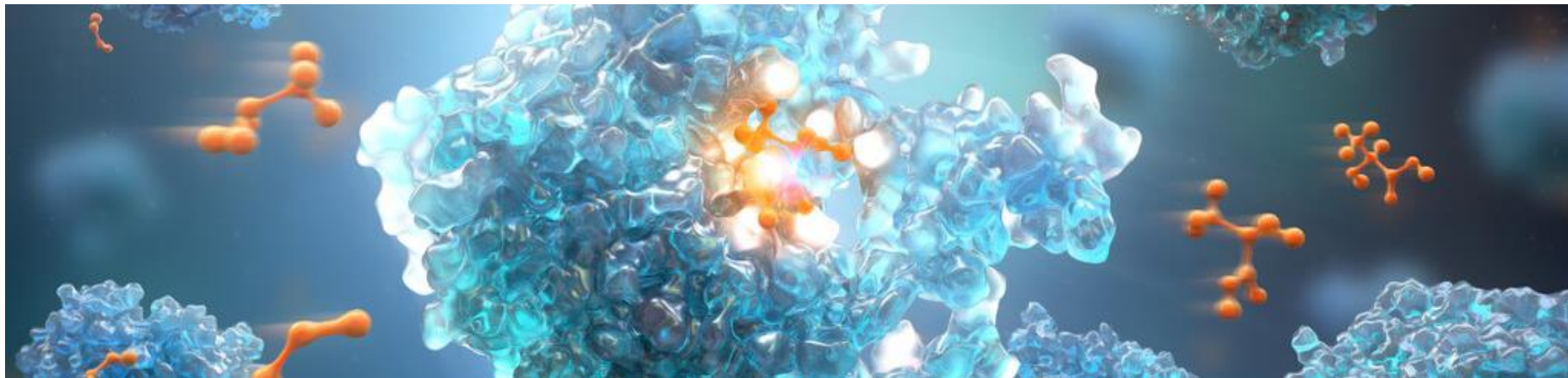


Modelling Macrocycles with Machine Learning and Molecular Dynamics

Dr Graeme Robb, AstraZeneca, Oncology R&D
UKQSAR Autumn Meeting, Sygnature, Nottingham

