J Comput Aided Mol Des (2010) 24:575–586
DOI 10.1007/s10822-010-9345-5

Annular tautomerism: experimental observations and quantum mechanics calculations

Aurora J. Cruz-Cabeza · Adrian Schreyer · William R. Pitt

Received: 31 January 2010 / Accepted: 17 March 2010 / Published online: 3 April 2010

Abstract The use of MP2 level quantum mechanical (QM) calculations on isolated heteroaromatic ring systems for the prediction of the tautomeric propensities of whole molecules in a crystalline environment was examined. A Polarisable Continuum Model was used in the calculations to account for environment effects on the tautomeric relative stabilities. The calculated relative energies of tautomers were compared to relative abundances within the Cambridge Structural Database (CSD) and the Protein Data Bank (PDB). The work was focussed on 84 annular tautomeric forms of 34 common ring systems. Good agreement was found between the calculations and the experimental data even if the quantity of these data was limited in many cases. The QM results were compared to those produced by much faster semiempirical calculations. In a search for other sources of the useful experimental data, the relative numbers of known compounds in which prototropic positions were often substituted by heavy atoms were also analysed. A scheme which groups all annular tautomeric transformations into 10 classes was developed. The scheme was designed to encompass a comprehensive set of known and theoretically possible tautomeric ring systems generated as part of a previous study. General trends across analogous ring systems were detected as a result. The calculations and statistics collected on crystallographic data as well as the general trends observed should be useful for the better modelling of annular tautomerism in the applications such as computer-aided drug design, small molecule crystal structure prediction, the naming of compounds and the interpretation of protein—small molecule crystal structures.

Keywords Quantum mechanics · Semiempirical · Crystallographic data · Chemical nomenclature · Molecular modelling · Chemical structure enumeration · Tautomers · Tautomeric equilibria

A. J. Cruz-Cabeza
The Pfizer Institute for Pharmaceutical Materials Science, The Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK

A. Schreyer · W. R. Pitt
Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1GA, UK

W. R. Pitt (✉)
Department of Medicinal Chemistry, UCB Celltech, 216 Bath Road, Slough SL1 3WE, UK
e-mail: will.pitt@ucb.com

Introduction

The study of tautomerism, either using experimental techniques or computational modelling, has been carried out in great detail on many model systems. This has been an extremely valuable process contributed to by very many workers. An extensive review of their results for prototropic tautomerism of heteroaromatic rings was published in 1963 by Katritzky and Lagowski [1–4] and updated recently by Katritzky and co-workers [5–9]. It is clearly a highly involved subject with many complicating factors. For instance, substituents and environmental factors can affect tautomeric propensities [8, 10]. However, the approach often adopted when applying quantum mechanics (QM) techniques to the problem is to put most of these factors to one side. Calculations are usually performed on isolated fragments, such as heteroaromatic ring systems, in

vacuo. Elguero et al. [8] comment that, within this restricted domain, these sorts of calculations can now exceed the accuracy and precision of most experiments. However, they also highlight the problem of comparing results from different papers in the absence of a standard level of theory. With a desktop computer, a MP2 level QM calculation with a standard basis set can now be carried out on a molecular system of the size of a monocyclic or bicyclic heteroaromatic ring system in a matter of hours. Semi-empirical calculations on the same sorts of molecules take only seconds to run and have also been used for predicting tautomeric equilibria [8]. Although there are some studies in which the theoretical stability and aromaticity of a considerable number of tautomeric families are compared [11–13], results have not been published for an extensive set of inflexible heterocycles with a high occurrence in experimental databases.

In the paper entitled "Let's not forget tautomers" [10], Martin emphasises the need for more experimental data, for cheminformatics databases to store information about tautomers, and for the validation of computational methods. Here, the Cambridge Structural Database (CSD) [14] and the Protein Data Bank (PDB) [15] are explored as sources of data on annular tautomerism. Statistics on chemical product availability are also examined to see whether they can, perhaps against expectation, provide indirect evidence of tautomeric propensities. The focus of this paper is on annular tautomerism, a type of prototropic tautomerism found in aromatic ring systems. All results are compared to those obtained by MP2/6-311++G** level QM calculations, on the assumption that if experimental results and calculations agree in most cases, at least qualitatively, the usefulness of both types of data is confirmed. The approach is similar to that taken by Hao et al. in their paper comparing QM calculated torsional energy barriers of chemical functional groups with those found in the PDB [16]. The results, although by no means comprehensive, could provide the basis for a database of knowledge on annular tautomeric propensities. Previous work on a comprehensive enumeration of heteroaromatic ring systems [17] is also exploited here for an examination and classification of the universe of annular tautomers. A logical extension of collecting ring system substructures together independent of the context in which they are found (substituents attached, and interactions formed etc.) is to further group them together by the tautomeric substructures they contain. In this way general trends can be identified. The results produced in this paper should be very useful to those modelling, predicting interactions and cataloguing small molecules containing potentially tautomeric heteroaromatic ring systems.

Methods

The comprehensive set of annular tautomers and subset studied in detail

The VEHICLe database [17] provided all the structures of ring systems studied in this paper. This database was designed to contain all possible heteroaromatic mono and bicyclic ring systems, within certain restrictions. All entries are neutral, contain only carbon, nitrogen, oxygen or sulfur, and the only exocyclic functionality considered are carbonyls. Tautomers were treated as separate ring systems. Importantly for this paper, only annular tautomers are present. For instance, lactam–lactim transformations were not studied and nor are they here.

Tautomeric pairings were identified using Pipeline Pilot (Accelrys Inc.). Groupings of tautomeric equivalents are referred to as clusters. There are 3288 tautomeric ring structures which fall into 1544 (mistakenly quoted as 772 by Pitt et al. 2009) clusters present in VEHICLe. Substructure searches for each ring system were carried out as part of the original work in databases of known molecules (referred to as the training dataset in that paper). Only 558 (out of a possible 3288) different ring systems were found. Below, the focus is on the most commonly found ring systems and their tautomers. This set was collated using the following procedure. Ring systems were identified that had at least 100 substructure hits within a known compound dataset. Added to these ring systems were all their tautomeric equivalents. When this was done, 84 ring systems in 34 clusters were collected. These ring systems are shown in Fig. 1 and will henceforth be referred to as the standard set.

The chemical product dataset (MCD & LCD)

It has been recognised that substituted tautomer derivatives where the Kekulé bond order is effectively fixed can be used as models of otherwise transient tautomeric structure [18]. Here derivative compounds are used for a different purpose. The relative prevalence of compounds containing different tautomeric ring systems within a cluster are compared with predictions of the most abundant tautomer. There are many reasons why these two values may not agree. For instance, ring systems may be built up using reagents that preincorporate substituents. When alkylating an aromatic ring, reactivity may be a stronger influence or independent of tautomeric equilibria [19]. Reaction conditions, such as the solvent and temperature, can also affect tautomeric equilibria [8]. Nevertheless, compound structure is such an abundant and diverse source of experimental data that it might be helpful for those wishing to predict the most abundant tautomeric alternative.

