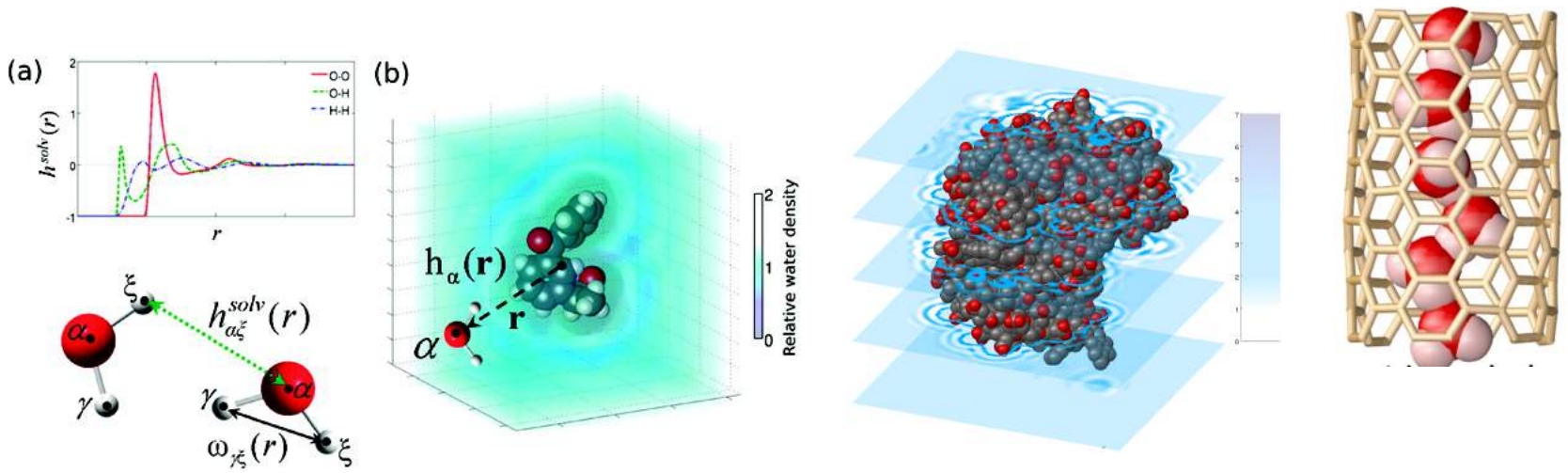
PhysChem Property Prediction from Molecular Theories of Solvation



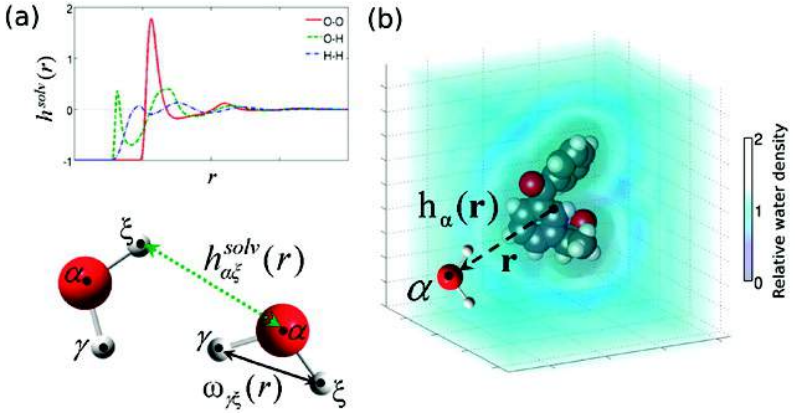
David S. Palmer

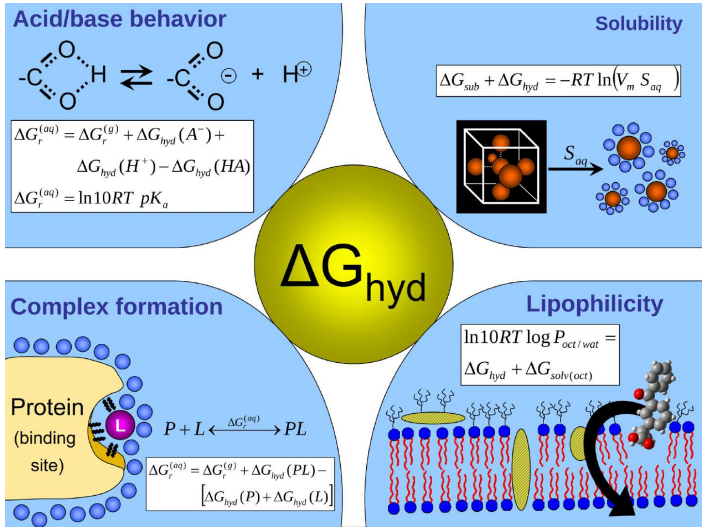
Department of Pure and Applied Chemistry, University of Strathclyde, Glasgow, UK.

University of Strathclyde Glasgow

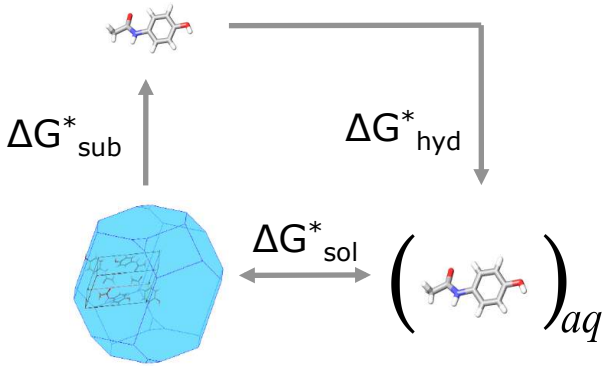
Overview

1. Introduction to Molecular Integral Equation Theory



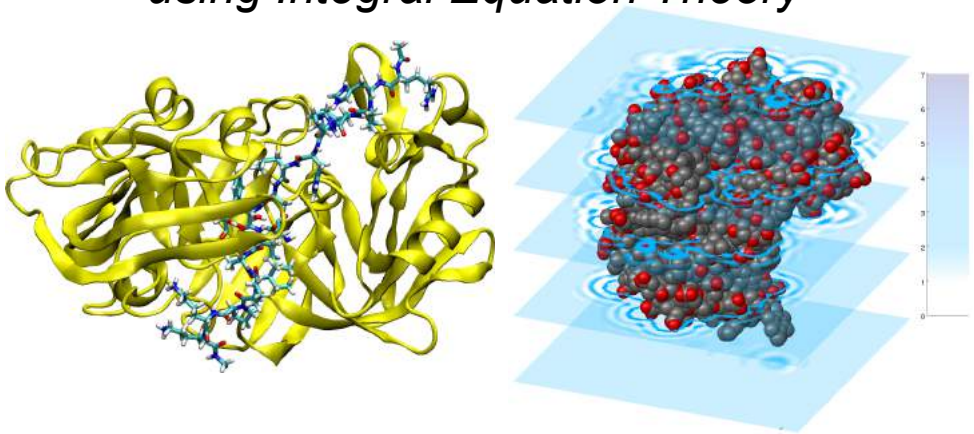


2. Prediction of the Intrinsic Aqueous Solubility of Druglike Molecules

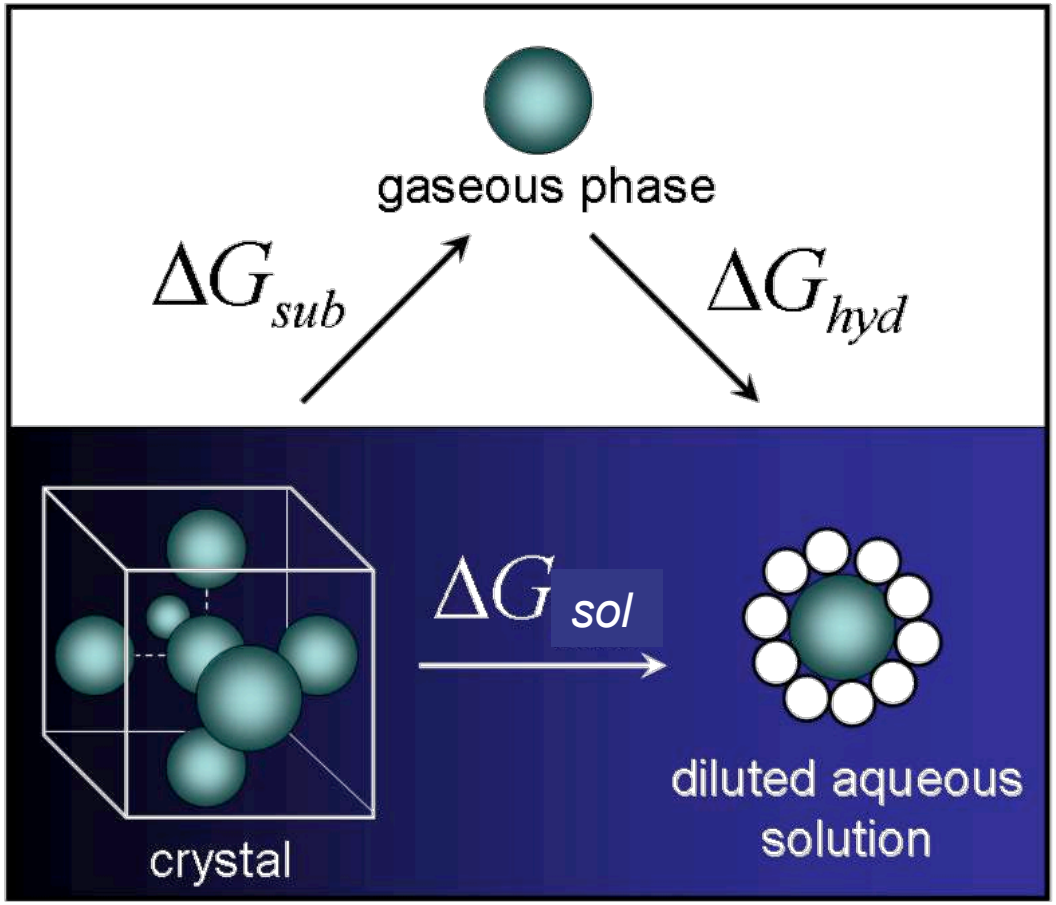


$$\Delta G^*_{sol} = \Delta G^*_{sub} + \Delta G^*_{hydr} = -RT \ln S_o V_m$$

3. Fast and General Property Prediction using Integral Equation Theory



Hydration Free Energy

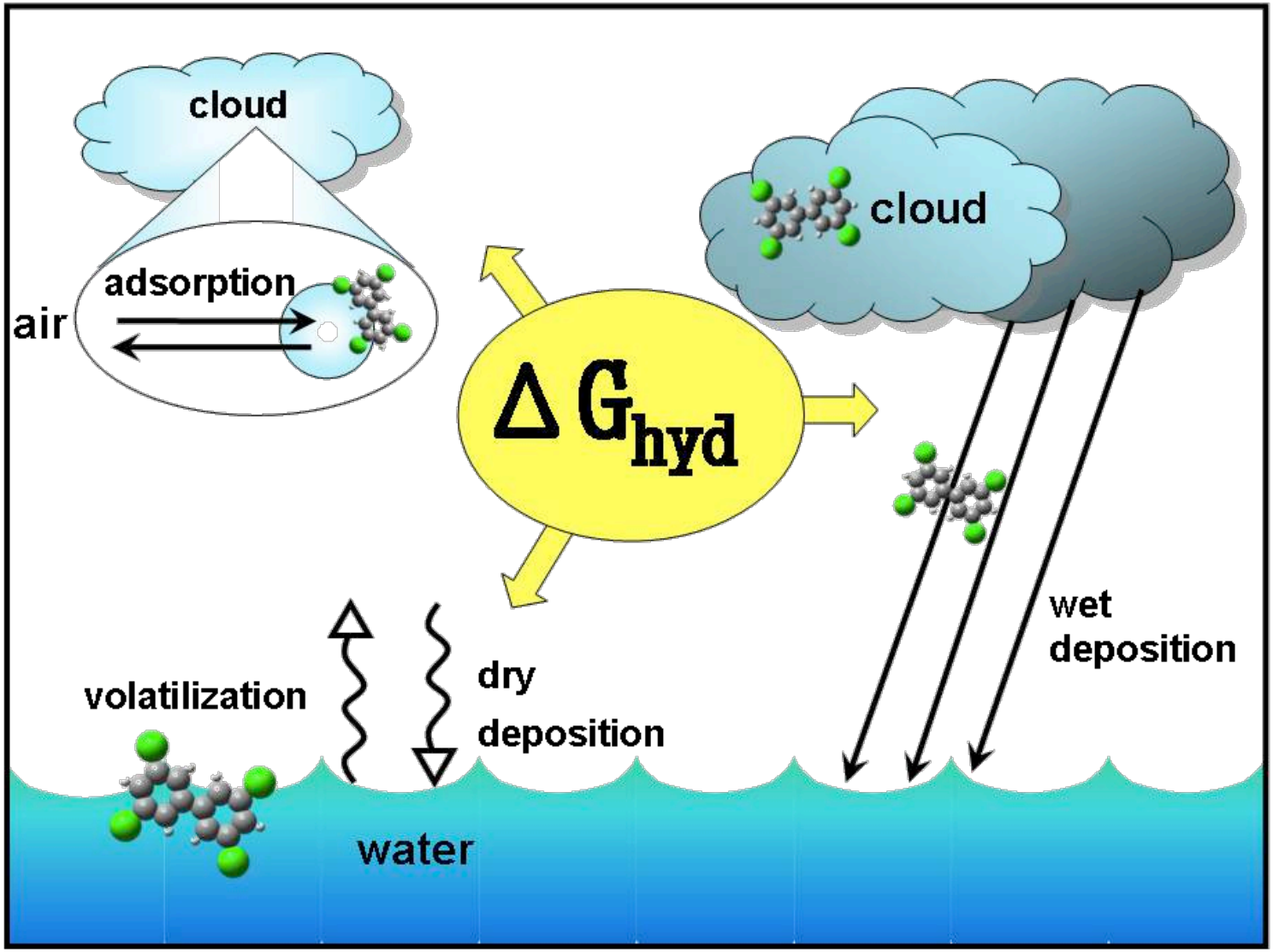


$$\Delta G_{hyd}$$

Hydration free energy is the change of the Gibbs free energy that accompanies the transfer of 1 mole of solute from gaseous phase to aqueous solution

$$\Delta G_{hyd} = -RT \ln K = -RT \ln \frac{[X]_{aq}}{[X]_g}$$

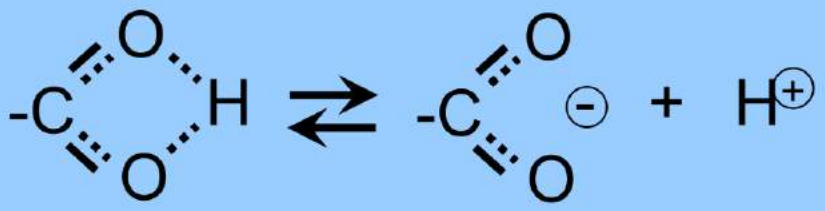
Ben-Naim A and Marcus Y 1984 *J. Chem. Phys.* ***81 2016–27***



