More adaptive sampling – Bayesian bandits for multi-optimization design

Acknowledgements

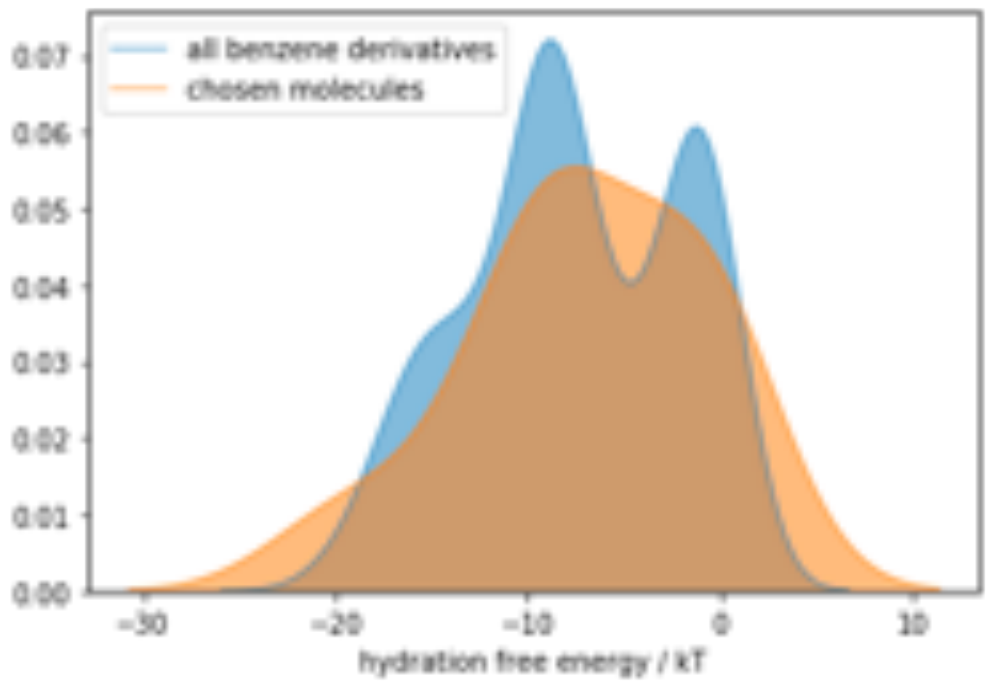
- John Chodera – *supervision, ideas, money*
- Dominic Rufa – *help on this project*
- Patrick Grinaway – *original perses developer*
- Josh Fass – *discussions on adaptive sampling*
- Chodera lab – *for everything*
- MSKCC – *resources*
- MolSSI – *funding*