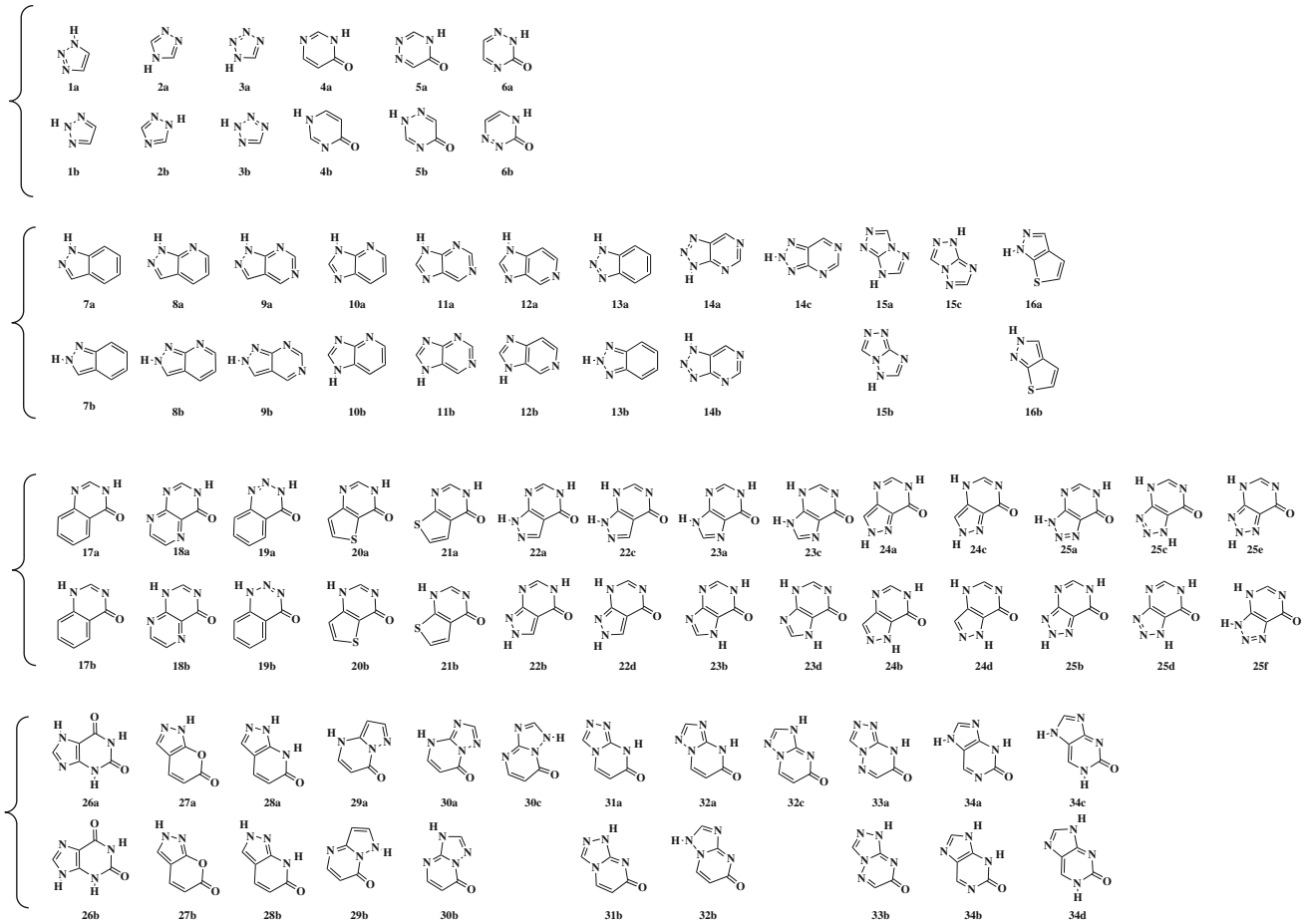

Fig. 1 Heteroaromatic ring substructures used in this study grouped in four main classes: (1) monocycles (**1–6**), (2) bicyclic derivatives of diazine and triazine (**7–16**), (3) bicyclic derivatives of the **4**, **5** and **6** monocycles (**17–25**) and (4) the remaining bicyclic substructures

If all prototropic hydrogens are substituted with heavy atom substituents the tautomer is referred to as 'locked', or 'partially locked' if not all tautomeric alternatives are fixed (see Fig. 2). We used the results of the searches done within a set of about 3 million commercial compounds, patented examples and literature structures published previously [17]. This dataset is a mixture of locked, partially locked and unlocked examples adding further noise to the data. In order to obtain a cleaner set of data, a fresh search for only locked examples was carried out within approximately 5 million unique commercial compounds. The database used for this was a proprietary relational system called IDAC (Philip Ashworth, unpublished). Substructure queries were designed so that each available hydrogen bonded to a ring nitrogen was substituted by a non-hydrogen atom and no extra aromatic rings or carbonyls were allowed. These queries were created automatically using Pipeline Pilot for each of the ring systems in the standard set. The 3 and the 5 million compound datasets are referred to as the mixed (MCD) and locked compound dataset (LCD) respectively.

Search in Cambridge Structural Database (CSD)

Initial subsets of crystal structures containing any of the 84 standard ring system substructures were retrieved from the CSD (version 5.31 2009 + November update) using the software Conquest [20]. This data was then analysed and filtered using Pipeline Pilot protocols. The data given in this paper passed the following structural filters: (1) tautomeric substructures must not be charged, (2) structures must contain no metals and (3) the ring substructures must be isolated (i.e. they cannot be a substructure of a bigger aromatic ring system).

Tautomers in the Protein Data Bank (PDB)

A further method of analysing tautomer propensities is to look at the occurrences of tautomers in structurally characterised protein–ligand complexes. The resolution of protein crystal structures is rarely sufficient to observe hydrogens; it is, however, possible to infer the tautomeric form by looking at the distribution of amino acid atoms

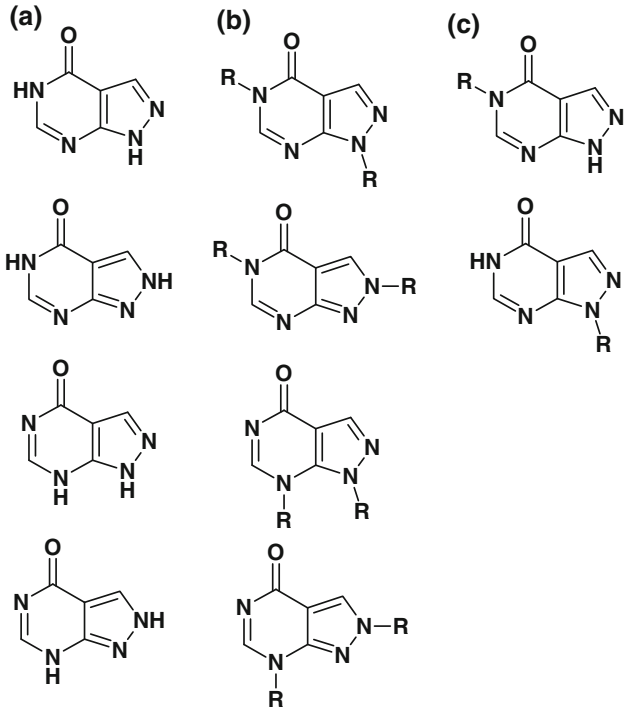

Fig. 2 **a** 'unlocked' tautomeric alternatives **b** 'locked' substructures **c** 'partially locked examples'. R ≠ H

surrounding a tautomer fragment in a protein binding pocket. For most amino acid residue atoms the protonation state is known, the principal exceptions being isoelectronic atoms such as the imidazole side chain of histidine.

