

The SAMPL2 blind prediction challenge: introduction and overview

Matthew T. Geballe · A. Geoffrey Skillman ·
Anthony Nicholls · J. Peter Guthrie ·
Peter J. Taylor

Received: 7 January 2010 / Accepted: 22 March 2010 / Published online: 9 May 2010
© Springer Science+Business Media B.V. 2010

Abstract The interactions between a molecule and the aqueous environment underpin any process that occurs in solution, from simple chemical reactions to protein–ligand binding to protein aggregation. Fundamental measures of the interaction between molecule and aqueous phase, such as the transfer energy between gas phase and water or the energetic difference between two tautomers of a molecule in solution, remain nontrivial to predict accurately using current computational methods. SAMPL2 represents the third annual blind prediction of transfer energies, and the first time tautomer ratios were included in the challenge. Over 60 sets of predictions were submitted, and each participant also attempted to estimate the error in their predictions, a task that proved difficult for most. The results of this blind assessment of the state of the field for transfer energy and tautomer ratio prediction both indicate where the field is performing well and point out flaws in current methods.

Keywords Solvation · Transfer energy · Tautomers · Blind prediction · Validation

Introduction

The way a molecule interacts with its aqueous environment underpins any process that occurs in solution, from simple chemical reactions to protein–ligand binding to protein aggregation. Fundamental measures of the interaction between a molecule and water, such as the transfer energy between gas phase and water or the energetic difference between two tautomers of a molecule in solution, remain nontrivial to predict accurately using current computational methods. Accurate calculations of more complex properties or events will remain over the horizon until these more basic values can be predicted with greater accuracy.

Importance of transfer energies and tautomer ratios

The free energy of transferring a molecule from a vacuum environment to an aqueous environment, i.e. the transfer energy or energy of solvation, is a fundamental value for any system involving molecules in solution (Fig. 1). Transfer energies represent an important physical quantity that plays a role in predictions of solubility, propensity to aggregation, binding affinity to a protein, and other qualities important in pharmaceutical, environmental, and materials science. Likewise, the existence of and energetic differences between different tautomeric forms of a molecule in solution presents a complicating factor that, although often ignored, can have a large effect on understanding protein–ligand interactions, solution-phase properties, and electrostatics. Computational predictions involving any of these scenarios require careful analysis of tautomers or risk developing large errors.

Transfer energies and tautomer ratios are properties of a small molecule in solution, which are less complex systems than many of the oft-tackled problems involving proteins

M. T. Geballe (✉) · A. G. Skillman · A. Nicholls
OpenEye Scientific Software, Inc., Santa Fe, NM 87508, USA
e-mail: mattg@eyesopen.com

J. P. Guthrie
Department of Chemistry, University of Western Ontario,
London, ON, Canada

P. J. Taylor
AstraZenica Pharmaceuticals, London, UK

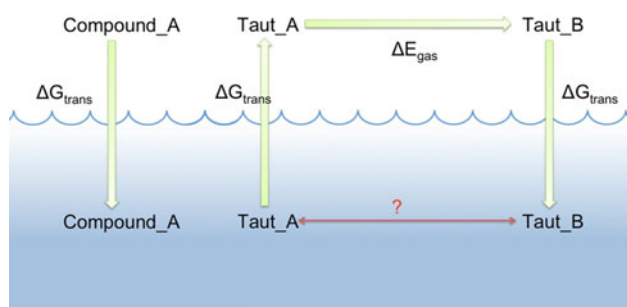


Fig. 1 Schematic of Transfer Energy and Tautomer Ratio, including the thermodynamic cycle often used for calculating tautomer ratios

or lipid bilayers. Performing the necessary sampling for calculation of transfer energies or tautomer ratios is a tractable problem, and thus allows for a wide variety of methods to be applied to these problems. If new, computationally intensive, or radically different methods or force fields can prove their worth in these calculations, then the time spent on performing such calculations on larger systems could be justified. In this way, addressing these questions can be a testing or proving ground for new methods or force fields.

Why is blind prediction important?

SAMPL is designed as a blind challenge, where the actual experimental values of the compounds (i.e. the answers) are withheld from participants until after predictions have been made. Blind prediction is the gold standard by which to assess the true predictive ability of a method, removing opportunities for the fitting of free parameters, observer bias, or other influences that may skew results. Many previous evaluations have not been blinded, and some have involved fitting parameters of some form to the experimental data. Fitting parameters to data leads to a reduction of error in prediction for that data, however it often results in over-fitting, which reduces the prediction power of the method when presented with new data [1]. As such, these evaluations often result in average errors that are artificially low when compared to prospective predictions [2].

Any and all participants are welcome to participate in SAMPL, and participants are able to submit multiple prediction sets if desired. Predominantly the participants are either the developers of, or expert users of, the method utilized. Even if non-expert predictions are submitted, there is no barrier to other submissions using the same method. Previous evaluations may have compared expert use of some methods to non-expert use of others, resulting in inaccurate estimates of error and comparisons. A recent example of this involved several of the methods used in SAMPL2 [3–5].

Existing experimental data on transfer energies is limited to a relatively small number of compounds and has already been well studied and parameterized by current methods. With only a small and largely static pool of known data, method development can be prone to over-fitting, which can increase the performance on the known data while reducing performance on future data. Additionally, most of the known transfer energies are for compounds that don't reflect the complexity common in pharmaceutical chemistry. Training methods on smaller, mono-functional molecules facilitates simple additive parameters that break down when applied to polyfunctional molecules. These factors can lead to large errors when current methods trained on existing data are applied to new data on drug-like compounds.

Comparing method performance between different published evaluations is fraught with uncertainty for several reasons: differences in dataset composition, method changes or improvements, differing skill level in method application, and public knowledge of experimental data. The design of SAMPL attempts to avoid some issues that have plagued previous evaluations. While all attempts are made to remove bias from the challenge, the lack of availability of blinded data constrains our ability to select an ideal problem set with which to test methods. Nevertheless, the SAMPL challenge is designed to provide a rigorous test of methods on complex, relevant molecules, and results may not be directly comparable to retrospective published studies with aforementioned systemic flaws.

Brief history of SAMPL

SAMPL2 represents the third annual blind prediction of transfer energies, and the first time tautomer ratios were included in the challenge. Previous SAMPL challenges presented difficult molecules for prediction and exposed problematic areas of chemistry for various techniques (See Fig. 2). The first SAMPL challenge, SAMPL0, involved a comparison between PB and explicit MD simulation on a small set of 17 compounds. Both methods achieved RMS errors of less than 2 kcal/mol for the best performing submission, and the data set, although small and somewhat

Challenge	Num. Compounds	Range of Transfer Energies	RMSE of Best Method
SAMPL0	17	-11.0 to 1.1	1.33
SAMPL1	56	-15.5 to -1.5	2.44
SAMPL2	40 (23 blinded)	-25.5 to -7.0	1.56

Fig. 2 Summary of challenge composition and subsequent best performance for the transfer energy portion of the past and present SAMPL challenges

overpopulated with esters, represented larger or more complex molecules than commonly examined in prior studies. SAMPL1 exposed participants to an increased level of difficulty, with 56 molecules covering significantly more and diverse functionality. Given the problematic dataset, RMS errors were above 2 kcal/mol for every method, although not as high as some had feared. SAMPL1 was instrumental in revealing shortcomings for several of the methods that were applied, and cemented the value of a regular comparison between prediction and experimental results [6, 7].

SAMPL2 challenge design

This year SAMPL2 challenged participants with datasets for both transfer energy and tautomer ratio prediction. The molecules included in each dataset are divided into three sections: obscure, explanatory, and investigatory. The obscure section is analogous to a traditional blind challenge; experimental values are known but withheld from the participants. The explanatory section contains compounds with unusual experimental values. The participants were provided with the experimental data for these molecules and asked to provide an explanation for their unexpected values. Lastly, the investigatory section contains structures for which no experimental data is available, yet are representative of common chemistry that is not well explored by experimental data. It is our desire that these compounds provide an opportunity to test consensus within the computational field as well as providing an impetus to experimentalists to provide experimental data for these compounds.

Methods

Selection of participants

The SAMPL2 challenge was open to all participants through the SAMPL website (<http://sampl.eyesopen.com>), although registration was necessary to download the data. Participants were able to indicate a desire to remain anonymous in subsequent published analysis during the data submission process. Additionally, invitations to participate were sent to many individuals and organizations from the academic, pharmaceutical, and commercial computational chemistry software fields. An invitation was also posted on the computational chemistry list (CCL).

Challenge and submission timeline and format

The data for each challenge (including the compounds, both as isomeric smiles and SDF, and an explanatory

document by the organizer of each section) was made available for download on February 20th from the SAMPL website. The deadline for submission of predictions was May 18th, and only the six submissions from Ribeiro et al. [8] were received after the deadline. Submitted data was uploaded as a text file to the SAMPL website, where it was parsed and participants were given the opportunity to double-check their predictions before submission was made final.

Results returned to participants

The experimental results were not revealed until after the submission deadline. Upon receipt of submitted predictions and the experimental data, an initial statistical analysis was performed for each submission to estimate the effects of experimental error (see “Choosing Appropriate Methods” below). After this was completed, each participant was sent their results, including the experimental data and both text files and plots describing the results of the analysis. Also included were the results of the application of the statistical analysis protocol applied to the experimental data to provide an example of the effects of bootstrapping.

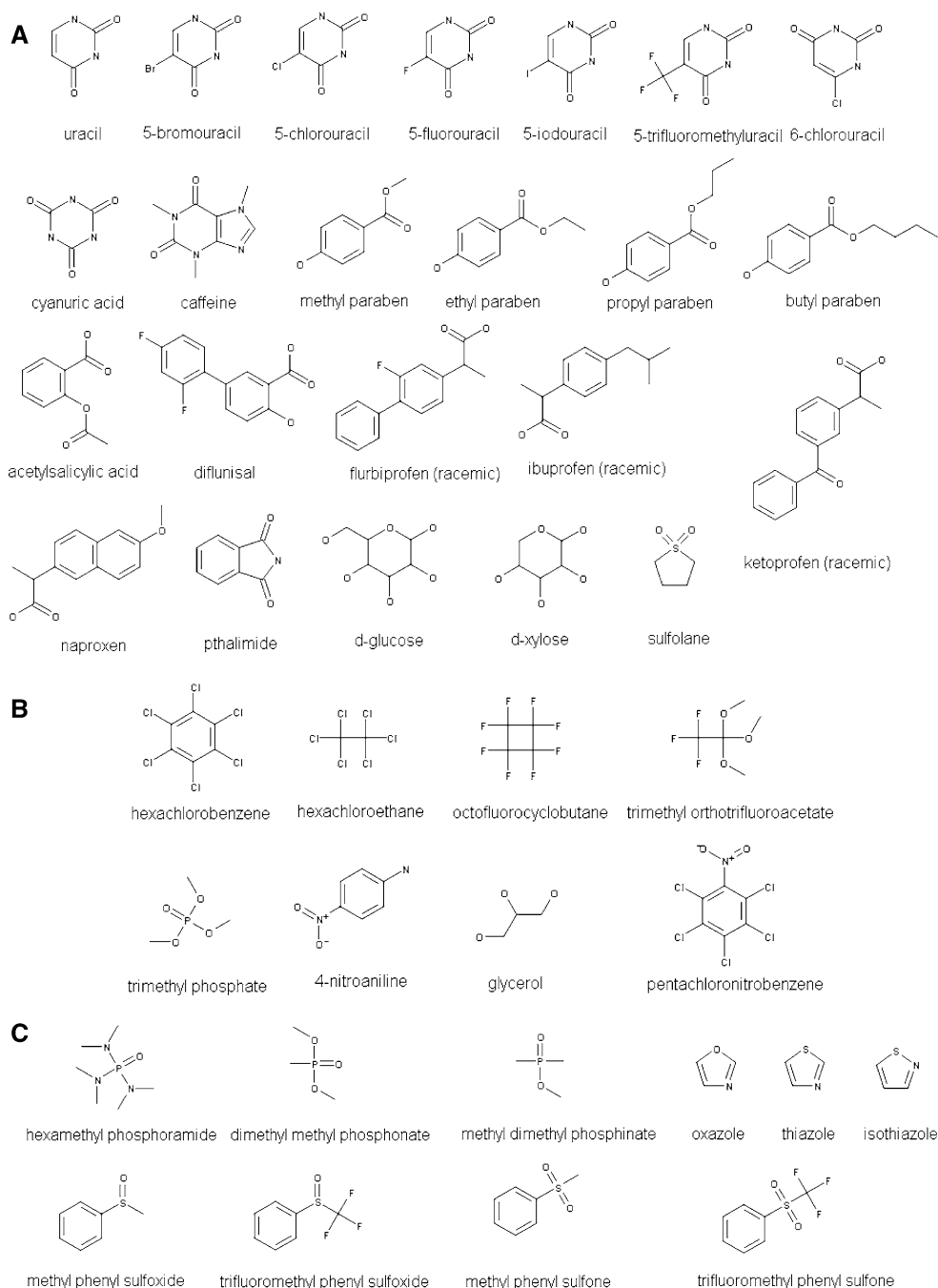
Transfer energy dataset (J. P. Guthrie)

Solvation energies (free energies of solvation) refer to the process gas (1 M) to aqueous solution (1 M). These values are based on all the data for the compounds involved (Fig. 3) which are available at this time from the ongoing construction of a critical data base of solvation energies, including a new batch of references added during this past summer.

