Acta Crystallographica Section B
Structural Science
ISSN 0108-7681

Universal prediction of intramolecular hydrogen bonds in organic crystals

Peter T. A. Galek,[a,b]* László Fábián[a,b] and Frank H. Allen[a,b]

[a]Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England, and [b]Pfizer Institute for Pharmaceutical Materials Science, Department of Materials Science and Metallurgy, University of Cambridge, Pembroke Street, Cambridge CB2 3QZ, England

Correspondence e-mail: galek@ccdc.cam.ac.uk

Received 19 November 2009
Accepted 1 February 2010

A complete exploration of intramolecular hydrogen bonds (IHBs) has been undertaken using a combination of statistical analyses of the Cambridge Structural Database and computation of *ab initio* interaction energies for prototypical hydrogen-bonded fragments. Notable correlations have been revealed between computed energies, hydrogen-bond geometries, donor and acceptor chemistry, and frequencies of occurrence. Significantly, we find that 95% of all observed IHBs correspond to the five-, six- or seven-membered rings. Our method to predict a propensity for hydrogen-bond occurrence in a crystal has been adapted for such IHBs, applying topological and chemical descriptors derived from our findings. In contrast to intermolecular hydrogen bonding, it is found that IHBs can be predicted across the complete chemical landscape from a single optimized probability model, which is presented. Predictivity of 85% has been obtained for generic organic structures, which can exceed 90% for discrete classes of IHB.

1. Introduction

The ability to predict the likelihood of hydrogen-bond formation in crystal structures can be advantageous to many fields such as crystal engineering (Aakeröy, 1997; Desiraju, 1995; Etter, 1991), crystal structure prediction (CSP; Day & Motherwell, 2006; Price, 2008), the solution of crystal structures from powder diffraction data (*e.g. DASH*: David *et al.*, 2006), and the prediction of protein–ligand docking (Böhm & Klebe, 1996). Our method which computes a *propensity* for hydrogen-bond formation (Galek *et al.*, 2007), *i.e.* a likelihood of occurrence in specific molecular environments, uses knowledge extracted from existing crystal structures in the Cambridge Structural Database (CSD; Allen, 2002). Recently, the method has been applied as an assessment of structural stability of a given crystal form by virtue of the hydrogen bonds it may or may not possess (Galek, Fábián & Allen, 2009), a result of significant relevance to the pharmaceutical community in providing a chemoinformatic aid to experimental polymorph screening.

Successful application of the approach requires confidence in the propensity model, whose predictivity can be assessed using a range of statistical tests (see §4). Inaccurate models, while they can be identified, can nonetheless call for extra effort in order to obtain reliable predictions. Systematic introduction of uncertainties may naturally arise from the neglect of any important physical influences on the proposed outcomes: for intermolecular hydrogen bonds, perhaps the most significant barrier to an expected interaction is the precursory formation of a counterpart between a donor and acceptor (D and A) of the same molecule, an intramolecular

doi:10.1107/S0108768110003988

hydrogen bond (herein IHB). This interaction removes (at least in part) the atoms' potential participation in bonding in the intermolecular domain. IHBs are also observed to stabilize particular molecular conformations and thus there is a notable influence on conformational polymorphism, which may present differing hydrogen-bonding networks (Galek, Fábián & Allen, 2009). A prototypical example is *o*-acetamidobenzamide (ACBNZA, ACBNZA01; Errede *et al.*, 1981) in which a six-membered amide–amide IHB is observed in only the former polymorph.

A confident prediction of any intermolecular interaction in a crystal structure can therefore only be made once any potential IHBs have been accounted for. Whereas in previous analyses this issue was deferred by training predictive models using only structures in which IHBs are absent, we now turn toward their prediction so that a complete methodology can be developed for the prediction of hydrogen bonds in organic crystal structures. An analogous technique to our existing methodology is applied (Galek *et al.*, 2007), however, the nature of IHBs has been found to depart significantly from their intermolecular counterpart, which has directed the addition of a unique, independent probability function. In particular, this function was prepared by gaining an understanding of IHB formation and the development of a set of corresponding descriptors. Most significantly, IHBs have been observed to behave quite uniformly across the range of organic species encountered in the CSD, which enabled generic models to be developed. This differs from our method for intermolecular hydrogen bonds, which has proved most effective when applied using small training sets composed of only the most chemically relevant crystal structures. It will be shown that our new models can predict the propensity for IHBs to form to a most satisfactory accuracy for the entire set of organic structures in the CSD.

The nature of IHBs varies quite dramatically with characteristics such as their ring size (Etter's notation proves useful, *Rn*, where *n* denotes the covalent bond count connecting the donor H to the acceptor atom +1; Etter, 1991), donor or acceptor chemistry, and other electronic and geometric properties of the bonding fragment (discussed in detail below). To more fully understand these influences, we undertook an investigation of IHB types in terms of both *ab initio* calculations on prototype IHB fragments and statistical analyses of typical IHB geometries and frequencies of occurrence in the CSD. In combination with the extensive literature available, we have been able to explore the physical effects on IHB likelihood and set up a framework for the types of IHB we aim to predict. In the context of our method, it is necessary to provide an algorithmic definition of an IHB, *i.e.* the proposed observation must be identifiably present or absent (see §3). In a wider sense, these studies also allow us to address some topical issues such as the somewhat ambiguous interaction between *D* and *A* separated by three covalent bonds (a potential *R*5 IHB; are they really hydrogen bonds?) or the nature of IHBs between topologically well separated *D* and *A* atoms.

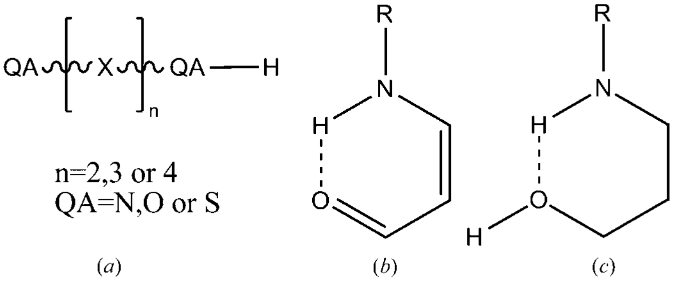

Figure 1
(*a*) The form of the CSD search query to locate compounds with the potential for forming generic intramolecular hydrogen bonds, (*b*) enaminone query fragment, (*c*) propanolamine query fragment.

The paper begins with an energetic and statistical assessment leading to the definition of our model IHBs. Next, a summary of the probability modelling is given and a definition of the model descriptors applied. A selection of IHB logit models is then presented which vary by the specific type of intramolecular bond they predict. Some example predictions are then presented which apply the new IHB propensity model functions. A discussion of the predictivity of our method and its performance for some best/worst cases is then presented, followed by some concluding remarks.

2. Data preparation

A dataset of organic crystal structures was prepared by searching the CSD (Version 5.30, plus the November 2008 and February 2009 updates) for structures that exhibit at least one IHB, as defined using our *H-BOND SURVEYOR* code (Galek *et al.*, 2007). Hydrogen-bond distance (r_{DA}) and angle (θ_{DHA}) tolerances categorize a true or false hydrogen-bond observation with settings of $r_{DA} < \Sigma r_{\rm vdW} + 0.1$ Å, and a $\theta_{DHA} > 90^\circ$, where $r_{\rm vdW}$ denotes the atomic van der Waals radius. Wood and co-workers (Wood *et al.*, 2009) have recently shown that 120° provides a realistic lower bound for hydrogen-bond angles, however, they noted that such a limit would preclude a proportion of IHBs. This trend is also observed in these studies (see below) leading to the chosen limit. All bifurcated (or further subdivided) hydrogen bonds are regarded as two (or more) observations. Intramolecular contacts are further specified by *D* and *A* sharing a covalently bonded unit and separated by more than three chemical bonds. Structural duplicates were also removed using a structure overlay method (Chisholm & Motherwell, 2005). These are infrequent but can occur, *e.g.* due to structure redeterminations. The resulting set of 22 041 crystal structures has been used in a statistical study of the bonding character of IHBs.

A second dataset was prepared containing structures with the potential for IHB formation (but not necessarily exhibiting IHBs, *i.e.* covering all true and false observations) to allow development of our predictive model function. Relevant CSD structures were obtained using *ConQuest* (Bruno *et al*, 2002; Cambridge Crystallographic Data Centre, 2009) searches for a generic IHB motif query (Fig. 1), which consists of N, S or O donor/acceptor atoms linked *via* unspecified atom types. Investigations of the frequency of IHB types in the CSD reveal that more than 95% of all observed IHBs belong to three ring sizes: *R*5, *R*6 or *R*7 (detailed discussion follows below), hence these constraints are specified. Structures with determined three-dimensional coordinates and $R < 0.1$ were retained, whereas those solved from powder X-ray diffractograms, or those that contain metal atoms, residual errors, disorder or polymeric (*catena*) bonds were discarded. The resultant dataset contained 32 550 structures.

Identification of the true IHBs from the set of potential observations was carried out using the criteria as above. Note that this dataset and the first set contain an almost equivalent set of true IHB observations, differing only by a small fraction of structures having only IHB motifs of ring size larger than 7. Next, false IHBs are recorded, that is those potential *D*–*A* pairs which do not interact within the geometric criteria defined above. To exclude fragments which could not possibly adopt the conformation required for intramolecular ring formation, torsion angles about non-rotatable multiple bonds and cyclic single bonds were restricted to values of −50 to +50°, and as a further constraint, the sum of the torsion angles for such inflexible bonds was restricted to < 180°. Mutually exclusive observations, *i.e.* an IHB disallowed by the presence of an alternative, were ignored [*e.g.* if amide(N—H)···(O═)carboxylic acid is observed then amide(N—H)···(OH)carboxylic acid is excluded]. Analogously to the true observations, the set of descriptor values are recorded for each false observation. The resultant data is then amenable to model regression techniques and further analysis.

3. The bonding character of IHBs

Thanks to comprehensive previous studies (as summarized by Buemi, 2006), the general nature of IHBs is well understood. IHB strength is most significantly correlated with the extent of π-electron delocalization in the region separating the donor and acceptor. Thus, IHBs are closely associated with the class of resonance-assisted hydrogen bonds (RAHBs). IHBs occur with and without resonance assistance, although the former (which we denote as resonance-assisted intramolecular hydrogen bonds; RAIHBs) are often much the stronger: the conjugated π-electrons associated with those environments offer a degree of charge transfer and favourable electrostatics at the donor and acceptor. In addition, a delocalized electron cloud consistent with π-type molecular orbitals supplies conformational constraints which can enforce an effective intramolecular hydrogen-bonding arrangement. The chemistry at the donor and acceptor is critical to potential bond strength, as *e.g.* electron-withdrawing or electron-donating substituents can affect donor acidity and acceptor basicity, respectively (Allen *et al.*, 1997).

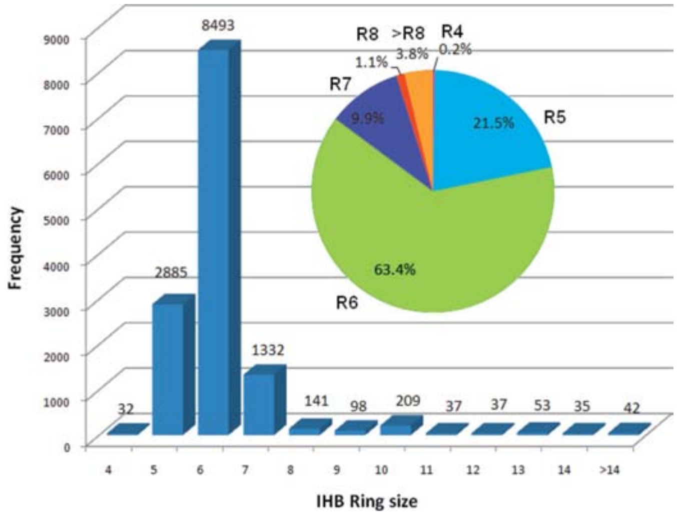

Figure 2
Frequency of intramolecular hydrogen-bond ring motif size (*Rn*) observed in the CSD. Inset shows relative percentages of the total set of observations.