SMARTS [21] queries were utilised to identify two common tautomer classes within small molecule ligands in the PDB. A pyrazole-like fragment (SMARTS = [#7;D2]1[#6][#6;x3][#6;x3][#7;D2]1) and a pyrimidone-like fragment (SMARTS = [#7;D2]1[#6](=O)[#6;x3] [#6;x3] [#7;D2] [#6;D2]1) were selected in part because of their asymmetry. Unambiguous superimposition of matching ligands would not be straightforward for various other tautomeric classes, for instance those that contain an imidazole-like fragment. The pyrazole-like fragment matches clusters **7, 8, 9, 16, 22, 24, 27,** and **28** in Fig. 1, the pyrimidone-like **17, 18, 20, 21, 22, 23, 24,** and **25**. The two queries were used for substructure pattern matching against all PDB chemical components with the help of the OEChem toolkit (OpenEye Scientific Software). For each matching chemical component, the chemical component atom names, which are unique and in the same order as the atoms in the query molecule, were recorded. Bound ligands, i.e. occurrences of each chemical component in protein–ligand complexes, were fetched from the CREDO protein–ligand interaction database [22] including all interacting residues within 6 Å. The ligand atoms matching the fragment were labelled with their match index to preserve their proper order. In addition, the Structural Classification Of Proteins (SCOP) [23] family identifier is stored for the proteins that the ligands are in contact with in order to enable grouping the fragments, including the binding site environment, by protein family. Only ligands with all heavy atom positions determined and with no clashes with other atoms were considered. After all occurrences for a fragment were found in the PDB and the tautomer substructure labelled, ligands as well as the surrounding amino acid residues, were superimposed on the tautomeric ring using the OEChem toolkit. Match pairs were created using the (renamed) atom names of the tautomer substructure and subsequently the whole structure rotated and translated for best fit. The results were visualised in PyMOL (DeLano Scientific) with all amino acid hydrogen bond donor (N, NE1, NE, NH1, NH2, ND2, NE2, NZ) and acceptor atoms (O, SD, OE1, OE2, OD1, OD2, ND1) within 3.6 Å of the fragment nitrogen atoms being shown. Amino acid residues were protonated with the OEChem toolkit.

Tautomer classification

All the tautomeric transformations were divided into groups of paired substructures. These substructures were sketched in ISIS/Draw (Symyx Technologies), exported as MDL mol files [24] and iteratively refined until all but 2 ring systems were matched and correctly classified. Matches were only recorded if a tautomer cluster contained ring systems that match both substructures or one substructure matched a differing set of atoms. This procedure required numbering atoms consistently within a cluster. This step was achieved by alignment to a version of the ring systems with undefined bond orders and no hydrogens. All this was carried out in Pipeline Pilot.

Quantum mechanics calculations

Initial molecular geometries with implicit hydrogen atoms were generated from SMILES [25] strings using the "molecule from smiles" component in Pipeline Pilot. The substructures in Fig. 1 were treated as such and all the substituents were defined as hydrogen atoms. Each of the 84 ring systems were then geometry optimised in gas phase using the Möller-Plesset MP2 method and the 6-311++G** basis set with the program GAUSSIAN03. Single point energy calculations using a Polarisable Continuum Model (PCM model), and the MP2/6-311++G** level of theory, were then performed on the gas-phase optimised molecular geometries. Three different values for dielectric constants (ε) were compared (1) $\varepsilon = 1$ (vacuum), (2) $\varepsilon = 3$ (average dielectric in organic crystals) and (3) $\varepsilon = 80$ (water dielectric). The longest geometry optimisation took 19 h of computer time, whereas the single-point calculations with the PCM model very rarely took over 1 h of computer time on a single 2.4 GHz CPU desktop machine.

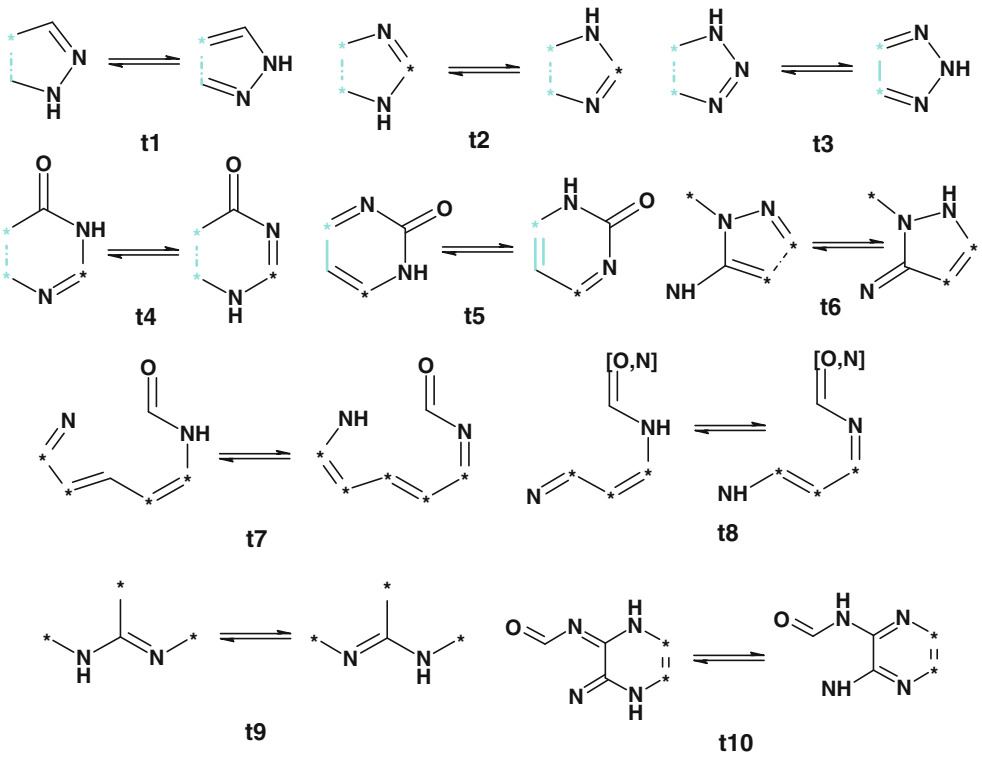

Fig. 3 Substructures representing the types of tautomeric transformations found in the VEHICLe database. Transformations t1–t5 can occur in monocycles or bicycles, cyan bonds can be single or double and indicate where ring fusion is allowed; t6–t10 only occur in bicycles. The * indicates carbon or nitrogen. The variation that is drawn on the left of the equilibrium arrows is alternative 1 and on the right is alternative 2, although often the choice is arbitrary due to symmetry