In general available experimental data were averaged, using the weighted average or grand mean, to obtain solvation energies. The data used, depending on availability, were aqueous solubility, limiting aqueous activity coefficient, vapor pressure and gas–water partition coefficient or Henry’s law constant. In the case of some compounds capable of ionizing in near neutral solution (e.g. ibuprofen) the solubilities reported by Avdeef [9], after careful attempts to correct for ionization and aggregation, were used to the exclusion of others that would be subject to systematic errors. In the case of caffeine, the solubility reported by Cesaro [10] with careful correction for aggregation was used.

In many cases data at multiple temperatures have been reported; in such cases all data were fitted to a van’t Hoff equation, modified so that the parameters were the vapor pressure at 298 K and the heat of vaporization. This means that the uncertainty in the vapor pressure at 298 K is directly obtained from the least squares fit with no need for additional calculations. The normal van’t Hoff equation is shown in Eq. 1. This was transformed into Eq. 2, where

Fig. 3 **a** The Obscure transfer energy set. **b** and **c** The Explanatory and Investigatory categories on the transfer energy challenge



$\theta = 298.15$ K and p_θ is the vapor pressure at 298 K. Analogous equations were used for temperature dependent solubility data. Where free energies of solvation at different temperatures were available they were fitted to Eq. 3.

$$\ln p = A + B/T \quad \text{or} \quad p = e^A e^{B/T} \quad (1)$$

$$p = p_\theta e^{B(\frac{1}{T} - \frac{1}{\theta})} \quad (2)$$

$$\Delta G_s = \Delta H_s - T\Delta S_s \quad (3)$$

For diflunisal a recent report [11] lists four crystal forms with different solubilities, with a range of two orders of

magnitude in solubility. The melting points of these crystalline forms were not given so it is not possible to correlate this report with other studies of these compounds. The lowest solubility was used, presumably corresponding to the most stable crystalline form.

For uracil the vapor pressure data reported [12] show serious inconsistencies, defining several distinct, though close, lines. All data were used to extrapolate to 298 K.

For cyanuric acid there were two sets of vapor pressure data which differed by a factor of 1.5–1.7 at various temperatures. Kozyro [13] gave the experimental vapor

pressure values but his paper is in Russian which I have difficulty reading to check the experimental. de Wit et al. [14] gave no experimental vapor pressures, but only smoothed values based on their data treatment. In this case I did two separate van't Hoff plots to extrapolate to 298 K, and then took the grand mean of the solvation energies.

For glycerol and sulfolane, data for vapor pressures over various mixtures of organic liquid and water at 298 K have been reported. On the assumption that the contribution to the total vapor pressure from the organic liquid is negligible, the activity coefficient for water can be calculated as $\gamma_w = p/x_w p^\circ$, where x_w is the mole fraction of water in the mixture, p is the vapor pressure above the mixture, and p° is the vapor pressure of water at 298 K. These data were fitted to $\ln \gamma_w = \alpha_2 x_s^2 + \alpha_3 x_s^3 + \alpha_4 x_s^4$ where x_s is the mole fraction of the organic solvent in the mixture, and α_i are three coefficients to be determined by fitting to the data and then γ_s was calculated using Eq. 4. This equation is equivalent to using the four suffix Margules method. [15]

$$\begin{aligned} \ln \gamma_s &= \left(\alpha_2 + \frac{3}{2} \alpha_3 + 2\alpha_4 \right) x_w^2 - \left(\alpha_3 + \frac{8}{3} \alpha_4 \right) x_w^3 + \alpha_4 x_w^4 \\ &= \alpha_2 + \frac{1}{2} \alpha_3 + \frac{1}{3} \alpha_4 \quad \text{at } x = 1 \end{aligned} \quad (4)$$

Non linear least squares fitting produced both the α_i parameters and the covariance matrix, which was used to calculate the statistical uncertainty in γ_s . My standard program for solvation database calculations converts γ_s into a hypothetical ideal solubility in moles per liter, $c = 55.5/\gamma_s$, which can then be used along with the vapor pressure of the pure liquid to calculate the solvation energy. Vapor pressure data for the pure liquids were fitted to the van't Hoff equation in the usual way to determine the vapor pressures at 298 K, and then the solvation energy was calculated from ideal solubility and vapor pressure.

Caffeine has a solid–solid phase transition at 414 K [16]; only vapor pressures below this temperature were used.

There has been only one report of a solvation energy for trimethyl orthotrifluoroacetate [17]. This value depended on a vapor pressure estimated from the observed boiling point; the vapor pressure at 298 K has been recalculated using a more recent estimation procedure [18]. The new value is very similar to the literature value.

Table 1 gives references to all of the data used, and weighted average values for each quantity for which there were data. Uncertainties in data were taken from the sources where available or assigned where they were not. In general errors were assumed to be 10% in each experimental value, and 14% when two experimental results had to be combined, as in Henry's law constant determinations. The not uncommon disagreement between different sets of data shows that this is often an optimistic assumption of uncertainty, mainly because of the

difficulty in assessing systematic error; scatter within a set of data is generally less than 10%. Errors in the final grand mean of ΔG_s were never assigned a value less than 0.1 kcal/mol, even though the standard calculation led to a smaller value of uncertainty; such situations only arose when there was only a single source for at least one input value. In view of the disagreements often seen when there are multiple data sets for a quantity, such pessimism seems justified. In a few cases, errors in the calculations were caught, leading to discrepancies between the values used as the experimental value in the statistical analysis and the current best values.

Cyanuric acid has been a problem, exacerbated by an error in my initial calculation of the solvation energy. An independent route to a solvation energy is available by a computation of the free energy of formation of gaseous cyanuric acid (-100.20 using G3MP2B3 in Gaussian03 [19]). Kozyro [13] provided values for $\Delta H_f(\text{solid}) = -168.15 \pm 0.35$ kcal/mol and $S^\circ(\text{solid}) = 33.99$ cal/K/mol, leading to $\Delta G_f(\text{solid}) = -120.73$ kcal/mol. Combined with the solubility value cited in Table 1 this leads to $\Delta G_f(\text{aq}) = -118.41$ and $\Delta G_s = -16.31$ kcal/mol, with an uncertainty of about 2 kcal/mol. This is in good agreement with the value in the Table. The value for cyanuric acid was challenged because sublimation might be accompanied by decomposition. There is no such concern for the solid state measurements at room temperature.

Tautomer ratios (P. J. Taylor)

The twelve sets of problems submitted for SAMPL2, encompassing 68 tautomer pairs, were intended to provide an overview of at least some types of tautomer (see Fig. 4; Table 2). They are by no means exhaustive and many others, related and unrelated, may prove of interest in the future.

One cluster of problems attaches to amide-iminol tautomerism. For the 6-membered ring oxoheterocycles, tautomer ratio in 2-pyridone (**1**) is supplemented by the effects of benzofusion to give the three possible products (**2**), (**3**) and (**4**). Here a major focus of interest consists in the experimental evidence that, while very different results are found for these three compounds, the *incremental change* $\Delta \log K_T$ at each position is nearly a constant when different monocycles are used as the starting point. Explicitly, when NH (for the oxo-tautomer) is placed α to the point of benzofusion, the known increments are 1.0, 0.8, 1.0 and 1.0; when placed β , 1.7 and 1.8; and when the oxo-form is forced into being quinonoid, -1.8 and -2.0 results; while the last is supplemented by -1.7, -1.8 and -2.1 for three cases in which the starting point for benzofusion is not an oxoheterocycle. It is difficult to see these

Table 1 Solvation energies for the obscure and explanatory datasets (All for aqueous solution at 298 K. All ΔG_s values (1 M/1 M) are given to two decimal places for uniformity although for many the uncertainties mean that the second decimal place is not significant)

Compound	Aqueous solubility (M)	Refs	Vapor pressure (atm)	Refs	ΔG_s (kcal/mol)	ΔG_s direct	Refs	ΔG_s grand mean (kcal/mol)	ΔG_s answer
Acetylsalicylic acid	1.78(0.13)E-02	[20]	4.11(1.27)E-10	[21]	-12.33(0.19)			-12.33(0.19)	-9.94(0.18)
5-bromouracil	1.91(0.06)E-02	[22]	1.61(0.21)E-14	[23]	-18.39(0.10)			-18.39(0.10)	-18.17(0.55)
Butyl paraben	1.32(0.06)E-03	[24]	1.29(0.04)E-08	[25]	-8.74(0.03)			-8.74(0.10)	-8.72(0.27)
Caffeine	3.90(1.00)E-02	[10]	3.94(0.17)E-10	[26, 27, 16]	-12.82(0.15)			-12.82(0.15)	-12.64(0.74)
5-chlorouracil	1.71(0.06)E-02	[22]	3.77(1.15)E-14	[23]	-17.82(0.24)			-17.82(0.24)	-17.74(0.78)
6-chlorouracil	3.35(0.02)E-02	[22]	1.59(0.79)E-12	[23]	-16.00(0.30)			-16.00(0.30)	-15.83(1.22)
Cyanuric acid	2.01(0.20)E-02	[28]	3.71(0.01)E-14	[14, 13]	-17.92(0.10)			-18.39(0.42)	-18.06(0.27)*
			8.96(1.02)E-15		-18.77(0.09)				
D-glucose	17.63(5.22)	[29]	1.01(0.21)E-16	[30]	-25.44(0.21)			-25.44(0.21)	-25.47(0.22)
D-xylose	10.78(4.65)	[29]	2.59(0.30)E-13	[30]	-20.50(0.26)			-20.50(0.26)	-20.52(0.27)
Diflunisal	1.26(0.60)E-06	[11, 31]	8.01(0.44)E-11	[32]	-7.63(0.28)			-7.63(0.28)	-9.4(0.2)
Ethylparaben	5.43(0.13)E-03	[24]	2.45(0.04)E-08	[25]	-9.20(0.10)			-9.20(0.10)	-9.2(0.3)
5-fluorouracil	1.38(0.003)E-01	[22]	1.19(0.29)E-12	[23]	-17.01(0.15)			-17.01(0.15)	-16.92(0.88)
Flurbiprofen	4.35(0.43)E-05	[33]	4.67(0.15)E-10	[32]	-8.68(0.06)			-8.68(0.10)	-8.42(0.16)
Glycerol	134(48.)	[34]	2.10(0.07)E-07	[35–38]	-13.92(0.22)			-13.92(0.22)	-13.43(1)*
Hexachlorobenzene	2.22(0.75)E-08	[39–42]	2.46(0.45)E-09	[43–46]	-3.212(0.23)	-2.58 (0.34)	[47–51]	-3.01(0.29)	-2.33(1.16)
Hexachloroethane	3.24(2.76)E-05	[40]	5.02(0.16)E-04	[52]	-0.28(0.51)	-0.64 (0.01)	[53, 54]	-0.64(0.10)	-1.41(0.1)
(±)ibuprofen	9.70(1.94)E-05	[9]	1.76(0.04)E-08	[55]	-7.01(0.12)			-7.01(0.12)	-7(0.64)
5-iodouracil	2.07(0.05)E-02	[22]	1.28(0.23)E-14	[23]	-18.57(0.11)			-18.57(0.11)	-18.72(0.64)
Ketoprofen	3.62(0.59)E-04	[9]	1.05(0.03)E-10	[56]	-10.83(0.10)			-10.83(0.10)	-10.78(0.18)
Methyl paraben	1.48(0.05)E-02	[24]	3.84(0.08)E-08	[25]	-9.52(0.02)			-9.52(0.10)	-9.51(0.26)
Naproxen	9.12(1.30)E-05	[9]	5.90(0.25)E-11	[57]	-10.35(0.09)			-10.35(0.10)	-10.21(0.18)
4-nitroaniline	3.80(0.04)E-03	[58, 59]	6.00(0.31)E-09	[60–64]	-9.82(0.03)	-12.90 (0.23)	[47]	-9.88(0.42)	-9.45(0.11)
Octafluorocyclo- butane	1.36(0.03)E-04	[65, 66]	1.00 atm	-	3.38(0.01)			3.38(0.10)	3.43(0.03)
Pentachloronitro- benzene						-5.22 (0.06)	[67]	-5.22(0.10)	-5.22(0.25)
Phthalimide	2.45(0.25)E-03	[68]	4.53(0.19)E-09	[69]	-9.73(0.06)			-9.73(0.10)	-9.61(0.5)
Propyl paraben	1.99(0.08)E-03	[24]	6.91(0.16)E-09	[25]	-9.35(0.03)			-9.35(0.10)	-9.37(0.22)
Sulfolane	4.67(0.25)	[70, 71]	2.23(0.05)E-05	[70, 72]	-9.17(0.03)			-9.17(0.10)	-8.61(0.31)
5-trifluoromethyl- uracil	1.49(0.01)E-01	[22]	1.76(0.46)E-11	[73]	-15.46(0.15)			-15.46(0.15)	-15.46(0.16)
Trimethyl orthotrifluoroacetate	8.9(1.5)E-03	[17]	6.9(1.3)E-02	16, 18]	-0.69(0.15)			-0.69(0.15)	-0.8(0.2)
Trimethyl phosphate						-8.76 (0.08)	[74]	-8.76(0.10)	-8.7(0.1)
Uracil	3.783(0.70)E-02	[22, 75]	1.62(0.45)E-12	[23, 12]	-16.06(0.20)			-16.06(0.20)	-16.59(0.28)

