Journal Name

ARTICLE

Competitor Analysis of Functional Group H-bond Donor and Acceptor Properties Using the Cambridge Structural Database

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

James McKenzie and Christopher A. Hunter*

ABSTRACT: Intermolecular interactions found in the Cambridge Structural Database (CSD) are analysed as the outcomes of competitions between the different functional groups that are present in each structure: the most energetically favourable interactions are expected to win more often than weaker interactions. Tracking winners and losers through each crystal structure in the CSD provides data that can be analysed using paired comparison algorithms to rank functional group H-bonding properties based on how frequently they outcompete other functional groups in the crystal. This treatment is superior to simple statistical analyses of whether functional groups H-bond or not, because the distribution of H-bond donors and acceptors in the structures of the molecules found in the CSD is non-random. Most organic molecules contain more acceptors than donors, so that all H-bond donors are almost always H-bonded in all crystal structures, and most acceptors are not. The rankings of H-bond acceptors obtained by applying the TrueSkill paired comparison algorithm to the CSD agree well with the corresponding experimentally determined solution phase H-bond acceptor parameters β, but there is insufficient data to corroborate H-bond donor rankings calculated in the same way. The method is used to make predictions of the H-bond acceptor properties of functional groups for which solution phase measurements are not available.

Introduction

H-bonding is one of the most important non-covalent interactions in determining the properties of biopolymers, synthetic supramolecular assemblies and the organisation of organic molecules in the solid state. Prediction of the behaviour of such complex molecular ensembles requires a quantitative understanding of the magnitudes of the different interactions involved. We have developed a universal scale that describes all non-covalent interactions in terms of two functional group parameters, α, a H-bond donor parameter, and β, a H-bond acceptor parameter.[1] The free energy change associated with the pairwise interaction between two functional groups can be estimated as the product α x β (in kJ mol^{-1}). The α-β scales were originally derived from experimental measurements of the formation of 1:1 complexes in non-polar organic solvents, but the parameters can also be calculated from *ab initio* molecular electrostatic potential surfaces.[1,2] The α-β parameters have been shown to successfully predict the association constants for formation of H-bonded complexes in a variety of solvents and in solvent mixtures.[3] When the entire surface of a molecule is described by a complete set of α and β parameters, it is possible to accurately calculate solubilities and partition coefficients.[2,4] Although the α-β parameters were originally derived from solution phase measurements, we have shown that they can also be used to describe non-covalent interactions between surfaces and in the solid state.[5,6] Specifically, the α-β description of molecular interactions provides a method for assessing the relative stability of different solids based on calculation of functional group pairing energies. A powerful *in silico* cocrystal screening tool has been developed and validated based on this approach.[7]

The development of the α-β scale relied on experimental data to calibrate the *ab initio* calculations. A database of about one thousand experimentally determined solution phase association constants provided information about the H-bonding properties of a range of functional groups, but this information was limited to simple compounds that contained a single functional group, dissolved in non-polar organic solvents, and were polar enough to form stable complexes.[2] In the search for additional information to improve calibration of the α-β scale, we turned to the Cambridge Structural Database (CSD), which represents a huge repository of experimental data on non-covalent interactions.

The Cambridge Structural Database (CSD) has been used extensively to study the properties of H-bonding interactions.[8] The propensity for a functional group to be involved in a H-bond can be estimated from the frequency of occurrence of H-bonds involving specific donors and acceptors in the solid state, and these parameters have been used to rank the probabilities of formation of different H-bond combinations

[a] *Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW (UK). Email: herchelsmith.orgchem@ch.cam.ac.uk*

Electronic supplementary information (ESI) available: Functional group definitions, SMARTS strings and solution phase H-bond parameters (PDF). See DOI: 10.1039/x0xx00000x

for the purposes of crystal engineering and protein ligand design. Mills first calculated H-bond donor and acceptor properties as the ratio of the number of times a functional group was observed to form a H-bond, N_{obs}, and the total number of times the functional group was found in the CSD, N_{total} (Equation 1).[9] This metric (P) is the probability that a specific functional group will form a H-bond in the solid state and is related to the average strength of H-bonding interactions involving that functional group. H-bond probabilities have been used to compare the properties of thioether and ether acceptors,[10] different acceptors in a range of competitive environments,[11,12] intermolecular[13] and intramolecular ring motifs,[14] and for the rational design of cocrystals.[15] A more comprehensive statistical analysis has been developed using logistic regression to survey all H-bond arrangements found in the CSD to calculate logit H-bonding propensities for pairwise combinations of functional groups in a crystal.[16] Taylor has proposed a related approach that normalises the H-bond probability based on the exposed surface areas of the atoms involved.[17]

$$P = \frac{N_{obs}}{N_{total}} \qquad \text{Eq. 1}$$

There have been some attempts to compare solid state H-bond probability values with solution phase free energies of H-bond formation. Mills compared solid state H-bond acceptor and donor probabilities calculated using Equation 1 with solution phase H-bond parameters derived from 1:1 association constants measured in trichloroethane.[9,18] Although qualitative agreement between the two sets of parameters was found, no attempt was made to establish a formal relationship due to the large range spanned by the solution phase parameters for individual functional groups. Ziao made a similar comparison of solid state H-bond acceptor probabilities with the solution phase pK_{HB} scale for a series of aminonitriles.[19] We have shown that solution phase H-bond parameters can be used to predict H-bonding outcomes in crystal structures of simple compounds, where two different H-bond acceptors compete for a single H-bond donor.[6] Here, we describe a more comprehensive comparison of the occurrence of H-bonds in the CSD with free energy parameters that quantify the relative strengths of solution phase H-bonding interactions. We show that H-bond probabilities based on Equation 1 fail to accurately describe the H-bond properties of functional groups, because the effects of competition with other H-bond donors and acceptors varies from one crystal structure to the next. We show that the effect of competing functional groups can be assessed using paired comparison algorithms developed for ranking competitors in multi-player games, and this new approach provides accurate solid state H-bond acceptor parameters that agree well with solution phase measurements.

Approach

For a set of functional groups for which solution phase H-bond parameters have been experimentally determined, two different methods were used to calculate solid state H-bond parameters using the CSD. The first method is based on functional group H-bond probabilities calculated using a modified version of Equation 1.[9] The idea is that stronger H-bond acceptors and donors should more frequently be observed to H-bond in crystal structures, resulting in larger probability values. This work is an update on previous literature studies. The CSD is now significantly larger in size (> 850,000 crystal structure), which reduces the sampling error associated with the calculation of probability values, and the larger dataset allows functional groups to be defined more specifically. The second point is particularly important, since solution phase measurements show that H-bond properties vary significantly for the same functional group in different molecular contexts.

