Journal of Pharmaceutical Sciences 112 (2023) 435–445

Contents lists available at ScienceDirect

Journal of Pharmaceutical Sciences

journal homepage: www.jpharmsci.org

Drug Discovery–Development Interface

Exploring the CSD Drug Subset: An Analysis of Lattice Energies and Constituent Intermolecular Interactions for the Crystal Structures of Pharmaceuticals

Cai Y. Ma[a,*], Alexandru A. Moldovan[b], Andrew G.P. Maloney[b], Kevin J. Roberts[a]

[a] Centre for the Digital Design of Drug Products, School of Chemical and Process Engineering, University of Leeds, Leeds, LS2 9JT, UK
[b] The Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1EZ, UK

ARTICLE INFO

Article history:
Received 14 June 2022
Revised 25 November 2022
Accepted 25 November 2022
Available online 1 December 2022

Keywords:
Pharmaceutical materials
Drug subset
HABIT98
Lattice energetics
Crystal structures
Intermolecular interactions

ABSTRACT

Intermolecular (synthonic) modelling is used for a statistical analysis of crystal lattice energies, together with their contributing intermolecular interactions for the crystallographic structures selected from the CCDC's Drug Subset (https://doi.org/10.1016/j.xphs.2018.12.011). Analysis of this selected subset reveal similarities in packing compared to other organic crystals in the CSD with linear relationships between molecular weight and unit cell volume, void space, and packing coefficient. Crystal lattice energy calculations converge within a 30 Å intermolecular radius characterised by a mean lattice energy of ca. -36 kcal mol^{-1} with ca. 85% and 15% due to dispersive and electrostatic interactions, respectively. The distribution of the strongest synthons within the individual structures reveals an average strength of −5.79 kcal mol^{-1}. The diversity of chemical space within the drug molecules is in agreement with the analysis of atom types across the selected subset with phenyl groups being found to contribute the highest mean energy of −11.28 kcal mol^{-1}, highlighting the importance of aromatic interactions within pharmaceutical compounds. Despite an initial focus on Z' = 1 structures, this automated approach enables rapid and consistent quantitative analysis of lattice energy, synthon strength and functional group contributions, providing solid-form informatics for pharmaceutical R&D and a helpful basis for further investigations.

© 2022 The Authors. Published by Elsevier Inc. on behalf of American Pharmacists Association. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

Introduction

The Cambridge Structural Database[1] now contains over 1.2 million entries of organic and metal-organic small molecule crystal structures, providing an invaluable resource in crystallography, structural chemistry, and drug discovery. The use of computational and statistical approaches that use CSD entries can provide insight into pharmaceutical development and manufacturing issues and has led to increasing interest in the fields of solid-form informatics[2] and particle informatics[3] as *in silico* tools for drug development. A wide-range of different approaches has been developed for the routine use of data-driven modelling of molecular and crystallographic properties including polymorph stability,[4-6] solubility,[7-12] and crystal morphology.[13-16]

Alongside a suite of molecular scale informatics tools, the CSD can provide a thorough understanding of solid-form properties.[2] However, limited studies exist to develop a detailed understanding of the relationship between a material's inherent particle and surface properties at a molecular level with its processing and manufacturability. Molecular modelling of the intermolecular interactions, or "synthons", in the bulk and at the surface of a crystalline particle has led to the prediction of properties such as morphology,[17-20] surface energy,[21-23] crystallisability,[24-27] and particle cohesivity/adhesivity.[28,29] Much wider use of such methods[30] could potentially lead to a deeper understanding of the molecular mechanisms that underpin a range of problematic particulate properties, such as poor flowability, tendency for agglomeration, and "fines" production, which can cause significant problems downstream during the various processing steps involved in drug product formulation and subsequent manufacture.[31] Bryant *et al.* [3] highlighted the importance of particle informatics approaches when revisiting the structure of lamotrigine through a reassessment of its solid-form within the context of the crystal structures in the CSD. The work highlighted

Abbreviations: a, b, c, Crystal unit cell parameters (Å); α, β, γ, Crystal unit cell parameters (°); API, Active pharmaceutical ingredient; Coul, Coulombic energy contribution (kcal mol^{-1}); CSD, Cambridge Structural Database; Disp, Dispersive energy contribution (kcal mol^{-1}); Elec, Electrostatic energy contribution (kcal mol^{-1}); Latt, Lattice energy (kcal mol^{-1}); R^2, Goodness-of-fit; vdW, van der Waals; Z'/Z, Number of molecules in the asymmetric unit / unit cell.

* Corresponding author at: University of Leeds, Centre for the Digital Design of Drug Products, School of Chemical and Process Engineering, Leeds LS2 9JT, UK.
E-mail address: c.y.ma@leeds.ac.uk (C.Y. Ma).

https://doi.org/10.1016/j.xphs.2022.11.027
0022-3549/© 2022 The Authors. Published by Elsevier Inc. on behalf of American Pharmacists Association. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

advances made in the analysis of particulate properties, and, through this, stressed the utility of such approaches in pharmaceutical product design and manufacturing work.

Such an analysis can provide an improved understanding of the specific problems or bottlenecks in a drug's development or manufacturing process. However, a more general and comparative analysis of a subset of pharmaceutical structures can provide wider insights into overall trends in behaviours, both in terms of crystal lattice energies and packing together with their constituent intermolecular interactions that confer stability to the solid-state. Hence, an analysis of the CSD Drug Subset,[32] which encompasses all available crystal structures from the CSD that contain an approved drug molecule, can provide potential insights into the nature of pharmaceuticals and how they differ to each other with respect to a wider range of organic crystal structures. Such an approach can also be used to develop more drug-oriented predictive tools notably through building statistical models based on solid-state structural descriptors. In the paper[32], various statistical analyses of the CSD Drug Subset were performed including molecular weight, number of rotatable bonds, number of aromatic rings, element composition, partition coefficient (cLogP), number of hydrogen-bond acceptors, number of hydrogen-bond donors, space group, packing coefficient, etc. with comparisons to other organic molecules in the CSD and also to the internal crystal structure databases from two pharmaceutical companies.