Henry's Law Constant (K_H): $\Delta G_{hyd} = RT \ln K_H$

ΔG_{hyd}

Acid/base behavior

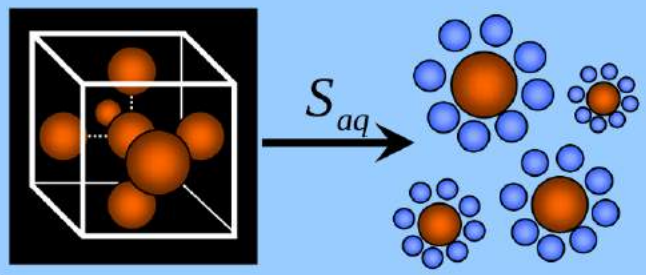


$$\Delta G_r^{(aq)} = \Delta G_r^{(g)} + \Delta G_{hyd}(A^-) + \Delta G_{hyd}(H^+) - \Delta G_{hyd}(HA)$$

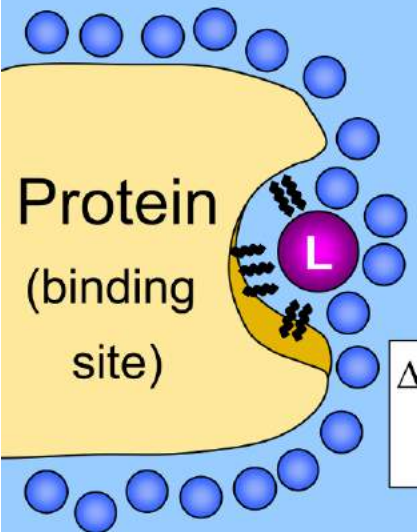
$$\Delta G_r^{(aq)} = \ln 10 RT \; pK_a$$

Solubility

$$\Delta G_{sub} + \Delta G_{hyd} = -RT \ln\left(V_m \; S_{aq}\right)$$



Complex formation

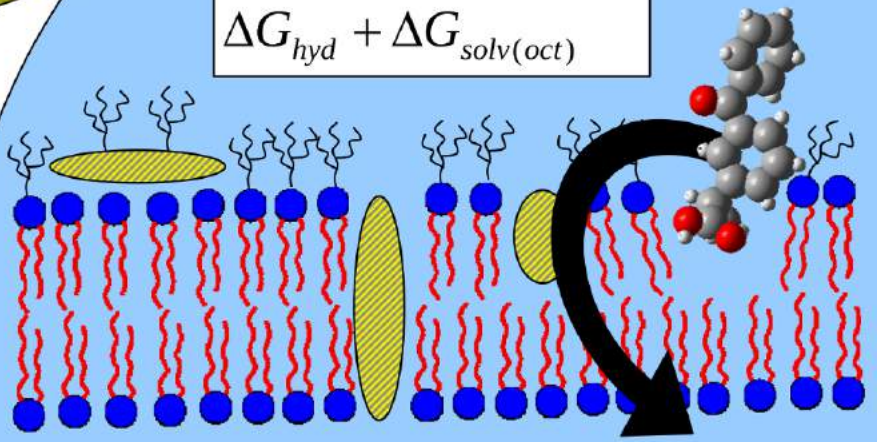


$$P + L \xleftrightarrow{\Delta G_r^{(aq)}} PL$$

$$\Delta G_r^{(aq)} = \Delta G_r^{(g)} + \Delta G_{hyd}(PL) - \left[\Delta G_{hyd}(P) + \Delta G_{hyd}(L)\right]$$

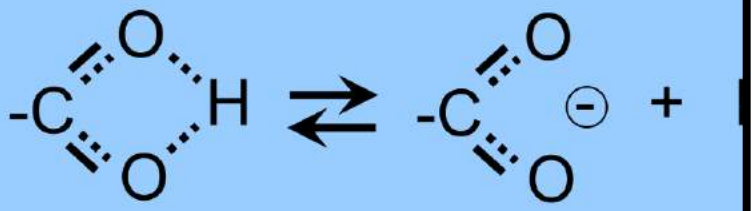
Lipophilicity

$$\ln 10 RT \log P_{oct/wat} = \Delta G_{hyd} + \Delta G_{solv(oct)}$$



ΔG_{hyd}

Acid/base behavior

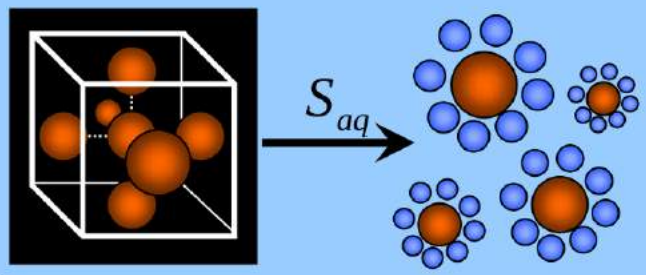


$$\Delta G_r^{(aq)} = \Delta G_r^{(g)} + \Delta G_{hyd}(A^-) + \Delta G_{hyd}(H^+) - \Delta G_{hyd}(HA)$$

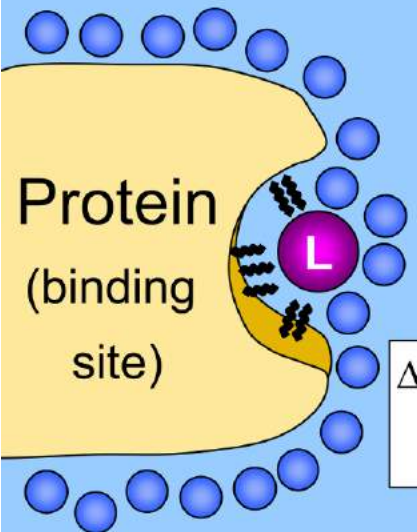
$$\Delta G_r^{(aq)} = \ln 10 RT \; pK_a$$

Solubility

$$\Delta G_{sub} + \Delta G_{hyd} = -RT \ln\left(V_m \; S_{aq}\right)$$

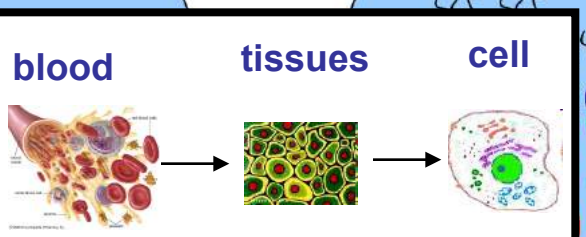


Complex formation



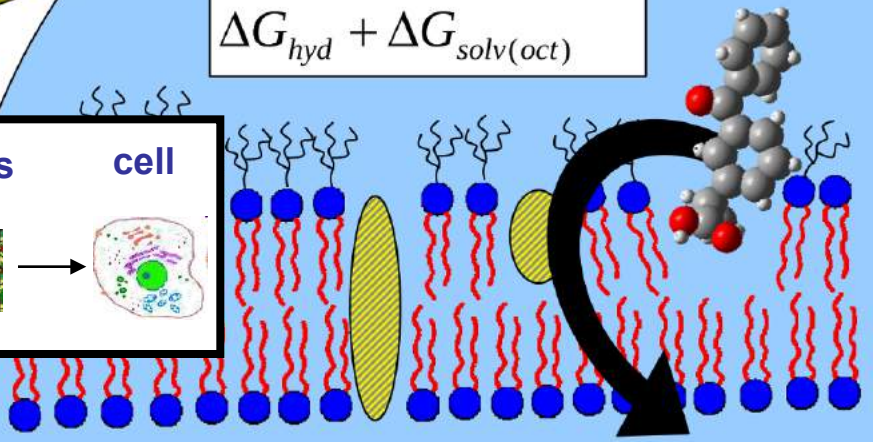
$$P + L \leftarrow \Delta G_r^{(}$$

$$\Delta G_r^{(aq)} = \Delta G_r^{(g)} + \left[\Delta G_{hyd}(P) + \Delta G_{hyd}(L)\right]$$



Lipophilicity

$$\ln 10 RT \log P_{oct/wat} = \Delta G_{hyd} + \Delta G_{solv(oct)}$$



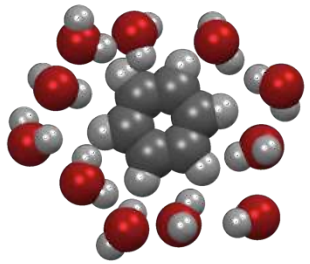
Methods

Methods for calculating thermodynamic parameters differ by description of the solvent structure

Solvent Models

Explicit Solvent (MD,MC)

Each solvent molecule is a separate atomistic object



accurate (?) but computationally expensive

Implicit Solvent

- Solvent distribution functions

Continuum electrostatics

fast but not always accurate enough

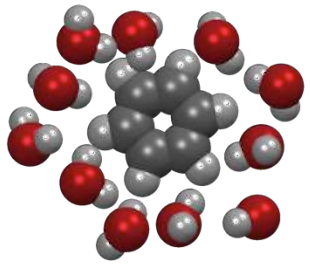
Methods

Methods for calculating thermodynamic parameters differ by description of the solvent structure

Solvent Models

Explicit Solvent (MD,MC)

Each solvent molecule is a separate atomistic object



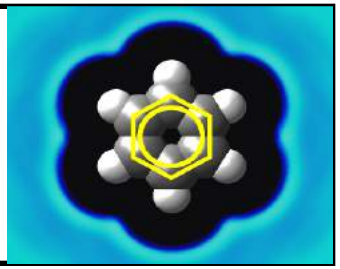
accurate (?) but computationally expensive

Implicit Solvent

- Solvent distribution functions

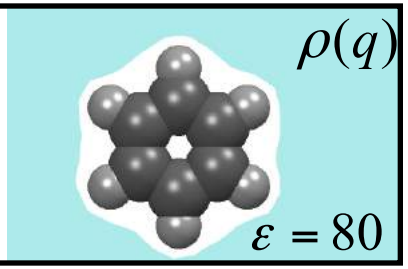
Integral Equations Theory of Liquids

we are working to make it fast and accurate



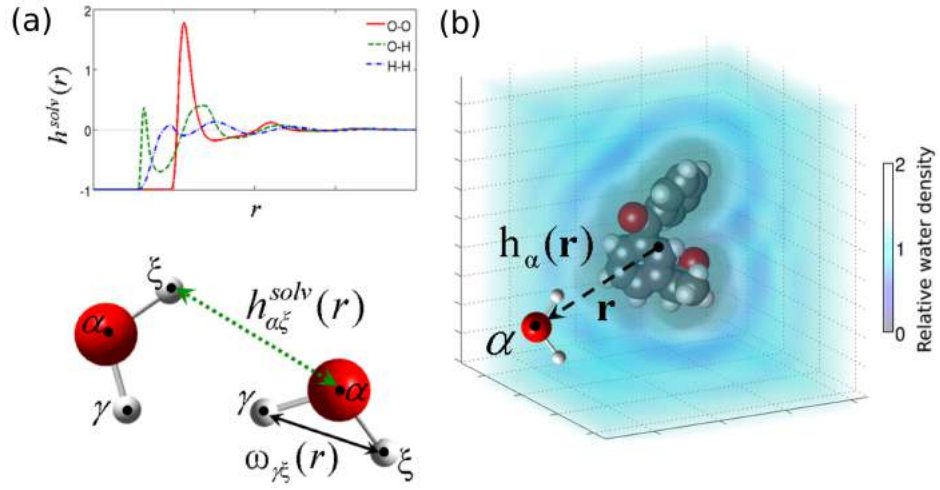
Continuum electrostatics

fast but not always accurate enough



3D Reference Interaction Site Model (3D RISM)

Solvent **Solution**



$$\chi_{\xi\alpha} = \omega_{\xi\alpha}(r) + \rho h_{\xi\alpha}(r)$$

$$h_\alpha(\mathbf{r}) = g_\alpha(\mathbf{r}) - 1$$

3D-RISM equation:

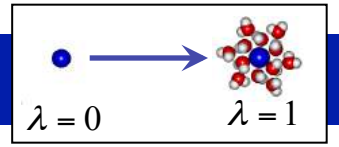
$$h_\alpha(\mathbf{r}) = \sum_{\xi=1}^{N_{solv}} \int_{R^3} c_\xi(\mathbf{r}-\mathbf{r}')\chi_{\xi\alpha}(|\mathbf{r}'|)\,d\mathbf{r}'$$

3D-RISM closure relationship:

$$h_\alpha(\mathbf{r}) = \exp\left(-\beta u_\alpha(\mathbf{r}) + h_\alpha(\mathbf{r}) - c_\alpha(\mathbf{r}) + B_\alpha(\mathbf{r})\right) + 1$$

bridge function

End-Point Calculations



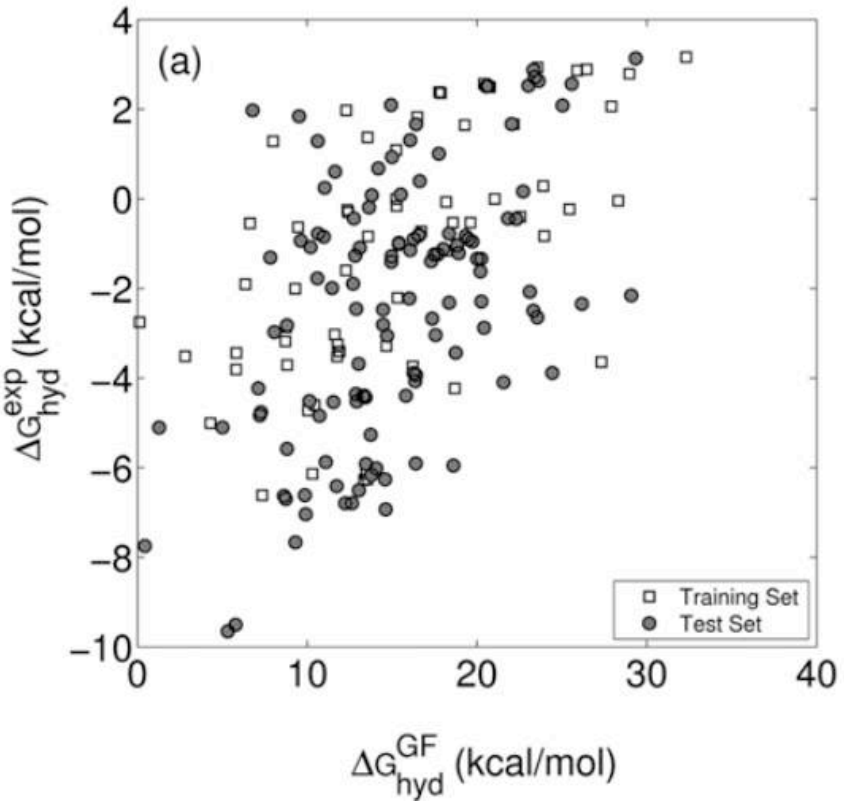
Hydration Free Energy:

$$\Delta G^{GF}_{hyd} = k_B T \sum_{\alpha=1}^{N_{solv}} \rho_\alpha \int_{R^3} \left[-c_\alpha(\mathbf{r}) - \tfrac{1}{2} c_\alpha(\mathbf{r}) h_\alpha(\mathbf{r})\right] d\mathbf{r}$$