The second approach uses a class of algorithms developed for the mathematical problem of paired comparison.[20-24] Paired comparison is widely used for ranking competitors in games such as chess, tennis and more recently, online multi-player computer games. The algorithms use the outcome of pairwise competitions to assign ratings to competitors. The expected outcome of a competition is computed as a function of the difference in believed abilities of the two competitors, and after each game, the actual outcome is used to update the ratings of the two competitors. To apply these algorithms to the problem of ranking the H-bonding properties of functional groups in the solid state, we consider each crystal structure in the CSD as the result of a competition between all of the functional groups present for the formation of H-bonds. Any functional group that is found to form a H-bond is considered a winner and those that do not are considered losers. This analysis means that the CSD holds a large amount of pairwise competition data that can be used to develop quantitative scales of H-bond donor and acceptor ratings.

Figure 1a illustrates the paired comparison approach using the crystal structure with CCDC reference code, AFUMAR. AFUMAR contains three H-bond acceptors (imidazole, nitrile and ester) and one H-bond donor (imidazole). Inspection of the close contacts in AFUMAR shows that the imidazoles form H-bonded chains. The other H-bond acceptors all form close contacts with CH groups in the crystal structure. Although contacts between H-bond acceptors and CH groups may be considered a form of H-bonding,[25] for the purposes of this work, we assume that these interactions are significantly weaker than conventional H-bonds. Therefore the H-bonding outcomes in AFUMAR give the results of two distinct competitions: imidazole versus nitrile, and imidazole versus ester. The imidazole acceptor was the winner in both competitions, because it H-bonded to the only good donor, whilst the other acceptors were losers, because they interact with CH groups.

Figure 1b shows a more complicated example, the crystal structure with CCDC reference code FIQNUP. In this crystal structure, there are two H-bond donors (carbamate and alcohol) and two H-bond acceptors (carbamate and alcohol). There are two types of H-bond in the crystal structure: chains of carbamate–carbamate H-bonds, and chains of alcohol–

alcohol H-bonds. In this case, some prior knowledge is required to assign winners and losers. For example, if we assume that the carbamate as a better H-bond acceptor than the alcohol, FIQNUP gives the result of the competition between a carbamate donor and an alcohol donor for a carbamate acceptor. The carbamate donor was the winner, because it H-bonds to the best acceptor, and the alcohol donor is the loser, so it ends up with the weaker H-bond acceptor, the alcohol. Using this logic over the entire set of crystal structures in the CSD gives outcomes for a large number of competitions, and these data can be used in paired comparison algorithms to determine ratings for the H-bonding properties of the functional groups in the CSD.

Methods

Functional Group Definitions and Solution Phase H-Bond Parameters.

The CCDC application program interface (API)[26] allows the user to search the CSD using a string representation of a molecular fragment (SMARTS).[27] CSD searches and analyses of crystal structures were performed using the CSD Python API, version 1.3.0 and the Python programming language, version 2.7.[28] A library of functional group SMARTS strings was written. Experimental values of α and β were used to calculate average values and standard deviations of solution phase H-bond parameters for each functional group.[2] In cases where experimental α and β values varied due to some obvious feature in a family of molecules, the functional group was subdivided into different categories, and a specific SMARTS string was written for each. For example, Table 1 shows the SMARTS strings for a primary aliphatic amine and a primary benzyl amine, as well as the available primary amine containing compounds with experimentally determined β values. It is clear that primary benzyl amines have a reduced β value compared with primary aliphatic amines, and these two classes were therefore considered as separate functional groups. The full list of SMARTS strings and experimental data used to calculate average functional group α and β values are provided in the ESI.

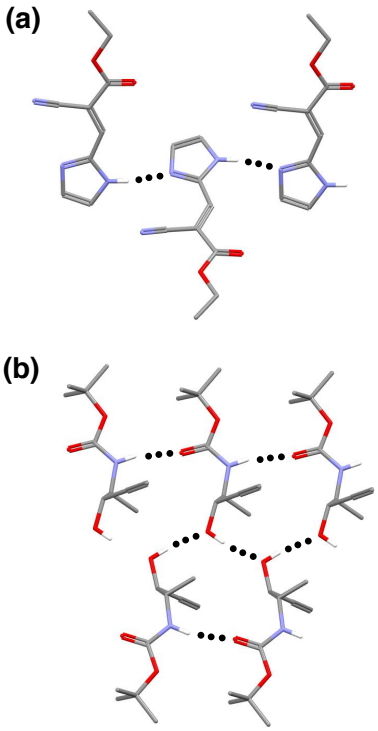

Figure 1. Sections of the crystal structures with CCDC reference codes (a) AFUMAR and (b) FIQNUP. Dotted lines are H-bonds.

Table 1. Primary amine SMARTS string definitions and solution phase β values.

Functional Group SMARTS string	Compounds	β	Average β (sd)
primary aliphatic amine	adamantan-1-amine	8.19	7.99 (0.16)
N([H])([H])C([H])([H])[CX4]	cyclohexylamine	8.15	
	n-octadecylamine	8.14	
	n-octylamine	8.11	
	t-butylamine	8.08	
	n-hexadecylamine	8.08	
	i-propylamine	7.99	
	n-butylamine	7.97	
	ethylamine	7.88	
	methylamine	7.84	
	n-hexylamine	7.73	
	n-heptylamine	7.73	
primary benzyl amine	3-methylbenzylamine	7.44	7.34 (0.14)
N([H])([H])C([H])([H])c	benzylamine	7.24	

Calculation of Solid State H-Bond Probabilities.

Solid state H-bond acceptor and donor probabilities (P_i) were calculated using Equation 2 for functional groups that had more than 50 occurrences in the CSD.

$$P_i = \frac{N_{i,obs}}{S\, N_{i,total}} \qquad \text{Eq .2}$$

where i defines the functional group, $N_{i,obs}$ is the number of times the functional group was observed to form a H-bond in the CSD, $N_{i,total}$ is the number of times the functional group was found in the CSD, and S is a normalisation factor accounting for the number of H-bond donor sites present in the functional group. For example, primary aniline has two donor hydrogens and therefore is twice as likely to form a H-bond in a crystal structure, so S = 2.