Bilton and co-workers (Bilton *et al.*, 2000), in a study of organic crystal structures in the CSD, identified the significant influence of the ring size of the potential IHB. In particular, certain six-membered ring motifs are almost 100% likely to form in structures where they are possible, whereas the expected probability for motifs of other ring sizes can be much more variable. This study also highlighted the influence of the chemistry of the fragment containing the donor and acceptor. It is notable that subtle differences between certain IHB motif types changed the average frequencies of occurrence quite significantly. In our studies, the general effects of electron delocalization and ring size on the character of IHBs have been systematically investigated in a threefold approach: (*a*) geometrical analysis from the CSD, (*b*) *ab intio* quantum chemical calculations and (*c*) comparison with exisitng literature data.

3.1. CSD geometry analysis

The frequency of occurrence of IHB motifs of various ring sizes was extracted by identifying all IHBs in organic CSD structures and recording the shortest path linking *D* and *A* within the molecular graph. The results, Fig. 2, reveal that *R*6 motifs represent more than 60% of all IHBs encountered. *R*5 and *R*7 in combination with *R*6 then make up 95% of the observations. Larger motifs, up to *R*20 are observed, however, beyond *R*10 they occur very infrequently. This observation appears characteristic of the small-molecule organic compounds in this study, in contrast to the well known 13- and ten-membered IHBs that form β-helices and α-turns of polypeptides and proteins (Jeffrey & Saenger, 1991). We note that the CSD contains small biomolecules such as tripeptides which may account for slight peaks in the histogram of Fig. 2 at *R*10 and (less so) *R*13. Interestingly, a handful of *R*4 motifs are identified. While they would seem to be incapable of forming a favourable $D—H\cdots A$ geometry, a few examples fall inside the r_{D-A} and θ_{DHA} tolerances. Such four-atom IHBs cannot realistically be thought of as a stabilizing interaction: *R*4 IHBs were shown not to exist within the carboxylic acid functional group, for example (Hermida-Ramon & Mosquera, 2006).

Collectively, the *R*5–7 IHB motifs form a natural categorization of the interaction we wish to predict. Any further unique characteristics they may offer for the modelling can be

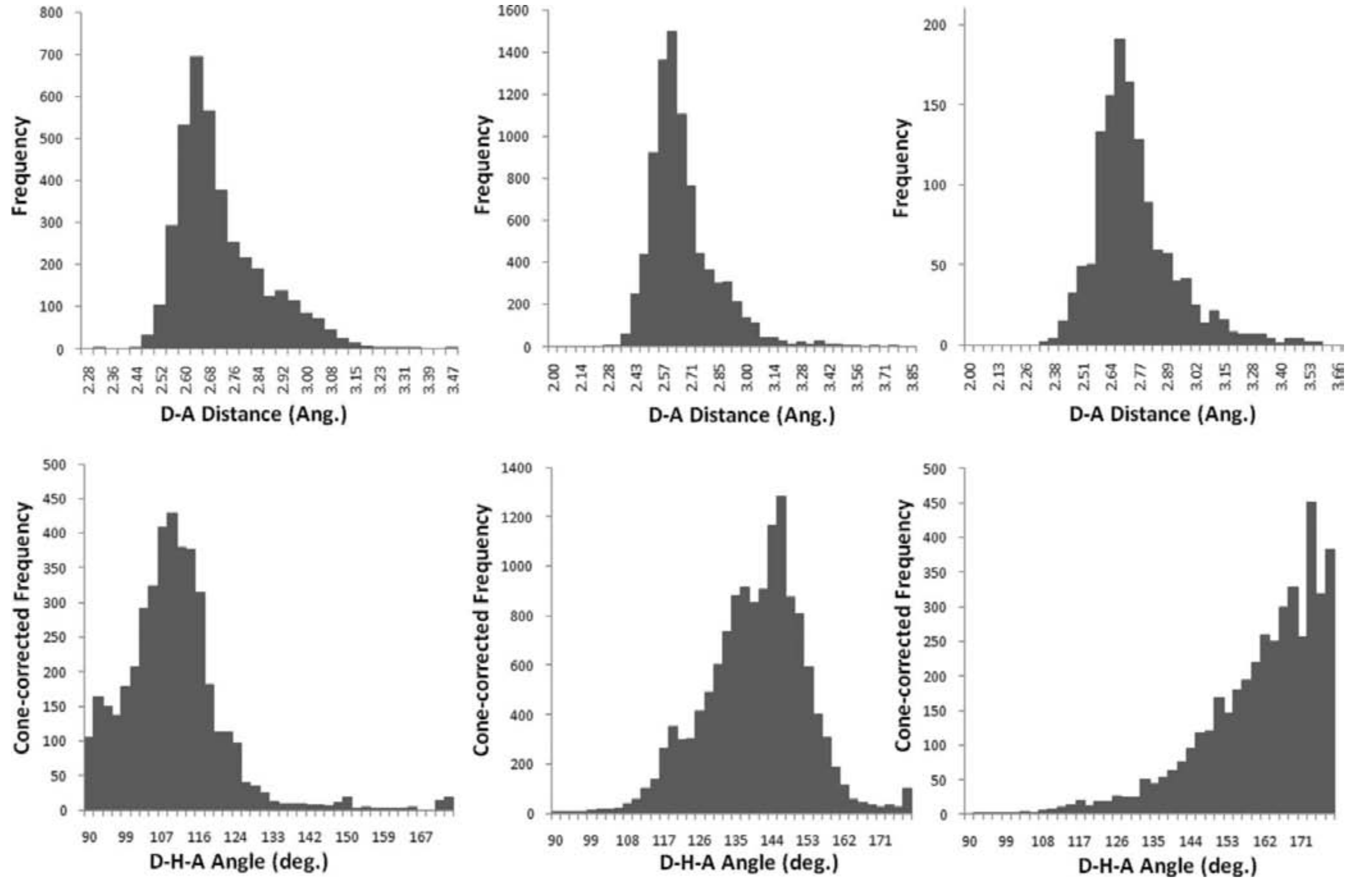

Figure 3
Distributions of donor–acceptor distances (top row) and donor–hydrogen-acceptor angles (bottom row) for observed (*a*) *R*5, (*b*) *R*6 and (*c*) *R*7 intramolecular hydrogen-bond motifs in the CSD.

Table 1
Mean values for distance and angle distributions of observed IHBs in the CSD for motifs of size 5, 6 and 7 (mode values in parentheses).

Motif	r_{DA}		r_{HA}		θ_{DHA}	
*R*5	2.735	(2.637)	2.321	(2.225)	109.4	(109.3)
*R*6	2.692	(2.616)	1.932	(1.884)	137.8	(146.2)
*R*7	2.768	(2.681)	2.184	(2.048)	151.9	(173.3)

probed by statistical analysis of the CSD. Distributions of $r_{DA,}$ $r_{\rm HA}$ and $\theta_{D{\rm HA}}$ in CSD structures can reveal contrasts in the behaviour of the various IHB types (Fig. 3). A cone correction (Kroon *et al.*, 1975) was applied to $\theta_{D{\rm HA}}$ in order to remove sampling bias. Histograms of a normalized $D-A$ distance $r_{\rm norm} = r_{DA}/(r_D + r_A)$ were also prepared to remove any effect of variable van der Waals radii, which appeared qualitatively identical (see supplementary material[1]). These observations are also represented as scatter plots (Fig. 4) which show the observed A position after a transformation which places D at the origin and the $D-$H bond on the crystallographic z axis, to allow consistent overlay. Comparing the resulting plots, the contrast between the hydrogen-bond geometries of the three major ring sizes is quite apparent (see Table 1 for selected statistics). *R*5 IHBs show a mode at a relatively short r_{DA} = 2.637; slightly longer than that for *R*6 motifs: r_{DA} = 2.616. *R*7 has a longer modal r_{DA} which tails off more slowly to higher r_{DA}. More significant contrasts lie with the IHB angles. The effect of constraints in the *R*5 ring is clear, with a low peak in the $\theta_{\rm DHA}$ distribution at around 109°. The modal observation in *R*6 IHBs is higher, θ_{DHA} = 146°, but still a long way short of the ideal linear 180° geometry. Nonetheless, this difference from *R*5 to *R*6 is most likely associated with a significant increase in hydrogen-bond strength. The distribution of $\theta_{D{\rm HA}}$ for *R*7 IHBs is typical of a classic intermolecular hydrogen bond with a maximum as $\theta_{D{\rm HA}}$ approaches 180°. Motifs of this size and above have enough flexibility to allow a linear $D-$H$\cdots A$ relationship, which the smaller motifs do not.

*R*6 IHBs were investigated in more detail by plotting equivalent histograms as above for CSD structures containing one of two prototypical fragments: either (*a*) β-enaminone, which can form an RAIHB, or (*b*) 1-aminopropan-3-ol, its saturated analogue (Figs. 1*b* and *c*). The distributions (Fig. 5) reveal significant contrasts in the geometries of the RAIHB *versus* the saturated IHB (Fig. 3; Table 2). Without resonance assistance, the mean $r_{\rm HA}$ is 2.234, significantly longer than that

[1] Supplementary data for this paper are available from the IUCr electronic archives (Reference: SO5033). Services for accessing these data are described at the back of the journal.

for the RAIHB, $r_{\mathrm{HA}} = 1.931$. θ_{DHA} is also more linear by 8–10° for the RAIHB on average. Interestingly, the spread of observed angles is much wider for the saturated IHB; the small variance in the distribution for the RAIHB indicates a more definite interaction and would support the suggested stronger bonding.

3.2. *Ab initio* energies

While there have been many previous theoretical studies on individual systems that exhibit IHBs, here we collect data for comparison between a selection of related prototypical systems with the potential to form five-, six-, and seven-membered IHBs, both with and without conjugated covalent bonding in the motif. In this way, relative interaction strengths, geometries and other parameters can be systematically compared to complement our statistical analysis of IHBs in the CSD. Calculations were performed with the *SPARTAN* program (Carpenter *et al.*, 1980; Wavefunction Inc., 2008). An estimate of the stabilizing energy due to IHB formation, E_{IHB}, is computed by taking the energy difference between the minimum energy conformation and the so-called 'open' conformation, which is related by a rotation of the donor H atom by 180°. As noted by previous authors, energetic influences other than the hydrogen bond, *e.g.* steric constraints, also affect this relative energy and so the value can only be an estimate. However, using the same approach for a series of fragments does provide a systematic comparison. For these reasons, extra effort to include the minor effects of either zero point-energy correction or intramolecular basis set superposition error has not been made. Cases in which the global minimum does not contain an IHB are assigned a value of 0.0 kJ mol^{-1}. For reference, energies obtained in this work are compared with literature values where available.

Table 2
Mean values for distance and angle distributions of observed IHBs in the CSD for β-enaminone and 3-aminopropan-1-ol fragments (RAIHBs *versus* saturated ring IHBs; Figs. 1*b* and *c*); mode values in parentheses.