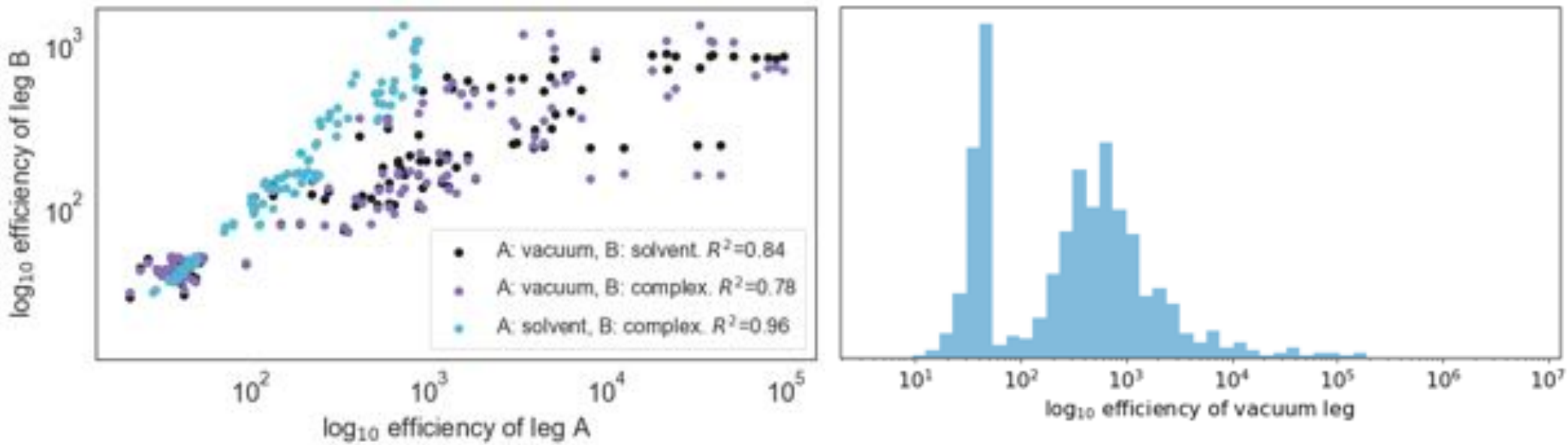


Email: Hannah.brucemacdonald@choderalab.org
Twitter: @hannahbruce

Memorial Sloan Kettering Cancer Center

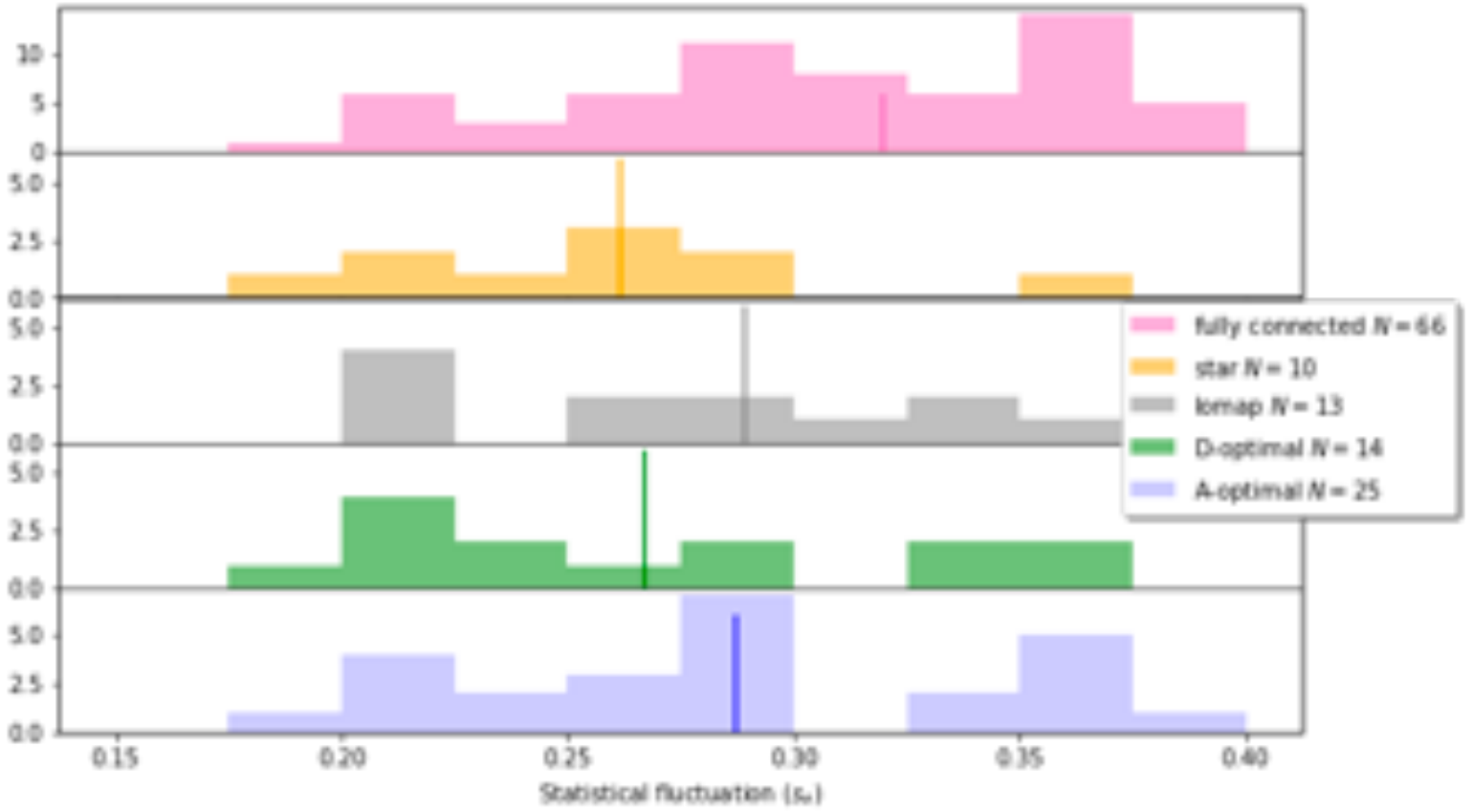
MolSSI


all benzene derivatives
chosen molecules
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
−30
−20
−10
0
10
hydration free energy / kT