26th September 2019



Why it pays to be organised

The entropy cost of dis-organisation in ligand binding

Simulating the system

A case study in preorganisation with macrocycles

Step beyond

Machine learning to find hidden features and improve prediction

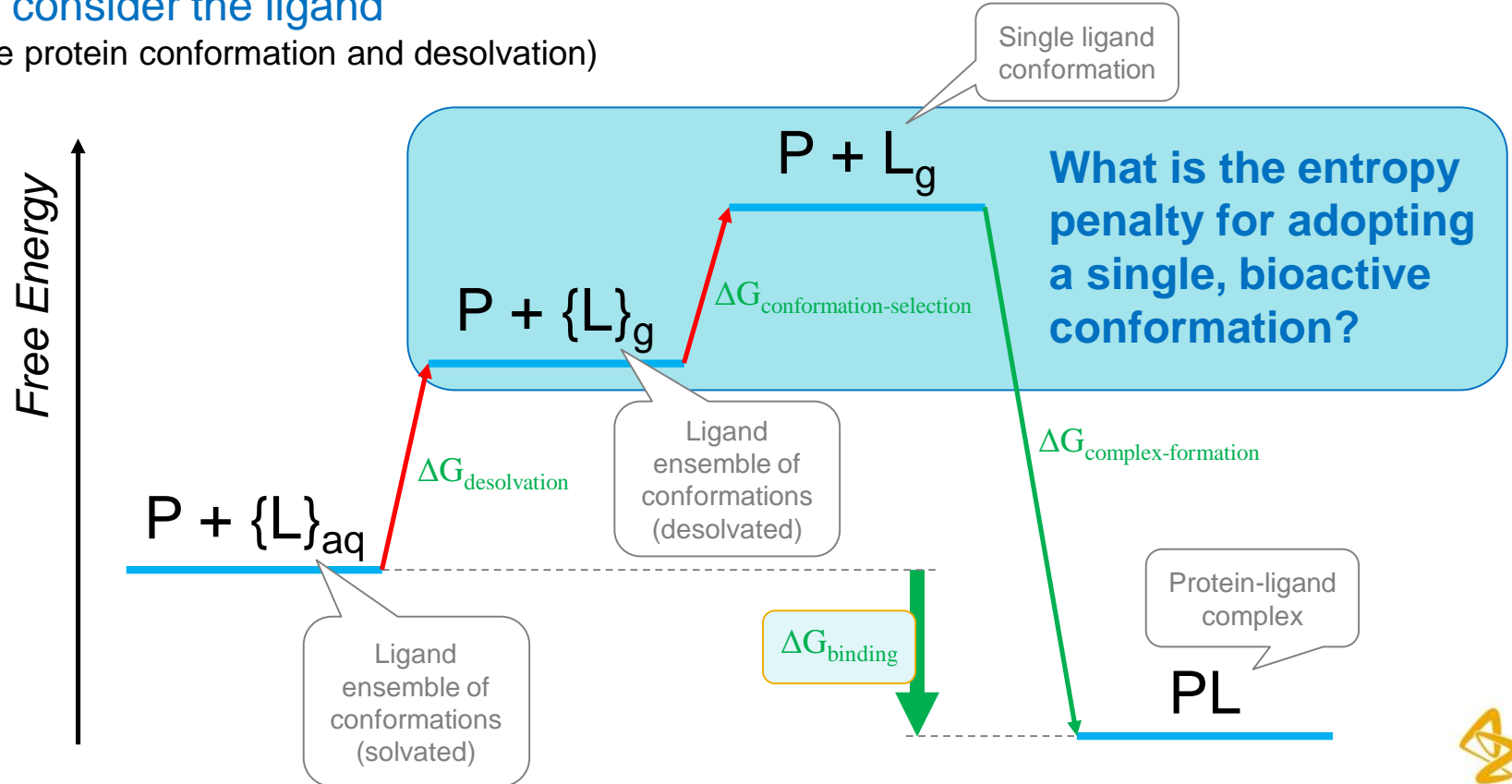
Validation

Is it robust? Is there a more simple model?



Ligand binding free energy

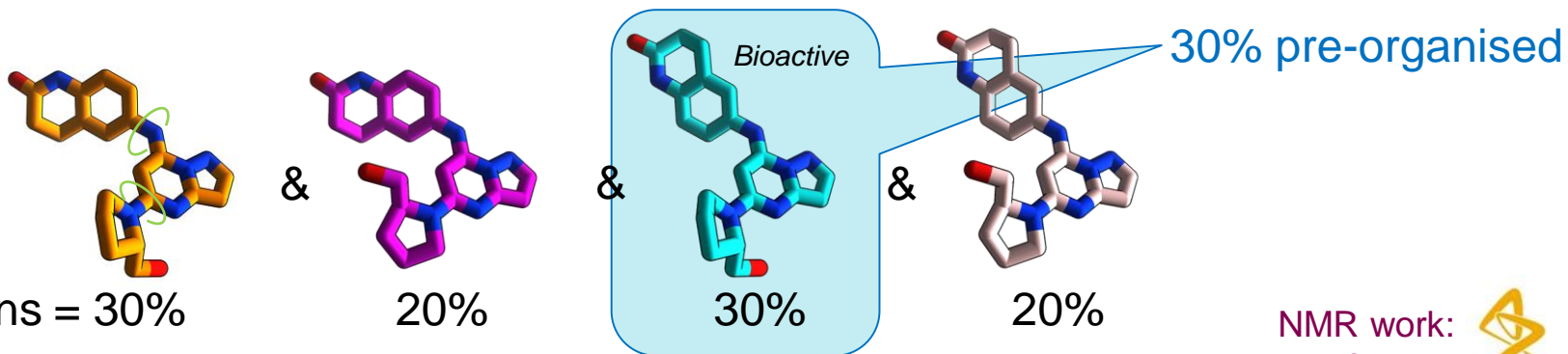
- Only consider the ligand
(ignore protein conformation and desolvation)



Pre-organisation and entropy

- Quantifying changes in entropy is hard
- Quantifying pre-organisation is easier
 - Pre-organisation varies as $1/\Delta\text{Entropy}$
 - Degree of pre-organisation is proportion of ensemble already in the bioactive conformation

e.g. A compound has 4 major conformers in solution, by NMR



Predicting preorganisation

- Paul Dirac (1929)

The underlying physical laws necessary for the mathematical theory of a large part of physics and the **whole of chemistry** are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble. It therefore becomes desirable that **approximate practical methods** of applying quantum mechanics should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computation.

- Ligand ensemble calculation: two philosophies

(i) Follow the Rules

e.g. Omega (QM)

Very quick (seconds)

- Inaccurate energies*
- Non-Boltzmann population

(ii) Physical simulation

e.g. Molecular Dynamics

- Much slower (hours)
- Accurate forces*
- Boltzmann population (if sampled sufficiently)

Rule-based methods fail for macrocycles (*i.e.* compounds with rings >10 atoms)

* QM-parameterised forcefields, do not give accurate energies, but do correctly approximate potential energy surfaces, especially around the minima.



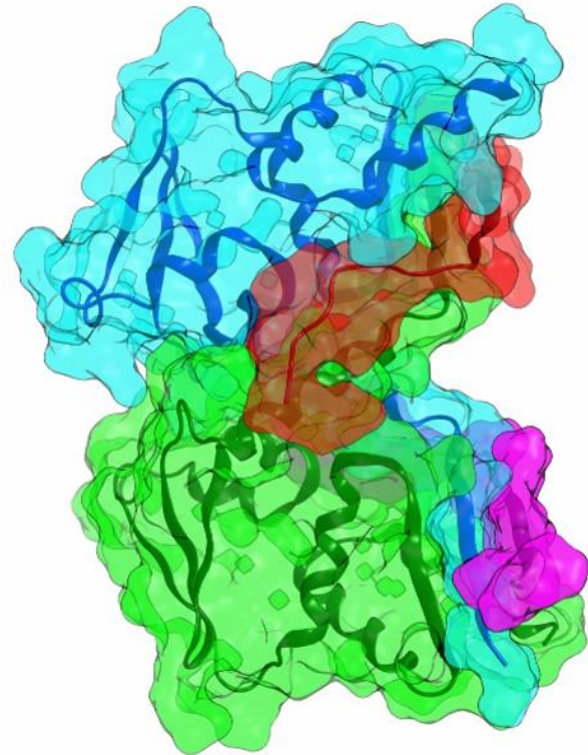
Case Study: BCL6 (B-Cell Lymphoma 6 Protein)