Motif	Search Fragment	r_{DA}		r_{HA}		θ_{DHA}	
*R*6	β-Enaminone	2.660	(2.670)	1.931	(1.925)	136.2	(138.3)
	3-Aminopropan-1-ol	2.866	(2.790)	2.234	(2.108)	128.7	(122.0)

Table 3
Ab initio geometrical data for global minimum ground-state conformations of selected prototypical IHB fragments, and estimated intramolecular hydrogen-bond energies, E_{IHB}, where applicable.

Values computed at the B3LYP/6-311++G(2d,2p) level in this work.

Motif	Compound (Fig. 6 ref.)	Reference	r_{DA}	r_{HA}	θ_{DHA}	Est. E_{IHB} (kJ mol^{-1})
*R*5	Ethylene glycol (*h*)	Present work	2.819	2.389	106.2	0.0
		Howard & Kjaergaard (2006) [QCISD/6-311++G(2d,2p)]	–	2.361	107.5	–
	Ethanolamine (*e*)	Present work	2.823	2.289	114.0	19.06
		MacLeod & Simons (2003) [MP2/6-311+G**]	–	2.251	114.9	–
	Aminoethanone (*a*)	Present work	2.769	2.356	103.0	0.0
*R*6	Malonaldehyde (*i*)	Present work	2.578	1.687	146.4	52.99
		Grabowski (2001) [MP2/6-311++G**]	2.585	1.687	148.4	50.63
		Buemi (2006) [B3LYP/6-311++G(d,p)]	2.587	–	–	54.14
		Buemi & Zuccarello (2004) [B3LYP/aug ccpVQZ)]	2.571	–	–	53.03
		Hargis *et al.* (2008) [B3LYP/DZP++]	2.546	–	–	–
	1,3-Propanediol (*j*)	Present work	2.821	2.048	137.2	14.73
		Howard & Kjaergaard (2006) [QCISD/6-311++G(2d,2p)]	–	2.045	137.5	–
		Mandado *et al.* (2006) [B3LYP 6-311++G(2d,2p)]	–	2.059	–	22.47
	β-Enaminone† (*b*)	Present work	2.720	1.966	128.4	18.24
		Gilli *et al.* (1994) [B3LYP/6-31+G(d,p)]	2.713	1.956	128.8	18.54
		Gilli *et al.* (2000) [MP2 6-31+G(d,p)]	2.702	1.951	128.5	18.95
	3-Aminopropan-1-ol† (*f*)	Present work	2.834	2.005	141.0	18.20‡
	3-Aminopropan-1-one (*b*)	Present work	3.048	2.353	116.5	9.83
*R*7	4-Aminobutan-1-one (*d*)	Present work	3.098	2.217	143.8	0.0
	4-Aminobutan-1-ol (*g*)	Present work	2.823	1.882	160.6	24.25‡

† Refer to Table 2 for the mean geometry from CSD analyses. ‡ Value computed using a global minimum which involves the OH···N hydrogen bond.

Initially MP2 and DFT/ B3LYP levels of theory were compared, using a 6–31 + G* basis with consistent results which also agreed with literature references where available. Some small discrepancies for very strong IHBs between particular levels of theory have been noted (Klein, 2002*a*,*b*; Buemi & Zuccarello, 2004), which might be explained by a failure of DFT methods to correctly describe an energetic dispersion component. In addition, polarized basis sets might also offer an improved description for these systems. However, equivalent results from less and more expensive levels of theory have been previously reported, *e.g.* for a series of aliphatic diols (Howard & Kjaergaard, 2006). To explore any potential effect in the present work we also investigated computed energies using the aug-ccPVTZ and 6-311++G(2dp,2p) bases. Subsequent results are from calculations performed at the B3LYP/ 6-311++G (2d,2p) level, observed to be most consistent over the series of fragments studied.

To complement the CSD surveys above, a series was studied containing an N—H donor and an O acceptor: aminoethanone, β-enaminone, 3-aminopropan-1-one, 4-aminobutan-1-one (Figs. 6*a*–*d*). The related compounds ethanolamine, 3-aminopropan-1-ol and 4-aminobutan-1-ol (Figs. 6*e*–*g*) were also investigated, also of interest owing to their relation to beta-blockers such as ephedrine and adrenaline (MacLeod & Simons, 2003). For comparison, and owing to the extensive previously published literature, a second series was studied containing O as both donor and acceptor:

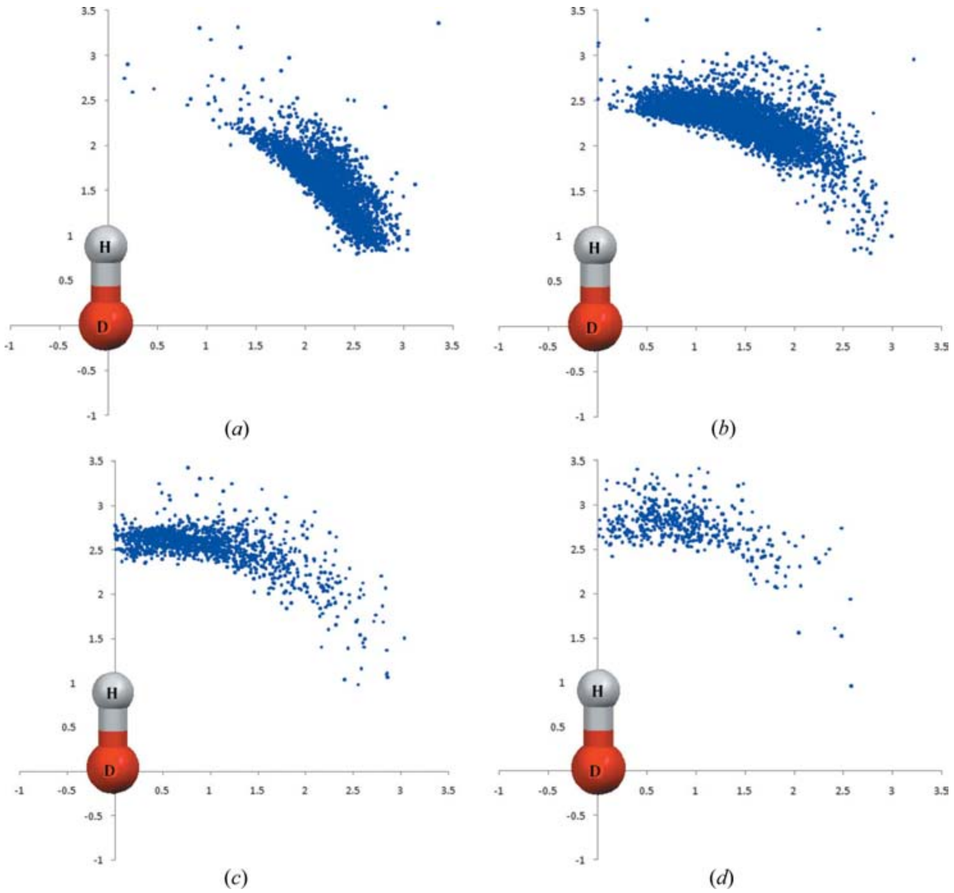

Figure 4
Scatterplots showing observed acceptor atom positions for (*a*) *R*5, (*b*) *R*6, (*c*) *R*7 and (*d*) combined *R*8, *R*9 and *R*10 IHBs in the CSD in relation to the vector defining the donor–hydrogen covalent bond.

ethylene glycol, malonaldehyde, and 1,3-propanediol (Figs. 6*h*–*j*).

Table 3 compares computed hydrogen-bond geometries and estimated IHB energies for the prototypes. First, note the excellent general agreement between the CSD geometry statistics (Tables 1 and 2) and the computed hydrogen-bond geometrical parameters in Table 3. This is a significant link between IHBs in the crystalline state and in the gas phase. As from the earlier CSD statistics, the difference between various IHB ring sizes is quite apparent from the calculations. The computed r_{HA} values are clearly the shortest for the unsaturated *R*6 fragments, with these prototypes exhibiting the largest estimated hydrogen-bond energies. The hydrogen bonds of the saturated *R*6 systems are longer than the unstaturated systems, *e.g.* r_{HA} for propanediol is ~ 0.35 Å longer than in malonaldehyde, and the IHB angle is slightly more bent. More importantly, the corresponding estimated $E_{\rm IHB}$ is significantly lower for the unsaturated fragment. This trend is noticeable, to a lesser extent between the heteroatomic IHBs in enaminone and aminopropanol. An increased hydrogen-bond strength for homoatomic O—H$\cdots$O interactions *versus* heteroatomic N—H$\cdots$O IHBs has been commented on previously by Gilli *et al.* (2000).

Strain in the potential motif ring is clearly influential: although it forms a planar geometry to enable hydrogen-bond formation in its minimum-energy conformation, the potential *R*6 IHB in 3-aminopropan-1-one is barely evident according to $E_{\rm IHB}$. The issue is worse for the *R*5 aminoethanone and ethylene glycol fragments, and the *R*7 4-aminobutan-1-one, which have global minimum-energy conformations without any D—H$\cdots A$ interaction. Another factor here could also be of a less favourable NH$\cdots$O interaction compared with an OH$\cdots$N IHB, as evidenced by the observed IHB in the equilibrium conformations of aminopropanol and aminobutanol. This observation may also explain why the aminoalcohols tend to form better IHBs than the aminoketones in Table 3, as the latter only have the option of NH$\cdots$O IHBs.

Some small variations in the data in Table 3 are dependent on the chosen level of theory and basis set, arising from the ability to describe electron correlation, *e.g.* compare calculation of the β-enaminone ground-state energy, $E_0 = -246.537217$ a.u. (MP2/6-31G*) *versus* −247.300733 a.u. (B3LYP/6-311++G[2d,2p]). Discrepancies become much less significant to energy differences and related properties, however, for example in malonaldehyde, all computed values of r_{HA} agree well with the experimentally determined 1.68 Å result for the deuterated system (Baughcum *et al.*, 1981). Thus, other absolute energy values are not reported (also these are often unavailable in the works cited in Table 3).

3.3. Discussion

The nature of IHBs of various size can now be discussed in greater detail. *R*6 and *R*7 IHBs are clearly structurally relevant, stabilizing interactions with a true bonding character, but what of five-membered IHBs? We commonly observe conformational arrangements of molecules in the CSD which form a five-ring intramolecular geometry. They are much less common than *R*6 IHBs, but occur more frequently than *R*7 IHBs. Bilton and co-workers (Bilton *et al.*, 2000) identified 26 unique IHB motifs that occur with >50% frequency in crystals in which they can potentially form. Of those 26, six motifs were unique *R*5 IHBs. We find a similar proportion, 21%, of all IHB interactions are *R*5 motifs. Hence geometry considerations indicate a structurally relevant role. A theoretical framework for the identification of chemical bonds is provided by QTAIM (quantum theory of atoms in molecules; Bader, 1990, 1991; Bader & Laidig, 1992), requiring a bond-critical point (BCP) in the electron density characterized by a line of maximum curvature passing through the BCP linking the involved nuclei. IHBs can be further identified by a ring-critical point (of positive curvature) which must arise internal to the IHB motif as a consequence of the BCP. Electron-density maps computed in the $D—H—A$ planes for a selection of the prototypes studied are shown in Fig. 7. Local maxima corresponding to hydrogen bonding can clearly be seen in Figs. 7(*a*)–(*d*) and (*f*). Moreover, more pronounced maxima appear to correlate with computed stronger interactions (compare malonaldehyde, Fig. 7*a*, with 3-aminopropan-1-ol, Fig. 7*d*; 52.99 and 18.20 kJ mol^{-1}, respectively).

