TAUTOMERIC TEASERS : SAMPL 2009

P. J. Taylor, AstraZeneca (retired)

INTRODUCTION

The aim of this exercise is to predict tautomer ratio in aqueous solution, for which purpose isolated molecule calculations are unlikely to be sufficient. The reliabilility of the solvational model will frequently be crucial, notably for case (2) *(q.v.)*, which may prove useful as a test bed.

Three categories of puzzle are presented here. In the first, cases (1) – (3), the answer is known quantitatively but is not revealed at this stage; a certain degree of surprise may attach to the results. In the second, (4) – (6), the answer is known at least qualitatively and is revealed, and what may sometimes appear a surprising result depends on one or more chemical phenomena which are not always as well recognised as they should be. Their identification is invited. In the remainder, (7) – (12), the answer is known and/or predictable qualitatively on good chemical grounds, but quantitative data are mostly absent and the purpose of this exercise is to attempt their prediction.

Those who participate are entitled to enquire how accurate are the data they are asked to emulate. In the rare cases where both (all) the tautomers are simultaneously visible, *e.g.* as their UV spectra, the answer is very accurately indeed, but this is confined to the range $0.1 < K_T < 10$ or not much outside it. Leaving aside the work initiated by Kresge[1] which allows much more than stoichiometric quantities of the minor tautomer to be generated, most work has been carried out in organic solvents whose results are not necessarily relevant to aqueous solution; nevertheless, if all solvents and the solid state agree on the dominant tautomer, water is unlikely to be *qualitatively* out of line. (Kresge's technique[1] is suited to aqueous solution but is confined necessarily to slow proton transfers on and off carbon; it cannot be used for the much more widespread and important fields of N ↔ N and N ↔ O proton transfers). The most generally applicable method for dealing with high tautomer ratios in aqueous solution is the 'basicity method' originated by Angyal and Angyal[2] in 1952 and which set off the explosion in quantitative studies that dominated the next 30 years. In principle, it can be applied to any tautomeric system in which the pK_a values of the rival tautomers are linked by a common cation. I hope to speak on this topic at the SAMPL workshop since only an inadequate outline is possible here. Briefly, NMe and OMe derivatives of NH and OH, in which there are no mobile protons to transfer, are used as models for the pK_a values of the individual tautomers; either the minor tautomer alone, or both. Despite a clear demonstration as long ago as 1971 that this assumption can give highly misleading results,[3] because the real and model pK_a values do not necessarily coincide, this methodology has been used by almost everyone who has worked in the field without regard to its limitations. In qualitative terms there is little need to criticise this approach, but for quantitative reliability as the foundation of some future computer-based algorithm, such distortions are unacceptable. My first task in attempting to take this subject in hand was their elimination. This has been accomplished by deriving empirical correction factors, now numbering almost 30, that can be applied across the board. In consequence most of the K_T values I shall quote are not those you will find in the literature, but an unexpected bonus has resulted. It is possible, for the first time, to detect *quantitative* regularities in the data that have been masked till now, and whose presence should make the task of writing such an algorithm much easier.

The internal consistency of the results provides compelling evidence for the assumptions employed. In the text that follows log K_T is quoted to the nearest 0.1 which implies Δ log K_T not greater than ± 0.2. This should be good enough for present purposes. Every quantitative value I quote has been re-calculated from the original data, before if necessary applying a correction factor, and leaving that feature aside, I have found scarcely any mistakes in this mostly ancient and certainly venerable literature.

The other methodology that allows prediction of aqueous tautomer ratios is LSER (linear solvation energy relationships), in which a multiple regression technique involving the use of empirical solvent parameters is used to derive solute-specific equations for rate or equilibrium processes.[4] In principle this can predict behaviour for solvents outside the experimental set, such as water, and except for certain types of compound whose misbehaviour is starting to be predictable, these estimates seem to be reliable where they can be tested. The chief problem with LSER is the lavish experimentation entailed, a minimum of six or eight strongly contrasted solvents being required for reliable extrapolation. Unsurprisingly, there have been few such studies, but their potential is very high in handling multiple equilibria to which the 'basicity method' is commonly inapplicable, as in case (5) below. Their accuracy is more difficult to assess, but the best appear to rival that of the 'basicity method' and the worst, excluding the special cases noted above, probably lie within Δ log K 0.3 – 0.5 of the target.