regularities as accidental and I was (am) hopeful that high level computation may throw some light on their origin.

Similarly, for the 5-membered ring oxoheterocycles the exceptionally low values of $\log K_T$ shown by compounds (**10**–**16**) require explanation. An empirical approach suggests two main factors to be present. One is aromatisation: if my estimate of $\log K_T$ 7.6 for 2-piperidinone (**22**) is correct, aromatisation in 2-pyridone (**1**) if responsible for the whole effect results in $\Delta \log K_T$ –4.1. However, effects are much greater here: relative to an (estimated) value of $\log K_T$ 7.0 for 2-pyrrolidinone (**23**), $\Delta \log K_T$ –6.1 results for (**10D**) → (**10C**) for example, and –7.2 for (**14D**) → (**14C**). Given the lower aromaticity of 5-membered vs. 6-membered ring heterocycles in general, some extra factor must be present. I believe this to be a strong repulsion between contiguous hetero-atoms of the same hybridisation type, present in the oxo-tautomers, which for want of a more precise description I describe as ‘dipolar repulsion.’ I was (am) looking for comments on this point, including, if possible, a more accurate description of the phenomenon concerned.

A better characterised intramolecular repulsive force is lone pair repulsion, here exemplified by the iminol (**6A**) and responsible, in my view, for the exceptionally high value of $\log K_T$ 6.8. It may also be responsible for the unusual importance of the zwitterion (**6Z**). Here the azanitrogen atoms of (**6A**) are of the same hybridisation type, and as e.g. for (**10D**) and (**14D**), the more stable grouping = NZ – (Z = NR or O) comes to the rescue in (**6B**) and (**6Z**). There appears to be an important distinction between these two forms of intramolecular repulsion. While lone pair repulsion is absent in non-cyclic compounds through bond rotation, and its magnitude varies with ring size—it is much less in 5-membered rings—I have accumulated scattered but convincing evidence that ‘dipolar repulsion’ is not so readily avoided this way; it is still present in open-chain compounds, and even perhaps to a comparable extent. This too is a topic on which I was (am) hoping for elucidation.

The final cluster of amide-like compounds consists in (**22**–**26**). All are ‘unknowns,’ but certain trends can be anticipated. Chief of these is that an imide differs from an amide in carrying an extra, electronegative substituent, so that on all precedent, K_T should be lower for (**24**) than for (**23**). However (**26**), whose iminol is anti-aromatic, should show a sharp rise in K_T relative to the latter. Less certainly, I was expecting to see a (slightly) higher K_T value for (**22**) than for (**23**). I was also expecting to see a higher K_T for the 2-C=O than for the 4-C=O group of (**25**), the first being ‘urea-type’ and the second ‘imide-type,’ but messed this up through forgetting to specify that both were required.

Enolisation receives less emphasis, since the recent methodology of Kresge and his collaborators has been so successful, and its results so comprehensive (for a summary see: J. R. Keefe and A. J. Kresge, *Chem. Soc. Rev.*, 1996, **25**, 275) that calculation for simple ketones is all but redundant. The exceptions all possess some complication. One concerns the α -diketones (**7**) and (**8**), for the first of which (**7B**) is invisible, and for the second of which (**8A**) is nearly so. This difference is most obviously due to lone pair repulsion in the latter, and I was hoping to see calculation quantify this. Also I have my own (rough) guesses as to K_E values. The β -diketones (**19**–**21**) show some parallels with (**24**–**26**) and were chosen partly for this reason. Here (**20**, Z = CH₂) is pivotal: there is a possible ring size effect vs. (**19**, R = H); in (**21**) like (**26**) the OH tautomer is anti-aromatic; and some effect on K_T is likely if Z = CH₂ is replaced by Z = O, S or NR, but its magnitude and direction are unknown. Still more intriguing are the β -triketones (**27**) where three chief factors enter: *endo*- vs. *exo*-C=C; competitive intramolecular hydrogen bonding; and possible complications due to the nature of Z=O, S or NH. In addition, much of the published (NMR) work comes to conclusions that are all but unbelievable, and computation might provide a counterweight. Finally, the most important but least understood category in this collection, the 5-membered ring 2-oxoheterocycles (**35**), poses formidable experimental difficulties that have led so far to its total neglect, and any clues that computation can provide will be welcome.

The remainder concern NH/N tautomerism. Simple annular tautomerism is missing since quite well charted, and all apparent examples are really concerned with something else. Tautomers (**5B**) and (**5C**) are almost evenly balanced in water, but nowhere else; even a slight drop in polarity, e.g. to aqueous ethanol, provides a rapid shift in the direction of (**5B**), so I have felt this to be a potential test-bed for solvational models. The extreme improbability of iminol formation in xanthine (**17**) leaves annual tautomerism in the imidazole moiety the only realistic subject for study, and here the main test for computation is whether it can detect the *peri*-NH/NH dipolar repulsion that must make (**17B**) a much less likely proposition than (**17F**). Again, the point of (**28**) → (**31**) concerns the varying electron-deficiency of the 5-membered rings, which as it increases will, I believe, make the **B**-tautomer progressively more favoured. Finally, while the diazepine (**32**) exists overwhelmingly as the **B**-tautomer, this becomes anti-aromatic in (**33**) and (perhaps less so) in (**34**), providing a further way, along with (**21**) and (**26**), of assessing a phenomenon which is known to be important but whose quantification has proved elusive.

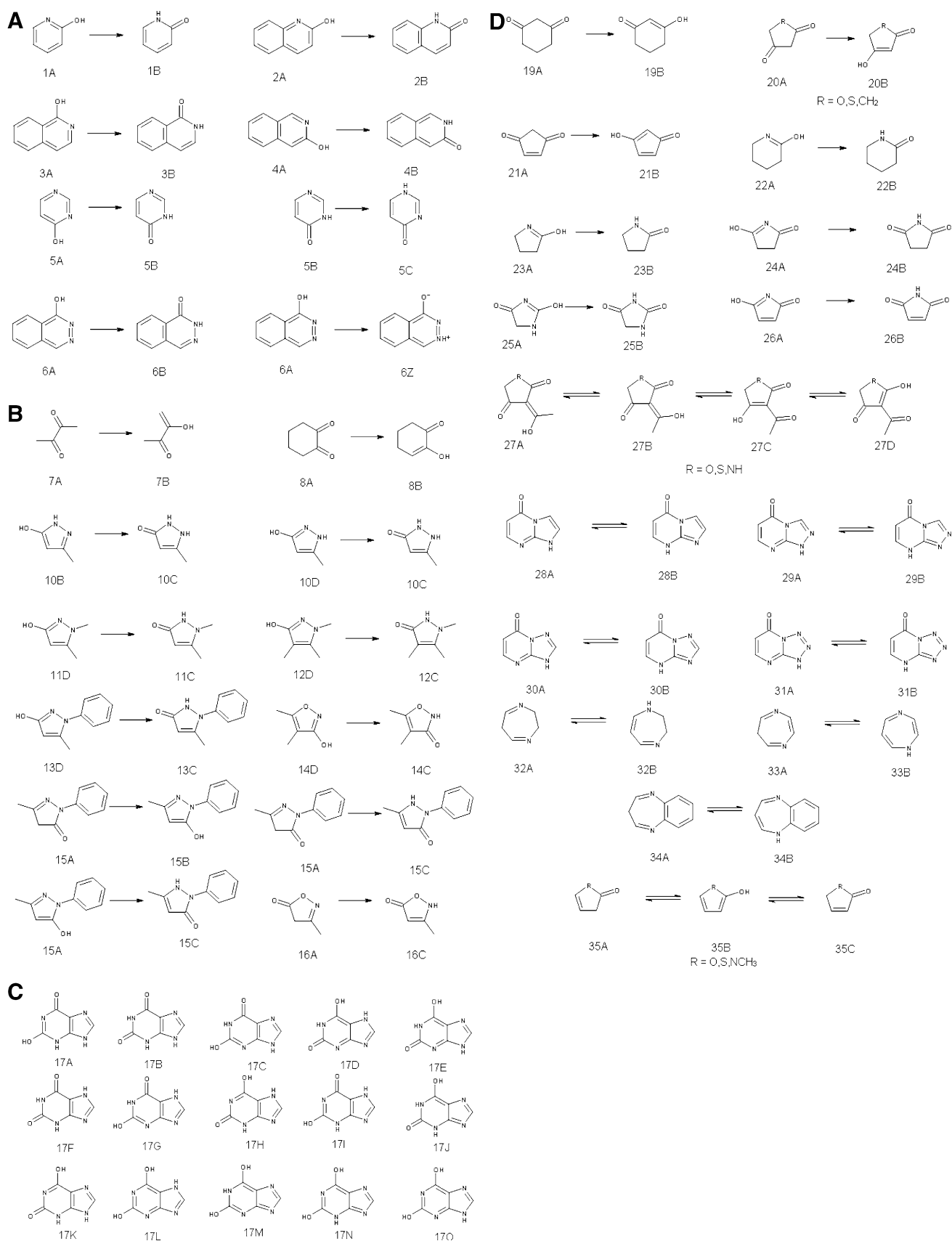


Fig. 4 Structures for the obscure section (a) and explanatory section (b) of the tautomer ratio challenge, along with the 15 tautomers of xanthine (c), which was also part of the explanatory section. (d) Tautomers in the Investigatory category of the challenge

Table 2 Tautomer pair energies for the obscure and explanatory datasets

Tautomer pair	Energy difference (kcal/mol)	Error estimate (kcal/mol)
1A_1B	-4.8	0.3
2A_2B	-6.1	0.3
3A_3B	-7.2	0.3
4A_4B	-2.3	0.4
5A_5B	-4.8	0.5
5B_5C	0.5	0.2
6A_6B	-9.2	0.4
6A_6Z	-2.4	0.3
7A_7B	7	1.5
8A_8B	-3	3
10B_10C	-2.9	0.4
10D_10C	-1.2	0.2
11D_11C	-0.5	0.2
12D_12C	-1.8	0.7
13D_13C	0.1	0.1
14D_14C	0.3	0.3
15A_15B	0.9	0.3
15A_15C	-1.2	0.3
15B_15C	-2.2	0.3
16A_16C	0.5	0.1
17A	11	
17B	3.4	
17C	8	
17D	10	
17E	6	
17F	0	
17G	8	
17H	7	
17I	8	
17 J	6	
17 K	8	
17L	>50	
17 M	>50	
17 N	>50	
17O	>50	

Predicting error

Knowledge of when a value can be predicted with little error and when it cannot provides valuable information when these calculations are used to inform decision-making. To test the ability of participants to predict the

error in their calculations they were asked to submit two estimates of error for each molecule or tautomer pair in addition to the predicted transfer energy or difference in tautomer energy: an estimate of the standard error of their method for that compound, and integer value from 1 to 5 representing their confidence in their prediction for this molecule (5 indicating highest confidence).