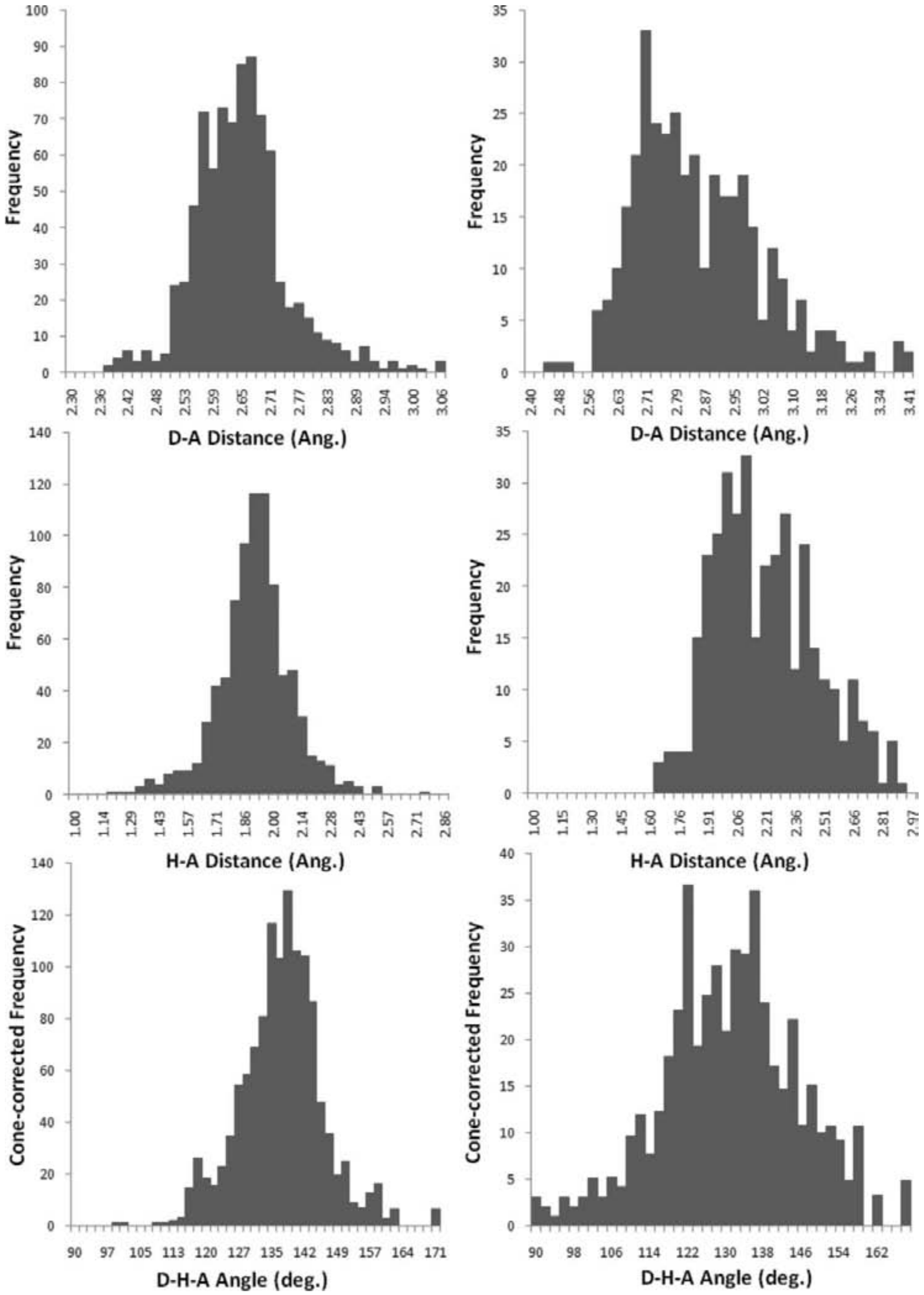

Figure 5
Comparing distributions of donor–acceptor distances (top row), hydrogen–acceptor distances (middle row) and donor–hydrogen–acceptor angles (bottom row) for observed resonance-assisted *versus* saturated *R*6 IHB motifs in the CSD (left *versus* right, respectively).

Our calculations did not reveal the presence of a BCP in the electron densities of the potential *R*5 IHB formers 1-aminoethan-2-one and ethylene glycol. An electron density surface for the former can be seen in Fig. 7(*e*), which neither a bond-critical point nor a ring-critical point are visible. Other authors have made the same observation for the latter (Klein, 2002*a*,*b*; Mandado *et al.*, 2006). It would seem that the strain in the five-ring is too high for hydrogen bonding. Nonetheless, according to QTAIM theory, true five-membered IHBs have been reported in both the ground-state conformation of glycine (Pacios & Gómez, 2001) and glycolic acid (Kassimi *et al.*, 2002; Roy *et al.*, 2005). The enhanced planarity of the molecule due to an sp^2 carbon adjacent to the acceptor would seem to facilitate the hydrogen bond in these cases. An estimate of $E_{\rm IHB}$ can be obtained from the difference in computed energies of the hydrogen-bonded and open conformers. For glycine, in the work of Pacios & Gómez (2001) $E_{\rm IHB}$ can be calculated as 5.27 kJ mol^{-1} (the conformers are denoted Ip and IVn). Equivalently, the relevant rotamers of glycolic acid (denoted G and V; Roy *et al.*, 2005) differ by approximately 18.41 kJ mol^{-1}, which is similar to $E_{\rm IHB}$ for ethanolamine calculated in the present work. Hence, neither are very strong interactions. Interestingly in the latter case, the intramolecular hydrogen bonding is broken in both known crystal structures in favour of intermolecular hydrogen bonds (GLICAC01, Pijper, 1971; GLICAC10, Ellison *et al.*, 1971). This observation serves as a reminder of the potential contrasts due to packing effects, other energetic influences and entropy in organic crystal structures at finite temperatures compared with 0 K gas-phase models.

It can be concluded that *R*5 IHBs are, in some cases, true hydrogen bonds. It is worth noting that whether *R*5 bonding interactions are formed or not, their commonality in the CSD can be interpreted to represent a not insignificant contribution to lattice stabilization. It would, however, appear minimal compared with larger-ring IHBs. In the context of HB propensity modelling, one might expect they have a less predictable behaviour than their larger siblings. Conversely, the other prototype fragments studied can have clear bonding interactions involving an hydrogen donor and acceptor which appear to vary in a systematic way. We emphasize that an aim of this work is to provide predictive models from the starting point of a molecular diagram of a target (*i.e.* with no three-dimensional structural information): the statistical and energetic trends observed here would suggest that these IHBs are indeed amenable to such prediction.

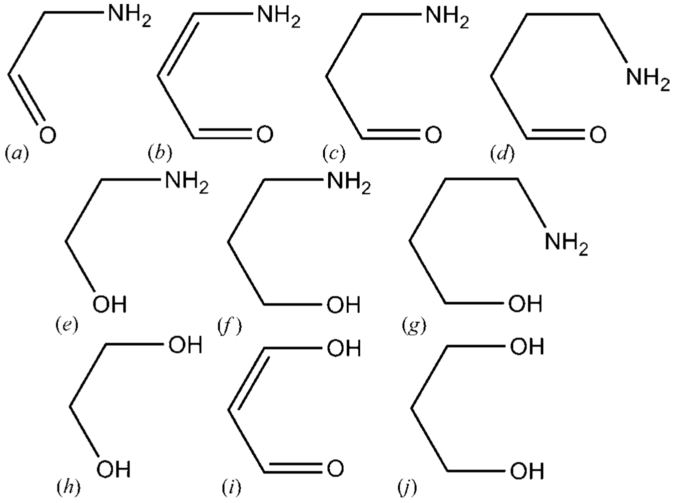

Figure 6
Prototype IHB fragments analysed in the *ab initio* theoretical calculations: (*a*) aminoethanone, (*b*) β-enaminone, (*c*) 3-aminopropan-1-one, (*d*) 4-aminobutan-1-one, (*e*) ethanolamine, (*f*) propanolamine, (*g*) butanolamine, (*h*) ethylene glycol, (*i*) malonaldehyde and (*j*) 1,3-propanediol.

4. Logit hydrogen-bond propensity methodology

In previous work (Galek *et al.*, 2007; Galek, Fábián, Allen & Feeder, 2009) hydrogen-bond likelihood has been successfully modelled as a two-state outcome: true or false using a *logit* function (one of a category of discrete choice models for binary variables; Agresti, 1990; Hosmer & Lemeshow, 2000). Crucially, the probability of an outcome, denoted π, is linked to a linear combination of explanatory variables through the relation

$$\operatorname{logit}(\pi^i_{c,k}) = \log\left(\frac{\pi^i_{c,k}}{1-\pi^i_{c,k}}\right) \qquad (1a)$$

$$= \alpha + \sum_k x^i_k \beta_k \qquad (1b)$$

such that underlying factors influential to a particular outcome, x_k, can direct a prediction. The superscript i denotes an individual *D*–*A* pairing, the subscript c notation is consistent with the formalism denoting a discrete choice model, and the k index runs over all model descriptors. The set β must be determined during model optimization using logistic regression to a set of known hydrogen-bond true/false outcomes and

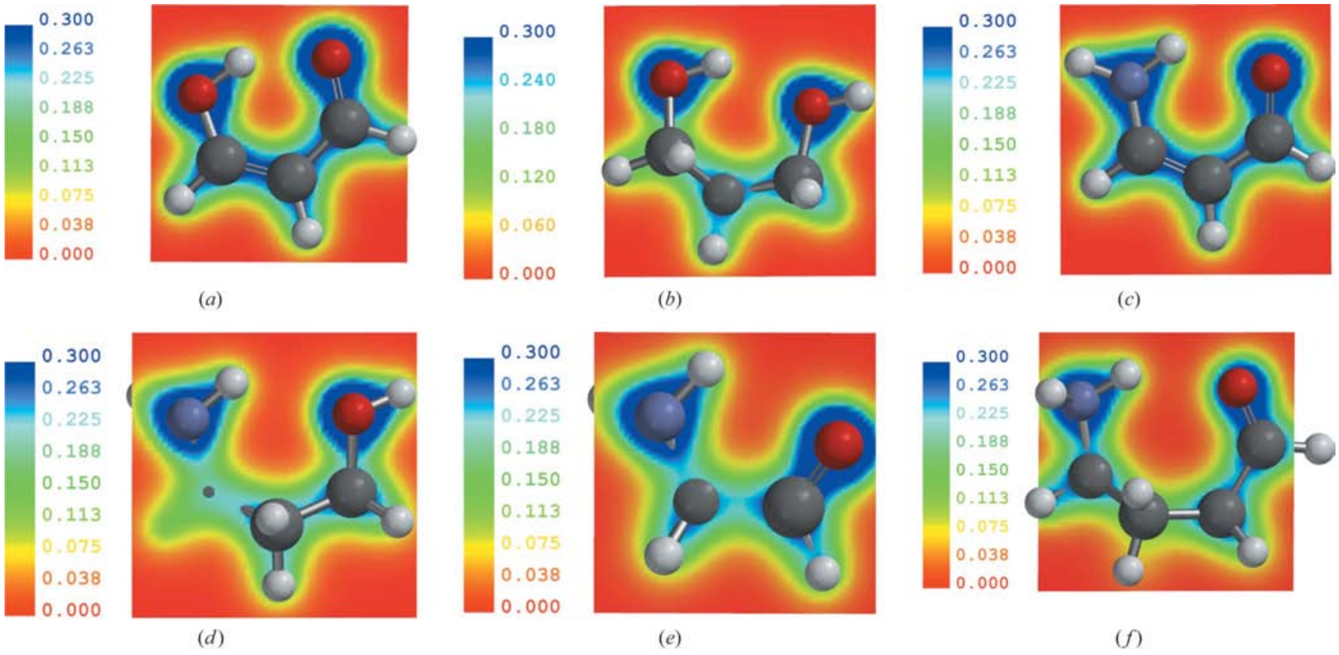

Figure 7
Electron-density map through the donor–hydrogen–acceptor plane in various IHB motifs, calculated at the B3LYP, 6-311++G(2d,2p) level: (*a*) malonaldehyde, (*b*) 1,3-propanediol, (*c*) β-enaminone, (*d*) propanolamine, (*e*) aminoethanone and (*f*) 4-aminobutan-1-one.