- Transcription repressor implicated in cancer (Diffuse Large B-cell lymphoma)
- Functional activation occurs through a peptide co-repressor binding
- Inactivation can be achieved by inhibition of the protein-protein interaction

“The BCL6 proto-oncogene suppresses p53 expression in germinal-centre B cells”

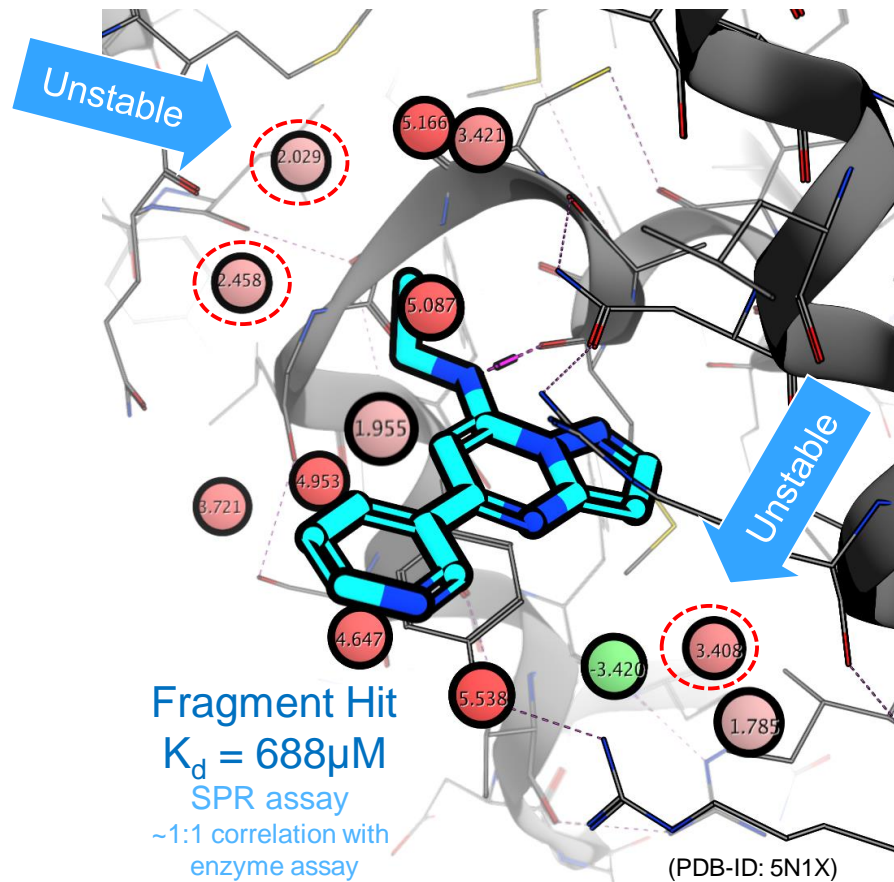
Phan & Dalla-Favera, *Nature*, 432 (7017), 635, 2004

BTB homodimer (PDB 1R2B)



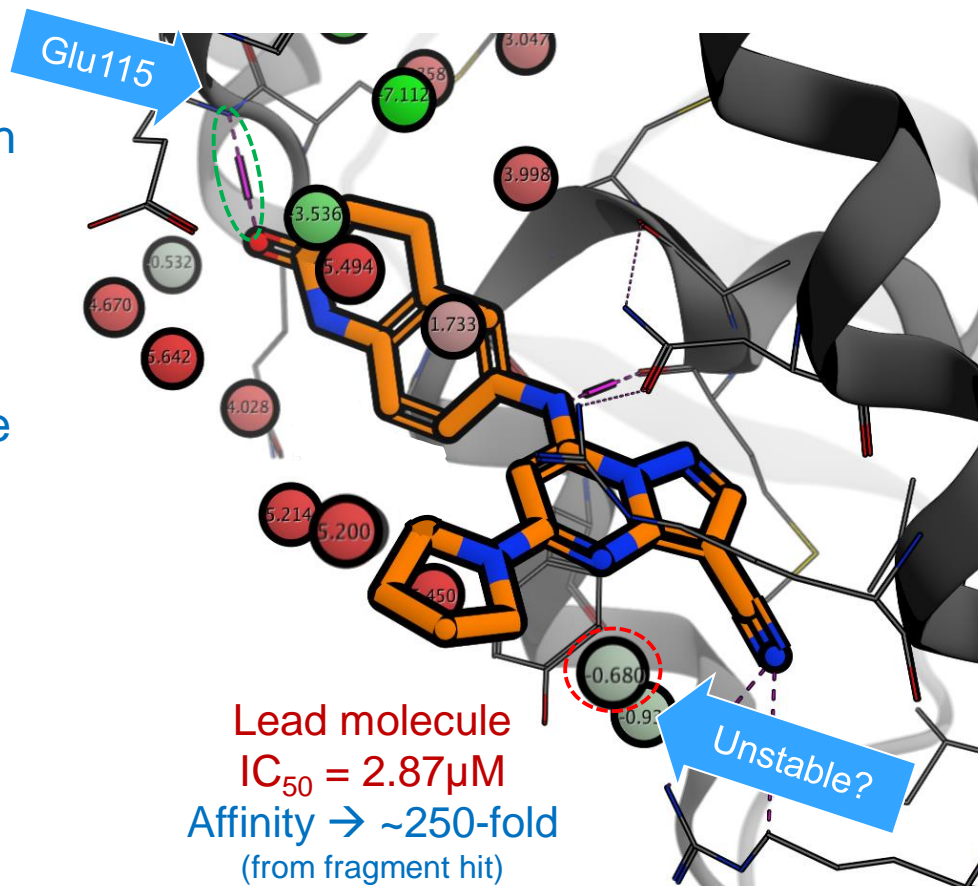
BCL6: Structure-based design, from Hit to Lead

