Building blocks of crystal engineering: a large-database study of the intermolecular approach between C-H "donor" groups and O, N, Cl or F "acceptors" in organic crystals

Angelo Gavezzotti[1] and Leonardo Lo Presti*[1,2,3]

1. Università degli Studi di Milano, Department of Chemistry, Via Golgi 19 I−20133 Milano, Italy
2. Center for Materials Crystallography, Aarhus University, Langelandsgade 140, 8000 Aarhus, Denmark
3. CNR-ISTM, Istituto di Scienze e Tecnologie Molecolari, Via Golgi 19 I−20133 Milano, Italy

* To whom correspondence should be addressed: leonardo.lopresti@unimi.it, Phone: +39-02-50314242, Fax: +39-02-50314300

Abstract

The nature of CH···X interactions in organic crystals, X being an electronegative atom, has been the subject of extensive consideration with sometimes contradictory results and ensuing opinions. We perform statistical analysis on large databases of crystal structures retrieved from the Cambridge Structural Database. Crystals containing only C-H donors only are considered, in conjunction with each of the title acceptors in turn. The analysis of Coulombic-polarization and dispersion components reveals that the lattice energies of these crystals are largely dominated by dispersive interactions. The frequency of short H···X contacts decreases through the series CHO > CHN > CHCl > CHF, being just sporadic in the latter. The presence of such contacts is positively correlated with the Coulombic contribution to molecule-molecule interaction energies, but do not generally determine the pair energy. Short CH···O or CH···N contacts are often relegated to weakly bound pairs; their minor energy contributions might be relevant in driving crystal packing of small molecules, where the contact energy is a substantial part of the lattice energy. In reproducible crystal engineering, and even more in crystal structure prediction, weak CH···X contacts are seldom responsible for the whole picture and the wider context of competing energies should be considered.

1. Introduction

The nature and strength of molecular interactions in crystals is the key issue in solid-state organic chemistry. The benefits of a consistent understanding, prediction and control of such interactions, in all fields of pure and applied chemistry, are evident. The mainstream approach to such a problem is a quantitative evaluation of the cohesive energies resulting from cooperation phenomena among the electron distributions interacting entities. This is nowadays possible by widely accessible quantum chemical methods with benchmark calculations,[1–3] and also by carefully calibrated semiempirical schemes such as PIXEL[4] which has been shown[5] to provide results comparable by high-level quantum chemical methods, or q-GRID,[6] based on the use of electron densities directly derived from the crystal. Moreover, the wealth of information provided by the Cambridge Structural Database[7] allows efficient and robust statistical studies on large numbers of crystals.

The present paper deals with a systematic analysis of the structural and energetic properties of selected groups of crystal structures containing only C, H, O, or C, H, N, or C, H, F, or C, H, Cl atoms, for the general aspects of dispersive *vs.* Coulombic intermolecular bonding, and with special attention to the possible relevance of C-H⋯heteroatom contacts. For these contacts, an extensive literature exists under the broad name of "weak bond" interaction,[8–12] with sometimes reliable, sometimes controversial[13–15] results. In this work, by means of principal component analysis (PCA), multivariate correlations are obtained between global descriptors such as molecular mass, overall polarity, relative volumes, densities and packing coefficients, as well as lattice energies and their Coulombic, polarization-dispersion and "Pauli-repulsion" components. A survey is then made of intermolecular distances between CH_n (n = 1-3) groups, called "donors", and the heteroatoms called "acceptors" in analogy with the traditional hydrogen bond nomenclature. The contacts are rather arbitrarily classified as worth of attention when shorter than the sum of H and X random contact radii, or of some percent thereof, and are then matched to the intermolecular interaction energies between the molecular pairs that bear them in search of correlation between geometry and energetic relevance.

The results should provide the theoretical and applied chemist with at least some guidelines for understanding when or whether an atom-atom close contact may imply a true bonding relationship with decisive consequences on crystal packing, as opposed to conditions in which such a contact may result from randomness or even be the unexpected consequence of other, stronger requirements from other parts of the molecular electron density.

2. Methods and procedures

2.1 Crystal structure retrieval. Data sets are retrieved from the Cambridge Structural Database (CSD)[7] with the following specifications: no more than one chemical unit in the asymmetric unit (Z'≤ 1), number of atoms with complete coordinates ≤ 30, $R \leq 5\%$, one chemical residue per structure, no disorder, no errors, no ions, no powder structures. Duplicates are eliminated by keeping the structure determination with the lowest R-factor; hydrogen atom positions are renormalized to the usual C-H distance of 1.08 Å, mediated from

recent results[16] into our general-purpose program manifold.[17] The CHO-*all* dataset, with C, H and O atoms only, consists of 1059 crystal structures, and a subset, the CHO-*pix* dataset of 376 crystals with smaller molecules, is selected for the calculation of lattice energies and molecule-molecule energies by PIXEL.[4] The CHN-*all* dataset, with C, H and N atoms only, consists of 478 crystal structures, and the CHN-*pix* subset of 174 structures. The CHCl-*all* dataset with C, H and Cl atoms only has 221 crystal structures, and the CHCl-*pix* subset has 70 structures. The CHF-*all* dataset with C, H and F atoms only has 117 crystal structures, but no PIXEL calculations were carried out. For CHCl and CHF the maximum number of atoms was set at 35, and the maximum R at 7.5% in order to increase the data number for better statistics. Compounds with a very low hydrogen/halogen ratio (e.g. pentachlorobenzene or perfluorohalocarbons, with typically H/X ≤ 1/5) were sorted out by hand and discarded to ensure a proper donor-acceptor ratio. Moreover, neither oxygen- nor nitrogen-bonded H donors were included in the datasets, as disentangling the weight of CH···X weak interactions in determining pairing and lattice energies when stronger XH···Y (X, Y=O, N) interactions are present might be problematic.

2.2 Evaluation of short contacts. Standard Atomic Radii (SAR) $R°$ are taken[17] as 1.10 (H), 1.58 (O), 1.64 (N), 1.46 (F) and 1.76 (Cl) Å. Recently proposed new values differ marginally.[18] Crystal structures are scanned for C-H···X intermolecular distances with $R_{HX} < P(R°_H+R°_X)$, where the quantity in parenthesis is the sum of SAR (SSAR). P is a cutoff tuning factor. Since there is no unique absolute definition for shortness, P values of 0.9, 0.95 or 1.0 were considered. Any H···X distance satisfying the above inequality is called an "extremely short" ($P = 0.9$), "very short" ($P = 0.95$) or "short" ($P = 1.0$) contact according to the selected level of the P cutoff. Obviously, categories with smaller P are subset of the P=1 category, where the term "short" includes anything below the sum of radii. In addition, a short-bond index B_s was prepared as:

$$B_s = \sum_k \left[100 \cdot \frac{\left(R^0_{HX} - R_k \right)}{R^0_{HX}} \right] \qquad (1)$$

where R_k is a short H···X distance, $R°_{HX}$ the corresponding SSAR, and index k runs on all short contacts found in a crystal structure. B_s is thus a summation of percent reductions in contact distance with respect to SSAR and is supposed to describe the frequency and importance of atom-atom contacts.

2.3 Association of short contacts with molecule-molecule energies. Each H···X contact is associated with a pair of molecules in the crystal, one carrying the donor and the other the acceptor. Each molecular pair may feature one or more contacts. Since apportioning energies over pairs of atoms is physically meaningless in our approaches, only the total interaction (pairing) energy for each molecular pair is evaluated by PIXEL in the CHO-*pix*, CHN-*pix* and CHCl-*pix* datasets. Molecular pairs in each crystal are then ranked in descending order of interaction energy. Each H···X atom-atom contact is therefore associated with the Coulombic (E_c), polarization (E_p), dispersion (E_d), repulsion (E_d) and total pairing energy (E_t) of the molecular pair to which it belongs, as well as to its energy rank. This procedure is an attempt to assess the importance of CH···X contacts in determining a high energetic rank of the corresponding molecular pair, a way of weighing geometry with energy.

2.4 Principal component analysis. The PIXEL calculation also produces total lattice energies with subdivision in the various contributions. The *-pix* datasets were used for the principal component analysis (PCA), a statistical non-parametric method employed to extract relevant information from redundant and noisy data sets.[19] In practice, an orthogonal transformation is applied to observations to define a new set of (possibly) uncorrelated variables called principal components (PC). The latter are eigenvectors of the data covariance matrix and are weighted linear combinations of the original variables. PCA is just a change of basis in the hyperspace of variables, so that the new reference system is aligned with the independent directions where top variance is expected. The number of significant PC is always much smaller than the number of original variables, so PCA identifies the factors that account for predominant data variability. Such factors highlights the interplay of the underlying physics, largely reducing statistical noise.

