Predictions of Intrinsic Aqueous Solubility of Crystalline Drug-like Molecules

NSCCS

University of St Andrews

EaStCHEM
The Edinburgh and St Andrews Research School of Chemistry

SULSA
Scottish Universities Life Sciences Alliance

James McDonagh

Supervisors Dr John Mitchell and Dr Tanja van Mourik

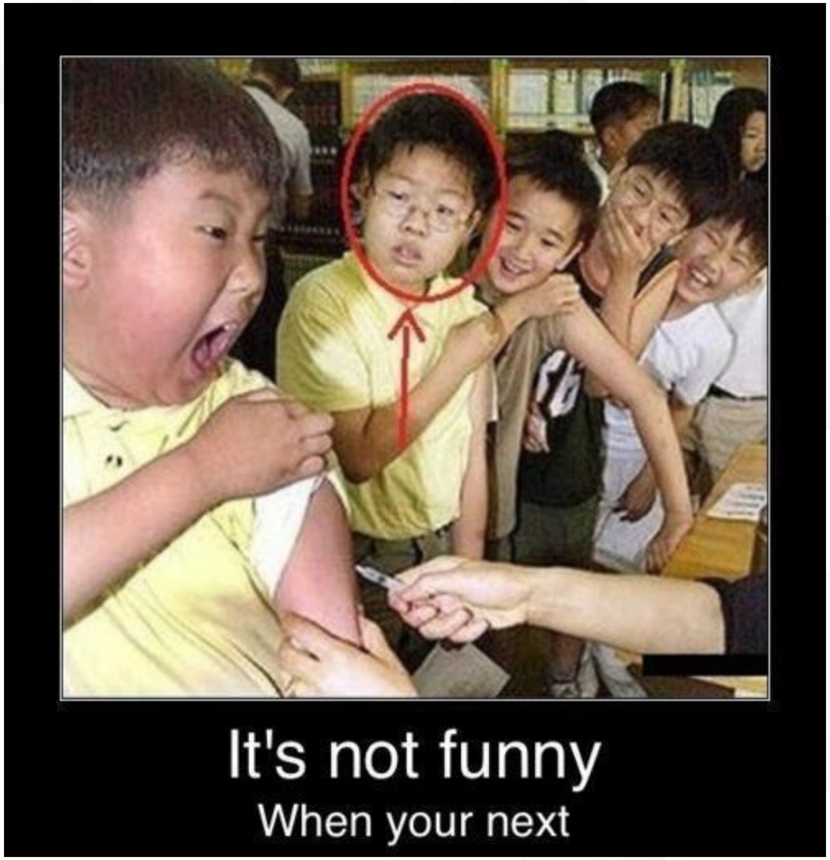
Overview

- **Introduction.**
- **Solution model** – Our solubility prediction model.
- **Results** – The performance of the solubility predictions.
- **Further models** – Work following on from the initial model.
- **Current/future work** – Where we are currently and where we are going.

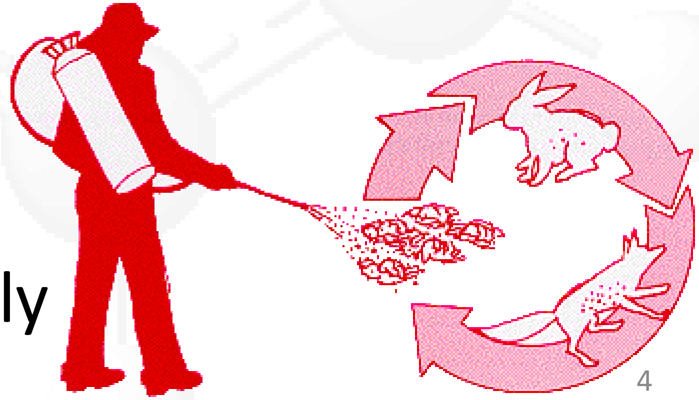
Why is solubility prediction important?

- Crucial factor to control bioavailability of drug candidates.
- Critical component in determining the environmental impact of pesticides.
- Accurate *in silico* predictions of solubility can save time and money.

Why should we care about solubility ?



- New drug candidates are often more insoluble than their predecessors.
- Formulation of these drugs often involve less pleasant administration methods.



- Certain pesticides can cause extensive damage and potentially enter the water cycle.

Existing Theoretical Approaches

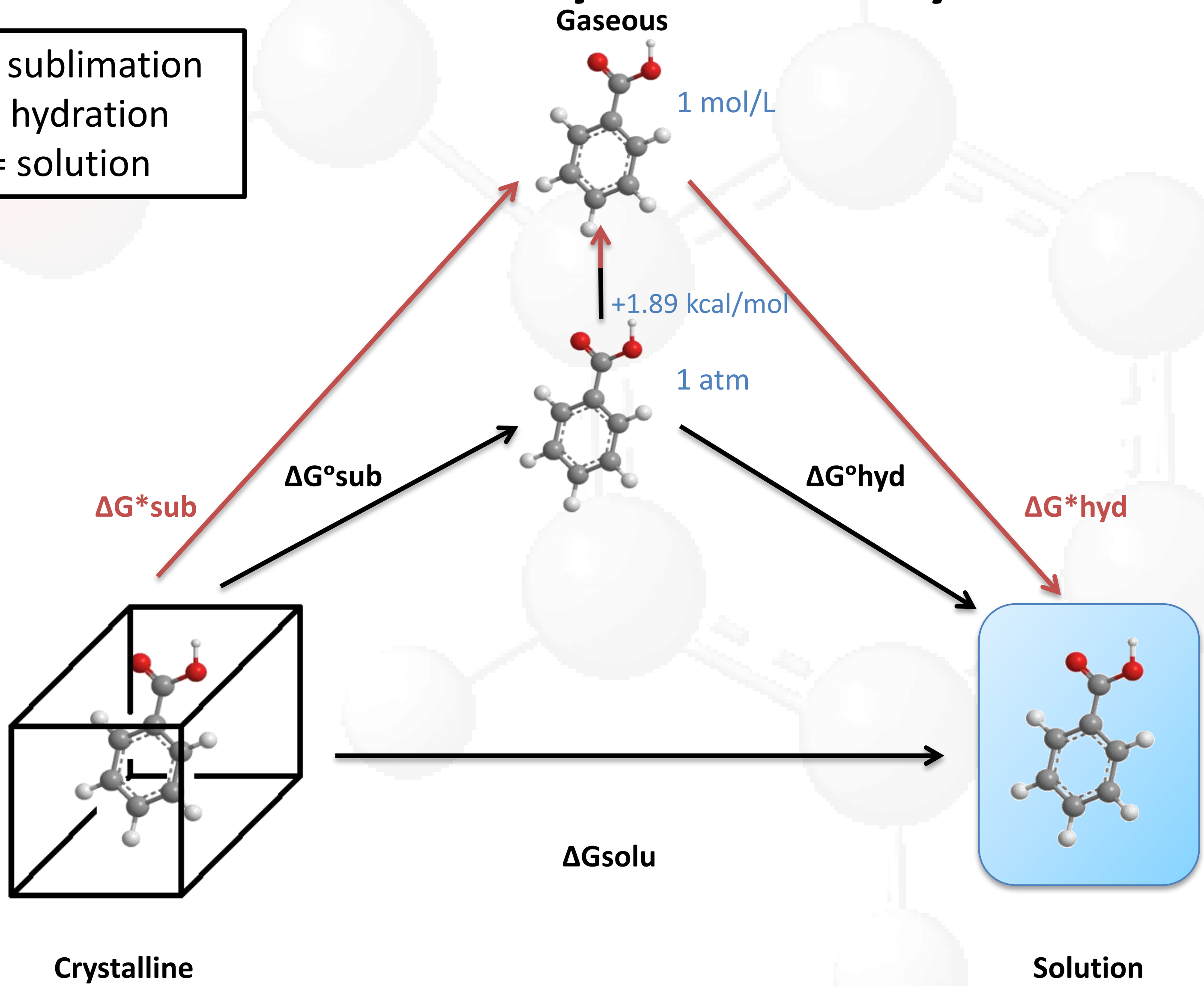
- So far, although theoretical methods have shown promise, they have not matched the accuracy of QSPR. Theoretical methods do have the advantage of being physically tractable.
- Industry also requires high through put methods. QSPR models are generally much faster than computational chemistry models.
- There are many theoretical models to make solubility predictions.

Our Methodologies

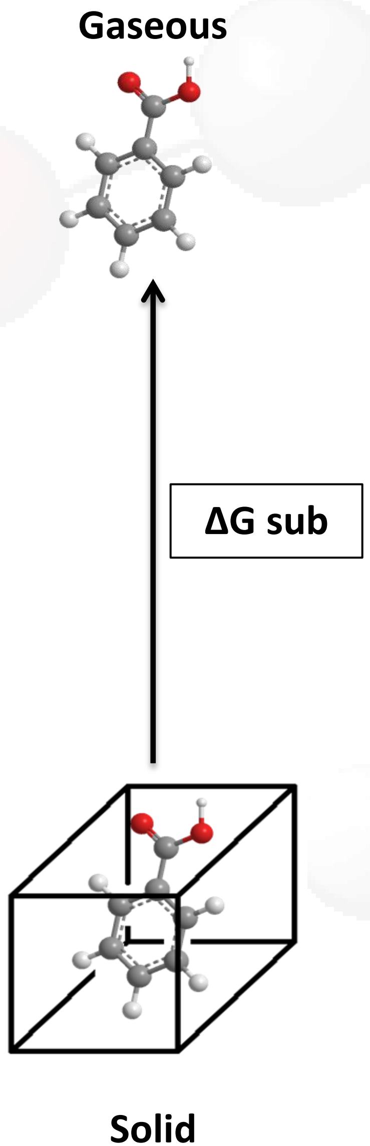
- We have decomposed the solution (solu) free energy prediction in to two distinct steps.
  - Sublimation (sub)
  - Hydration (hyd)
- We have applied a range of methodologies to each step.
- Methodologies include simulation, QM calculation and machine learning.