To calculate H-bond probabilities, the SMARTS strings contained in the SMARTS library were organised into two sets: those that contain acceptor functional groups, and those that contain donor functional groups. SMARTS strings of functional groups that are both acceptors and donors appeared in both sets. Each SMARTS string was used in conjunction with the CSD Python API to search the CSD for structures containing the molecular substructure defined by the SMARTS string. Crystal structures that satisfied the following criteria were accepted for use in the experiment:

1. 3D atomic coordinates
2. R factor < 7.5%
3. No disorder
4. No errors
5. Contain no metals
6. No duplicate entries
7. No missing hydrogen atoms
8. Contain no intramolecular H-bonds
9. At least one donor and one acceptor functional group

Criteria 1-6 were specified using the settings attribute of a substructure search, whereas criteria 7-9 were enforced manually post search. For structures containing missing hydrogen atoms, i.e. structures without explicit coordinates for some or all hydrogen atoms in the atomic coordinates file, the hydrogen atoms were added using the 'add_hydrogens' method.[29]

Each structure was screened for intramolecular H-bonds using the 'hbonds' method, which allows a user to search a crystal structure for either intramolecular or intermolecular close contacts with user defined H-bond definitions. An intramolecular H-bond was defined as a close contact between a donor, DH, and an acceptor, A, where D, A ≠ C, H and where the donor heavy atom, D, and the acceptor are separated by at least three atoms in the molecular structure. Additionally, the distance between D and A was required to be less than the sum of the van der Waals radii plus 0.5 Å, and the D-H•A angle was required to be greater than 100°. These rather relaxed criteria were chosen following a series of manual tests on structures known to have intramolecular H-bonds, where the distance and angle criteria were changed until all intramolecular H-bonds were detected by the API in these structures. Crystal structures found to contain intramolecular H-bonds were then removed from the experiment. The exclusion of structures containing intramolecular H-bonds is based on previous observations in the literature that, when possible, intramolecular interactions form with high propensity in crystal structures.[11,30]

Crystal structures containing at least one good donor and one good acceptor were used in the experiment. This criterion removes the possibility of retrieving structures with no possibility of H-bond formation (of which there are numerous examples in the CSD). For searches involving a SMARTS string that was an acceptor only, an additional substructure search was performed over each retrieved crystal structure, using an XH substructure, where X = O, N, F, S, P, Cl, Br, I. If no XH substructure was found in a crystal structure then it was removed from the experiment.

String matching using CCDC atom labels from the mol2 file was used to identify whether a functional group had formed a H-bond in the crystal structure. The idea is to use the atom labels of atoms matched to a SMARTS string and the atom labels matched to a H-bond to deduce whether an acceptor or donor has formed a H-bond. To achieve this, the CCDC atom labels corresponding to the substructure(s) that had matched the SMARTS string were obtained using the 'match_atoms' method of a hit object. The 'hbonds' method of a crystal object was then used to identify intermolecular H-bonds in the crystal structure. Intermolecular H-bonds were defined as any close contact between a donor, DH, and an acceptor, A, where D, A ≠ C, H and where DH and A are in different molecules in the crystal structure. Additionally, the distance between D and A was required to be less than the sum of the van der Waals radii plus 0.1 Å, and the D-H•A angle was required to be greater than 120°. The CCDC atom labels of the atoms involved in each H-bond were retrieved and then cross-referenced against the atom labels matched to the substructure. For every substructure detected (which can be more than one per crystal structure) the value of $N_{i,total}$ was increased by one, and if a substructure was found to be involved in a H-bond then $N_{i,obs}$ was increased by one.

The process is exemplified in Figure 2 using the crystal structure with CCDC reference, ABABAI. ABABAI contains two potential H-bonding functional groups, a pyridine and a primary alcohol highlighted in blue and red in Figure 2(b), and was retrieved from the CSD in three separate SMARTS substructure searches: searches for pyridine acceptor, for primary alcohol acceptor and for primary alcohol donor. For each search the 'match_atoms' method returned the CCDC atom labels of the substructures that matched the SMARTS string: C13, C18, H5, H6, H11, H12, O1, H19 for the primary alcohol, and C1, C2, C3, C4, C7, N1 for the pyridine. In the crystal structure of ABABAI, there is only one H-bond which matched the definition used in this experiment and that was between the pyridine acceptor and the primary alcohol donor (Figure 2(a)). The atom labels of the atoms involved in the H-bond were retrieved: O1, H19, N1.

In the case when ABABAI was retrieved from the CSD using the pyridine SMARTS string, it can be deduced that the pyridine was involved in a H-bond as an acceptor, because the nitrogen

with CCDC label N1 was identified as one of the atoms in a H-bond as well as in the pyridine substructure. This crystal structure therefore increases $N_{i,obs}$ and $N_{i,total}$ by one for the pyridine acceptor.

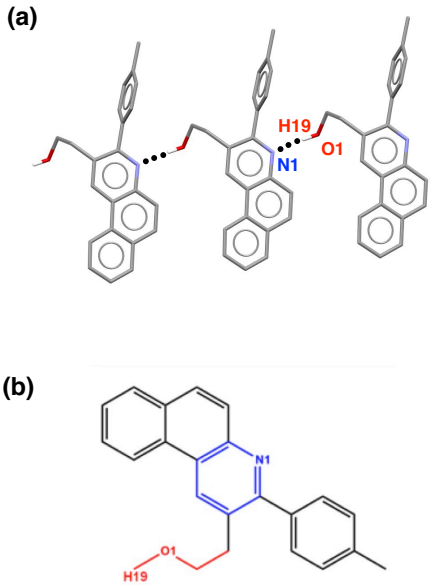

Figure 2. a) A section of the crystal structure with CCDC reference code ABABAI. Intermolecular H-bonds are shown as dotted lines. b) Functional groups detected using the SMARTS library: pyridine in blue, and primary alcohol in red.

In the case when ABABAI was retrieved from the CSD using the primary alcohol SMARTS string from the list of acceptor SMARTS strings, it can be deduced that the primary alcohol was not used as an acceptor. Although the primary alcohol oxygen atom label, O1, is matched to an atom label in the H-bond list, the primary alcohol hydrogen, H19, is also matched. It is therefore possible that the primary alcohol oxygen could be the donor heavy atom in the H-bond and not the acceptor. In this situation, the third atom in the H-bond is checked. If the third atom also came from the primary alcohol then the primary alcohol would be the acceptor, but in this case, the third atom label is from another functional group so the primary alcohol was not the acceptor. This structure therefore increases the value $N_{i,total}$ by one for a primary alcohol acceptor, but $N_{i,obs}$ remains the same.