The correlation coefficients between PC and original variables are called loadings,[19] and provide the weights by which the original variables are combined into the PC. The scores are the coordinates of the original data in the PC reference frame. Scores and loadings are somewhat prone to subjective interpretation, but when carefully considered they provide an understanding of the degree of cooperation of the original physical parameters. PCA was here applied to the three *-pix* databases (see above) for which PIXEL energies were computed. Observations were always standardized so that each variable had 0 mean and unit variance. A total of eleven potentially structure-defining continuous variables were considered for each crystal structure with its constituent molecule, including packing efficiency and polarization descriptors, plus usual PIXEL energy terms (see Section 3.6 below). PC were defined, according with the procedure above sketched, as suitable linear combinations of former variables running along the independent directions of maximum data variance. Moreover, a further transformation of the PC reference frame according to the varimax technique[20] was carried out. It consists of a rotation of the orthogonal PC basis, so that the sum of the squared loadings is maximized in the new reference frame. The purpose is to split the original variables into (possibly) disjoint sets, whose loadings are high for one PC and low for all the others so that a given PC is defined by fewer variables and interpretation in terms of recognizable physical effects becomes easier.

2.5 Detail of programs and reproducibility. All procedures for the retrieval and checking of crystal structures, for the reassignment of H-atom positions,[17] for finding short atom-atom distances and for atom-atom energy calculations are carried out with in-house software, as usual available for download from www.angelogavezzotti.it, plus standard CSD software (Mercury[21]). Electron densities for PIXEL calculations are obtained at MP2/6-31G** level from GAUSSIAN03.[22] The TANAGRA data-mining tool[23] is employed for PCA. Lists of the crystal structure CSD refcodes are available in the Supporting Information (SI, Tables S1–S4).

3. Results and discussion

3.1 Overall properties. Table 1 shows the size and composition of compounds considered in this study. As concerns the analysis of short contacts, all databases provide a sufficient number of acceptors to ensure significant averaging. There is a dearth of acetylenic hydrogens due to some (practical, or experimental) bias in the crystal structure determination and deposition. As appears from the high percent of $C(sp^2)$ and H-

C(sp^2) hydrogens, aromatic systems are predominant in the nitrogen database, where also sp^2 nitrogen is frequent due to the conspicuous presence of nitrogen heterocycles. The chlorine database seems more balanced between aliphatic and aromatic composition, while the oxygen database has a relatively high percent of sp^3 CH groups. Such observations may be important because the flat shape of aromatic systems may facilitate intermolecular contacts of their peripheral monovalent atoms (H, F, Cl), less prone to steric screening than in globular or flexible molecules. The same kind of purely geometric considerations suggest that carbonyl oxygen is more accessible than ether oxygen, and that the order of accessibility to nitrogen is ≡N > =N- > aliphatic N. These geometrical biases should be kept in mind when judging statistical results.

Table 1. Chemical composition of the databases: molecular size and presence of each atom type.

	database			
	CHO	CHN	CHCl	CHF
total structures	1059	478	221	117
total atoms	29733	11909	5507	2703
donor/acceptor ratio	46%/11%	42%/14%	37%/16%	30%19%
$0<N_{atom}\leq 11$	0.9%	2.5%	6.3%	9.4%
$12\leq N_{atom}<30$	85%	91%	72%	67%
	average number of atoms per molecule			
H- [C≡]	0.05	0.03	-	0.1
H- [C=]	3.5	5.4	3.9	4.6
H-[C-]	9.2	4.6	5.4	2.3
C(sp)	0.1	0.9	0.1	0.5
C(sp^2)	5.1	7.7	6.4	8.4
C(sp^3)	5.3	2.3	4.5	2.1
=O	1.8	-	-	-
-O-	1.4	-	-	-
>N-	-	0.8	-	-
=N-	-	1.9	-	-
≡N	-	0.8	-	-
F	-	-	-	4.3
Cl	-	-	4.0	-

3.2 Geometry: number of contacts and average contact distances. Table 2 shows that the number of observed contacts below P(SSAR) decreases very sharply with decreasing P. H···O(ether, –O–) contacts decrease much faster than H···O(carbonyl, O=) contacts, confirming the former guesses on relative accessibility. As expected, the average distances between oxygen and the more acidic acetylene hydrogen are shorter. Contacts involving H(sp^2) are marginally shorter than contacts to H(sp^3) and contacts to carbonyl oxygen are marginally shorter than contacts to ether oxygen. These differences are at the borderline of significance and always comparable with the amount of spread of the distributions.

Table 3 shows the percentage of structures that contain no contacts, as a function of P. For F and Cl contacts virtually disappear already at 90% SSAR. Table 4 gives the average number of contacts per acceptor <N>, *i.e.* the total number of contacts divided by the total number of acceptor atoms of each species present in the database. This shows that carbonyl oxygen is a much better acceptor than ether oxygen, and the that order of acceptor strength for nitrogen is sp nitrogen > sp^2 nitrogen >> sp^3 nitrogen, the last species being no acceptor

at all. At 100% SSAR, carbonyl oxygen and *sp* nitrogen are at the center of a coordination pattern including on average more than one contact (average > 1); these species are also the ones for which the contact frequency better survives the decrease in threshold distance.

Table 2. Number of contacts (N_c) according to type of hydrogen, type of acceptor, and % SSAR (sum of standard atomic radii) cutoff. The quoted R_{av} are the average H···X distances over the sample truncated at the quoted P(SSAR). H···N(sp^3) contacts are too few for statistics. From CHX-*all* datasets.

	P	$P = 0.9$		$P = 1.0$	
Donor	Acceptor	N_c	R_{av}	N_c	R_{av}
[≡C]-H	-O-	9	2.22	12	2.29
[≡C]-H	O=	22	2.21	24	2.24
[=C]-H	-O-	49	2.35	375	2.54
[=C]-H	O=	196	2.34	887	2.50
[-C]-H	-O-	88	2.35	1021	2.56
[-C-]-H	O=	227	2.35	1680	2.54
[=C]-H	-N=	120	2.38	578	2.57
[=C]-H	N≡	45	2.39	308	2.58
[C-]-H	-N=	13	2.42	235	2.63
[C-]-H	N≡	27	2.41	245	2.60
CH··Cl		$P = 0.95$		$P = 1.0$	
[=C] H	Cl	5	2.66	87	2.80
[C-] H-	Cl	18	2.69	118	2.78
CH···F					
[=C] H-	F	30	2.38	139	2.48
[-C] H-	F	9	2.38	53	2.49

Table 3. Contact frequencies: percent of crystal structures with zero CH···X contacts in each dataset, as a function of threshold distance *R*.

|	CHO		CHN		CHCl		CHF	
	%	R	%	R	%	R	%	R
100% SSAR	2.5	2.680	13.2	2.740	49.8	2.860	24.8	2.560
95% SSAR	16.2	2.546	32.6	2.603	91.4	2.717	82.1	2.432
90% SSAR	60.1	2.412	71.8	2.466	100.0	2.574	100.0	2.304

Table 4. Acceptor capability: for each type of acceptor, ratio of the total number of CH···X contacts to the total number of that acceptor in the database.

	>O···H	=O···H	>N-···H	=N-···H	≡N···H	F···H	Cl···H
100% SSAR	0.76	1.74	0.13	0.91	1.50	0.38	0.23
95% SSAR	0.31	0.92	0.03	0.47	0.72	0.08	0.03
90% SSAR	0.08	0.30	-	0.15	0.20	-	-

The combined results in Tables 2-4 show that CH···O contacts are very frequent, and CH···N contacts are less frequent but still numerous; 40% and 28% of the structures, respectively, contain extremely short contacts below 0.9 SSAR. These findings are in agreement with elementary reasoning based on relative electronegativities. On the other hand, CH···Cl and CH···F contacts are sparse or just sporadic, and disappear already at 0.95 SSAR. This finding in itself does not encourage further speculation on a special nature of

these contacts, as they are likely even less influential than other CH···X, X=O, N, Cl ones in determining crystal cohesion and definitely not worthwhile of further investigation.