This paper is the first attempt to extend the 2019 CSD Drug Subset paper[32] to study a subset of drug crystal structures on not only the distributions of their molecular and crystal properties but also their energetic profile and its distribution in 3D including lattice energy, synthonic interactions, and functional group contributions, utilising the molecular modelling software HABIT98 and the inter-relationships between these properties and descriptors. As a result, the study intended to focus on the development of a workflow with a selected set of drug structures under the following conditions/limitations:

1) The study was restricted to a selected set of 487 crystal structures of small molecule pharmaceuticals after excluding the structures with $Z' \neq 1$ (including co-crystals, hydrates, salts and solvates), structures containing certain higher atomic mass elements (B, Br, I, Si, P), and structures having lattice energy > −20 kcal mol^{-1} (see Fig. S1 in Supplementary Material) from a selected set of drug crystal structures based on the best representative CSD organic set in 2019 CSD Drug Subset paper.[32]

   Although the crystal structures of discovered and approved drugs over the last decade have become more complex,[33-35] the selected drug structures (487) represent small molecule pharmaceutical crystals (93.6% have a molecular weight < 500) with the majority of them being found to meet the "Lipinski rule of 5" criteria (>85%) as shown in Fig. S5 as well as being solved before 2010 (Fig. S2).
2) The crystal structures in this study were taken directly from the CSD without pre-processing (although hydrogen atom positions were automatically assigned if missing within the structures):

   Not optimising hydrogen positions in molecules

   As shown in Fig. S3, a small set of 14 structures was tested with and without the optimisation of hydrogen atom positions in the crystal structures. It was found that based on the small set of structures, the lattice energy differences between the calculations with and without hydrogen optimisation are less than 16.6% with an average of 5%. This indicated that the application of the developed workflow for the energetic analysis with the crystal structures directly from the deposited ones in the CSD would not generate big differences from the same structures after hydrogen atom optimisation. Despite this, one needs to be mindful that for more complex structures, the whole structure would need to be minimised.

   Not including conformation in molecules

   The effects of conformational flexibility and torsion angles would make this study much more complicated, hence further deviating the workflow development and this proof-of-concept study. We fully understand the importance of conformational flexibility and torsion angles as shown in e.g.,[35] hence these factors will be addressed in further work and be reported in due course.
3) The use of Dreiding 2 forcefield:

   This study is the first attempt to analyse both geometrical and energetic properties/descriptors of the selected drug structures deposited in the CSD using the widely used molecular modelling software HABIT98. The Dreiding 2 forcefield has been widely used for HABIT98 calculations with reasonable results. This study seeks to explore a rapid computational workflow, and as such higher level calculations, which are not implemented in HABIT98, were not investigated.

In this study, and subject to the above-mentioned conditions/limitations, the selected drug subset has been used to investigate the diversity of lattice energies within the crystal structures of pharmaceuticals, and to understand diversity in terms of the strength and chemical nature of the constituent intermolecular synthons associated with their crystal chemistry. In this, molecular and crystallographic descriptors within the selected drug subset have been re-examined based on interatomic energy calculations using empirical forcefield inter-molecular modelling revealing trends in lattice energies, energy convergence, synthon strength distribution, as well as characterising the contributions that different functional groups make within the lattice energy balance.

Methodology

Dataset Preparation

The CSD Drug Subset[32] contains a total of 8632 crystal structures of 785 unique drug compounds in a variety of solid forms. Based on one of the refined subsets[32], the best representative CSD organic set, a total number of 628 structures was used as a starting point for preparing the dataset examined in this study. Duplicate structures, structures where Z' is not equal to 1 (including co-crystals, hydrates, salts and solvates), and structures containing certain higher atomic mass elements (B, Br, I, Si, P), were all removed, which led to a reduced drug subset of 577 structures. Intermolecular packing energies were directly calculated with the molecules in the crystal unit cell being based upon their published molecular coordinates and crystal structures. In this, the energetic contributions were based upon the synthons within the structures as calculated through the summation of intermolecular interactions between the molecules within the 3D crystallographic structure. Intramolecular interaction were not analysed in this study.[23,36-38] A further 90 structures where the lattice energy was calculated to be greater than -20 kcal mol^{-1} were not considered so as not to bias subsequent analysis. All the structures, including those not considered, did not undergo any optimization of their structures. The final dataset for which the analysis was presented therefore contained a selected drug subset of 487 crystal structures (see Fig. S1 for the distribution of the 628 structures from the best representative CSD organic set in the 2019 Drug Subset paper).[32]

Molecular and Crystallographic Descriptors and Functional Groups

The CSD Python API[1] was used to calculate a range of molecular and crystallographic descriptors for each unique drug structure in our dataset. If some structures did not have coordinates for hydrogen

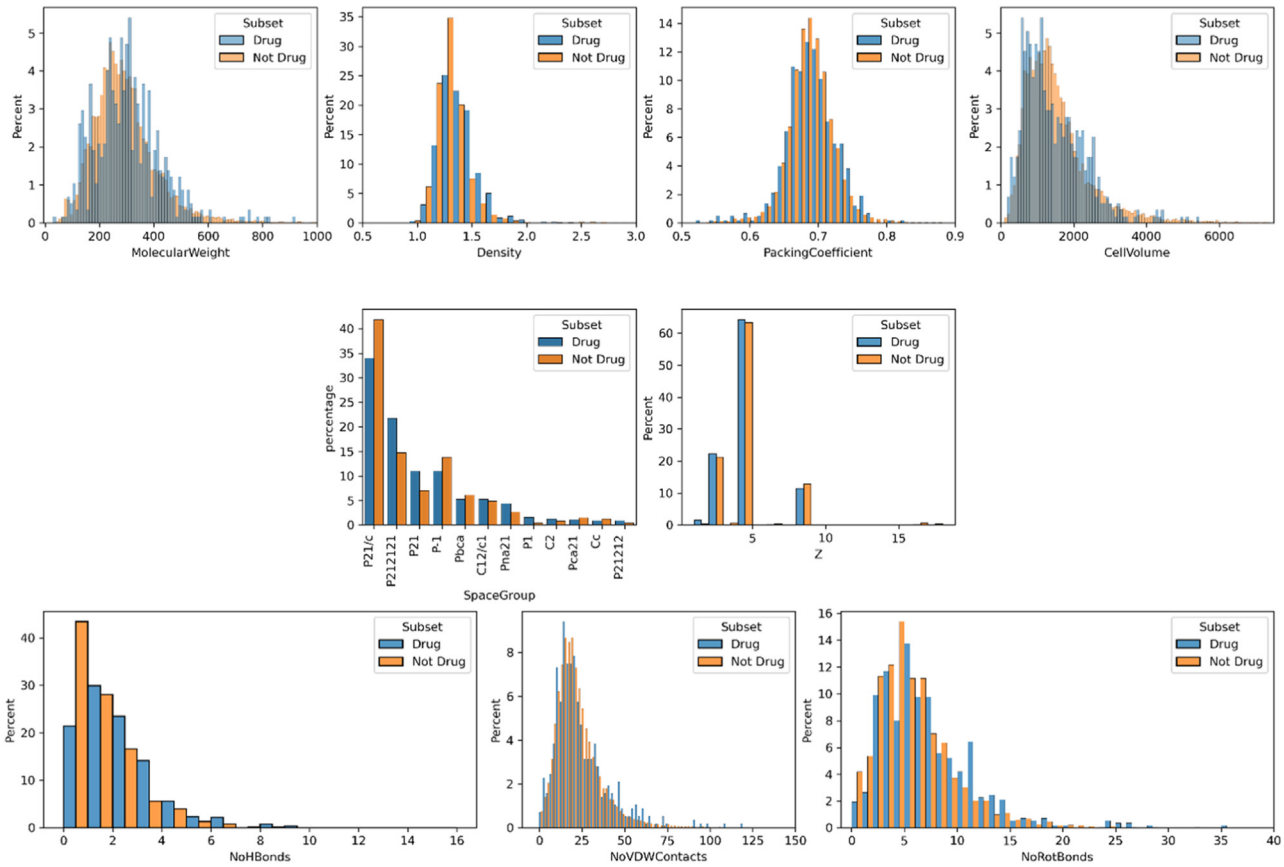

Figure 1. Distributions of molecular and crystallographic descriptors within the selected drug subset. From top-left: molecular weight; calculated density; packing coefficient; unit cell volume; space group; Z value; number of hydrogen bonds; number of short contacts and number of rotatable bonds.