Thermodynamic cycle

Sub = sublimation
Hyd = hydration
Solu = solution

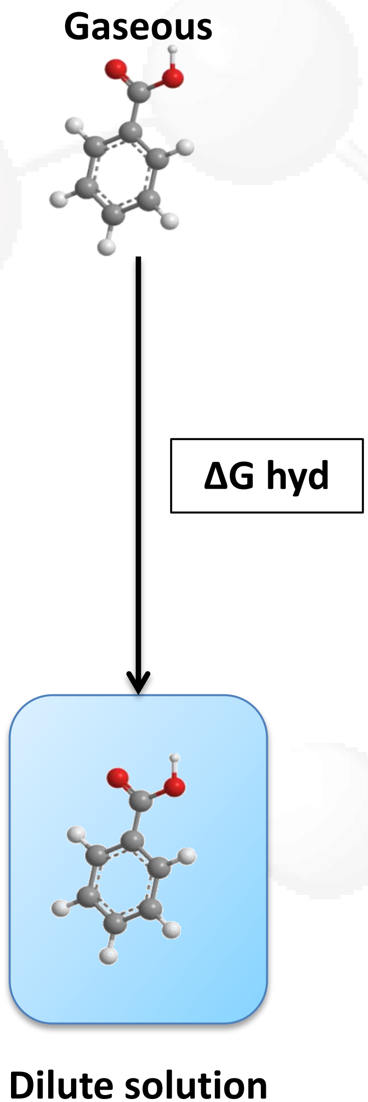


Sublimation : Predictions by DMACRYS



- DMACRYS a periodic lattice simulation program.
- Electrostatics, from distributed multipoles.
- Buckingham potential to account for repulsion and dispersion.
- Calculates lattice energy and crystal entropy from phonon modes.
- Gas phase contributions calculated in Gaussian 09.

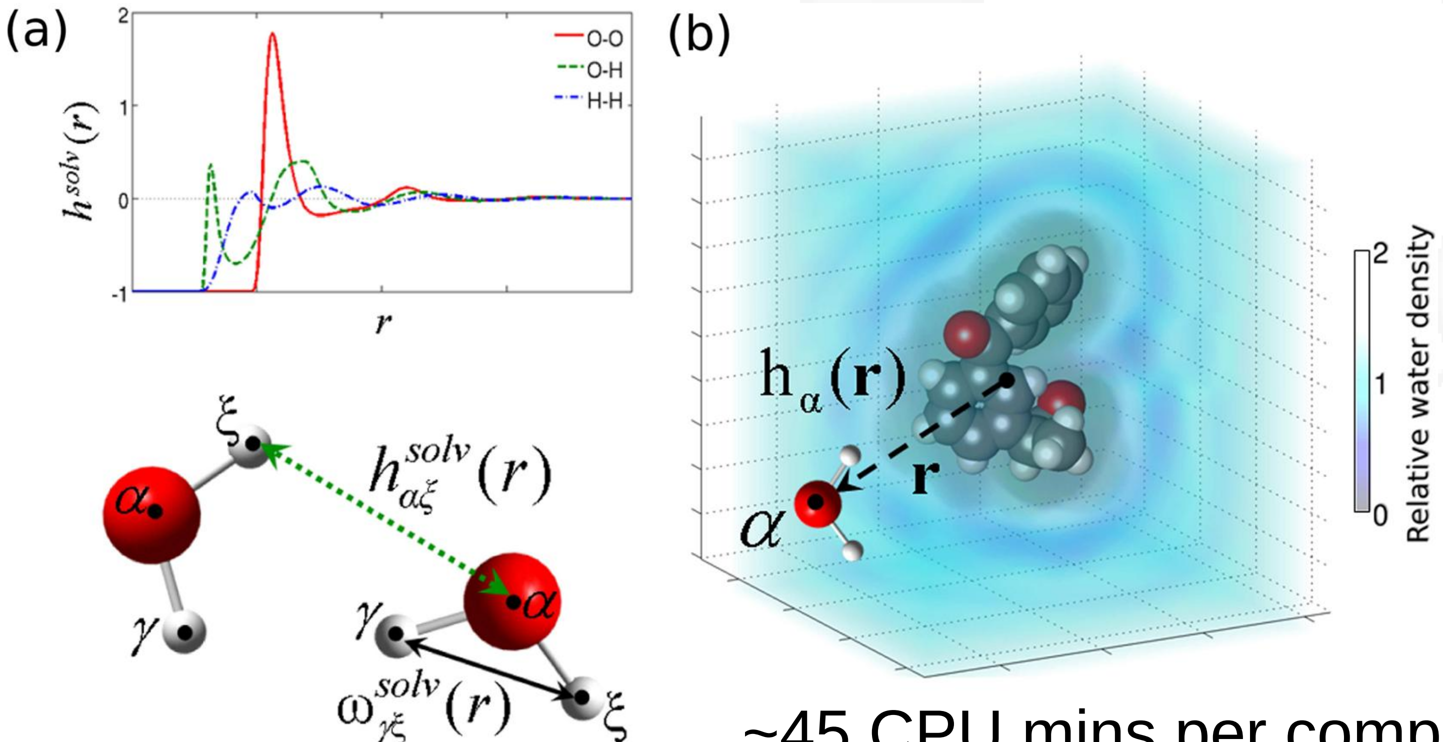
Hydration: Solvation models



- Continuum solvent, solvation model based on density (SMD).
- An integral equation theory (IET) of molecular liquids methodology the Reference Interaction Site Model (RISM).

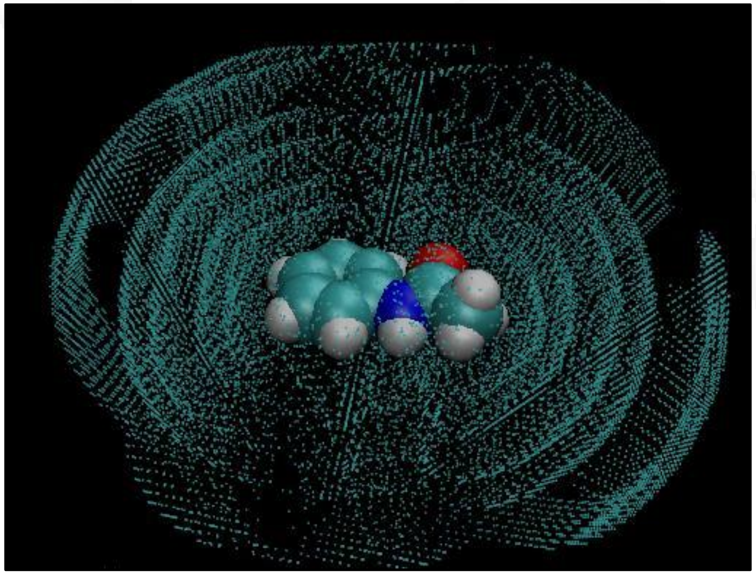
Hydration: RISM

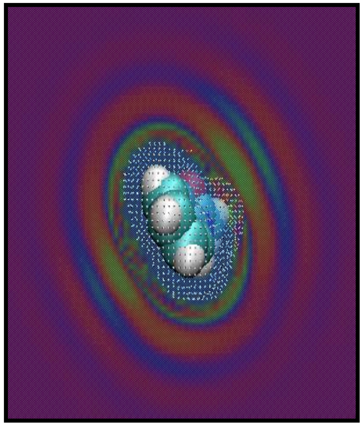
- Combines features of explicit and implicit solvent models.
- Solvent density is modelled, but no explicit molecular coordinates or dynamics.

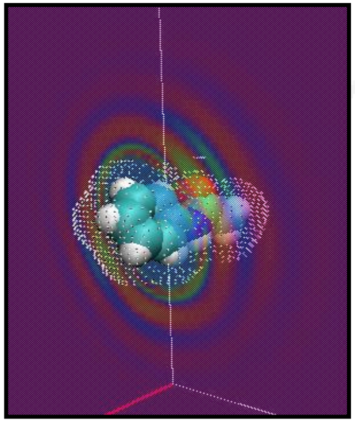


~45 CPU mins per compound

Hydration: RISM







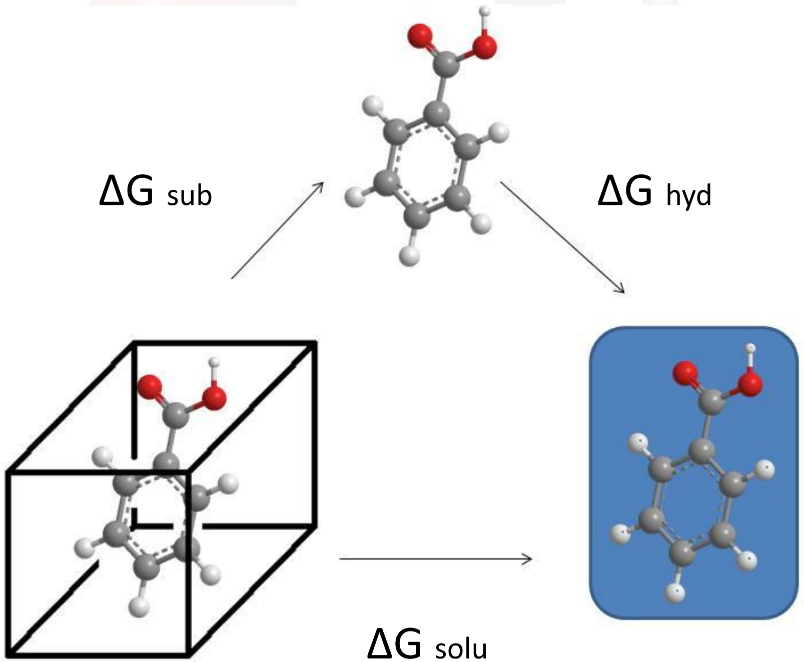
- We use the 3D RISM variation.
- We employ the Kovalenko-Hirata (KH) closure and Gaussian fluctuation free energy functional.
- We also employ the universal correction (UC).

$$\Delta G_{hyd}^{3DRISM-KH/UC} = \Delta G_{hyd}^{3DRISM} + a(\rho V) + b$$

- The methodology we used will be referred to as 3DRISM-KH/UC.

David, S. Palmer., Andrey I Frolov, Ekaterina L Ratkova and Maxim V Fedorov,. *Journal of Physics: Condensed Matter* **2010,** *22* (49), 492101.