Semi-empirical calculations

Molecules were generated from SMILES strings using the "molecule from smiles" component in Pipeline Pilot. Hydrogen atoms where added to the structures and 3D coordinates were generated. The semi-empirical enthalpies of formation were calculated using the "Semi empirical QM descriptors" component in Pipeline Pilot. This component uses the VAMP semi-empirical molecular orbital code. The enthalpies of formation were calculated using the MNDO and AM1 methods in vacuum and in water. All calculations for the 84 ring systems were completed in 6 min.

Results

All possible annular tautomerisms

There are 3288 annular tautomers in the VEHICLe database which fall into 1544 clusters. As noted by Martin [10], consideration of all possible tautomers is not necessarily a time or disk space consuming task. The vast majority of clusters (2762) contain only 2 members, with 399, 100, 15 and 12 clusters of size 3, 4, 5 and 6 respectively. Clusters with more than 2 members occur when pairs of tautomeric alternatives combine. These pairs were classified into 10 classes t1–t10 (see Fig. 3). Only classes t2, t4, t9 and t10 are of the type $HX-Y=Z \rightleftarrows X=Y-ZH$ which the International Union of Pure and Applied Chemistry (IUPAC) Gold Book defines as tautomerism. Classes t1 and t3 are of the form $W-X(H)-Y=Z \rightleftarrows W=X-Y(H)-Z$; classes t5, t6 and t8 are $HV-W=X-Y=Z \rightleftarrows V=W-X=Y-ZH$; class t7 is $HT-U=V-W=X-Y=Z \rightleftarrows T=U-V=W-X=Y-ZH$.

Extraction of data from the CSD

The CSD contains crystallographic information on over 500,000 small molecule crystal structures determined by either neutron or X-ray diffraction techniques. Determination of the hydrogen atom positions is possible with high quality neutron diffraction data. With X-ray data, however, erroneous location of hydrogen atom positions in tautomers is not unknown [26, 27]. For an experienced crystallographer, a careful analysis of heavy atom bond distances and angles should point to the exact location of hydrogen atoms in the majority of the crystal structures without disorder. In our dataset of CSD crystal structures containing tautomers, we acknowledge that a small proportion of hydrogen atom positions could have been wrongly positioned.

The number of hits in the CSD is considerably lower than those found in the compound datasets but provides a reliable source of information on tautomeric states. Small molecule crystal structures that matched one of the substructure queries are either locked or partially locked (CSD_R) or show genuine tautomeric potential (CSD_H; see Fig. 2). A histogram of the occurrences in the CSD is given

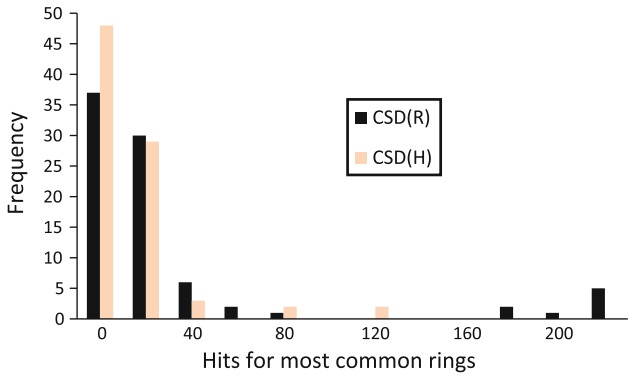

Fig. 4 Histogram representing the frequency of occurrence of the 84 standard ring systems in the CSD

in Fig. 4. Over all the standard ring systems, and after filtering, the total number of matching structures were 2468 and 583 for CSD_R and CSD_H respectively. CSD_R structures are more abundant than CSD_H (see Fig. 4). There are 37 and 48 substructures, for CSD_R and CSD_H respectively, for which there are no hits in the CSD. Those missing were mostly bicyclic substructures shown in the bottom half of Fig. 1. Substructures with the most hits in the CSD correspond to ring systems **11a**, **2b**, **1a**, **2a**, **13a**, **3a**, **23a** and **17a**.

Extraction of data from the PDB

Tautomeric representatives of ring system **7**, **8** and **22** could be identified for the pyrazole-like fragment and **23** and **24** for the pyrimidone-like in the PDB. Remaining hits were tricyclic (or larger) ring systems not explicitly studied in this work. The pyrazole-like and pyrimidone-like fragments matched 58 and 17 different small molecules respectively. Of the 136 and 157 PDB structures with matching ligands, 86 and 79 were for unique proteins. The two substructure queries matched ligands in contact with 35 and 29 unique chains (in terms of UniProt [28] accession number) and 6 and 10 unique SCOP families. The position of potentially prototropic hydrogens and hence the identity of the tautomeric alternative bound to protein was manually inferred by the location of neighbouring hydrogen bonding groups. In the case of the pyrazole-like fragment, all examples in the standard set were found to be in tautomeric form with the lowest energy as calculated by the QM calculations (t1, form 1 in Fig. 3). Interestingly, there were 4 ligands (2QHN-582, 3COH-83H, 2B55-D31) in which t1, form 2 was found. However, in each case the ring system involved was made up of 3 or more rings. Although more ligands were found for the pyrimidone-like fragment, the distribution of amino acid atoms resulting from the alignment of ligands was sparse and contained many acidic amino acid side chains. However, as can be seen in Fig. 5b, virtually all hydrogen bond acceptor groups were clustered around the lactam nitrogen indicating that t4 form 1 predominated. There were no cases where t4 form 2 could be unambiguously assigned from neighbouring protein hydrogen bonding groups.

