

Knowledge-based model of hydrogen-bonding propensity in organic crystals

Peter T. A. Galek,^{a,b*} László Fábrián,^{a,b} W. D. Samuel Motherwell,^a Frank H. Allen^a and Neil Feeder^c

^aCambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England, ^bPfizer Institute for Pharmaceutical Materials Science, Department of Materials Science and Metallurgy, University of Cambridge, Pembroke Street, Cambridge CB2 3QZ, England, and ^cPfizer Global R&D, Ramsgate Road, Sandwich, Kent CT13 9NJ, England

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

Received 23 March 2007

Accepted 25 June 2007

A new method is presented to predict which donors and acceptors form hydrogen bonds in a crystal structure, based on the statistical analysis of hydrogen bonds in the Cambridge Structural Database (CSD). The method is named the logit hydrogen-bonding propensity (LHP) model. The approach has a potential application in identifying both likely and unusual hydrogen bonding, which can help to rationalize stable and metastable crystalline forms, of relevance to drug development in the pharmaceutical industry. Whilst polymorph prediction techniques are widely used, the LHP model is knowledge-based and is not restricted by the computational issues of polymorph prediction, and as such may form a valuable precursor to polymorph screening. Model construction applies logistic regression, using training data obtained with a new survey method based on the CSD system. The survey categorizes the hydrogen bonds and extracts model parameter values using descriptive structural and chemical properties from three-dimensional organic crystal structures. LHP predictions from a fitted model are made using two-dimensional observables alone. In the initial cases analysed, the model is highly accurate, achieving ~90% correct classification of both observed hydrogen bonds and non-interacting donor–acceptor pairs. Extensive statistical validation shows the LHP model to be robust across a range of small-molecule organic crystal structures.

1. Introduction

The Cambridge Structural Database (CSD), Version 5.27, November 2005 (Allen, 2002) provides a repository of > 355 000 small molecule organic and organometallic crystal structures. The existence of each structure in the database may be considered as testament to the practicalities of that structure's own crystallization, or more precisely, each database entry adds to a collective *a posteriori* understanding of the kinetics and thermodynamics of similar crystal formation. Related molecular and structural properties may be subjected to statistical analysis such that this *latent information* in the CSD may be revealed. The work presented here focuses on hydrogen bonds, which play a dominant role in the supra-molecular arrangement within a crystal structure. Knowledge of such interactions is vital in crystal engineering (Aakeröy, 1997; Desiraju, 1995; Etter, 1991), crystal structure prediction (CSP; Motherwell, 1999; Day & Motherwell, 2006), solution of crystal structures from powder diffraction data (e.g. *DASH*: David *et al.*, 2006) and prediction of protein–ligand docking (Böhm & Klebe, 1996).

An important category is the occurrence of polymorphism, particularly in the pharmaceutical industry. Unusual hydrogen-bonding interactions can indicate a metastable crystalline form, when identification of possible thermodynamically more stable form(s) is sought, or conversely, when rationalization of an obtained (believed) stable form is required. Polymorph prediction techniques are widely used to provide potential screening of metastable crystal structures, which can dramatically increase the efficiency of the drug-development process (Price, 2004). These powerful methods have a proven track record, however, their current effectiveness quickly decreases as the system complexity increases, *e.g.* multiple component crystals, flexible molecules and salts (Day *et al.*, 2004; Nowell & Price, 2005). A measure of likelihood of the dominant potential intermolecular interactions, *i.e.* hydrogen bonding, whilst not describing the complete set of interactions in the structure, may be effective in identifying the presence (or absence) of significant structure-directing factors. Furthermore, a knowledge-based method as such may potentially be highly practical as a precursor to the computationally intensive polymorph-prediction methods.

There have been many previous hydrogen-bonding analyses of organic crystal structures (Braga *et al.*, 1997; Bilton *et al.*, 2000; Haynes *et al.*, 2004; Infantes & Motherwell, 2004; Chisholm *et al.*, 2006; Parkin *et al.*, 2006). Much of the recent work has been in identifying and applying knowledge of common motifs in the extended three-dimensional crystal structure which, it is indicated, suggests routes of crystal formation *via* supramolecular synthons (Desiraju, 1995). This approach has proved very useful and shows much promise.

The primary aim of this study is to describe pairwise interactions alone. While we acknowledge the advantage of a more detailed extended three-dimensional structural description, especially for more complex examples, we suggest that given the right data, the most likely pairwise hydrogen bonds in a crystal structure will be both predictable and prolific across structures containing similar molecules. Indeed, as will be shown, the initial description proves to be extremely successful. This success suggests that there is a dominant role of pairwise hydrogen bonding in molecular recognition and organic crystal growth. That being said, a prediction of non-

covalent bonding could well be improved by the inclusion of a description of the extended three-dimensional crystal structure (and future developments in this direction may be carried out).

In our new approach, the descriptive properties of hydrogen donor and acceptor atoms and their molecular environments are accumulated per crystal. These are used as both qualitative and quantitative parameters to construct a dichotomous (two-state) probability model to predict the propensity of a hydrogen bond to form. Owing to the discrete nature of the variable, the logit probability distribution is used as a model curve by which to fit parameters *via* the logistic regression technique (Agresti, 1990).

Currently, the CSD suite of software provides a wide range of tools for searching the database for specific interactions and structural properties, such as *ConQuest* (Allen & Taylor, 2005; Bruno *et al.*, 2002), *Isostar* (Bruno *et al.*, 1997) and *Mercury* (Macrae *et al.*, 2006). From the outset, however, the work has required that new surveyor and analysis methods be constructed, since the specific descriptive parameters associated with each donor and acceptor group must be obtained. A new survey and categorization application has been constructed, named *H-Bond Surveyor* that conducts complete analysis of hydrogen donor and acceptor atoms, and the hydrogen bonds in which they participate in crystal structures. For increased flexibility, functional groups can be defined in the *Quest* (Allen *et al.*, 1991) query format that may be of specific detail for each survey, to describe the environment surrounding an identified donating or accepting atom.

We begin this paper by describing the database survey and statistical methodology. An example application is then presented and a specific structure is analysed in detail. Results and discussion of the application, and in particular a statistical assessment of the model, are presented in the following section, and finally, conclusions are drawn. Specific theoretical details are provided in the *Appendix*.

2. Methodology

In this approach, hydrogen donor and acceptor atoms, and the hydrogen bonds in which they participate are identified per crystal structure. We define the existence of a hydrogen bond between potential donor and acceptor atoms using distance and angle criteria, and the observation of any given pair as being hydrogen bonded is recorded as a two-state variable, *True* or *False*. Descriptive properties are then accumulated that are based on the atoms' molecular environments. These are used as both qualitative and quantitative parameters to construct a dichotomous (two-state) probability model. The model is applied to compute a probability measure termed *propensity*, denoted π , for the formation of a hydrogen bond between a specified donor and acceptor atom. The resulting description is termed the logit hydrogen-bonding propensity (LHP) model. The key assumption in the LHP model is that the *actual* hydrogen bonds directing the formation of a crystal structure will be those with the *highest likelihood* of forming among all possible donor–acceptor pairs.

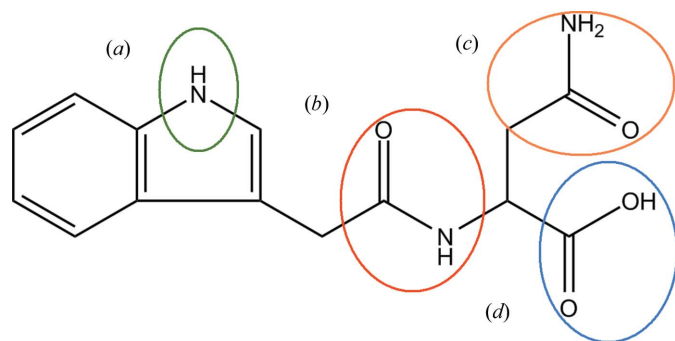


Figure 1
Chemical diagram for the MIFCEJ molecule used for LHP model example propensity predictions, showing potential hydrogen-bonding functional groups: (a) NH, of indole; (b) secondary amide, CONH; (c) primary amide, CONH₂; (d) carboxylic acid, COOH.

The general mode of application of the model is to identify a target molecular structure, for which one would like to predict propensity values for each possible donor–acceptor pairing within the crystal structure. One aims to combine the information in the CSD for structures and/or chemical functionality related to the structure being modelled to form an indicative representation of that structure. Thus, despite the possibility of there being no precedent structures in the CSD with all the target functionality, knowledge of interactions between all target structure sites is obtained collectively.

In order to illustrate the presented methodology, an example crystal structure has been selected from the CSD. The example chosen is *N*²-(indol-3-ylacetyl)-L-asparagine, a molecule of moderate size, which contains strong hydrogen-bonding functionality and is also of biological interest, displaying growth promoting activity in plant tissue. The crystal can be found in the CSD under reference code MIFCEJ (Antolic *et al.*, 2001; Fig. 1). In order to maintain a fair assessment, data from this structure are kept separate from any model training sets. As such, the example forms a ‘blind test’ for the approach that may subsequently be critically assessed. Further discussion of the chemical properties of the example and the associated descriptive properties to be used in model prediction can be found in §§3 and 4. Once a target structure is identified, the LHP model procedure involves three stages, as follows.