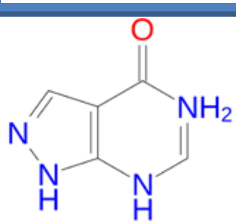
Solution Free Energy

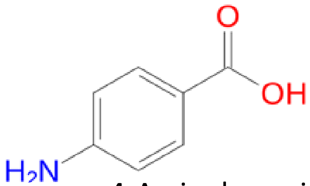


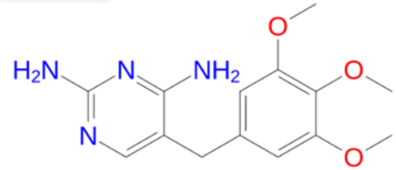
- The sum of our predictions of ΔG_{sub} and ΔG_{hyd} produce a ΔG_{solu} prediction.
- These methods were carried out for 25 chemically diverse drug-like molecules.
- Chemical accuracy ~4 kJ/mol or ~ 1 LogS unit.
- Useful predictions are within the standard deviation (SD) of the experimental values.

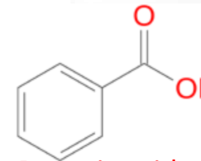
Data set selection

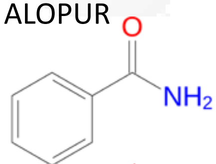
- The thermodynamically most stable polymorph was selected
- 10 had available experimental sublimation and hydration free energies. (in red).

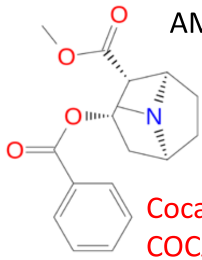

Allopurinol
ALOPUR

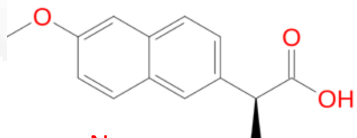

4-Aminobenzoic acid
AMBNAC04

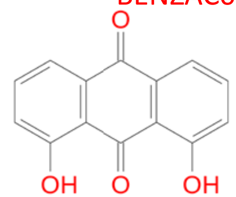

Trimethoprim
AMXBPM10

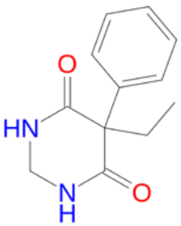

Benzoic acid
BENZAC02

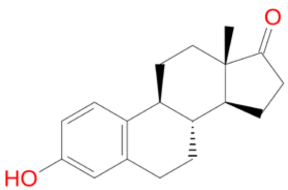

Benzamide
BZAMID02

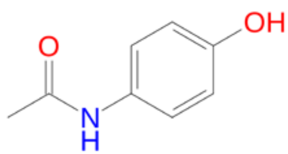

Cocaine
COCAIN10

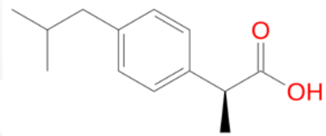

Naproxen
COYRUD11

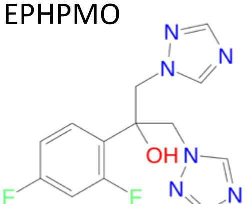

Danthron
DAHNQU06

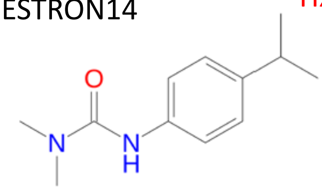

Primidone
EPHPMO

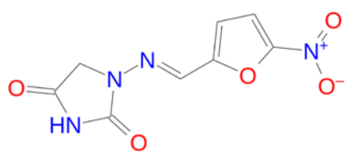

Estrone
ESTRON14

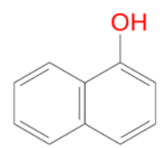

Acetaminophen
HXACAN04

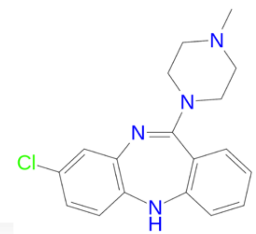

Ibuprofen
IBPRAC01

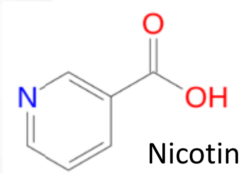

Fluconazole
IVUQOF

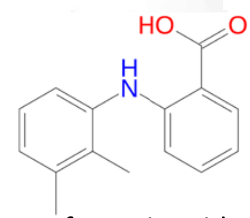

Isoproturon
JODTUR01

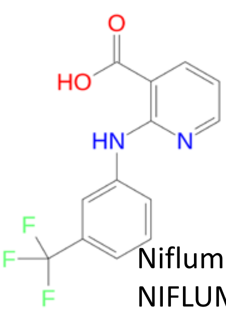

Nitrofurantoin
LABJON01

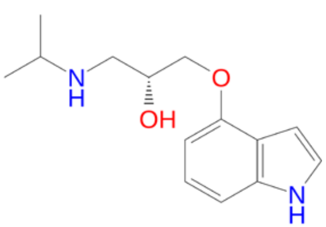

1-Naphthol
NAPHOL01

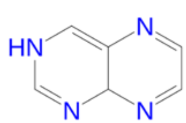

Clozapine
NDNHCL01

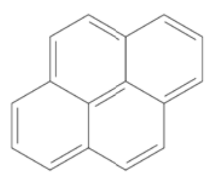

Nicotinic acid
NICOAC02

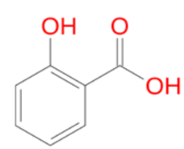

Mefenamic acid
XYANAC

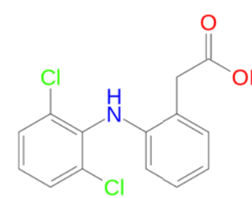

Niflumic acid
NIFLUM10


Pindolol
PINDOL

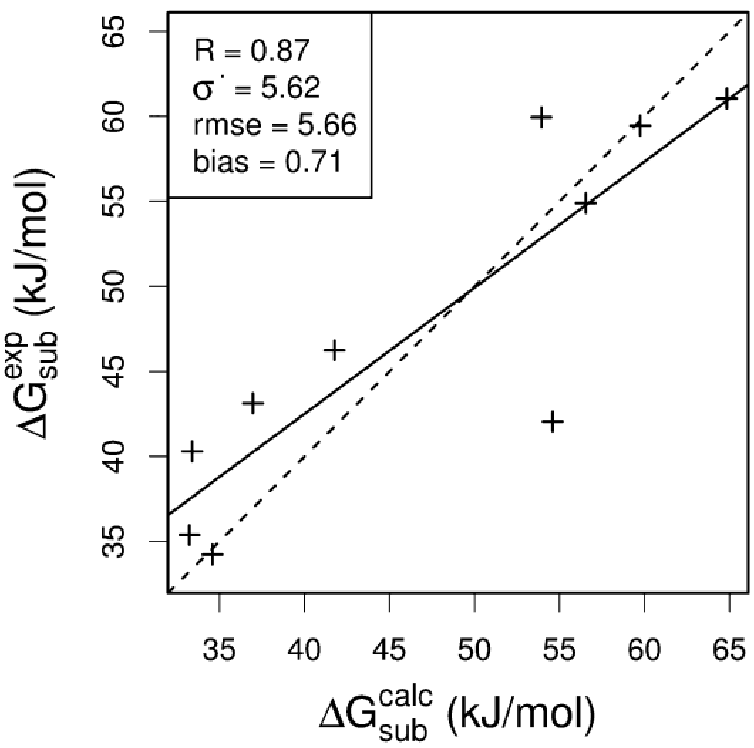

Pteridine
PTERID11


Pyrene
PYRENE07


Salicylic acid
SALIAC

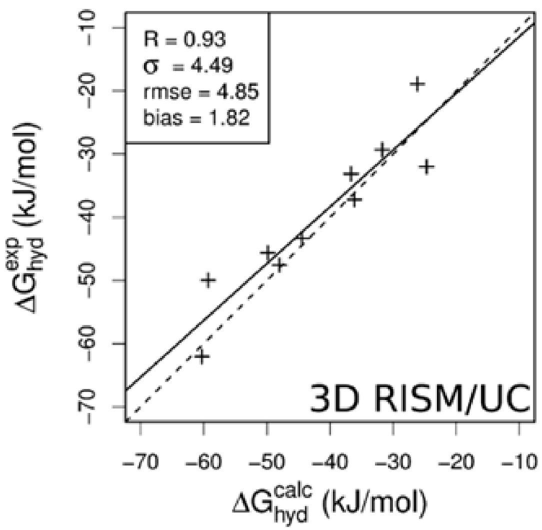

Diclofenac
SIKLIH01

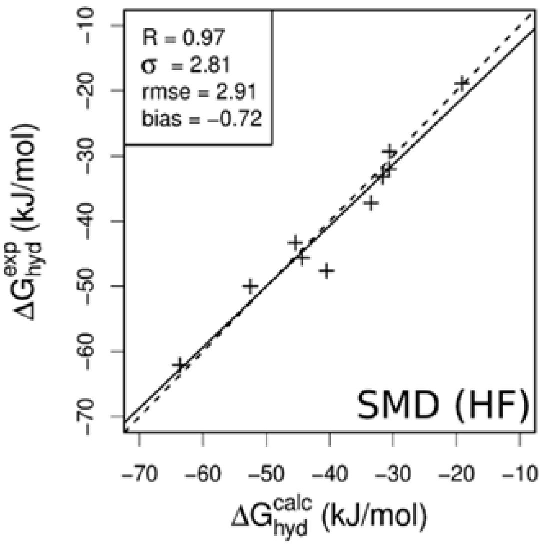
Sublimation free energy predictions



- Validation of the sublimation free energy prediction.
- B3LYP/6-31G(d,p) multipole.
- FIT repulsion dispersion potential.
- Correlation coefficient (R) 0.87.
- RMSE 5.66kJ/mol.