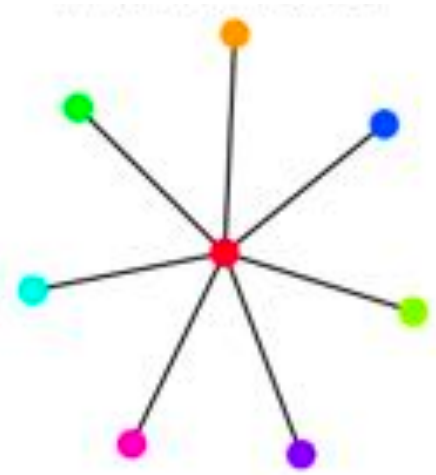

log10 efficiency of leg B
10^3
10^2
A: vacuum, B: solvent. R^2=0.84
A: vacuum, B: complex. R^2=0.78
A: solvent, B: complex. R^2=0.96
10^2
10^3
10^4
10^5
log10 efficiency of leg A
10^1
10^2
10^3
10^4
10^5
10^6
10^7
log10 efficiency of vacuum leg

Results – DiffNet

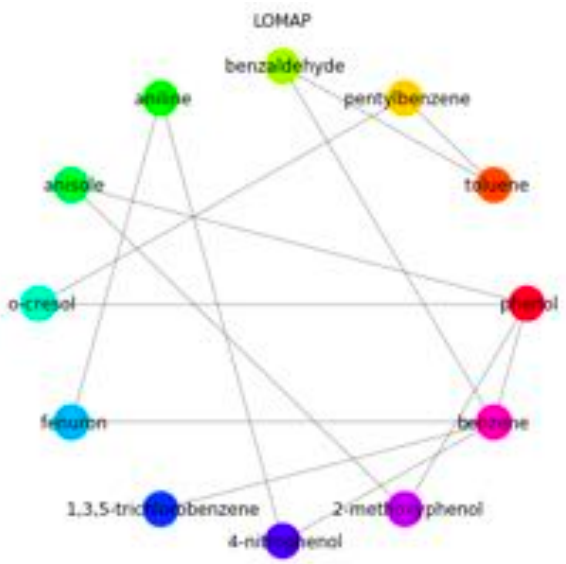
- Are these graphs better?