Table 4
Optimized model parameters: general IHB logit model.

The 'Value' column displays coefficients (α and the set β_k) which form the propensity model according to equation (2), obtained following logistic regression on training data detailed in the text. Odds ratios represent influence towards a true or false prediction in the logit function (see text). $\Pr > \chi^2$ is an indication of parameter significance. Coefficient bounds are also provided.

Descriptor	Category	Value	Odds ratio	$\Pr > \chi^2$	Lower bound (95%)	Upper bound (95%)
α		2.553	n.a.	< 0.0001	2.208	2.897
β						
Donor	N.3	0.000	1.00	–	–	–
	N.4	−0.032	0.97	0.848	−0.356	0.292
	N.am	−0.788	0.45	< 0.0001	−0.983	−0.593
	N.pl3	−0.117	0.89	0.281	−0.330	0.096
	O.3	−1.113	0.33	< 0.0001	−1.306	−0.920
	Other	−0.410	0.66	–	–	–
Acceptor	Other	0.000	1.00	–	–	–
	Cl	−0.583	0.56	0.001	−0.941	−0.225
	F	−7.406	0.00	< 0.0001	−10.205	−4.607
	N.1	−2.286	0.10	< 0.0001	−2.761	−1.810
	N.2	0.657	1.93	< 0.0001	0.347	0.968
	N.3	1.241	3.46	< 0.0001	0.858	1.625
	N.ar	2.085	8.04	< 0.0001	1.748	2.422
	O.2	0.780	2.18	< 0.0001	0.496	1.065
	O.3	0.120	1.13	0.407	−0.164	0.404
	O.co2	0.834	2.30	< 0.0001	0.429	1.238
	S.3	−1.508	0.22	< 0.0001	−1.855	−1.160
Ring size	6	0.000	1.00	–	–	–
	7	−0.910	0.40	< 0.0001	−0.993	−0.826
	5	−1.705	0.18	< 0.0001	−1.808	−1.603
Path constraint	True	0.000	1.00	–	–	–
	False	−1.985	0.14	< 0.0001	−2.064	−1.905

parameter values (the training set). The size and sign of the β coefficients effectively control the influence of each parameter on the model prediction. π is a strict probability of forming a hydrogen bond: a maximum of $\pi = 1$ indicates a true predicted hydrogen bond and a minimum $\pi = 0$ denotes a false predicted hydrogen bond (*i.e.* no interaction). Note that in this formalism, a false prediction is not concerned with any other interaction D or A may or may not be involved with, only that for the specified D–A pair, an IHB will not form.

A set of chemical/molecular descriptors have been implemented to capture the influences on IHB formation (the set x_k, see below). To build the model equation, the training set is obtained from the chosen crystal structures using our *H-BOND SURVEYOR* algorithm (Galek *et al.*, 2007). Logistic regression and subsequent statistical quality assessment was carried out using *XLSTAT* (Addinsoft, 2008), a statistical software plug-in application to *Microsoft Excel*. Providing a satisfactory model can been obtained, assessment of a target compound is then performed, *i.e.* a set of π values is computed for all potential hydrogen bond $D-A$ pairs. A prediction is made using a rearrangement of (1)

$$\pi^{i}_{c,k} = \frac{1}{1 + \exp(-\alpha - \sum_{k} x^{i}_{k}\beta_{k})}. \qquad (2)$$

Each computation requires an evaluation of the x^i_k parameters associated with a chosen $D-A$ pairing. Examples applying (2) are provided in §5.1. All model descriptors require at most two-dimensional connectivity data (*e.g.* by way of a chemical diagram) enabling this step to be truly predictive for a chosen compound.

4.1. Influences on intramolecular hydrogen-bond formation

Our investigations have yielded a set of distinct influences on the likelihood of IHB formation. The first three are:

(i) the extent of π-electron conjugation in the region separating the donor and acceptor,

(ii) the chemical groups comprising donor and acceptor, and

(iii) the size of the potential IHB ring motif.

Further to these, competition from other donors has been observed to decrease a particular IHB's potential, which increases with donor numbers. However, the effect of other acceptors was not seen to be systematically influential. The form of the descriptors is now given. In most cases, one descriptor sufficed for each described influence, however, a number of solutions were trialled to represent the influence of electron delocalization in the fragment separating D and A. After investigating a variety of descriptions, we found that a separate description of the bond rotatabilities and the bond multiplicities was most effective.

4.1.1. Hydrogen-bond motif ring size. Potential IHB ring size is computed from the number of covalent bonds separating D and A plus one $D-H$ bond plus one potential $H\cdots A$ bond. Etter's Rn notation denotes a ring motif of size n; for

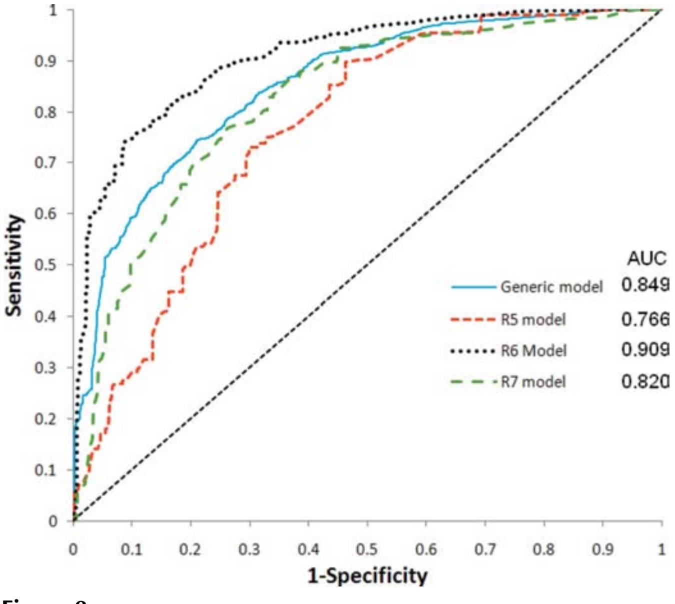

Figure 8
ROC curves for the 25% validation sets of the four IHB model functions developed during the study.

cyclic fragments, the most direct covalent path separating D and A contributes $n - 2$. This variable has three categories, $R5$, $R6$ and $R7$. As discussed previously, motifs larger than $R7$ occur less systematically, are more rare, and are correspondingly not modelled in this work.

4.1.2. Path conjugation descriptor. To describe electron conjugation in the $D-A$ intramolecular path, a text string to denote the covalent bond type is employed, *e.g.* *SUS* denotes a potential $R5$ intramolecular motif consisting of saturated–unsaturated–saturated bonds. The label is easily constructed using bond-type information stored within each CSD entry.

4.1.3. Path constraint descriptor. A two-state flag to mark any constraint in the fragment, denoted C, has values of either $C = 1$ or 0 marking constrained/unconstrained fragments. C has a value of 0 only if all bonds in the fragment are single, rotatable and acyclic, and 1 otherwise.

4.1.4. Donor and acceptor chemical typing. Our previous work on intermolecular hydrogen bonds (Galek *et al.*, 2007) employed a descriptive functional group assignment using a library of 85 predefined fragments, however, its application in this case generated a prohibitively high number of unique categories. Here, the D and A environments are recorded using *SYBYL* atom types (Clark *et al.*, 1989). The description is based on element type and atom connectivity, *e.g.* O.3 represents an sp^3 hybridized O atom. This proved effective in discriminating separate types of donor/acceptor while maintaining a manageable number of discrete categories.

Table 5
Optimized model parameters: five-membered IHB logit model (details as in Table 4).

Descriptor	Category	Value	Odds ratio	Pr > χ^2	Lower bound (95%)	Upper bound (95%)
α		0.794	n.a.	0.002	0.280	1.308
β						
Donor	N.4	0.000	1.00	–	–	–
	N.am	−0.483	0.62	0.004	−0.810	−0.157
	N.pl3	−0.042	0.96	0.815	−0.393	0.309
	O.3	−0.857	0.42	< 0.0001	−1.190	−0.523
	N.3	0.612	1.84	0.002	0.230	0.994
	Other	−0.155	0.86	–	–	–
Acceptor	Other	0.000	1.00	–	–	–
	Cl	0.515	1.67	0.028	0.054	0.975
	F	−6.461	0.00	< 0.0001	−9.269	−3.653
	N.2	0.359	1.43	0.087	−0.052	0.770
	N.3	0.712	2.04	0.005	0.211	1.213
	O.3	−0.053	0.95	0.780	−0.426	0.319
	O.co2	0.325	1.38	0.215	−0.189	0.839
	S.3	−1.217	0.30	< 0.0001	−1.655	−0.779
	N.1	−4.457	0.01	0.002	−7.307	−1.607
	N.ar	1.525	4.60	< 0.0001	1.045	2.005
	O.2	0.197	1.22	0.368	−0.232	0.626
Path conjugation	S-S-S	0.000	1.00	–	–	–
	S-S-U	0.218	1.24	0.177	−0.099	0.536
	S-U-S	0.505	1.66	< 0.0001	0.311	0.700
	Other	0.822	2.28	0.028	0.091	1.553
Path constraint	True	0.000	1.00	–	–	–
	False	0.233	1.26	< 0.0001	−1.453	−1.204
Donor count		0.101	n.a.	< 0.0001	0.078	0.125

4.2. Statistical model assessment and validation

Rigorous statistical assessment ascertains the quality and applicability of the choice of model function. A range of metrics is used for this assessment, as discussed in previous work (Galek *et al.*, 2007; Galek, Fábián, Allen & Feeder, 2009). Some details are presented here for clarity. Assessment of model predictivity is achieved by way of an ROC curve (receiver operating characteristics). It plots the *sensitivity* (a fraction of correct positive predictions) and 1 − *specificity* (the fraction of correct negative predictions) over the range of potential propensity values. The diagonal is the outcome of a purely random model, whereas a curve above this marks a degree of predictivity which can be quantified by the area under that curve (AUC). AUC ranges from 0.5 for random outcomes to 1.0 for perfect prediction of all observations (AUC may be less than 0.5, but would indicate a worse model than purely random assignment). AUC > 0.8 is considered excellent and > 0.9 is outstanding (Hosmer & Lemeshow, 2000). AUC provides a universal, objective and non-parametric measure of predictivity, unlike related measures such as BedROC (Boltzmann-enhanced discrimination of ROC) and RIE (robust initial enhancement) popularized in virtual screening in the field of protein–ligand docking (see *e.g.* Nichols, 2008).

Checking the versatility of a model beyond the data provided for its own training is achieved by validation techniques. Hold-out validation is a variant in which the original set of true/false hydrogen-bond outcomes is split into two subsets, using a proportion to fit a propensity model, and reserving a proportion external to the training with which to test predicted outcomes. As for the full model, predictions for the validation set can be assessed using ROC curves. Optimal models should not suffer a significant decrease in predictivity during hold-out validation. For these studies hold-out validation has been applied using an approximate 75:25% ratio of training to validation set (specific values for each model are presented in the following section).