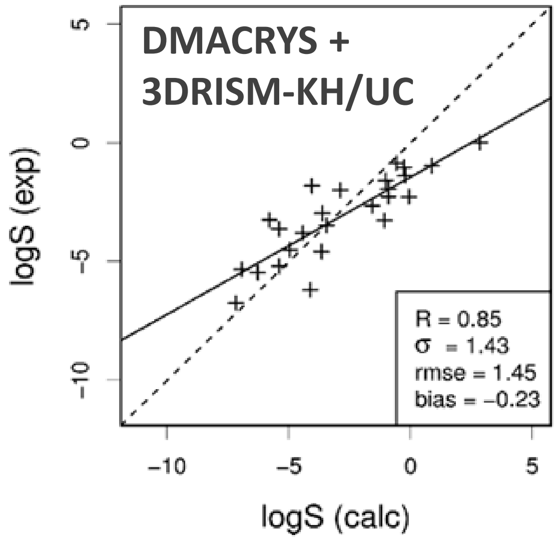
Hydration Free Energy Predictions

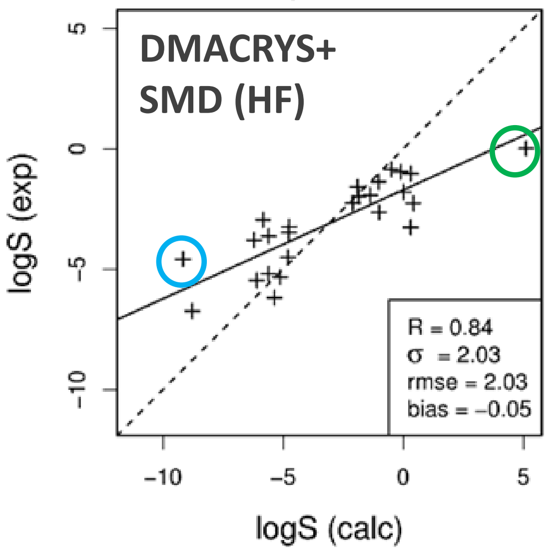




- Validation of the hydration free energy predictions.
- SMD HF/6-31G(d,p).
- Both have strong R values 0.93 RISM, 0.97 SMD.
- RISM has a significantly higher RMSE 4.85kJ/mol RISM, 2.91kJ/mol SMD.

Solution Free Energy Predictions





- The full 25 molecule set is compared to experiment.
- Chemical accuracy ~ 1logS unit. Experimental SD 1.79 LogS.
- Reasonable correlation R 0.85 RISM, R 0.84 SMD.
- RISM method provides best RMSE, RMSE of 1.45LogS RISM, RMSE of 2.03LogS SMD.
- SMD outliers Niflumic Acid and Pteridine.

Palmer, D. S.; McDonagh, J. L.; Mitchell, J. B. O.; van Mourik, T.; Fedorov, M. V., *Journal of Chemical Theory and Computation* **2012**.

To sum up

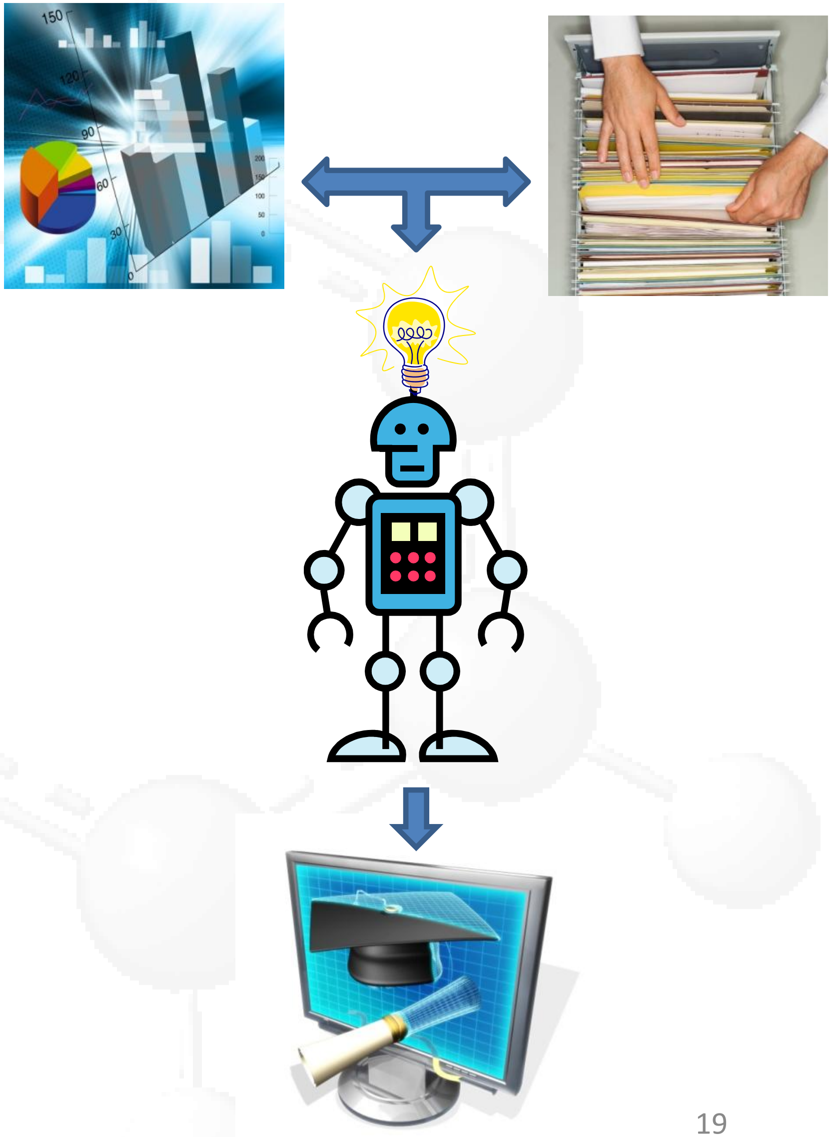
- From these results we concluded it was possible to make predictions of a reasonable accuracy.
- In our methodology a larger portion of error could be attributed to the sublimation free energy prediction.
- Larger datasets were required to fully validate the methodology.

Further models

- We took two approaches to follow up this work:
  - Parameterisation and machine learning approaches to predict ΔG solution.
  - Systematic theoretical improvements in Sublimation free energy predictions. (work currently ongoing).

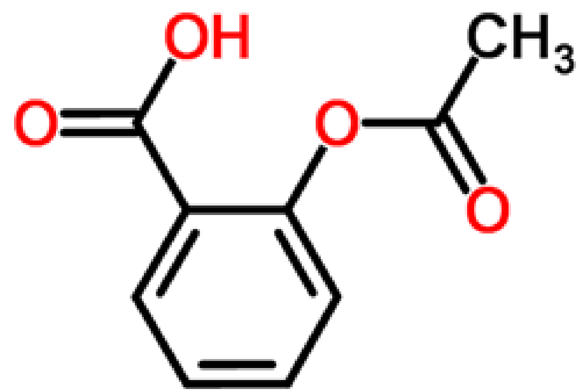
Informatics and machine learning

- We selected a dataset of 100 molecules.
- Calculated descriptors using the chemistry development kit (CDK).
- We followed our previously laid out theoretical methodology for the 100 molecules.
- We combined descriptors and theoretically calculated energies.



Descriptors

- SMILES are input into CDK.
- Structural and some predicted properties are output for use as descriptors.
- These cheminformatics descriptors were used as part of the input for the machine learning methods.

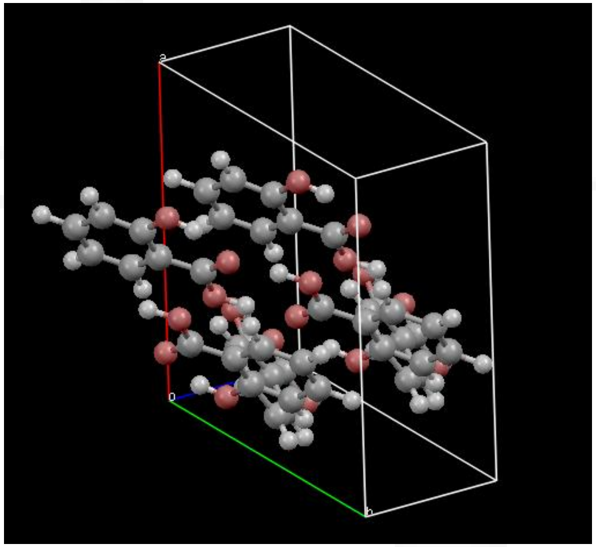


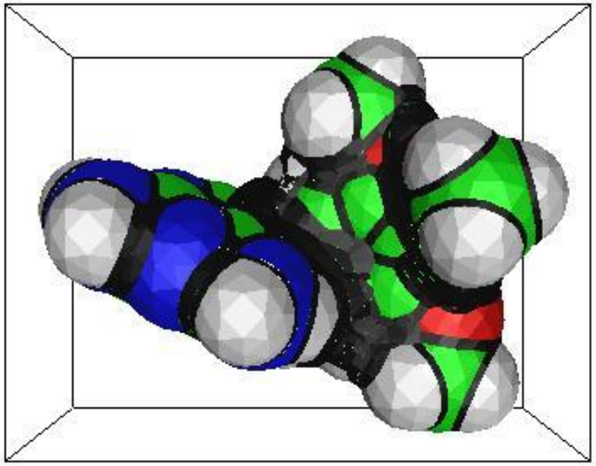
CC(=O)Oc1ccccc1C(=O)O

Descriptor	Value
Molecular Weight	280
Molecular Formula	C9H8O4
XLogP	1.24
Freely rotatable bonds	3
H bond acceptor	4
…..	…..
……	……