atoms, the python API automatically added them to the structures. The properties selected were: molecular weight; density; space group; unit cell volume; packing coefficient; void volume (calculated with a probe radius of 0.2 Å and a grid spacing of 0.2 Å);[39] number of hydrogen bonds; number of short contacts (defined where the interatomic distance is less than the sum of the atomic van der Waals radii + 0.5 Å, which is long for regular analysis, but was used here to approximate the "coordination shell", hence effectively identifying packing in the crystal structures); number of molecules in the unit cell (*Z*-value); number of molecules in the asymmetric unit (*Z'*); number of rotatable bonds. Additionally, the functional groups in each molecule were automatically identified based on contact atom or central atom groups[40,41] in the library of information about non-bonded interactions implemented in the CSD.

Intermolecular Interactions and Lattice Energies

All intermolecular interactions were calculated using the Dreiding 2 force field parameters.[42] Atomic charges were calculated using the semi-empirical quantum mechanics program MOPAC[43] with the Austin Model 1 (AM1) approach, which has been widely used for simulations of molecular crystals.[44] Polarization energies were not calculated separately in this study. Lattice energies and intermolecular interaction energies were calculated using HABIT98, an enhanced version of HABIT95[45] developed from HABIT.[14] Completion of each calculation was confirmed by increasing the limiting radius in 5 Å steps to 35 Å until a difference of less than 0.1% of total energy between the final two steps was achieved. Individual intermolecular interactions, or "synthons", were classified based on the major contributing term to the total energy

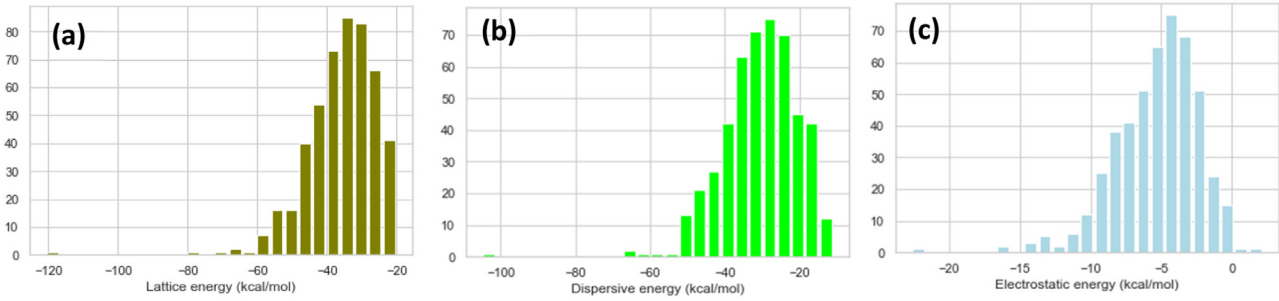

Figure 2. Number density distributions of (a) lattice (b) dispersive, and (c) electrostatic energies at a 30 Å limiting radius in the selected drug subset. Note that the y-axis represents the counts of crystal structures.