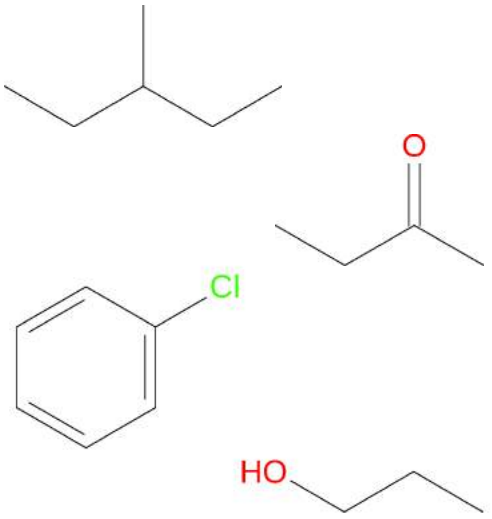
Partial Molar Volume:

$$\rho V = \rho k_B T \eta \left(1 - \rho \sum_{\alpha=1}^{N_{solv}} \int_{R^3} c_\alpha(\mathbf{r})\, d\mathbf{r}\right)$$

3D-RISM

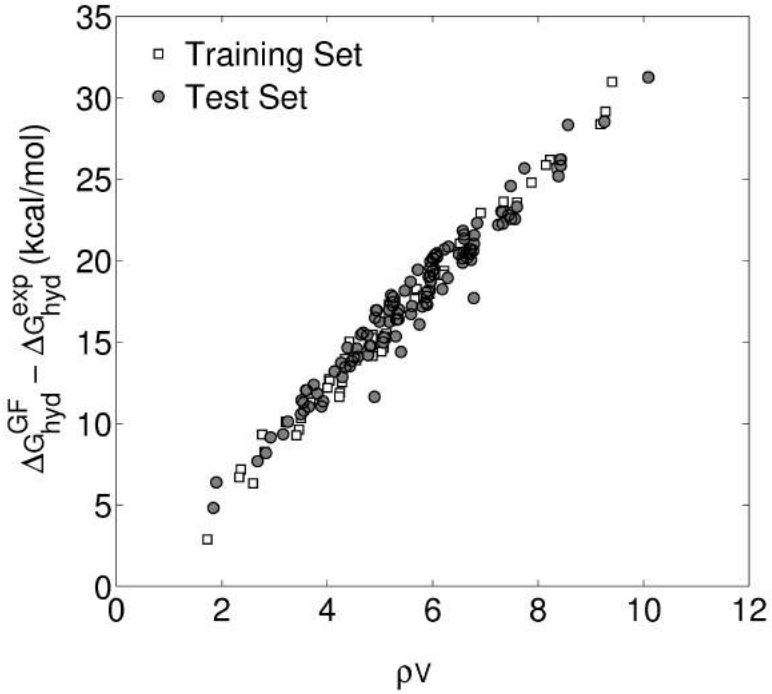


Simple Organic Molecules

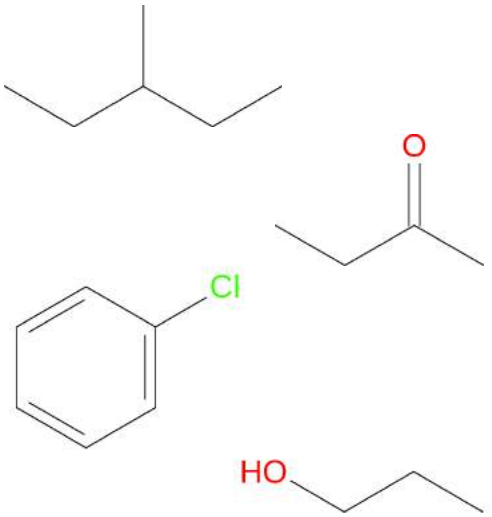


- 185 small organic molecules from multiple different chemical classes

3DRISM-UC

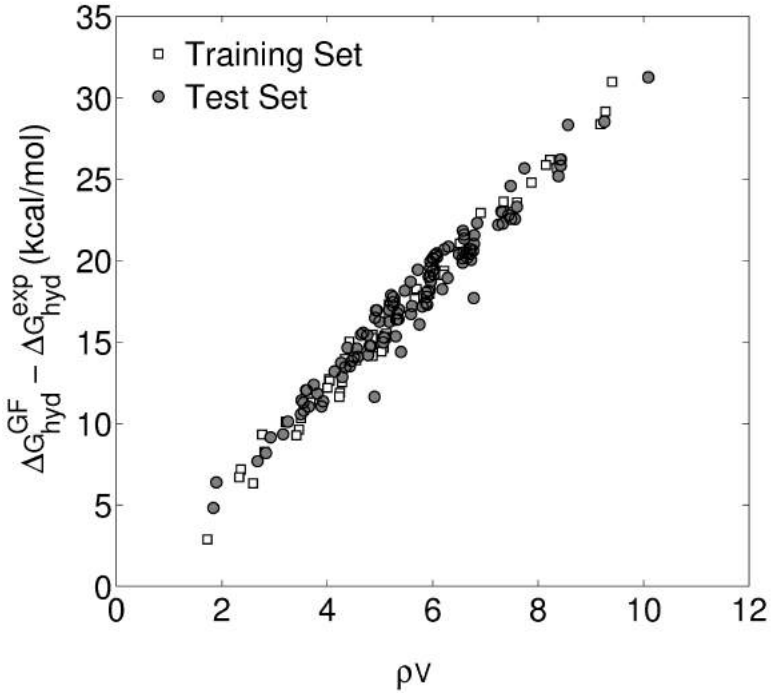


Simple Organic Molecules

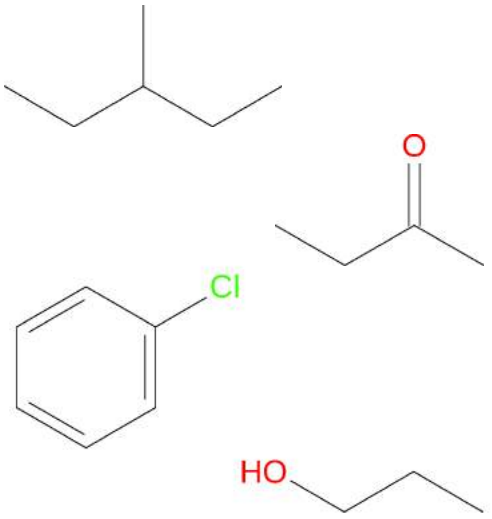


J. Phys. Cond. Matt., **2010**, *22*, 492101

3DRISM-UC



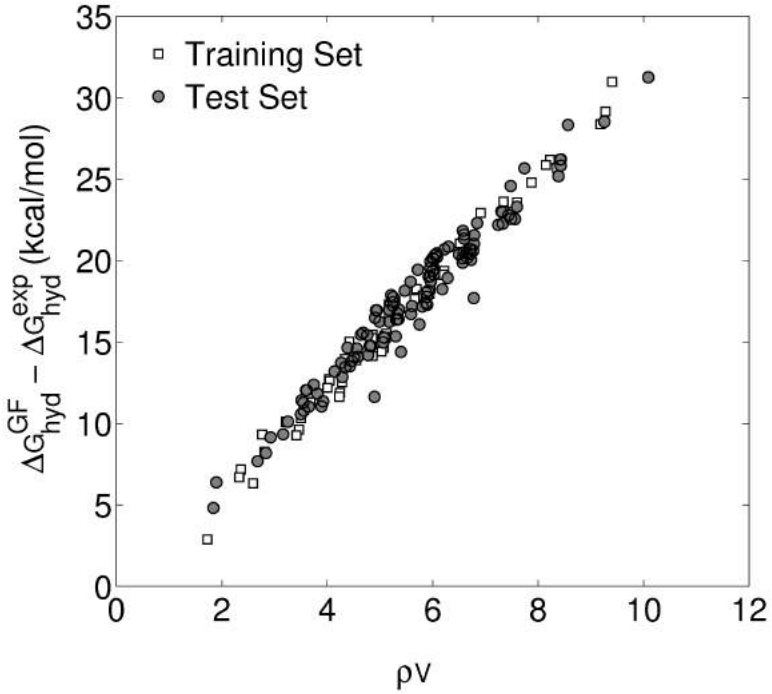
Simple Organic Molecules



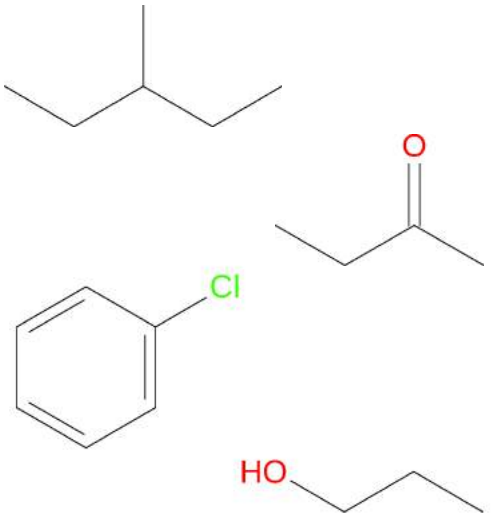
3DRISM-UC $\Delta G_{hyd}^{\text{3DRISM-UC}} = \Delta G_{hyd}^{\text{3DRISM-GF}} - 3.222(\rho V) + 0.578$

J. Phys. Cond. Matt., **2010**, *22*, 492101

3DRISM-UC



Simple Organic Molecules

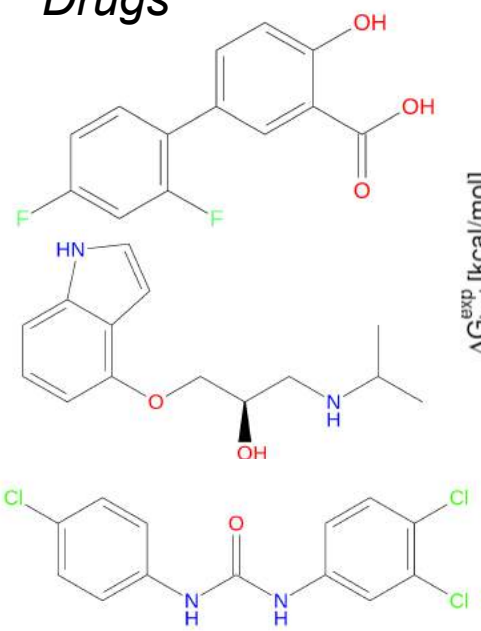


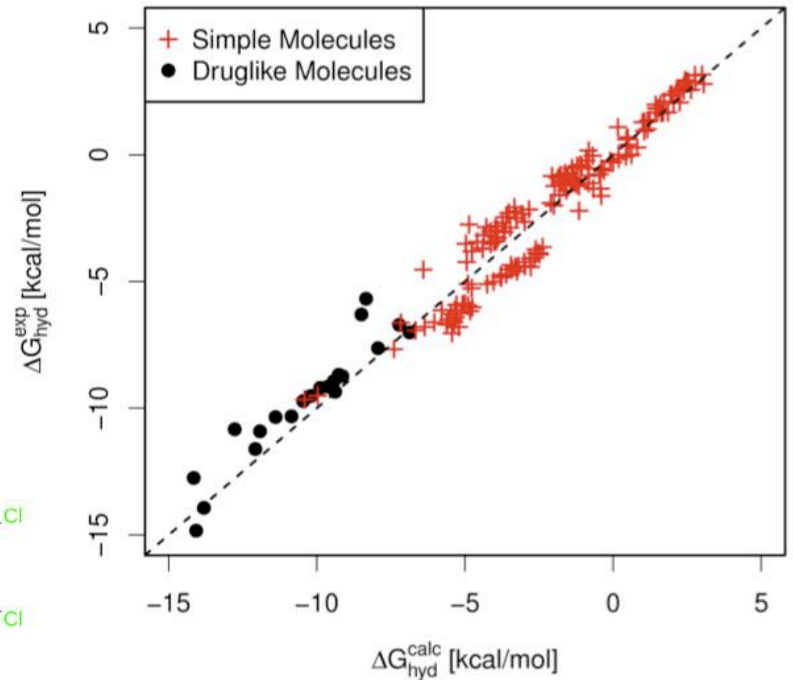
	Training dataset (n=65)						Test dataset (n=120)		
Model	R (tr)	σ (tr) (kcal mol^{-1})	Bias (tr) (kcal mol^{-1})	R (cv)	σ (cv) (kcal mol^{-1})	Bias (cv) (kcal mol^{-1})	R (te)	σ (te) (kcal mol^{-1})	Bias (te) (kcal mol^{-1})
RISM-UC	0.96	0.76	0.00	0.96	0.78	0.00	0.94	0.99	0.11

J. Phys. Cond. Matt., **2010**, *22*, 492101

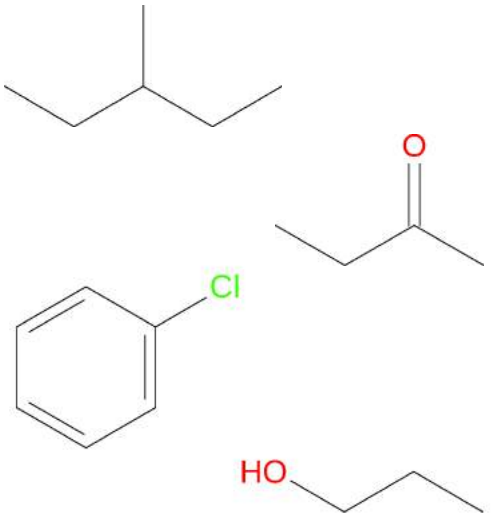
3DRISM-UC Druglike Molecules

Drugs





Simple Organic Molecules

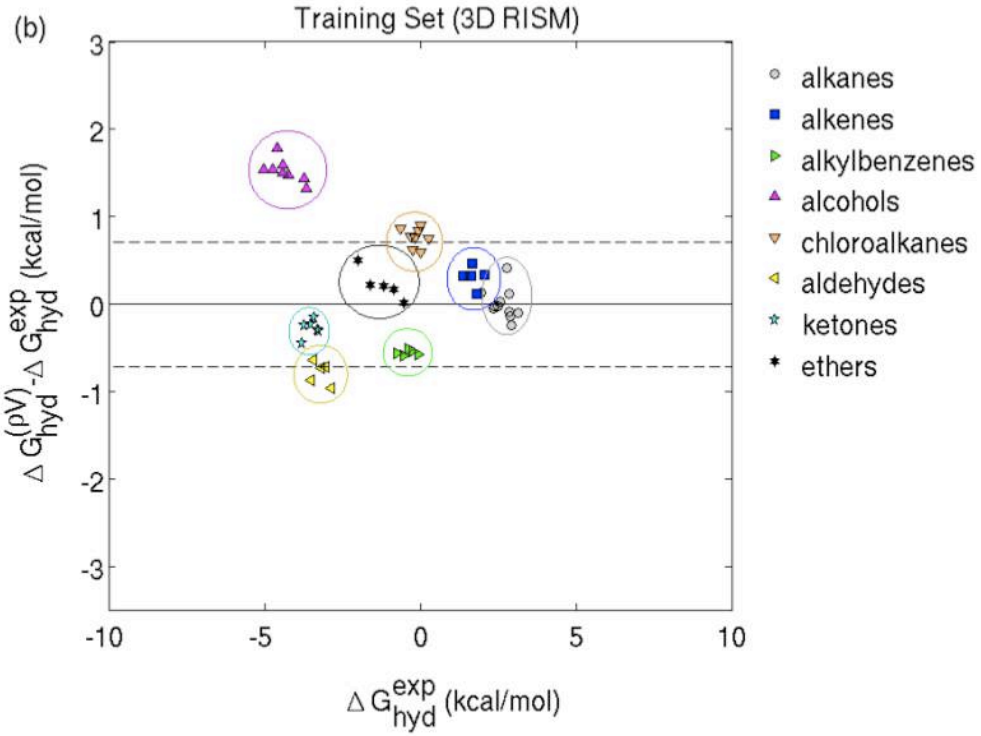


Method	R	RMSE (kcal/mol)	σ (kcal/mol)	bias (kcal/mol)
3DRISM/UC	0.94	1.06	0.78	-0.72