Choosing appropriate metrics

Several common statistical measures were calculated for each submission, and smooth bootstrapping was performed to achieve a measure of the precision of these statistics given experimental error. Mean and median raw error, mean and median absolute error, and root mean square error (RMSE) were calculated over every member of the obscure set that was predicted. Additionally, a Kendall's Tau rank correlation was determined, as well as the slope, intercept and *R*-squared for a linear fit to the data. Each of these calculations was performed for the predicted values against over 10,000 iterations of smooth bootstrapping on the experimental values. Smooth bootstrapping involves comparing the predicted values to the experimental values over many iterations, where during every iteration a value for each experimental data point was selected by adding random noise to the experimental value. The noise was chosen from a normal distribution with a standard deviation equal to the experimental error. The aggregate results for all rounds of bootstrapping were then used to determine median and 95% confidence intervals for each metric.

Predicted error was compared to the actual error in submitted predictions to determine participants' accuracy in estimating errors in their predictions. The differences in error were calculated for a single submission by subtracting the predicted standard error from the actual absolute error for each prediction in the obscure set. Thus, a negative error difference implies that the predicted error was larger than the actual error, while a positive difference indicates that the actual error was larger than the predicted error.

Similarly, the confidence estimates were compared to the actual errors through the following treatment. The mean and standard deviation of the actual absolute errors in one submission were determined, and then each prediction in that submission was assigned an error rank based on how far the actual error was from the mean (see Eq. 5). The cutoffs between ranks were designed such that each rank would be assigned to 20% of the predictions if the absolute errors were distributed normally (see Eq. 6). Thus, the ranks were independent of the actual size of the error, and depended only on the deviation in error from the mean absolute error. These error ranks were then compared to the confidence estimates using Kendall's Tau rank correlation. This correlation, and thus this analysis, was only valid in

cases where the participant did not estimate the same confidence value for all predictions.

$$ErrDev_i = \frac{(AbsErr_i - MeanAbsErr)}{StDev_{AbsErr}} \quad (5)$$

$$ErrorRank_i = \begin{cases} 5 & ErrDev_i \leq -0.8416 \\ 4 & -0.8416 < ErrDev_i \leq -0.2533 \\ 3 & -0.2533 < ErrDev_i \leq 0.2533 \\ 2 & 0.2533 < ErrDev_i \leq 0.8416 \\ 1 & 0.8416 \leq ErrDev_i \end{cases} \quad (6)$$

New metrics, derived from information theory, can incorporate effects of differences in experimental error and prediction confidence, and should provide a better measure of the predictive value of a method. If the experimental value (with experimental error) and the predicted value (with either predicted error or a confidence estimate) are interpreted as two probability distributions, these can be compared using the Kullback–Leibler divergence (KLDiv, see Eq. 7), which provides a measure of the loss of information when an approximate probability distribution, Q , is used in place of the true distribution, P . The KLDiv is always greater than or equal to zero, and is zero only if $Q=P$. The KLDiv between two gaussian distributions can be calculated analytically given their respective means and variances (Eq. 8).

$$KLDiv = \int p \ln \left(\frac{p}{q} \right) du \quad (7)$$

$$KLDiv = \frac{1}{2} \left(2 \ln \left(\frac{err_{pred}}{err_{exp}} \right) + \left(\frac{err_{exp}}{err_{pred}} \right)^2 + \frac{(E_{pred} - E_{exp})^2}{err_{pred}^2} - 1 \right) \quad (8)$$

The Kullback–Leibler divergence for each prediction was calculated by considering the experimental value and error as the mean and standard deviation of the “true” distribution, and the predicted value and predicted error as the mean and standard deviation of the “model” distribution. In this case, the method performance depends on the difference between the experimental error and predicted error as well as the differences between experimental value and predicted value. Thus a perfect prediction would recreate the experimental value *and* the experimental error, and predicted error smaller than experimental error would result in a less favorable KLDiv, even if the predicted value was equal to the experimental value. To avoid “penalizing” situations where the predicted error smaller than the experimental error, in those cases the experimental error was used for both standard deviations. A second calculation of the KLDiv was performed using the experimental error as the predicted error in all cases (Eq. 9). This provides a metric that is only dependent on the differences between the

predicted and experimental values, yet factors in the effect of differences in experimental error.

$$KLDiv_{ExpErr} = \frac{1}{2} \left(\frac{(E_{pred} - E_{exp})^2}{err^2} \right) \quad (9)$$

The connection between the KLDiv and optimal betting strategies provides a useful transformation of KLDiv into a more numerically appealing metric, referred to here as Expected Loss (ExpLoss, see Eq. 10). While attractive scientifically as a measure of the distance between a model and reality, the KLDiv faces the same numeric problems as many other metrics, namely that is a number between 0 and infinity. However, the exponential function of the negative KLDiv represents the reduction in capital from optimal betting on a predicted probability distribution rather than the true probability distribution. This converts the range of the KLDiv into the unit interval, resulting in a score resembling a similarity score, where 1 indicates perfect agreement with experiment. Another interesting feature is that the exponential form provides an effective cap to the penalty of any large KLDiv, providing a measure that is much less influenced by a few large outliers. The average of these values over all predictions is referred to as the expected loss, and represents the expected performance of the method on a future compound.

$$ExpLoss = \langle e^{-KLDiv} \rangle \quad (10)$$

Results

Transfer energies

Participation

Participation in the transfer energy portion of SAMPL2 surpassed previous years, with 47 submissions from 13 participants. The methods used to generate submissions for the transfer energy challenge were separated into three groups: implicit solvent methods using a single conformer, implicit solvent methods using multiple conformers, and molecular dynamics simulations using explicit solvent (Fig. 5).

Successful compounds

Over the entire set, half of the compounds were predicted within 2 kcal/mol of the experimental transfer energy by at least 75% of the submissions. With the exception of uracil, all of these compounds had experimental transfer energies of -11 kcal/mol or smaller. Only one of the molecules (uracil) with a transfer energy more negative than -11 kcal/mol was predicted with this accuracy, compared to

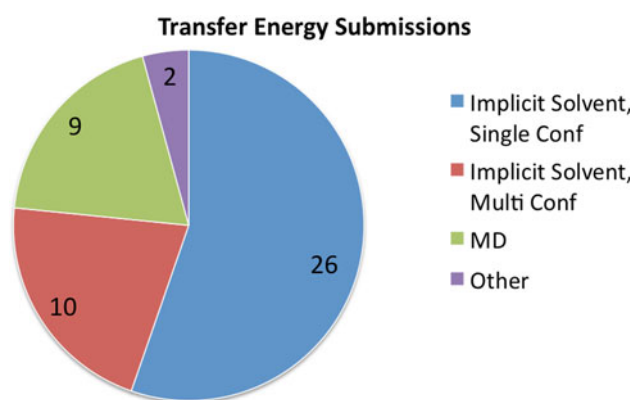
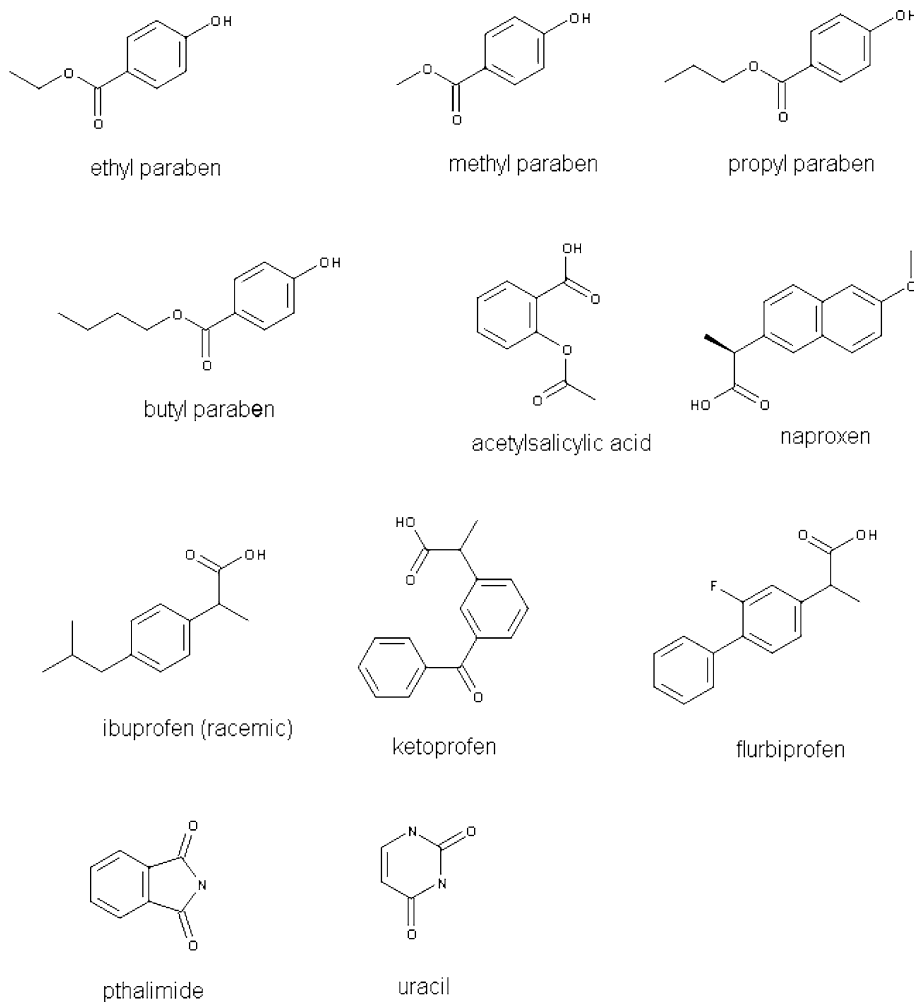


Fig. 5 Chart illustrating the distribution of methods used for the transfer energy portion of SAMPL2

all but two of the molecules with transfer energies less negative than -11 . Sulfolane and diflusalin were the only poorly predicted compounds in this range, while every member of the paraben series, the ibuprofen series, and pthalimide, naproxen, and acetylsalicylic acid were accurately predicted (Fig. 6). An example of the distribution of predictions is shown for ethyl paraben in Fig. 7.

Fig. 6 Structures of the compounds that were well predicted by a majority of the methods. Uracil is the only compound with a transfer energy more negative than -11 kcal/mol



Problem compounds

The sugars D-glucose and D-xylose, as well as glycerol from the explanatory set, were both hydroxyl-rich and the most flexible compounds in SAMPL2. The transfer energies of both sugars and glycerol were under-predicted on average (see Fig. 8). In addition to the conformational complexity of the sugar rings, rotations of the many hydroxyls in these compounds can lead to large swings in the predicted transfer energies. This conformational complexity can be problematic for both implicit solvent models, where choice of one or a few conformers may skew results, and dynamics-based calculations, where adequate sampling may be difficult to achieve. Furthermore, the initial reported value for glycerol was inaccurate. The first value was derived from experimental measurements of glycerol that were performed more than 70 years ago, and subsequent measurements have moved the experimental value closer to the average prediction.