2.1. Dataset extraction

Stage 1 is to extract a subset of crystal structures from the CSD that contain the relevant chemical functional groups. Primarily, one would like to have data for only the most similar crystal structures, in an effort to obtain discriminatory data. Thus, it is desirable to screen away any chemically incomparable structures (*e.g.* charged *versus* uncharged species). This stage is also a matter of efficiency, since many of the crystal structures contain no hydrogen bonding, or contain only bonds between atoms unrelated to the target crystal, thus not contributing any useful data for the particular case of interest. This subset selection is achieved using *ConQuest* via the design and selection of donor/acceptor functional groups, and the application of screens and/or search combinations.

The set is then screened to remove duplicate entries. Often in the CSD there is more than one entry for the same crystal structure. These are often due to repeated refinements, republications and redetermination under different experimental conditions (van de Streek & Motherwell, 2005). Polymorphs, solvates and cocrystals are not removed since they may contain different extended three-dimensional structures. The screening is achieved using the comparison algorithm available in *Compack* (Chisholm *et al.*, 2006), which overlays a pair of structures to identify those that are identical. The procedure for data-set extraction is illustrated for the example structure in §3.1.

2.2. Hydrogen-bond surveyor

Stage 2 is the extraction of the descriptive properties per donor–acceptor pair (*X–Y*) for every crystal structure, using the H-Bond Surveyor application. In each crystal structure, hydrogen bonds are located as atom–hydrogen–atom contacts (*X–H··Y*), and the appropriate donating and accepting atoms are identified. The analysis employs user-variable contact criteria (*X–Y* atom distances and *X–H–Y* atom angles are assessed) by which to accept or discard possible contacts. Iteratively over the set of contacts, the pair of functional groups and the set of properties related to the bond are then identified (see below), to act as input model parameters describing the hydrogen-bonded pair. Once the bonded pairs are surveyed, all permutations of observed potential donors and acceptors not found to be hydrogen bonded (according to the survey) are then paired together and the model properties are extracted. Thus, all combinations of donor–acceptor pairs in a structure, either interacting or non-interacting, are analysed.

2.2.1. Model parameters. The role of the parameters is to discriminate between hydrogen bonds, *i.e.* to assess whether one bond is statistically favoured over another within the chosen CSD subset. This is achieved by assigning all descriptors a model coefficient that changes magnitude according to how influential the descriptor is to the outcome. Two types of parameters are used: quantitative and qualitative. Quantitative parameters scale the predicted propensity by the influence felt from the specific magnitude of that parameter. The value of the quantitative parameters must be calculated for a chosen donor–acceptor pair, and then its influence on the model equation is applied through its model coefficient. Qualitative parameters separate groups of sample data according to their label, *e.g.* all the pairs which involve a carboxylic acid as the hydrogen donor. Different groups may feel a greater or lesser effect from other variables in the model and so this distinction

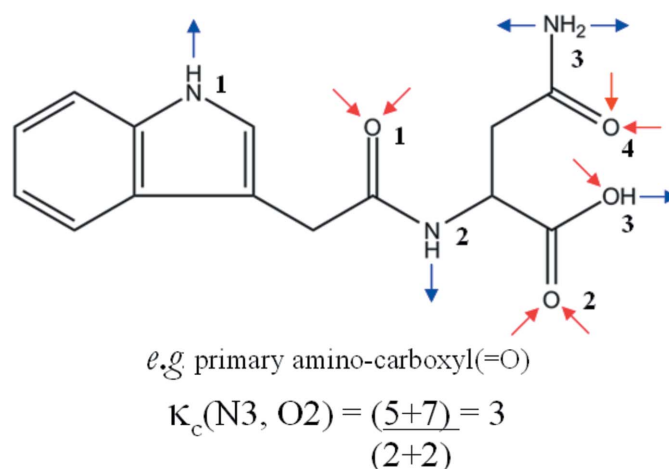


Figure 2

Illustrated calculation of the competition function on the MIFCEJ molecule for a choice of donor and an acceptor site (N3 of the primary amido group and O2 of the carboxylic acid group). The arrows show the potential donor hydrogen atoms (outward) and potential acceptor lone pairs (inward).

provides more flexibility. Qualitative parameters act like a switch in the model. They provide an influence from their model coefficient to the model equation only if the parameter is relevant to the specific donor–acceptor pair chosen.

Qualitative parameters: (a) Donor and acceptor functional group

Different hydrogen-bonding behaviour is well known to be separated qualitatively by the functional group, denoted Γ herein, to which the donor/acceptor atom belongs, *e.g.* carboxyl, hydroxyl, ether *etc.* Indeed, the concept of ‘functional group’ is an empirical description from the result of discriminatory chemical observation. A survey of the hydrogen bonding in a set of structures using the functional group as a model parameter obtains a complete set of atom pairs which did and did not form contacts, categorized by donor/acceptor group membership. This label can then weight the propensity score up or down (which may or may not depend on other parameters) in the case of a group which may donate or accept more or less than the other alternatives. Thus, for each possible functional group, i , a fitted model contains the binary $\Gamma_A(i)$ or $\Gamma_D(i)$ (or perhaps both, given an acceptor group which may also have donor functionality, *e.g.* alcohol OH) to act as a switch. In this way, the influence of the appropriate fitted coefficient is included in the model calculation.

Previously, a similar quantitative approach has been taken in which the frequencies of hydrogen bonds in the CSD for a designated set of functional groups are expressed as a fraction of the total possible occurrences of the groups both being present (Allen *et al.*, 1999; Infantes & Motherwell, 2004). In the current approach, applying the functional group as a lone model parameter, a similar score is accessible from the H-Bond survey. We define this percentage as a *participation* of specific functional groups in hydrogen bonds within a crystal structure subset. In the model scheme, however, the functional group parameter is qualitative and the effects on the propensity of a pair remain relative to different groups, as will be described later. This then in the model scheme can be

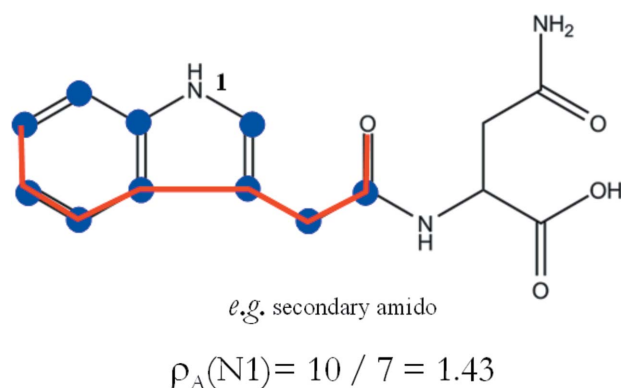


Figure 3
Illustrated calculation of the steric density function on the MIFCEJ molecule for the secondary amino group. The sum of the marked atoms represents the region bounded by the other donors and acceptors, and the line marks the most direct path between the most separated atoms.

described as *relative participation*. We also define further discriminatory variables in this new method, as will be presented in the next section, that give a *specific* expectation, amending the *average* observed behaviour for the formation of hydrogen bonds between groups.

Quantitative parameters: (b) Competition function

The competition function assesses the number of H atoms available for donation, D , and the number of available acceptor lone pairs, A , expressed as a fraction of the total on the discrete covalently bonded unit(s) [and potentially also ion(s)] in the asymmetric unit. Thus, the formation of a particular hydrogen bond may be thought of as a function of the different donor and acceptor atoms competing for a partner. If a functional group has more than one donor/acceptor site the competition function accounts for this potential advantage over other groups. Thus, permutations of non-bonded donor–acceptor pairs are analysed per group, not per potential donor/acceptor atom. The competition function is

$$\kappa_c(i, a) = \frac{\sum_c D_c + \sum_c A_c}{D_i + A_a} \quad (1)$$

It is expressed in terms of alternative a for individual i , belonging to choice set c . For example, given an individual donor atom, a potential pair with an acceptor atom as one of the alternatives has an associated competition value due to all the donor and acceptor atoms in the set of choices. (Note the description is identical if one considers the symmetric situation of an individual acceptor’s ‘choice’ per potential donor.) The function has a minimum of 1 and an unlimited maximum.

A computation of the competition function is illustrated in Fig. 2. We use the example MIFCEJ molecule, as introduced in the previous section. A donor and an acceptor site are chosen (N3 of the primary amido group and O2 of the carboxylic acid group). The arrows show the potential donor H atoms and potential acceptor lone pairs. We see the total $D_c + A_c$ on the molecule is 12 and the sum on the chosen sites is 4, thus the competition as a quotient of the two scores is 3.

(c) Steric Density Function

If the functional group is considered to be a description of the primary environment around a donating/accepting atom, the steric density function is an attempt to describe the secondary or more long-range environment. A measure of *steric density* aims to account for situations when a normally preferable hydrogen bond pairing may not form owing to steric accessibility reasons, *i.e.* how crowded is the region around a potential donor/acceptor.