- Identified region of unstable solvent with potential for protein interactions
 - Initiated library chemistry



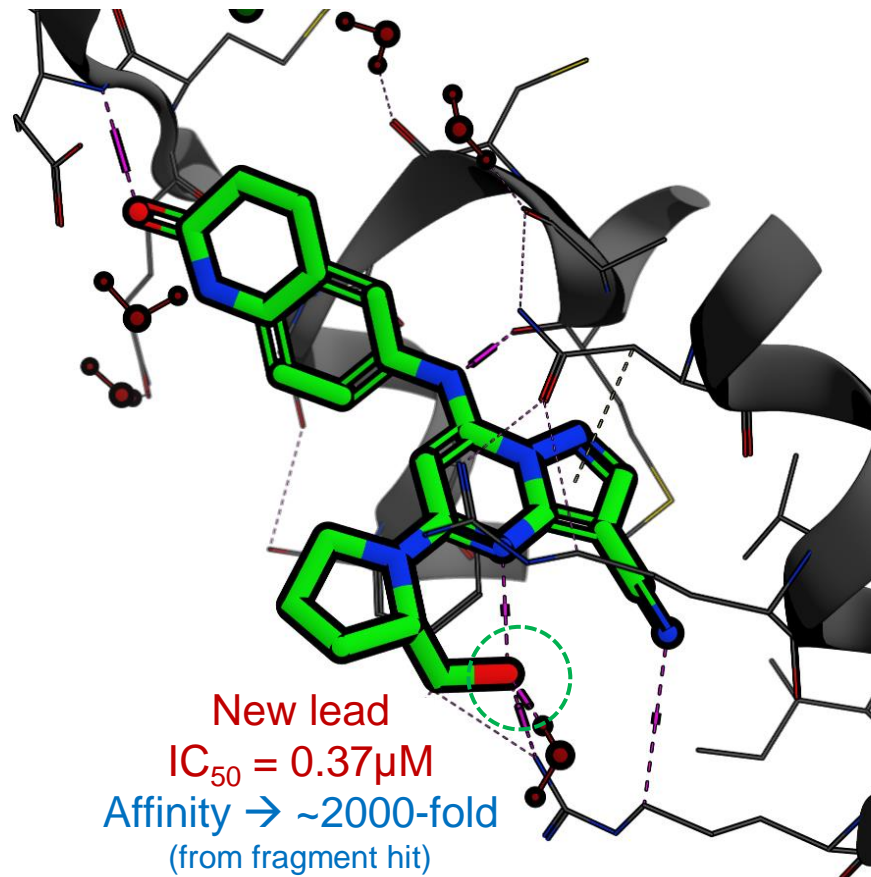
BCL6: Structure-based design, from Hit to Lead

- Identified region of unstable solvent with potential for protein interactions
 - Initiated library chemistry
 - Interaction with Glu115
- Additional meta-stable solvent molecule
 - Careful to replicate H-bonding



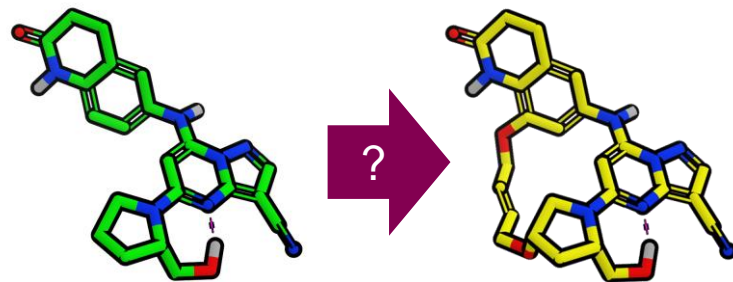
BCL6: Structure-based design, from Hit to Lead

- Identified region of unstable solvent with potential for protein interactions
 - Initiated library chemistry
 - Interaction with Glu115
- Additional meta-stable solvent molecule
 - Careful to replicate H-bonding
 - Interaction with Arg28
- Successful strategy, but exhausted
 - **Not potent enough**
 - **NMR shows only 30% pre-organised**