References

1 See, *e.g.*, A. J. Kresge, *Acc. Chem. Res.*, 1990, **23**, 43; J. R. Keefe and A. J. Kresge, *Chem. Soc. Rev.*, 1996, **25**, 275.

2 S. J. Angyal and C. L. Angyal, *J. Chem. Soc.*, 1952, 1461.

3 E. Spinner and G. B. Yeoh, *J. Chem. Soc. (B)*, 1971, 279.

4 M. J. Kamlet, J.-L. M. Abboud, M. H. Abraham and R. W. Taft, *J. Org. Chem.*, 1983, **48**, 2877.

5 P. W. Kenny and P. J. Taylor, *The Prediction of Tautomer Preference in Aqueous Solution*, in preparation.

6 M. J. Field and I. H. Hillier, *J. Chem. Soc., Perkin Trans. 2*, 1987, 617.

7 M. J. Cook, A. R. Katritzky, A. D. Page and M. Ramaiyah, *J. Chem. Soc., Perkin Trans. 2*, 1977, 1184.

8 (a) K. Hagen and K. Hedberg, *J. Am. Chem. Soc*, 1973, **95**, 8266; (b) P. J. Taylor, unpublished observations; (c) N. Sleszynski and P. Zuman, *J. Org. Chem.*, 1987, **52**, 2622; (d) G. Schwartzenbach and Ch. Wittwer, *Helv. Chim. Acta*, 1947, **30**, 669; (e) J. P. Segretario, N. Sleszynski and P. Zuman, *J. Electroanal. Chem. Interfacial Electrochem.*, 1986, **214**, 259.

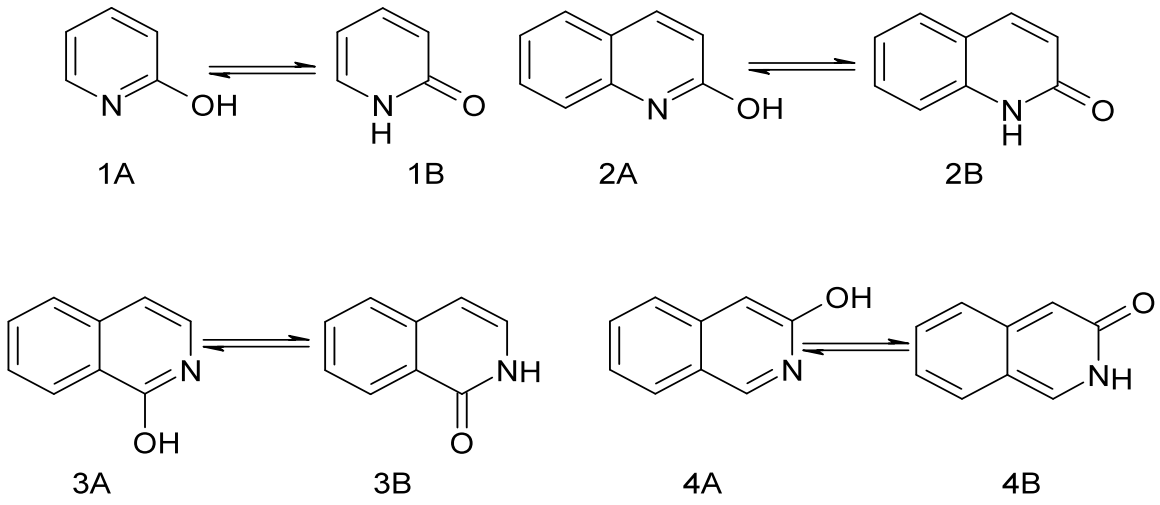
9 (a) A. R. Katritzky and F. W. Maine, *Tetrahedron*, 1964, **20**, 299, 315; (b) J. Elguero, R. Jacquier and G. Tarrago, *Bull. soc. chim. Fr.*, 1967, 3772, 3780; (c) O. G. Parchment, D. V. S. Green, P. J. Taylor and I. H. Hillier, *J. Am. Chem. Soc.*, 1993, **115**, 2352.

10 (a) A. J. Boulton, A. R. Katritzky, A. M. Hamid and S. Øksne, *Tetrahedron*, 1964, **20**, 2835; (b) A. R. Katritzky, S. Øksne and A. J. Boulton, *Tetrahedron*, 1962, **18**, 777.