   As is implicit in the definition of standard (random) contact radii, at the full value of their sum (100% SSAR) there are very frequently 2-5 contacts in each crystal structure (Figure 1a), not necessarily denoting a particular chemical relationship. The distribution curves obviously shift sharply to the left (fewer contacts per structure) on decreasing the contact distance threshold. At 1.00 (0.95) SSAR, 47% (26%) of the acceptor oxygens have more than one donor atom (Figure 1b), so that the CH···O contact has a good amount of multiple-coordination character. The corresponding estimates for CH···N (Figure S1 SI) are quite similar and amount to 37 and 13%. Deep discussion of such results would require a difficult sorting of true bonding effects and steric accessibility effects.
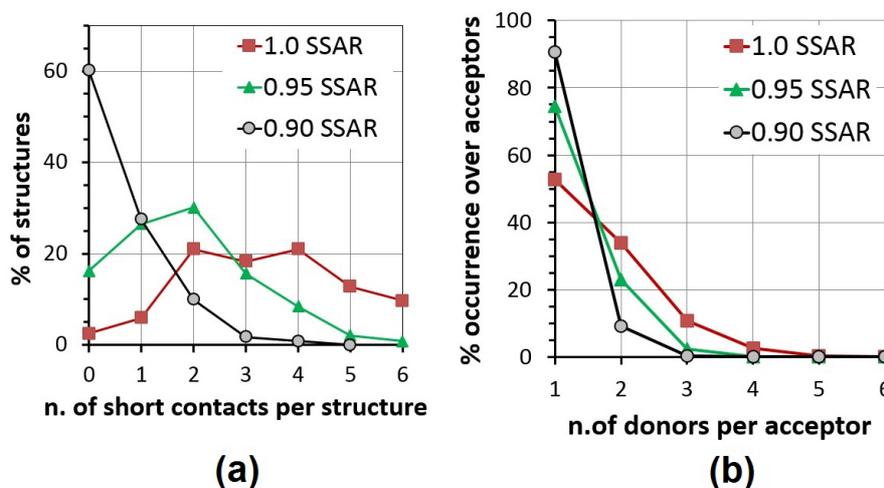
Figure 1. (a) Distribution of short CH···O contacts and (b) occurrences of single and multiple donors per acceptor, or bifurcation of the interaction in the CHO-*all* dataset. Red, green, black: 1.0, 0.95, 0.90 SSAR.

3.3 Energies: CH-X contacts and molecule-molecule cohesion. If CH···X contacts were a general, permanent dominant factor in the definition of a crystal structure, and assuming that shorter contacts are more stabilizing, the shortest contacts should appear in the top-ranking (*i.e.*, most cohesive) molecular pairs. Figure 2 shows the percent of short CH···X contacts that belong to molecular pairs of a given rank. Only about 45-50% of short contacts belong in the top two molecular pairs, the rest belong in lower-ranking pairs. Above rank 3 there is hardly any difference in the distribution with different distance limits. Thus, short CH···X contacts do not seem to produce higher total molecule-molecule energies, which are more likely determined by the dispersive contributions, much more massive than the minor Coulombic advantage of H···X confrontation. It is not surprising that individual short contacts do not correlate with interaction energies in dispersion-dominated structures, as dispersive effects are mainly related with the polarizability of the whole molecular charge density. This implies that quantities depending on the whole structure (such as, among others, density and packing coefficient, see Section 3.6 below) are probably better estimators of the overall crystal stabilization in this kind of materials. Besides, for extremely short C-H···O distances (<0.9 SSAR) the percent of contacts in the top-ranking molecular pair decreases rather sharply. This seems to suggest that very short contacts do not always imply more stabilizing pairing energies. In a tentative

explanation, some (many?) short contacts may be secondary effects introduced by the energetic requirements of other molecular groups or factors ('compressed' or 'tolerant' short contacts[24]). This peculiar result does not appear in the CH···N distribution, where the very short contacts mostly belong in the top-ranking pairs.
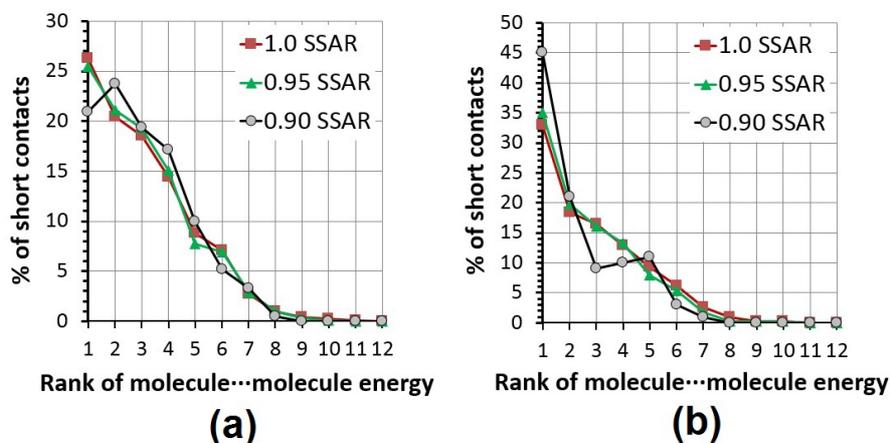
**Figure 2**. (a) CHO-*pix* dataset: rank in cohesive energy of the molecular pair in which short contacts appear at 100% (red) 95% (green) or 90% (black) SSAR. (b) Same for CHN-*pix*. For example, at 1.00 SSAR 25% of the CH···O contacts and 33% of the CH···N contacts are in the top-ranking molecular pair in the crystal, at 0.9 SSAR the numbers are 20 and 45%. CH···Cl contacts are quite insensitive to rank (see Figure S2 SI).

Figure 3 shows plots of the molecular pair interaction energy corresponding to each detected short contact. For CHO, the distribution shows, if anything, a slight tendency to smaller molecular interaction energies for pairs with shorter CH···O distances; this is in agreement with the dip in the black curve in Figure 2a. For CHN, one can hardly discern any real trend in the distribution as a function of distance. In these plots many very short contacts appear in scarcely cohesive pairs; for example, all of the shortest C-H···O contacts are between terminal acetylenic groups and carbonyl oxygen, so that the resulting pair is stretched out horizontally. The same C≡C–H···O arrangement is found in both the NEQBUJ and the NEQSEJ structures (highlighted by a circle in Figure 3) and corresponds to that shown in Figure 4a for NEQBUI. The energy components for this pair are $E_c+E_p =$ -25.1, $E_{disp} =$ -6.7, $E_{rep} =$ +17.3, $E_{tot} =$ -14.5 kJ mol^{-1} or a significant Coulombic term flanked by repulsion, and a smaller dispersion term. Figure 4b, on the other hand, displays an example of a π-stacked dispersion-dominated pair in NEQBUJ, without specific short contacts, but with a pair energy even more negative than the hydrogen-bonded system in Figure 4a (energy components: $E_c+E_p =$ -20.2, $E_{disp} =$ -37.9, $E_{rep} =$ +30.3 and $E_{tot} =$ –28.1 kJ mol^{-1}). Many other pairs are strongly cohesive without short H···O contacts; molecular-pair interaction energies in crystals do not necessarily correlate with the presence or absence of short contacts, because dispersion is higher in vertically stacked molecular pairs and therefore competes with C-H···X contacts which must be perforce coplanar. Short contacts are influential but again can hardly be considered as the unique and-or predominant factor in crystal packing.
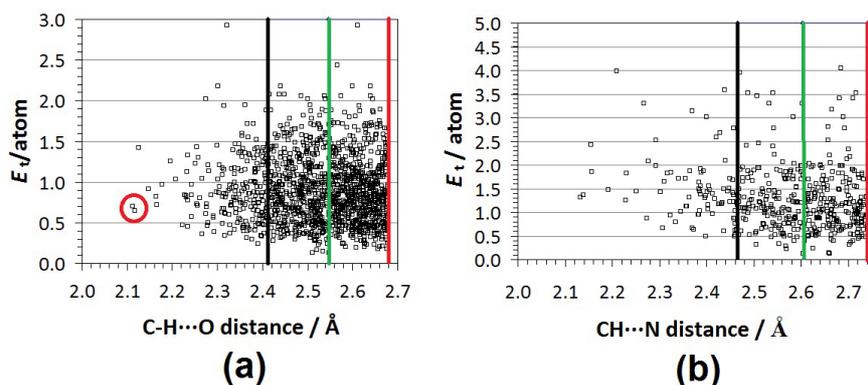
Figure 3. A plot of the molecular pair energy (kJ/mol; normalized to the number of atoms in the molecule) as a function of the CH···X distance of each 'short' contact. (a) oxygen contacts, (b) nitrogen contacts. The red circle marks some of the shortest distances (crystal structures with CSD refcodes NEQBUJ and NEQSEJ, see text). CH···Cl contacts are too few for reliable statistics. Cutoffs for 'short', 'very short' and 'extremely short' hydrogen bonded contacts (Table 3) are highlighted as red, green and black vertical bars, respectively.
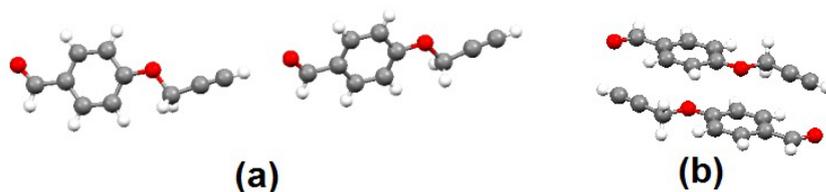
Figure 4. Pairing motifs in NEQBUI,[25] (a): the Coulomb-dominated pair, one of the shortest H···O contacts in Figure 3a, total pairing energy -14.5 kJ·mol⁻¹; (b): the dispersion-dominated, stacked pair, -28.1 kJ·mol⁻¹.