Uracil and its several halogenated derivatives were somewhat overweighted in the obscure set, forming over 30% of the blinded molecules. Accurately capturing the

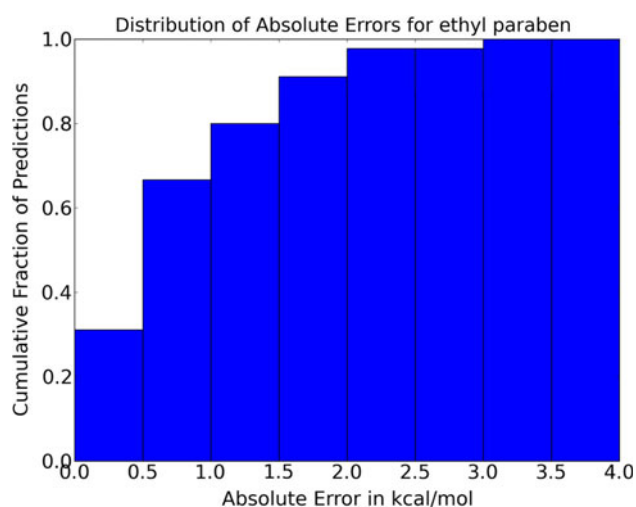


Fig. 7 Cumulative histogram of absolute error in transfer energy predictions for ethyl paraben. Over 60% of predictions were within 1 kcal/mol of the experimental value, and over 90% were within 2 kcal/mol

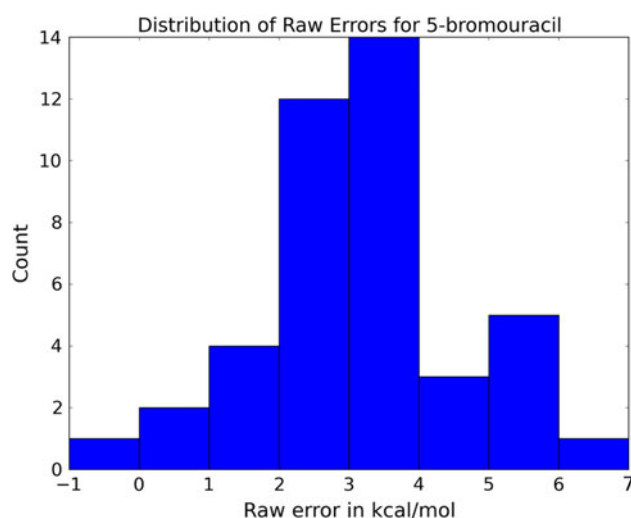


Fig. 9 Distribution of raw errors for predictions of 5-bromouracil. Note the heavy preponderance of positive raw errors, indicating that almost all predictions were more positive than the experimental value

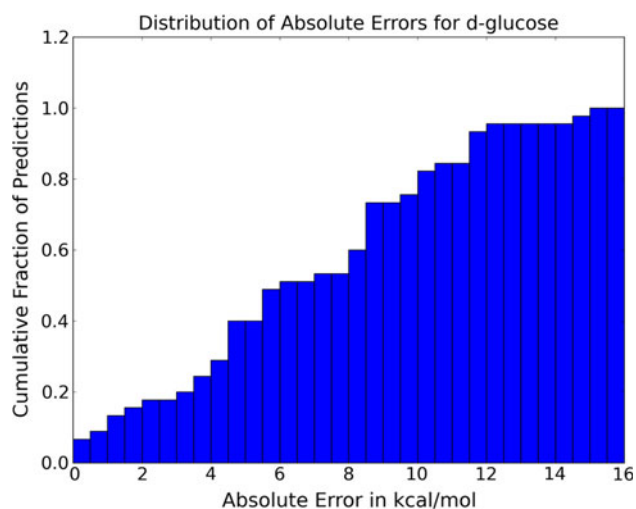


Fig. 8 Cumulative histogram of the absolute error in D-glucose predictions. Half of the predictions have errors of 6 kcal/mol or greater

trend in the transfer energies of the uracil series proved difficult for many methods, and success on the uracils was a large component of overall metrics. However, even methods that had low absolute error on the uracil set did not do well at rank ordering the compounds. Most methods consistently under-predicted transfer energies for this series, as exemplified in the histogram of raw errors in predictions for 5-bromouracil (Fig. 9).

While not part of the obscure set, the two heavily chlorinated members of the explanatory set proved important in revealing missing polarization effects in most

methods. Although experimental values were provided for these compounds, many participants submitted unmodified predictions. Thus both hexachlorobenzene and hexachloroethane were largely predicted to have positive transfer energy even though their experimental transfer energies are negative. The average prediction for each compound was over 2 kcal/mol higher than the reported value. These small molecules have no dipole moment, and so the driving force for a favorable interaction with water seems to be their large propensity for polarization imparted by the many chlorine atoms. Most methods underestimate this contribution.

Cyanuric acid, which has twice as many nitrogen and oxygen atoms as it does carbon atoms, appears to be a molecule that interacts readily with water on visual analysis. However, the initial value of the experimental transfer energy for cyanuric acid was the most positive of the obscure set, at -6.44 kcal/mol. Perhaps unsurprisingly, cyanuric acid originally proved a large outlier for every prediction method, with an average error in prediction of over 10 kcal/mol. However, the predictions of cyanuric acid were primarily in the -16 to -20 kcal/mol range (see Fig. 10), and overall the predictions of the field showed less variation than four other compounds that were predicted with greater accuracy yet less precision. The initial value was in error and the correct experimental value was -18.39 kcal/mol, which falls roughly in the center of the predicted values.

High variance compounds

High variance in the predicted results was another indicator of difficulty: D-glucose, D-xylose, caffeine, and diflunisal

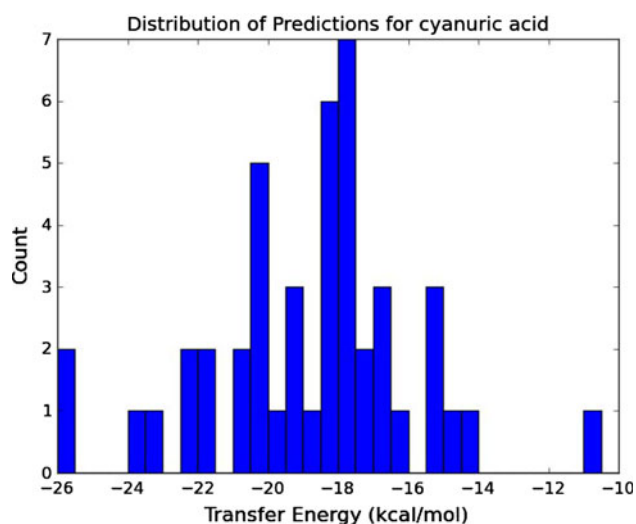


Fig. 10 Distribution of predictions for cyanuric acid. The reported experimental value was -6.44 kcal/mol, which was later revised to -18.06

all had raw error standard deviations of 3 kcal/mol or greater, while cyanuric acid had a raw error standard deviation of just less than 3 kcal/mol (see Fig. 11). The average error and standard deviations for each of the sugar compounds were roughly equivalent: 5.5 kcal/mol for D-glucose and 4.2 kcal/mol for D-xylose. Perhaps unsurprisingly for these flexible molecules, predictions of these compounds had the highest variance of any compound in SAMPL2, likely due to issues of conformation choice or the difficulty in achieving complete sampling. More surprising were the high standard deviations for diflunisal (3.3) and caffeine (3.2), despite having smaller transfer energies and only one and zero rotatable bonds respectively. The average raw errors for these compounds were

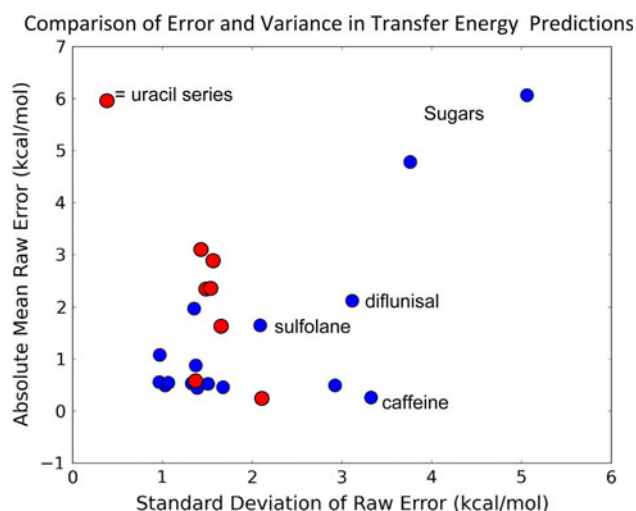


Fig. 11 Comparison between prediction error (*offset*) and prediction variance for all transfer energy submissions

less than 1 kcal/mol, implying the predictions as a whole were accurate but not precise. This is especially surprising for caffeine, which despite similarity to the uracil series, was predicted with less error but greater variance.

Method comparison

There was tight clustering among the top-performing submissions, with 10 submissions having a median absolute error within the 95% confidence interval of the submission with the lowest error. The best-performing submissions of each type of method performed within error of each other when compared by several metrics with one multi-conformer implicit solvent submission edging out the others. As shown in Fig. 12, the 95% confidence intervals of the median absolute error overlap for the best submissions using implicit solvent with single or multiple conformers, as well as MD simulations using explicit solvent treatments. When RMSE, a metric subject to bias from a small number of large outliers, is used, the methods are separated more, with the overlap in confidence intervals nearly eliminated between the multi-conformer implicit solvent method and the other approaches.

Tautomer ratios

Participation

The 20 submissions from the 7 tautomer ratio participants were all implicit solvent calculations combined with QM or DFT to calculate the energy difference between tautomers; mostly methods employed a single conformer of each tautomer, with only three submissions using multiple conformations.

Problem compounds

Although the relatively small obscure set of the tautomer ratio challenge contains almost exclusively examples of an amide-iminol tautomerization, submissions consistently incorrectly predicted only one of these pairs. Case 4 of the obscure set had a mean raw error of almost 3 kcal/mol, but all of the other similar tautomer pairs had mean raw errors within 1 kcal/mol. Given that the experimental value was

Method Type	Median Absolute Error	RMSE
Implicit: single conf*	0.79 (0.67 – 0.98)	2.07 (1.99 – 2.17)
Implicit: multi conf	0.74 (0.57 – 0.88)	1.56 (1.48 – 1.65)
MD*	0.99 (0.76 – 1.18)	2.48 (2.41 – 2.54)

Fig. 12 Performance of submissions from each method type with the smallest error as measured by median absolute error (MAE) and root mean squared error (RMSE). Asterisks denote where the lowest MAE and RMSE were obtained by different submissions

–2.30 kcal/mol, this meant the average prediction favored the tautomer that was higher in energy.

Obscure set (P. J. Taylor)

These comprise compounds (1–6). The tautomeric ratios for these compounds are known with exceptional accuracy, so that I have felt justified in applying very strict criteria to the notion of “success.” In the event, a gratifying number of attempts have passed with flying colours.

Eight equilibria are involved, seven of which proved tractable. Three methodologies from the Luque group gave really excellent results, expressed as $\Delta \log K_T = \log K_T$ (calc) – $\log K_T$ (obs):

$$334 : 0.2 \pm 0.2 (n = 7)$$

$$338 : -0.05 \pm 0.2 (n = 6)$$

$$335 : -0.1 \pm 0.4 (n = 6).$$

Methodologies 320 (Klamt) and 330 (Luque) also receive honourable mention. Of this set, 334 wins by a head through its unique ability (320 comes second) to fit the process **6A** → **6Z** in which a zwitterion is formed—I have to confess I never expected anyone to get this right, so I have to grovel. Almost as impressive is the ability of all in this set but 320 (338 comes closest) to get almost the right answer for **5B** → **5C**, whose huge difference in dipole moment—2.24D and 7.50D, respectively for the fixed NMe forms—helps to demonstrate the solvational problems that any calculation has to face. To emphasise this point, as ethanol is added to water the proportion of **5C** rapidly diminishes, to the extent that, in pure ethanol, it becomes invisible. Given that 334 predicts equimolar proportions of **5B** and **5C** it may also be of interest that this ratio is attained when electron donor 2-substituents (NH₂, OEt) are present.

Except for certain methodologies (332, 336) which persistently overpredict $\log K_T$, none come near to getting it right for **4A** → **4B**. Typically, the expected “quinonoid penalty” comes out at $\Delta \log K_T - 3.5$ to -4 , or about double its real value in free energy terms. Again, this of course is for water; it may well be greater in less polar media. There is one intriguing exception. For **1**, **2**, **3**, **4**, **5A** → **5B** and **6A** → **6B**, method 319 overpredicts $\log K_T$ by 2.8 ± 0.3 ($n = 6$); that is, all including **4** are overpredicted by about the same amount. Is there any obvious reason for this, some aspect of the calculation whose removal could lead to more accurate prediction?