In the case when ABABAI was retrieved from the CSD using the primary alcohol from the list of donor SMARTS strings, it can be deduced that the primary alcohol donor was used in H-bonding, because the hydrogen atom found in the H-bond labels list matches a hydrogen from the primary alcohol substructure. This structure therefore increases $N_{i,obs}$ and $N_{i,total}$ by one for the primary alcohol donor. The rules used to deduce whether a functional group has been used in a H-bond, as either a donor or acceptor, in a crystal structure where coded in a Python script to allow the process to be fully automated.

Calculation of Solid State H-Bond Ratings

The use of paired comparison algorithms requires that the data in the CSD is first transformed into the results of competitions between H-bond donors and acceptors. The competition data needed for the calculation of H-bond acceptor and donor ratings use crystal structures with slightly different requirements. For acceptors, all crystal structures found in the CSD that meet the nine requirements described above for H-bond probabilities were used. The results of acceptor competitions were evaluated based on which acceptors form a H-bond in the crystal structure (winners) and which do not (losers). For donors, the results of competitions were evaluated based on which acceptors the donors interact with: one donor is considered to have out-competed another if it H-bonds to an acceptor with a higher rating. This analysis requires that every H-bonding functional group in the crystal structure must be present in the SMARTS library, because an unidentified H-bond acceptor will not have a rating that can be used to evaluate the result of the competition.

H-bond acceptor competition data were gathered using the CSD Python API to retrieve crystal structures from the CSD containing the substructure defined by each acceptor SMARTS string in the SMARTS library. In each crystal structure, the atom labels of the intermolecular H-bonds were retrieved using the 'hbonds' method of a crystal object using the intermolecular H-bond definition described above for H-bond probabilities. A further substructure search was then performed to identify the atom labels of any other acceptor functional groups contained in the crystal structure. The same procedure described above for H-bond probabilities was then used to deduce which acceptors had formed H-bonds and which had not.

The data for cyclic secondary amines were removed in this experiment, because in many crystal structures, the nitrogen is in a sterically hindered environment. There are bulky substituents on the adjacent carbon atoms that prevent H-bonding interactions with the nitrogen acceptor in these crystal structures. The average solution phase β value for the cyclic secondary amine functional group is derived from three compounds, pyrrolidine, azetidine and piperidine, none of which have substituents on these carbon atoms. The SMARTS string definition of secondary cyclic amine should therefore require that the carbon atoms attached to nitrogen only have hydrogen substituents, but this constraint resulted in too few competitions for the calculation of a rating.

H-bond donor competition data were gathered by constructing a database of crystal structures where every functional group and H-bonding outcome could be identified. Unidentified functional groups were defined as a collection of atoms that were not defined by a SMARTS string in the SMARTS library but have the potential to form H-bonds (*i.e.* any substructure containing combinations of the elements, B, O, N, F, Al, Si, P, Cl, Br). The database was constructed by searching the CSD for crystal structures using a donor SMARTS string from the SMARTS library and then performing further substructure searches on each crystal structure. To remove crystal structures containing unknown functional groups, the 'heavy_atoms' attribute of a crystal hit object was used to retrieve a list of the non-hydrogen atoms contained in the crystal structure. The list of heavy atoms was then used to find

the total number of non-hydrogen and non-carbon atoms in a structure, and these numbers were checked against the expectation based on the detected functional groups.

For example, in the crystal structure with CCDC reference code, ABABAI, shown in Figure 2, the 'heavy_atoms' property of the crystal object shows that the molecule in the structure contains one oxygen and one nitrogen atom. A SMARTS substructure search on this molecule with the SMARTS strings library detected a primary alcohol substructure, which is known to contain one oxygen atom, and a pyridine substructure, which is known to contain one nitrogen atom. Since the heavy atoms in this crystal structure are in agreement with the detected functional groups, this crystal structure is allowed for use in this experiment. Any structure where the heavy atom counts did not match expectation based on the atoms contained in matched substructures was removed. The H-bond atom labels for each H-bond and functional group atom labels were then used to deduce which functional groups had formed H-bonds in each crystal structure, using the string matching process described above for H-bond probabilities. For each crystal structure, a Python script was used to assess which H-bond donors had out-competed which using a lookup table of acceptor ratings.

The TrueSkill algorithm developed by Microsoft was used in the publically available Python implementation.[31] Trueskill is based on the Elo ranking system used in chess.[21] In Elo, players A and B are initially assigned ratings R_A and R_B, where typical values span from 0-2500. When A and B compete, the expected outcome for player A, E_A, is given by Equation 3.

$$E_A = \frac{1}{1+10^{\frac{R_B - R_A}{400}}} \qquad \text{Eq. 3}$$

Following the match, ratings are updated using Equation 4, where the new rating of player A, R'_A, is the sum of the original rating and an update term.

$$R'_A = R_A + K(S_A - E_A) \qquad \text{Eq. 4}$$

where S_A = 0 for a loss, 1 for a win and 0.5 for a draw, and K is a scaling factor which can be changed depending on the game. The difference between the expected outcome and the actual game outcome determines by how much the ratings change, and the amount that is added to the winners rating is the same as the amount subtracted from the losers rating. Glickmann improved the Elo algorithm by assigning ratings a probability given by a normal distribution, $N(\mu, \sigma)$, with mean μ and standard deviation σ.[22] In this Glicko algorithm, updates in ratings are a function of the σ values of the two players. This allows for a greater update to be applied for a player who had a larger uncertainty in their rating compared with a player whose ability was well-established. The update algorithm uses Bayesian inference, whereby the prior probability distribution functions of the two players is multiplied by a likelihood, which quantifies how likely the result was given the two players ratings, and results in a posterior probability distribution functions for the two players. TrueSkill is also based on a Bayesian ranking system, but was developed to specifically handle team games, where the skill of a team is governed by the skills of the individual players.[23,24] Following the result of a team game, the update algorithm uses factor graphs to allow for efficient computation of individual player's posterior skills.

The Trueskill algorithm was implemented by assigning an initial default rating value and uncertainty to each functional group. The results of a competition are then passed to the algorithm and new ratings for the competing groups are calculated. This process is repeated iteratively until all the results of all possible competitions have been used. The rating can be tracked as a function of the number of competitions to monitor convergence of the final value.

Results and Discussion

Solid State H-bond Probabilities

Figure 3 shows the relationship between the solid state H-bond probabilities for H-bond donor and acceptor functional groups, P_D and P_A, and the corresponding solution phase H-bond parameters, α and β. Although the correlation coefficient is not very high, there is a clear correlation between the two H-bond acceptor parameters, P_A and β (Figure 3(b)), which shows that strong acceptors are more likely to form H-bonds than weak acceptors in a crystal structure. In contrast, Figure 3(a) shows that there is no relationship between the two H-bond donor parameters, P_D and α. The majority of donors have P_D values close to one, which means that almost every instance of a donor functional group in the CSD is involved in a H-bond.