Mol. Pharmaceutics, **2011**, *8*, 1423

3DRISM-SDC

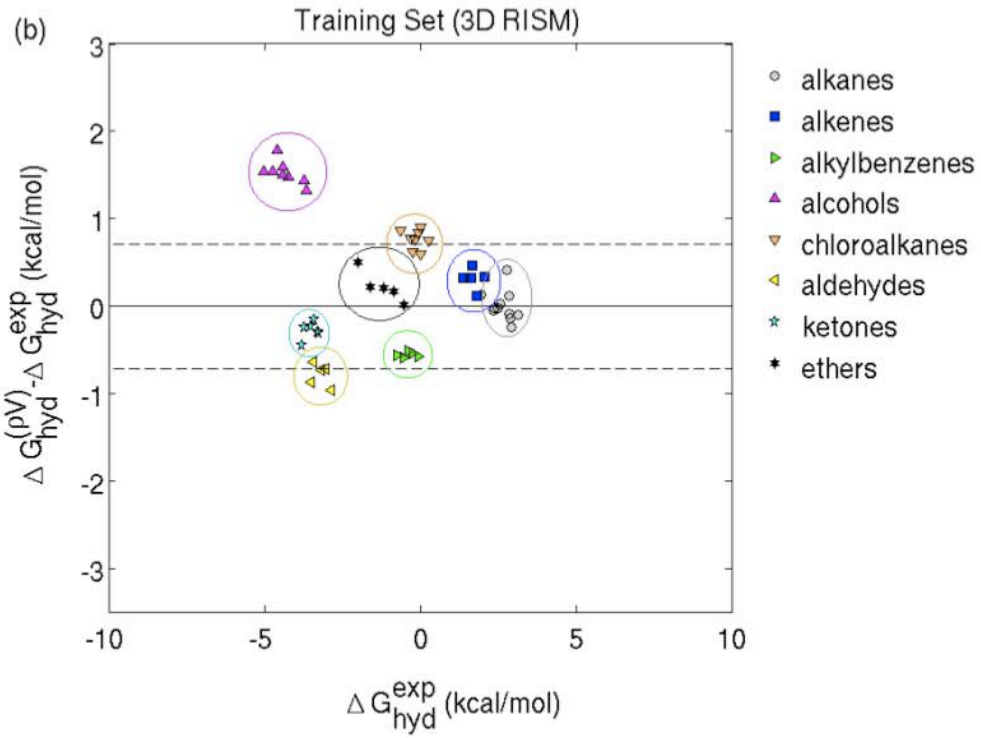
3DRISM-UC



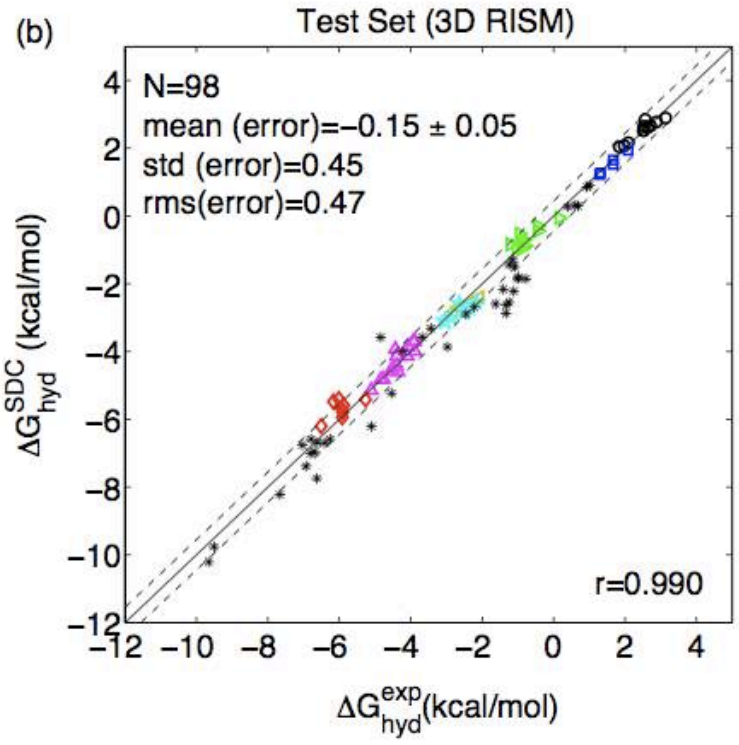
J Phys Chem B, **2011**, *115*, 6011–6022

3DRISM-SDC

3DRISM-UC



3DRISM-SDC



$$\Delta G_{hyd}^{\text{3DRISM-SDC}} = \Delta G_{hyd}^{\text{3DRISM-UC}} + \sum_i a_i x_i + c$$

functional group counts

J Phys Chem B, **2011**, *115*, 6011–6022

Several groups have suggested explanations for why 3D RISM/UC works

3D RISM / UC

$$\Delta G_{hyd,UC} = \Delta G_{KH} + a\rho\Delta V + b$$

2-P

"Cavity Correction" [1]

$$\Delta G_{hyd,NgB} = \Delta G_{KH} + a\frac{k_B T}{\rho}\int_{V_{in}} c_o^{np}(\vec{r})\,d\vec{r}$$

1-P

"Ensemble Correction" [2]

$$\Delta G_{hyd,ISc*} = \Delta G_{3DRISM} + \rho^2\frac{k_B T}{2}\hat{c}(k=0)\Delta V$$

0-P

"PC+ Correction" [3]

$$\Delta G_{hyd,PC} = \Delta G_{3DRISM} + \rho^2\frac{k_B T}{2}\hat{c}(k=0)\Delta V - \rho k_B T\Delta V$$

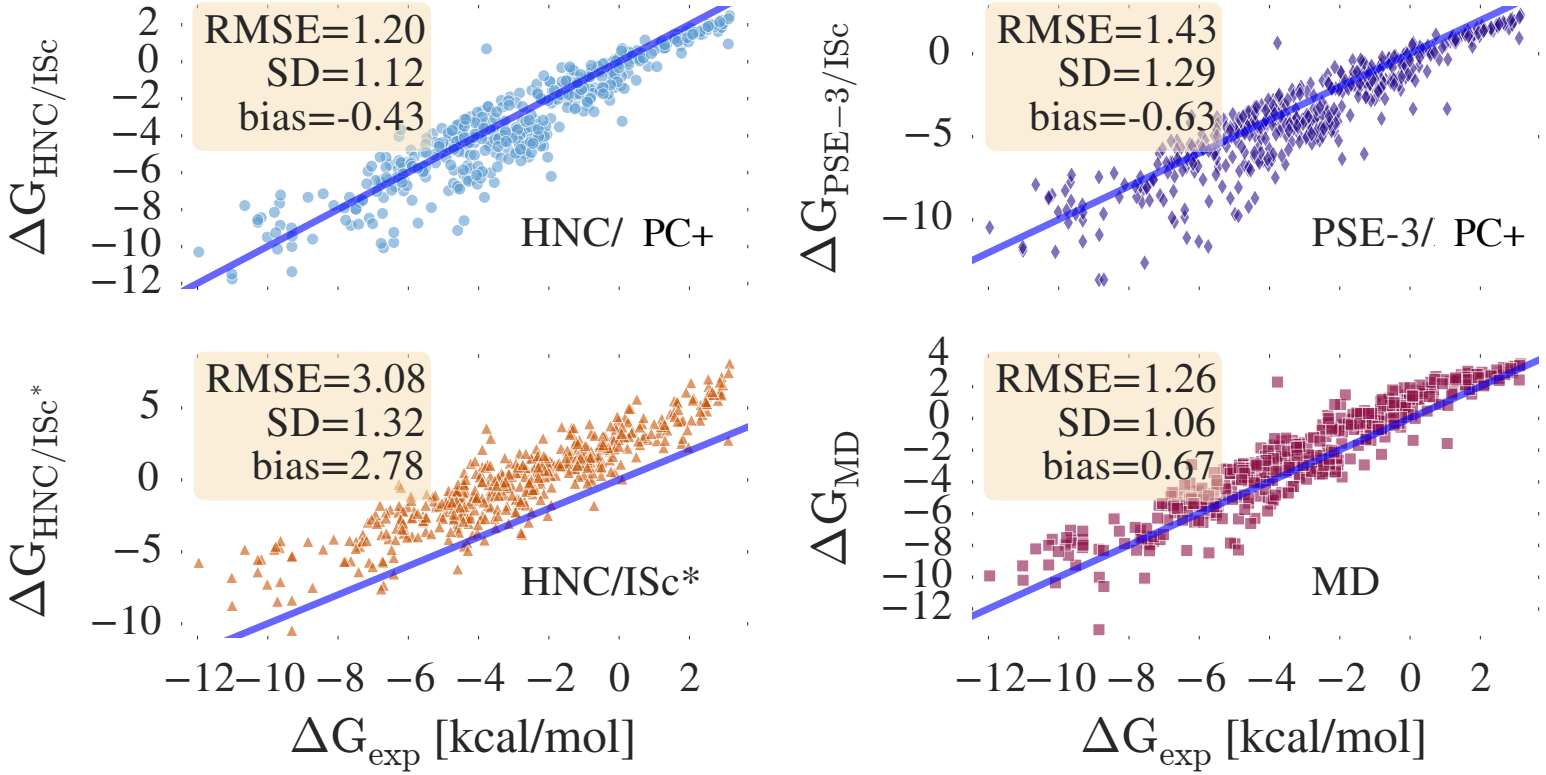
0-P

[1] Truchon, J-F; Pettitt, B. M; Labute, P; J. Chem. Theory Comput., ***2014***, *10*, 934

[2] Sergiievskyi, V. P.; Jeanmairet, G.; Levesque, M.; Borgis, D.; J. Phys. Chem. Lett., **2014**, *5*, 1935

[3] Mišin, M.; Fedorov, M. V.; Palmer D. S. J. Chem. Phys., **2015**, *in press.*

Accurate hydration free energies without empirical parameters using 3DRISM/PC+

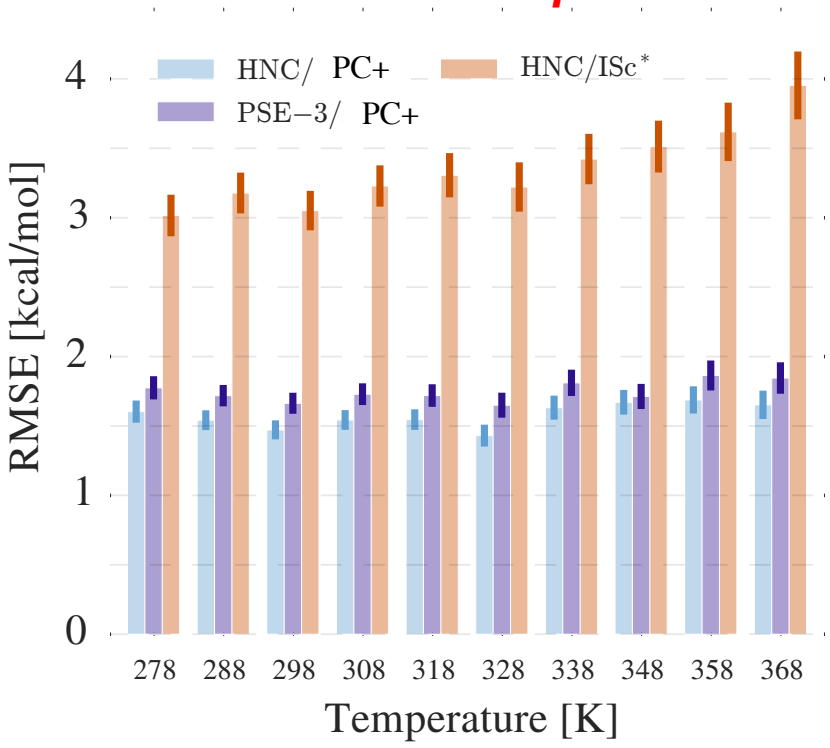


Hydration free energy data for 502 small organic molecules

J. Chem. Phys., **2015**, 142, 091105.

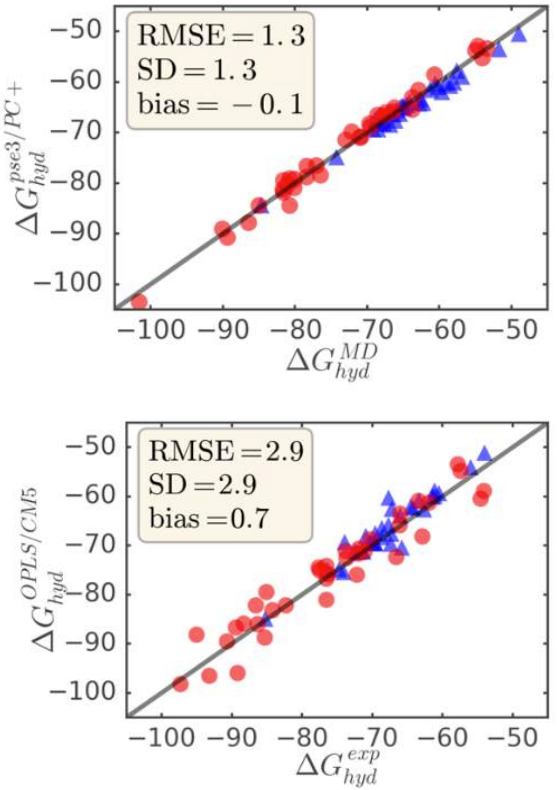
3DRISM/PC+ gives accurate hydration free energies for *neutral* and *ionized* solutes at *ambient* and *non-ambient* temperatures

Non-ambient temperatures



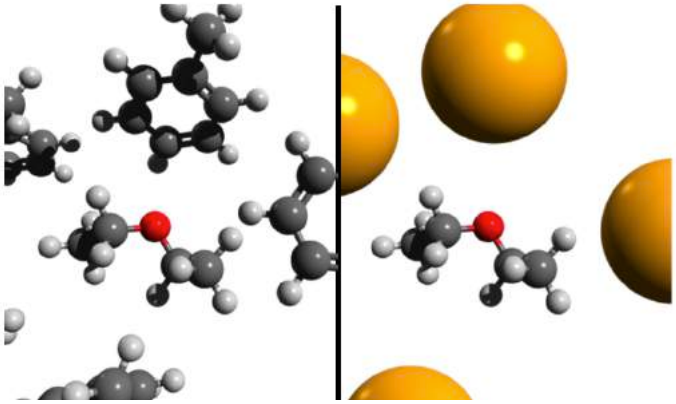
J. Chem. Phys., **2015**, 142, 091105.