Some of the trends revealed have considerable chemical interest:

- (i) One oddity revealed by experiment is that benzofusion with N in the β -position to the position of ring

fusion has a much larger effect than with N in the α -position in enhancing K_T , and this is now backed up by calculation. The oddity is that, on any simple chemical argument, the opposite would have been expected. Does any feature of any calculation throw light on this? Or are such calculations essentially impenetrable? In case this helps it is interesting to note, for 1–3, that most of the change in pK_a on benzofusion attaches to the amide tautomer; the iminols (strictly, α -OMe forms) show little difference.

- (ii) The considerable enhancement in K_T in going from 3 to 6, bucking the usual trend that electronegative substituents reduce this, is plausibly attributable to lone pair repulsion between the aza-nitrogen atoms of **6A**. (Sceptics on lone pair repulsion are welcome to put forward alternative explanations!). So is formation of **6Z**, another way of relieving it.
- (iii) On a personal note, it is very gratifying to find $\log K_T$ 6.8 for **6**, which I have derived by the use of correction factors, preferred by calculation to the published 7.8: 6.5 ± 0.5 is found for the five chosen methodologies highlighted above. The original investigation was one of the most complex ever carried out (M. J. Cook, A. R. Katritzky, A. D. Page and M. Ramaiyah, *J. Chem. Soc., Perkin Trans 2*, 1977, 1184) and its re-examination was something of a nightmare.

Xanthine analysis (P. J. Taylor)

The 15 xanthine tautomers specified and predicted in the poster by Szegezdi and Cszmadia [76] were assessed in 16 sets of calculations, all relative to **17F** taken by me as the dominant tautomer on well researched experimental evidence. It is clear that not all the original calculations were intended to be taken seriously and we may start by eliminating the total no-hopers: **17D** and **17E** as iminols and quinonoid; **17J**, **17M** and **17N** as possessing a double dose of one or the other. Nevertheless I consider later certain other no-hopers which might in principle throw light on the (small) degree of iminol formation present. I concentrate on methodologies 334 and 301 since, at this stage, these have appeared the most efficient at handling lactam tautomerism.

The only viable rival to **17F** is **17B**, placed second at 33% by Szegezdi and Cszmadia, but nobody assessed it as actually the more important, though for 301 these tautomers were evenly balanced, while 334 predicts 20% of **17B**. Most other calculations clustered around $\Delta G \sim 1.5$, i.e. about 10% of **17B**, which cannot be discounted on present experimental evidence, though I hope to have some later opportunity to describe the strong circumstantial

evidence that points to $\log K$ ca. 2.5 in favour of **17F** ($\Delta G \sim -3.5$). The ‘basicity method,’ employed on fixed NMe tautomers at N-7 and N-9, would sort this problem out with ease (and without the need for any correction factor!).

Simple iminols fall into two groups, according to the position of the imidazole proton. For H at N-7 as I propose, these are **17G** and **17I** for 2-OH, or **17H** for 6-OH. For 334 and 301, **17G** is disfavoured by $\log K - 6.2$ and -6.7 respectively, while **17I** is disfavoured by ca. -8.4 for both. This is quite an interesting result, in that **17I** is likely to be less stable than **17G** through the effect of annular tautomerism between N-1 and N-3, probably about $\Delta \log K \sim 1$ in favour of **17G** (the countervailing effect of the *peri*-lone pair interaction in **17G** should be small). These estimates compare with mine of $\log K_T$ 5.8 that results from applying basicity corrections to the published work on uracil; xanthine as less ‘aromatic’ should give a slightly higher value. The values predicted by 334 and 301 for **17H**, -8.1 and -8.4 , respectively, are anomalous in being higher than for the 2-position (my estimate for uracil, $\log K_T$ 4.6, reflects its character as formally imide carbonyl), but a *peri*-interaction in the amide tautomer between C=O and NH may help to account for this, though I have difficulty in believing that it could be so great.

For H at N-9, the corresponding values of $\log K$ for 334 and 301 are -4.6 and -6.2 in the case of **17C**, -9.9 and -9.3 in the case of **17A**, and -7.8 and -8.7 in the case of **17K**. The proton switch for the imidazole moiety should be favourable for **17C**, have no effect for **17A**, and be favourable for **17K**. The first prediction is fulfilled but the others are not; there is essentially no difference between **17H** and **17K**, while **17A** is more disfavoured, for 334, than the di-iminol **17O**, though 301, at $\log K - 12$, registers the most negative value for either methodology in the set. It may be recalled that, according to the calculations of Szegezdi and Csizmadia, **17A** should be the most important tautomer, with about 60% of the total.

As noted above, the values of $\log K$ for **17G** predicted by 334 and 301 look reasonable in the light of a postulated $\log K_T$ 5.8 for the corresponding position of uracil, and while those predicted for **17H** look too great on parallel grounds, this may be an unexpectedly large effect of ring fusion so cannot be ruled out a priori. More important is the apparent insensitivity of these methodologies and possibly all to the disfavoured *peri*-interaction involving two NH’s, especially when contrasted with the apparent over-reaction of all methodologies bar 301 to that between contiguous NHNR (or NHO) in compounds **10–16** above. Is it possible that these entail in actuality different types of specific interaction with water molecules, which continuum models unfortunately fail to pick up?

Method comparison

One submission combining extremely high-level quantum calculations with an implicit solvent model outperformed all other submissions, while the rest of the top-performing submissions were tightly clustered with similar levels of error.

This application of an MP2 complete basis set extrapolation and electron correlation correction derived from the differences between MP2 and CCSD levels of theory resulted in a bootstrapped MAE of 0.47 and RMSE of 1.1. However, the 95% confidence interval of many of the other submission overlapped with that of the top-performing method.

Predicting error

For both the transfer energy and tautomer ratio sections of SAMPL2, participants were asked to estimate the error in their predictions in two ways: by predicting the standard error and by assigning an integer confidence value from 5 to 1. Predicting which molecules would be large sources of error for their methods, either directly or by assigning a confidence value, was difficult for most participants. Many participants assigned one error or confidence value to all predictions, often based on previous performance. Only a little over a quarter of submissions included differing values for error or confidence estimates. Additionally, several of the participants who used MD methods reported sampling errors as predicted error, which were much smaller than reasonable error estimates.

Comparison of predicted error and the actual absolute error reveals that over half the submissions had average errors within 1 kcal/mol of the predicted errors (Fig. 13). However, when compared to the cases where participants actually attempted to predict different error values for different molecules, the difference in error skews higher. This may be biased by the small error estimates from many of the MD submissions. Comparison of the average error difference to the median (Fig. 14) reveals that most of the submissions had a few large discrepancies between predicted and actual error, which shifts the means to higher values than the medians.

Correlation between confidence and actual error was explored by placing molecules into 5 bins based on the deviation of each actual error from the average error (see Eqs. 5 and 6 in Methods). In this way, the correlation to estimated confidence is independent of the overall performance of the submission, and depends only on the variance in performance within the submission. Almost half of the 20 submissions that included varying confidence estimates showed little, no, or even negative correlation between the predicted confidence and the extent actual error that

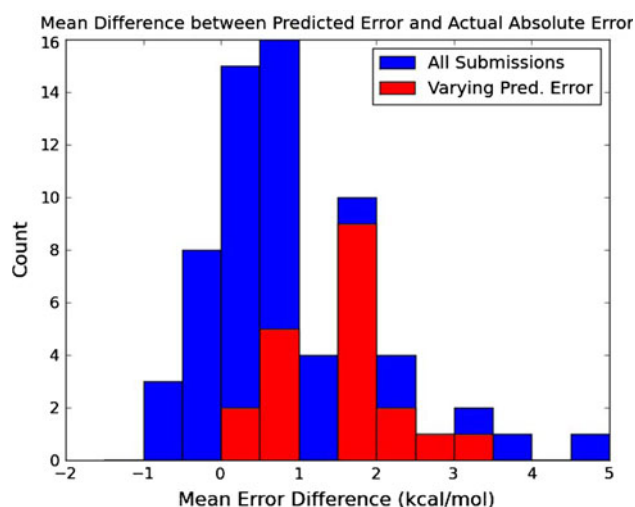


Fig. 13 Histogram of the mean difference between predicted errors and actual absolute errors for all submissions. Those submissions where the participant varied their predicted error are indicated in red

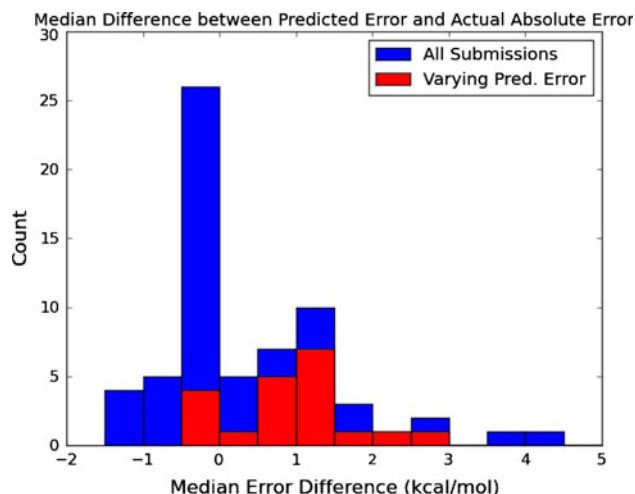


Fig. 14 Histogram of the median difference between predicted errors and actual absolute errors for all submissions. Note that the median is shifted towards smaller differences than the mean

occurred in the prediction deviated from the average error. In a few cases a reasonable correlation did exist, with 3 submissions having a correlation greater than 0.5 (see Fig. 15). These three submissions were from two different participants using different types of methods (MD with explicit water versus QM charges with implicit solvent), and high correlation between the predicted confidences for these two participants indicate they both identified molecules that were difficult independent of the method used.

Discussion

The results of this blind assessment of the state of the field for transfer energy and tautomer ratio prediction both

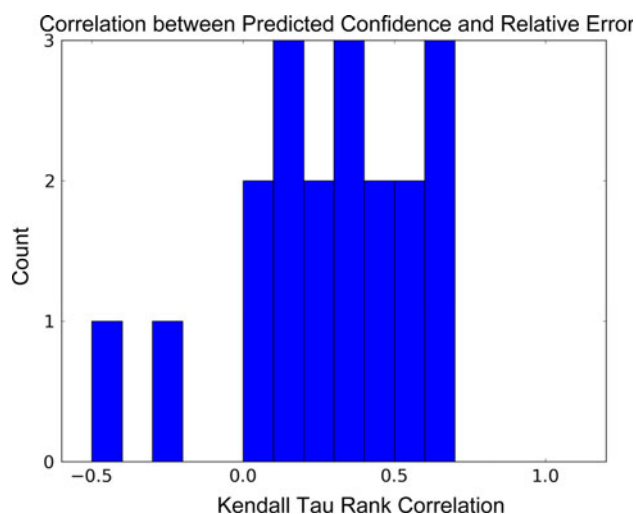


Fig. 15 Histogram depicting the correlation between predicted confidence and the actual error that occurred. Note this can only be calculated in cases where the participant submitted varying confidences

indicate where we are performing well and point out flaws in current methods. As is often the case, the difficult compounds provide the most information about where our methods require improvement.

State of the field: transfer energies

Analysis of the degree of agreement in predictions of the top-performing methods indicates overall the field does well on most compounds. It is striking that there is a clear distinction in the good performance on compounds with smaller transfer energies and the poor performance on those with larger energies. This may be a reflection of the lack of dynamic range in available and widely used transfer energy datasets, or the experimental problems that arise when dealing with compounds with such large transfer energies. Of the difficult cases some, such as caffeine and diflunisal, are characterized by large variations in the predictions, which occur even when the values are evenly distributed around the experimental value. However, other examples exist where the spread in predictions is small, but is significantly offset from the experimental values, as seen for many of the uracil derivatives. While the consistent offset seen in the uracils may indicate a systematic error either in theory or in measurement, the large variance of predictions, especially for a rigid molecule like caffeine, is more difficult to interpret.

Although we can't examine the average raw error for the investigatory set, the variance in their predicted values may shed some light on their predictive tractability. As discussed previously, low variance in predicted values was a necessary but not sufficient criterion for good prediction.