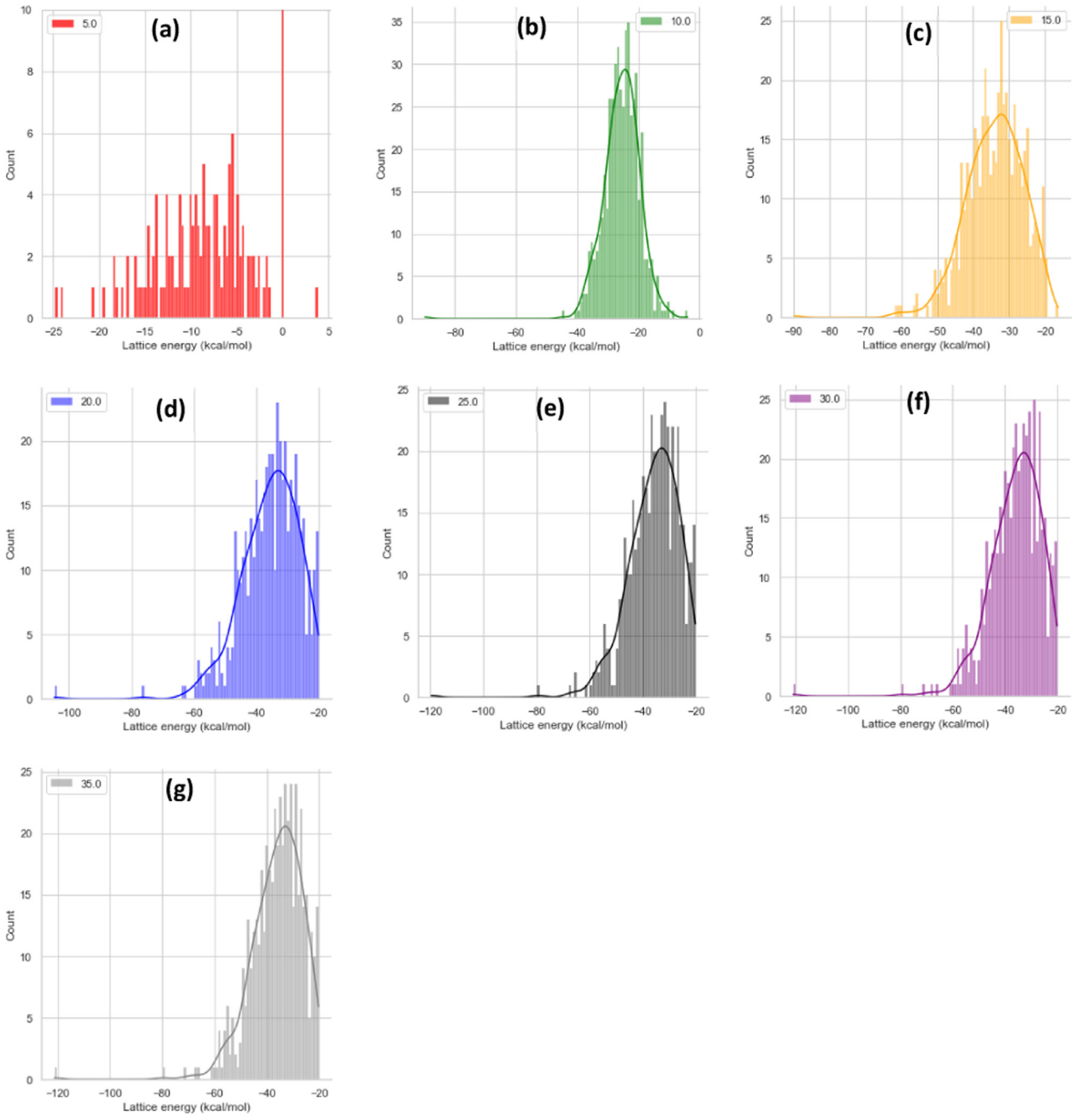

Figure 3. Distributions of lattice energies across increasing limiting radii (5 – 35 Å, a – g).

(electrostatic, van der Waals, or hydrogen bonding) and ranked based on the magnitude of the total interaction energy.[21,46,47]

Note that in this paper, the energetic values of crystal lattice energy together with their dispersive/electrostatic contributions, synthon strength and functional group contributions are by convention negative. A compound with higher lattice energy, a synthon with higher strength and a functional group making larger contribution were identified by their higher absolute energy values. In this work, intra-molecular energy contributions to the crystal lattice energies were not considered.

Data Processing

Following on from generation of molecular and crystallographic descriptors, and the calculation of lattice and intermolecular interaction energies, all subsequent data analysis and presentation was performed in Python using the Pandas,[48] Matplotlib,[49] and Seaborn[50] libraries.

Discussion and Analysis

Distributions of Molecular and Crystallographic Descriptors

Fig. 1 shows the distributions of a variety of molecular and crystallographic descriptors across the selected drug subset. "Non-drug" molecules are here defined[32] as all molecules of organic structures in the CSD that are not included in the CSD Drug Subset. The distributions are not dissimilar from those generated in the 2019 study[32] and, perhaps unsurprisingly, indicate that drug molecules pack in similar ways to small non-drug molecules in the solid-state. A

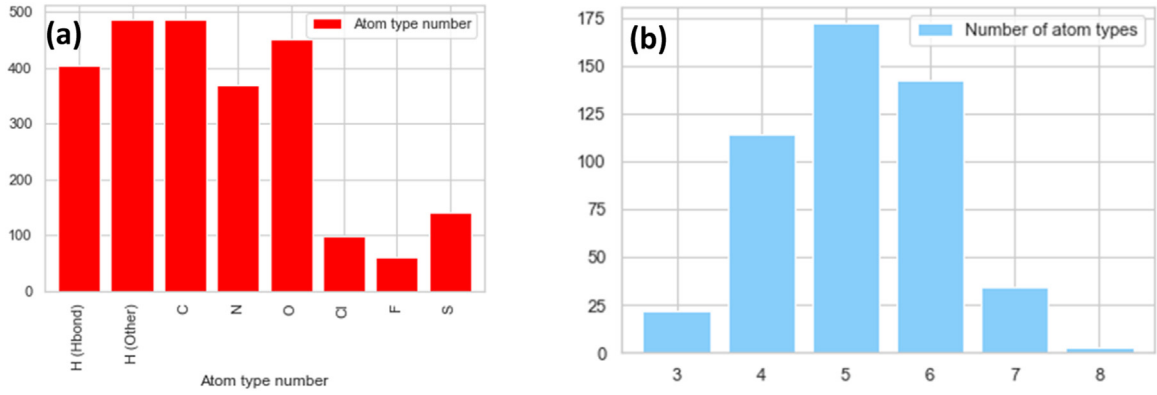

Figure 4. Distribution of different atom types across the selected drug subset (a) and distribution of number of different atom types across the structures in this dataset studied (b). Note that the y-axis represents the counts of crystal structures.

relatively higher incidence of Sohncke space groups[32,51] in the selected drug subset reflects the often-chiral nature of drug molecules, and it was noteworthy that a larger proportion of drug crystal structures had a *Z*-value of 1 compared to non-drug molecules.[32] The latter, when combined with the larger percentage of structures in the space group *P*1, perhaps indicates lower symmetry in the crystal structures of drugs. The relative differences are modest, however, and do not detract from the overall similarities observed between the two subsets.

The relationships between these descriptors have been plotted in the Supplementary Material (**Figs. S6** and **S7**). Some clear, especially linear, correlations were identified and further analysed including the linear relationships between molecular weight and crystal cell volume (**Fig. S8**), cell volume normalized by Z and molecular weight (**Fig. S9a**), crystal void volume and packing coefficient (**Fig. S9b**). Of the above structures, Probucol (HAXHET) has the highest void volume (45.92%), lowest packing coefficient (0.52) with the lowest density (1.05 g cm^{-3}) characterized by a structure without any hydrogen bonds and with only a small electrostatic contribution (0.04 kcal mol^{-1}) to the lattice energy (-40.67 kcal mol^{-1}) with a crystal chemistry associated with only pure vdW interactions. Note that these trends are only for the particular entries within the selected subset of drug structures. It was found that within these structures, about 82% had fewer than 10 rotatable bonds (Fig. 1). Generally, the higher the molecular weight, the higher number of rotatable bonds exist in a drug molecule of the selected drug subset with no clear trend with the other descriptors (**Fig. S6**).

Distributions of Lattice Energies and Their Components

Fig. 2 shows the distributions of lattice energies with the component electrostatic and dispersive energies, across the crystal structures in the selected drug subset. Subsequent analysis of these distributions reveals a mean lattice energy of -36.15 kcal mol^{-1} with a standard deviation of 10.08 kcal mol^{-1}, a mean dispersive energy of -30.78 kcal mol^{-1} with a standard deviation of 9.92 kcal mol^{-1}, and a mean electrostatic energy of -5.36 kcal mol^{-1} with a standard deviation of 3.02 kcal mol^{-1}. The magnitudes of the standard deviations are unsurprisingly large reflecting the diversity of molecular and crystal properties within the subset. However, the distributions themselves give an indication of the average "stability" of a drug-like molecular solid.

The distributions in Fig. 2 also show that the dispersive component makes up most of the lattice energy. However, there does not exist any clear trend between dispersive energy (lattice energy) and the crystal descriptors. A linear fit across all structures in the dataset gives a straight line with the equation $E_{latt} = 0.96 \times E_{disp} - 0.652$, with an R^2 value of 0.91, indicating the importance of weaker intermolecular interactions between neutral drug molecules in the solid-state. Further distributions of crystal descriptors and number of polymorphs against lattice energy were plotted in **Figs. S11** and **S12** in the Supplementary Material.