Ionized small organic molecules



J. Phys. Chem. B. **2016**, *120*, 975–983

Non-polar solvents can be modelled as Lennard-Jones spheres



Lennard-Jones potential (solvent):

$$U(r) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$

According to the corresponding states principle, for classical fluids with orientation independent interaction potentials, reduced critical temperature and reduced critical density are given by:

$$T_c^* = \frac{kT_c}{\epsilon}$$

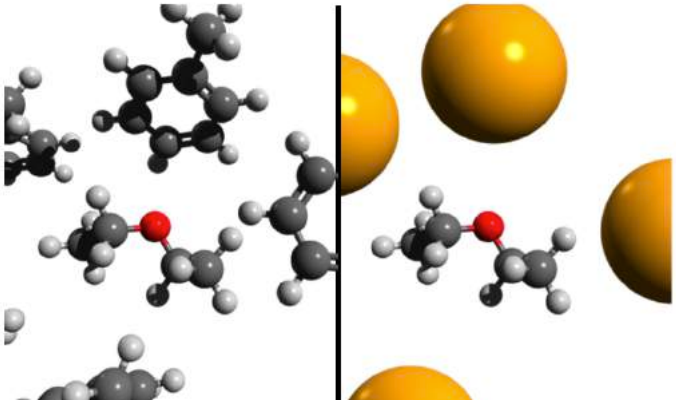
$$\rho_c^* = \rho_c\sigma^3$$

We take values of T* and ρ* from the work of Okumura and T and ρ from standard data tables.

$$T_{c,LJ}^* = 1.313,\ \rho_{c,LJ}^* = 0.304$$

Okumura, H.; Yonezawa, F. *J. Chem. Phys.* **2000**, *113*, 9162-9168

Non-polar solvents can be modelled as Lennard-Jones spheres



Lennard-Jones potential (solvent):

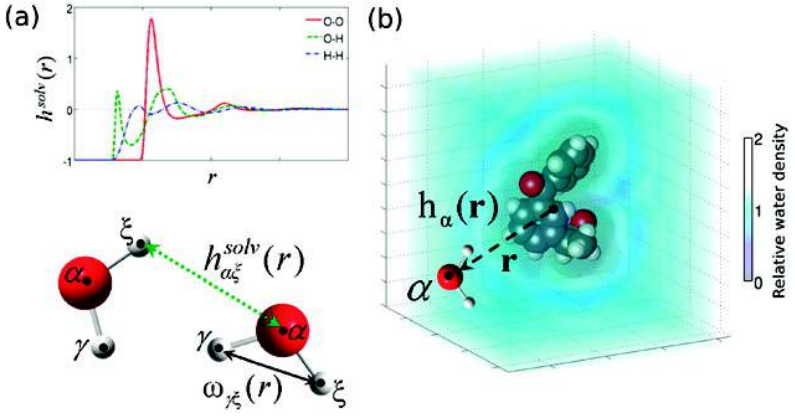
$$U(r) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$

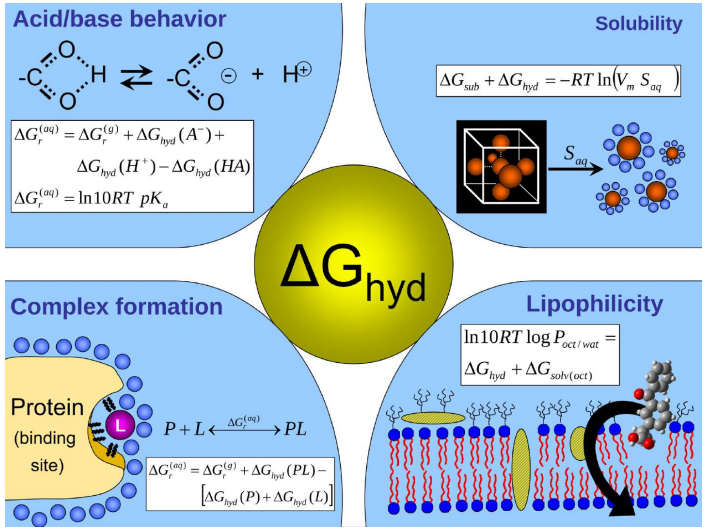
Solvent	N	RMSE	SDE	bias
	Apolar			
1,2-dichloroethane	39	1.16	1.07	0.47
benzene	71	1.28	1.28	0.04
bromobenzene	27	1.17	1.15	-0.23
carbon disulfide	15	0.94	0.89	-0.30
carbon tetrachloride	79	0.85	0.84	-0.11
cyclohexane	103	1.01	0.75	-0.67
isooctane	32	0.98	0.68	-0.70
n-decane	39	1.70	1.23	-1.17
n-decane (4-mer)	39	0.68	0.56	-0.38
n-heptane	67	0.95	0.86	-0.42
n-heptane (3-mer)	67	0.74	0.74	0.05
toluene	51	1.00	0.99	0.08
xylene (mixture)	48	1.00	0.99	-0.10

Misin, M.; Fedorov, M.V.; Palmer, D.S. *J. Phys. Chem. Lett.* **2016**, *submitted.*

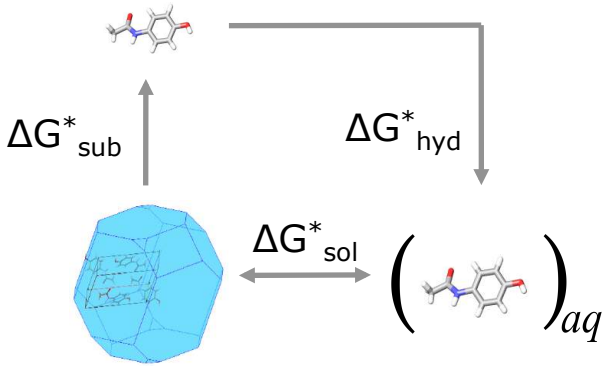
Overview

1. Introduction to Molecular Integral Equation Theory



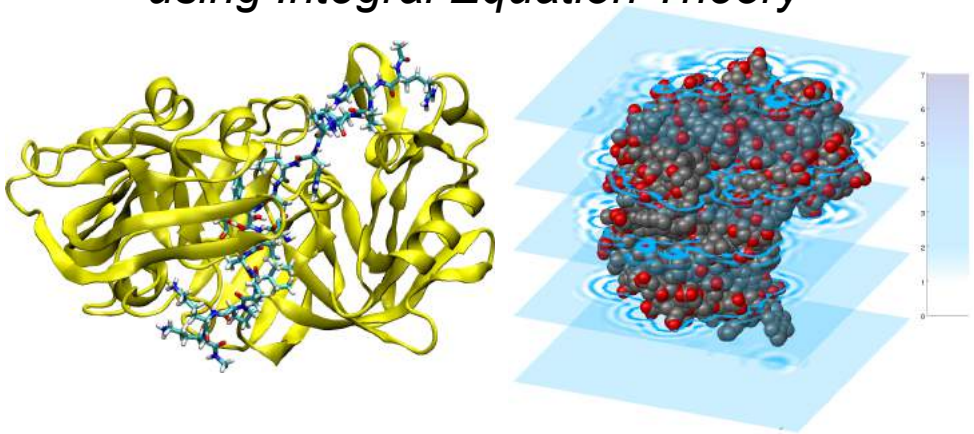


2. Prediction of the Intrinsic Aqueous Solubility of Druglike Molecules

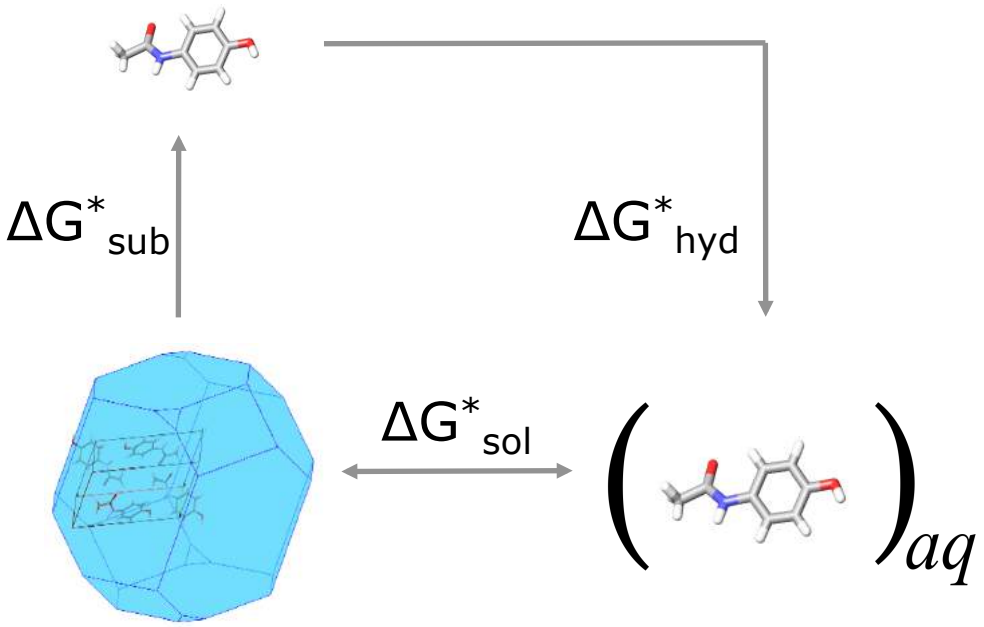


$$\Delta G^*_{sol} = \Delta G^*_{sub} + \Delta G^*_{hydr} = -RT \ln S_o V_m$$

3. Fast and General Property Prediction using Integral Equation Theory



Calculation of Solubility using 3D RISM / UC



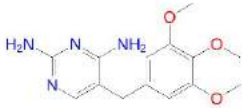
$$\Delta G^*_{sol} = \Delta G^*_{sub} + \Delta G^*_{hydr} = -RT \ln S_o V_m$$

ΔG^*_{sub} calculated by crystal lattice energy minimization and lattice dynamics.
ΔG^*_{hyd} computed using 3D RISM/UC
Dataset of 25 druglike molecules

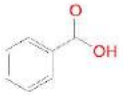
Palmer, DS; McDonagh, J; Mitchell, JBO; van Mourik T; Fedorov MV; J. Chem. Theory Comput., **2012**, *8*, 3322
Palmer DS, Llinàs A, Morao I, Day GM, Goodman JM, Glen RC, Mitchell JBO, Mol Pharmaceutics, **2008**, *5*, 266