The function is defined by assessing the size of the hydrophobic or non-hydrogen-bonding region around a donating/accepting atom using the molecular graph connectivity data of the structural component. The method accounts for all the atoms in this region by way of a directed walk around the molecular graph until the next potential hydrogen-bonding atom is met. In this way a sub-graph for the hydrophobic region is built. The method identifies the total number of atoms in the region, \sum_{cfj} , that are not part of j identified functional group(s), and a distance measure named the

furthest direct-atom path length, $r_{c\neq j}$ in terms of covalent bond count. This is the maximal shortest path in the set of shortest paths between the atoms of the sub-graph. The method returns a ratio of the total atom count and the furthest direct-atom path length. In this way a type of ‘steric density’ $\rho_c(i)$ is defined

$$\rho_c(i) = \frac{\sum_{c\neq j} r_{c\neq j}}{r_{c\neq j}}. \quad (2)$$

The function applies a graph-traversal technique (see *e.g.* Cormen *et al.*, 1989) using a recursive algorithm, beginning at a donor or acceptor atom and analysing adjacent non-hydrogen-bonding atoms. The algorithm may walk to the next non-hydrogen-bonding atom, stop at an atom with hydrogen-bonding functionality, or reach a branch in the connectivity graph. At a branch point the algorithm begins a new lower level of recursion and begins the walk again from this point in the new direction. At a terminal atom, the algorithm steps back up to the next higher level. Thus, the algorithm performs an exhaustive walk from the donor/acceptor around the connected hydrophobic region of the molecule. Counted atoms are flagged which ensures no recounting. This helps the efficiency of the algorithm and ensures cyclic fragments can be dealt with. ρ_c has a minimum of zero and is increasingly positive as the number of hydrophobic atoms in the molecular region increases.

Computation of this function is illustrated again using the MIFCEJ example molecule, Fig. 3. The total atom count, $\sum_{c\neq j}$, of the region bounded by the other donors and acceptors is 10, as displayed by the marked atoms, and the most direct path highlighted between the most separated atoms is 7. Thus, the steric density is 1.43 as a quotient of the scores.

2.2.2. Summary. We have thus far extracted a set of unique crystal structures containing functionality of relevance to the target molecule, and located their hydrogen bonds and combinations of non-interacting potential donor–acceptor atom pairs. For each pair we compute model parameters comprised of chosen molecular and chemical descriptors. Once the data is extracted, the next step is to derive a model as an attempt to account for the hydrogen-bonding behaviour that is observed.

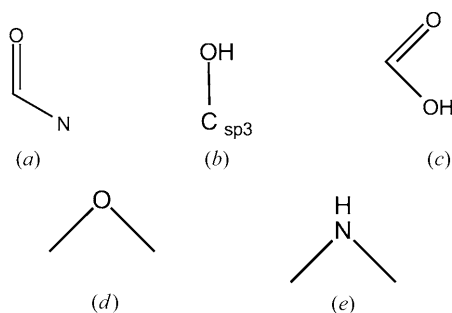


Figure 4
Functional groups as defined in H-Bond Surveyor for the example model survey. Specific groups: (a) amido; (b) aliphatic hydroxyl; (c) carboxyl; (d) ether; (e) secondary amino.

2.3. Logistic regression

The logistic regression technique is a widely used tool for the analysis of discrete events (see *e.g.* Agresti, 1990; Hosmer & Lemeshow, 2000). Following initial adoption in the field of epidemiology, the technique has found applications in the fields of biomedical research, ecology, finance, criminology and linguistics, to name a few examples. The application of logistic regression to the properties of crystal structures, as presented in this work, is a novel approach. For this reason, much of the theory behind the regression modelling presented herein will be discussed alongside the associated results. The general goal of the regression model is to find the best fitting and most economical model, whilst preserving a scientifically (chemically) reasonable description. For the interested reader, a more detailed theoretical background of the approach may be found in the *Appendix*. An overview of the approach now follows.

2.3.1. Application. Categorization of hydrogen bonds and descriptive data are obtained as discussed previously in the section for all permutations of possible hydrogen-bonding atom pairs in a selected subset of crystal structures. Model fitting is then carried out *via* logistic regression with a linear description of the variable parameters. The data set contains three types of information to be used in the model. The response variable is the key information per donor–acceptor pair which gives the two-state outcome under the settings of the survey, that the pair was involved in a hydrogen bond or it was not. The quantitative and qualitative variables (as defined above) then act as discriminating factors to account for the frequency of occurrence of one type of bond over another.

The essence of the model is a function of probability (the log of the odds of the probability, P , of a pair forming a bond), represented by a linear model of the descriptive parameters, x_k^i

$$\log\left(\frac{P}{1-P}\right) = \alpha + \sum_k x_k^i \beta_k. \quad (3)$$

The function of the log of the odds of P is also known as the logit of P . The rationale for this choice of functional form is described in the *Appendix*. α is the intercept or baseline variable, and the β_k coefficients vary according to the degree of influence of their corresponding parameter, x_k^i . Assuming only partial information, an approximation to P may be obtained, which is denoted π . Specifically for the LHP model, with the parameters defined in §2.2, the model is given by

$$\begin{aligned} \log\left(\frac{\pi}{1-\pi}\right) &\equiv \text{logit}(\pi) \\ &= \alpha + \kappa_c(i, a)\beta_k + \rho_{D,c}(i)\beta_{\rho_D} + \rho_{A,c}(a)\beta_{\rho_A} \\ &\quad + \Gamma_{D,c}(i)\beta_D + \Gamma_{A,c}(a)\beta_A. \end{aligned} \quad (4)$$

Model fitting (optimization of the β coefficients) is achieved using a form of non-linear regression using the logistic function and a *maximum likelihood* algorithm (see *Appendix*). Given a fitted model, the estimated likelihood of a pair to form a hydrogen bond is then measured from the relation

$$\pi_c^i(a) = \frac{\exp(\alpha + \kappa_c(i, a)\beta_k + \rho_{D,c}(i)\beta_{\rho_D} + \rho_{A,c}(a)\beta_{\rho_A} + \Gamma_{D,c}(i)\beta_D + \Gamma_{A,c}(a)\beta_A)}{1 + \exp(\alpha + \kappa_c(i, a)\beta_k + \rho_{D,c}(i)\beta_{\rho_D} + \rho_{A,c}(a)\beta_{\rho_A} + \Gamma_{D,c}(i)\beta_D + \Gamma_{A,c}(a)\beta_A)} \quad (5)$$

$$= \frac{1}{1 + \exp(-(\alpha + \kappa_c(i, a)\beta_k + \rho_{D,c}(i)\beta_{\rho_D} + \rho_{A,c}(a)\beta_{\rho_A} + \Gamma_{D,c}(i)\beta_D + \Gamma_{A,c}(a)\beta_A))} \quad (6)$$

obtained from a rearrangement of (4). This measure has been termed the *propensity* for a hydrogen bond to form, or more completely, the *logit hydrogen-bonding propensity*. For any hypothetical hydrogen bond with a donor–acceptor pair, it is thus possible to calculate a propensity score using (6), taking:

- (i) a fitted set of LHP model coefficients;
- (ii) values of the quantitative parameters: competition and steric density about the donor and acceptor for the residue on which the pair resides;
- (iii) identified qualitative parameters: the specific donor and acceptor group for the selected pair.

Thus, by way of the coefficients' values, the model equation gives a result for a particular donor–acceptor pair in their specific molecular environment.

3. Example application

The following section details the extraction of the dataset, analysis of hydrogen bonding and model fitting with regard to forming a prediction for the example molecule, MIFCEJ (Fig. 1). The intention is to form a hierarchy of predicted propensities for potential hydrogen-bonded atom pairs in the structure using the methodology set out in the previous section. Once obtained, the predictions may then be compared with the interactions found in the known three-dimensional structure. Aside from this illustration using the example, we note that statistical assessment of the regression procedure allows a complete description of the predictive power of models, and so such treatment will also be presented. This forms, in effect, a global view of the success of the model for the entire set of crystal structures.

3.1. Details of the CSD survey

The CSD survey was conducted to obtain a relevant dataset to the example. Search specifications (using *ConQuest*) located any structure that contains at least one amide fragment and one carboxylic acid fragment, with resolved three-dimensional coordinates (including all H-atom three-dimensional coordinates determined). The amine group proved to be prevalent in the dataset without explicitly searching for this fragment. To generalize, adjacent atoms to the amide group were not specified and the C-atom type was unspecified, leading to the general *amido* fragment description (see Fig. 4a). To screen against potential organometallic structures, salts and other ionic species, specific queries were defined, including element screens. The CSD survey yielded 1182 structures. Following the removal of duplicate entries, yielding

Table 1
Hydrogen-bond surveyor summary.

Item	Count
Total structures surveyed	1083
Total bonds observed	4769
Total intermolecular bonds observed	4331
Total unique donating groups	5
Total unique accepting groups	6

1083 structures, the dataset was analysed with the H-Bond Surveyor to obtain the hydrogen-bonding information.

The survey specifications were as follows. The minimum and maximum donor–acceptor distance cut-offs were the sum of the atomic van de Waals radii -5 and $+0.1$ Å, respectively. The minimum donor–hydrogen–acceptor angle cut-off was 90° (the optimum angle being 180°). Intramolecular and poly-furcated bonds were not considered individually, however, in the latter case the presence of such interactions would count as multiple separate *true* hydrogen-bond observations. (Future applications may include such interactions as unique options for the model outcomes.) The details of the survey are given in Table 1.

Other commonly occurring groups were specified as qualitative parameters, namely the aliphatic hydroxyl, carboxyl, ether and secondary amino groups. These groups, as defined in the survey, are displayed in Fig. 4. The remaining less-common groups were cumulatively classified as 'other'. Frequency statistics for the specific pairs are discussed in the next section.

3.2. LHP model regression

3.2.1. Training dataset statistics. The surveyor results listing the true/false hydrogen-bond observations and model parameter values for every potential donor–acceptor atom pair were fitted using the logistic regression technique available in the statistics package *XLSTAT* (Addinsoft, 2006, available as an add-in to the Microsoft *Excel* spreadsheet software). There were 17 558 potential pairs as raw data for the model fitting. Of all the potential hydrogen-bond pairs, 76% were observed not to be hydrogen bonded (permutations in possible pairs tend to create a much greater number of pairs that are not bonded).