11 W. Freyer, H. Köppel, R. Radeglia and G. Malewski, *J. prakt. Chem.*, 1983, **325**, 238.

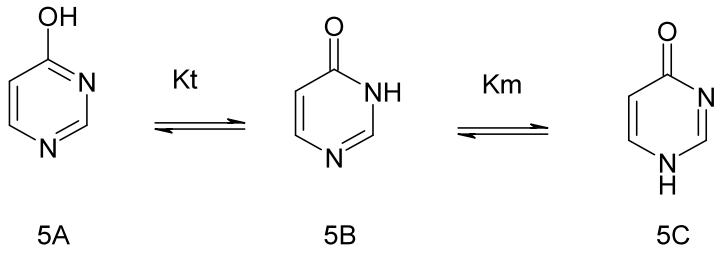
12 (a) J. Szegezdi and F. Csizmadia, ChemAxon Ltd., Budapest, Hungary, paper (poster) presented at ACS Fall National Meeting, August 2007; http://www.chemaxon.com/conf/Tautomer_generation_A4.pdf (b) D. Lichtenberg, F. Bergmann and Z. Neiman, *J. Chem. Soc, Perkin Trans. 1*, 1971, 1676; (c) B. Pullman and A. Pullman, *Adv. Heterocyclic Chem.*, 1971, **13**, 77.

Case (1) 6-Membered Aromatic Oxoheterocycles



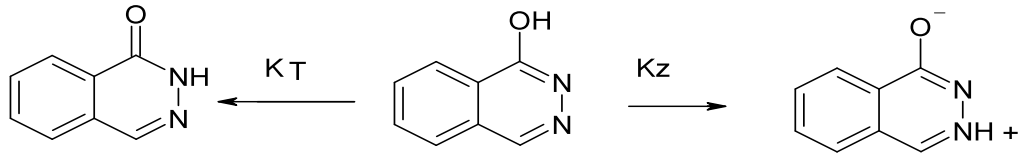
These four examples show 2-pyridone **1B**, its tautomeric iminol **1A**, and the corresponding tautomers of the three benzofused oxoheterocycles that can be constructed from it. For **1 – 3** the log K_T values (**A** → **B**) have been obtained from literature data with correction factors added, while the value for **4** has been derived by LSER analysis (unpublished[5]). The results of calculation should show widely differing positional effects for each of which there is a rational chemical explanation. There has been an attempt to compute K_T for 2-pyridone[6] but none, so far as I am aware, for any of the others.

Case (2) The 3-way Tautomerism of 4-Pyrimidone



The three-way tautomerism of 4-pyrimidone is shown above (note that we[5] use K_T and K_M to distinguish usefully between N ↔ O and N ↔ N tautomerism respectively; see later entries also). Literature data for the 2-methyl derivative have been processed, with the assistance of correction factors, to obtain both values; our estimate is that K_T for the parent compound should be less by Δ log K_T *ca*. 0.2. The value of K_M is known to be insensitive to the 2-substituent but extremely sensitive to solvent variation and the principal aim of putting forward this example is to provide a test bed for solvational models. I am unaware of any published computations.

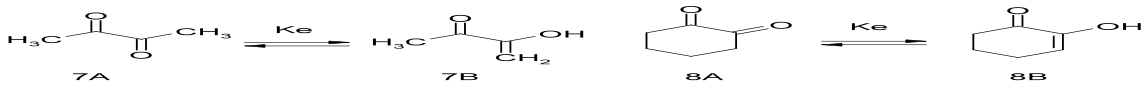
(3) The Tautomerism of 4-Phthalazinone



6B 6A 6Z

Disentangling the tautomer ratios for this compound was an incredibly complex process,[6] involving the use of six model compounds each requiring its own pK_a correction factor. It is perhaps not surprising, therefore, that its 'corrected' log K_T value differs more from that published than for any other oxoheterocycle. Here the abnormal value of K_T , and the existence of **6Z**, both probably are due to the same structural feature. It will be interesting to see whether computation can identify it.