allopurinol
ALOPUR

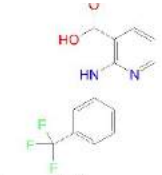
4-aminobenzoic acid
AMBNAC04



trimethoprim
AMXBPM10



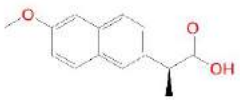
benzoic acid
BENZAC02



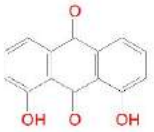
niflumic acid
NIFLUM10

benzamide
BZAMID02

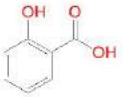
cocaine
COCAIN10



naproxen
COYRUD11

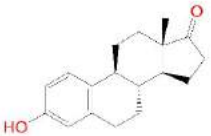


danthron
DHANQU06

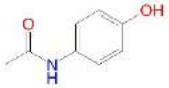


salicylic acid
SALIAC

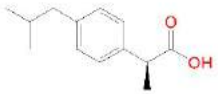
primidone
EPHPMO



estrone
ESTRON14



acetaminophen
HXACAN04



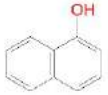
ibuprofen
IBPRAC01

clozapine
NDNHCL01

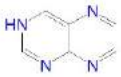
fluconazole
IVUQOF

isoproturon
JODTUR01

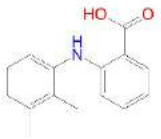
nitrofurantoin
LABJON01



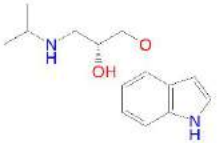
1-naphthol
NAPHOL01



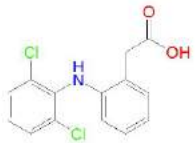
pteridine
PTERID11



mefenamic acid
XYANAC

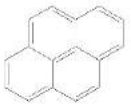


pindolol
PINDOL



diclofenac
SIKLIH01

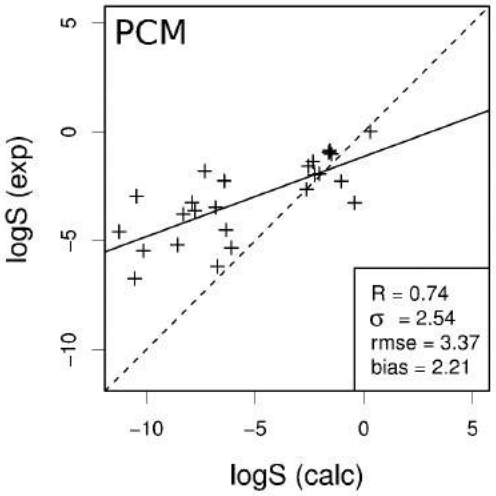
nicotinic acid
NICOAC02

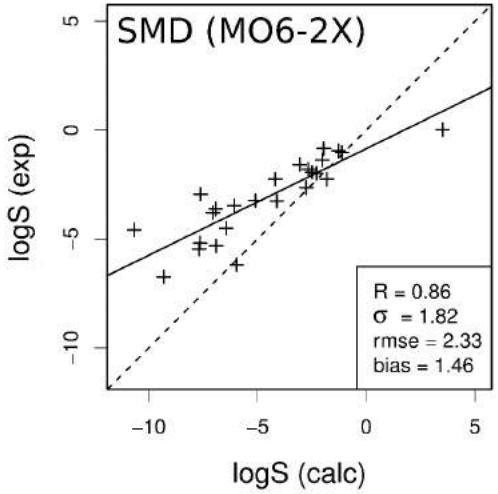


pyrene
PYRENE07

Solubility

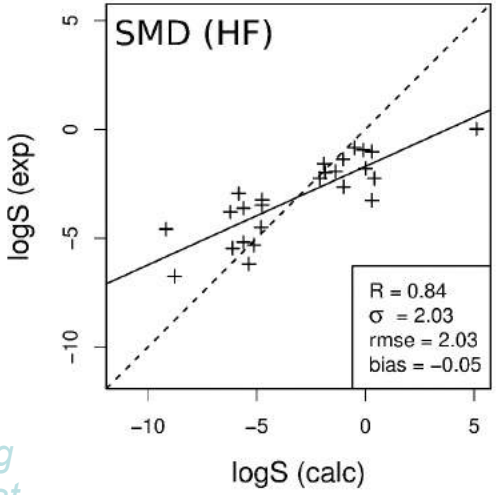
*HF/6-31G**
PCM

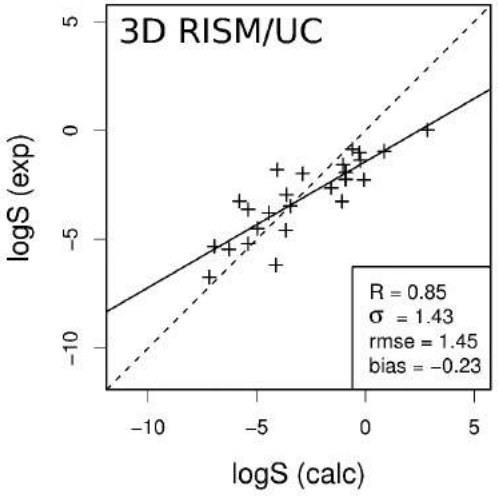




*M06-2X/6-31G**
SMD

*HF/6-31G**
SMD



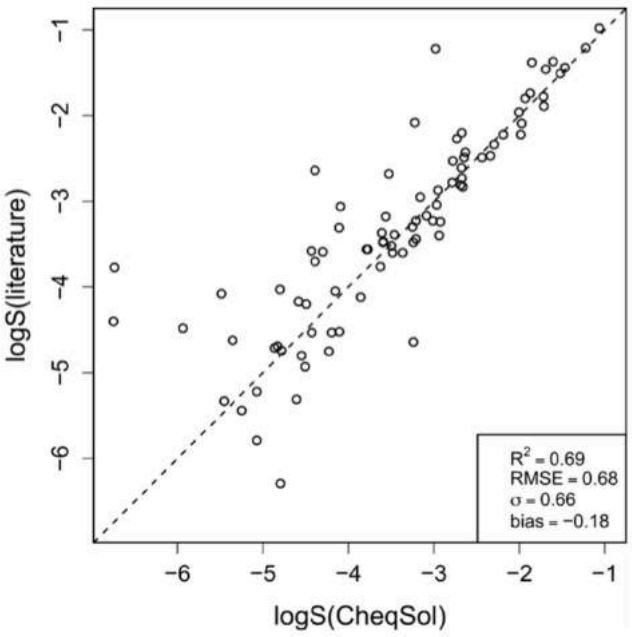


3D RISM/UC

*ΔG_{sub} calculated using multipoles evaluated at the B3LYP/6-31G** level*

σ ($logS_{exp}$) = 1.8

Is experimental data quality the limiting factor in predicting the aqueous solubility of druglike molecules?



n=85

We tested *Random Forest* quantitative structure−property relationships (QSPR) that were trained upon the same druglike molecules but two different sets of solubility data:

(i) data extracted from several different sources from the published literature, for which the experimental uncertainty is estimated to be 0.6−0.7 log S units (referred to mol/L);

(ii) data measured by a single accurate experimental method (CheqSol), for which experimental uncertainty is typically <0.05 log S units.

Palmer D.S.; Mitchell, J. B. O. *Mol. Pharmaceutics* **2014**, *11*, 2962−2972

Is experimental data quality the limiting factor in predicting the aqueous solubility of druglike molecules?

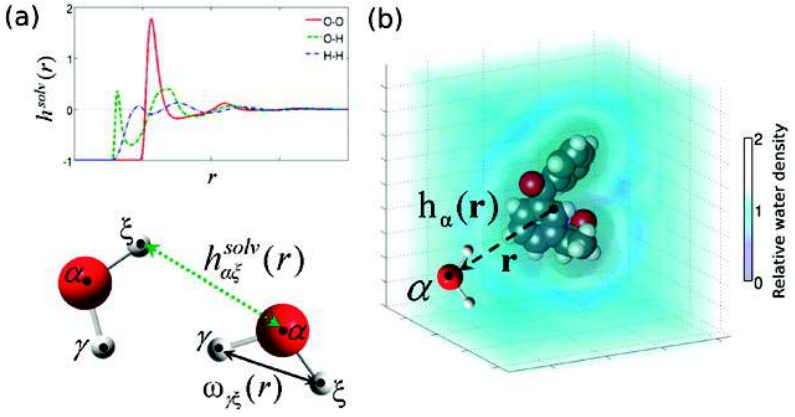
	r^2(oob)	RMSE(oob)	bias(oob)	r^2(te)	RMSE(te)	bias(te)	σ(tr)	σ(te)	mean(tr)	mean(te)
CheqSol	0.45	0.90	0.03	0.45	0.88	0.09	1.23	1.24	−3.48	−3.44
Literature	0.42	0.85	0.01	0.44	0.89	0.08	1.13	1.22	−3.32	−3.29

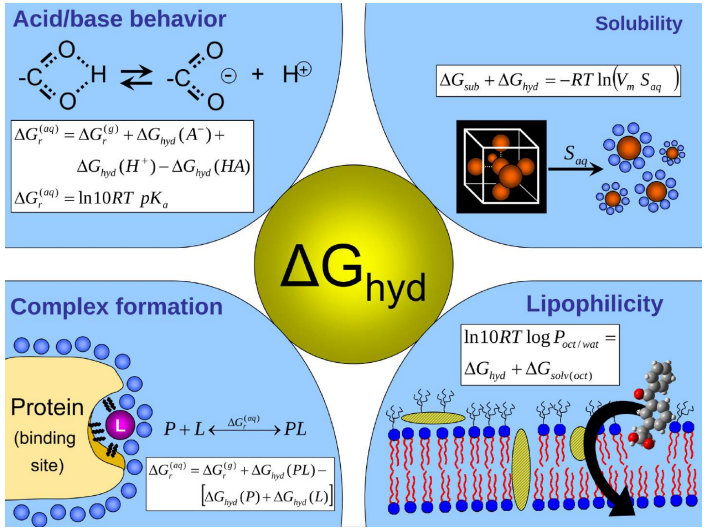
The results suggest that, at the present time, it is the deficiency of QSPR methods (algorithms and/or descriptor sets), and not, as is commonly quoted, the uncertainty in the experimental measurements, which is the limiting factor in accurately predicting aqueous solubility for pharmaceutical molecules.

Palmer D.S.; Mitchell, J. B. O. *Mol. Pharmaceutics* **2014**, *11*, 2962−2972

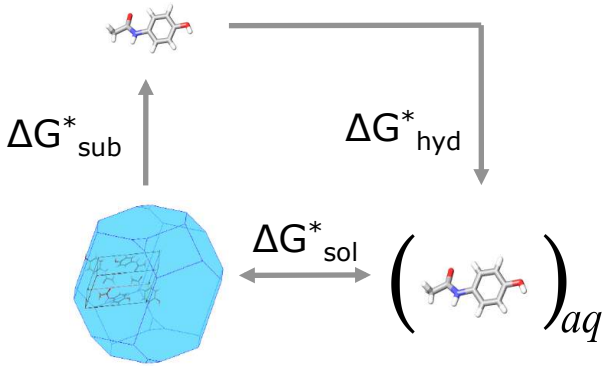
Overview

1. Introduction to Molecular Integral Equation Theory



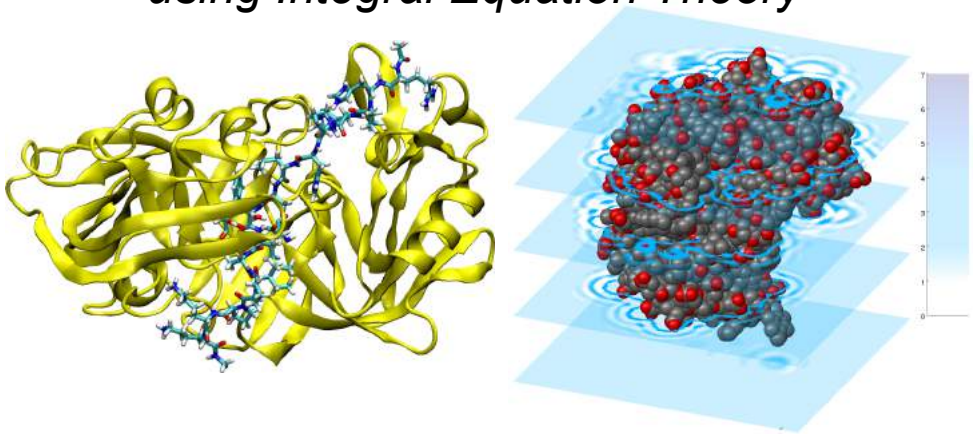


2. Prediction of the Intrinsic Aqueous Solubility of Druglike Molecules



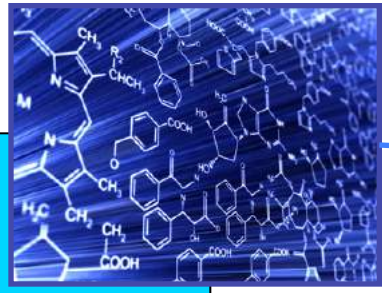
$$\Delta G^*_{sol} = \Delta G^*_{sub} + \Delta G^*_{hydr} = -RT \ln S_o V_m$$

3. Fast and General Property Prediction using Integral Equation Theory



Database screening in computational biomolecular design

Predictions should be:
- accurate
- *fast*



Databases

(molecular structures)

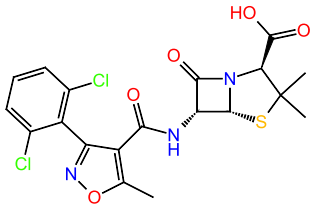
Filters

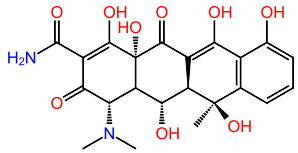
selected molecules are **analyzed** to find new molecules with similar properties

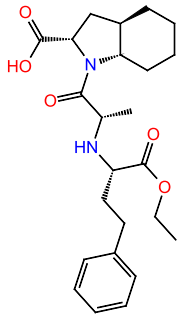
Predictions of biophysical/biomedical properties

(blood brain barrier permeability, human intestinal absorption rate, solubility, partition coefficients, toxicity, etc)

Quantitative Structure-Property Relationship (QSPR)



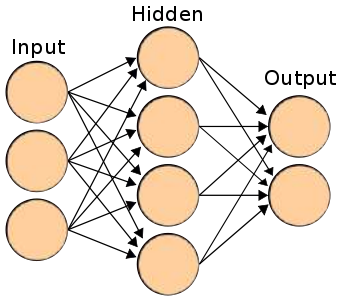


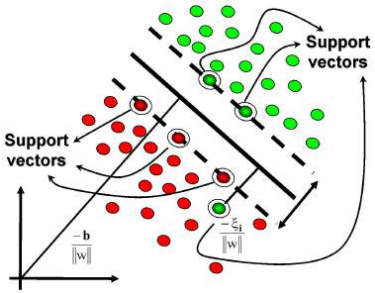


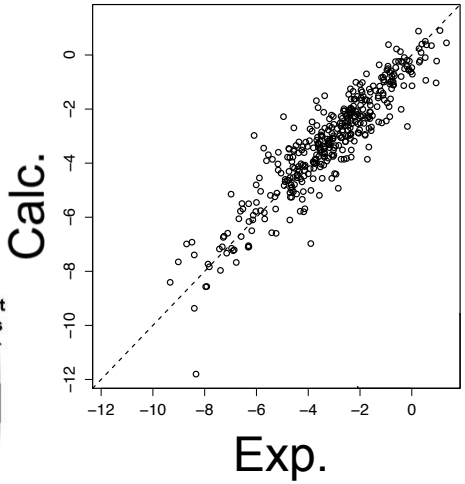
Mol	Y	X_1	...	X_N
1				

Mol	Y	X_1	...	X_N
2				

Mol	Y	X_1	...	X_N
3				





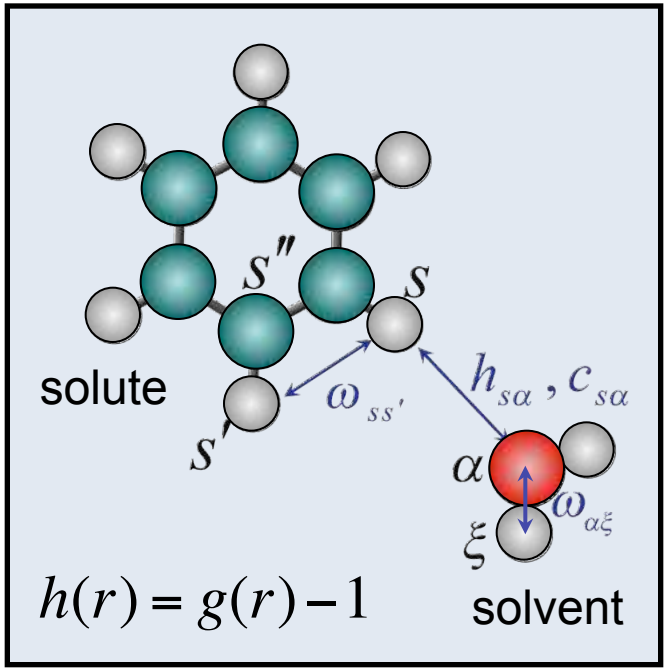


Data Set → **Variables** → **Model** → **Prediction**

Experimental Data + Molecular Structures

Molecular Descriptors

MLR, PLS, SVM, RF, etc



RISM equations:

$$h_{s\alpha}(r) = \sum_{s'=1}^{N_{solute}} \sum_{\xi=1}^{N_{solvent}} \int_{R^3} \omega_{ss'}(r) * c_{s'\xi}(r) * \chi_{\xi\alpha}(r)$$

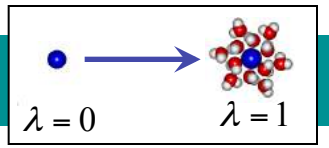
Closure relations:

$$h_{s\alpha}(r) = \exp\left[-\beta u_{s\alpha}(r) + h_{s\alpha}(r) - c_{s\alpha}(r) + B_{s\alpha}(r)\right] - 1$$

site-site bridge function

$$\chi_{\alpha\xi} = \omega_{\alpha\xi}(r) + \rho h_{\alpha\xi}(r)$$

end-point calculations



$$\Delta G_{hyd}^{HNC} = \frac{2\pi\rho}{\beta} \sum_{s=1}^{N_{solute}} \sum_{\alpha=1}^{N_{solvent}} \int_0^{\infty} \left[-2c_{s\alpha}(r) - c_{s\alpha}(r)h_{s\alpha}(r) + h_{s\alpha}^2(r)\right] r^2 dr.$$

(RISM) Chandler, Singh, J. Chem Phys. (1984), 81, 1975
(MD) Matubayasi, Nakahara, *J. Chem. Phys.* (2002), 117, 3605

Solvation Free Energy Functionals

- Each h(r) or c(r) function is defined between one solute atom (S) and one solvent atom (α)
- Solving the 1D RISM equations gives $N_s \times N_\alpha$ h(r) and c(r) functions.
- To compute molecular descriptors, we would like a single function that defines the solvation behavior of the molecule
- We note that the standard 1D RISM free energy formulae have the same general form:

Hydration Free Energy

Hyper-netted-chain:

$$\Delta G_{\mathrm{HNC}} = 2\pi\rho kT \sum_{s\alpha} \int_0^\infty [-2c_{s\alpha}(r) - h_{s\alpha}(r) \times (c_{s\alpha}(r) - h_{s\alpha}(r))] r^2 dr$$

Gaussian-Fluctuations:

$$\Delta G_{\mathrm{GF}} = 2\pi\rho kT \sum_{s\alpha} \int_0^\infty (-2c_{s\alpha}(r) - c_{s\alpha}(r) h_{s\alpha}(r)) r^2 dr$$

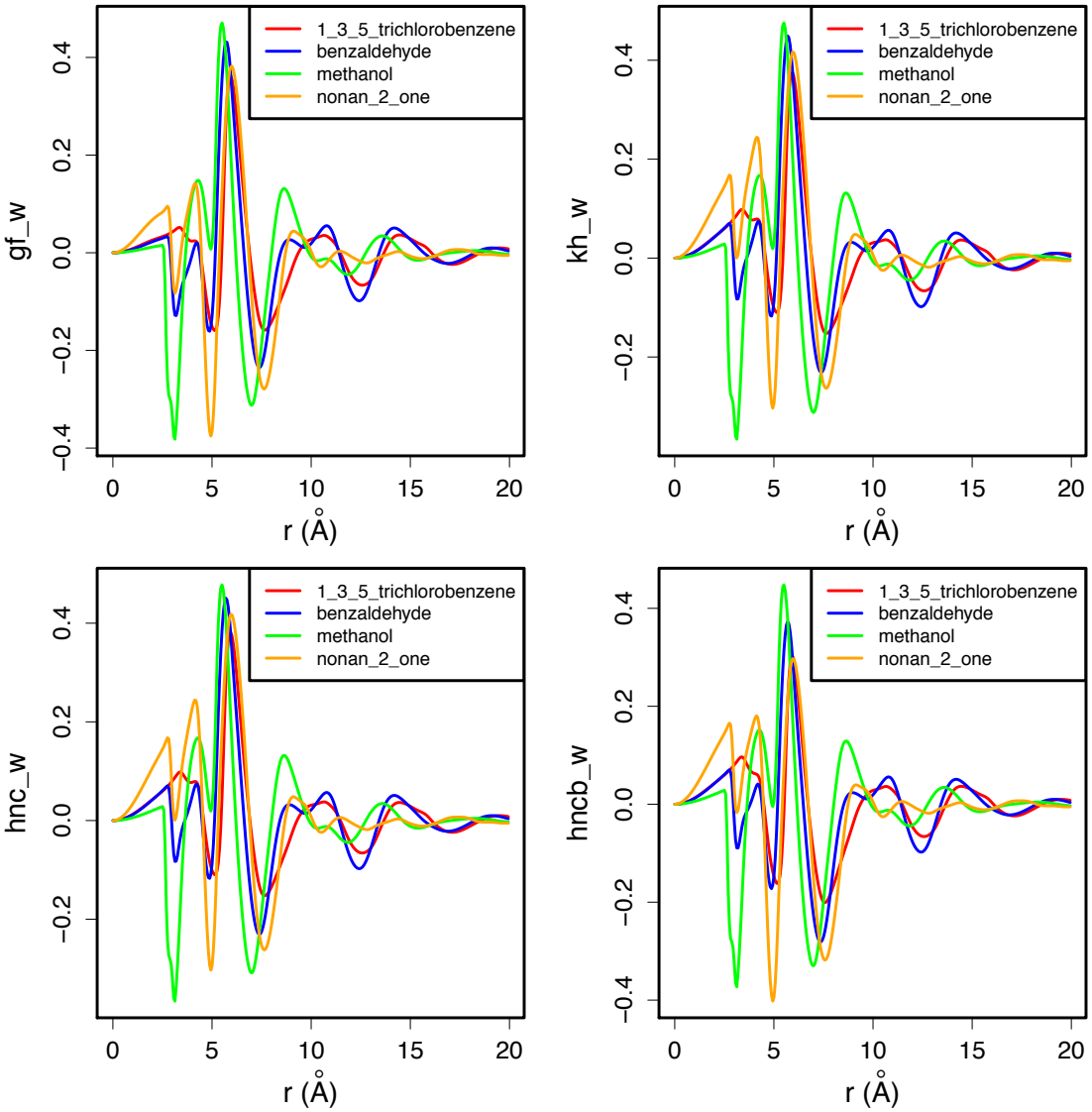
Kovalenko-Hirata:

$$\Delta G_{KH} = 2\pi\rho kT \sum_{s\alpha} \int_0^\infty [-2c_{s\alpha}(r) - h_{s\alpha}(r)(c_{s\alpha}(r) - \Theta(-h_{s\alpha}(r)))] r^2 dr$$

General Form:

$$\Delta G = c \sum_{S\alpha} \int_0^\infty f(r) r^2 \, dr$$

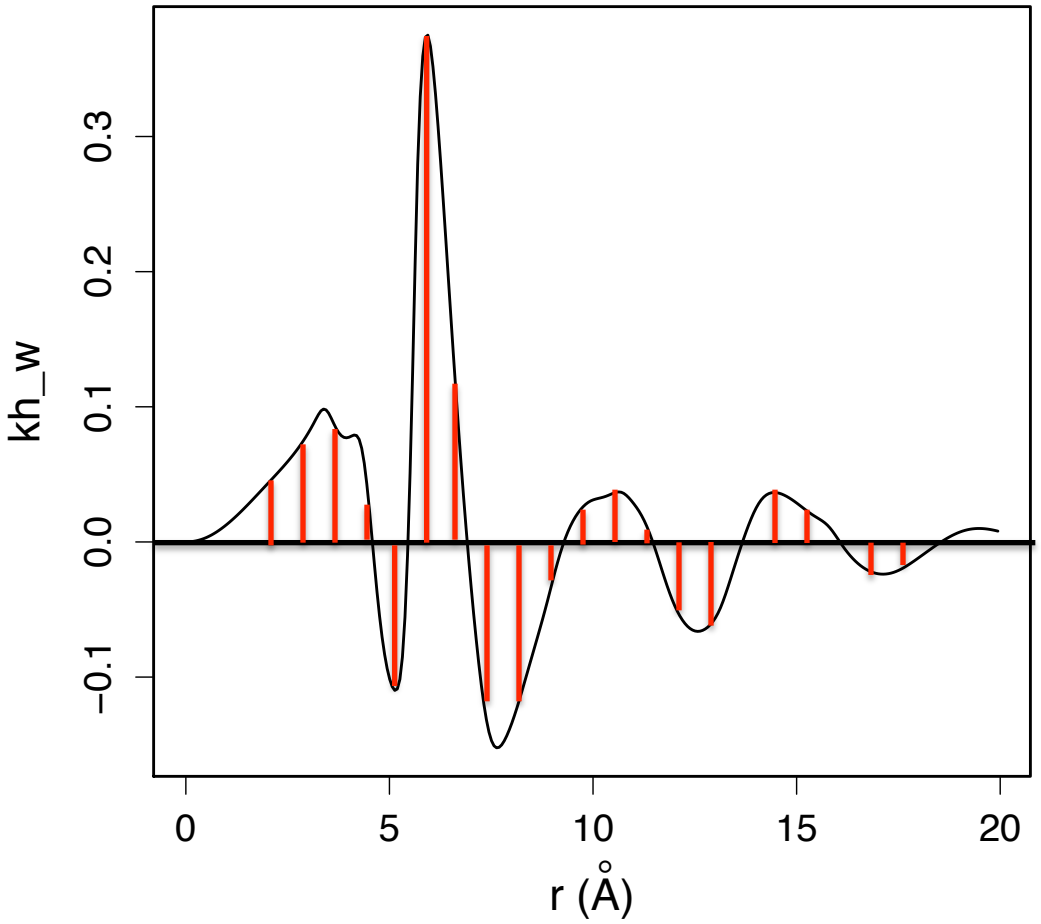
RISM-MOL-INF-CONF



- The graphs show the result of combining the $N_s \times N_\alpha$ h(r) and c(r) functions using the standard 1D RISM free energy functionals, but with the integration over r omitted.
- For example, based on the HNC functional, we define a functional hnc_w(r) (where the suffix, "_w" is used to indicate that the solvent is water_, e.g.

$$hnc_w(r) = 2\pi\rho kT \sum_{s=1}^{N_{\text{solute}}} \sum_{\alpha=1}^{N_{\text{solvent}}} \left[-2c_{s\alpha}(r) - c_{s\alpha}(r)h_{s\alpha}(r) + h_{s\alpha}^2(r) \right] r^2$$

RISM-MOL-INF descriptors

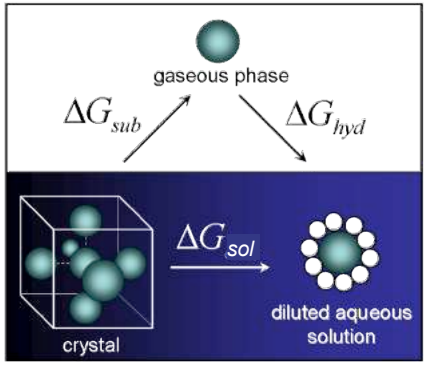


Datasets

Hydration Free Energy

(Ratkova, EL, J. Phys. Chem. B **2010,** *114,* 12068–12079)

Training/Testing: 65/120 small organic molecules

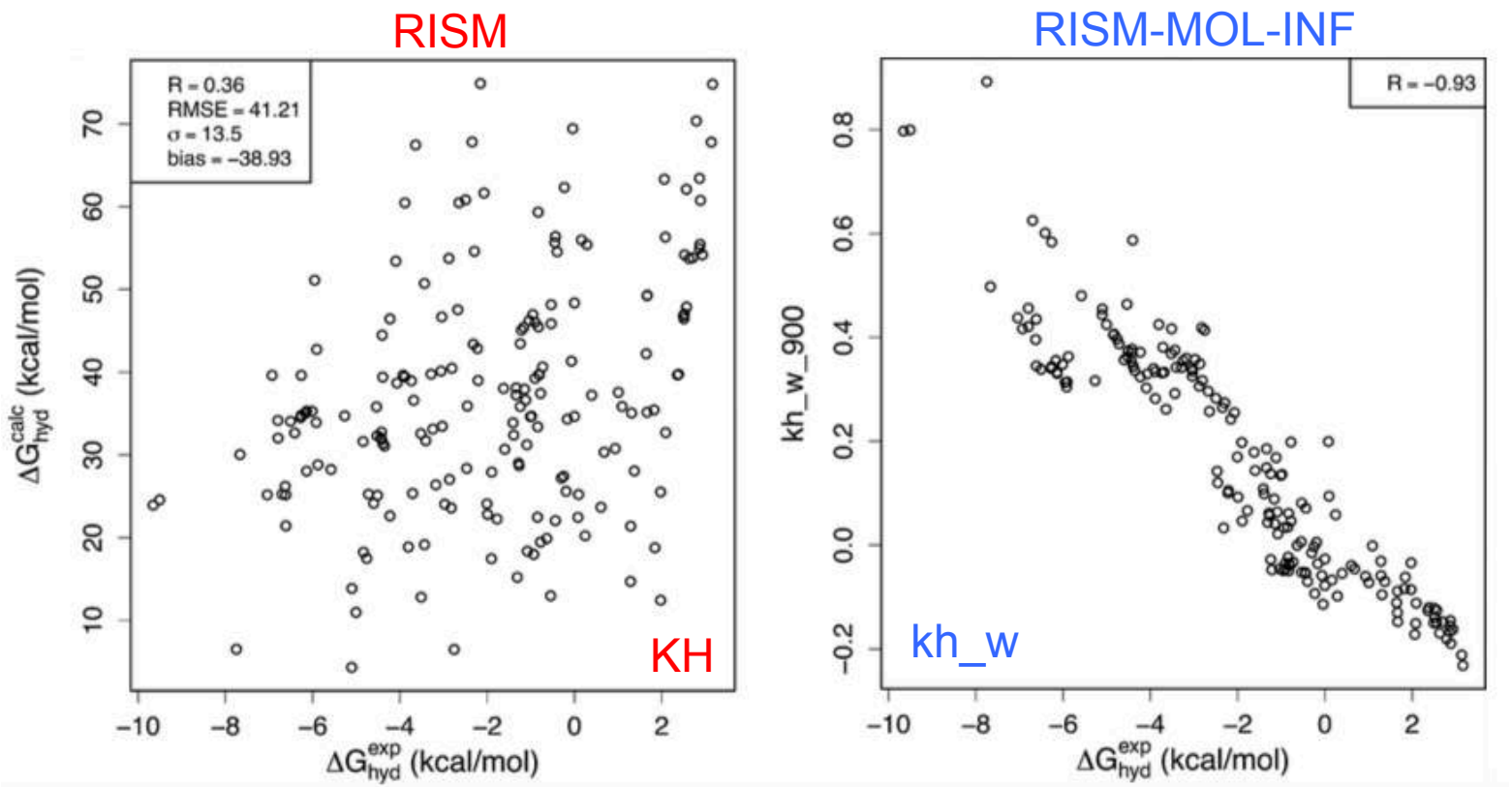


Caco-2 Permeability

*(*Hou, TJ, et al. *J. Chem. Inf. Comput. Sci.* **2004,** *44,* 1585-1600)

Training/Testing: 77/23 druglike molecules

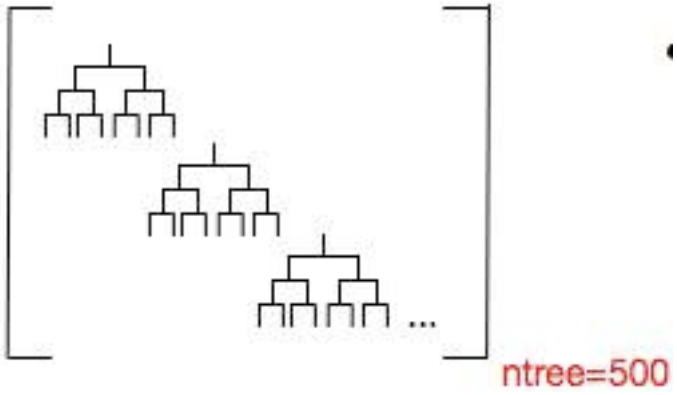
RISM-MOL-INF descriptors are correlated with experimental hydration free energy



Machine Learning

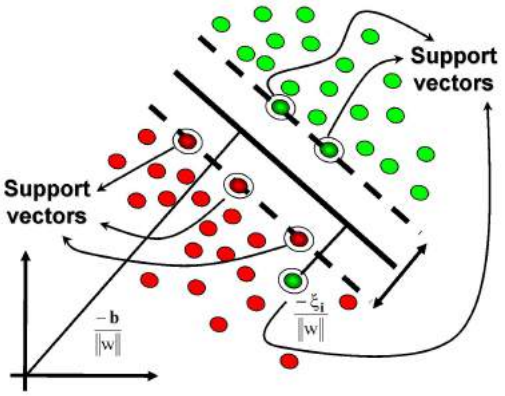
Random Forest

- *a coupled ensemble of regression trees*
- *each tree is trained:*
  - *from a bootstrap sample of the data*
  - *without pruning back (nodesize=5)*
  - *from subset of descriptors at each split (mtry=no.of descriptors/3)*
- *incorporates descriptor selection*
- *robust to over-fitting*