As qualitative variables, five functional groups were identified and one category named 'other' for those miscellaneous individually less frequent groups, resulting in six variables. Five of these could possibly donate (all but the ether group) and five have potential accepting ability (all but the secondary amino group).¹ Itemized group frequency data are given in Table 2. It is not surprising, given the origin of the data from

¹ All six groups were, in fact, observed to accept hydrogen bonds, although the frequency of the secondary amino group accepting an H atom with respect to the total number of observations was prohibitively low for convergence of the regression. Manual analysis revealed that most of these few observations were aziridine groups, which would perhaps, due to a sufficiently different chemical nature, warrant a separate model parameter from the general description of the secondary amino group used in this survey. For these reasons, the group was not considered as a parameter in the acceptor category, and the corresponding observations were removed from the dataset.

Table 2

Summary statistics of the donor–acceptor pair data.

(a) Pairwise hydrogen-bonding observations

Variable	Category	Frequency	%
Hydrogen bond exists	FALSE	13 278	75.6
	TRUE	4280	24.4

(b) Quantitative parameters

Variable	Minimum	Maximum	Mean	Standard deviation
Competition function κ	0.000	124.000	12.722	14.443
Donor steric density function ρ_D	0.000	8.333	1.106	1.045
Acceptor steric density function ρ_A	0.000	9.833	1.495	1.345

(c) Qualitative parameters. Frequency is the total observations of a category in pairwise combinations, forming potential hydrogen-bonded donor–acceptor groups, throughout the set of crystal structures. Percentages express a fraction of group frequency with the combined frequency of all identified categories.

Variable	Category	Frequency	%
Donor, Γ_D	Amido	7337	40.8
	Other	3861	21.4
	Amino, secondary	2974	16.5
	Hydroxyl, aliphatic	2369	13.2
	Carboxyl	1464	8.1
Acceptor, Γ_A	Other	7315	40.7
	Amido	4724	26.3
	Ether	2931	16.3
	Carboxyl	1280	7.2
	Hydroxyl, aliphatic	1315	7.3

the CSD survey, that the most common donor group was the amido group, occurring in 41% of all possible pairs observed. However, this group was less dominant in the set of potential acceptor groups, 27%. There is also frequent occurrence of the remaining five group categories, giving a well balanced data set.

3.2.2. Regression statistics. The quality of the model, how well it can predict the hydrogen-bonding likelihood of a donor–acceptor pair, can be measured in terms of both the training dataset and optionally a separate validation set, that is, similar donor–acceptor pairs from crystal structures not used in the model fitting. Both analyses have been conducted for the example regression, and the observed model quality is discussed in §4, including fitting statistics and model validation.

3.3. Example comparison: model propensities versus observed hydrogen bonds

Model validation provides a global view of the predictive power of an optimized model. In order to illustrate this power in a specific case, we apply the optimized LHP model to the MIFCEJ example (Fig. 1). The values for the model parameters will be derived for the structure and then LHP π values will be calculated for all possible donor–acceptor pairs. These

may subsequently be compared with the actual bonds observed in the CSD structure. We restate that this forms a true ‘blind test’, as would be a molecule whose crystal structure was unknown, since the structure was not included in the training set for the model.

The example molecule currently has no other polymorphic forms in the CSD. It is pertinent to state that this is no guarantee that other thermodynamically more stable structures may not exist. When comparing the hydrogen bonding of an existing structure with predicted propensities for its donor–acceptor pairs, predictions and the observations should thus be rationalized together. Low predicted propensities may indicate that if the hydrogen bond is observed, the structure to which it belongs could be thermodynamically metastable (or *vice versa*).

In the molecule, there are four potentially hydrogen-bonding functional groups (labelled in Fig. 1):

- (a) NH, of indole;
- (b) secondary amide, CONH;
- (c) primary amide, CONH₂; and
- (d) carboxylic acid, COOH.

The observed hydrogen bonding, as surveyed, is presented in Table 3. Firstly, the observed intermolecular hydrogen bonds are listed, followed by combinations of groups which were not observed to be interacting. The parameter details, as model input for each pair, are also presented here. Note that when identifying the non-interacting pairs, the survey considers only unique functional-group pairs. That is, irrespective of which group donated or accepted an H atom, an XY pair will not be in the list of *False* observations if it is observed as *True*, either as X–H–Y or as Y–H–X. Thus, although both cases can, and do, occur simultaneously (giving two unique *True* observations), the hydrogen bond is maintained as a pairwise interaction in the set of *False* observations. The model results are compared with these observations in the following section.

4. Results and discussion

4.1. Model characteristics

The fitted LHP model equation for the amide dataset (1083 structures) is

$$\begin{aligned} \text{logit}(\pi) = & 2.50 - 0.21\kappa_c - 0.14\rho_D - 0.95\rho_A \\ & + 0.45\Gamma_{D \text{ amido}} + 0.02\Gamma_{D \text{ other}} \\ & + 0.42\Gamma_{D \text{ hydroxy, aliph.}} - 1.35\Gamma_{D \text{ carbox.}} - 2.16\Gamma_{A \text{ other}} \\ & - 1.72\Gamma_{A \text{ ether}} - 0.90\Gamma_{A \text{ carbox.}} - 0.36\Gamma_{A \text{ hydroxy, aliph.}} \end{aligned} \tag{7}$$

The baseline variables (of coefficient = 0) in the model are the secondary amino group as the donor and the amido group as the acceptor (chosen at random at the beginning of the regression procedure). Negative or positive influence is with respect to this baseline and the coefficients are adjusted during the regression to reproduce correctly the observed effect of each parameter.

Table 3
Model parameters of MIFCEJ.

The donors/acceptors are categorized as identified survey groups, with the functional description on the example molecule in parentheses.

Donor, Γ_D	Acceptor, Γ_A	Hydrogen-bonding observation, label (see Fig. 8a)	Competition value, κ	Donor steric value, ρ_D	Acceptor steric value, ρ_A	Hydrogen-bond distance (\AA)
Amido (secondary amide)	Amido (secondary amide)	Intermolecular, α	4	2	1.43	2.837
Amino, secondary (indole)	Carboxyl(=O)	Intermolecular, β	4	5	1	3.061
Amido (primary amide)	Amido (primary amide)	Intermolecular, γ	3	1	1	2.880
Amido (primary amide)	Carboxyl(=O)	Intermolecular, δ	3	1	1	2.987
Carboxyl(OH)	Amido (primary amide)	Intermolecular, ϵ	4	1	1	2.627
Amino, secondary (indole)	Amido (primary amide)	Not bonded	4	5	1	N.a.
Amino, secondary (indole)	Amido (secondary amide)	Not bonded	4	5	1.43	N.a.
Carboxyl(OH)	Carboxyl(=O)	Not bonded	4	1	1	N.a.
Carboxyl(OH)	Carboxyl(OH)	Not bonded	12	1	1	N.a.

Standardized coefficients allow for direct comparison of the relative influence of the model parameters on the likelihood of a hydrogen bond forming. They account for the dependence on different parameter magnitudes by scaling the equation coefficients by their respective estimated standard deviations. The standardized model coefficients are presented in Fig. 5, along with associated uncertainties, calculated with 95% statistical confidence. Comparison of the coefficients reveals that the most influential model parameter is the hydrogen-bond competition function. It is large and negative, thus a donor–acceptor pair with high competition will be much less likely to bond. The steric density function around the acceptor group is also influential in the model, but less so than the competition function. They are both much more influential than the steric density on the donor group. The most positively influential donor groups are the amido and aliphatic hydroxyl groups. More discrimination is observed between particular acceptor groups: the ether and remaining ‘other’ individually less numerous groups are strongly and negatively influential – much less favoured to accept a hydrogen bond than the secondary amino group, and the baseline, amido group.

For a concise quantitative comparison of the categorical model parameters (donor and acceptor type) it is possible to

calculate relative odds ratios using the parameter coefficients. The odds ratio is obtained using the exponent of each coefficient (as is each coefficient’s influence on the resulting logistic function), and so all odds ratios are relative to the parameters assigned zero coefficient magnitude, the baseline variables. The odds ratios for the example donor and acceptor group variables in the model are given in Table 4.

The values show that, within the 95% confidence interval (CI), the amido group and the hydroxyl group are the most prolific hydrogen donors. They are roughly 1.5 times more likely to donate than the baseline, secondary amino, having odds ratios of 1.56, CI (1.28, 1.89) and 1.52, CI (1.21, 1.90), respectively. The carboxyl group is 0.26 times less prolific than the secondary amino group with a CI of (0.20, 0.35). The ‘other’ miscellaneous donors show a similar potential donating ability as the baseline with an odds ratio of 0.98, CI (0.79, 1.22). A similar shared odds ratio such as this indicates that, with all other quantitative parameters being equal, the two groups may be described as of a similar hydrogen-bonding propensity in the set of structures surveyed.

The disparity within the set of acceptors is greater. All groups have odds ratios lower than 1 and hence are less prolific acceptors than the baseline, amido group. The hydroxyl is the next most prolific acceptor, as shown by an odds ratio of 0.70, CI (0.59, 0.83) and the carboxyl group is ranked next, 0.41, CI (0.33, 0.50). The ether and ‘other’ groups are between five and eight times less likely to accept than the amido group, with respective ratios 0.18, CI (0.14, 0.23) and 0.12, CI (0.10, 0.13).