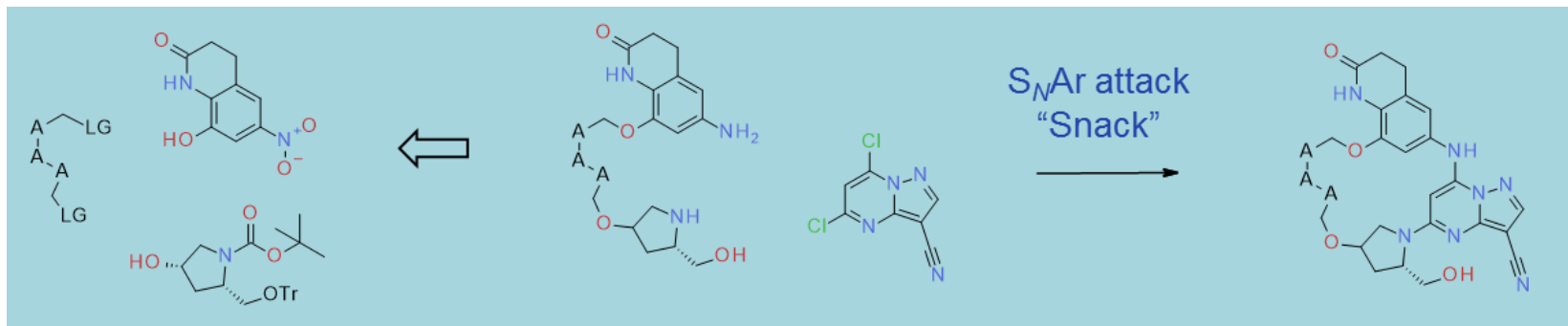


Constrain the shape to bioactive only

- Link the two sides of the molecule to rigidify
– Macrocycle!



- But complex chemistry, can we predict the best?

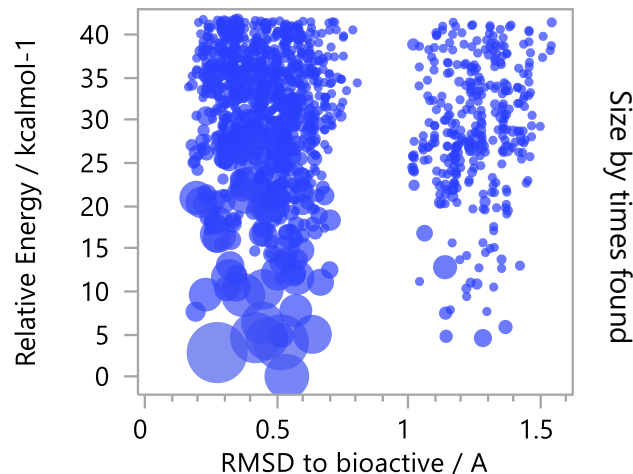


Piotr Raubo

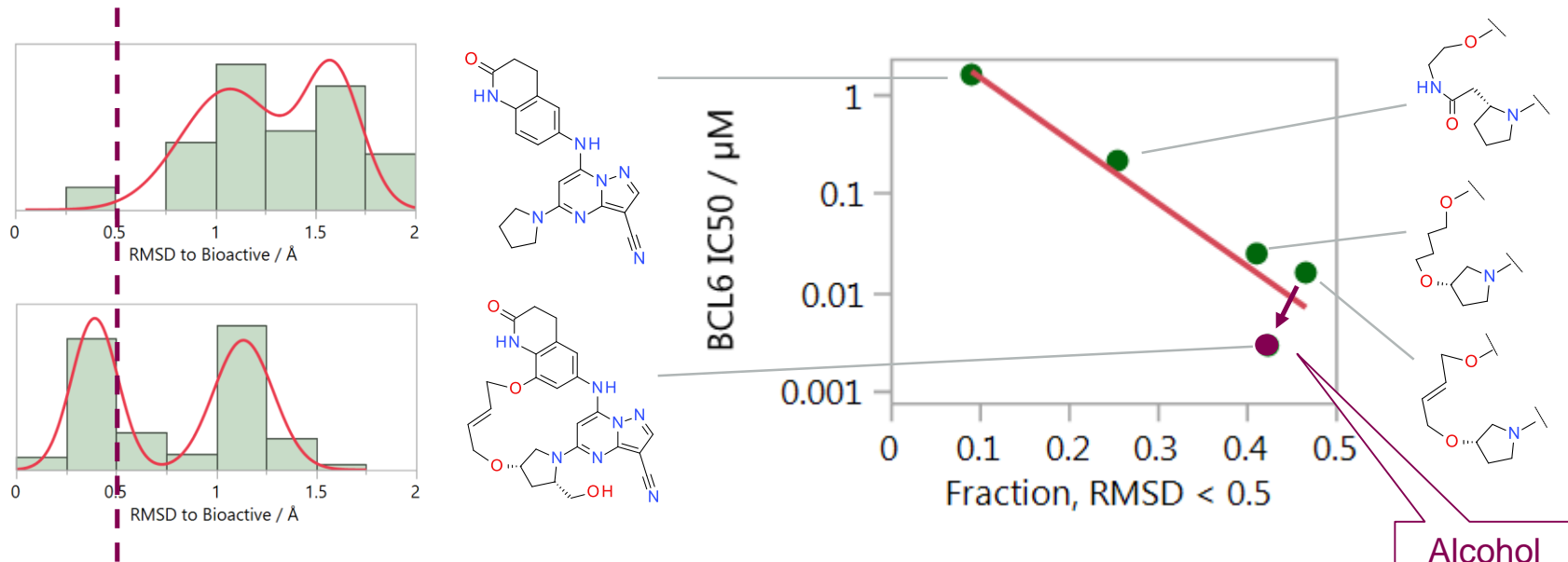


Macrocycle conformation prediction

- Accurate populations requires Molecular Dynamics (MD)
 - Macrocycles present a specific problem of inefficient/incomplete sampling
- Mixed Monte Carlo – Molecular Dynamics (MCMD)
 - 10,000 starts from Monte Carlo
 - 10,000 simulation/equilibration steps
 - Record unique conformations (RMSD < 0.01Å)
- What fraction of the conformational population is like the bioactive conformation?
 - And how similar to bioactive does it need to be?