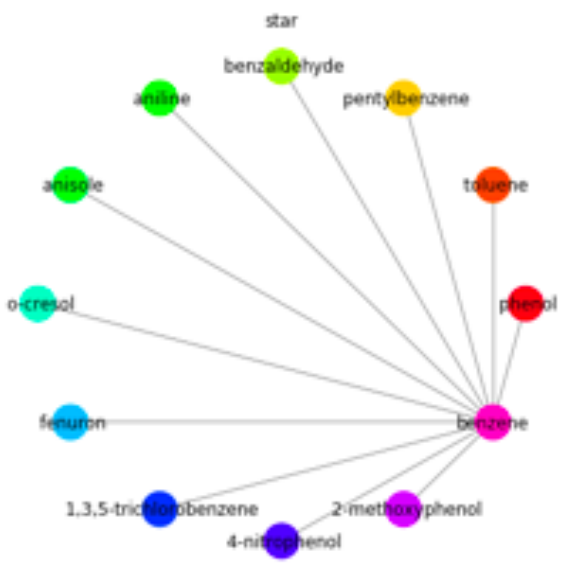
star



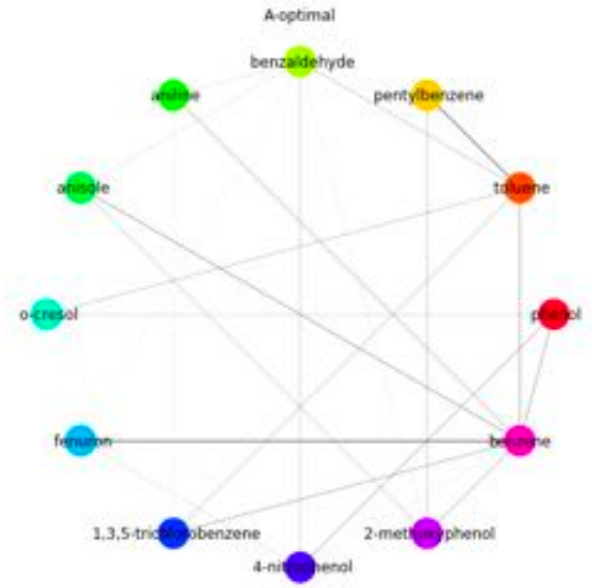
Results – DiffNet



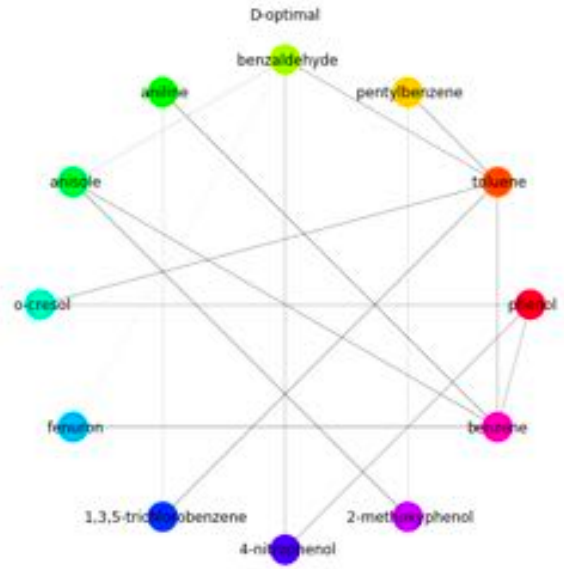
D: -57.93
A: 0.09



D: -52.28
A: 0.13



D: -57.51
A: 0.07



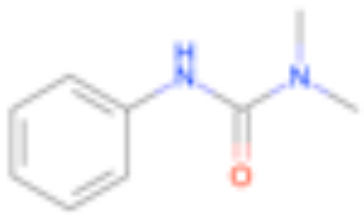
D: -59.19
A: 0.07

- These results are generated from the results… could we do this prospectively?
- Would need an estimate for the variance

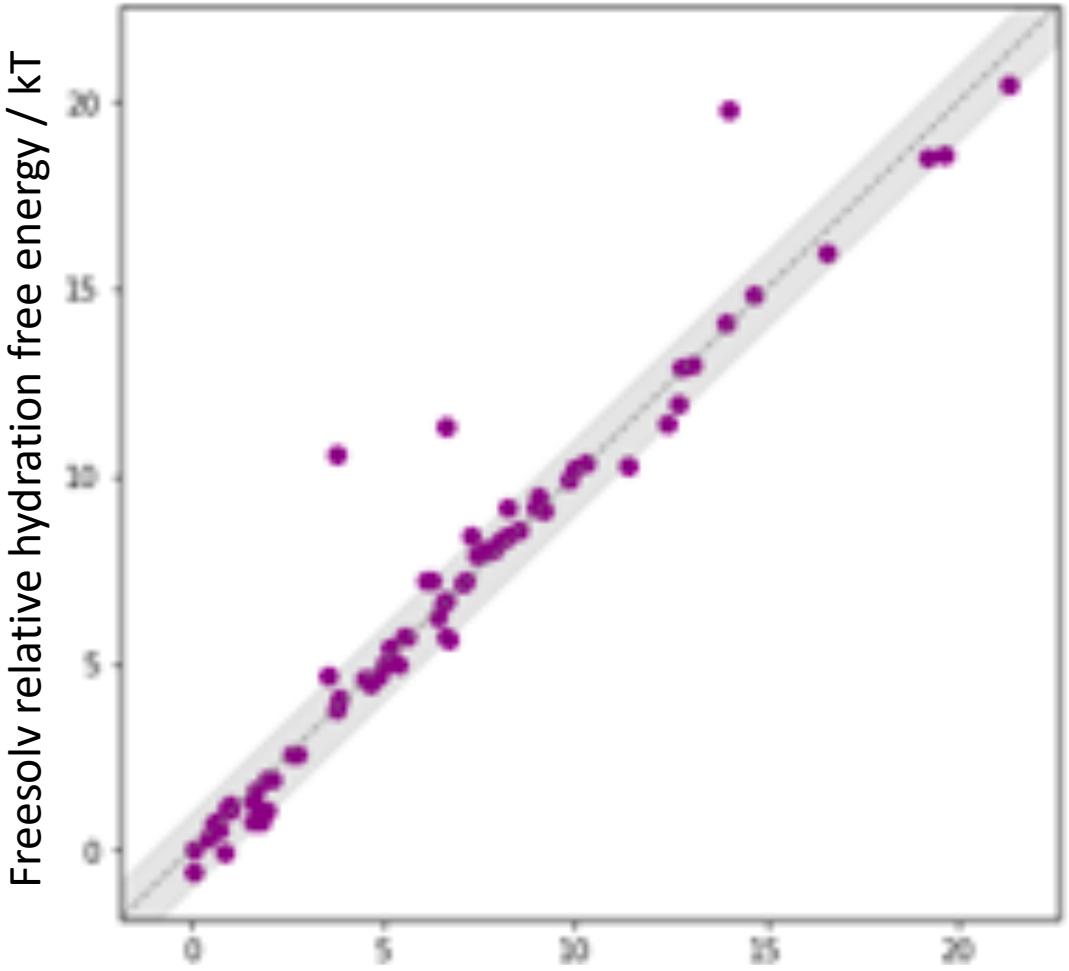
Results – validation

- 12 benzene derivatives
- Error bars *are* drawn
- 3 outliers (all involve fenuron)
  - Closer to experiment for all 3