Compound	Average Pred	Std. Dev.
Thiazole	-3.9	0.88
Oxazole	-4.7	1.13
Isothiazole	-4.4	1.45
Trifluoromethyl phenyl sulfone	-5.4	1.97

Fig. 16 Four compounds with unknown experimental values were predicted with standard deviations under 2 kcal/mol

The standard deviations of predicted values were less than 2 for only four of the investigatory compounds (see Fig. 16): thiazole, oxazole, isothiazole, and trifluoromethyl phenyl sulfone. Average predicted transfer energies for these compounds were all smaller than -6 , well within the range of good performance in the obscure set. These compounds might provide desirable targets for future experimental measurement. Unsurprisingly, predictions for all of the phosphorous-containing molecules had high standard deviations, as did many of the compounds with a sulfur atom, likely due to problems with accurate partial charges or solvation radii, or due to polarizability.

The best performing submissions of all four methods highlighted in the SAMPL1 overview [6] achieved an average RMSE almost 1 kcal/mol lower against the SAMPL2 obscure set than obtained last year (see Fig. 17). This improvement in prediction accuracy from SAMPL1 to SAMPL2 is likely a result of both improvements in methods and differences in difficulty between the molecules (Refs to other SAMPL2 papers). The SAMPL2 obscure set, at 23 molecules, is half the size of the SAMPL1 set, whose 56 compounds were drawn primarily from pesticides, and contains only one sulfur-containing and no phosphorous-containing compounds. The SAMPL1 set had 26 molecules containing either a sulfur or

Method	SAMPL1 RMSE	SAMPL2 RMSE
All-Atom MD (Mobley)	3.53	2.73
FISH (Purísima)	3.63	2.24
COSMO-RS (Klamt)	2.76	1.56
ZAP (Nicholls & Ellingson)	2.44	2.17
Average of Submissions	3.09	2.18
“Perfect” Submission	-	0.18

Fig. 17 Performance comparison between transfer energy portion of SAMPL1 and SAMPL2 for methods which were used in both challenges

phosphorous atom, including 14 structures with both elements in addition to many with several conformation-dependent polar interactions. These factors likely contributed to the average decrease in RMSE from SAMPL1 to SAMPL2. The all-atom MD method applied by Mobley et al. which was executed using exactly the same protocol as the submission to SAMPL1, showed a 0.8 kcal/mol RMSE improvement on the SAMPL2 obscure set. The other three methods, which likely included modifications from the previous year’s submissions, had varied improvements from ~ 0.3 kcal/mol for ZAP, a PB/SA approach to about 1.4 kcal/mol for the FiSH methodology. The larger reductions in RMSE were likely at least partially due to improvements in the methods, while the smaller improvement in ZAP may be due to the fact that the ZAP submission in SAMPL1 omitted some of the most complex molecules for which required QM calculations did not finish [7]. It should be noted that the bootstrapping procedure, when performed on a “perfect” submission using the experimental values as predicted values, resulted in a bootstrapped RMSE of almost 0.18 kcal/mol, while averaging of all the predictions performs quite well, returning a bootstrapped RMSE of 2.18 kcal/mol.

Uracils were very prevalent in the SAMPL2 obscure set, and were some of the more difficult compounds. Interestingly, there were two cases of uracil derivatives in the SAMPL1 set, and these showed surprising differences in experimental transfer energies to the uracils in SAMPL2 (Fig. 18). While all the uracils in SAMPL2 had transfer energies between -15 and -19 kcal/mol, the two in SAMPL1 had experimental values of -9.73 and -11.14 kcal/mol. Each of the uracil derivatives in SAMPL1 had two additions relative to the comparable uracil in

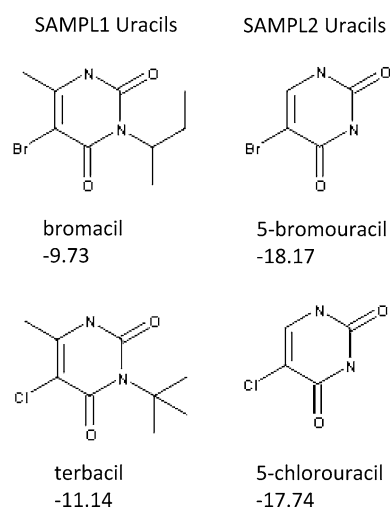


Fig. 18 Comparison between the two uracil analogues in SAMPL1 and the comparable molecules in SAMPL2

SAMPL2: a methyl group and a large, bulky, butyl group (sec-butyl and tert-butyl, respectively). Although these additions would certainly result in a less negative transfer energy, the experimental values were shifted by 8.4 and 6.6 kcal, which may be unduly large. Although the experimental values in SAMPL1 could not be traced back to reports with experimental details, and thus been assigned large uncertainties of 1.93 kcal/mol, this discrepancy in experimental values is particularly interesting given the uracil series as a class was generally predicted to have more positive energies by the field.

State of the field: tautomer ratios (P. J. Taylor)

Tautomer preference, let alone tautomer ratio, was for decades a minority interest among chemists and the main incentive for investigating either has come from the biological interface. J. D. Watson has described, in ‘The Double Helix’ (1968), how elucidating the structure of DNA was held up for years by the attempt to fit its binding pattern to the wrong tautomeric match. Even when chemists knew that, e.g. ‘2-hydroxypyridine’ as required by the systematic chemical nomenclature of the time did not fit the facts, they often failed to explain it to biochemists and others on the assumption that it did not matter. And, before the development of modern physical methods, tautomer preference was infernally difficult to establish. Tautomers are slippery customers. Chemical methodologies known to be reliable elsewhere are liable to fail spectacularly for compounds with mobile protons; for example, 2-pyridone on methylation gives mostly 2-methoxypyridine. Mechanistic theories of chemical reactivity were simply not well enough developed, till the 1950s, to explain this apparent paradox. Even physicochemical methodologies can sometimes be fooled; for example, tautomer interconversion is generally too fast for NMR time-averaging to handle. And physical chemists, who like to work with ultrapure compounds of unequivocal structure, have tended to shun this messy field.

So the subject has been left to physically minded organic chemists. For quantitative results the work is time consuming and a lot of preliminary synthesis may be involved. It is scarcely surprising if few such investigations have been duplicated, hence conventional statistics can rarely be employed. When I started to collate the known information I had no idea what I would find. However, there has been a heartening and quite unexpected development. I have explained elsewhere how model compound pK_a values frequently fail to reproduce those of the (usually unobservable) minor species they are supposed to model, so that I had to start by deriving correction factors based, for the most part, on compounds for which the typically minor tautomer had, for some reason, become the major or

at least a comparable one. In some of these cases, if not in all, it is possible to apply statistics to the correction factors. When their use then produces a sharp improvement in the *regularity* of incremental effects, e.g. for benzofusion, it becomes difficult to believe this regularity to be accidental, which helps to generate confidence in the original data: the Reverend Thomas Bayes and his “bundle of sticks” comes to mind. Or so I personally believe.

It is a fortunate fact, though not a coincidence, that most quantitative work has been carried out in aqueous solution. It is now becoming widely accepted that computation, to be of most use, will have to address the problem of solvation by water. At the same time I find it encouraging that at OpenEye, at least, experiment is understood to be the stone on which computational methodologies have to be sharpened and that more experimental results are needed, carefully chosen to plug the gaps. If this advice is heeded I think we should be in for a very fruitful time.

Information-based metrics

Most metrics (e.g. RMSE, MUE) applied to predictions are pure comparisons between the actual (experimental) value and the predicted value, often ignoring experimental error. However, every reputable experiment has error bars to describe the precision with which the experiment was performed. Predictions likewise can have an associated error, whether through simple estimation of confidence or more rigorous calculations of the propagation of error. Both experimental and predicted error provide important guidance on how much trust to place in either value. Metrics derived from information theory incorporate experimental error and value more accurate a priori assessments of the error in predictions, which are of importance whenever predictions are used to advise costly decisions. Consider the case of the prediction of an experimental value that is off by 0.5 kcal/mol. If the experimental measure has tight error bars, it is clear the prediction is incorrect. Alternatively, if the experiment has large error bars it is not clear that the prediction is inaccurate, in fact it is a successful prediction within the limits of what is known experimentally. Similarly, consider two predictions, one estimated to have high confidence and one with low confidence. If they are both off by a large amount, clearly the prediction with low confidence was superior. Both of these examples demonstrate situations where information-based metrics would favor the more useful case while commonly used metrics would not distinguish between the two.

The information-based metrics used in this study identified a few submissions with better performance than was evident by more conventional metrics. The inclusion of effects of error estimates (experimental and predicted) as well as the capping effect of the functional form of

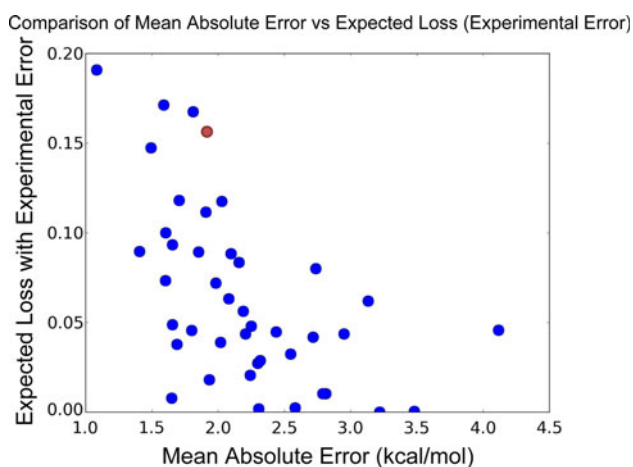


Fig. 19 Comparison of Mean Absolute Error with the information-based Expected Loss metric. Note the Expected Loss with Experimental Error incorporates information about experimental error, not predicted error

Expected Loss leads to a significantly different measure of performance. This is indicated by the low correlation between Expected Loss and Mean Absolute Error. In fact, the fourth-best method by Expected Loss (the red datapoint in Fig. 19) is an explicit solvent submission with a mean absolute error of greater than 2 kcal/mol. We believe that information-based metrics are fundamentally appropriate for measuring performance in the computational chemistry field, and would precipitate a needed evaluation of error in our predictions.

Challenge to experimentalists

A prerequisite to improvement of computational predictions is the existence of high-quality experimental data to which our calculations can be compared. This challenge has, over the past 3 years, highlighted the need for further measurements of transfer energies. The number of measured transfer energies is thought to be less than 2,000. Many of these are well known to computational chemists, and compose training and test sets for the development of current methods, making them unsuitable for a blind challenge. Therefore, construction of this challenge required extensive examination of literature to obtain values for compounds likely not to have been used in parameterization of existing methods. The scarcity of recent experimental data requires delving further and further into older publications, which can leave questions about the quality of data relative to current techniques, and result in the use of obscure or unusual molecules. When the experimental values are not directly measured, there is always uncertainty in the veracity of the experiments. Additionally, the prospects of finding existing yet unknown or inaccessible data are becoming increasingly bleak in the

information age. This shrinking pool of candidates for blind challenges hampers the ability to test methods and creates a temptation for over-parameterization to boost performance against known data. Due to the paucity of existing, blinded data, our ability to improve the field through events like SAMPL will be severely hampered without new data.

Results from this challenge demonstrate that current methods perform much better on compounds with smaller transfer energies, and the range of good performance overlaps well with the energy range of most known transfer energies. To expand the useful range of these computational methods new data is needed on a wider variety of molecules. Experimental data from specific types of compounds are necessary to continue to extend and improve existing methods. Rather than small, monofunctional compounds, which comprise a majority of the known transfer energies, measurements for drug-like molecules or pollutants could expand the breadth of available data and provide further challenges to prediction methods. Ideally, new data might also span a series of similar compounds with changes in functional groups, providing comparisons that could test specific aspects of methods or probe intramolecular interactions of functional groups.

The predictions of compounds for which experimental data does not yet exist, along with the cases where predictions brings the experimental data into question, provide clear opportunities for experimentalists to enter into the debate. Predictions made in this challenge for the transfer energy of thiazole, isothiazole, and oxazole had relatively low variance, and were predicted within a range that should be accessible to experimental measure. Alternatively, many molecules in the investigatory set had large variations in predicted transfer energies, providing an opportunity for experiments to provide data on compounds that could differentiate between current methods. Lastly, there is an opportunity to examine cases where calculations do not agree with potentially questionable experimental results. These instances provide the possibility of either correcting an erroneous value in the published literature or proving an important shortfall in current methods. Even in this work, computational results led to a re-examining of several of the experimental values, sometimes resulting in corrections to values as seen in Table 1.