Macrocycle conformation predicts affinity



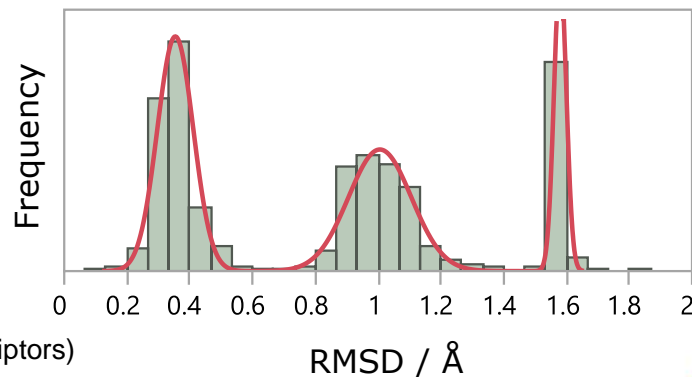
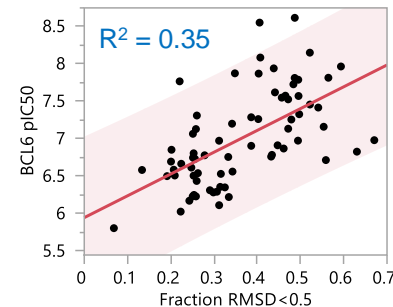
- Fraction with RMSD < 0.5Å (to bioactive) correlates with potency

Discovery of Pyrazolo[1,5-a]pyrimidine B-Cell Lymphoma 6 (BCL6) Binders and Optimization to High Affinity Macrocyclic Inhibitors. McCoull et al, *J. Med. Chem.* 2017, 6010, 4386-4402



Learning the features of an ensemble

- Can we improve upon our initial model?
 - Based on physical understanding, but empirical and simple
 - Shows poorer accuracy for more complex linkers
- Machine learning can discover a more complex relationship
 - Many features of conformational ensemble will determine potency
- What are the features of our ensemble?
 - Deconvolution of RMSD distribution into cluster contributions
 - In practise, binning the distribution at 0.1\AA gave the best descriptor set
(representative while avoiding the overfitting problem with excessive descriptors)



Machine Learning

- Anthony Nicholls, CEO OpenEye (2019)
“You think you’re using AI? You’re just making better curve-fitters.”
- Three major types of machine learning
 - Supervised Learning
 - Fit a function to inputs in order to predict output,
 - e.g. linear regression, neural nets and everything we call QSAR
 - Unsupervised Learning
 - Fit a function to organise your data, without prior output labels
 - e.g. clustering, rule-mining
 - Active Learning
 - Several rounds of supervised learning where the functional form is improved by the predictions and errors from previous rounds
 - *i.e.* the sexy stuff

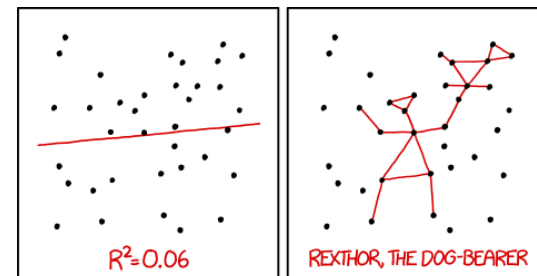


Need more
data!
(a lot more)



Methodology and Validation

- **The dataset:** 72 acyclic and macrocyclic compounds in two series
- **Methodology:** - too few data for many methods, keep it simple, avoid overfitting
 - Multi-Linear Regression (MLR)
 - Principle Component Regression (PCA-MLR)
 - Partial Least Squares (PLS)
 - Bayesian Neural Net (BNN)
- **Assessment:** K-fold cross validation
 - Average performance across 6 folds in the data



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.



Model Assessment

Method	Training Set mean	
	R^2	RMSE
MLR	0.79	0.31
PCA-MLR	0.71	0.35
PLS	0.61	0.40
BNN	0.71	0.34



Model Assessment

Method	Training Set mean		K-fold Validation mean	
	R ²	RMSE	R ²	RMSE
MLR	0.79	0.31	0.52	0.48
PCA-MLR	0.71	0.35	0.49	1.13
PLS	0.61	0.40	0.62	0.40
BNN	0.71	0.34	0.47	0.47