Lattice Energy Convergence

Fig. 3 shows the distribution of lattice energies across a range of limiting radii from 5 Å up to 35 Å across the selected drug subset. Mean values for each limiting radius, along with standard deviations on these mean values and the increment at each limiting radius (expressed as a percentage of the total energy) are given in Table 1. The distributions of the dispersive and electrostatic components at increasing limiting radii can be found in **Fig. S15** in the Supplementary Material.

The highest increase in lattice energy occurs when the limiting radius is increased from 5 Å to 10 Å. For most systems, this range is likely to correlate with filling the first inter-molecular co-ordination shell, those neighbouring molecules with the shortest and most energetically favourable intermolecular contacts to the central molecule. At a limiting radius of 25 Å the increase in lattice energy drops to around 1% of the total energy, as the strength of the interactions drops away due to increasing distance. The negligible difference upon increasing the limiting radius from 30 Å to 35 Å justifies the selection of the former value as the default for the selected drug structures studied in this paper.

Atom Type Contributions to Lattice Energies

Fig. 4a shows the distribution of different atom types across the selected drug subset. All 487 structures in our dataset contain carbon and "non-hydrogen bonded" hydrogen atoms. Around 83% of structures contain "hydrogen bonded" hydrogen atoms. The distribution of different heteroatoms across the dataset agrees with previous observations of the CSD Drug Subset.[32] Fig. 4b shows the diversity of atom types across individual structures in the dataset. 22 structures contain only three different atom types; oxygen is the most common heteroatom in these structures, the others being nitrogen and fluorine.

Fig. 5 shows the energetic contributions of the eight individual atom types considered in this study. Unsurprisingly, given its ubiquitous nature across the selected drug subset, carbon atoms provide the largest mean contributions to lattice energy. Some heteroatoms, particularly those involved in hydrogen bonding interactions, are observed to provide positive contributions to the otherwise negative

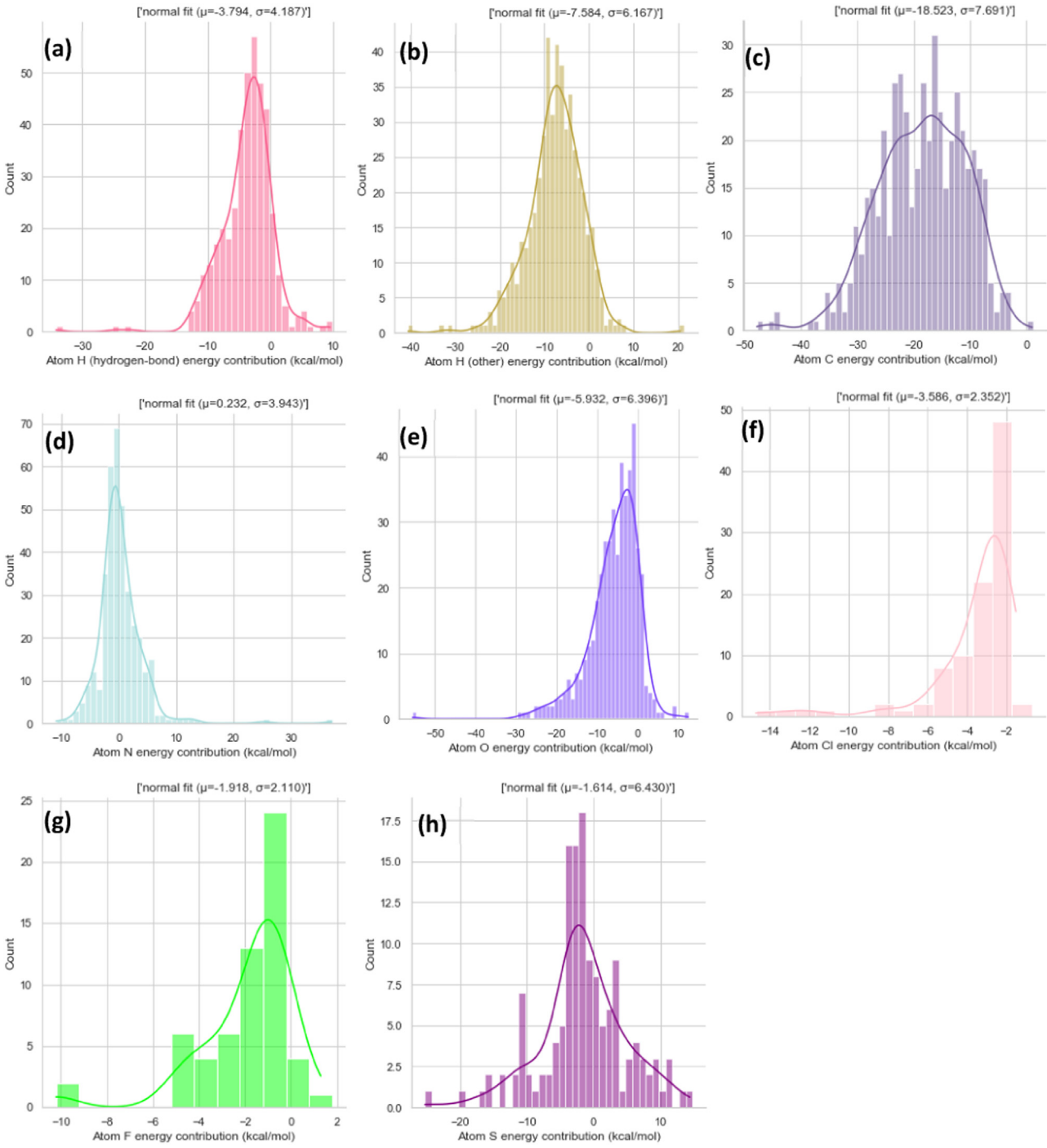

Figure 5. Distributions of the energetic contribution of eight individual atom types.

total energies. This indicates the repulsive nature of some short contacts that can exist between more electronegative atoms in crystal structures, but which are balanced by the overall stabilisation of close packing. Organic fluorine (Fig. 5g) shows a modest contribution towards the total energy, and is sometimes repulsive, in agreement with previous studies.[52]

Synthon Strength Distributions

Fig. 6a shows that, while the strongest intermolecular synthon in each structure in the selected drug subset can vary, the average energy is -5.79 kcal mol^{-1} with a standard deviation of 2.09 kcal mol^{-1}. Significant deviations away from these values may be indicators of the high or low stability of a given structure in the solid-state. Form II of Ritonavir (YIGPIO03) has the strongest individual synthon (i.e., the largest interaction energy of -17.31 kcal mol^{-1}) across the selected drug subset, an indication of its large molecular weight and stable hydrogen bonding network.[53] Initial studies on the full dataset highlighted the structure of pyruvic acid (PRUVAC01) which has the lowest-ranked synthon with an interaction energy of only -0.77 kcal mol^{-1}, consistent with its low melting point of 13.8 °C.[54]