Another test of the influence of short contacts in crystal structures is provided by an attempted correlation between the short-bond index of equation (1) and lattice energies. In bivariate analysis, this index shows a systematic weak correlation with the Coulombic energy density, as shown in Figure 5, but generally neither with the total lattice energy density nor the corresponding dispersive contributions (see Section 3.6 below). This implies that B_s is expected to be a poor predictor for dispersive or total energies, especially in dispersion-dominated crystals. As discussed above, indeed, the short contacts have a net Coulombic character but their action is in competition with the dispersive contribution. The point is discussed further with results shown in Figure 6.
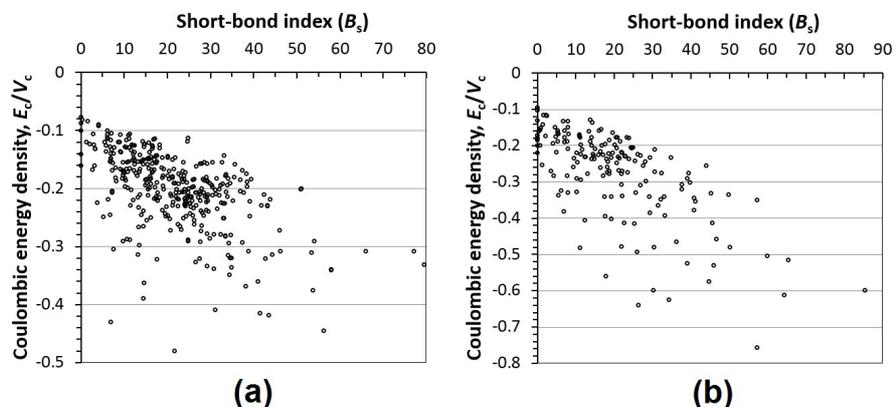
Figure 5. The weak correlation between short-bond index B_s, a measure of the number and shortness of H···X contacts, and the Coulombic energy density. (a) CHO-*pix*, (b) CHN-*pix* datasets.

3.4 Energies: physical types of interaction. Coulombic and polarization energy contributions are large when molecular regions of opposite permanent polarization come into close contact: in our case, positively charged hydrogen zones and negatively charged oxygen or nitrogen molecular regions. Dispersive contributions arise from diffuse, transient polarization in the whole molecular electron densities, and are larger for highly polarizable electrons, in our case presumably aromatic-ring electrons and valence electrons in chlorine atoms. In H···O or H···N contact zones dispersive contributions are very small due to the scarce polarizability of the contact termini. To characterize the nature of the interaction one can define a Coulombic contribution to the cohesive stabilizing energy, and a dispersive overall stabilizing ratio, $Q(C)$ and $Q(d,\text{stab})$, as follows:

$$E_t = E_c + E_p + E_d + E_r = E_s + E_r \tag{2}$$

$$Q(C) = 100 \cdot \left[\frac{E_c + E_p}{E_s} \right] \tag{3}$$

$$Q(d,\text{stab}) = 100 \cdot \frac{E_d}{E_t} \tag{4}$$

Meanings of energy terms in equations (2)–(4) are the same defined in Section 2.3 above, while E_s collects all the contributions but short range repulsions (E_r) and might be thought as a measure of the strength of the stabilizing part of the crystal field. Figure 6 shows that in CHO crystal structures, and even more in CHN or CHCl structures, the dispersive contribution provides the major part of the total stabilizing lattice energy. Figure 7 shows details at the level of single molecule-molecule pairs. When CH···X contacts are shorter, the Coulombic contribution $Q(C)$ in each molecular pair increases and the dispersive contribution $Q(d,\text{stab})$ decreases, but irrespective of the length of the contacts present, the dispersive ratio is nearly always greater than 50%. The same trends appear even more clearly in the CH···N contacts. Short contacts generate a considerable amount of Coulombic stabilization, but also a repulsion due to molecular overlap, which needs a substantial dispersive contribution to achieve overall stabilization. In some cases ($Q(d,\text{stab}) > 100\%$) the dispersive contribution exceeds the total stabilizing energy and thus takes the form of a net overcompensation for repulsion. These findings issue some warning against a consideration of atom-atom contacts alone as structure determinants.
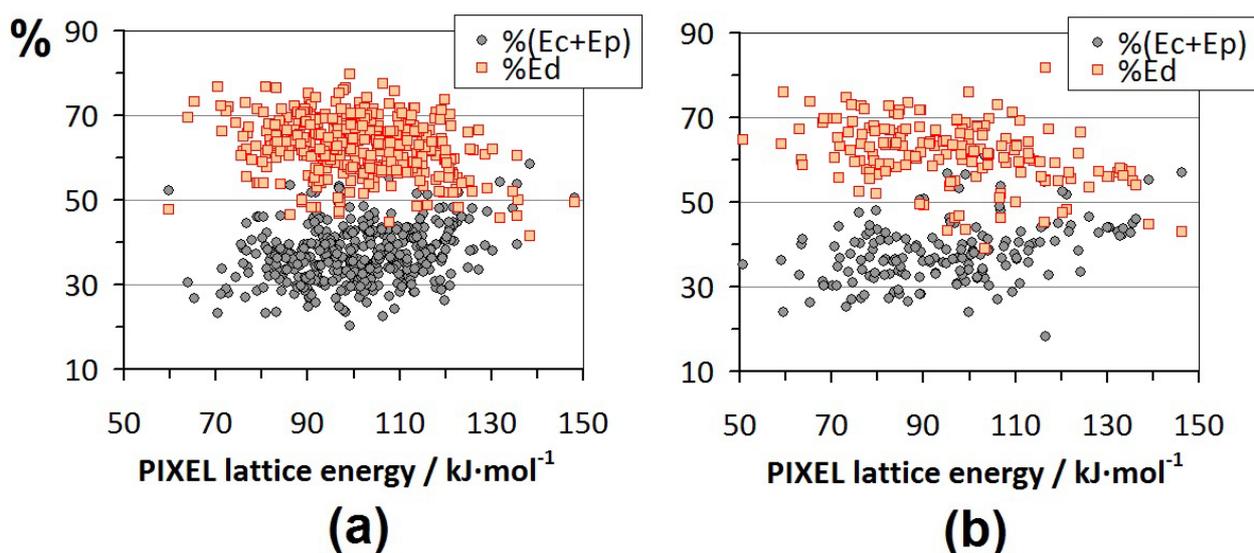
Figure 6. The percent of Coulombic+polarization contribution, Q(C) (black) to the total cohesive lattice energies of crystals in the CHO-*pix* and CHN-*pix* datasets. The percent dispersive is in this case 100-Q(C) (red), reckoned on the stabilizing energy and not on the total energy and not to be confused with Q(d, stab) appearing in Figure 7. The corresponding graph for the CHCl dataset shows an even sharper separation with dispersive contributions Q(d) always > 70%. (see Fig. S3 SI).
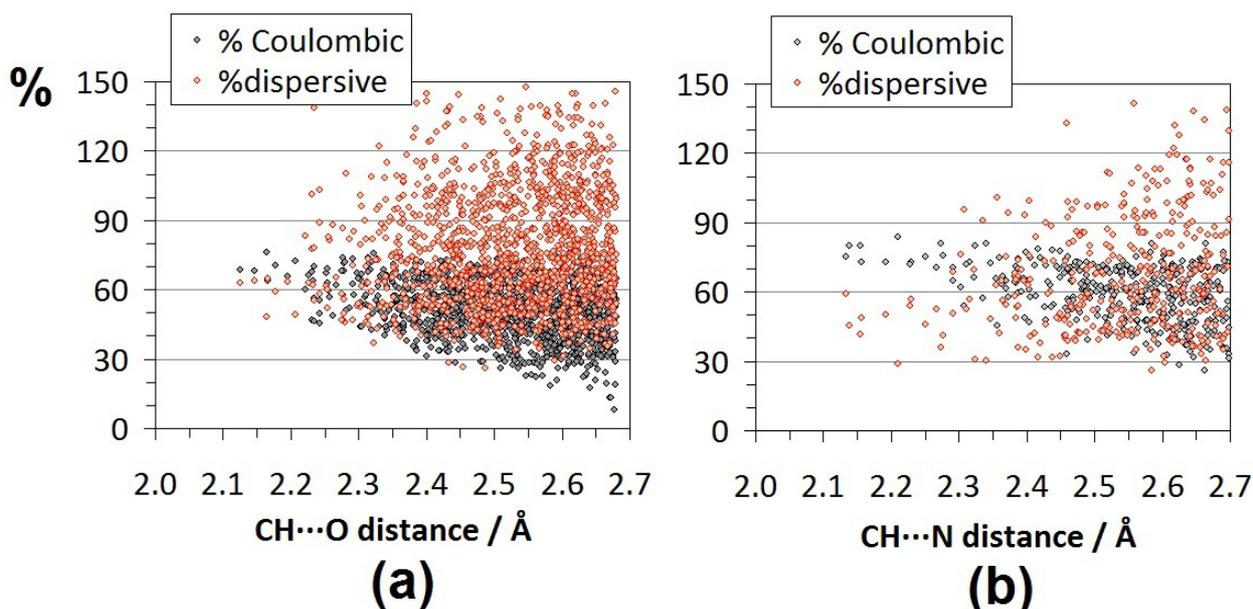
Figure 7. (a) Percent coulombic+polarization contribution to stabilization, Q(C) (red), and dispersive contribution to total energy, Q(d,stab) (black) in the molecule-molecule cohesive energy of CH···O contacts. The horizontal axis is the distance of the O···H contacts present. (b), as above, for the CHN dataset. For the CHCl dataset the data are too few for significant statistics. Only molecular pairs with a substantial total pairing energy (> 10 kJ mol⁻¹) are included.