4.2. Model quality

This section discusses the statistical assessment of the example model. Such assessment is vital to ascertain the quality and applicability of one’s choice of model function. A variety of techniques are applied to assess the accuracy

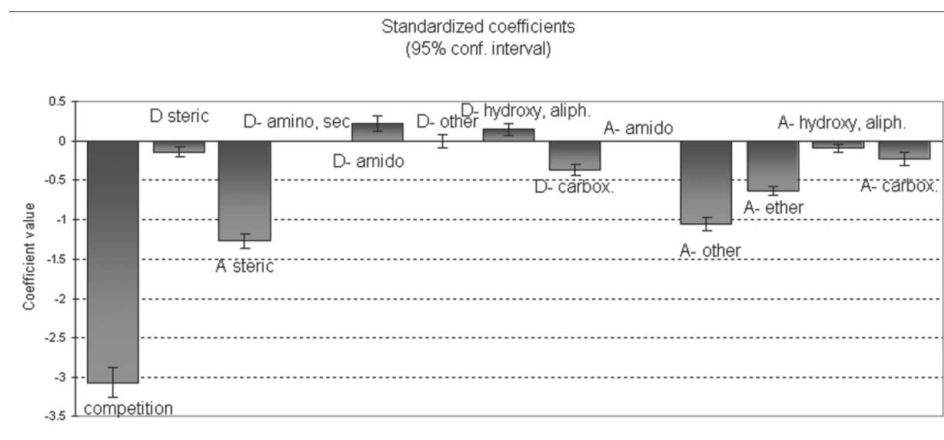


Figure 5
Standardized model coefficients. *D* = functional group as hydrogen donor, *A* = functional group as hydrogen acceptor. The baseline model variables (*D*-amino, *sec*; *A*-amido) are fixed in the regression procedure and the categorical β coefficients are adjusted relative to them. They appear as zeroes with no error bars.

Table 4

Odds ratios for the example qualitative model parameters using baseline donor secondary amino and acceptor amido groups.

Parameter	Model coefficients	Lower bound (95% confidence)	Upper bound (95% confidence)	Odds ratio	CI Lower	CI Upper
Donor group, Γ_D						
Amino, secondary	0.000	–	–	1.000	–	–
Amido	0.445	0.250	0.639	1.560	1.284	1.894
Other	–0.021	–0.238	0.196	0.979	0.788	1.216
Hydroxyl, aliphatic	0.416	0.190	0.643	1.517	1.209	1.903
Carboxyl	–1.345	–1.627	–1.064	0.260	0.197	0.345
Acceptor group, Γ_A						
Amido	0.000	–	–	1.000	–	–
Other	–2.155	–2.266	–2.045	0.116	0.104	0.129
Ether	–1.717	–1.958	–1.476	0.180	0.141	0.228
Carboxyl	–0.903	–1.098	–0.708	0.405	0.333	0.492
Hydroxyl, aliphatic	–0.360	–0.533	–0.187	0.700	0.587	0.830

Table 5

Correlation matrix of the quantitative parameters shows the degree of independence of each model parameter to the model prediction.

Variables	κ	ρ_D	ρ_A
κ	1.000	0.225	0.104
ρ_D	0.225	1.000	0.195
ρ_A	0.104	0.195	1.000

Table 6

Goodness-of-fit statistics comparing the independent LHP model with the optimal model following regression.

The methods are introduced in §4.2.1.

Statistic	Independent model	Converged model
Observations	17 558	17 558
Degrees of freedom	17 557	17 546
–2 log(likelihood)	19 502	11 617
R^2 (McFadden)	0.000	0.404
R^2 (Cox and Snell)	0.000	1.000
R^2 (Nagelkerke)	0.000	1.000
AIC	19 506	11 641
SBC	19 522	11 734
Iterations	0	11

with which the model predicts the *true/false* observations from the CSD survey.

4.2.1. Goodness-of-fit and parameter significance. The model is firstly assessed by the extent to which each variable is independent (*i.e.* providing unique information to the outcome). The correlation matrix, Table 5, represents how similarly the independent quantitative parameters (competition and steric density around both the donor and acceptor) vary with the likelihood of either of the two outcomes. The correlation between the donor steric density function and competition is the largest correlation between different parameters, 0.225. This value is not negligible, but is of an acceptable magnitude. The correlation between competition and acceptor steric density is much lower, 0.104. This may be expected as the competition is more related to donor sites, which may often also be able to accept hydrogen bonds. This would also explain the correlation coefficient of 0.195 between donor steric density and acceptor steric density.

The goodness-of-fit statistics are given in Table 6. They represent how much better the model equation fits the sample data, compared with the independent model (which is the model before optimization of the coefficients in subsequent iterations). $-2 \text{Log}(\textit{Likelihood})$ is often termed the deviance. The R^2 values measure how well the model is adjusted compared with the independent data. They result from a point biserial correlation between predictions and observations. Three methods for estimating this

value are given: McFadden (1973), Cox & Snell (1989) and Nagelkerke (1991). They have a value between 0 for uncorrelated outcomes and 1 for completely correlated outcomes. *AIC* is Akaike’s information criterion and *SBC* is Schwarz’s Bayesian criterion, used to show the relationship between the goodness-of-fit and the number of model parameters used. The above tests may also be referred to in Agresti (1990).

The null hypothesis test, Table 7, checks if the fitted model is significantly more accurate than the null model (a model which gives the ‘null probability’, $P_0 = 0.244$, whatever the value of the training set explanatory variables). If the null model is proved to be significantly worse than the fitted model at reproducing the training data, we conclude the fitted model, known as the *alternative hypothesis* is supported by the data. Three statistics are calculated, all of which follow a χ^2 distribution. If the $\text{Pr} > \chi^2$ value is less than a confidence interval (5% in this example) then the model is said to be significantly better than the null hypothesis. The outcome is $\text{Pr} > \chi^2$ of $< 0.01\%$ in each case, demonstrating the significance of the model fit.

Type III analysis, Table 8, is used to observe the significance of each variable parameter in the model. It recalculates the predictions removing one variable at a time from the model and observes any significant reduction in predictive power. Again, the confidence interval is 5% and we see that all the parameters in this example are significant with a $\text{Pr} > \chi^2$ value of $< 0.01\%$, using two different estimations: Wald’s χ^2 and the LR χ^2 values. Details of both estimates may be found in Agresti (1990). (Note that the significance of the qualitative model parameters has been discussed with regard to odds ratios in the previous section.)

4.2.2. Classification of training set. The ROC curve (receiver operating characteristics) gives a measure of how well classified are the predictions using the training data as a test (Agresti, 1990). The description originates from the field of signal processing, but now finds application throughout the field of statistical analysis. It calculates percentage classification using a variable cut-off, either side of which a propensity is considered positive or negative. The *sensitivity* is the fraction of correct positive predictions and the *specificity* is the fraction

Table 7

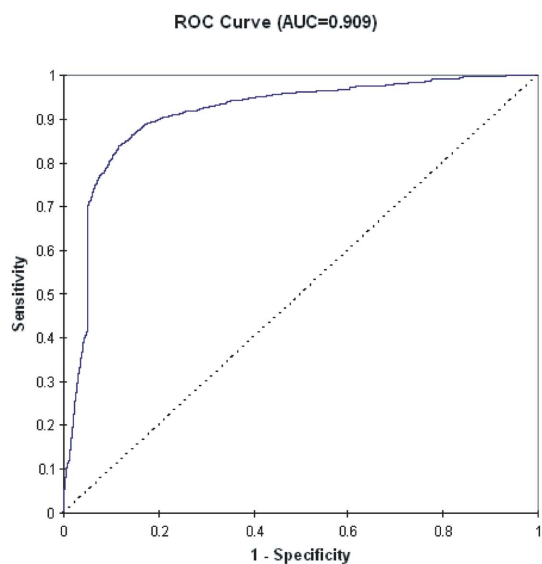
Test of the null hypothesis.

Comparison of the optimal model following regression with a model independent of any parameters (the 'null hypothesis'). DF = degrees of freedom. The methods are introduced in §4.2.1.

Statistic	DF	χ^2	Pr > χ^2
-2 log(likelihood)	12	7885	< 0.0001
Score	12	4960	< 0.0001
Wald	12	3302	< 0.0001

of correct negative predictions. The diagonal dotted line (see Fig. 6) is the outcome of a purely random model as there is an equal number of correct and incorrect predictions. An AUC (area under the curve) greater than 0.5 indicates the model predictions are correct more frequently than a random choice. An AUC of 1 indicates a perfect model: correct every time. An AUC above 0.8 is considered excellent and above 0.9 indicates an outstanding model (Hosmer & Lemeshow, 2000). The difficult middle section around sensitivity/specificity = 0.5 needs a well discriminating model in less extreme cases, and must be well described by the model parameters in order to obtain a high AUC. One may observe that the example LHP model gives an outstanding classification of the training data and achieves an AUC of 0.909.

4.2.3. LHP model validation. Validation of the model is crucial to check that it is not highly dependent on the particular sample data, but also describes model behaviour in independent systems. In order to test the validity of the LHP model to the particular dataset, *holdout validation* was performed. The original dataset of 17 558 potential pairs was split into two subsets: 8297 random pairs were removed from the set to leave 9261 remaining pairs (a ratio of 47% to 53% of the complete dataset). Those remaining pairs were used as

**Figure 6**

ROC curve using the model to predict the training set outcomes. Sensitivity and specificity are defined in the text (§4.2.2). The diagonal dotted line indicates the curve of a model with no predictive power: there is equal likelihood of a correct and an incorrect prediction.

training data for the regression and following convergence, the resulting model equation was used to check the predicted outcomes for the 8297 data in the validation set.