Conclusion

We must walk before we can run. If computational chemistry, as a field, is ever to be able to predict solubilities, binding affinities or other critically important physical properties, we must first assess our ability to predict more fundamental values. This requires continued critical

evaluation of our methods, ideally through blind prediction, against experimental data unknown to the participants. Significant advancement will only be possible with ongoing measurement of new experimental data for relevant and informative compounds.

The results from SAMPL2 highlight the success and the shortcomings of current methods for transfer energy predictions. We expect that the proper application of any of the following seven methods (All-atom MD [77], multi-conformer implicit solvation models such as COSMO-RS [78] and EPIC [79], and the FiSH [80], IEF-PCM [81], SM8 [8], or ZAP [82, 83] implicit solvation methods which used single conformers) should lead to, with 95% confidence, transfer energies predictions within 2.0 kcal/mol of the experimental value for new unknowns. However, the average error for each of these methods has a greater than 95% chance of being larger than 1 kcal/mol, indicating there is still room for improvement.

We expect that if you want to reliably calculate tautomer ratios with an error of less than 1.5 kcal/M, you should calculate the gas-phase electronic energy difference with at least MP2/pVDZ level of theory. Using a complete basis set extrapolation with higher levels of electron correlation was able to push the error below 1 kcal/mol with 95% confidence. Because the tautomer challenge involved only a few similar transitions, it is difficult to draw general conclusion about tautomer prediction from this exercise alone. For the aqueous transfer portions of the thermodynamic cycle, several solvation models may prove reliable (vide supra), but the methods of Luque [81], Klamt [78], and Kast [84] have all been demonstrated to perform equivalently well.

SAMPL3 will occur in the fall of 2010 with further prospective evaluations. We plan to again have a transfer energy component to the challenge, although availability of blinded data for suitable compounds will likely again be a limiting factor. Additionally, there will be a component to SAMPL3 derived from a significant amount of data related to fragment-based drug discovery. More details can be found at the SAMPL website (sampl.eyesopen.com), and all groups are welcome to participate. We believe that only through continued blinded, public evaluation and analysis of our methods can we perceive and promote improvement in our field.

References

- Hastie T, Tibshirani R, Friedman J (2009) *The elements of statistical learning: data mining, inference, and prediction*, 2nd edn. Springer, Berlin
- Bordner A, Cavasotto C, Abagyan R, Phys J (2002) Chem. B 106:11009–11015
- Cramer C, Truhlar D (2008) *Acc Chem Res* 41:760–768
- Klamt A, Mennucci B, Tomasi J, Barone V, Curutchet C, Orozco M, Luque F (2009) *Acc Chem Res* 42:489–492
- Cramer C, Truhlar D (2009) *Acc Chem Res* 42:493–497
- Guthrie J (2009) *J Phys Chem B* 113:4501–4507
- Nicholls A, Wlodek S, Grant J (2009) *J Phys Chem B* 113:4521–4532
- Ribeiro RF, Marenich AV, Cramer CJ, Truhlar DG (2010) *J Comput Aided Mol Des* 24. doi:10.1007/s10822-010-9333-9
- Avdeef A (2007) *Adv Drug Deliv Rev* 59:568–590
- Cesaro A, Russo E, Crescenzi V (1976) *J Phys Chem* 80:335–339
- Hopfinger AJ, Esposito EX, Llinàs A, Glen RC, Goodman JM (2009) *J Chem Inf Model* 49:1–5
- Bardi G, Bencivenni L, Ferro D, Martini B, Cesaro SN, Teghil R (1980) *Thermochimica Acta* 40:275–282
- Kozyro AA, KABO GY, Soldatova TV, Simirskii VV, GOGO-LINSKII V, Krasulin AP, Dudarevich NM (1992) *Russ J Phys Chem* 66:1374–1377
- De Wit HGM, Van Miltenburg JC, De Kruif CG (1983) *J Chem Thermodyn* 15:651–663
- Reid RC, Prausnitz JM, Poling BE (1987) *The properties of gases and liquids*. MacGraw-Hill, New York, p 256
- Emel'yanenko VN, Verevkin SP (2008) *J Chem Thermodyn* 40:1661–1665
- Guthrie JP (1976) *Can J Chem* 54:202–209
- Guthrie JP (1986) *Can J Chem* 64:635–640
- Frisch MJ, Trucks GW, Schlegel HB, Scuseria GE, Robb MA, Cheeseman JR, Montgomery Jr JA, Vreven T, Kudin KN, Burant JC et al (2003) *Gaussian 03*, Revision B. 04, Gaussian, Inc., Pittsburgh
- Bergström CAS, Norinder U, Luthman K, Artursson P (2002) *Pharm Res* 19:182–188
- Perlovich G, Kurkov S, Kinchin A, Bauer-Brandl A (2004) *AAPS J* 6:22–30
- Szterner P (2008) *J Chem Eng Data* 53:1738–1744
- Szterner P, Kaminski M, Zielenkiewicz A (2002) *J Chem Thermodyn* 34:1005–1012
- Allexander KS, Laprade B, Mauger JW, Paruta AN (1978) *J Pharm Sci* 67:624–627
- Perlovich GL, Rodionov SV, Bauer-Br A (2005) *Eur J Pharm Sci* 24:25–33
- Boller A, Wiedemann HG (1998) *J Therm Anal Calorim* 53:431–439
- Kaminski M, Zielenkiewicz W (1985) *Calorim Anal Therm* 16:281
- Belaj F, Tripolt R, Nachbaur E (1990) *Monatshefte Für Chemie/Chem Mon* 121:99–108
- Goldberg RN, Tewari YB (1989) *J Phys Chem Ref Data* 18:809
- Oja V, Suuberg EM (1999) *J Chem Eng Data* 44:26–29
- Perlovich GL, Kurkov SV, Bauer-Brandl A (2006) *Eur J Pharm Sci* 27:150–157
- Perlovich GL, Kurkov SV, Bauer-Brandl A (2003) *Eur J Pharm Sci* 19:423–432
- Avdeef A, Berger CM, Brownell C (2000) *Pharm Res* 17:85–89
- To EC, Davies JV, Tucker M, Westh P, Trandum C, Suh KS, Koga Y (1999) *J Solution Chem* 28:1137–1157
- Ross GR, Heideger WJ (1962) *J Chem Eng Data* 7:505–507
- Cammenga HK, Schulze FW, Theuerl W (1977) *J Chem Eng Data* 22:131–134
- Filosofo I, Merlin M, Rostagni A, Nuovo Cimento II (1943–1954) 7 (1950) 69–75
- Tang IN, Munkelwitz HR (1991) *J Colloid Interf Sci* 141:109–118
- Miller MM, Ghodbane S, Wasik SP, Tewari YB, Martire DE (1984) *J Chem Eng Data* 29:184–190
- Ruelle P, Kesselring UW (1997) *Chemosphere* 34:275–298

41. Shiu WY, Wania F, Hung H, Mackay D (1997) *J Chem Eng data* (print) 42:293–297
42. Weil L, Dure G, Quentin KE (1974) *Z Wasser-Abwasser-Forsch.* 7:169–175
43. Verevkin SP, Emel'yanenko VN, Klamt A (2007) *J Chem Eng Data* 52:499–510
44. Farmer WJ, Yang MS, Letey J, Spencer WF (1980) *Soil Sci Soc Am J* 44:676–680
45. Sears GW, Hopke ER (1949) *J Am Chem Soc* 71:1632–1634
46. Wania F, Shiu WY, Mackay D (1994) *J Chem Eng Data* 39: 572–577
47. Altschuh J, Brüggenmann R, Santl H, Eichinger G, Piringer OG (1999) *Chemosphere* 39:1871–1887
48. Atlas E, Velasco A, Sullivan K, Giam CS (1983) *Chemosphere* (Oxford) 12:1251–1258
49. Jantunen LM, Bidleman TF (2006) *Chemosphere* 62:1689–1696
50. Hellmann H (1987) *Fresenius Zeitschrift Fuer Analytische Chemie ZACFAU* 328:475–479
51. Ten Hulscher TE, Van Der Velde LE, Bruggeman WA (1992) *Environ Toxicol Chem* 11:1595–1603
52. Ivin KJ, Dainton FS (1947) *Trans Faraday Soc* 43:32–35
53. Warneck P (2007) *Chemosphere* 69:347–361
54. Ashworth RA, Howe GB, Mullins ME, Rogers TN (1988) *J Hazard Mater* 18:25–36
55. Perlovich GL, Kurkov SV, Hansen LK, Bauer-Brandl A (2004) *J Pharm Sci* 93:654–666
56. Perlovich GL, Kurkov SV, Kinchin AN, Bauer-Brandl A (2003) *J Pharm Sci* 92:2502–2511
57. Perlovich GL, Kurkov SV, Kinchin AN, Bauer-Brandl A (2004) *Eur J Pharm Biopharm* 57:411–420
58. Brisset JL (1985) *J Chem Eng Data* 30:381–383
59. LePree JM, Mulski MJ, Connors KA (1994) *J Chem Soc, Perkin Trans* 2:1491–1497
60. Ferro D, Piacente V (1985) *Thermochemica Acta* 90:387–389
61. Majury TG (1956) *Chem Ind* 349–350
62. Malaspina L, Gigli R, Bardi G, Maria GD (1973) *J Chem Thermodyn* 5:699–706
63. Sawanoi Y, Shimbo Y, Tabata I, Hisada K, Hori T (2002) *Dyes Pigm* 52:29–35
64. Shimizu T, Ohkubo S, Kimura M, Tabata I, Hori T (1987) *J Soc Dyers Colour* 103:132–137
65. Clever HL (2005) *J Phys Chem Ref Data* 34:2347–2349
66. Scharlin P, Battino R (1994) *Fluid Phase Equilibria* 95:137–147
67. Kawamoto K, Urano K (1989) *Chemosphere* (Oxford) 19: 1223–1231
68. Lunden H, Chim J (1907) *Physique* 5:145–185
69. Ribeiro da Silva MA, Santos CP, Monte MJ, Sousa CA (2006) *J Therm Anal Calorim* 83:533–539
70. Benoit RL, Choux G (1968) *Can J Chem* 46:3215–3219
71. Tommila E, Lindell E, Virtalaine M, Laakso R (1969) *Suom Kemistil B* 42:95
72. Steele WV, Chirico RD, Knipmeyer SE, Nguyen A (1997) *J Chem Eng Data* 42:1008–1020
73. Zielenkiewicz W, Szterner P (2004) *J Chem Eng Data* 49: 1197–1200
74. Wolfenden R, Williams R (1983) *J Am Chem Soc* 105: 1028–1031
75. Herskovits TT, Harrington JP (1972) *Biochemistry* 11:4800–4811
76. Szegezdi J, Csizmadia F (2007) Tautomer generation. pKa based dominance conditions for generating dominant tautomers. In: American Chemical Society Fall National Meeting, ChemAxon Ltd., Budapest
77. Klimovich PV, Mobley DL (2010) *J Comput Aided Mol Des* 24. doi:10.1007/s10822-010-9343-7
78. Klamt A, Diedenhofen M (2010) *J Comput Aided Mol Des* 24. doi:10.1007/s10822-010-9354-4
79. Meunier A, Truchon J-F (2010) *J Comput Aided Mol Des* 24. doi: 10.1007/s10822-010-9339-3
80. Purisima EO, Corbeil CR, Sulea T (2010) *J Comput Aided Mol Des* 24. doi:10.1007/s10822-010-9341-9
81. Soteras I, Orozco M, Luque FJ (2010) *J Comput Aided Mol Des* 24. doi:10.1007/s10822-010-9331-y
82. Ellingson BA, Skillman AG, Nicholls A (2010) *J Comput Aided Mol Des* 24. doi:10.1007/s10822-010-9355-3
83. Nicholls A, Wlodek S, Grant JA (2010) *J Comput Aided Mol Des* 24. doi:10.1007/s10822-010-9334-8
84. Kast SM, Heil J, Güssregen S, Schmidt KF (2010) *J Comput Aided Mol Des* 24. doi:10.1007/s10822-010-9340-x