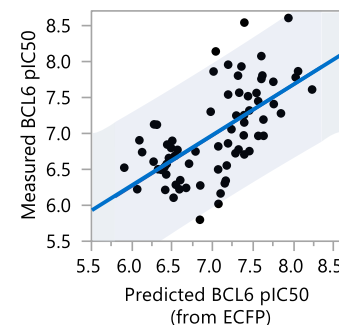
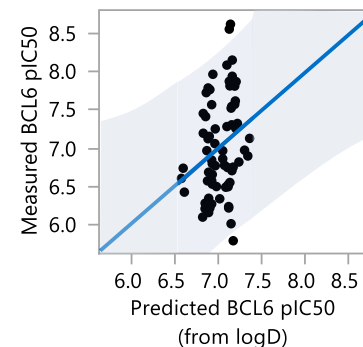
- Partial Least Squares (PLS) is the most robust method
 - Significant improvement over initial model
 - Demonstrates the importance of model validation
 - Often challenging or ignored (!) for more complex machine learning



Null Model Comparison (PLS)

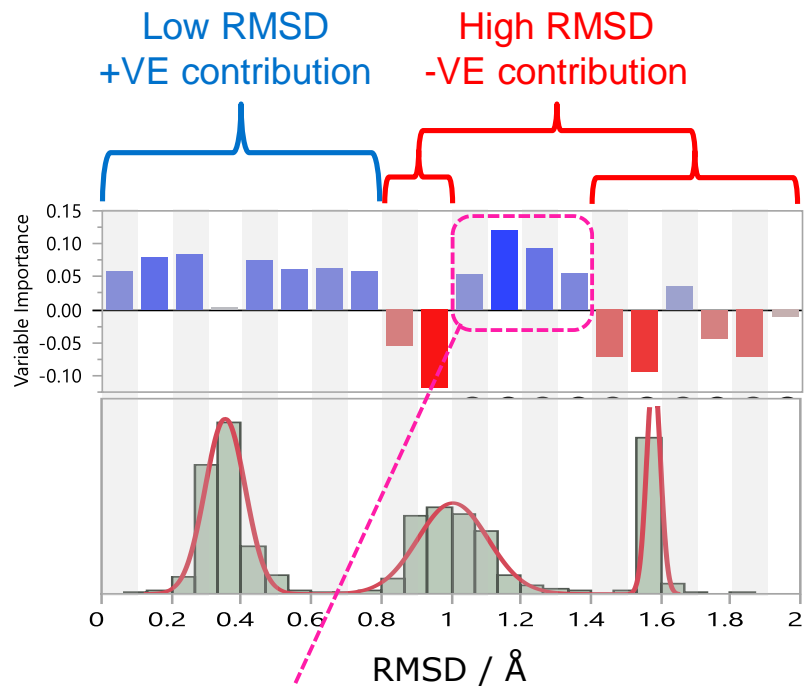
c.f. Conformation model
Validation Set $R^2=0.62$, $RMSE=0.40$

- The Medicinal Chemist Test – “it’s probably all logD”
 - False
 - K-fold validation, mean $R^2 = 0.05$
- The CompChemist Test – “it can probably be modelled with 2D fingerprints”
 - Using 2048-bit ECFP fingerprints (with variable selection)
 - Less good model
 - Model shows overfitting (poor for validation set)
 - K-fold validation, mean $R^2 = 0.46$, mean $RMSE = 0.48$



Model Interpretation

- Can visualise the contribution of the original descriptors to the model
- Interpretation makes physical sense for our knowledge of the system
- ML-set contributions
 - Reveals new details
 - Removes human bias



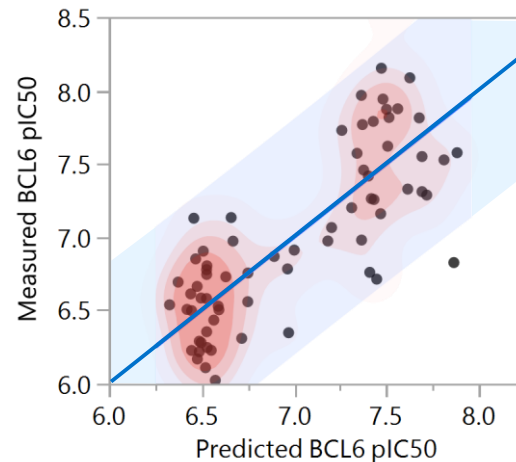
Wait! What does this mean?

Potentially due to secondary conformations that can easily collapse into the bioactive conformation (unproven hypothesis)