Computational Chemistry Calculations

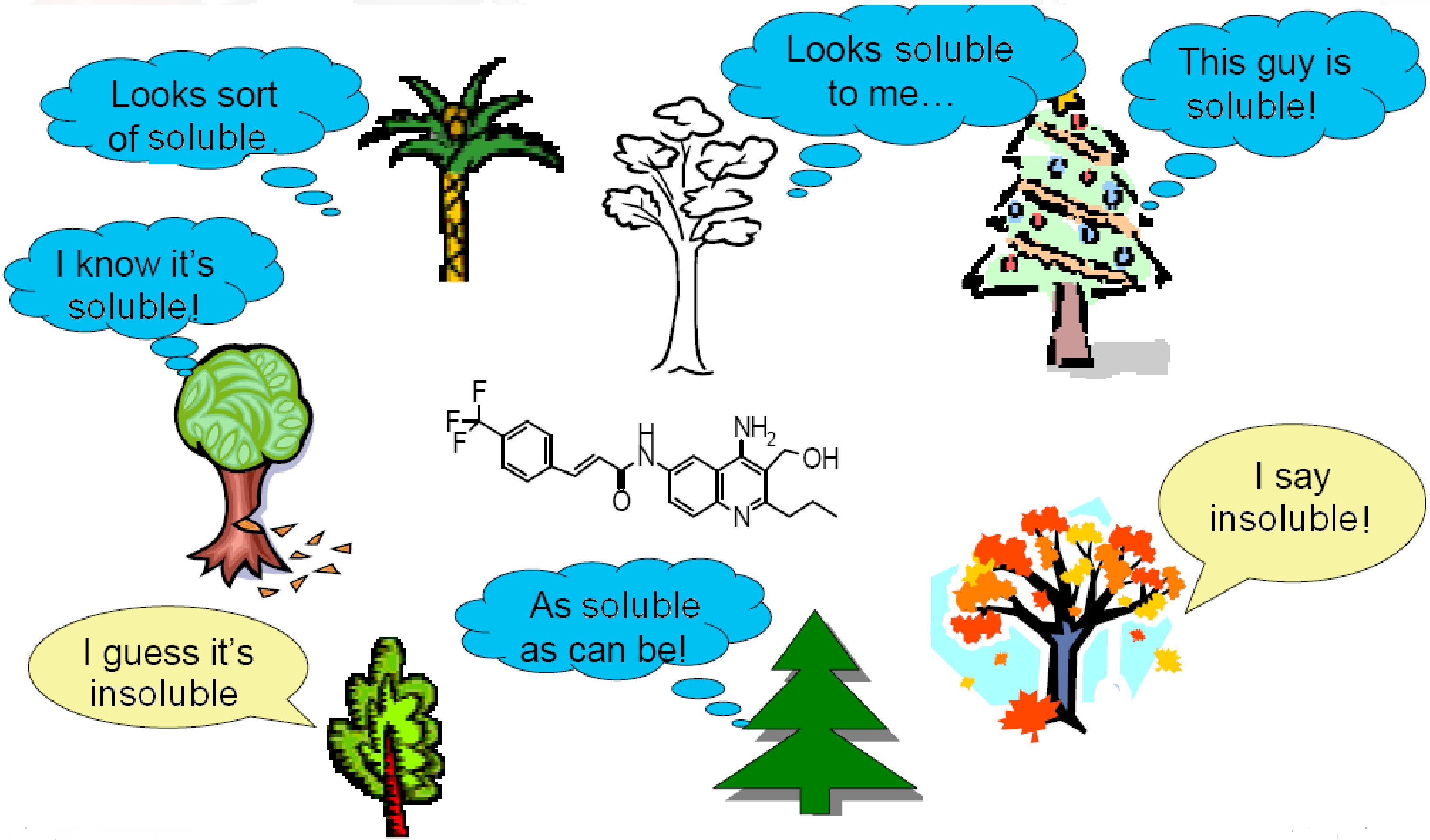
- DMACRYS –B3LYP/ 6-31G(d,p), FIT potential.
- SMD with HF/6-31G(d,p)
- The same level of theory was used in the gas phase as the solution phase.
- SMD selected over RISM as it provided a better correlation in the previous work.



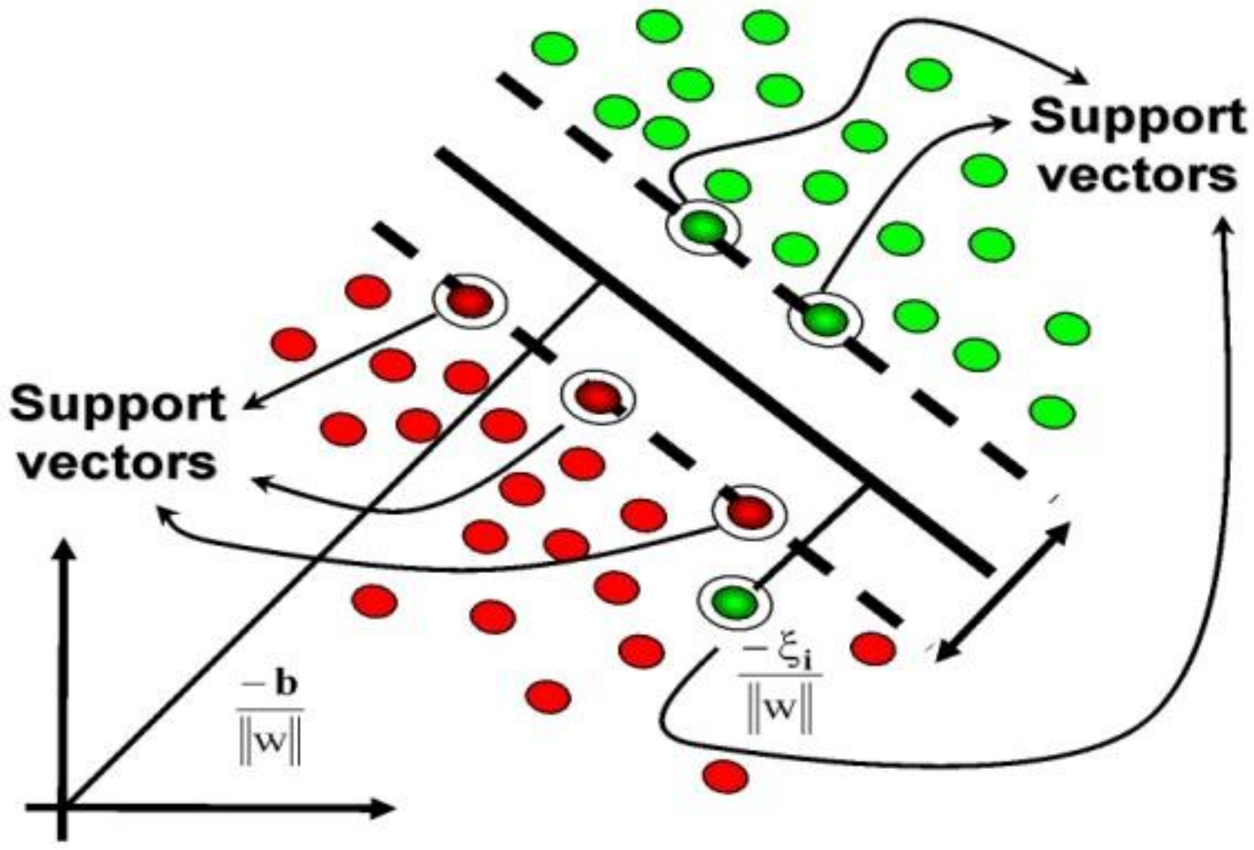


Machine Learning Models

- Random Forest (RF) – A forest of decision trees.



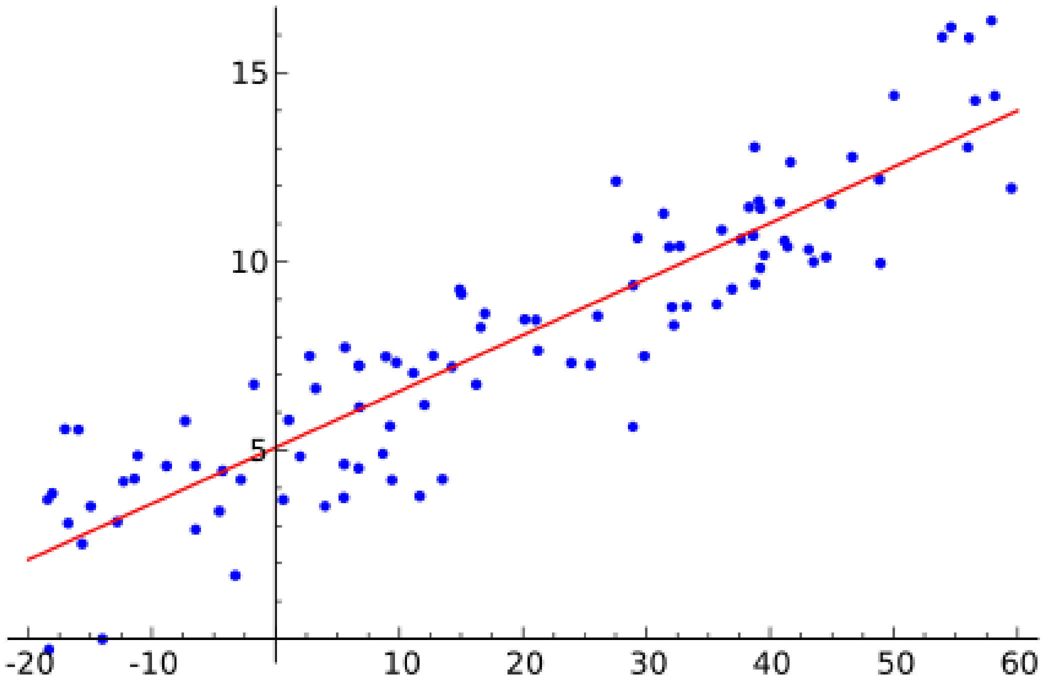
- Support Vector Machines (SVM) – Classification by projection into a higher space and separation by a hyperplane.