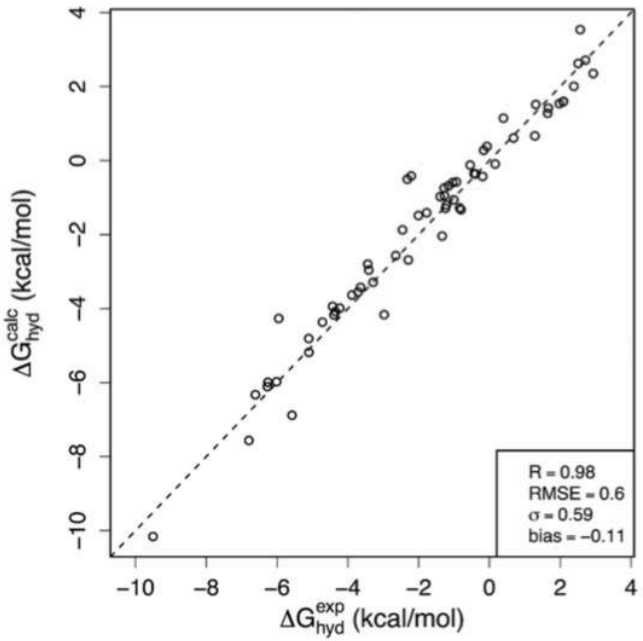


Support Vector Machine

- *kernel based learning algorithm*
- *radial basis function kernel used here*
- *requires careful training:*
  - *three training parameters (ε, γ, C)*
  - *some sensitivity to irrelevant descriptors*



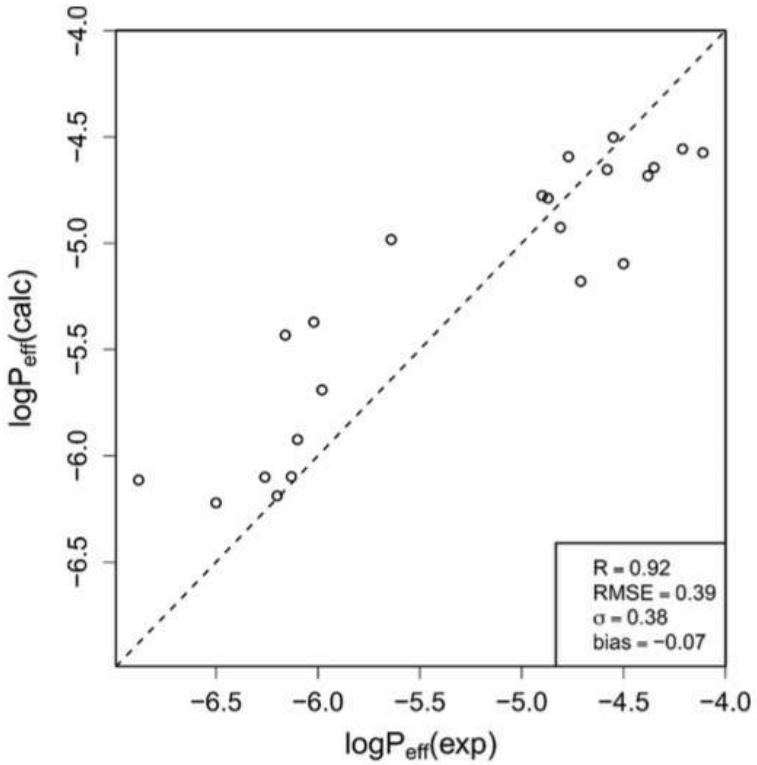
Hydration Free Energy



method	R(te)	RMSE(te)	bias(te)
gf_w	0.98	0.61	−0.12
hncb_w	0.98	0.63	−0.13
hnc_w	0.98	0.60	−0.11
kh_w	0.98	0.60	−0.11
GF	0.83	3.70	−2.49
HNCB	0.82	5.79	4.46
HNC	0.23	42.18	−40.17
KH	0.23	40.95	−38.90
QM	0.97	0.78	−0.31
3D RISM/UC	0.96	0.84	−0.16

Mol. Pharmaceutics **2015**, *12*, 3420−3432

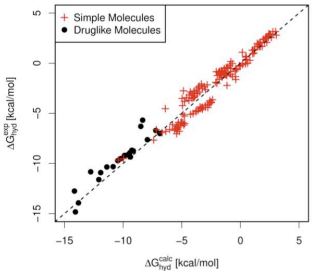
Caco-2 Permeability



descriptors	model	*R*(te)	RMSE(te)	bias(te)
gf_w	RF	0.91	0.39	−0.07
hncb_w	RF	0.91	0.41	−0.08
hnc_w	RF	0.91	0.41	−0.09
kh_w	RF	0.91	0.41	−0.09
Padel	RF	0.75	0.60	−0.13
ACD		0.79	0.56	0.17
Hou et al.[62]		0.77	0.58	0.04
Ponce et al.[80]		0.80	1.43	0.52

Mol. Pharmaceutics **2015**, *12*, 3420−3432

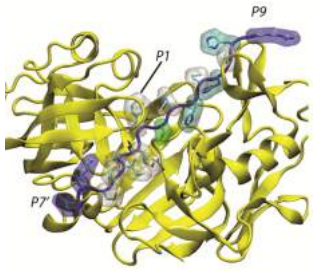
Conclusions



Fast and accurate predictions of hydration free energy are possible using 3D RISM.

- *The 3D RISM/UC and 3D RISM/PC+ functionals are easily implemented using existing software (e.g. AMBER, RISM-MOL3D)*

Solubility of crystalline drugs can be computed from a thermodynamic cycle via the gas-phase, but this method is not yet as accurate as informatics approaches (and currently requires that crystal structure is known)



RISM-MOL-INF is a promising fast and general method to predict physico-chemical and ADMET properties of druglike molecules. Further testing is ongoing ...

Review articles

- Solvation Thermodynamics of Organic Molecules by the Molecular Integral Equation Theory: Approaching Chemical Accuracy, *Chem. Rev.*, **2015**, *submitted.*

- Computing Intrinsic Aqueous Solubility of Crystalline Druglike Molecules from a Thermodynamic Cycle via the Gas Phase. In: Abramov. Y. A. ed. *Computational Approaches in Pharmaceutical Solid State Chemistry,* Wiley, **2015**, *in press.*

Thank you!

Maksim Mišin

***Collaborators*:**

Prof. Maxim Fedorov (Strathclyde)

Dr Antonio Llinas (AstraZeneca)

Dr. Karsten Bruun Qvist (Chr. Hansen)

Prof. Gennady Chuev (ITEB RAS,Russia)

Dr. Ekaterina Ratkova (ISC RAS, Russia)

Dr. Andrey Frolov (ISC RAS, Russia)

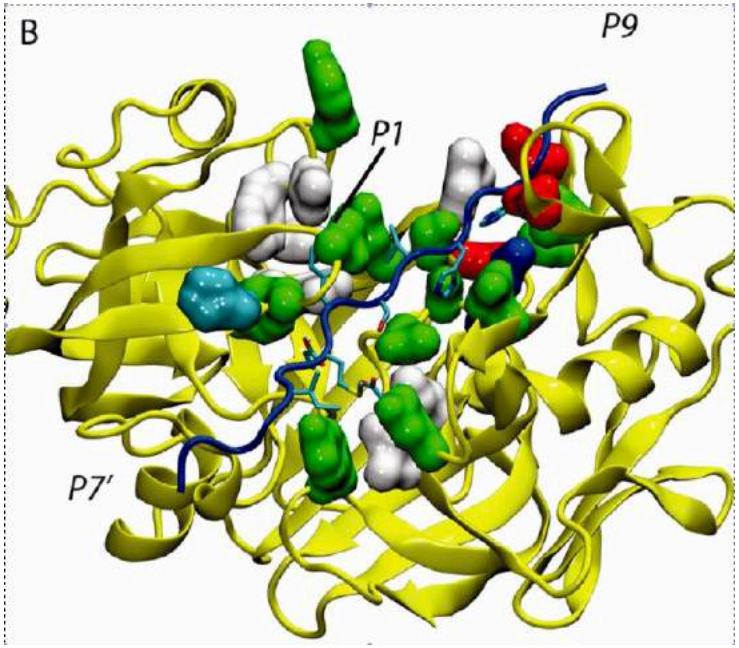
Dr. Volodya Sergiievskyi

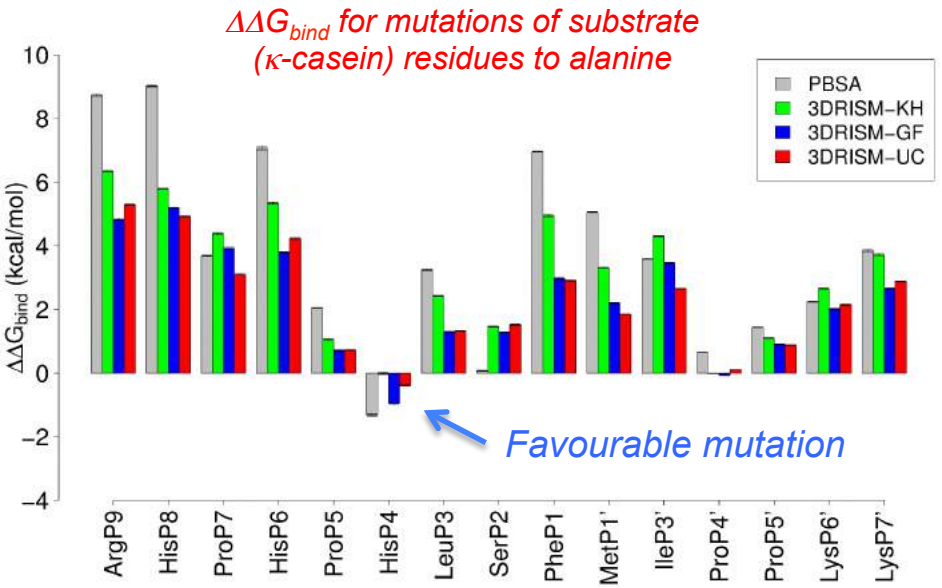
Dr John Mitchell (St Andrews)

Dr Tanja van Mourik (St Andrews)

Dr James McDonagh (Manchester)

Computational Alanine Scanning (MM-3DRISM) to probe the binding thermodynamics in Bovine and Camel Chymosin – κ-Casein complexes





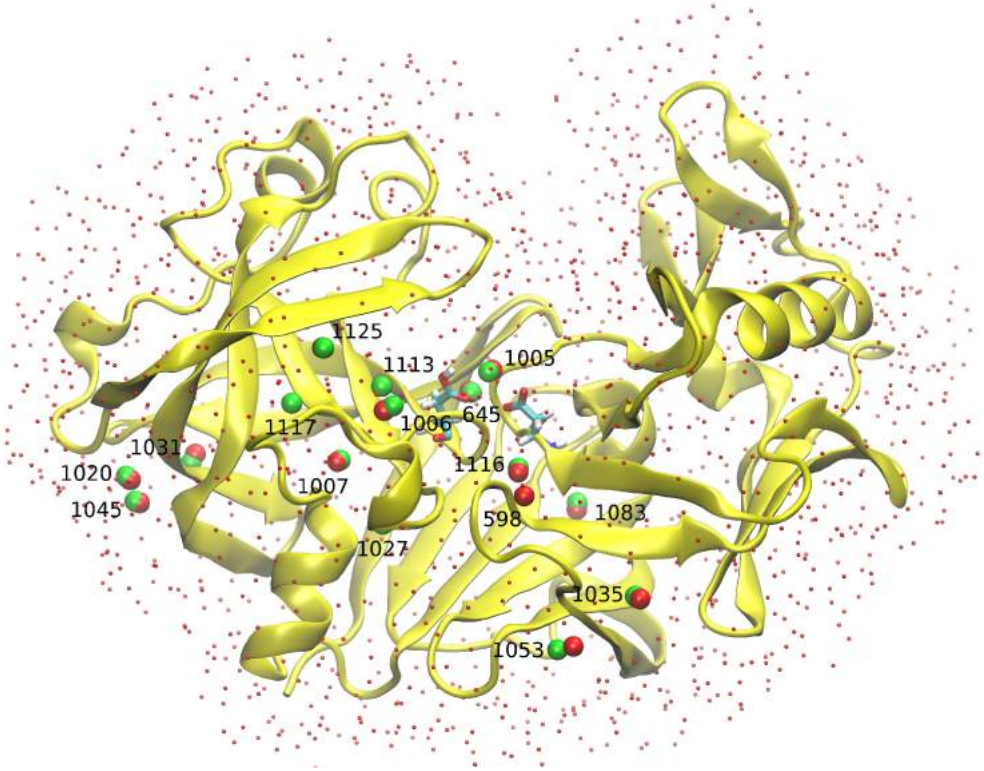
- Bovine and camel chymosin are aspartic proteases used to clot milk in cheese manufacturing
- High sequence identity (85%), but different catalytic properties.

- MM-3DRISM computational alanine scanning to systematically investigate the sequence-dependence of binding thermodynamics

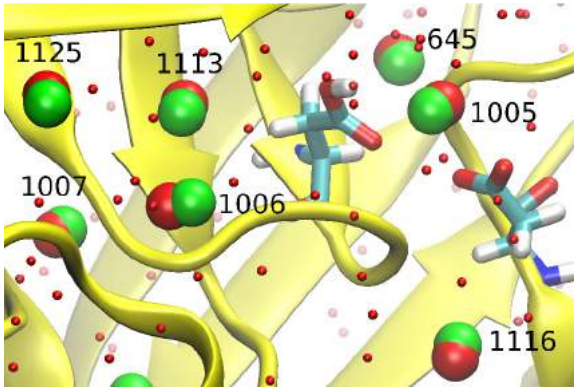
J. Chem. Theory Comput., **2013**, 9, 5706−5717 .

Peaks in the 3D RISM Solvent Distribution Functions reveal the Location of Bound Solvent Molecules

Bovine Chymosin (3CMS)



Binding Site



exptl. water (ID)	nearest calcd. water (ID)	R (separation, Å)	$\sum_2 R <$ Å	$g(r)_0$
1113	1	0.62	1	17.27
598	2	0.31	1	17.07
645	10	0.22	1	14.96
1007	13	0.28	1	14.50
1116	20	0.41	1	13.71
1125	25	0.7	1	13.05
1006	47	0.77	1	11.77
1005	126	0.45	1	8.78

● Experiment[1] ● 3D RISM[2]

1. Prasad BVLS, *et al.*, Acta Cryst., 2002, D58, 250-259
2. Palmer DS, *et al.* J. Chem. Theory Comput., **2013**, *ASAP.*