Case (4) : *cis*- and *trans*- α-Diketones



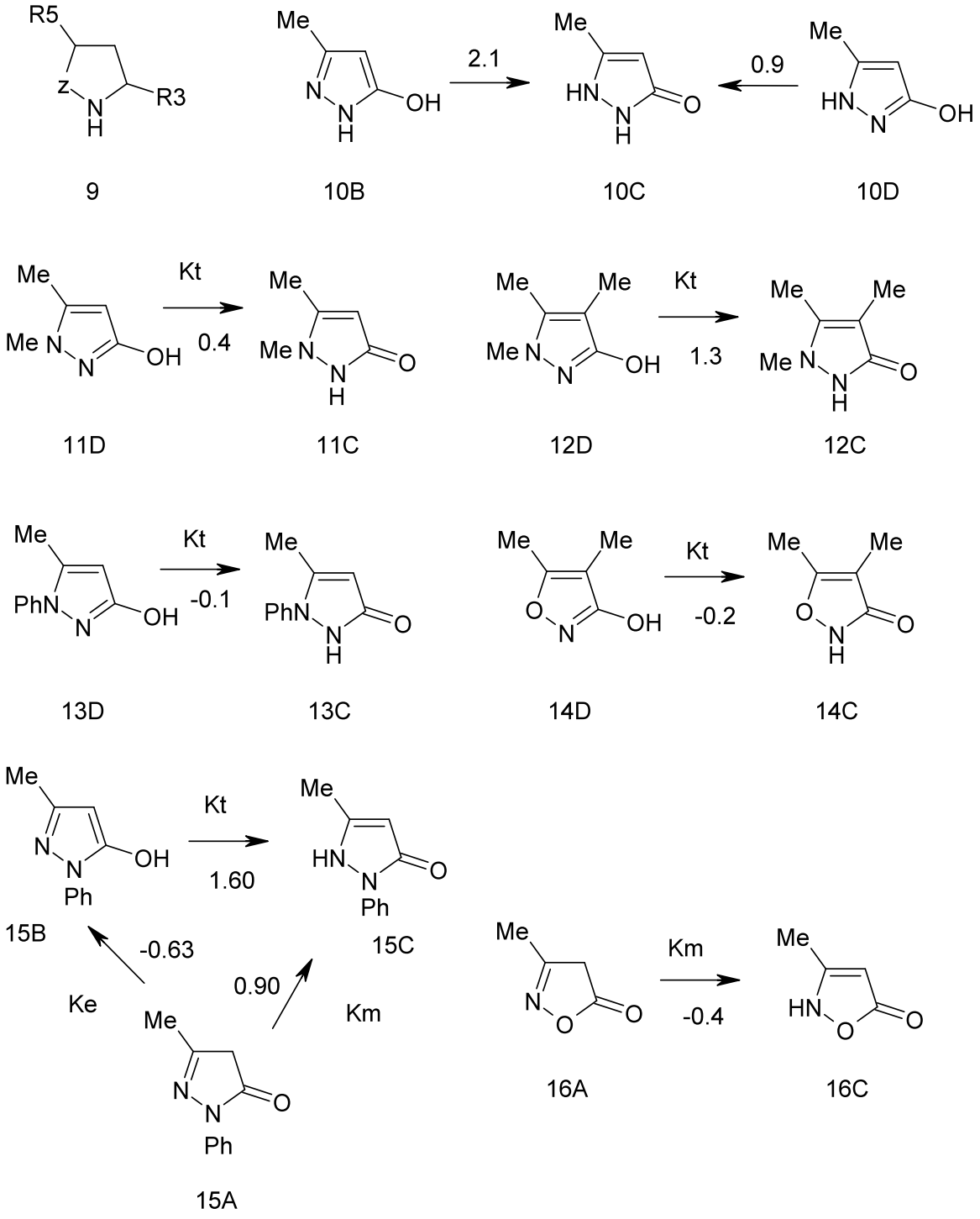
7A 7B 8A 8B

$$\log K_E = -7.39(0.22) + 1.97(1.01)\,\sigma_I + 8.11(1.25)\,\sigma_R \quad (n = 7\ r^2 = 0.977\ s = 0.14\ F = 86) \quad \text{(i)}$$