Chemical compound datasets

Searches of the MCD retrieved representatives for 157 clusters out of a possible 1544, with 40 clusters having more than 50 unique molecules containing at least one ring system (the most abundant tautomer). There were usually far fewer example compounds containing the second most abundant tautomer. For three quarters of these 40 clusters there were at least 10 fold less (see Fig. 6). Similar results were found for the locked compound dataset but more

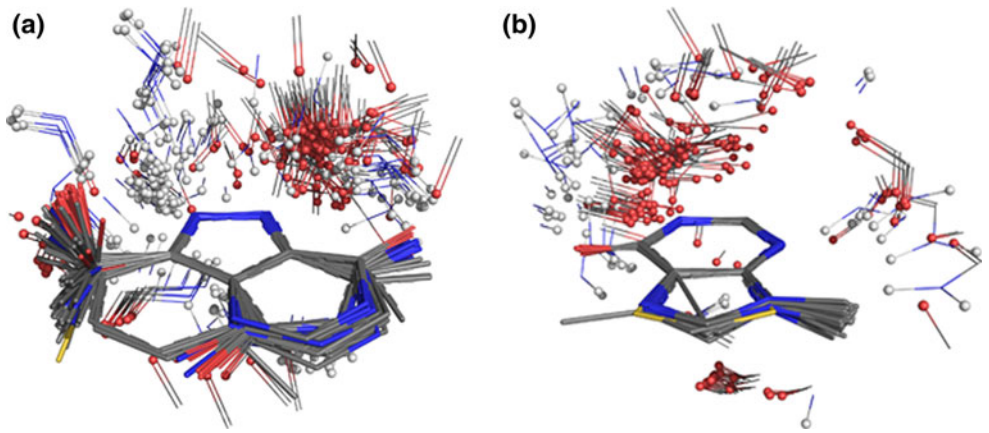

Fig. 5 Hydrogen bond donor and acceptor distribution within 3.6 Å of two fragments in the PDB. Hydrogens attached to donor atoms as well as acceptors are shown as spheres. The donor/acceptor oxygen atoms of the carboxylic acids in Glu and Asp are coloured in a dark shade of red. Created with PyMOL **a** Overlay of the indazole-like fragment (extended by two bonds) of 136 ligands including binding site atoms **b** Overlay of the pyrimidone-like fragment found on 157 ligands in the PDB

clusters were found with a highly pronounced relative abundance with 8 clusters having a 1000 fold difference.

Quantum mechanics calculations

Whilst most theoretical calculations are performed in gas-phase, aqueous media are more relevant to those studying biological systems. Although we do not account for complicating factors such as substituent effects in the calculations, we have accounted for environment effects in the theoretical stability of tautomers by using a Polarisable Continuum Model (PCM) in the QM calculations. The polarity of the environment can considerably affect the stability of tautomers and gas-phase calculations are far from the reality of most biological and physical systems.

In Fig. 7 we represent a histogram of the frequency (N) of the energy differences between the two most stable tautomers (ΔE_t) for each of the 34 clusters. The histogram shows the results calculated in vacuo ($\varepsilon = 1$), in a crystal environment ($\varepsilon = 3$) and in water ($\varepsilon = 80$). At higher dielectric constants the relative energy of the tautomers become smaller and can, in some cases, even reverse the order of stability. For example, **3b** is more stable than **3a** in gas phase (by $\sim$6 kJ/mol) but as soon as solvent effects are accounted for in the calculations, **3a** becomes more stable than **3b** by $\sim$4 kJ/mol in a crystal environment, and by $\sim$14 kJ/mol in water. In water, ΔE_t is always less than 40 kJ/mol. In fact, 32 of the 34 most stable pairs of tautomers have a $\Delta E_t < 30$ kJ/mol, the energy of a strong hydrogen bond. From now on, we will refer to the ΔE_t with calculated with $\varepsilon = 80$, because water is the most common solvent, the ΔE_t is generally smaller and is a more realistic model than "ideal" gas-phase calculations.

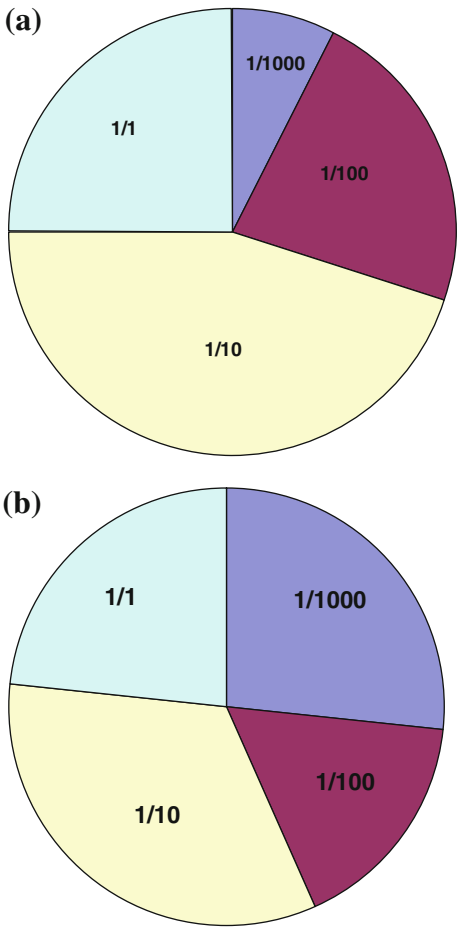

Fig. 6 Ratio of 2nd/1st most abundant tautomer within the compound datasets for the 40 tautomer clusters with at least 50 examples containing the most abundant tautomer. Where no examples were found, 0.1 was used for ratio calculation **a** mixed compound dataset (MCD) and **b** locked compound dataset (LCD)

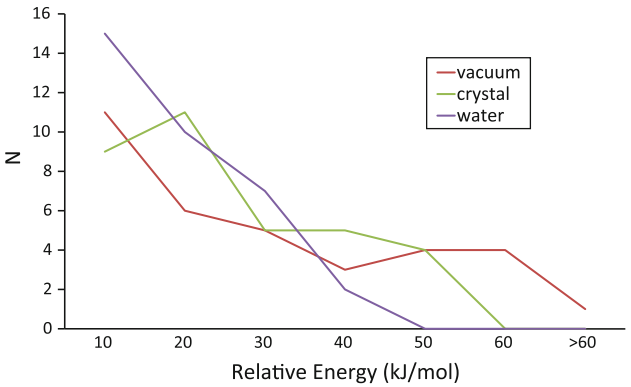

Fig. 7 Frequencies (N) of the two most stable tautomers relative energies (MP2/6-311++G** + PCM model, ΔE_t) per cluster in three different environments: (1) vacuum, (2) an organic crystal environment and (3) water

Semi-empirical calculations

Figure 8 shows the relative energies produced by the semi-empirical method compared to those from the MP2 calculations. There is some qualitative agreement for some ring systems which share certain similarities. The stable indazole-like tautomers (**7–9**) and the pyrimidone (**4**) and its analogues when there is no other tautomeric competition (**17–21**) are, at least, predicted qualitatively. Alkorta and Elguero have also compared QM calculations with semi-empirical calculations in indazoles and they concluded that the semi-empirical methods could be used as "an exploratory tool" for this family of molecules [29]. For more challenging tautomers, the quality of the semi-empirical methods is clearly insufficient.