5. Hydrogen-bond propensity modelling results

A significant contrast in the bonding character of IHB ring motifs of various sizes has been observed. As a consequence, four propensity model functions have been developed with the aim to best predict potential IHB formation. First, a general IHB propensity model has been prepared that can predict all potential $R5$, $R6$ and $R7$ IHBs (Table 4), and secondly, indi-

Table 6
Optimized model parameters: six-membered IHB logit model (details as in Table 4).

Descriptor	Category	Value	Odds ratio	Pr > χ^2	Lower bound (95%)	Upper bound (95%)
α		−1.604	n.a.	< 0.0001	−1.809	−1.399
β						
Donor	O.3	0.000	1.00	–	–	–
	N.am	0.774	2.17	< 0.0001	0.603	0.946
	N.pl3	0.895	2.45	< 0.0001	0.659	1.130
	N.3	0.871	2.39	< 0.0001	0.439	1.303
	N.4	1.383	3.99	0.000	0.684	2.083
	Other	0.785	2.19	–	–	–
Acceptor	O.3	0.000	1.00	–	–	–
	O.2	0.454	1.57	0.061	−0.022	0.930
	N.2	0.473	1.60	0.002	0.168	0.779
	F	−6.526	0.00	< 0.0001	−9.322	−3.730
	N.3	1.497	4.47	< 0.0001	0.959	2.035
	N.1	−3.932	0.02	< 0.0001	−4.614	−3.250
	S.3	−1.872	0.15	< 0.0001	−2.295	−1.448
	N.ar	2.248	9.47	< 0.0001	1.575	2.921
	Cl	−1.918	0.15	< 0.0001	−2.345	−1.490
	O.co2	−0.350	0.70	0.394	−1.155	0.455
	Other	−1.108	0.33	< 0.0001	−1.658	−0.557
Path conjugation	S-S-S-S	0.000	1.00	–	–	–
	S-S-S-U	0.576	1.78	0.016	0.105	1.047
	S-U-S-U	2.490	12.06	< 0.0001	1.989	2.991
	S-S-U-S	−0.148	0.86	0.241	−0.395	0.100
	S-U-S-S	−0.162	0.85	0.358	−0.507	0.183
	U-S-U-S	3.066	21.46	0.033	0.254	5.878
	Other	1.053	2.87	0.024	0.137	1.970
Path constraint	True	0.000	1.00	–	–	–
	False	2.117	8.31	< 0.0001	1.935	2.300
Donor count		0.233	n.a.	< 0.0001	0.184	0.281

Table 7
Optimized model parameters: seven-membered IHB logit model (details as in Table 4).

Descriptor	Category	Value	Odds ratio	Pr > χ^2	Lower bound (95%)	Upper bound (95%)
α		−0.618	n.a.	0.121	−1.399	0.163
β						
Donor	Other	0.000	0.46	–	–	–
	N.am	−0.766	1.96	0.007	−1.324	−0.208
	N.pl3	0.672	0.59	0.031	0.061	1.283
	O.3	−0.523	1.00	0.058	−1.064	0.017
Acceptor	N.2	0.000	0.75	–	–	–
	O.3	−0.292	0.04	0.327	−0.877	0.292
	S.3	−3.285	0.85	< 0.0001	−4.537	−2.033
	O.2	−0.161	0.57	0.559	−0.702	0.380
	Other	−0.560	1.00	0.063	−1.150	0.030
Path constraint	False	0.000	18.32	–	–	–
	True	2.908	1.00	< 0.0001	2.658	3.159
Path conjugation	S-S-S-S-S	0.000	0.33	–	–	–
	S-S-S-S-U	−1.107	0.13	0.000	−1.704	−0.510
	S-S-S-U-S	−2.018	0.10	< 0.0001	−2.570	−1.465
	S-S-U-S-S	−2.263	0.22	< 0.0001	−2.697	−1.828
	S-S-U-S-U	−1.503	0.46	< 0.0001	−2.111	−0.896
	S-U-S-S-S	−0.775	0.44	0.012	−1.383	−0.168
	S-U-S-S-U	−0.831	0.08	0.017	−1.515	−0.147
	S-U-S-U-S	−2.510	0.45	< 0.0001	−2.907	−2.112
	Other	−0.789	0.46	0.065	−1.628	0.050

vidual models have been prepared for the separate motif sizes (Tables 5–7). Allowing model parameters to vary independently for each ring size can account for distinct influences from descriptors, and achieve improved predictivity. One can also include individual descriptors which may seem appropriate for one motif and not another. This exercise therefore allows the comparison of any gains over the general IHB model and also informs us about the influence of various descriptors toward each motif type. We note first that for the four models presented below, the fitting of the model parameters has worked very well. Parameter significance is described by the Pr > χ^2 statistic (Tables 4–7, column 4), with a high value indicating uncertainty or that the parameter is unnecessary. For each model all values are close to zero. Further statistical assessment of model fitting may be found in the supplementary material, Tables S1–4.

The generic IHB model (Table 4) is obtained using a categorical ring-size descriptor as a logit model parameter (with potential value 5, 6 or 7). Here, five common donor and 11 common acceptor types are identified, and a binary path constraint variable is included. The path conjugation and donor count descriptors were not effective during the optimization of this model, and were excluded. It is observed that model predictivity is most acceptable; AUC = 0.851 indicating ~ 85% probability that any known true IHB is preferred over any false IHB. More significantly, we observe virtually no change in the AUC when analysing a held-out subset of data for 5300 *D*–*A* pairs (of a total 21 396); AUC = 0.849 (see Fig. 8, further discussion below).

For each of the *R*5, *R*6 and *R*7 models, the types of true/false IHB observations and descriptor data vary. Table 8 lists the frequencies of various *SYBYL* atom types observed for the IHB ring types. For *R*5 and *R*6 motifs (Tables 5 and 6), there is an equivalent set of five common donor and ten common acceptor atom types. The frequency with which they are observed differs however, *e.g.* the N.pl3 donor is roughly twice as prevalent in the *R*6 data as it is in the *R*5 data (13.9:27.1%). The list of potential *R*7 IHB formers (Table 7) is much reduced compared with *R*5 and *R*6. Here, only three common donor and four common acceptor types are observed with sufficient frequency to form effective categorical parameters in the model. However, there are several atom types observed less frequently which can be categorized under the *other* type label (a total of 82 observations). Interestingly, there are very few such observations as donors when constructing the *R*5 and *R*6 models, meaning atom types other than the five common types listed are

Table 8
Frequencies of occurrence of *SYBYL* atom types in the *R*5, *R*6 and *R*7 model training data.

IHB Model		*R*5		*R*6		*R*7	
		Count	%	Count	%	Count	%
Donor *SYBYL* atom type	N.3	510	6.2	213	2.1	–	–
	N.4	229	2.8	73	0.7	–	–
	N.am	2620	32.0	2068	20.1	700	23.9
	N.pl3	1137	13.9	2785	27.1	270	9.2
	O.3	3688	45.1	5152	50.1	1871	64.0
	Other	–	–	–	–	82	2.8
Acceptor *SYBYL* atom type	Cl	250	3.1	127	1.2	–	–
	F	303	3.7	88	0.9	–	–
	N.1	56	0.7	151	1.5	–	–
	N.2	485	5.9	1224	11.9	119	4.1
	N.3	197	2.4	195	1.9	–	–
	N.ar	527	6.4	330	3.2	–	–
	O.2	2065	25.2	5696	55.3	1420	48.6
	O.3	3498	42.7	2169	21.1	1116	38.2
	O.co2	177	2.2	46	0.4	–	–
	S.3	460	5.6	164	1.6	56	1.9
	Other	166	2.0	101	1.0	212	7.3

Table 9
ROC analysis for the predicted models and subsets of training data used in statistical validation.

Model type	No. of training observations	*AUC*	No. of validation observations	*AUC*
Generic	21 396	0.851	5300	0.849
*R*5	8184	0.780	2000	0.766
*R*6	10 291	0.905	2500	0.909
*R*7	2923	0.826	700	0.820

exceptional as donors. Consequently, optimizing an *other* donor category is not practical due to the lack of data. Nonetheless, for completeness, a parameter to capture any rare donor type is desirable for future applications. As such, a coefficient value has been assigned for each *R*5 and *R*6 model which is simply the mean of the donor coefficient values in the model. The required values are β(other) = −0.410 for the generic model, β(other) = −0.155 for *R*5 and β(other) = 0.785 for *R*6.

The predictivity achieved for model training and validation can be compared for each model (Table 9). Of all IHBs, *R*6 motifs are the most common and are highly probable in certain structures, thus we might expect improvements over the general IHB model. Indeed, we find this more specific model achieves an improved AUC score = 0.905, and again performs well under the scrutiny of hold-out validation: AUC = 0.909 computed for 2500 *D*–*A* pairs from a total of 10 291 in the dataset (at the 95% confidence level the two AUC scores are essentially equivalent). This is very encouraging given that this value indicates we can predict with more than 90% confidence not only the likely interactions (*i.e.* the strong RAIHBS), but also the unlikely interactions (the weaker interactions we may wish to discount). The *R*5 model in general is not as predictive: the ROC curve indicates a predictivity of approximately 0.780, the lowest value of the four models. Recall, in our earlier investigations, the *R*5 IHBs behaved more irregularly and it was less clear that they had an influential energetic contribution toward the structures in which they were observed. The model performs well in hold-out validation, with an AUC of 0.766 for 2000 *D*–*A* pairs from 8184 in the complete dataset, which is a slight decrease from that for the complete model. This is perhaps the best we can achieve with the chosen set of descriptors for *R*5 IHBs. Finally, the AUC of the *R*7 model ROC curve is 0.826 which indicates it is also less predictable than *R*6, but the model is nonetheless most acceptable. Once again the difference between this and the AUC score for the validation set (0.820) is minimal (computed for 700 observations of a total of 2923 in the dataset).

Before we move on to some example predictions, we demonstrate how the optimized models reveal a quantitative influence on the propensity for IHB formation, through their molecular descriptors. That is, the methodology allows for *feedback*, which can be a valuable route to new scientific understanding. In crude terms it answers ‘what effect does this property have on the propensity for hydrogen bonds?’. A concise quantitative comparison of the influence of the model descriptors can be achieved through relative odds ratios (Hosmer & Lemeshow, 2000) using the parameter coefficients β. Comparing each coefficient to the parameter-assiged zero-coefficient magnitude, the so-called baseline variable, provides a simplification as the odds ratio is then the exponent of each coefficient [which follows from a rearrangement of equation (1) and returns the contribution of each coefficient to the logistic function]. Tables 4–7 also contain calculated odds ratios relative to the baseline parameter in each category. Significantly, poor donors and acceptors may be noticed, *e.g.* the F acceptor in the *R*6 model has a ratio of 0.0014 compared with the baseline O.3. Strong candidates may also be seen, *e.g.* N.3 has an odds ratio 4.462 relative to O.3. Inspection of Tables 4–7 reveals how these values nicely correlate with a classical understanding of donor and acceptor strength. Finally, differences in the models can be clearly seen using this statistic. The variation in influence of *D*—*A* path descriptors is clear, *e.g.* in the *R*6 model β(S-U-S-U) = 2.490 corresponding to a 12.061 odds ratio compared with β (S-S-S-S) = 0, whereas in the *R*5 model, β(S-U-S) = 0.505, an odds ratio of 1.656 compared with β(S-S-S). Such contrasts might be said to have a geometric origin, given that strain in an *R*5 IHB is likely irrespective of the nature of the connectivity linking *D* and *A*, whereas in the larger rings the presence or absence of flexibility in the *D*—*A* path has much greater influence on IHB potential.