Each donor–acceptor pair (hydrogen bonded or not) in the training set can be considered a unique source of explanatory data to go toward the model fit. However, groups of pairs may originate from the same crystal structure. Thus, in order to avoid any information cross-over from validation set to training set, groups of pairs originating from a particular crystal structure were kept together in one of the two sets. Thus, no crystal structure was represented in both the model fit and validation.

Classification tables (also sometimes referred to as confusion matrices) show the extent of correct *versus* incorrect model predictions with respect to the reference data. These tables for both the training set and validation set are presented in Tables 9(a) and (b).

Classification of the training data is presented in Table 9(a), which lists correct and incorrect results following a model fit on data from 2355 hydrogen-bonded pairs and 6906 non-hydrogen-bonded pairs. This model was then used to predict propensities for the validation set, of which there were 1925 pairs observed to be hydrogen bonded in the sample, and 6372 pairs not hydrogen bonded. Classification of this validation data is then shown in Table 9(b). None of the information for the latter set of pairs was used in the model regression. The resulting propensity values for 87% (7220 of 8297) of the samples had the correct result, with a cut-off of $\pi = 0.35$. [This value was chosen as a value that gave balance in the percentage classification of both positive (sensitivity) and negative outcomes (specificity).] (A more informative representation in general is the ROC curve which shows how classification varies with cut-off; see Fig. 6.)

As the model is fitted, within the potential of the parameters in the model, to reproduce the input observations, this validation is necessary to check that the model is not highly dependent on the particular sample data. Since the model set reproduces in the validation set the percentage correct classification of the training set, to well within the uncertainties of the coefficients in the model, one may be satisfied that such dependence on input data does not exist in this case. Thus, the underlying trends modelled in the example would appear to be more universal and re-observable. Qualitatively identical results were also obtained using a random choice validation set with 12 000 data, leaving 5558 data for the training set, and also using variations in the random sampling of validation data.

4.3. Results: example comparison

Possible hydrogen bonds between donor and acceptor atoms in the example molecule, Fig. 1, have been given propensity scores using the fitted LHP model, and compared against the observed behaviour in the crystal structure, MIFCEJ. Visualizations of the three-dimensional structure of MIFCEJ, using the *Mercury* program (Macrae *et al.*, 2006), are given in Fig. 8. Specific bonding groups and labelled bonds can

Table 8

Type III analysis.

The significance of the model parameters is tested by sequentially removing one and observing any reduction in model quality. DF = degrees of freedom.

Source	DF	χ^2 (Wald)	Pr > Wald	χ^2 (LR)	Pr > LR
Competition function	1	1304	< 0.0001	1304	< 0.0001
Donor steric function	1	18.39	< 0.0001	18.39	< 0.0001
Acceptor steric function	1	817.2	< 0.0001	817.2	< 0.0001
Donor-amino, secondary	4	265.0	< 0.0001	265.0	< 0.0001
Donor-amido	4	1555	< 0.0001	1555	< 0.0001

Table 9

Classification tables comparing hydrogen-bond prediction *versus* observation using the validation model, fitted with a training set of 9261 pair data, and validated with a separate set of 8297 pair data.

(a) Model predictions for training set.

From/to	Predicted no hydrogen bond	Predicted hydrogen bond	Total pairs	% correct
Not hydrogen bonded in sample	6022	884	6906	87.20
Hydrogen bonded in sample	364	1991	2355	84.54
Total	6386	2875	9261	86.52

(b) Model predictions for validation set.

From/to	Predicted no hydrogen bond	Predicted hydrogen bond	Total	% correct
Not hydrogen bonded in sample	5602	770	6372	87.92
Hydrogen bonded in sample	307	1618	1925	84.05
Total	5909	2388	8297	87.02

Table 10

Propensity predictions for donor–acceptor atom pairs in MIFCEJ.

The donor–acceptor are categorized as identified survey groups, with the functional description on the example molecule in parentheses.

Donor, Γ_D	Acceptor, Γ_D	Hydrogen-bonding observation, label (see Fig. 8a)	Propensity, π
Amido (secondary amide)	Amido (secondary amide)	Intermolecular, α	0.614
Amino, secondary (indole)	Carboxyl(=O)	Intermolecular, β	0.367
Amido (primary amide)	Amido (primary amide)	Intermolecular, γ	0.772
Amido (primary amide)	Carboxyl(=O)	Intermolecular, δ	0.579
Carboxyl(OH)	Amido (primary amide)	Intermolecular, ϵ	0.314
Amino, secondary (indole)	Amido (primary amide)	Not bonded	0.589
Amino, secondary (indole)	Amido (secondary amide)	Not bonded	0.488
Carboxyl(OH)	Carboxyl(=O)	Not bonded	0.157
Carboxyl(OH)	Carboxyl(OH)	Not bonded	0.033

be found in Fig. 8(a). The calculated propensities are displayed in Fig. 7, and the combination of the predictions and the observations may also be viewed in Table 10.

With a cut-off criterion of $\pi = 0.35$ to distinguish hydrogen bonds as likely or not (this is a value which gives a similar percentage of correct classification of both outcomes), all the observed bonds are predicted to form (although there is some uncertainty in two predictions, see below). The calculations reveal, given no prior three-dimensional geometric information, that the most likely hydrogen bond to form is the primary amide–primary amide bond [denoted amido(1°)–amido(1°) on the chart], with a propensity of 0.772, *i.e.* a likelihood of

greater than 77%. The second amido pair [amido(2°)–amido(2°)] involving the secondary amide fragments is predicted to be less likely to form, but is ranked the second most likely pair, 61%. From the structural survey we can see that both hydrogen bonds are observed (see Fig. 8a, bonds α and γ). There are three remaining observed hydrogen bonds. The primary amide and carboxyl groups (carbonyl oxygen) have a predicted propensity of 0.58 (δ in Fig. 8a), the secondary amino and carboxyl groups have a propensity of 0.37 (β in Fig. 8a) and the carboxyl and primary amide [amido(1°)] have a propensity of 0.31 to form a hydrogen bond (ϵ in Fig. 8a). In the latter two cases, the error bars span the cut-off, and although this was an arbitrary choice to give a degree of equality in the fraction of correct positive and negative predictions, this reflects an uncertainty in predictions using the propensities of a mid-ranged value.

Considering the non-interacting pairs in the structure, the model accurately predicts the propensities of bonding for the carboxyl–carboxyl pair, with either the carbonyl oxygen [carboxyl–carboxyl(=O)] or hydroxyl oxygen [carboxyl–carboxyl(OH)] as an acceptor, of 15% for the former and 3% for the latter, *i.e.* very unlikely for either acceptor site in the functional group. This ability to identify pairs that are unlikely, and perhaps unfeasible as a hydrogen-bonded pair, is of great value, for example, in reducing the size of the search space in a crystal-structure prediction procedure, when iterating through permutations of hydrogen-bonded networks in sets of equi-

energetic candidates.

The remaining pairs observed not hydrogen bonding, which involve the amido fragments as an acceptor, are predicted to be quite likely to bond, despite this not being the case in the structure. Both the secondary amino with either primary or secondary amides are predicted to bond; their predicted propensities being 0.59 and 0.49, respectively. The amide functionality in this structure is clearly a strong hydrogen-bond former, with the other pairings involving this group observed to be hydrogen bonded and scoring highly. We may observe, however, that all but one of the predicted propensities for pairs involving the amido group not found to be

hydrogen bonded are lower than those that are observed. Such simultaneous hydrogen-bond formation is much less likely.

Of course, treating possibilities in this pairwise manner ignores combinatorial aspects. That is, if a pair is considered as being hydrogen bonded, the likelihood of a second hydrogen bond forming is dependent on new factors, and could possibly be much lower. (This behaviour itself could be measured in the CSD in future work). It is relevant to restate that the propensity score is a *potential* pairwise probability considering a discrete choice involving all alternatives. For mid-ranged predictions only speculative assertions can be made.