Bisacetyl (butane-2,3-dione) exists in the gas phase exclusively as the planar *trans*-conformation **7A**[8a] and an IR study[8b] failed to find any trace of the enol **7B** in solvents that range from CCl_4 to DMSO. The alkali-catalysed dehydration of its hydratr (see below) has been studied in water and its enol content stated to be < 1% (*i.e.* unobservable) but K_E was not determined.[8c] An early value[8d] of log K_E -4.25 must be discounted since in technique it predates the 'Kresge era' and several accompanying values are seriously in error. Early work on keto-enol tautomerism tends to exaggerate the enol content. Equation (i) above, based on reliable values for compounds of type $RCOCH_3$ in water,[5] would predict log K_E -4.85 but this involves a long extrapolation and the data base is skewed, so it must not be taken too seriously. By contrast, cyclohexane-1,2-dione **8** exists overwhelmingly as **8B** with an estimated[8e] <1% of **8A** present. You are invited to identify the phenomenon responsible for this difference and, if possible, quantify it.

Both compounds are heavily hydrated in aqueous solution[8c,8e] (one carbonyl becomes a diol), to the extent for of 77% for **7** and of 40% for **8**. This is an irrelevancy in the present context and one advantage of computation, which I shall be happy to see exercised, is that it can be ignored.

(5) Pyrazolones and Isoxazolones



These compounds provide some of the most convoluted puzzles in the whole field of tautomerism. In terms of the generic structure **9** they consist in two series: the 3-substituted,

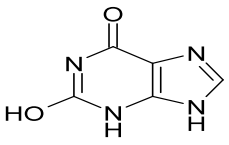
exemplified here by compounds **11** – **14**, and the 5-substituted, exemplified by **15** and **16**. For **10** with Z = NH, uniquely, the two series overlap; here the original data of Katritzky[9a] and Elguero,[9b] base on NMe and OMe model compounds, were processed using correction factors to give the log K_T values shown.[9c] On this basis, **10C** is estimated to contribute 88% to the tautomeric mixture where 90% is found. Other agreements are equally satisfactory. For **11**, Katritzky and Maine[9a] found 70% of **11C** by direct UV observation where the corrected basicity method gave an estimate[9c] of 72%; for **13**, the corresponding values[9a,9c] for **13C** are 50 and 45%. The case of the isoxazole **14** is vexing. On simple 'basicity method' assumptions Katritzky[10a] estimated a 5:1 ratio of **14D** to **14C** but believed the ratio to be much higher in reality, basing this view on its great predominance in CCl_4 and DMSO which, however, is readily explained on other grounds. In fact, our own estimations[5] lead to the slimmer margin of log K_T -0.2, a roughly 60:40 split, while near-equality in water is also suggested on UV grounds: λ_{max} 224 nm for **14** itself neatly splits the difference between 218 nm for its OMe and 230nm for its NMe model compounds.[10a] The value of log K_T 1.3 given for **12** is an isodesmic extrapolation based on the effects, elsewhere in the series, of *C*-methylation, coupled with the unquantified 'predominance' of **12C** found experimentally for its UV spectrum in water.[9b]

$\log K_E = -2.78(0.82) + 0.60(1.19)\pi^* + 0.05(0.56)\alpha + 3.61(1.06)\beta$ ($n = 9$ $r^2 = 0.80$ $s = 0.42$ $F = 6$) (ii)
$\log K_M = -3.78(0.52) + 1.79(0.54)\pi^* + 1.33(0.16)\alpha + 2.93(0.37)\beta$ ($n = 9$ $r^2 = 0.94$ $s = 0.21$ $F = 28$) (iii)
$\log K_T = -0.41(0.80) + 0.98(1.02)\pi^* + 1.29(0.44)\alpha - 1.42(0.58)\beta$ ($n = 9$ $r^2 = 0.76$ $s = 0.32$ $F = 5$) (iv)

Much less quantitatively is known concerning the 5-series and by far the most valuable work in this respect is an LSER study carried out by Freyer *et al.*[11] on **15**, which allowed the derivation[5] of the (revised) equations (ii) – (iv). Extrapolation to water then gave values for log K_E, K_M and K_T of -0.63, 0.90 and 1.60 respectively.[5] Whatever the limitations of the statistics these equations are genuinely independent since each was forced to use a different selection of the 12 solvents through absence of data – the results[11] were expressed as % component, and for any solvent in which 0% is given for one of them, two of the three possible ratios disappear. In these circumstances, the near-fit of these three ratios (see accompanying Scheme) to the requirement that $\Sigma \log K = 0$ is comforting and helps vindicate the methodology. It becomes even better if account is taken of log K_T 1.4 for the same compound by use (after correction) of the 'basicity method.'[9a] Since the **A**-tautomer does not share a cation in common with **B** or **C**, LSER is currently the only experimental technique available for complex equilibria of this sort and it would be enormously helpful if computational chemistry were able to fill the gap. Outside this set the result for **16**, obtained by direct UV observation,[10b] is almost the only quantitative datum we possess for K_M, with none at all (in water) for K_E.