Comparisons between experimental and calculated results

Table 1 is a summary of the data collected from compound and crystallographic databases and the calculated relative

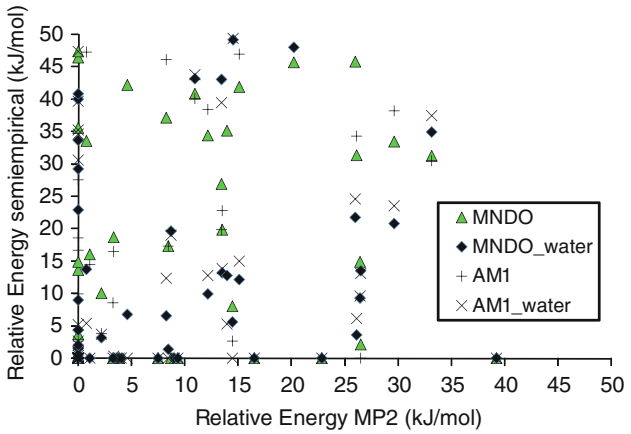

Fig. 8 Lack of correlation between the relative energies of the tautomers calculated with MP2 and semi-empirical methods

energies in water (ΔE_t) compared to the lowest energy tautomer per cluster. Overall, there is a good agreement between the relative number of experimental observations and the calculated ΔE_t. The results for 'unlocked' ring systems in the CSD (CSD_H in Table 1) are the cleanest dataset, providing the best experimental evidence for the most abundant tautomeric form of each ring system. There are 24 clusters out of the 34 in our standard set which have 'unlocked' structures in the CSD, and in 22 (92%) of these cases the lowest energy tautomer is also the one most usually observed. However, in 15 out of 24 clusters the number of different structures in the CSD is in single figures. When using results derived from the mixed compound dataset and the 'locked' commercial compound dataset, in around 70% of cases the most abundant tautomeric substructure matched the lowest energy tautomer. These data are larger in number and are available for more clusters but are not reliable as a source of information on the most abundant tautomer.

Figure 9 shows the agreement between the MP2 derived prediction of the most abundant tautomer and the most abundant tautomer found in the three databases sources, ordered by ΔE_t. It can be seen from this plot that the two clusters where QM and CSD observations contradict each other have low ΔE_t values (**10** = 1.07; **1** = 4.0 kJ/mol). While numbers are low, this result is consistent with the hypothesis that if the energy difference between tautomers is small, crystal packing forces and hydrogen bond formation are the dominant influences. The effect of substituent, other environmental factor and the quality of the QM calculations might also affect tautomeric preference in these cases. When comparing the MP2 predicted order of stability to the compound databases findings, discrepancies for clusters with higher ΔE_t (>10 kJ/mol) values are found. This can be explained by the dominance of one or more synthetic route for which tautomeric equilibria are an irrelevance or subordinate to reactivity.

Pairs of tautomers with a large energy difference, $\Delta E_t > 20$ kJ/mol (**6**, **29**, **19**, **7**, **8**, **9** and **21**), show clear experimental preferences towards the most stable tautomer in all databases but this could be partially coincidental. This can be observed more clearly in Fig. 9 where ΔE_t is plotted together with database occurrences of the different ring systems (in %) for the most populated dataset (MCD).

Comparison of results by tautomer classes

When comparing tautomer preferences within classes, we also see some patterns. For example, **7**, **8** and **9** include indazole and indazole analogues (class t1) in which changes in the second aromatic ring are introduced by substituting *C* by *N*. These changes in structure do not affect the tautomeric energy difference or experimental observations: tautomer **a** is always favoured in all three cases by over 26 kJ/mol. The results are consistent with previous studies in which indazole in its tautomeric form **a** always dominate independently on the solvent, and substitution, due to the low aromaticity of the **b** tautomer [5].

The third section of tautomers in Fig. 1 (**17–25**) and its common substructure in tautomer 4, also show commonalities. In all cases, the MP2 calculations predict the **a** tautomers (O = C–NH) favoured over the **b** tautomers (O = C–N = C–NH), transformation t4, independently of the presence and nature of the second ring. The occurrences of tautomers **a** (almost always >90%) in the databases are also noticeably favoured over the tautomers **b** of the same kind. It is not unknown for tautomers **a** and **b** to coexist within a single crystal structure where multiple hydrogen bonds are formed (e.g. see Fig. 10).

Table 2 shows the frequency of ring systems by class and alternative form (see Fig. 3) within the complete set of 3288. All but 2 ring systems contain at least one class of tautomerisation. The exceptions are 6,6 bicyclic rings containing 8 nitrogens and 2 carbons, which are probably not stable in any case. As can be seen, the pyrimidone class t4 is present in at least twice as many ring systems as any other class. However, within the known ring systems found within the mixed compound dataset, no one class dominates. Class t10 is special in that it involves the simultaneous relocation of 2 hydrogens and can only occur in 12 ring systems (6 tautomer clusters), 2 of which are known. In most cases neither tautomeric form is predominant overall. This is unsurprising for t2, t3 and t9 which are symmetrical about a central axis. Class t1 (pyrazole-like) shows a hint of form 1 being most commonly predominant but only t4 (pyrimidone-like) shows a clear bias in both database hits and energy calculations. It is quite possible that such biases could exist for classes t6, t7, t8 and t10 but not enough examples are known or energy calculations carried out thus far.

Table 1 The number of hits for the lowest energy ring system (LERS) and its tautomeric ring systems (RS) in the different databases and their relative energy (MP2/6-311++G**, PCM model $\varepsilon = 80$, ΔE_t) in kJ/mol

LERS	MCD	LCD	CSD_R	CSD_H	PDB	RS	MCD	LCD	CSD_R	CSD_H	PDB	ΔE_t
1b	770	1877	53	11	5	1a	7925	15173	249	27	23	4.00
2b	13044	25440	345	116	12	2a	148082	261727	242	7	8	7.47
3a	33901	56596	184	103	11	3b	10308	12627	73	10	6	13.44
4a	8412	8959	58	80	3926	4b	2118	1106	4	26	292	8.25
5b	202	18	3	4	5	5a	2310	739	19	0	15	9.32
6a	390	158	1	1	0	6b	4	9	1	0	15	39.19
7a	2653	2334	30	31	52	7b	146	127	28	0	0	26.46
8a	1892	13204	26	3	6	8b	25	393	0	0	0	26.43
9a	6537	22735	32	10	0	9b	5	93	12	3	4	26.06
10a	2238	5655	11	3	0	10b	179	70	4	4	3	1.07
11a	4287	2816	379	62	145	11b	307	63	16	13	98	3.31
12b	19	1422	6	1	2	12a	126	55	0	0	3	3.81
13a	8744	11066	226	11	3	13b	6443	6811	27	0	0	20.17
14a	2260	6711	1	0	17	14c	5	5	0	0	0	14.49
						14b	7	0	0	0	2	26.87
15c	6	2	0	0	0	15a	244	149	0	0	0	16.46
						15b	32	0	0	0	0	52.38
16b	0	0	0	0	0	16a	2879	4624	0	0	0	0.74
17a	66445	45524	168	14	33	17b	372	82	1	1	37	12.10
18a	85	421	4	4	1391	18b	30	1	1	0	33	13.48
19a	1816	3088	5	1	0	19b	4	0	0	0	0	29.58
20a	1730	4585	0	0	4	20b	9	0	0	0	0	15.09
21a	13377	23377	5	4	9	21b	13	1	0	0	0	25.96
22a	3094	8768	24	1	0	22b	86	22	1	0	0	8.97
						22d	2	0	0	0	0	28.73
						22c	9	0	0	0	0	33.12
23a	671	12	173	5	40	23b	73	6	9	2	63	2.14
						23d	20	1	3	1	21	20.85
						23c	181	0	0	0	11	34.79
24b	145	66	20	3	2	24a	10	346	2	0	0	3.26
						24d	22	0	0	0	0	17.10
						24c	0	0	0	0	0	23.69
25a	1338	2868	8	6	17	25b	5	0	1	1	0	10.9
						25c	3	0	0	0	0	16.71
						25d	3	0	1	0	0	33.20
						25f	0	0	0	0	17	33.52
						25e	0	0	0	0	0	38.82
26a	31888	19636	0	0	28	26b	434	610	0	0	27	8.39
27b	2	0	0	0	0	27a	186	100	1	0	0	13.96
28b	8	0	0	0	0	28a	226	1808	2	0	0	4.59
29a	486	374	2	6	0	29b	3	2	0	0	0	33.11
30a	471	493	1	4	0	30b	58	9	0	0	0	14.41
						30c	9	339	0	1	0	54.92
31a	216	7	1	0	0	31b	11	10	0	0	0	8.70
32a	496	3	0	3	0	32c	0	0	0	0	0	22.82
						32b	2	0	0	0	0	70.08
33b	16	13	0	0	0	33a	169	56	0	0	0	6.69