8- fenuron

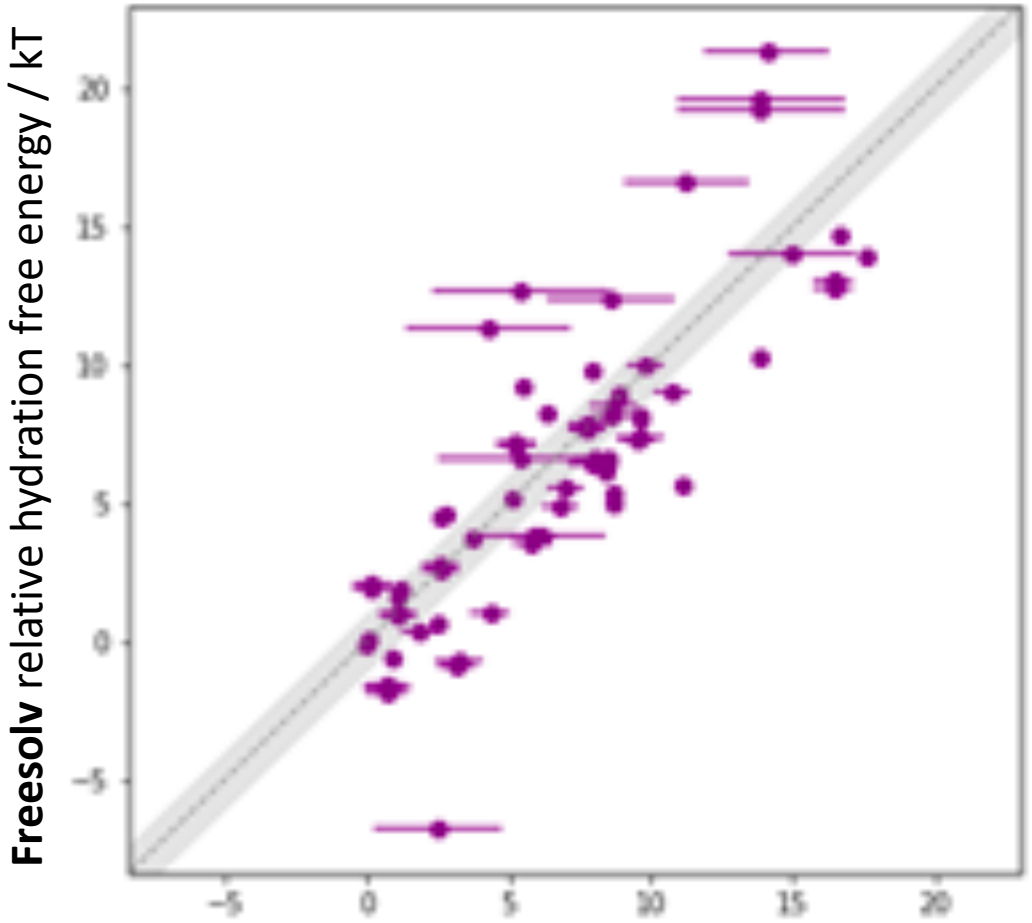


RMSE : 1.32 [95%: 0.52, 1.90] kcal/mol
MUE : 0.61 [95%: 0.36, 0.90] kcal/mol
R^2 : 0.93 [95%: 0.82, 0.99] kcal/mol
ρ : 0.97 [95%: 0.93, 0.99] kcal/mol



Results – benchmarking

```
RMSE :      3.02 [95%:  2.46,  3.61] kcal/mol
MUE  :      2.28 [95%:  1.82,  2.79] kcal/mol
R²   :      0.57 [95%:  0.30,  0.74] kcal/mol
ρ    :      0.84 [95%:  0.77,  0.89] kcal/mol
```




```
RMSE :      2.90 [95%:  2.36,  3.40] kcal/mol
MUE  :      2.22 [95%:  1.79,  2.70] kcal/mol
R²   :      0.60 [95%:  0.37,  0.73] kcal/mol
ρ    :      0.85 [95%:  0.78,  0.90] kcal/mol
```