It is worth noting the degree of success of the model with this example. Five of the nine pairings have been correctly predicted; three observed as hydrogen bonded and two otherwise. Two of the nine pairings are wrongly classified: both not observed as hydrogen bonded. There are two predictions whose outcome is uncertain at a 95% confidence interval and a cut-off of 0.35. Thus, five of a discernable seven predictions, 71%, are classified correctly, and the model then succeeds less decisively for this example case than the average of 87% in the validation set of > 8000 pairs. Other structures in the validation set are then likely to show more accurate discrimination over the full set of donor–acceptor pairs. Nonetheless, it is the difficult cases for which the model does not comparatively succeed by which one may learn and refine any model description. An advantage of the logit method is that predictions for ambiguous cases may be improved by further model descriptors (*e.g.* the presence of precursory intramolecular

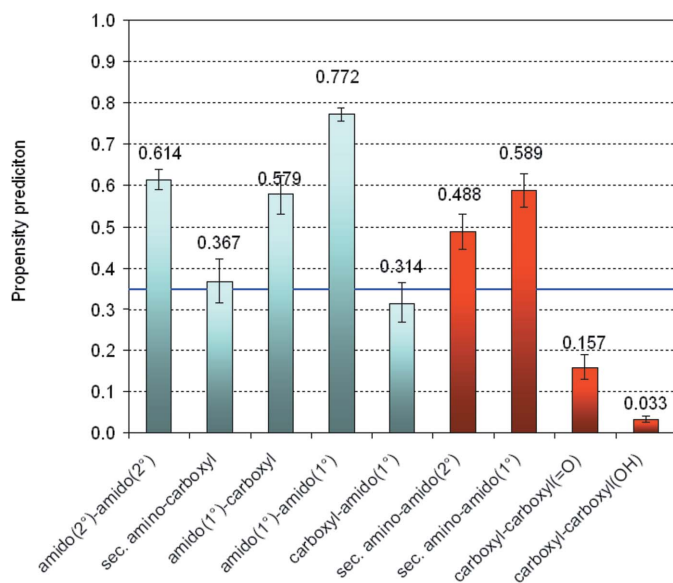


Figure 7

Histogram showing predicted hydrogen-bond propensities for each donor–acceptor pair on the molecule *N*²-(indol-3-ylacetyl)-*L*-asparagine. Each pair is labelled by its associated model group, Γ , with any required distinction given in parentheses. The parameters for each pair and corresponding functional group on the molecule are displayed in Table 3. Predictions for pairs observed to be hydrogen bonded in the structure are coloured light blue and pairs observed not to be hydrogen bonded are coloured red. The criterion as to whether a hydrogen bond between the pair is more or less likely is above or below a propensity value of 0.35 (see text, §4.3), displayed as a horizontal blue line to aid comparison.

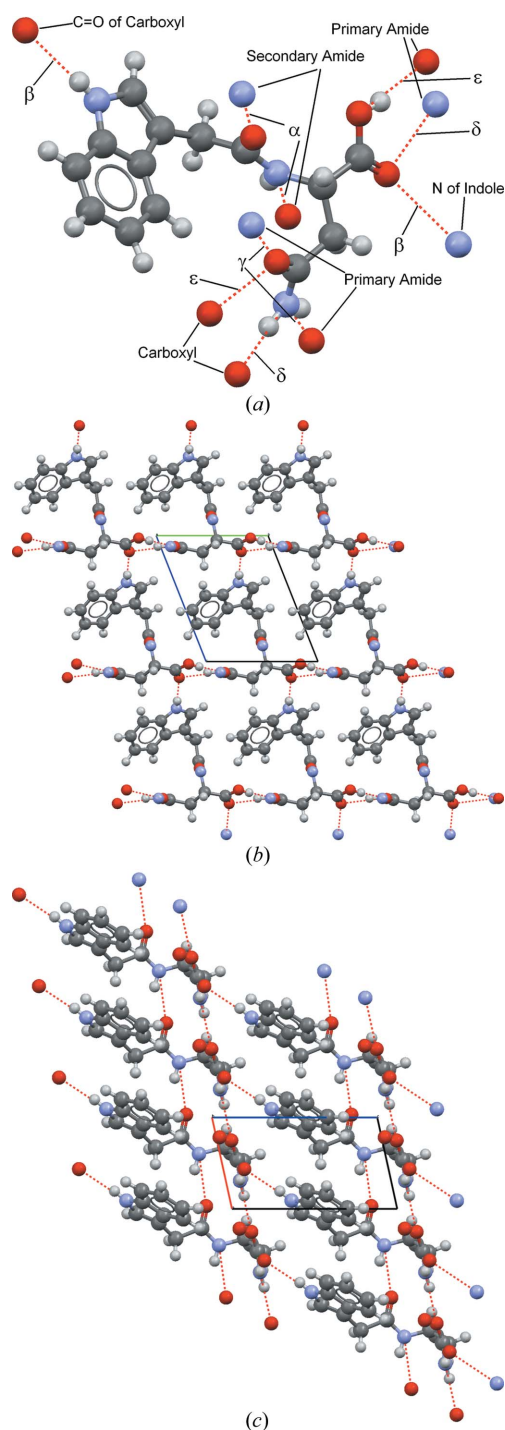


Figure 8

The three-dimensional structure of MIFCEJ, displayed using *Mercury* (Macrae *et al.*, 2006). (a) The hydrogen bonds (shown as red dashed lines) in the MIFCEJ structure. The bonds are labelled (α – ϵ) as described in Table 3. Adjacent hydrogen-bonding donor and acceptor groups to the central molecule are labelled. (Donor H atoms from the adjacent groups are not displayed). (b) Packing of MIFCEJ viewed in the *a* projection. Hydrogen bonds show continuous chains in the *b* (carboxyl primary amide links) and *c* directions (indole C=O of carboxyl links). (c) Packing of MIFCEJ viewed in the *b* projection. Hydrogen bonds show two types of continuous chain in the *a* direction (primary amide–primary amide and secondary amide–secondary amide links).

hydrogen bonds). Given the form of the model, these can be included straightforwardly, and their added influence may be systematically observed.

Because both the most likely and unlikely donor–acceptor pairings are revealed as correct predictions on analysis of the structure in the example, MIFCEJ, we may conjecture that this structure is the thermodynamically stable crystal form. There are some mid-ranged propensity predictions that are less decisive and we suggest that these potential bonds, although not as likely, may be favourable in an alternative conformation. If the formation of a majority of the set of likely interactions is possible, other metastable polymorphs cannot be ruled out. Subsequent polymorph screening on the molecule may in the future verify these conjectures.

5. Conclusions

A new hydrogen-bond survey method has been presented, based on the CSD, which extracts descriptive molecular and chemical properties. The properties are designated as discriminatory variables in the formation of a hydrogen bond between a specified donor and acceptor atom pair. A new predictive statistical model has been developed, denoted the logit hydrogen-bonding propensity (LHP) model, which applies the variables to define a binary likelihood: the propensity of bond formation between the pair. Initial results are promising. The model gives correct classification for ~90% of sample donor–acceptor pairings, of a set of 17 558 pairs from 1083 organic crystal structures in the CSD.

A specific example structure, MIFCEJ, from the surveyed set of crystals was analysed in detail. The observed hydrogen bonds in the structure were compared with the model propensities, and an important discrimination was observed between likely and unlikely pairs to form a bond. This strong attribute of the model quickly identifies unlikely bond pairs and may be used, for example, to identify less-common structural factors.

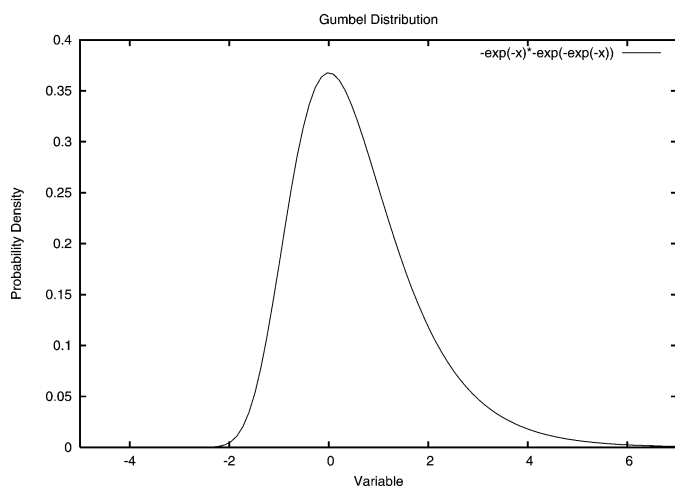


Figure 9

The Gumbel distribution. The distribution is that assumed for the error in the utility in the logit model.

The model was validated using a smaller training set, as a subset from the original data, and assessed for goodness-of-fit using the remaining excluded subset for validation. The validation shows that the strong discrimination of likely *versus* unlikely hydrogen bonds is maintained for similar crystals, despite no information from these being used as input in the model. We conclude that this initial LHP model is quite successful, in that:

- (i) it has few parameters, all of which are chemically pertinent,
- (ii) it has a high degree of correct classification and
- (iii) it reproduces successful discrimination for new or predictive cases, extra to cases which form the model data.

The model has much flexibility, improvements and refinements being quite accessible in the current framework. Work is being undertaken to study the success of fitted LHP models in a bid to better understand those structures that form the most notable failures. New or improved model parameters may be designed based on these studies. Future publications are planned presenting our progress and application of the approach to a variety of examples from different chemical families.

APPENDIX A

Summary of statistical theory

This appendix contains a short summary of the mathematical theory which comprises logistic regression. As well as providing further information for the interested reader, it provides some justification for the present use of the theory, given the nature of the sample data studied here.

A1. Assumptions on the hydrogen bond

The (potential) hydrogen-bonded pairs in the sample data are considered to satisfy only one of two outcomes: either they are hydrogen bonded in a crystal structure or they are not.

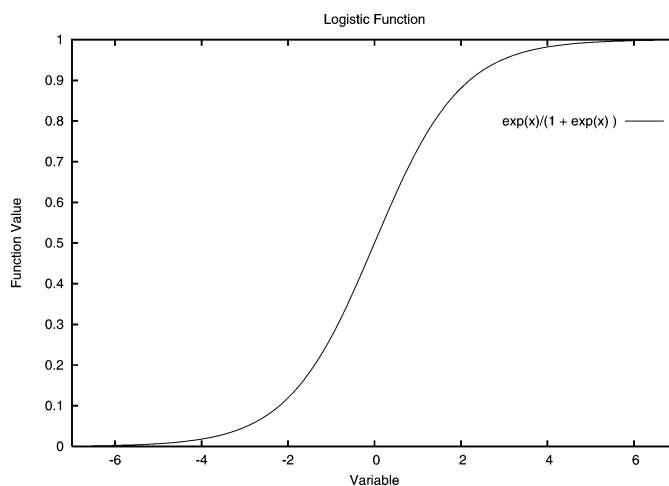


Figure 10

The Logistic function. The ordinate axis variable is the value of the utility and the abscissa value gives a prediction of propensity.