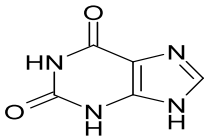
These results taken together reveal a picture very different from that of a typical amide, even when aromatic *e.g.* 2-pyridone, and comments are invited. Certain overall features are evident. One is the enormous positive effect of ring *C*-alkylation on K_T, a sure sign of severe electron-deficiency in the ring itself, and one consistent with abnormally low K_T values. The other, on similar evidence, is a trend in increasing electronegativity for Z in both forms of **9** which clearly lies in the order NH < NMe < NPh < O (the NH/NMe difference is solvational), as might be expected. For K_T, the only equilibrium that can be compared for both series, this trend is clearly steeper in the 3-(amide) than 5-(vinylogous amide) series; comments on both features, and calculations if possible, would be welcome. And it may be worth noting that the solvent dependence of K_M is exceptionally steep, posing a severe challenge to any solvational model used for that quantity.

(6) The Xanthine Saga



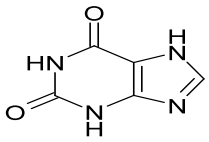
17A

60%



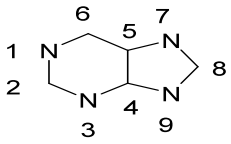
17B

33%



17F

0.55%



18

At an ACS meeting last year, Szegezdi and Csizmadia[12a] presented a paper (poster) on xanthine tautomerism of which Yvonne Martin has kindly provided me with a two-page print-out. This document lists the 15 possible tautomers of xanthine **17**, explicitly for water, and claims to do so using "pK_a based dominance conditions" which, other examples in it make clear, are an attempt to derive tautomer ratios using computed energies of neutral and their common anionic species (a legitimate variation on the 'common cation' of the 'basicity method'). The first, second and sixth most important of xanthine's tautomers according to their calculations are shown as **17A**, **17B** and **17F** above, as fractions of the total. These fractions are very difficult to read off the poster and magnification does not much help, but they are set out in order of importance and I think I have their values roughly right. To see for yourself, check http://www.chemaxon.com/conf/Tautomer_generation_A4.pdf, figure 4. There is no mention in their paper of any preceding literature.

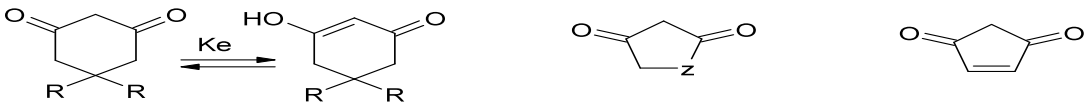
In 1971, Lichtenberg *et al*.[12b] produced the definitive paper on this subject. I show as **18** the confusing and illogical numbering system used for purines as a help in understanding what follows. Xanthine and some judiciously selected mono- and multi-N-methyl xanthines were examined by NMR in D_2O + dmso-d_6 and assignments were based on proton shifts of H(8) and the N-methyl groups. It was decisively shown that the mobile protons are found at N1, N3 and N7, which identifies **17F** as the dominant tautomer. UV spectroscopy in water confirmed this assignment and, indeed, failed to detect any other component. Pullman and Pullman[12c] calculated, at about the same time, an energetic advantage of 7 kcal mol^{-1} for **17F** over **17B**; this of course was an isolated molecule calculation and the energy gap in water is likely to be much smaller.

I shall postpone detailed discussion to the SAMPL 2009 workshop, but a few comments are in order. The tautomeric problem splits into two, one for each ring, and there is evidence which I hope to quote that appreciable interaction between them is unlikely. The tautomeric advantage of 2-C=O and 6-C=O for uracil over their respective iminols, at log K_T 5.8 and 4.6 respectively,[5] is so enormous that any species of type **17A** must be of negligible importance. And while the degenerate tautomers of the fused imidazole moiety cannot *directly* be affected by the uracil ring, the dipolar repulsion represented by the *peri*-NH groups of **17B** is likely to be horrendous: I estimate it, in this context and some others,[5] as an effective penalty of Δ log *K ca*. 2.5. On this basis I estimate *ca*. 99.7% for the fraction of **17F**, with *ca.* 0.3% for **17B** and no other tautomer worth any consideration at all.