3.5 Rank of CH···X energies versus lattice energies. The existing crystal structures of organic compounds show a hierarchy of intermolecular interactions that can be readily evaluated by quantum chemical or semiempirical modeling. The relative energetic intensity of the intermolecular approaches found in the molecular coordination shell is the result, and possibly a trace, of the events occurring at crystal formation time.

To provide a quantitative assessment of the importance of the coplanar approaches resulting from C-H···X contacts, the pairing energies of some model compounds (Scheme 1) have been calculated and the results are given in Table 5.
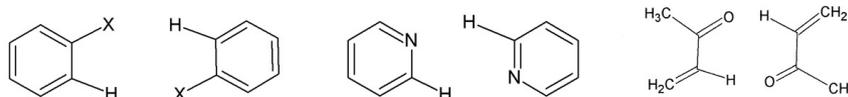

Scheme 1

Note that these numbers are upper-limit ones, being obtained for very favorable arrangements and for very acidic hydrogen atoms. The energetic relevance of a single contact is estimated as the difference in cohesive energy from that of an identical nonbonding approach involving only C-H···H contacts in the equivalent benzene dimer. Figure 8 shows the distribution of lattice energies in the CHO-*pix* crystal sample, ranging from 70 to 130 kJ mol^{-1}, essentially in relationship with the size of the molecule.[26] A single C-H···O contact can provide up to 15% of the total energy in very small molecules, but no more than 3% in crystals of the largest molecules. The results are quite similar for the CHN-*pix* database. One sees here a clear indication that short atom-atom contacts between hydrogen and oxygen or nitrogen atoms can definitely be relevant to crystal packing only in special cases when the contact distance is 90-95% of the sum of atomic radii, and for small molecules where the influence of bulk dispersion is less dominant. At the same time, Table 5 shows that there is little ground for postulating a similar role for H···Cl or even less for H···F contacts.

Table 5. PIXEL results for the binding energies (kJ·mol^{-1}) of coplanar dimers at the minimum of their interaction energy curves.

X	$d_{H··X}$ / Å	E_c	E_p	E_d	E_r	E_t	E/H···X[a]
H	2.32	0.3	-1.1	-8.0	5.5	-3.3	0.0
F	2.50	-4.3	-1.1	-6.4	5.0	-6.7	-1.7
Cl	2.79	-3.2	-3.2	-16.3	13.8	-8.9	-2.7
O=	2.40	-13.2	-4.9	-8.9	12.9	-14.0	-5.3
-N=	2.56	-13.0	-5.4	-11.5	15.7	-14.1	-5.4

[a] Energy per single contact relative to the energy of the benzene dimer.
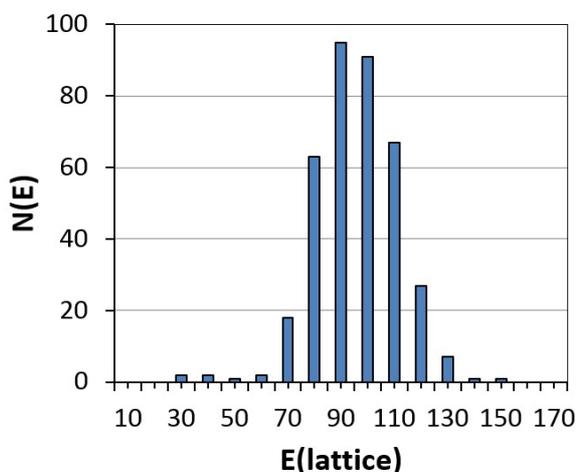
Figure 8. Distribution of the total PIXEL lattice energies (kJ·mol^{-1}) in the CHO-*pix* database (377 structures).

3.6 Principal component analysis. In general, one may wonder what crystal or molecular properties are significant in the description of crystal packing and cohesion. In the literature several global variables have been employed, related to molecular size and constitution, along with cohesive energies.[24,27-30] Attempts were also made to relate packing properties with crystal symmetry[31] and, obviously, with the occurrence of a network of H···X interactions of variable strength.[32-35] We focus here of the following continuous variables:

- for molecular size: N_{at}, global number of atoms in the molecule; M_w, molecular weight; V_c/Z, the cell volume per molecule or unitary volume;

- for molecular polarity: Q, sum of positive values of atomic charges from a Mulliken population analysis (equal to the absolute value of the sum of negative charges), a measure of the charge separation within the molecule;

- for packing quality: crystal density (d_x) and packing coefficient, C_k, defined as the total volume of all molecules in the unit cell, computed as usual through a summation over overlapping atomic van der Waals spheres, divided by the total cell volume;

- for energies: the usual PIXEL lattice terms, i.e. Coulombic (E_c), polarization (E_p), dispersion (E_d), repulsion (E_r) and total energy (E_t).

In all cases the first three PC's, labeled 1-3 in decreasing order of importance and accounting for 86(2)% of the original variance, were found to satisfy the Kaiser-Guttman significance test[36] while just the first two are also within the Karlis-Saporta-Spinakis eigenvalue tolerance criterion [37] (Table S5 SI, Fig. S4 SI). The appearance of three leading PC's was already noticed in an early study on hydrocarbon crystals.[38]

Figure 9 shows the representation of the three datasets in the PC1-PC2 reference frame. The vectors are the projections (loadings) of the original variables onto the PC1-PC2 plane. Each point is the score for one crystal structure. Orthogonal projections of each score onto the loading vectors give the predicted values of the corresponding variables for that crystal structure. PC1 represents the covariance of molecular size variables and stabilizing energies, recalling that the latter are negative numbers so their vectors point in a direction opposite to that of the positive size indicators.