Model Limitations

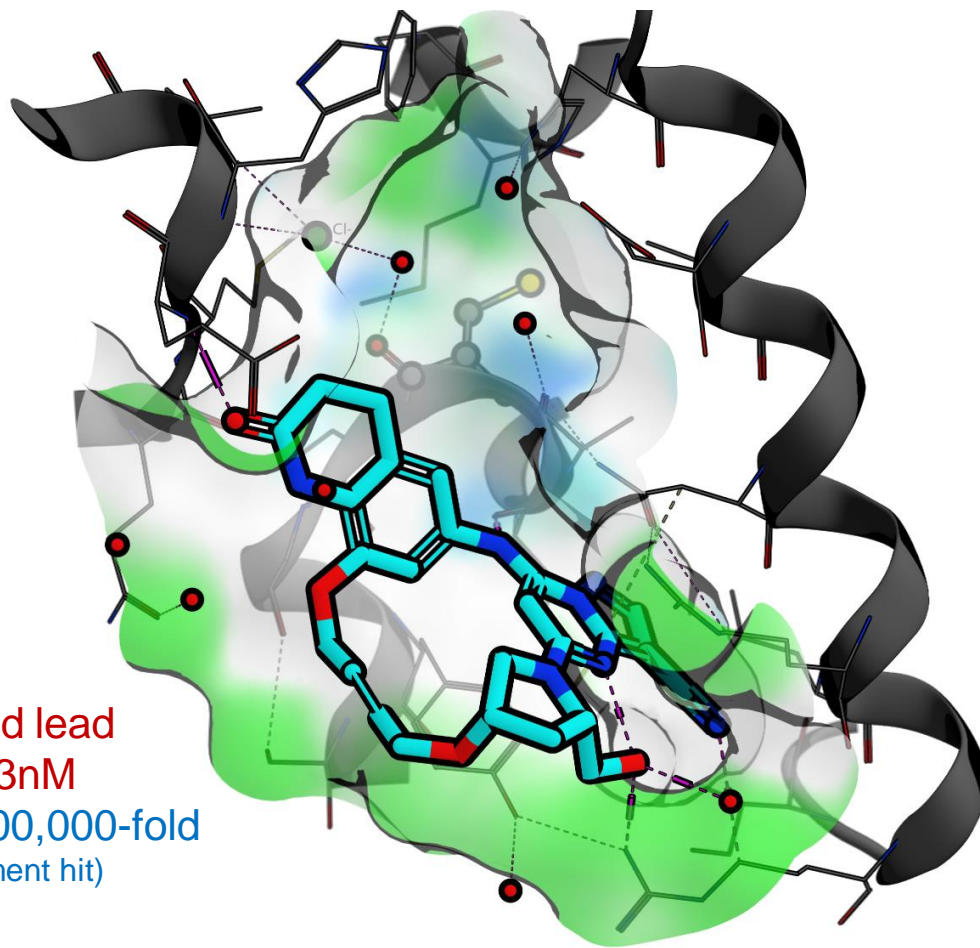
- While improved, the model is not perfect
 - Predictions deviate ± 0.4 on average
- Evident clustering within the data
 - No obvious structural reason
- Re-examine our initial assumptions
 - Only contributes to entropy of conformation selection
 - No contribution to enthalpy of binding
- As always, reality is more complex



Achieved significant potency boost

- Method used to help prioritise macrocycle linkers
- Delivered tool compound
- Solution NMR confirms 79% free population like bioactive

Optimised lead
 $IC_{50} = 3nM$
Affinity \rightarrow ~200,000-fold
(from fragment hit)



Conclusions

- Entropy is a key component of ligand binding
 - Controlling for it can increase potency significantly
- Machine learning can be used to draw out additional information from biophysical simulation results
- Don't ignore simpler modelling methods
 - More appropriate for small data-sets
 - Often more robust
 - Easier to interpret
- Validation is essential to finding a useful model



Acknowledgements

- Much as I might give the impression I did everything, this was a team project 😊

Chemistry

Graeme Robb
Roman Abrams
Kevin Blades
Pete Barton
Matthew Box
Jonathan Burgess
Qing Cao
Claudio Chuaqui
Rodrigo Carbajo
Shaun Fillery
Nathan Fuller
Matthew Grist
Paul Kemmitt

Richard Lonsdale
William McCoull
Jennifer Nelson
Piotr Raubo
Junjie Shi
Michael Waring
David Whittaker
Marta Wylot

DMPK

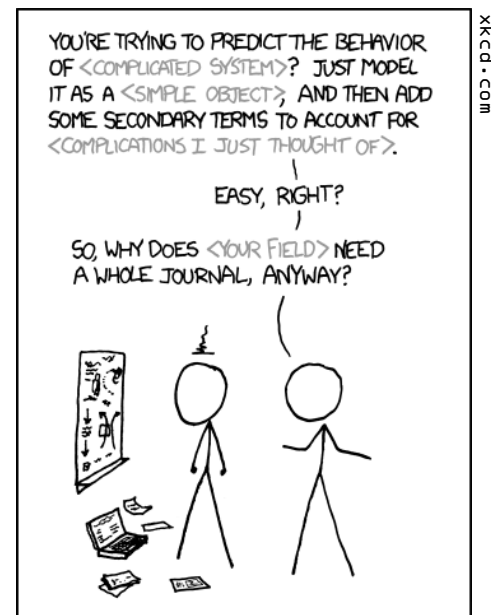
Eric Gangl
Martin Howard

Bioscience

Tony Cheung
Erica Anderson
Kate Byth (PL)

Discovery Sciences

Nichole O'Connell
Erin Code
Andrew Ferguson
Ning Gao
David Hargreaves
Jun Hu
Bryan Prince
Philip Rawlins
Xiahui Zhu



MEDICINAL CHEMISTS: MAY BE ANNOYING SOMETIMES, BUT THERE'S *NOTHING* MORE OBNOXIOUS THAN A DATA SCIENTIST FIRST ENCOUNTERING A NEW SUBJECT.



Confidentiality Notice

This file is private and may contain confidential and proprietary information. If you have received this file in error, please notify us and remove it from your system and note that you must not copy, distribute or take any action in reliance on it. Any unauthorized use or disclosure of the contents of this file is not permitted and may be unlawful. AstraZeneca PLC, 1 Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge, CB2 0AA, UK, T: +44(0)203 749 5000, www.astrazeneca.com