There is one other relevant point. The 'basicity method' and its anionic equivalent both depend, in their simple form, on *one* common cation (or anion) linking *two* neutral species. Uracil possesses *three* neutral species, entailing *two* tautomeric cations and one common *dication*. Incidentally, **3** is another such example. Szegezdi and Csizmadia[12a] never had any chance of success.

Given the large number of tautomers for xanthine, participants should report relative energies for only those tautomers predicted to contribute a fraction of 1% or greater to the ensemble.

(7) Cyclic β-Diketones

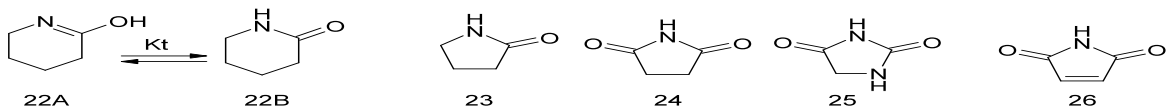


19A 19B 20 21

Z = CH_2, O, S or NR

For dimedone (**19**, R = Me), K_E is known; it is unlikely however to be appreciably different from that for (**19**, R = H) and I am happy for the latter to be calculated instead if this saves computer time. It is also known for **21**. This is an area where ring size is likely to make an appreciable difference and comparative values for (**19**, R = H) and (**20**, Z = CH_2) would be of great interest; with **21**, these form a logical sequence. The tetronic acids (**20**, Z = O) are an important class of biologically important compound; they are thought to be highly enolised but no quantitative data exist. Here and for Z = S and Z = NR it would be of some interest to know whether, and to what extent, the hetero-atom may influence K_E. In this respect I do not recommend that anyone should waste their time in trying to calculate the degree of enolisation for the carbonyl group adjacent to the hetero-atom; it will be negligible, at least by the side of the other. Incidentally the solvent dependence of dimedone has been studied by LSER and is large, being almost entirely determined by the proton acceptor ability of the solvent, so the choice of solvational model is likely to be crucial.

(8) Cyclic Lactams and Similar

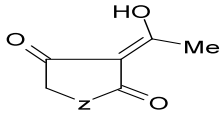


22A 22B 23 24 25 26

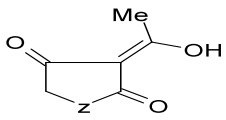
This is another area where ring size is likely to be important. No value of K_T (defined by **22**) is known, but I have guesstimates for **22** and **23** based on isodesmic extrapolation and I shall be very interested in the results of calculation. Compounds **24** and **26** are imides, a class about which

nothing is known except that the dominant tautomer remains the oxo-form, as everywhere else in this series. Compound **26** has the extra feature of unsaturation. Compound **25** is hydantoin, representative of another important class, and of special interest as a 5-membered ring, alicyclic relative of the 6-membered ring, unsaturated structure of uracil such as is present in **17** for example. All these compounds are in principle capable of supplying *transferable building blocks* of a type that should be capable of helping to build an algorithm capable of predicting tautomer preference.

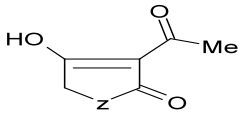
(9) Cyclic Triketones



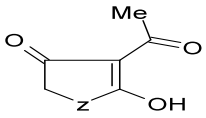
27A



27B



27C

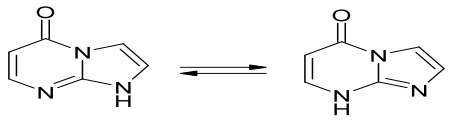


27D

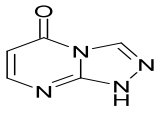
These compounds, for which Z = O, S or NH, present a puzzle in that the papers that exist on them show not only tautomeric preferences that are often bizarre, but also allege major changes in these to be triggered by bland groups such as alkyl in the (here) unsubstituted 5-position. They were investigated using ^{13}C-NMR in $CDCl_3$ and claims are made for the absence or avoidance of time-averaging which, again, are difficult to countenance. In addition, some of the assignments seem to me questionable, and the one compound of this type whose structure has been determined by X-ray diffraction has exactly the tautomer preference that would be expected of it.