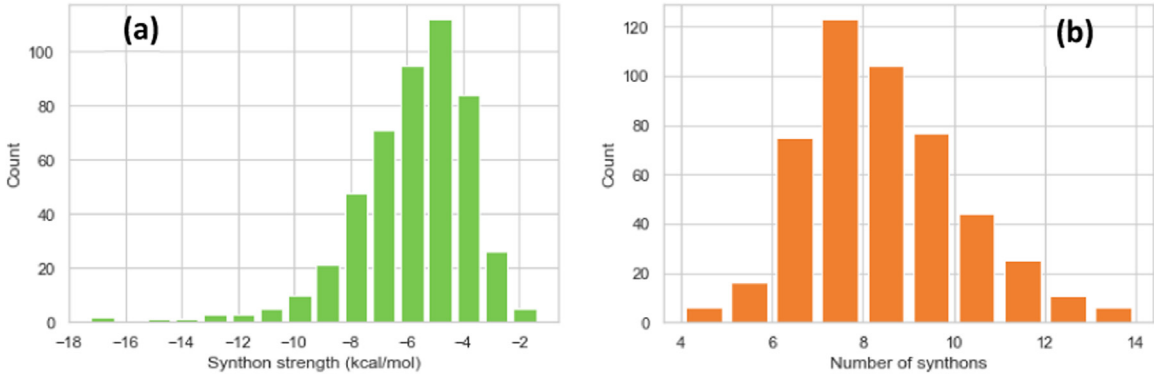

Figure 6. (a) Top-ranked synthon energy distribution and (b) number density distribution of unique synthons in the selected drug subset.

Table 1
Mean and standard deviations of energy distributions and increment (as a percentage of the final energy) at increasing limiting radii.

Radius (Å)	Lattice energy (kcal mol^{-1})		Dispersive energy (kcal mol^{-1})		Electrostatic energy (kcal mol^{-1})	
	Mean value	Increment (%)	Mean value	Increment (%)	Mean value	Increment (%)
5	-7.07 (4.80)	6.75 (14.41)	-6.27 (4.45)	7.37 (13.56)	-0.17 (1.00)	3.17 (53.88)
10	-25.35 (6.31)	63.37 (19.88)	-21.58 (5.81)	62.74 (17.09)	-3.77 (3.15)	67.16 (91.54)
15	-34.08 (8.19)	24.15 (14.29)	-28.95 (8.26)	23.94 (12.77)	-5.13 (3.06)	25.37 (49.79)
20	-35.68 (9.46)	4.43 (4.69)	-30.32 (9.36)	4.45 (3.80)	-5.36 (3.02)	4.29 (76.72)
25	-36.05 (9.98)	1.02 (1.46)	-30.68 (9.84)	1.17 (1.36)	-5.36 (3.02)	0.1 (7.62)
30	-36.15 (10.08)	0.28 (0.52)	-30.78 (9.92)	0.33 (0.23)	-5.36 (3.02)	0.1 (4.33)
35	-36.15 (10.10)	0.1 (0.22)	-30.78 (9.93)	0.1 (0.03)	-5.36 (3.01)	0.1 (3.05)

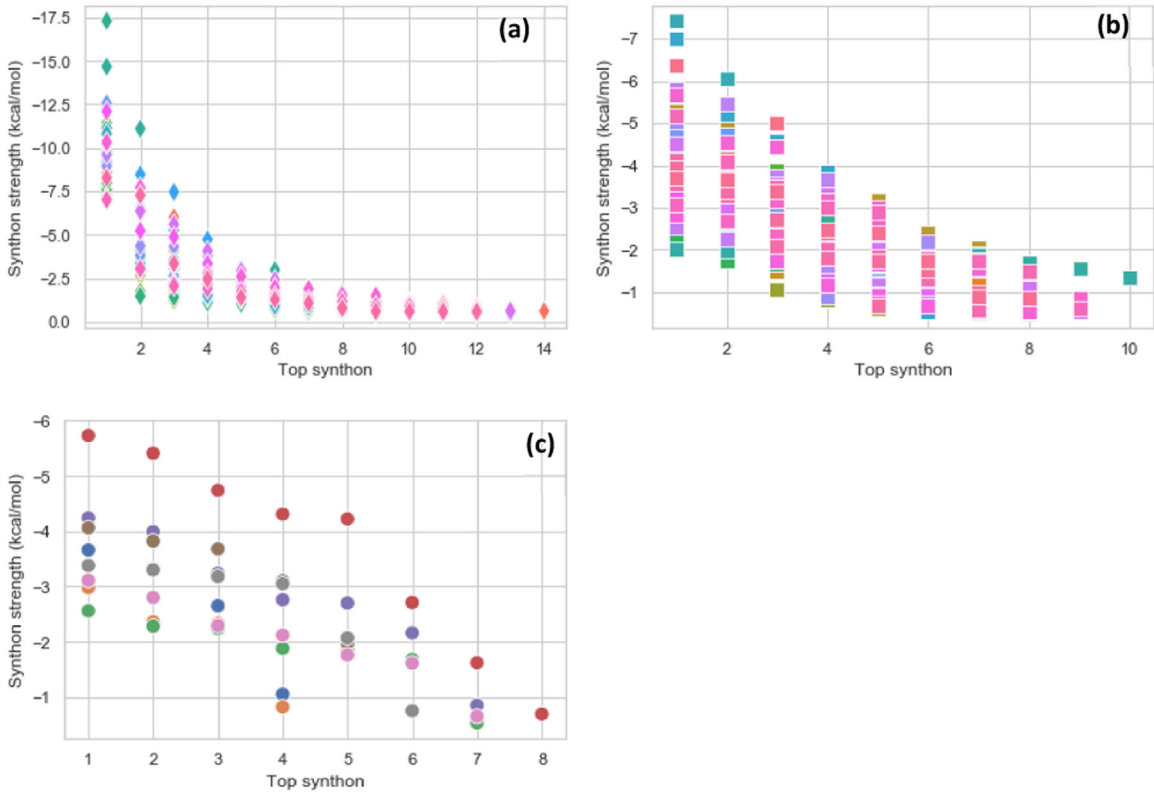

Figure 7. Decay in synthon energies across the selected drug subset, categorised as "fast" (27 structures) (a), "medium" (73 structures) (b), and "slow" (8 structures) (c).