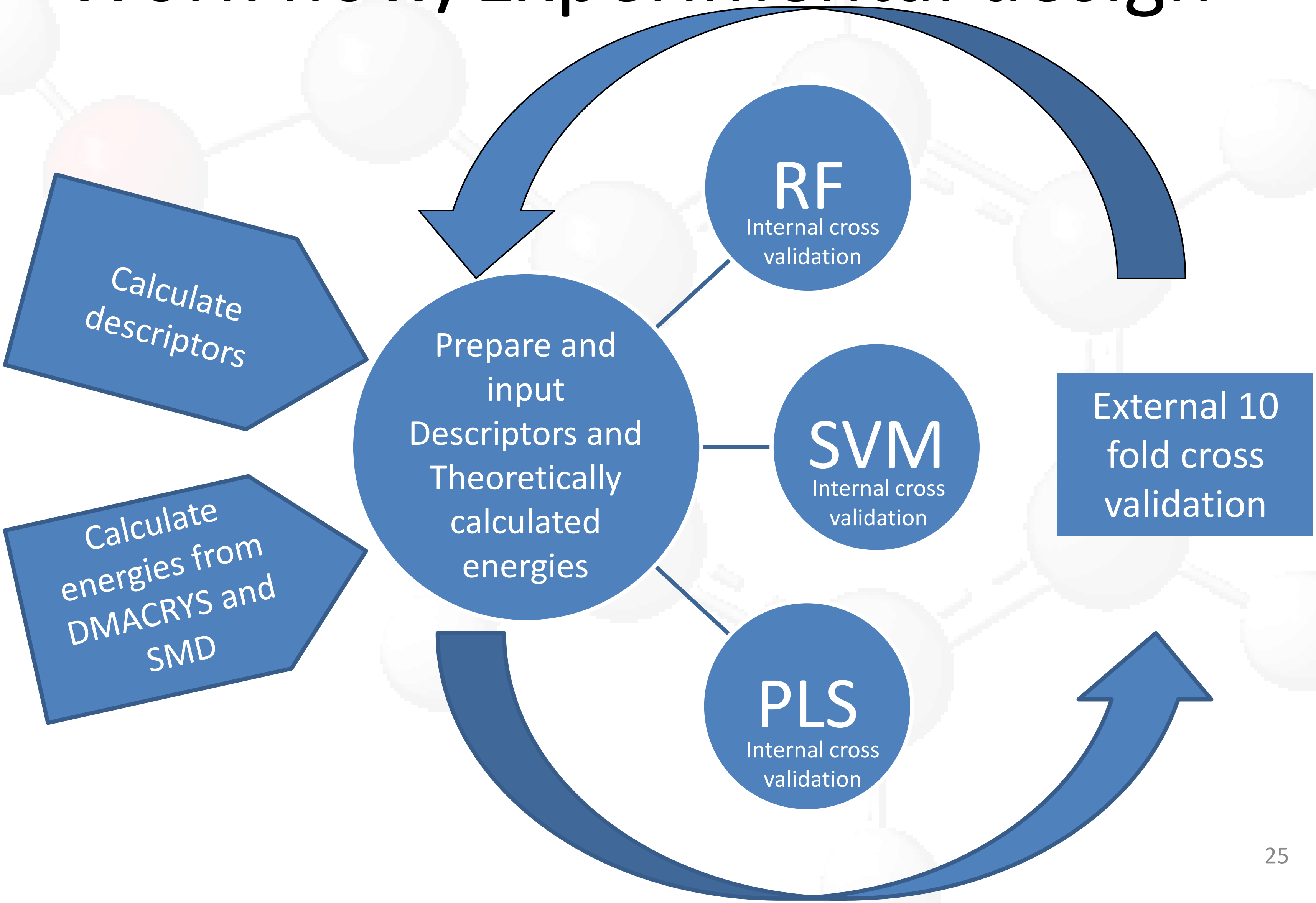
- Partial least squares Regression (PLS) – can be considered as classification by deflation.

$$y = X\hat{b} + \epsilon_t$$

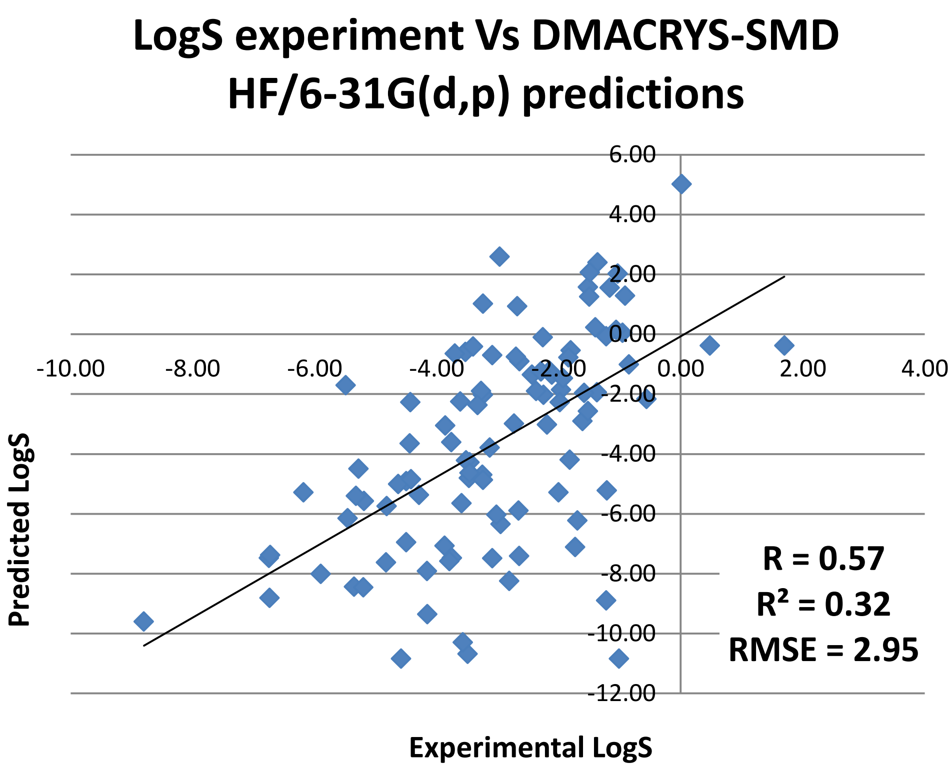
$$\begin{pmatrix} 1 & 0.5 & \color{red}{5} \\ 1 & 0.7 & \color{red}{7} \\ 1 & 0.9 & \color{red}{9} \end{pmatrix} \begin{pmatrix} 50 \\ 70 \\ \color{red}{2} \end{pmatrix}$$



Work flow/Experimental design

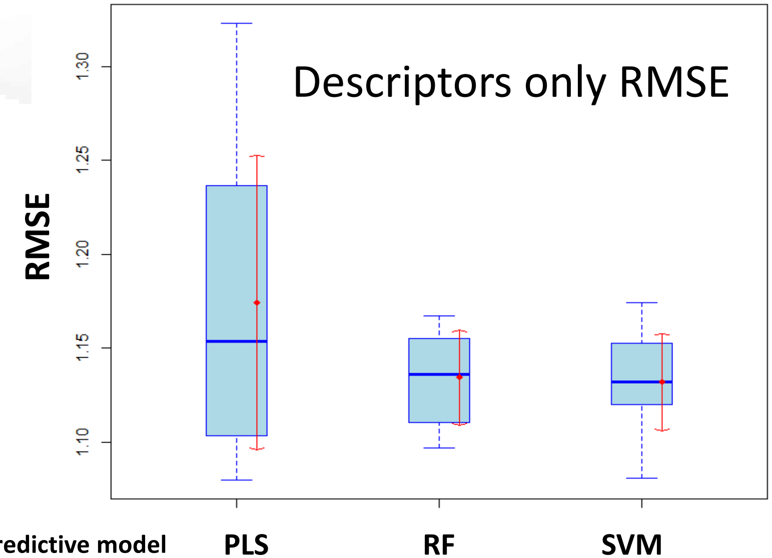


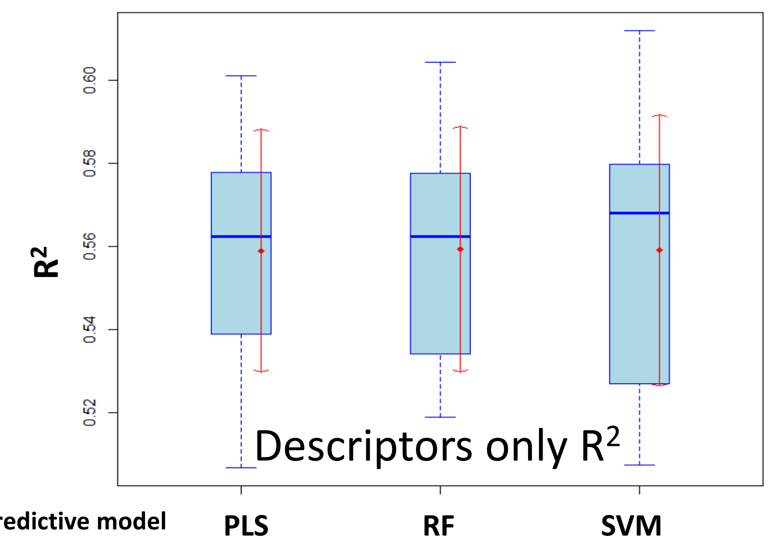
Results



- <u>Results of the purely theoretical prediction</u>.
- Our results are correlative.
- Standard linear regression is a poor fitting model.
- Chemical accuracy ~1logS unit.
- Experimental SD 1.71 LogS units.