There are two donor functional groups with P_D values significantly lower than one: secondary *N*-aryl aniline and primary aniline have P_D values of 0.70 and 0.79 respectively. Crystal structures where these functional groups where not detected to form a H-bond were inspected and were found to fall into one of two categories: the donor hydrogen atom is in a sterically hindered environment and not accessible to H-bond acceptors, or the donor forms a H-bond to a π-system. In this experiment, close contacts between a hydrogen and a carbon atom were not considered as H-bonds, which means that H-bonds to π-systems were excluded.

Some H-bond donors have P_D values that are greater than one. Crystal structures that were detected to have a greater number of H-bonds involving a donor functional group than the number of occurrences of that functional group were visually inspected. Figure 4 shows an example. The hydrogen of the alcohol donor sits between two ether oxygen atoms with similar OH•O distances (3.05 Å and 3.02 Å), so that both contacts fall within the definition of a H-bond used here.[32] Thus the value of P_D can exceed one due to the presence of these three-centre bifurcated H-bonding interactions.

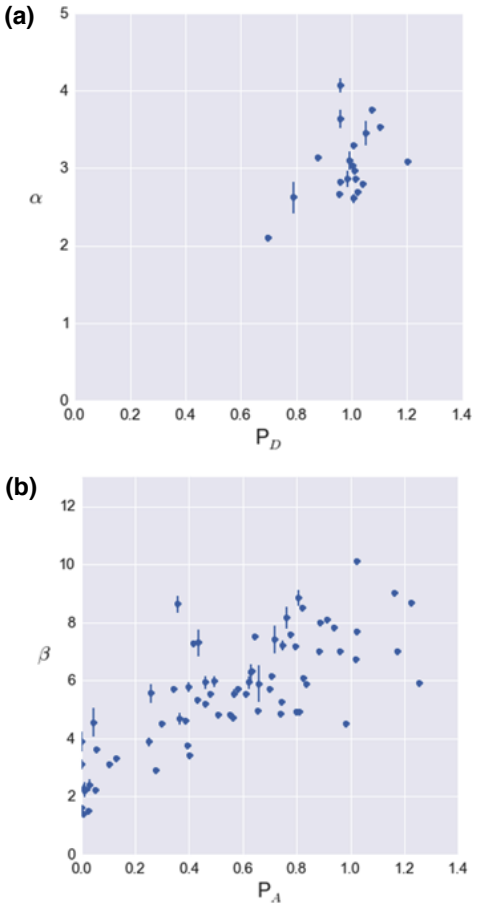

Figure 3. The relationship between (a) P_D and α (r^2 = 0.22) and (b) P_A and β (r^2 = 0.58).

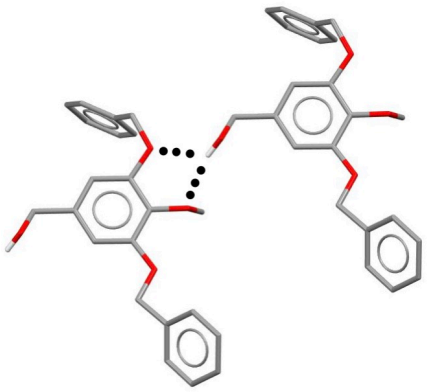

Figure 4. A section of the crystal structure with CCDC reference code RIPVUI. Intermolecular H-bonds are shown as dotted lines. The hydroxyl donor forms a bifurcated H-bond with two aryl ether acceptors.

The number of data points plotted in Figure 3(a) is relatively small, due to the limited number of donor functional groups that had both experimentally determined α values and more than 50 hits in the CSD. However, there are other donor functional groups for which it is possible to determine P_D values, and these results are listed in Table 2. In all cases, the P_D values are close to one. The results in Figure 3(a) and Table 2 show that it is not possible to use donor H-bond probabilities to rank donor functional groups. The reason is that H-bond acceptors are typically in excess in organic molecules, so that donor functional groups are almost always H-bonded and all have approximately the same P_D value of one.

Table 2. Solid state H-bond donor probabilities.

Functional Group	SMARTS string	P_D
N-aryl benzamide	C(=[OX1])(c)N([H])c	0.99
N-alkyl benzamide	C(=[OX1])(c)N([H])[CX4]	0.98
imidazole	N1(C(=[NX2]C(=C1[!F!Cl!Br!!O!N])[!F!Cl!Br!!O!N])[!F!Cl!Br!!O!N])[H]	1.05
α,β-unsaturated carboxylic acid	C(=[OX1])(O[H])C=C	1.04
primary sulphonamide	S(=[OX1])(=[OX1])(N([H])[H])([cC])	1.06
primary amide	C(=[OX1])(N([H])[H])[Cc]	1.01
sulphinamide	[SX3](=[OX1])(N([H])[cC])[cC]	0.83
alkylidene sulphonohydrazide	S(=[OX1])([Cc])(N([NX2]=C)[H])=O	0.95
primary *m*-haloaniline	c:1(:c(:c(:c(:c(:c:1[!F!Cl!Br!!O!N])[!F!Cl!Br!!O!N] N([H])[H])[!F!Cl!Br!!O!N])[FClBrI])[!F!Cl!Br!!O!N]	1.25
primary carbamate	C(=[OX1])(N([H])[H])O[cC]	1.01
alkylidene hydrazide	C(=[OX1])([cC])N(N=C)[H]	1.04
N,N'-disubstituted thiourea	C(=[SX1])(N([cC])[H])N([cC])[H]	0.96
imide	N(C(=[OX1])[!F!Br!Cl!!!O!N])([H])C(=[OX1])[!F!Br!Cl!!!O!N]	1.02
benzimidazole	c:1:c2:c(:c:c:c:1)[NX2]=CN2[H]	1.04

Solid State H-Bond Acceptor Ratings

The results above indicate that H-bond acceptors compete for a limited supply of H-bond donors in crystals of organic compounds. It is therefore possible to apply paired comparison algorithms to rank H-bond acceptor properties in a straightforward manner without any prior knowledge or ranking of H-bond donor properties: any acceptor that makes a H-bond with a donor is considered a winner and any acceptor that does not is a loser. This strategy cannot be used to rank donors, because there would be no losers, so a different treatment is required as explained below. Solid state H-bond acceptor ratings (R_A) were calculated using the TrueSkill algorithm, and Figure 5 shows the relationship with the solution phase H-bond acceptor parameter β. The correlation between the two parameters is significantly better than the correlation found for P_A with β (r^2 is 0.88 for Figure 5

compared with 0.58 for Figure 3(b)). The line of best fit in Figure 5 therefore provides a useful solid state method for estimating β values for functional groups for which solution phase data is not available (Equation 5).