Table 1 continued

LERS	MCD	LCD	CSD_R	CSD_H	PDB	RS	MCD	LCD	CSD_R	CSD_H	PDB	ΔE_t
34d	4	0	5	0	11	34c	4	0	0	0	21	11.05
						34a	1123	0	0	0	21	11.72
						34b	16	1	0	0	12	18.77

The databases are: (1) Mixed compound dataset (MCD), (2) Locked compound dataset (LCD), (3) CSD_R, (4) CSD_H and (5) PDB (results of SMILES based substructure searches)

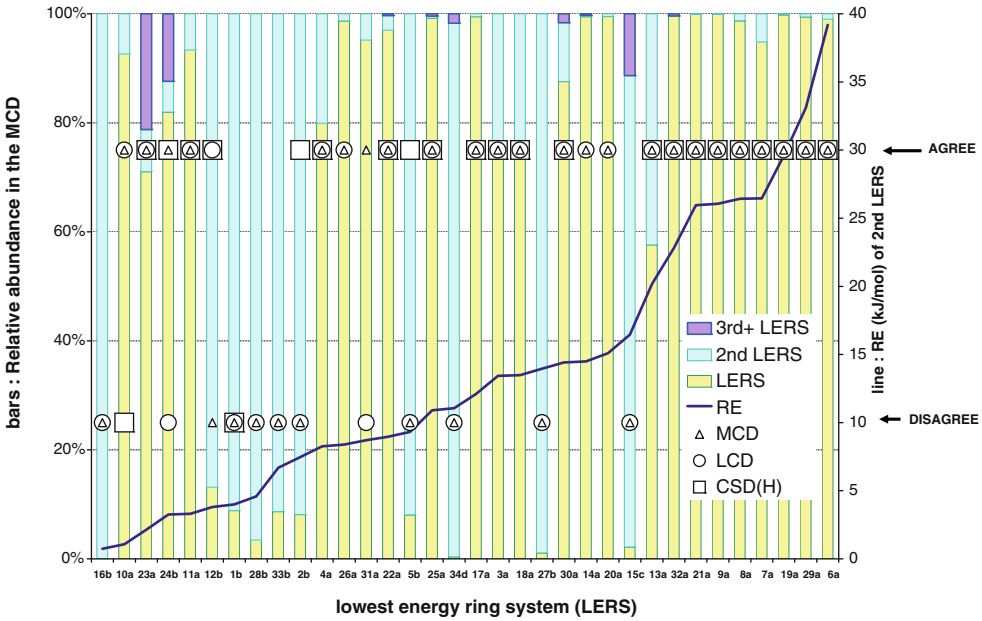

Fig. 9 Plot showing the agreement between the CSD_H (*squares*), compound databases (*triangles*) and 'locked' commercial compounds (*circles*), as a function of the relative energy of the second lowest energy ring system (LERS) in each cluster (*line*). Also shown is the relative abundance of the LERS (*yellow*), 2nd LERS (cyan) and ring systems in each cluster

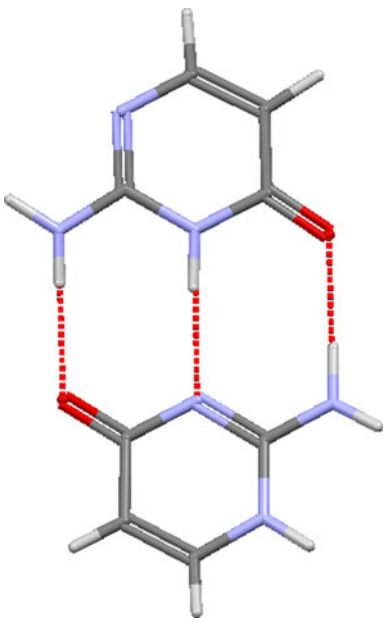

Fig. 10 Hydrogen bond dimer formed between tautomers of the type 4a/4b in the crystal structure with CSD reference code = ICYTIN01

Table 2 Ring system statistics by tautomer class

Class	Number of ring systems						Number predominant			
	All		Locked		RE		Locked		RE	
	1	2	1	2	1	2	1	2	1	2
1	174	173	12	9	14	15	8	4	8	4
2	316	316	9	12	16	16	3	8	2	8
3	333	103	6	3	10	5	5	0	4	1
4	700	691	14	6	16	16	11	0	10	1
5	302	304	2	1	4	2	1	1	2	0
6	72	56	4	2	4	4	3	0	4	0
7	29	17	0	0	0	0	0	0	0	0
8	64	64	1	0	1	1	1	0	1	0
9	119	119	4	5	5	5	2	3	1	4
10	6	6	0	0	0	0	0	0	0	0

Number of ring systems: total number of ring systems either observed or the subject of calculations; Number predominant: number of ring systems which are the predominant form, either the highest number of hits or the lowest energy/cluster; all: all in VEHICLe; locked: at least one substructure hit found in the locked compound dataset; class: as defined in Fig. 3; RE: count with the lowest calculated energy (QM $e = 80$); Yellow: form 1; Cyan: form 2 (see Fig. 3)