5.1. IHB predictions

To further explore the method’s application, a variety of structures has been selected from the complete training set of 32 550 CSD structures to be assessed in detail using the fitted model functions presented above. All examples have a potential to form at least one of the *R*5, *R*6 or *R7* IHB types. We also illustrate some of the better and poorer predictions of the complete list of potential *D*–*A* interactions in the training set, which allows any strengths or weaknesses to be discussed. Table 10 displays a list of selected CSD structures together with propensity predictions for each potential IHB. Diverse examples have been selected of varying chemical functionality, and although this selection represents only a small subset of structures encountered with IHB potential, the quality of the predictions is representative of the method's accuracy as revealed from the training set statistics and validation exercises.

A general inspection of the selected predictions (Table 10) reveals that there can be a distinct separation of likely and unlikely IHBs, *i.e.* the propensity values cover the full 0–1 range. The last column verifies whether or not these predic-

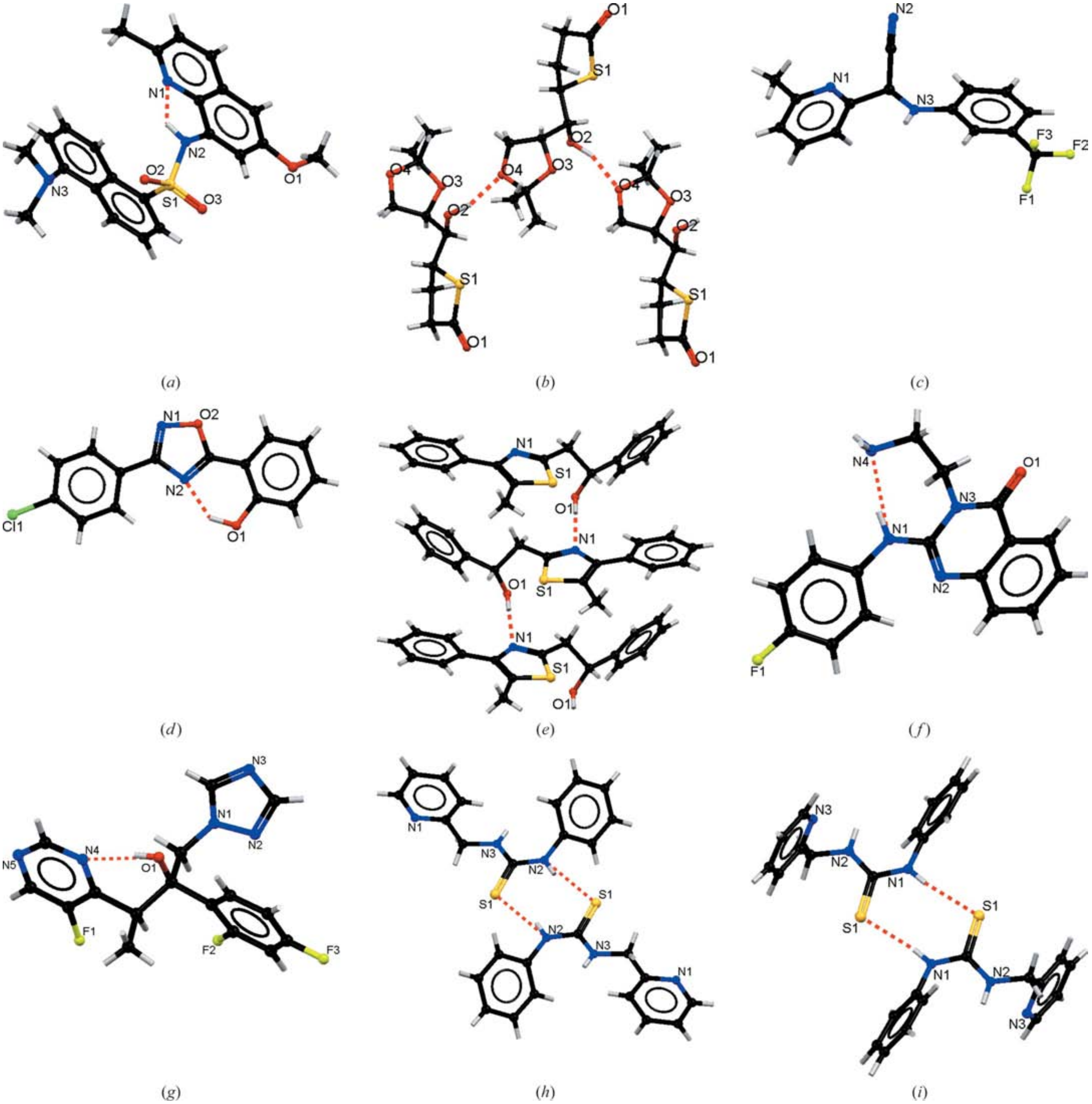

Figure 9
Selected CSD structures with potential for *R*5, *R*6 or *R*7 IHBs. (*a*) XULWUW, (*b*) NALDAF, (*c*) IFIROE, (*d*) SETGII, (*e*) EKEZUP, (*f*) IGAPIQ, (*g*) CEXMAU, (*h*) ACAFEQ and (*i*) ACAFEQ01. Hydrogen bonds are shown as dashed lines. Images produced using the *Mercury* program (Macrae *et al.*, 2008).

Table 10
IHB propensity predictions for selected CSD structures.

Potential IHB motif	Refcode	*D*—*A* label	π	IHB Observed?
*R*5	XULWUW	N2—N1	0.957	√
	NALDAF	O2—O3	0.207	×
		O2—S1	0.075	×
	IFIROE	N3—N2	0.009	×
		N3—N1	0.762	×
	ACAFEQ	N3—N1	0.674	×
	ACAFEQ01	N2—N3	0.674	×
*R*6	SETGII	O1—N2	0.976	√
		O1—O2	0.642	×
	EKEZUP	O1—S1	0.038	×
		O1—N1	0.420	×
	CEXMAU	O1—N4	0.810	√
		O1—N2	0.289	×
		O1—F2	0.026	×
*R*7	IGAPIQ	N1—N4	0.917	√
		N4—O1	0.735	×
	CEXMAU	O1—F1	0.307	×
	ACAFEQ	N2—N1	0.045	×
	ACAFEQ01	N1—N3	0.045	×

Table 11
Computation of the propensity for IHB formation between the N donor and N acceptor in XULWUW.

Refer also to equation (2) and Fig. 7(*a*). The null model descriptors for this *D*–*A* pair (those with $x = 0$) are not shown since they contribute zero to the sum.

Descriptor	Coefficient	x value	Contribution
Intercept, α	0.794	1	0.794
β			
Donor (N.pl3)	−0.042	1	−0.042
Acceptor (N.ar)	1.525	1	1.525
Path (S-U-U)	0.822	1	0.822
Constraint flag	0	1	0
Sum			3.099
Propensity			0.957

tions have been observed in the three-dimensional structure, indicating any predictive success. First we assess XULWUW (Kimber *et al.*, 2002; Fig. 9*a*) with a potential *R*5 IHB motif; the method shows this is very much expected ($\pi = 0.957$). Recall this value is determined through equation (2); for illustration, the steps in the computation are detailed in Table 11. The model picks up the constraints in the molecule around the N donor and N acceptor which assist correct geometrical arrangement. In NALDAF (Fava *et al.*, 1996; Fig. 9*b*) there are two potential IHBs: O—H···S is very unlikely, whereas O—H···O has a higher propensity to form, although is still not expected, and indeed, neither IHB is observed. IFIROE (Iovel *et al.*, 2001, Fig. 9*c*) represents a more problematic case. There are two potential *R*5 IHBs, the first, N3···N2, is correctly predicted not to form with almost zero propensity, however N3···N1 is likely according to the prediction ($\pi = 0.762$), but not observed either. We postulate in this case that the potential acceptor, a nitrile group, is incapable, owing to its fixed linear geometry, of forming the spatial arrangement required for an *R*5 IHB. Such subtleties are not accounted for in the model predictions. Including more delicate group-specific influences could further improve the methodology and may direct future research.

Turning to potential *R*6 IHBs, we see that despite their high average propensity to form, both likely and unlikely motifs can be predicted. SETGII (Ding *et al.*, 2006; Fig. 9*d*) displays the S-U-S-U path conjugation type and has no competing donors: its two potential IHBs could be expected to be highly probable. One of two hydroxyl(OH···N)oxadiazole is observed and is well predicted ($\pi = 0.976$), whereas the alternative hydroxyl(OH···O)oxadiazole bond is less likely ($\pi = 0.641$), owing to the weaker accepting ability of the sp^3 oxygen [recall from Table 6, $\beta(\mathrm{O.3}) = 0$ and $\beta(\mathrm{N.3}) = 1.497$, indicating the relatively greater propensity for the latter acceptor]. The second IHB is not observed: the interactions are mutually exclusive and the highest predicted IHB is that which is observed. Two IHBs are again possible in EKEZUP (Rybakov *et al.*, 2003; Fig. 9*e*), although neither are predicted to form, or are in fact observed [an intermolecular interaction is observed instead between hydroxyl(O—H···N)thiazole].

Finally, we turn to some potential *R*7 motifs. IGAPIQ (Yang & Wu, 2008; Fig. 9*f*) has a likely IHB involving a secondary and a primary amine ($\pi = 0.917$), which is seen. A second, less likely IHB ($\pi = 0.735$) is not realised in the crystal structure. Again, the potential IHBs are mutually exclusive, and the more likely is seen to take precedence over the less likely interaction. This structure, perhaps more than the other examples, demonstrates the practical use of the method with regard to the subtle differences in potential interactions. CEXMAU (Ravikumar *et al.*, 2007; Fig. 9*g*) can be considered one of the most successful cases. It has a potential *R*7 motif and three potential *R*6 IHBs. Only one of four possibilities is predicted as probable, *R*6-hydroxyl(OH—N)bipyridyl, $\pi = 0.810$. The remaining possible IHBs obtain $\pi < 0.31$, and the observed IHB is that predicted as most likely. Finally, ACAFEQ and ACAFEQ01 (Valdes-Martinez *et al.*, 2004; Ferrari *et al.*, 2007; Figs. 9*h* and *i*) are dimorphs of a compound that has two potential IHBs: *R*5 and *R*7 involving a choice of two separate N donors of thiourea and the N of pyridyl. The hydrogen bonding is in fact equivalent in the dimorphs (the structures differ by a twist of the pyridyl group of $\sim 180^\circ$). Neither IHB is formed in either dimorph: the *R*7 motif is neither expected ($\pi = 0.045$) nor observed and thus is an accurate prediction in both cases. The *R*5 IHB propensity is relatively high ($\pi = 0.674$) but is nonetheless absent. An intermolecular hydrogen bond is found involving the donor and acceptor sites of the would-be *R*5 IHB in both structures. Hence, in either case, the predicted *R*5 IHB is disfavoured over what can be considered as a strong intermolecular hydrogen bond. It is also noted that $\pi = 0.674$ is lower than other predicted propensities in the above examples. Still, the *R*5 model is formally incorrect in this case. The value of a predicted propensity (*i.e.* its closeness to 0 or 1) may well be a more important factor than whether it falls either side of 0.5 (or some other cut-off value), which would add prominence to the ROC metric for assessing model predictivity in these applications.