An interesting feature of these compounds is that, while they are highly polar (the parent triketonic tautomer is not found), their polarity is unlikely to differ much between tautomers. Hence it may be that, for once, *differential* solvation will be unimportant and a computational approach based largely on internal energy differences will prove successful.

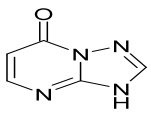
(10) Nitrogen-nitrogen Tautomerism in Compounds with Bridgehead Nitrogen



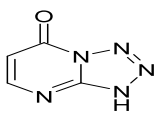
28A ⇌ 28B



29A



30A

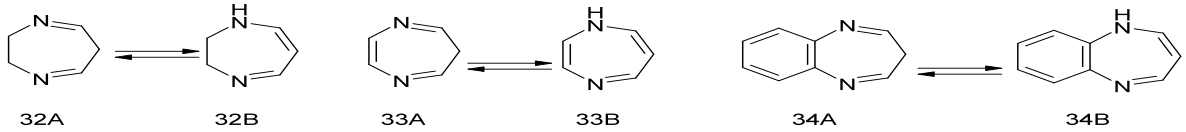


31A

Annular tautomerism in 5,5-ring fused heterocycles has been extensively studied by Elguero and his school and this is a variant on that theme in that one of the rings has become 6-membered

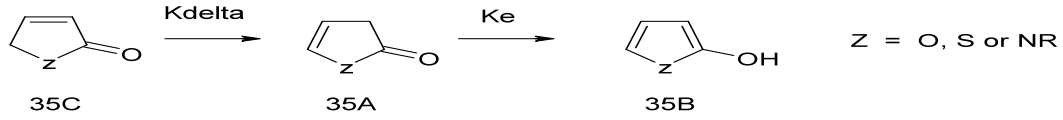
by insertion of carbonyl. It differs from the originals in three respects. Firstly, this type of structure is of some prominence in medicinal chemistry. Secondly, quantitative results in aqueous solution are known for one of these pairs.[8b] Thirdly, there is some reason to believe that a general principle exists which may be used to rationalise *quantitatively* the results, if known, for this series of compounds,[5] and could be extended without difficulty to those studied, and provisionally rationalised, by Elguero. The overall similarity of these four compound pairs raises the hope that computation will be able to place this series in an order that, at the least, can *qualitatively* rationalise the physical principles involved.

(11) Diazepines



This series explores another aspect of N,N-tautomerism and the effect of unsaturation and/or aromatisation on it. Nitrogen's extra valency by the side of oxygen and sulfur allows it to appear in endocyclic contexts not open to them, and these compounds may be considered as distant relatives of the β-diketones, which is their chief interest here. The preferred tautomer is known for **32** and **34** but **33** does not appear to have been examined.

(12) Five Membered Ring 2-Oxoheterocycles



While tautomerism in the corresponding 3-oxo series has been quantified, no attempt has been made to quantify either step of this twofold process. Possibly this stems from the sheer difficulty of disentangling them, which since **35C** is the dominant tautomer for all three parent compounds, is severe – I have to give this fact away as it is crucial to understanding the point of the investigation.

Although values for K_{delta} are unknown, there are empirical ways of getting at these and a correlation has been found with an independent set of parameters[5] which appears, tentatively, to put the quantities so derived on a sound basis. Nevertheless, a check is needed. It may help computation that all present evidence points to a minimal degree of solvent sensitivity for this equilibrium, as suggested for **27** above, and probably results from the same reason: while **35A** and **35C** are both polar species, there is not likely to be much difference betweeen them in this respect.

The second step involves aromatisation, but this is known to give a much lower degree of stabilisation than is found for the related 6-membered rings and it also involves loss of the resonance energy attaching to ZC=O. Empirically, substitution of electronegative groups or those capable of forming an intramolecular hydrogen bond – acyl combines both features – appears necessary for generation of **35B** in polar solvents. However, as the least polar of these three species the enol form is sometimes visible in nonpolar solvents even when the above criterion is not satisfied. Hence K_E is a very solvent-sensitive process and here, as mostly elsewhere, a good solvational model will be essential.