$$\beta = 0.2910\, R_A - 1.532 \qquad \text{Eq. 5}$$

In the calculation of the solid state ratings, three functional groups failed to converge: alkyl fluoride, aryl fluoride and tertiary phosphine. Figure 6 shows the behaviour of the rating value R_A as a function of number of competitions (n) used to determine the rating. The behaviour of the organic fluorides is compared with other organic halides. For the other organic halides, the rating rapidly decreases and then oscillates around a stable value, as the acceptor wins or loses competitions with other acceptors. The ratings plotted in Figure 5 are the final rating values reached at the right-hand side of Figure 6 after all possible competitions have been evaluated. The behaviour is different for alkyl fluoride and aryl fluoride, where the ratings decrease steadily with the number of competitions and do not converge to stable values. The reason is that organic fluorides are never observed to out-compete another acceptor in the set of crystal structures used in this experiment,[33,34] so there is no reference point to establish how low the rating values should go: the ratings for organic fluorides are simply less than of all the other ratings.

The failure of organic fluorides to converge is consistent with the solution phase H-bond parameters β for these functional groups, which indicate that they are exceptionally poor H-bond acceptors. However, the failure of the tertiary phosphine rating to converge is not consistent with the solution phase β value, which falls in the middle of the H-bond acceptor scale. None of the crystal structures used in this experiment contained a tertiary phosphine that had out-competed another acceptor for a H-bond, so it is not possible to assign a rating to this functional group: it is a very poor H-bond acceptor. A search of the CSD was performed to check the frequency of occurrence of H-bonds involving tertiary phosphines. There are 619 crystal structures that contain PX_3 (where X is aliphatic or aromatic carbon) and at least one YH donor (where Y is O, N, F, S, Cl, Br, I). The tertiary phosphine accepted a H-bond in only 35 cases (6%). For comparison, the same analysis was performed for alkyl nitrile, which has a similar solution phase β value to tertiary phosphine, and the alkyl nitrile accepted a H-bond in 38% of the crystal structures. The solution phase β value for tertiary phosphine is based on a single experimental value, which suggests that the discrepancy may be due to an error in the solution phase parameter. The solid state analysis indicates that tertiary phosphines are very poor H-bond acceptors, on a par with organic fluoride.

Solid state competition data were also gathered for functional groups with no experimentally determined solution phase β values, and the ratings were used in Equation 5 was to calculate solid state β values (Table 3). The SMARTS strings in Table 3 describe how the functional groups are defined.

Solid State H-Bond Donor Ratings

As explained above, H-bond donors almost always form H-bonds in crystal structures, due to the excess of H-bond acceptors present in organic molecules, so the same approach cannot be used to rank donors with paired comparison algorithms. However, if the H-bond acceptor ratings are already known, then a donor can be considered to have out-competed another donor if it H-bonds to an acceptor with a larger rating. Thus crystal structures that contain multiple donors provide the results of competitions between all pairs of donors present, and these data were used to calculate solid state H-bond donor ratings (R_D) using the TrueSkill algorithm. Figure 7 shows that there is no correlation between the solid state and solution phase H-bond donor parameters. This is a surprising result given that the H-bond acceptor parameters show such a strong correlation. There are two possible reasons: the number of functional groups for which solution phase H-bond donor parameters are available is relatively small, and the values fall in a narrow range; there was not enough solid state competition data for a rating to be calculated for some H-bond donors, because the calculation requires crystal structures with at least two H-bond donors and where every H-bonded acceptor has a rating. These conditions mean that the number of competitions acquired for the donors was much less than for the acceptors, so the poor correlation could be due to insufficient sampling.

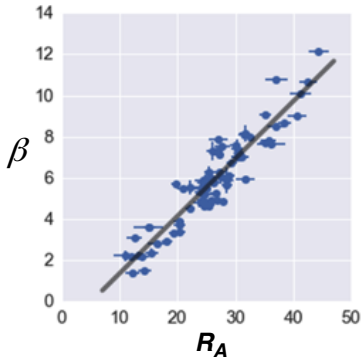

Figure 5. The relationship between the solution phase H-bond acceptor parameter β and the solid state H-bond acceptor rating R_A ($r^2 = 0.88$).

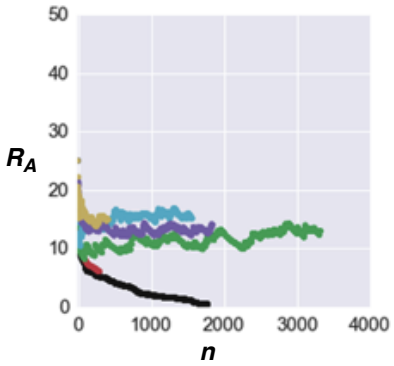

Figure 6. Solid state acceptor rating (R_A) as a function of the number of competitions (n) each acceptor is involved in (aromatic fluoride in black, aliphatic fluoride in red, aromatic chloride in green, aliphatic chloride in blue, aromatic bromide in purple, aromatic iodide in yellow).

Table 3. Solid state functional group *β* values.