Descriptors only

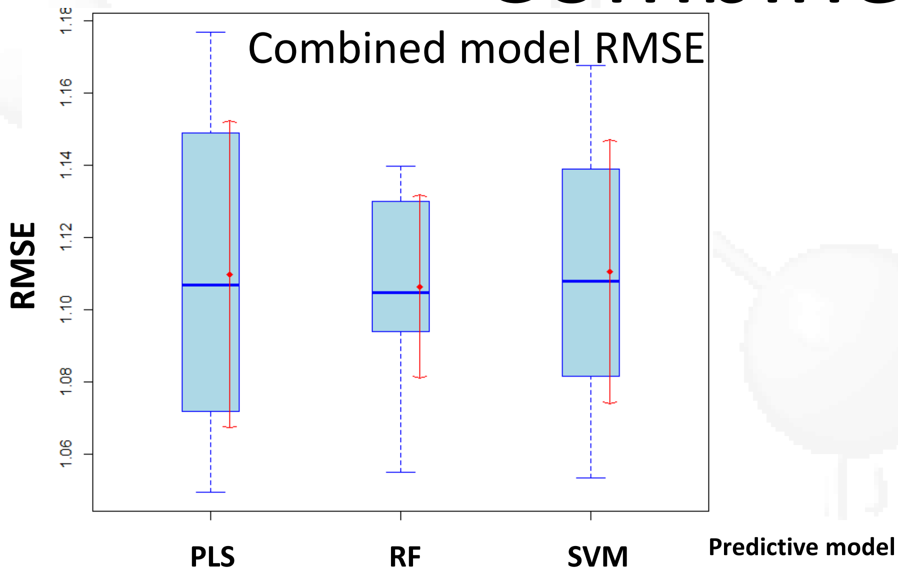


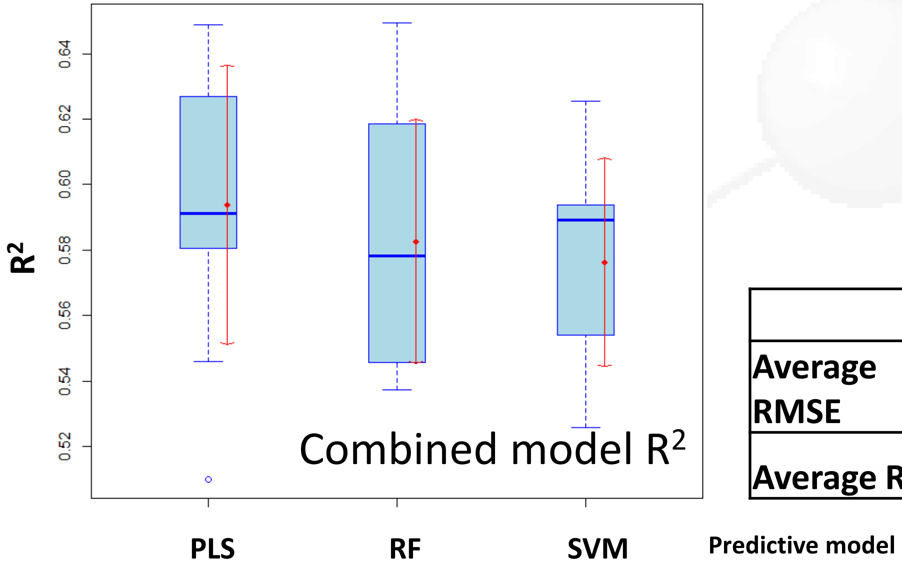


- Results of prediction exclusively using the CDK descriptors.
- All machine learning methods perform better than theory alone.
- Red bars show the SD and mean result.
- Boxes represent 75% of the predictions. Dark blue line shows the median.

	PLS	RF	SVM	Theory
Mean RMSE	1.174(±0.08)	1.134(±0.03)	1.132(±0.03)	2.95
Mean R^2	0.56(±0.03)	0.56(±0.03)	0.56(±0.03)	0.32

Combined model





- The model contains HF/6-31G(d,p) energies and descriptors.
- All show improvement over pure theory.
- Result are similar to those of the descriptors alone.
- Experimental SD 1.71 LogS units.

	PLS	RF	SVM	Theory
Average RMSE	1.110(± 0.04)	1.107(±0.03)	1.111(±0.04)	2.95
Average R^2	0.594(±0.04)	0.583(±0.04)	0.576(±0.04)	0.32

Summary

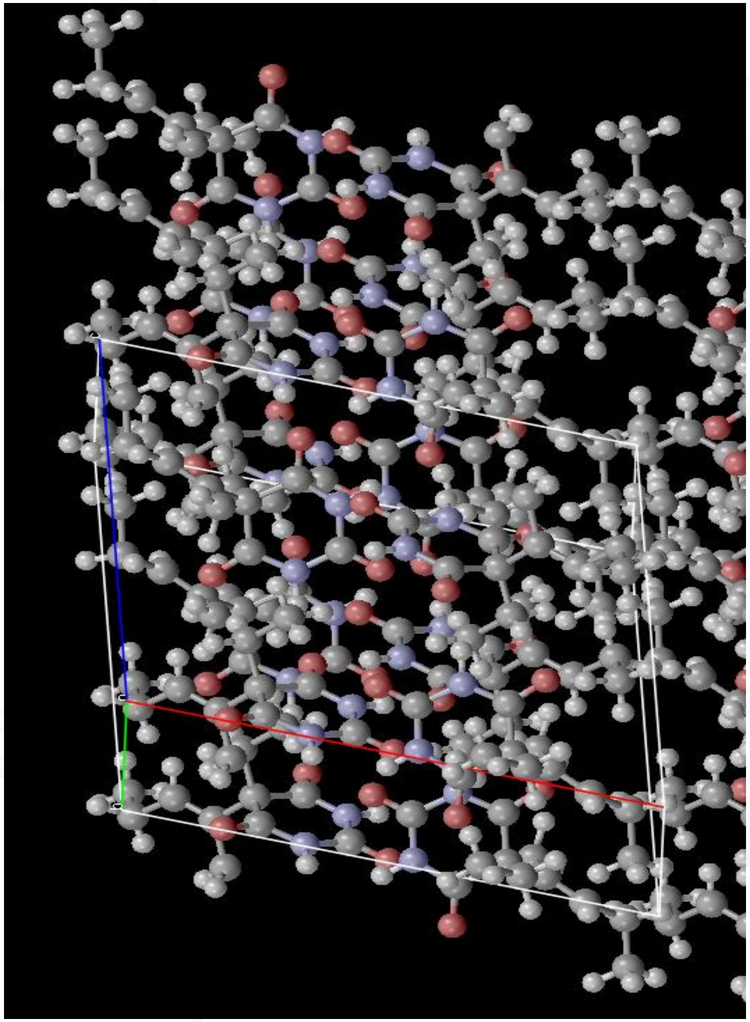
- The information from theoretical calculations at this level has a minor impact but does improve accuracy and correlation of the results.
- The descriptors already hold much of the information.
- Further exploration of models of this type could allow us to find information not held in the descriptors that is accessible by chemical calculation.

Exploration of sublimation free energy prediction

- The largest source of error in the initial method was the sublimation free energy prediction.
- We have a dataset of 60 molecules.
- We made sublimation free energy predictions using this dataset with our previously outlined method.
- We look to DFT methods to provide improved predictions.

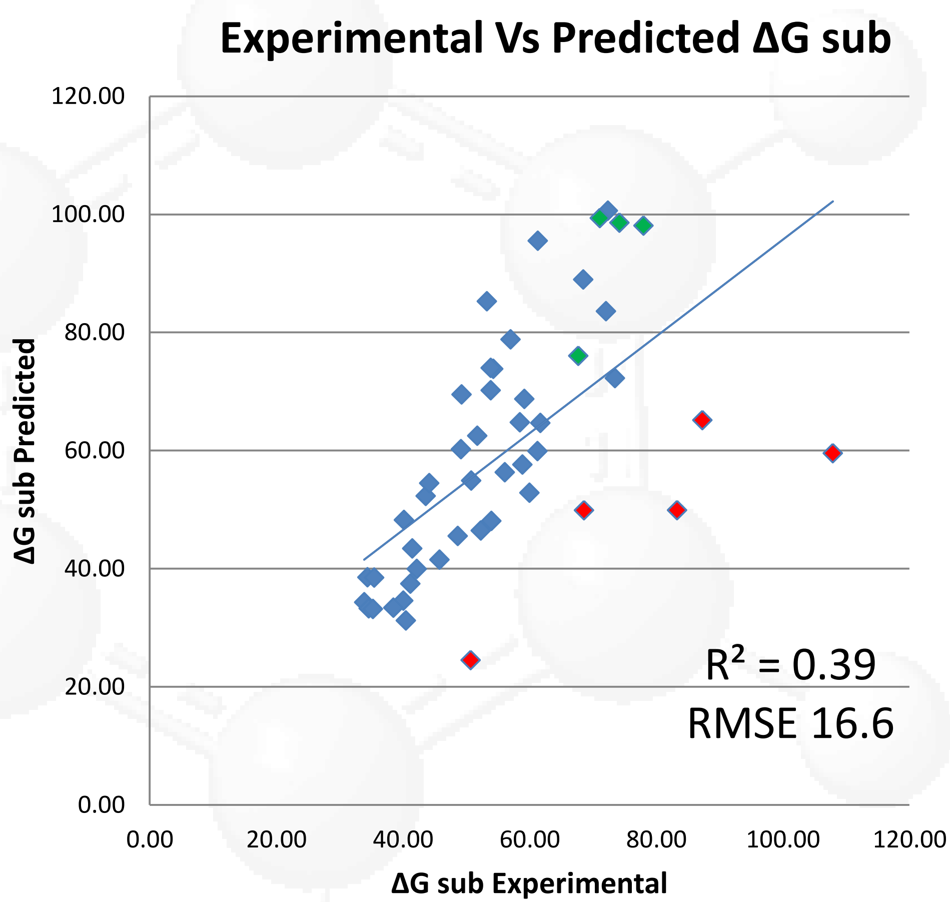
Periodic DFT

- We are using Periodic DFT to make sublimation free energy predictions.
- We are exploring dispersion corrections.
- We will look at the accuracy of prediction of the components of the free energy.



Free Energy of Sublimation

- From our initial methodology.
- A poor correlation.
- Significant RMSE.
- Outliers hold significant leverage.
- All outliers contain NO_2 groups (in red) which are known to be difficult to represent accurately in force fields.



Summary

- We have explored a purely theoretical methodology for predictions of solvation free energy.
- We have expanded from this to produce a combined computational chemistry - cheminformatics methodology.
- We have begun exploration of sublimation free energy, due to the large error it contributed in our original methodology.

Thanks for your attention

Thanks to the members of room 150 School of Chemistry University of St Andrews:


Dr Lazaros Mavridis


Rosanna Alderson


Neetika Nath

Dr Luna De Ferrari


Luke Crawford


Leo Holroyd


Dr Ludovic Castro


Ava Sih-Yu Chen

Many thanks to my Supervisors


Dr John Mitchell


Dr Tanja van Mourik

Many thanks to collaborators, co-workers and funders.