6. Conclusions

Our knowledge-based method to predict hydrogen-bond propensity has been applied specifically to intramolecular hydrogen bonds (IHBs). Statistical analyses on the CSD and *ab initio* energy minimizations were carried out in order to better understand IHBs and to categorize them based on general behaviour. Significantly, five-, six- and seven-membered hydrogen-bond motifs were found to contribute > 95% of all IHBs encountered in the CSD. The general trend of molecules in the CSD would therefore appear to contrast somewhat with larger biomolecules such as polypeptides which commonly exhibit *R*10 and *R*13 IHBs (α-turns and β-helices). It is perhaps the regularity of donor and acceptor groups along the chemical unit in such structures which affords that secondary structure, and which is not shared in general in small organic molecules. These might require a more specialized treatment given their individual behaviour (*e.g.* novel topological descriptors) which might provide interesting future study. We conclude that IHB motifs of size greater than 7 should not be considered individually owing to their rarity. They are much less likely to interfere with a good candidate for intermolecular hydrogen bonding.

*R*5 IHBs, from a geometric perspective, are found to be common ($\sim$ 21%), although in individual calculations a majority are not seen to have a true bonding character. Thus, these favourable geometries would seem to be directed by other influences such as packing forces favouring planar fragments (Brock & Minton, 1989). Some *R*5 IHBs are found to be true hydrogen bonds through a topological analysis of the electron density, although energy calculations suggest these motifs are rather weak (of the order of 5–20 kJ mol^{-1}). The ability for such fragments to form a bonding interaction seems largely to depend on the flexibility of the mediating covalent unit; for this small ring size, close approach of the *D* and *A* groups is not facile. For example, our CSD statistics reveal a mean $D-\mathrm{H}\cdots A$ angle of $\sim 110^\circ$, which is far from the ideal linear geometry. Of course, these influences affect all IHBs to some extent.

Having characterized influential descriptors on the formation of IHBs, four propensity models were developed: a generic IHB model and three further models tailored for individual IHB ring motifs of size 5, 6 and 7. The observed predictivity was excellent, especially for the six-membered IHB motifs (AUC $\simeq$ 91%). The success of our modelling demonstrates that with relatively primitive but well chosen two-dimensional topological descriptors, influences on IHB formation have been captured and can be accounted for predictively. These models will enhance the prediction of hydrogen bonding in general by allowing the potential interference of an IHB (especially with donor H) to be accounted for in an otherwise expected intermolecular hydrogen bond. A manuscript is currently in preparation detailing an example target, the antiallergic drug, Tranilast (Azuma *et al.*, 1976), which illustrates this potential.

We conclude by noting that the method could gain a potentially wider application, having been developed for generic organics. One avenue in which we would encourage exploration is the link between the occurrence of IHBs in the crystalline state and in solution. Substantial correlation in behaviour would direct the current theory to the prediction of solute–solvent interactions, of wide relevance, *e.g.* to the development of blood–brain barrier (BBB) permeability models. In particular, over-estimation of solute–solvent affinity can occur by an incorrect expectation of intermolecular hydrogen-bond formation, displaced by intramolecular hydrogen bonds in the solute (Hitchcock & Pennington, 2006). Other avenues open to exploration include the effect on molecular conformation and conformation prediction, and the development of knowledge-based scoring functions for protein–ligand docking, in which hydrogen-bond potential is a crucial component. We suggest that the simplicity of the models presented in this work will be a decisive factor in their ease of application and potential future utilization.

References

Aakeröy, C. B. (1997). *Acta Cryst.* B**53**, 569–586.

Addinsoft (2008). *XLSTAT*, Version 2008.6.05. Addinsoft USA, New York, USA; http://www.xlstat.com.

Agresti, A. (1990). *Categorical Data Analysis.* New York: Wiley.

Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.

Allen, F. H., Bird, C. M., Rowland, R. S. & Raithby, P. R. (1997). *Acta Cryst.* B**53**, 696–701.

Azuma, H., Banno, K. & Yoshimura, T. (1976). *Br. J. Pharmacol.* **58**, 483–488.

Bader, R. F. W. (1990). *Atoms in Molecules – A Quantum Theory.* Oxford University Press.

Bader, R. F. W. (1991). *Chem. Rev.* **91**, 893.

Bader, R. F. W. & Laidig, K. E. (1992). *J. Mol. Struct. Theochem*, **261**, 1–20.

Baughcum, S. L., Duerst, R. W., Rowe, W. F., Smith, Z. & Wilson, E. B. (1981). *J. Am. Chem. Soc.* **103**, 6296.

Bilton, C., Allen, F. H., Shields, G. P. & Howard, J. A. K. (2000). *Acta Cryst.* B**56**, 849–856.

Böhm, H.-J. & Klebe, G. (1996). *Angew. Chem. Int. Ed. Engl.* **35**, 2589–2614.

Brock, C. P. & Minton, R. P. (1989). *J. Am. Chem. Soc.* **111**, 4586–4593.

Bruno, I. J., Cole, J. C., Edgington, P. R., Kessler, M., Macrae, C. F., McCabe, P., Pearson, J. & Taylor, R. (2002). *Acta Cryst.* B**58**, 389–397.

Buemi, G. (2006). *Hydrogen Bonding, New Insights*, edited by S. J. Grabowski, pp. 51–107. Dordrecht, The Netherlands: Springer.

Buemi, G. & Zuccarello, F. (2004). *Chem. Phys.* **306**, 115–129.

Carpenter, J. E., Baker, J., Hehre, W. J. & Khan, S. D. (1980). *SPARTAN.* Wavefunction Inc., Irvine, CA, USA.

Cambridge Crystallographic Data Centre (2009). *ConQuest*, Version 1.11. CCDC, 12 Union Road, Cambridge, UK; http://www.ccdc.cam.ac.uk/products/csd_system/.

Clark, R. D., Cramer, R. D. & van Opdenbosch, N. (1989). *J. Comput. Chem.* **10**, 982–1012.

Chisholm, J. A. & Motherwell, S. (2005). *J. Appl. Cryst.* **38**, 228–231.

David, W. I. F., Shankland, K., van de Streek, J., Pidcock, E., Motherwell, W. D. S. & Cole, J. C. (2006). *J. Appl. Cryst.* **39**, 910–915.

Day, G. M. & Motherwell, W. D. S. (2006). *Cryst. Growth Des.* **6**, 1985–1990.

Desiraju, G. R. (1995). *Angew. Chem. Int. Ed. Engl.* **34**, 2311–2327.

Ding, W.-L., Shen, Y.-M., Xing, Z.-T., Wang, P.-L. & Wang, H.-B. (2006). *Acta Cryst.* E**62**, o5592–o5593.

Ellison, R. D., Johnson, C. K. & Levy, H. A. (1971). *Acta Cryst.* B**27**, 333–344.

Errede, L., Etter, M. C., Williams, R. C. & Darnauer, S. M. (1981). *J. Chem. Soc. Perkin Trans. 2*, p. 233.

Etter, M. C. (1991). *J. Phys. Chem.* **95**, 4601–4610.

Fava, G. G., Ferrari, M. B., Pelosi, G., Zanardi, F., Casiraghi, G. & Rassu, G. (1996). *J. Chem. Cryst.* **26**, 509.

Ferrari, M. B., Bisceglie, F., Cavalli, E., Pelosi, G., Tarasconi, P. & Verdolino, V. (2007). *Inorg. Chim. Acta*, **360**, 3233.

Galek, P. T. A., Fábián, L., Motherwell, W. D. S., Allen, F. H. & Feeder, N. (2007). *Acta Cryst.* B**63**, 768–782.

Galek, P. T. A., Fábián, L. & Allen, F. H. (2009). *Acta Cryst.* B**65**, 68–85.

Galek, P. T. A., Fábián, L., Allen, F. H. & Feeder, N. (2009). *CrystEngComm*, **11**, 2634–2639.

Gilli, P., Bertolasi, V., Ferretti, V. & Gilli, G. (1994). *J. Am. Chem. Soc.* **116**, 909.

Gilli, P., Bertolasi, V., Ferretti, V. & Gilli, G. (2000). *J. Am. Chem. Soc.* **122**, 10405.

Grabowski, S. J. (2001). *J. Mol. Struct.* **562**, 137–143.

Hargis, J. C., Evangelista, F. A., Ingels, J. B. & Schaefer III, H. F. (2008). *J. Am. Chem. Soc.* **130**, 17471–17478.

Hermida-Ramon, J. M. & Mosquera, R. A. (2006). *Chem. Phys.* **323**, 211–217.

Hitchcock, S. A. & Pennington, L. D. (2006). *J. Med. Chem.* **49**, 7559–7583.

Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression.* New York: Wiley.

Howard, D. L. & Kjaergaard, H. G. (2006). *J. Phys. Chem. A*, **110**, 10245–10250.

Iovel, I., Golomba, L., Belyakov, S., Kemme, A. & Lukevics, E. (2001). *Appl. Organomet. Chem.* **15**, 733.

Kassimi, N. E.-B., Archibong, E. F. & Thakkar, A. J. (2002). *J. Mol. Struct. Theochem*, **591**, 189–197.

Klein, R. A. (2002*a*). *Comput. Chem.* **23**, 585–599.

Klein, R. A. (2002*b*). *J. Am. Chem. Soc.* **124**, 13931–13937.

Kimber, M. C., Ward, A. D. & Tiekink, E. R. T. (2002). *Z. Kristallogr. New Cryst. Struct.* **217**, 349–350.

Kroon, J., Kanters, J. A., van Duijneveldt-van de Rijdt, J. G. C. M., van Duijneveldt, F. B. & Vleingenthart, J. A. (1975). *J. Mol. Struct.* **24**, 109–129.

Jeffrey, G. A. & Saenger, W. (1991). *Hydrogen Bonding in Biological Structures.* Berlin: Springer.

Macleod, N. A. & Simons, J. P. (2003). *Phys. Chem. Chem. Phys.* **5**, 1123–1129.

Macrae, C. F., Bruno, I. J., Chisholm, J. A., Edgington, P. R., McCabe, P., Pidcock, E., Rodriguez-Monge, L., Taylor, R., van de Streek, J. & Wood, P. A. (2008). *J. Appl. Cryst.* **41**, 466–470.

Mandado, M., Mosquerab, R. A. & van Alsenoy, C. (2006). *Tetrahedron*, **62**, 4243–4252.

Nichols, A. J. (2008). *J. Comput. Aided Mol. Des.* **22**, 239–255.

Pacios, L. F. & Gómez, P. C. (2001). *J. Comput. Chem.* **22**, 702–716.

Pijper, W. P. (1971). *Acta Cryst.* B**27**, 344–348.

Price, S. L. (2008). *Int. Rev. Phys. Chem.* **27**, 541–568.

Ravikumar, K., Sridhar, B., Prasad, K. D. & Bhujanga Rao, A. K. S. (2007). *Acta Cryst.* E**63**, o565–o567.

Roy, A. K., Hu, S. & Thakkar, A. J. (2005). *J. Chem. Phys.* **122**, 074313.

Rybakov, V. B., Liakina, A. Y., Popova, I. S., Formanovsky, A. A. & Aslanov, L. A. (2003). *Acta Cryst.* E**59**, o1293–o1295.

Valdes-Martinez, J., Hernandez-Ortega, S., Rubio, M., Li, D. T., Swearingen, J. K., Kaminsky, W., Kelman, D. R. & West, D. X. (2004). *J. Chem. Cryst.* **34**, 533–540.

Wavefunction, Inc. (2008). *SPARTAN*, Version. 1.2.0. Wavefunction Inc., Irvine, CA, USA; http://www.wavefun.com.

Wood, P. A., Allen, F. H. & Pidcock, E. (2009). *CrystEngComm*, **11**, 1563–1571.

Yang, X.-H. & Wu, M.-H. (2008). *Acta Cryst.* E**64**, o2240.