Functional Group	SMARTS string	*β*
diaryl alkyl phosphine oxide	[PX4](=[OX1])([c])([c])([C])	9.6
N-substituted benzimidazole	c:1:c2:c(:c:c:c:1)[NX2]=CN2([cCX4])	9.1
N-substituted imidazole	N1(C(=[NX2]C(=C1[!F!Cl!Br!I!O!N])[!F!Cl!Br!I!O!N])[!F!Cl!Br!I!O!N])[cCX4]	8.0
N,N',N'-trisubstituted urea	[CX4c]N([H])C(=[OX1])N([CX4c])[cCX4]	7.9
primary alkyl amide	C(=[OX1])(N([H])[H])[Cc]	7.5
alkylidene hydrazide	C(=[OX1])([cC])N(N=C)[H]	7.4
N-substituted alkylidene hydrazide	C(=[OX1])([cC])N([NX2]=C)[cC]	7.3
alkylidene sulfonamide	S(=[OX1])([Cc])(=[OX1])[NX2]=C	7.1
N,N'-disubstituted hydrazide	C(=[OX1])(C)N(N([H])(C))C	7.1
tetrazole	[NX2]1=[NX2][NX2]=CN1[H]	7.0
N-aryl *N*-alkyl benzamide	[c]C(=[OX1])N(c)[CX4]	6.6
benzimidazole	c:1:c2:c(:c:c:c:1)[NX2]=CN2[H]	6.5
carbonate	C(=O)([OX1-1])([OX1-1])	6.5
imide	N(C(=[OX1])[!F!Br!Cl!I!O!N])([H])C(=[OX1])[!F!Br!Cl!I!O!N]	6.4
sulphinamide	[SX3](=[OX1])(N([H])[cC])[cC]	6.4
N-substituted pyrazole	[NX2]1=CC=CN1[cC]	6.2
thiourea	C(=[SX1])(N([cC])[H])N([cC])[H]	6.2
α,β-unsaturated aldehyde	C(=[OX1])([H])C=C	6.2
α,β-unsaturated carboxylic acid	C(=[OX1])(O[H])C=C	6.2
alkylidene hydrazinecarbodithioate	C(S)(=[SX1])N([H])[NX2]=C	6.1
N-aryl secondary amide	C(=[OX1])([CX4])N([H])c	6.1
aryl aldehyde oxime	c:1:c:c(:c:c:c:1)C([H])=[NX2][OX2][H]	6.1
N,O-dialkyl secondary carbamate	O(C(=[OX1])N([H])[CX4])[CX4]	6.0
α,β-unsaturated alkyl ester	C=C[C!R1](=[OX1])O[CX4]	5.8
alkyl aryl sulphone	S(=[OX1])(=[OX1])([c])([CX4])	5.7
N-aryl secondary benzamide	C(=[OX1])(c)N([H])c	5.5
primary sulphonamide	S(=[OX1])(=[OX1])(N([H])[H])([cC])	5.4
N-aryl *O*-alkyl secondary carbamate	[OX2](C(=[OX1])N([H])[c])[CX4]	5.3
oxadiazole	[NX2]1=C[OX2]C=[NX2]1	5.3
N,O-dialkyl *N*-aryl oxime	C([CX4])(c)(=[NX2][OX2]([CX4]))	5.0
alkylidene sulphonohydrazide	S(=[OX1])([Cc])(N([NX2]=C)[H])=O	4.9
oxime	[NX2](=C([cCX4])[cCX4])[OX2][H]	4.9
dialkylidene hydrazine	C=[N!R][N!R]=C	4.8
sulphonate diester	S(=[OX1])(=[OX1])([cC])O[cC]	4.3
N,N'-dialkyl diazo	C[NX2]=[NX2]C	4.2
N,N'-diaryl diazo	c[NX2]=[NX2]c	4.0
enol	[OX2]([CX3]=C)[H]	3.6
N-alkyl *N'*-aryl diazo	C[NX2]=[NX2]c	3.5
organic azide	C[NX2]=[NX2]=[NX1]	3.3
α,β-unsaturated alkyl ether	[CX4][OX2]C=C	3.0
m-halophenol	c:1(:c(:c(:c(:c(:c:1[!F!Cl!Br!I!O!N])[!F!Cl!Br!I!O!N])[OX2][H])[FClBrI])[!F!Cl!Br!I!O!N])[!F!Cl!Br!I!O!N]	2.9
alkyl aryl thioether	[CX4][SX2]c	2.6
disulphide	C[SX2][SX2]C	2.3
thiol	[CX4][SX2][H]	1.8
diaryl thioether	c[SX2]c	1.5
N-alkyl secondary aniline	c:1(:c(:c(:c(:c(:c:1[!F!Cl!Br!I!O!N])[!F!Cl!Br!I!O!N])[NX3]([CX4])[H])[!F!Cl!Br!I!O!N])[!F!Cl!Br!I!O!N])[!F!Cl!Br!I!O!N]	1.3
α,β-unsaturated secondary amine	[NX3](C=C)([CX4])[H]	0.7

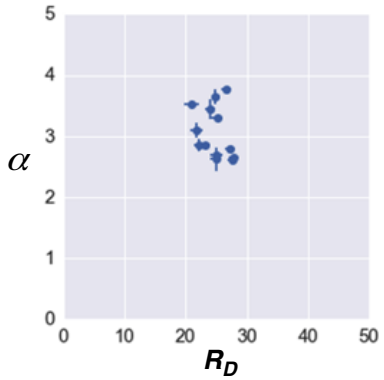

Figure 7. The relationship between the solution phase H-bond donor parameter α and the solid state H-bond donor rating R_D (r^2 = 0.08).

Conclusion

Two methods for using the H-bonding outcomes in crystal structures to quantify the H-bonding properties of functional groups have been explored. The first calculates donor and acceptor functional group probabilities of H-bond formation by comparing the frequency at which they formed H-bonds relative to the number of times the functional group is observed in the CSD. For donors, this method proved problematic, because it was found that all H-bond donors tend to be H-bonded in crystal structures, due to the excess of acceptor sites in most organic molecules. The result is that donor probabilities were all close to one and therefore did not correlate with the solution phase H-bond donor parameters. Acceptor probabilities were found to correlate with solution phase H-bond acceptor parameters, but the relationship between the two parameters was not sufficiently strong to allow for reliable calculation of solution phase β values using solid state acceptor probabilities. The source of the correlation between the acceptor probabilities and β values is ascribed to the fact that acceptors typically outnumber donors in organic molecules and must therefore compete with one another to form H-bonds in a crystal structure. The strongest acceptors more frequently out-compete weaker ones and therefore have higher H-bond probability values (and *vice versa*). A problem with the method is that whether an acceptor forms a H-bond or not in a crystal structure is a function of the competitor environment. Acceptors that frequently compete against weak acceptors will have larger probabilities and acceptors that frequently compete against strong acceptors will have smaller probabilities. Since it is impossible to standardise the competitive environments across all the crystal structures used to calculate probabilities, this is not a viable method to quantitatively assess and compare H-bond acceptor or donor properties.

The second method uses the assumption that any acceptor which has formed a H-bond in a crystal structure has outcompeted an acceptor which has not. Observations of the H-bonding outcomes in a large number of crystal structures give the results of many different acceptor competitions and can be used in paired comparison algorithms to rate each H-bond acceptor. The acceptor ratings calculated with the TrueSkill algorithm correlate very well with the corresponding solution phase H-bond acceptor parameters. This novel solid state approach therefore offers an alternative way to quantify the properties of H-bond acceptors which is complementary to solution phase measurements. It should be possible to use the acceptor ratings to rank H-bond donors in a similar manner. The large excess of H-bond acceptors relative to H-bond donors means that the donor that interacts with the best H-bond acceptor has outcompeted all other donors in that crystal structure. The results of these donor competitions were used in the TrueSkill algorithm to calculate donor ratings. However, the correlation between these donor ratings and α was poor, which is most likely to due the relatively small sample size available for this comparison. One approach to this problem would be to expand the range of donors to include CH groups, which would lead to an excess of donors over acceptors and a more competitive donor environment.