This is known as a *dichotomous response variable*, for which a *Bernoulli*-distributed probability estimate can be obtained. This type of estimate is a *discrete* distribution. A further assumption is that the *actual* hydrogen bonds observed in a crystal structure are considered to be those with the *highest likelihood* of forming among all possible donor–acceptor pairs.

The particular discrete choice model selected is the logit model. It is suited to the current application given its specific deterministic and stochastic probabilities for a choice. The logit model assumes the predicted likelihoods are *disaggregated*, *i.e.* based on individual attributes. There are a finite number of alternatives, each having a number of attributes that determine the *preference (indifference)* for a choice. A degree of *utility* for a choice is defined, given the model assumes the preference has zero uncertainty. Utility is represented by a random model, providing a deterministic decision process.

A2. Utility function

Within the logit model, the utility is modelled assuming that the choice will be made with perfect discrimination, but the analyst has limited or partial information (a limited extent of available descriptive data). The *utility function* is composed of a deterministic part, modelling the underlying choice discrimination, and a stochastic part for the uncertainty in the representation of the sample data.

The utility that individual *i* associates with alternative *a* is given by

$$U_a^i = V_a^i + \varepsilon_a^i, \quad (8)$$

where V_a^i is the deterministic utility and ε_a^i is a stochastic variable, capturing the uncertainty. The assumption is that the alternative with the highest utility is that which is chosen, *b*, thus the probability of the optimal choice P_c^i , from the choice set *c*, is given by

$$P_c^i(b) = P[\max_{a \in c}(U_a^i)]. \quad (9)$$

A2.1. Stochastic utility. The logit model is derived from the assumption that error terms in the utility function per choice are independent and identically Gumbel distributed. The probability distribution function (displayed in Fig. 9) arises in extreme value theory and is a specific case of the Fisher–Tippet distribution [see (10)], around mean $\mu = 0$ and width coefficient $\beta = 1$ (see also *e.g.* McFadden, 1973). It approximates the normal distribution around the mean, shares the same mode, but has large positive skew. This error-distribution estimate assists in maximization of the utility function.

$$P(x, \mu, \beta) = \frac{\exp\left(\frac{-x-\mu}{\beta}\right) \cdot \exp\left(-\exp\left(\frac{-x-\mu}{\beta}\right)\right)}{\beta} \quad (10)$$

A2.2. Deterministic utility. Given independently Gumbel distributed error terms in a utility prediction for alternative, *a*, one can build the underlying deterministic probability distribution. This is given by the logistic function

$$\pi_c^i(a) = \frac{\exp(U_a^i)}{1 + \exp(U_a^i)}, \quad (11)$$

which can be seen displayed in Fig. 10. A probability estimate is derived by an approximation to the deterministic utility, V_a^i , to complete the utility estimate. This is commonly expressed as a linear function of explanatory variables

$$V_a^i = \alpha + \sum_k x_k^i \beta_k. \quad (12)$$

It is defined by α as the baseline variable or intercept and x_k^i as a vector representing all *k*-independent attributes, whose influence on the model is controlled by the magnitude of the *k*th coefficient in the set, β_k . The parameters are incremental effects to an arbitrary baseline parameter. A deterministic utility function that is linear in the explanatory variables is much more flexible than it may first appear, as functions of the parameters within the sum can also be included, *e.g.* a quadratic function. Changes in observed likelihood with one parameter can reveal these non-linear relationships, and the effects of the resultant model fit should be examined to reveal any improvements from a linear model.

The complete probability estimate is written by inserting V_a^i into (11)

$$\pi_{c,k}^i(a) = \frac{\exp(\alpha + \sum_k x_k^i \beta_k)}{1 + \exp(\alpha + \sum_k x_k^i \beta_k)}. \quad (13)$$

A set of $\pi_{c,k}^i$ values can be determined from a sample data set to approximate P_c^i . Thus, the function is fitted on the logit scale. Inversion of (11) reveals the relation of the log odds (the ratio of positive to negative probability of an outcome, also denoted the logit) and the linear utility function of the explanatory variables

$$\text{logit}(\pi_{c,k}^i) = \log\left(\frac{\pi_{c,k}^i}{1 - \pi_{c,k}^i}\right) \quad (14)$$

$$= \alpha + \sum_k x_k^i \beta_k. \quad (15)$$

A3. Model regression

To fit the model one must maximize the *likelihood function* (which produces function parameters from given probabilistic outcomes). *Maximum likelihood* algorithms can be applied in this case to approximate the form of the probability function which is most likely to have generated a particular probability score. There is no analytical expression for the constraint in this model and so the procedure is done iteratively, often using Newton's method or related scheme. Once the model parameters are obtained, $\pi_{c,k}^i$ enables a comparison of fitted values on the probability scale, or $\text{logit}(\pi_{c,k}^i)$ enables a comparison of ratios of fitted variables. The regression output lists the values of the intercept and the β coefficients, defining the specific model equation for the training data.

The functional form of the competition function described in the text is developed from an original concept by Dr James Chisholm. The authors wish to thank Dr John Liebeschuetz for a critical reading of the manuscript.

References

- Aakeröy, C. B. (1997). *Acta Cryst.* **B53**, 569–586.
- Addinsoft (2006). *XLSTAT* 2006 v. 2006. <http://www.xlstat.com>.
- Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley.
- Allen, F. H. (2002). *Acta Cryst.* **B58**, 380–388.
- Allen, F. H., Davies, J. E., Galloy, J. J., Johnson, O., Kennard, O., Macrae, C. F., Mitchell, E. M., Mitchell, G. F., Smith, J. M. & Watson, D. G. (1991). *J. Chem. Inf. Comput. Sci.* **31**, 187–204.
- Allen, F. H., Motherwell, W. D. S., Raithby, P. R., Shields, G. P. & Taylor, R. (1999). *New J. Chem.* pp. 25–34.
- Allen, F. H. & Taylor, R. (2005). *Chem. Commun.* pp. 5135–5140.
- Antolic, S., Kveder, M., Klaić, B., Magnus, V. & Kojic Prodic, B. (2001). *J. Mol. Struct.* **560**, 223.
- Bilton, C., Allen, F. H., Shields, G. P. & Howard, J. A. K. (2000). *Acta Cryst.* **B56**, 849–856.
- Böhm, H.-J. & Klebe, G. (1996). *Angew. Chem. Int. Ed. Engl.* **35**, 2589–2614.
- Braga, D., Grepioni, F. & Desiraju, G. R. (1997). *J. Organomet. Chem.* **548**, 33–43.
- Bruno, I. J., Cole, J. C., Edgington, P. R., Kessler, M., Macrae, C. F., McCabe, P., Pearson, J. & Taylor, R. (2002). *Acta Cryst.* **B58**, 389–397.
- Bruno, I. J., Cole, J. C., Lommerse, J. P. M., Rowland, R. S., Taylor, R. & Verdonk, M. L. (1997). *J. Comput. Aided Mol. Des.* **11**, 525–537.
- Chisholm, J., Pidcock, E., Van de Streek, J., Infantes, L., Motherwell, W. D. S. & Allen, F. H. (2006). *CrystEngComm*, **8**, 11–28.
- Cormen, T. H., Leiserson, C. E. & Rivest, R. L. (1989). *Introduction to Algorithms*. Cambridge, MA, USA: MIT Press.
- Cox, D. R. & Snell, E. J. (1989). *The Analysis of Binary Data*, pp. 208–209. London: Chapman and Hall.
- 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., Chisholm, J., Shan, N., Motherwell, W. D. S. & Jones, W. (2004). *Cryst. Growth Des.* **4**, 1327–1340.
- 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.
- Etter, M. C. (1991). *J. Phys. Chem.* **95**, 4601–4610.
- Haynes, D. A., Chisholm, J. A., Jones, W. & Motherwell, W. D. S. (2004). *CrystEngComm*, **6**, 584–588.
- Hosmer, D. W. & Lemeshow, S. (2000). *Applied Logistic Regression*. New York: Wiley.
- Infantes, L. & Motherwell, W. D. S. (2004). *Chem. Commun.* pp. 1166–1167.
- Macrae, C. F., Edgington, P. R., McCabe, P., Pidcock, E., Shields, G. P., Taylor, R., Towler, M. & van de Streek, J. (2006). *J. Appl. Cryst.* **39**, 453–457.
- McFadden, D. (1973). *Frontiers in Econometrics*, edited by P. Zarembka, pp. 105–142. New York: Academic Press.
- Motherwell, W. D. S. (1999). *Nova Acta Leopold.* **79**, 89–98.
- Nowell, H. & Price, S. L. (2005). *Acta Cryst.* **B61**, 558–568.
- Nagelkerke, N. J. D. (1991). *Biometrika*, **78**, 691–692.
- Parkin, A., Barr, G., Dong, W., Gilmore, C. J. & Wilson, C. C. (2006). *CrystEngComm*, **8**, 257–264.
- Price, S. L. (2004). *Adv. Drug Deliv. Rev.* **56**, 301–319.
- Streek, J. van de & Motherwell, S. (2005). *Acta Cryst.* **B61**, 504–510.