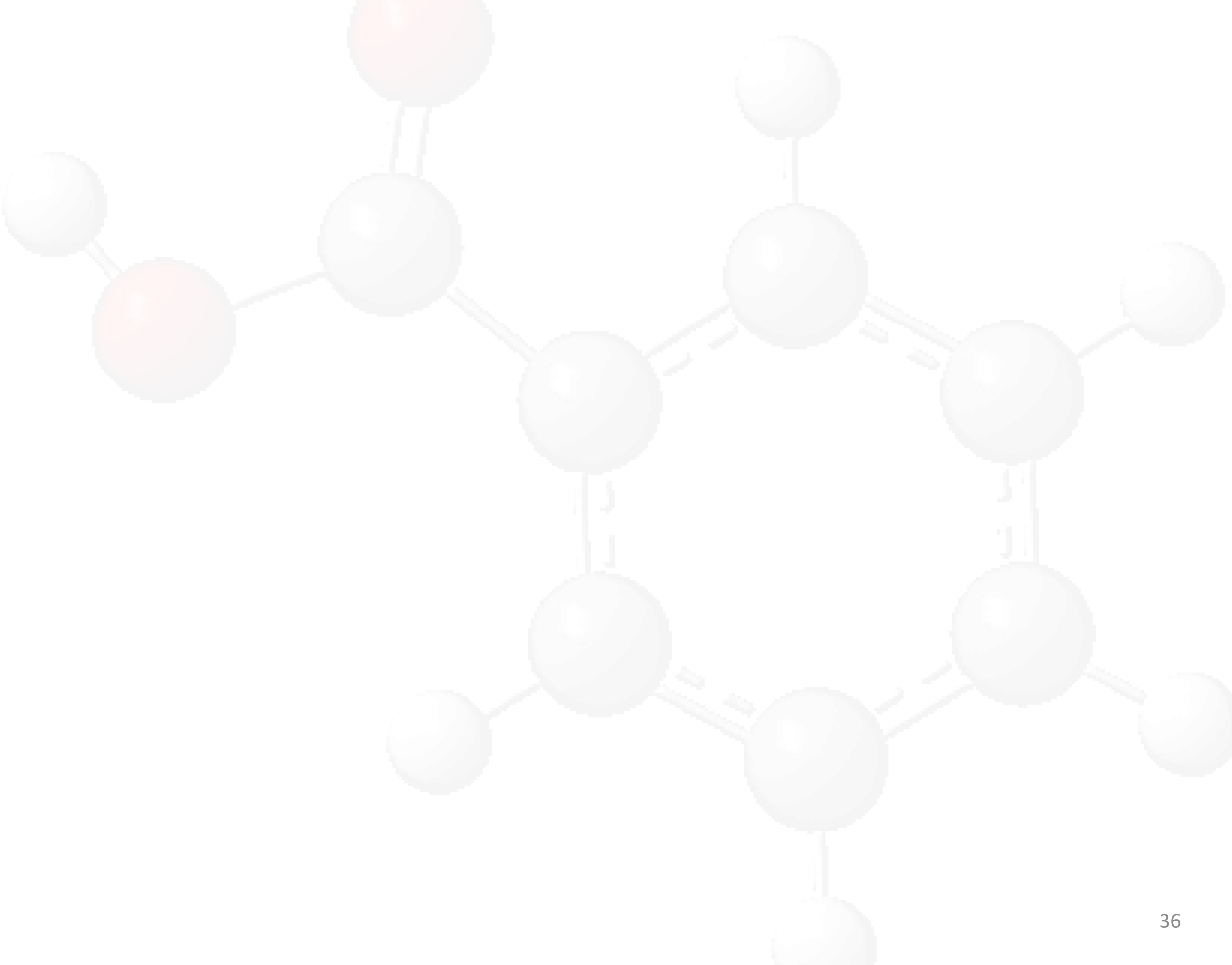
- Dr David Palmer (RISM)
- Professor Maxim Fedorov (RISM)
- Ms Neetika Nath (Machine Learning)
- Dr Luna De Ferrari
- Dr Herbert Früchtl
- Professor Michael Bühl

NSCCS

EaStCHEM
The Edinburgh and St Andrews Research School of Chemistry

University of St Andrews

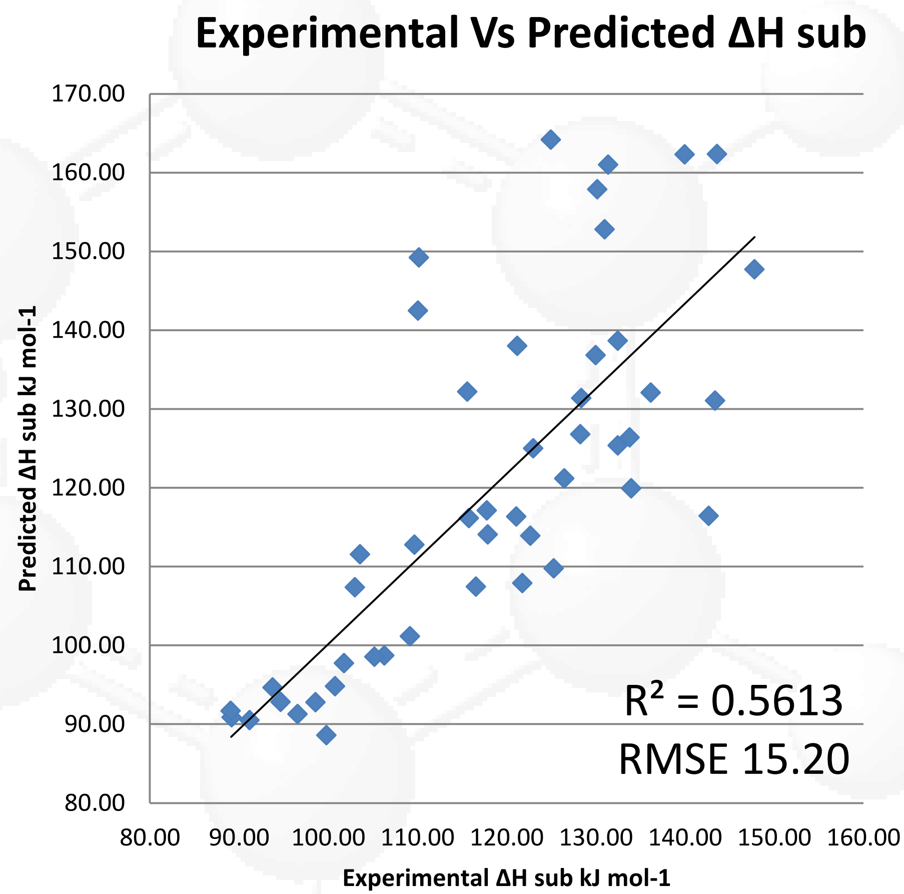
SULSA
Scottish Universities Life Sciences Alliance



Current Preliminary Results

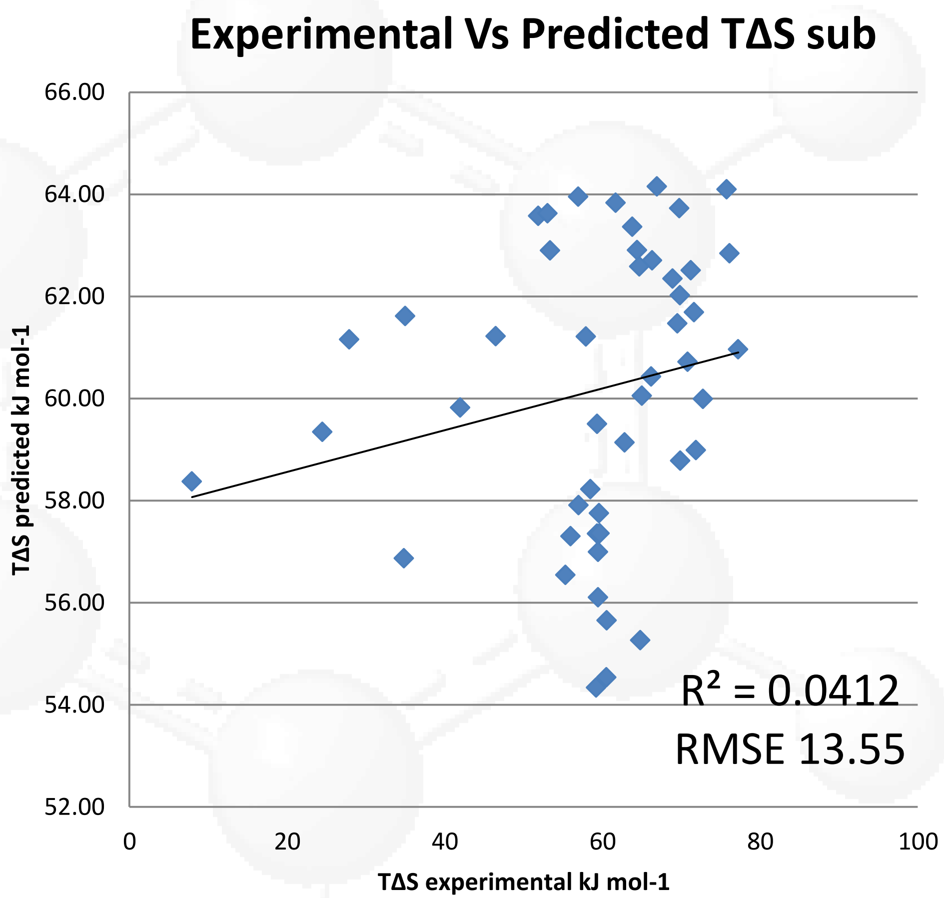
Enthalpy of Sublimation

- Using DMACRYS – B3LYP multipoles FIT potential and Gaussian 09.
- 48 molecules.
- A fair correlation but significant RMSE.



Entropy of Sublimation

- Crystal entropy calculated in DMACRYS, gas phase in Gaussian 09.
- No meaningful correlation.
- Significant RMSE.
- 48 molecules



Additional notes

- Buckingham potential
  $E_{lk} = Bexp(-Cr_{lk}) - Ar_{lk}^{-6}$
- Universal correction
  $\Delta G_{hyd}^{3DRISM-UC} =$
  $\Delta G_{hyd}^{3DRISM} + a(\rho V) + b$
- LogS
  $\frac{\left(\frac{\Delta G_{sol}}{-RT}\right)}{\ln(10)}=\log(S)$
  $\Delta G_{solu} = -RTlnS$
- 1logS unit = 5.71 in terms of ΔG

- Thermodynamics
- ΔH_{sub} = $-U_{latt}$ +2RT
- ΔS_{sub} = $(S_r+S_t) - S_{crys}$
- ΔG_{sub} = ΔH_{sub} - $T\Delta S_{sub}$
- ΔG_{hyd} = $E_{sol} - E_{gas}$
- ΔG_{solu} = $\Delta G_{sub} + \Delta G_{hyd}$
- $\Delta G^{GF} =$
  $K_B T \sum_{\alpha=1}^{Nsolvent} \rho_\alpha \int_{R^3} \left[c_\alpha(r) - \frac{1}{2} c_\alpha(r) h_\alpha(r)\right] dr$

$$\Delta G_{sol}^{o} = \Delta G_{sub}^{o} + \Delta G_{hyd}^{o} = -RTln(S_0 v_m)$$

Molar volume of the crystal Vm

Intrinsic solubility So