For the CHO and CHN datasets total cohesive energies are dominated by dispersion terms, while Coulombic terms play a secondary role. However, it is worth noting that E_c is more significant, in terms of both loading magnitude and covariance with total energy E_t, in the CHO dataset (Fig. 9b) than in the CHN and CHCl ones (Fig. 9a,c). In PC2, density and packing coefficient are dominant; moreover, packing coefficient C_k is almost orthogonal to the direction of dispersion and total stabilizing energies, implying that a given crystal structure does not necessarily achieve at the same time a high packing efficiency and a large stabilizing cohesive energy. C_k also correlates with the Coulombic term E_c, but not with molecular size. The repulsive energy term E_r is at 45°, being influenced both by size (larger molecules overlap more) and by packing efficiency (more closely packed molecules overlap more). It is worth noting that Coulomb contributions E_c are also weakly correlated with polarization (E_p), but this is appreciable just in the PC1, PC3 plane (see *infra*). Such correlation is not evident in the PC1, PC2 plot (Fig. 9), as the second principal component mainly accounts for close packing issues.
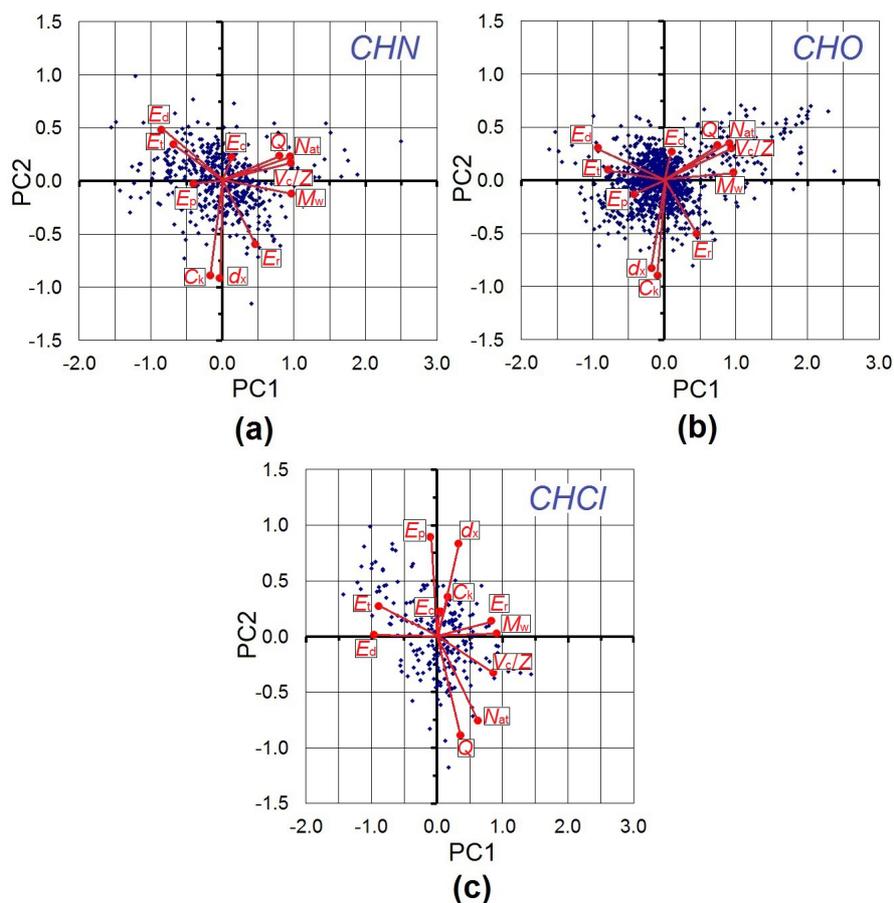
Figure 9. Biplot of the variable loadings (red lines) and scores (blue squares) against the first (horizontal) and second (vertical) principal component. A 0.1 factor was applied to all the scores to bring them to a scale comparable with loadings. E_c, E_p, E_d, E_r and E_t are the usual PIXEL energy terms, d_x is the crystal density, N_{at} the atom count, M_w the molecular weight, V_c/Z the cell volume per molecule, Q the amount of charge separation and C_k is the packing coefficient. (a) CHN dataset. (b) CHO dataset. (c) CHCl dataset.

The score distributions in CHO and CHN datasets (Fig. 9a, 9b) are nearly globular, meaning that the crystals can equally have all four combinations of larger/smaller than average size and larger/smaller than average packing tightness. Recalling that PC1 accounts for 50% of the total variance and PC2 for 25% (Table S5 SI and Fig. S4), the cluster of equioriented mass, number of atoms and cell volume loadings defines the direction of maximum spread; as expected, crystals of bigger molecules tend to have the most stabilizing total and dispersion energy contributions. Dispersion increases anyway with the number of electrons in the molecule.

The CHCl dataset (Figure 9c) shows a different picture, firstly because molecular mass M_w no longer correlates with the number of atoms, reflecting the different amounts of chlorination. Dispersion (E_d) has the lion's share of stabilizing terms, but also polarization (E_p) is now more important in defining the PC2 variable, along with the charge separation Q, in a joint effect depending mostly on the presence of the more polarizable chlorine 3s and 3p electrons. The picture is also somewhat fuzzier due to the smaller number of data points.

Table 6. Selection of representative structures with the PC2 coordinate significantly different from zero, in the direction of the C_k loading vector (see Fig. 9a,b). For each entry, the mean of the corresponding distribution is given in square brackets to highlight whether it is above or below the average expectation value for analogue structures. See Fig. 10 for the corresponding packing diagrams.

CSD code Database	KENDEM [a] CHO	DAYBIO [b] CHN
Variable		
PC1	+0.393	+0.409
PC2	−0.802	−1.156
N_{at}	24 [28]	20 [25]
V_c/Z / Å^3	250.6 [262]	228.1 [248]
M_w / amu	268.2 [209]	208.2 [194]
Q / e	2.76 [2.5]	1.89 [1.9]
d_x / g·cm^{-3}	1.777 [1.3]	1.515 [1.3]
C_k	0.83 [0.72]	0.83 [0.73]
E_c / kJ·mol^{-1}	−35.1 [−30]	−95.9 [−23]
E_p / kJ·mol^{-1}	−44.9 [−30]	−31.0 [−21]
E_d / kJ·mol^{-1}	−147.9 [−107]	−145.6 [−114]
E_r / kJ·mol^{-1}	82.3 [51]	135.5 [40]
E_t / kJ·mol^{-1}	−145.6 [−116]	−137.0 [118]

[a] $C_{14}H_4O_6$, $P2_1/c$, $Z = 2$
[b] $C_{10}H_4N_6$, $P4_2/mnm$, $Z = 2$

Two structures exhibiting high scores on PC2 along the direction of the packing coefficient loading were further examined in detail, to check whether efficient packing prevails over cohesive energy requirements, and whether the high score in that direction could be related to the presence of short CH···X contacts. Table 6 reports the detail of the loadings and of the deviations from averages, while the corresponding packing diagrams are shown in Figure 10. Both structures are for planar rigid molecules: KENDEM, naphthalene-tetracarboxylic anhydride, and DAYBIO, (1,4-pyrazinio)-bis-dicyanomethylide, formally a symmetric double zwitterion with null dipole moment. KENDEM forms a naphthalene-like zig-zag pattern of tightly stacked molecules, presumably maximizing π···π interactions. DAYBIO has an exceptionally large Coulombic component, due to its zwitterionic structure. Flat sheets of molecules are staggered so as to prevent proximity of groups of the same charge and to minimize the separation between groups of opposite charge. Both crystal structures show very short CH···O or CH···N contacts; nevertheless, however important these contacts may be in defining minor structural adjustments, the overall packing is clearly dominated by the large dispersion and/or Coulombic contributions, and a proper discussion of crystal packing should highlight first these major, global stabilizing factors.