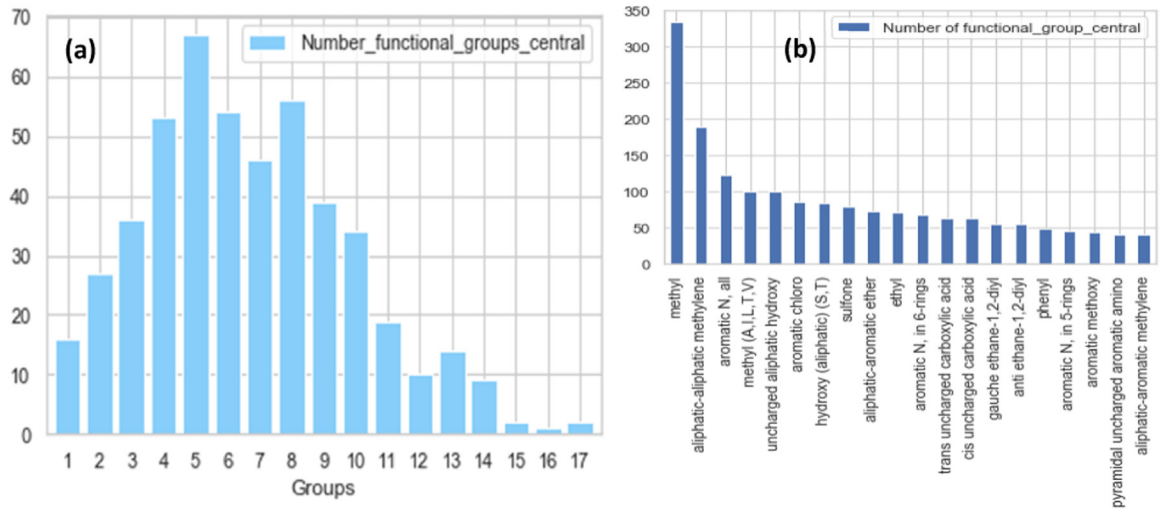

Figure 8. (a) Numbers of unique functional groups in structures from the selected drug subset. (b) Number of observations of the top twenty most common functional groups across the selected drug subset. Note that the y-axis represents the counts of crystal structures.

Table 2
Descriptor ranges for structures with different synthon "decay rates".

Descriptor	"Fast"	"Medium"	"Slow"
Molecular weight (g mol^{-1})	120 – 530	210 – 720	150 – 380
Density (g cm^{-3})	1.02 – 1.88	1.28 – 1.65	1.28 – 1.65
Packing coefficient	0.59 – 0.76	0.65 – 0.77	0.67 – 0.75
Cell volume (A^3)	460 – 3400	700 – 5100	600 – 2100
Number of hydrogen bonds	0 – 8	0 – 8	0 – 5
Number of vdW contacts	2 – 73	12 – 73	2 – 48
Z	1, 2, 4, 8	1, 2, 4, 8	2, 4, 8

Most structures (Fig. 6b) have several unique synthons (interactions not related by symmetry) commensurate with observed *Z* values and typical molecular close packing. The unique synthons were identified by their different individual energies; only interactions with energies below 0 kcal/mol were considered. Extremes in this distribution can be indications of molecular size and packing symmetry. For instance, Gefitinib (FARRUM02), a large molecule, whose form I polymorph packs in space group P-1, has 14 unique synthons.

The decay in synthon strength from the top-ranked interaction to the weakest unique synthon varies across the selected drug subset (Fig. 7), and this behaviour can be broadly categorised into those which decay "fast" (where the synthon strength has dropped by around 75% by synthon 6), "medium" (where the synthon strength has dropped by around 50% by synthon 6) and "slow" (where the drop is much more gradual). Analysis of the descriptors of these different categories (Table 2) do not show much variance. Smaller molecules with fewer interactions (either hydrogen bonding or van der Waals contacts) might be said to decay more slowly, although the strongest synthons in this category (which contains only 8 structures) are generally lower than those of the other categories.

Functional Group Analysis

To understand the variation in the calculated synthon energies, further analysis of the diversity of functional groups across the selected drug subset was performed. Fig. 8a shows the distribution of the number of different functional groups that are contained within each molecule in the structures of our dataset. The automatic assignment of functional groups based upon existing "central groups" means that some more complex functionality, such as the carbamoyl central group, appear as the only functional group in several structures such as oxcarbazepine (CANDUR01) and carbamazepine (CBMZPN12, CBMZPN16, CBMZPN18). Larger, more complex drug molecules, such as Lapatinib (OVAYOB) and Ritonavir (YIGPIO03), contain up to 17 unique functionalities. Fig. 8b shows the occurrence of the twenty most common functional groups across the selected drug subset. Terminal methyl groups and aliphatic methylene groups are unsurprisingly common, forming the backbone of numerous molecules, but we note the presence of various functionalities containing heteroatoms, such as aromatic nitrogen, hydroxy, and sulfone groups, that provide further insight into the distribution of atom types observed over the selected drug subset.

The energy distributions of the 15 most common functional groups (as defined in the CSD functional group library[32]) are shown in Fig. 9, and the mean energies and their standard deviations are listed in **Table S3** (Supplementary Material). In Fig. 8, one crystal structure, Lapatinib (OVAYOB), has 17 functional groups, the highest number of the structures studied. Sixteen structures have only one functional group (based on the current CSD library[32]): aliphatic chloro (3); aromatic N, all (1); aromatic N-H, uncharged, in 5-rings (1); carbamoyl (3); coumarin (1); methyl (1); nitro (2) and phenyl (4 structures). As shown in **Table S3**, Phenyl groups, with a mean energy of -11.28 kcal mol^{-1}, provide the most contribution to the total lattice energy of the functional groups studied, highlighting the potential importance of aromatic interactions in the crystal structures of drug molecules. Aliphatic hydroxy groups, with a mean value of -3.06 kcal mol^{-1}, make the lowest contribution of these 15 functionalities (Fig. 9 and **Table S3**). This may result from the weakness of hydroxyl oxygen-based hydrogen bonds, although Fig. 9d suggests that interactions involving this functional group may be slightly destabilising in some instances.

Figure 9. Energy (kcal mol^{-1}) distributions across 15 of the most common functional groups. Each functional group definition is given on each distribution (a) through (o). Note that the y-axis represents the counts of crystal structures.

Of the 15 common functional groups defined by the CSD functional group library,[32] 6 functional groups (Fig. 9e, f, i, j, m & n) make negative energy contributions to their lattice energies, hence stabilising the crystal structures. The other 9 functional groups (Fig. 9a-d, g, h, k, l & o) occasionally provide positive energy contributions to a few crystal structures: methyl (Lapatinib, Sumatriptan, Teniposide, Etoricoxib), aliphatic-aliphatic methylene (Lapatinib, Vemurafenib), aromatic N, all (Sulfasalazine, Almitrine, Adenosine), uncharged aliphatic hydroxyl (Chloramphenicol palmitate, Rapamycin, α-Flupenthixol, Cephalotaxine, Difluprednate, Teniposide, Quinidine, Levonorgestrel), hydroxy (aliphatic) (S,T) (α-Flupenthixol, Chloramphenicol palmitate, Rapamycin), sulfone (Famotidine), aromatic N, in 6-rings (Sulfasalazine), trans (or cis) uncharged carboxylic acid (Stearic acid (E), Deferasirox, Stearic acid, Vitamin A acid, Flufenamic acid (II), Sulfasalazine), aromatic N, in 5-rings (Nilotinib).