Conclusions

In this paper we have searched for experimental data on the existence of alternative annular tautomeric forms within crystal structure databases. These databases contain a huge amount of data but a limited amount on compounds containing potentially tautomeric heterocycles. For this reason, data for the same ring systems with different substituents were pooled. However, even after pooling, due to the power law distribution of the frequency of usage of heteroaromatic ring systems [17], there are a lot of data on a few ring systems and a small amount of data on many ring systems. Nevertheless, the small molecule crystallographic data extracted from the CSD and the PDB are entirely consistent with that calculated using MP2 calculations i.e. the most commonly observed tautomeric form always had the lowest energy, except where the energy difference was small (<5 kJ/mol). An attempt was made to supplement the crystallographic data with information extracted from large numbers of known chemical structures. While in general the most commonly derivatised tautomeric form of a ring system was also that calculated to be one with the lowest energy, there were several notable exceptions to this trend and we do not recommend using this sort of data for this purpose. Pooling the data further into classes of tautomeric transformation, revealed some general trends that could be used as rules of thumb: (1) bicycles containing a pyrimidone substructure (Class t4 in Fig. 3) are found to be almost always protonated at the lactam nitrogen and (2) indazole-like ring systems (Class t1) are more likely to be protonated at the nitrogen next to the bridge but perhaps surprisingly this is by no means a universal finding. In other types of tautomeric transformation, general rules remain to be discovered but may not be necessary. Our MP2 calculations are for 34 commonly observed tautomeric clusters but there is no reason why this set could not be extended to include all 57 clusters found in known compounds or, using a reasonably sized compute farm, to all possible clusters in the VEHICLe database (1544 clusters consisting of 3288 tautomeric alternative). We believe these calculations provide a useful resource for those seeking to generate better models of annular tautomerism in applications such as computer-aided drug design, small molecule crystal structure prediction, the naming of compounds (compounds should be named using the most abundant tautomer [8]) and the interpretation of protein–ligand crystal structures.

Acknowledgments AJCC thanks the Pfizer Institute for Pharmaceutical Materials Sciences for funding. WRP thanks UCB Celltech for funding his secondment to Professor Tom Blundell's group. Thanks also to Dr Yvonne Martin for her encouragement and invitation to submit a paper in this subject area

References

1. Katritzky AR, Lagowski JM (1963) Prototropic Tautomerism of Heteroaromatic compounds: I. General discussion and methods of study. Adv Heterocycl Chem 1:311–338
2. Katritzky AR, Lagowski JM (1963) Prototropic Tautomerism of Heteroaromatic compounds: II. Six-membered rings. Adv Heterocycl Chem 1:339–437
3. Katritzky AR, Lagowski JM (1963) Prototropic Tautomerism of Heteroaromatic compounds: III. Five-membered rings and one hetero atom. Adv Heterocycl Chem 2:1–26
4. Katritzky AR, Lagowski JM (1963) Prototropic Tautomerism of Heteroaromatic compounds: IV. Five-membered rings with two or more hetero atoms. Adv Heterocycl Chem 2:27–81
5. Minkin V, Garnovsk A, Elguero J, Katritzky A, Denisko O (2000) The Tautomerism of Heterocycles: Five-membered rings with two or more heteroatoms. Adv Heterocycl Chem 76:157–323
6. Stanovnik B, Tiler M, Katritzky A, Denisko O (2001) The Tautomerism of Heterocycles. Six-membered heterocycles: Part 1, Annular Tautomerism. Adv Heterocycl Chem 81:253–303
7. Stanovnik B, Tisler M, Katritzky A, Denisko O (2006) The Tautomerism of Heterocycles: substituent Tautomerism of six-membered ring Heterocycles. Adv Heterocycl Chem 91:1–134
8. Elguero J, Katritzky A, Denisko O (2000) Prototropic Tautomerism of Heterocycles: Heteroaromatic tautomerism. General overview and methodology. Adv Heterocycl Chem 76:1–84
9. Shcherbakova I, Elguero J, Katritzky A (2000) Tautomerism of Heterocycles: condensed five-six, five-five, and six-six ring systems with heteroatoms in both rings. Adv Heterocycl Chem 77:51–113
10. Martin YC (2009) Let's not forget tautomers. J Comput Aided Mol Des 23:693–704
11. Alkorta I, Blanco F, Elguero J (2008) Application of Free-Wilson matrices to the analysis of the tautomerism and aromaticity of azapentalenes: a DFT study. Tetrahedron 64:3826–3836
12. Alkorta I, Blanco F, Elguero J (2008) Heteropentalenes aromaticity: a theoretical study. J Mol Struct THEOCHEM 851:75–83
13. Alkorta I, Elguero J, Liebman J (2006) The Annular Tautomerism of imidazoles and pyrazoles: the possible existence of nonaromatic forms. Struct Chem 17:439–444
14. Allen F (2002) The cambridge structural database: a quarter of a million crystal structures and rising. Acta Crystallogr B 58:380–388
15. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov I, Bourne P (2000) The Protein Data Bank. Nucl Acids Res 28:235–242
16. Hao MH, Haq O, Muegge I (2007) Torsion angle preference and energetics of small-molecule ligands bound to proteins. J Chem Inf Model 47:2242–2252
17. Pitt W, Parry D, Perry B, Groom C (2009) Heteroaromatic rings of the future. J Med Chem 52:2952–2963
18. Platonov MO, Samijlenko SP, Sudakov OO, Kondratyuk IV, Hovorun DM (2005) To what extent can methyl derivatives be regarded as stabilized tautomers of xanthine? Spectrochim Acta Part A Mol Biomol Spectrosc 62:112–114
19. Alkorta I, Goya P, Elguero J, Singh SP (2007) A simple approach to the tautomerism of aromatic heterocycles. Natl Acad Sci Letts 30:139–159
20. Bruno I, Cole J, Edgington P, Kessler M, Macrae C, McCabe P, Pearson J, Taylor R (2002) New software for searching the Cambridge structural database and visualizing crystal structures. Acta Crystallogr B 58:389–397
21. James CA, Weininger D, Delay J SMARTS. http://www.daylight.com/dayhtml/doc/theory/index.html

22. Schreyer A, Blundell T (2009) CREDO: a protein–ligand interaction database for drug discovery. Chem Biol Drug Des 73: 157–167
23. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–540
24. Dalby A, Nourse J, Hounshell D, Gushurst A, Grier D, Leland B, Laufer J (1992) Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. J Chem Inf Comput Sci 32:244–255
25. Weininger D (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci 28:31–36
26. Idrissi MS, Senechal M, Sauvaitre H, Cotrait M, Garrigou-Lagrange C (1980) J Chim Phys 77:195
27. Claramunt RM, Lopez C, Garcia MA, Otero MD, Torres MR, Pinilla E, Alarcon SM, Alkorta I, Elguero J (2001) Untitled. New J Chem 25:1061–1068
28. Bairoch A, Apweiler R, Wu C, Barker W, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin M, Natale D, O'Donovan C, Redaschi N, Yeh LS (2005) The universal protein resource (UniProt). Nucl Acids Res 33:D154–D159
29. Alkorta I, Elguero J (2005) Theoretical estimation of the Annular Tautomerism of Indazoles. J Phys Org Chem 18:719–724