Figure 11 shows the picture that results from plotting the CHO and CHN datasets in the PC1-PC3 plane, recalling that PC3 accounts only for a modest 10% of the total variance and hence the conclusions drawn may be somewhat tentative (Table S5 SI, Fig. S4 SI). PC3 is related to the Coulombic and polarization factors (E_c and E_p), whose vectors are almost parallel to the PC3 axis, and hence can be interpreted as a general measure of the role of electrostatics, being also almost orthogonal to the dispersion factor (E_d) and to the close packing coefficient (C_k)..
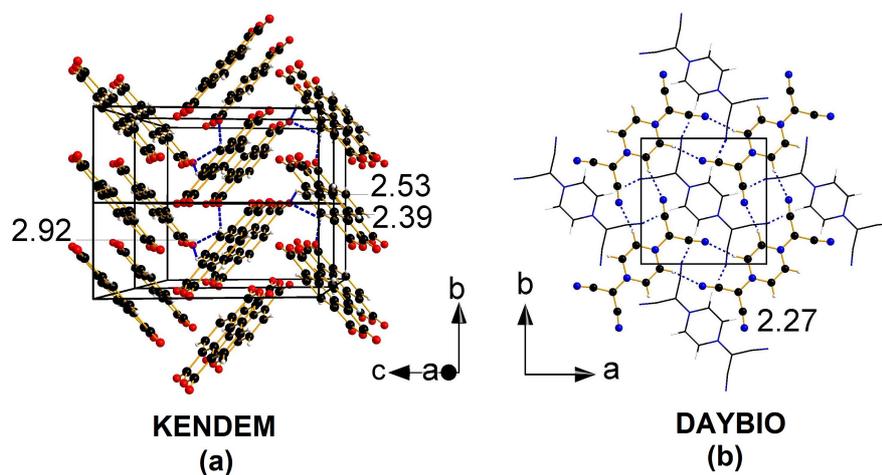
Figure 10. Packing diagrams of the structures shown in Table 6. Short H···X, X=O, N contacts are highlighted as dashed blue line, with symmetry-independent H···X distances specified in Å. For the DAYBIO structure, the ball-and-stick (wireframe) drawing refers to molecules in the top (underlying) layer. Figure realized by Diamond v3.2k © 1997-2014 Crystal Impact GbR, Bonn, Germany.

The southwest-northeast alignment of the bulk of the data is almost exactly along the Q/E_{tot} directions, or a larger total energy for a bigger charge separation; this can be interpreted as indicating that charge separation takes over the role of molecular size when electrostatic issues are at stake. The same plot for the CHCl dataset shows only a scarcely interpretable spread, an obvious consequence of the lack of frankly electronegative groups in chlorohydrocarbons.

In a separate PCA run, intensive energy variables were obtained by dividing energy terms by the unitary cell volume, thus taking the character of energy densities within the crystal structure, and the short-bond index was also included in the analysis. The interpretation of the resulting PC plots (Fig. S5–S7 and Table S6 SI) is less straightforward than for the plots in Figures 9 and 11. The safest general conclusions seem to be that (i) energy densities are significantly correlated and tend to cluster together, and (ii) no straight correlations can be found between the short-bond index and any of the size or energy variables. A more in-depth analysis of results for the CHO dataset shows however that the dispersion term (E_d) is nearly orthogonal to the electrostatic term (E_c), while the short-bond index B_s, the charge separation index Q/N_{at}, and the electrostatic terms E_c and E_p are almost covariant. This underscores the importance of electrostatics in the CHO dataset, due to interactions involving high-order moments of the charge density distribution in these more polar molecules including electronegative oxygen. On the one hand, a larger Coulombic attraction between whole molecules is more likely to produce also short atom-atom contacts; on the other hand, local contacts between H(δ^+) and O(δ^-) termini produces higher Coulombic stabilization. All considered, a cautious conclusion is that short contacts are local features that do not necessarily couple with bulk structure determinants. Once again, (see also Section 3.3) the results suggest that these contacts should be analyzed carefully and their importance justified in each case, only after the main energetic features of the crystal packing have been adequately covered.
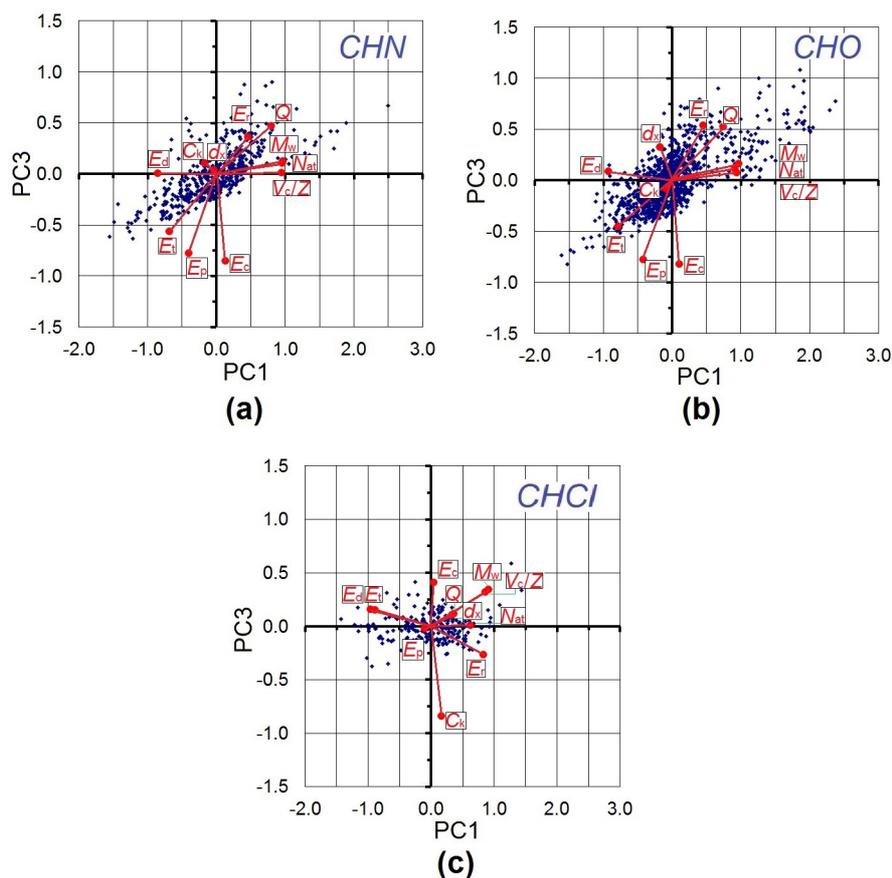
Figure 11. As Figure 9, for the PC1, PC3 plane.

Conclusions

Statistical results of a study for the first time carried out on very large databases of hundreds of crystal structures should provide solid-state chemists with a sound and extensive amount of hard and objective geometrical and energetic data on the title matter. The small size of the molecules in the database should be a bonus rather than a limitation, helping the focus of the study, alongside with the absence of O···H or N···H hydrogen bonds. The following is a list of main points made alongside.

1) H···X atom-atom contacts shorter than the sum of standard atomic radii are very frequent in CHO crystals, and less frequent but still numerous in CHN crystals. Acetylenic hydrogens form the shortest contacts, followed by aromatic and aliphatic hydrogens. When the analysis is carried out at the threshold of sum of atomic radii, most acceptors in these datasets show a multiform and non-specific coordination with many donors, suggesting that only really shorter contacts are worth considering. On the other hand, short CH···Cl contacts are sparse, and CH···F contacts are just sporadic: both disappear at the threshold of 95% of the atomic radii. This purely geometrical result does not encourage a further analysis of hydrogen-halogen contacts. It is worth noting, however, that in recent years the so-called 'halogen bonds' were introduced as possible structure-determinant interactions. The latter are different in nature with respect to classical hydrogen bonds, as in halogen bonds the donor places a region of charge concentration (e.g. a lone pair), and not a hydrogen atom, toward the σ-hole of an acceptor halogen atom.[35] Therefore, it remains to be seen

whether short X···Y contacts, X and Y being halogen atoms, could be more significant than CH···X interactions in determining pair and packing energies.

2) In all datasets dispersion plays the top role in crystal energy stabilization, its contribution being always > 50% both on lattice energies and on molecule-molecule contact energies. Short contacts also generate a significant repulsion due to molecular overlap, and a significant dispersive contribution is always indispensable to achieve a net stabilization. The balance between Coulomb forces and dispersion takes a clearer structural aspect in flat elongated molecules, because dispersion is higher in vertically stacked molecular pairs and therefore competes with coplanar C–H···X contacts.

3) Sample calculations on the dimerization energies of model compounds provide an estimate of the energy contribution of H···X contacts as 2 kJ mol^{-1} for fluorine, 3 kJ mol^{-1} for chlorine, and about 5 kJ mol^{-1} for oxygen or nitrogen, but only when the H···X distance is much shorter than the sum of atomic radii. These numbers compare with estimates of total lattice energies ranging from 60 to 140 kJ mol^{-1}. Therefore, in our databases of constant composition and free from H···O hydrogen bonding, the lattice energy mainly depends on molecular size. As an order of magnitude, a single CH···O or CH···N contact can thus provide up to 8% of the lattice energy in very small molecules, and no more than 4% of the lattice energy in larger molecules. The influence of these special contacts should then decrease with the increase of molecular complexity, as expected. Assigning an energetic relevance to contacts to halogens is indeed a problematic exercise.