The results described here show that despite the size of the CSD, the coverage of different functional group combinations is not sufficient to use raw statistics to draw conclusions about the relative strengths of the different intermolecular interactions present in the database. However, an analysis that treats each crystal structure as the outcome of a competition between functional groups for intermolecular interactions appears to be a promising approach to determining relative rankings of the strengths of non-covalent functional group interactions.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We thank the Engineering and Physical Sciences Research Council and the Cambridge Crystallographic Data Centre for funding, and Dr Neil Feeder for useful discussions.

Notes and references

1 C. A. Hunter, *Angew. Chem. Int. Ed. Engl.*, 2004, **43**, 5310.
2 C. S. Calero, J. Farwer, E. J. Gardiner, C. A. Hunter, M. Mackey, S. Scuderi, S. Thompson and J. G. Vinter, *Phys. Chem. Chem. Phys.*, 2013, **15**, 18262.
3 (a) J. L. Cook, C. A. Hunter, C. M. R. Low, A. Perez-Velasco and J. G. Vinter, *Angew. Chem. Int. Ed. Engl.*, 2007, **46**, 3706; (b) J. L. Cook, C. A. Hunter, C. M. R. Low, A. Perez-Velasco and J. G. Vinter, *Angew. Chem. Int. Ed. Engl.*, 2008. **47**, 6275; (c) R. Cabot and C. A. Hunter, *Org. Biomol. Chem.*, 2010, **8**, 1943; (d) R. Cabot, C. A. Hunter and L. M. Varley, *Org. Biomol. Chem.*, 2010, **8**, 1455; (e) V. Amenta, J. L. Cook, C. A. Hunter, C. M. R. Low and J. G. Vinter, *Org. Biomol. Chem.*, 2011, **9**, 7571; (f) V. Amenta, J. L. Cook, C. A. Hunter, C. M. R. Low and J. G. Vinter, *J. Phys. Chem. B*, 2012, **116**, 14433; (g) R. Cabot and C. A. Hunter, *Chem. Soc. Rev.*, 2012, **41**, 3485.
4 C. A. Hunter, *Chem. Sci.*, 2013, **4**, 1687.

5 (a) K. Busuttil, M. Geoghegan, C. A. Hunter and G. J. Leggett, *J. Am. Chem. Soc.*, 2011, **133**, 8625; (b) N. Nikogeorgos, C. A. Hunter and G. J. Leggett, *Langmuir*, 2012, **28**, 17709.
6 J. McKenzie, N. Feeder and C. A. Hunter, *CrystEngComm,* 2016, **18**, 394.
7 (a) D. Musumeci, C. A. Hunter, R. Prohens, S. Scuderi and J. F. McCabe, *Chem. Sci.*, 2011, **2**, 883; (b) T. Grecu, H. Adams, C. A. Hunter, J. F. McCabe, A. Portell and R. Prohens, *Cryst. Growth Des.*, 2014, **14**, 1749; (c) T. Grecu, C. A. Hunter, E. J. Gardiner and J. F. McCabe, *Cryst. Growth Des.*, 2014, **14**, 165.
8 C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, *Acta Cryst.*, 2016, **B72**, 171.
9 J. E. Mills and P. M. Dean, *J. Comput. Aided. Mol. Des.*, 1996, **10**, 607.
10 F. H. Allen, C. M. Bird, R. S. Rowland and P. R. Raithby, *Acta Cryst.*, 1997, **B53**, 696.
11 L. Infantes and W. D. S. Motherwell, *Z. Kristallogr.*, 2005, **220**, 333.
12 R. S. Rathore, Y. Alekhya, A. K. Kondapi and K. Sathiyanarayanan, *CrystEngComm.*, 2011, **13**, 5234.
13 F. H. Allen, W. D. S. Motherwell, P. R. Raithby, G. P. Sheilds and R. Taylor, *New J. Chem.*, 1999, **23**, 25.
14 C. Bilton, F. H. Allen, G. P. Sheilds, J. A. Howard, *Acta Cryst.*, 2000, **B56**, 849.
15 A. E. Elo, *Chess Life*, 1961, **16**, 160.
16 S. Eppel and J. Bernstein, *Acta Cryst.*, 2008, **B64**, 50.
17 P. T. Galek, L. Fabian, W. D. Motherwell, F. H. Allen, N. Feeder, *Acta Crystallogr B*, 2007, **63**, 768.
18 R. Taylor, *CrystEngComm.*, 2014, **16**, 6852.
19 M. H. Abraham, P. P. Duce, D. V. Prior, D. G. Barratt, J. J. Morris and P. J. Taylor, *J. Chem. Soc. Perkin Trans 2.*, 1989, 1355.
20 N. Ziao, J. Graton, C. Laurence, J. Y. Le Questel, *Acta Cryst.*, 2001, **B57**, 850.
21 R. A. Bradley and M. E. Terry, *Biometrika.*, 1952, **39**, 324.
22 M. E. Glickman, *Am. Chess J.*, 1995, 59.
23 R. Herbrich, T. Minka and T. Graepel, *Adv. Neural. Inf. Process. Syst.*, 2007, **19**, 569.
24 P. Dangauthier, R. Herbrich, R. Minka and T. Graepel, *Adv. Neural. Inf. Process. Syst.*, 2007, **20**, 337.
25 R. Taylor and O. J. Kennard, *J. Am. Chem. Soc.*, 1982, **104**, 5063.
26 F. H. Allen, *Acta Cryst.*, 2002, **B58**, 380.
27 SMARTS Theory Manual, Daylight Chemical Information Systems, Sante Fe, New Mexico.
28 G. van Rossum, Python tutorial, Technical Report CS-R9526, Centrum voor Wiskunde en Informatica (CWI), Amsterdam, 1995.
29 F. H. Allen, *Acta Crystallogr. Sect B Struct. Sci.*, 1986, **42**, 515.
30 M. C. Etter, *Acc. Chem. Res.*, 1990**, 23,** 120.
31 A Python implementation for TrueSkill written by Heungsub Lee can be found at http://trueskill.org/.
32 J. Donohue, *Acta Cryst*., 1957, **10**, 383.
33 J. D. Dunitz and R. Taylor, *Chem. Eur. J.*, 1997, **3**, 89.
34 J. R. Loader, S. Libri, A. J. H. M. Meijer, R. N. Perutz and L. Brammer, *CrystEngComm*, 2014, **16**, 9711.