4) A multivariate study using principal component analysis shows a covariance of molecular size indicators with dispersive and total lattice energies, but not necessarily with Coulombic terms, as expected because all electrons contribute to dispersion while only local polarizations contribute to Coulombic terms. In fact in CHO and CHN crystal structures, and even more in CHCl structures, the dispersive contribution provides the major part of the lattice energy. In-depth analysis of the stabilizing energy in molecule-molecule single interactions shows that the most stabilizing molecular-pair energies in the first molecular coordination shell neither require nor correlate with the presence or absence of special atom-atom contacts. Rather, the interplay of collective dispersive and electrostatic interactions determines the force field on which cohesion relies. Even when high packing efficiency is achieved, this is more likely related to steric requirements or shape effects rather than to the set-up of structurally determinant CH···X patterns. On the other hand, the Coulombic contribution is larger when very or extremely short CH···O or CH···N contacts are present. This feature of the stabilization arising from short H···O and H···N contacts is further demonstrated by the correlation between the short-bond index of eq. (1) and the Coulombic energy density (at least in the CHO dataset), and by the lack of correlation of the index with the total lattice energy in general (with the exception of strongly polar or zwitterionic structures, where the Coulombic energy is very large and obviously correlates with total energy).

In summary, CH···N and CH···O contacts can be influential and to some extent can be real structure determinants, but only when very short and especially in small molecules. Discussions of crystal stability and even more exercises in crystal engineering based on these contacts require a careful analysis in each case with comparative consideration of the overall requirements of other crystal forces. A large proportion of crystal structures seem to be well able to do without, so that these atom-atom interactions cannot be taken *a*

priori as reliable and solid crystal building blocks, nor can they be entrusted with a general status of reproducible chemical bonds. These caveats apply even more strongly to CH⋯Cl or CH⋯F contacts, where relevance to crystal packing seems to be more an exception than a, however weak, rule.

Acknowledgements

References

1 Thanthiriwatte, K.S.; Hohenstein, E. G.; Burns, L.A.; Sherrill, C.D. *J. Chem. Theor. Comp.* **2011**, *7*, 88–96.

2 Jurečka, P.; Šponer, J.; Černý, J.; Hobza, P. *Phys. Chem. Chem. Phys*. **2006**, *8*, 1985–1993.

3 Ehrlich, S.; Moellmann, J.; Grimme, S. *Acc. Chem. Res.* **2013**, *46*, 916–926.

4 Gavezzotti, A. *J. Phys.Chem*. **2003**, *B107*, 2344–2353.

5 Moggach, S. A.; Marshall, W. G.; Rogers, D. M.; Parsons, S. *CrystEngComm* **2015**, *17*, 5315–5328.

6 de Klerk, N. J. J.; van den Ende, J. A.; Bylsma, R.; Grančič, P.; de Wijs, G. A.; Cuppen, H. M.; Meekes, H. *Cryst. Growth Des.* **2016**, *16*, 662–671.

7 Allen, F. H. *Acta Cryst*. **2002,** *B58***,** 380–388.

8 Herrebout, W. A.; Suhm, M. A. *Phys. Chem. Chem. Phys*. **2011**, *13*, 13858–13859.

9 Bernstein, J. *Cryst. Growth Des.* **2013**, *13*, 961–964.

10 Leigh, D. A.; Robertson, C. C.; Slawin A. M. Z.; Thomson, P. I. T. *J. Am. Chem. Soc.* **2013**, *135*, 9939–9943.

11 Balamurugan, V.; Hundal, M. S.; Mukherjee, R. *Chem Eur. J.* **2004**, *10*, 1683–1690.

12 Chopra D.; Guru Row, T. N. *CrystEngComm* **2011**, *13*, 2175–2186.

13 Dunitz, J. *IUCrJ* **2015**, *2*, 157–158.

14 Dance, I. *New J. Chem*., **2003**, *27*, 22–27.

15 Dunitz, J. D.; Gavezzotti, A. *Angew. Chem. Int. Ed.*, **2005**, *44*, 1766–1787.

16 Allen, F. H.; Bruno, I. J. *Acta Crystallogr.* **2010**, *B66*, 380–386.

17 Gavezzotti, A. *CLP User's manual*, **2016**. Manual, executables and source codes are freely available at http://www.angelogavezzotti.it.

18 Alvarez, S. *Dalton Trans*. **2013**, *42*, 8617–8636.

19 Jolliffe, I.T. *Principal Component Analysis (2nd Edition)*, Springer Series in Statistics: Springer (USA) 2010; ISBN: 9781441929990.

20 Kaiser, H.F. *Psychometrika*, **1958**, *23*, 187–200.

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

22 Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheesman, J. R.; Montgomery, J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V. et al. GAUSSIAN03, Gaussian Inc.: Wallingford CT, 2004.

23 Rakotomalala, R. *TANAGRA: a free software for research and academic purposes* in *Proceedings of EGC'2005*, RNTI-E-3, **2005**, *2*, 697–702.

24 Gavezzotti, A. *Acta Crystallogr*. **2010**, *B66*, 396–406.

25 Doi, I.; Okuno, T. *Acta Crystallogr*. **2013**, *E69*, o125

26 Dunitz, J. D.; Gavezzotti, A. *Acc. Chem. Res*. **1999**, *32*, 677-684.

27 Gavezzotti, A. *Molecular Aggregation, Structure Analysis and Molecular Simulation of Crystals and Liquids*, Oxford University Press: Oxford (UK) 2007; ISBN: 9780198570806.

28 Lo Presti, L.; Sist, M.; Loconte, L.; Pinto, A.; Tamborini, L.; Gatti, C. *Crystal Growth Des*. **2014,** *14*, 5822–5833.

29 Gavezzotti, A.; Lo Presti, L. *Crystal Growth Des*. **2015**, *15*, 3792–3803.

30 Lo Presti, L.; Orlando, A. M.; Loconte, L.; Destro, R.; Ortoleva, E.; Soave, R.; Gatti, C. *Crystal Growth Des*. **2014**, *14*, 4418–4429.

31 Gavezzotti, A.; Filippini, A. *Acta Crystallogr*. **1992**, *B48*, 230–234.

32 Bertolasi, V.; Gilli, P.; Ferretti, V.; Gilli, G. *Acta Crystallogr*. **2001**, *B57*, 591–598.

33 Espinosa, E.; Molins, E.; Lecomte, C. *Chem. Phys. Letters*, **1998**, *285*, 170–173.

34 Lo Presti, L.; Soave, R.; Destro, R. *J. Phys. Chem. B* **2006**, *110*, 6405–6414.

35 Destro, R.; Sartirana, E; Loconte, L.; Soave, R.; Colombo, P.; Destro, C; Lo Presti, L. *Crystal Growth Des*. **2013**, *13*, 4571–4582.

36 Jackson, D. A. *Ecology*, **1993**, *74*, 2204–2214.

37 Karlis, D.; Saporta, G.; Spinakis A. *Communications in Statistics - Theory and Methods*, **2003**, *32*, 643–666.

38 Gavezzotti, A. *J. Am. Chem. Soc.* **1989**, *111*, 1835–1843.
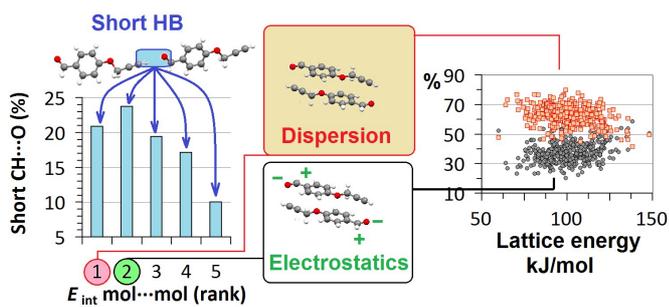
Synopsis: Top energy ranking molecular pairs in organic crystals containing only C–H hydrogen bond donors are mostly dominated by dispersion. No strong correlations are evident between the frequency of short CH···O or CH···N contacts and pair or lattice energies. These contacts provide minor stabilizing contributions, possibly relevant in driving crystal packing for small molecules, but not necessarily for medium-large ones.