The functional group, trans (or cis) uncharged carboxylic acid, in the structure of Stearic acid (form E) has a positive energy of 15.79 kcal mol^{-1} (Fig. 9l) with an 18-carbon long chain molecular structure, which is the most destabilising case. These observations are dependent on the quality of the molecular coordinates and unit cell geometries encompassed within the published crystal structures as well as the forcefield used for the calculations. Nevertheless, they do provide considerable insight into the nature of the chemistry and energies associated with the intermolecular interactions that drive the crystallisation and structural stability of organic solids.

Conclusions

An automated informatics and computational analysis to understand the molecular, crystallographic, and energetic properties of a selection of 487 structures taken from the CSD Drug Subset has been performed. While the crystal structures of drug molecules pack in similar ways to other organic molecules in the CSD, analysis of their lattice energies and synthons may provide additional insight into the solid-state structures of pharmaceuticals.

Mindful of the stated conditions and restrictions highlighted in the Introduction, the mean lattice energy across the selected drug subset structures of small molecule pharmaceuticals approximately -36 kcal mol^{-1}, with a standard deviation of 10.08 kcal mol^{-1}. Around 85% of the lattice energy comes from dispersive interactions, and the remaining 15% is attributable to electrostatics. Analysis of the convergence of the lattice energy calculation indicates that at a limiting radius of 30 Å, the calculated energy has reached 99.9% of its final value for the molecular systems investigated in this study. The average energy of the strongest synthon in each structure across the selected drug subset is -5.79 kcal mol^{-1} with a standard deviation of 2.09 kcal mol^{-1}, and molecules tend to display a varying number of unique synthons in accordance with conventional packing arrangements and symmetries.

Functional group analysis across the selected drug dataset demonstrates the diversity of chemical space in drug molecules, in agreement with analysis of atom types across the wider CSD Drug Subset. The highest mean energy of -11.28 kcal mol^{-1} for phenyl groups shows the importance of aromatic interactions in pharmaceuticals, while this approach highlights the sometimes-destabilising nature of close contacts in molecular crystals.

An extension of the approach used in this study to enable calculations on charged molecular species, multicomponent systems, and crystals with more than one molecule in the asymmetric unit is ongoing, and this will enable further analysis of the diverse solid form landscape of the whole CSD Drug Subset. Mindful that hydrogen atoms have lower electron densities compared to higher molecular weight atoms, their atomic positions within published crystal structures are not always optimum. Another extension of the approach could include the optimisation of hydrogen positions of the drug molecules and also conformational optimisation of the crystal structures within the selected drug subset. The latter could also take into account the deformation energies i.e., comparing conformation energy differences (deformation strain) between molecular equilibrium with that in the solid-state (e.g.[20,23,36]). Further studies are also planned to analyse the energy differences between polymorphs, particularly non-conformational polymorphs that display differences in the number of molecules in the asymmetric unit. Additionally, synthonic analysis of multicomponent systems might shed further light on the principles that govern their formation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors are grateful for the financial support of Innovate UK through the Digital Design Accelerator Platform Project (TS/T011262/1). This work also builds upon software developments funded through the ADDoPT and Synthonic Engineering Programs, supported by AMSCI (Grant No. 14060) in collaboration with AstraZeneca, Bristol-Myers Squibb, BRITEST, GSK, Perceptive Engineering, Pfizer, Process Systems Enterprise and the STFC Hartree Centre together with the Universities of Cambridge and Strathclyde, and the EPSRC (Grant EP/I028293/1) in collaboration with Pfizer, Boehringer Ingelheim, Novartis, and Syngenta, respectively.

Supplementary Materials

Supplementary material associated with this article (26 pages) can be found in the online version at doi:10.1016/j.xphs.2022.11.027.

References

1. Groom CR, Bruno IJ, Lightfoot MP, Ward SC. The Cambridge structural database. *Acta Crystallogr Sect B: Struct Sci, Cryst Eng Mater*. 2016;72(2):171–179.
2. Galek PTA, Pidcock E, Wood PA, Bruno IJ, Groom CR. One in half a million: a solid form informatics study of a pharmaceutical crystal structure. *CrystEngComm*. 2012;14(7):2391–2403.
3. Bryant MJ, Rosbottom I, Docherty R, Edge CM, Hammond RB, Peeling R, Pickering JP, Roberts KJ, Maloney AGP. Particle informatics": advancing our understanding of particle properties through digital design. *Cryst Growth Des*. 2019;19:5258–5266.
4. Galek PTA, Allen FH, Fábián L, Feeder N. Knowledge-based H-bond prediction to aid experimental polymorph screening. *CrystEngComm*. 2009;11(12):2634–2639.
5. Price SL. Computational prediction of organic crystal structures and polymorphism. *Int Rev Phys Chem*. 2008;27(3):541–568.
6. Wood PA, Olsson TSG, Cole JC, Cottrell SJ, Feeder N, Galek PTA, Groom CR, Pidcock E. Evaluation of molecular crystal structures using Full Interaction Maps. *CrystEngComm*. 2013;15(1):65–72.
7. Gozalbes R, Pineda-Lucena A. QSAR-based solubility model for drug-like compounds. *Bioorg Med Chem*. 2010;18(19):7078–7084.
8. Klamt A, Eckert F, Hornig M, Beck ME, Bürger T. Prediction of aqueous solubility of drugs and pesticides with COSMO-RS. *J Comput Chem*. 2002;23(2):275–281.
9. Bergström CAS, Larsson P. Computational prediction of drug solubility in water-based systems: qualitative and quantitative approaches used in the current drug discovery and development setting. *Int J Pharm*. 2018;540(1-2):185–193.
10. Hong RS, Mattei A, Sheikh AY, Bhardwaj RM, Bellucci MA, McDaniel KF, Pierce MO, Sun G, Li S, Wang L, Mondal S, Ji J, Borchardt TB. Novel physics-based ensemble modeling approach that utilizes 3D molecular conformation and packing to access aqueous thermodynamic solubility: a case study of orally available bromodomain and extraterminal domain inhibitor lead optimization series. *J Chem Inf Model*. 2021;61(3):1412–1426.
11. Abramov YA. Major source of error in QSPR prediction of intrinsic thermodynamic solubility of drugs: solid vs nonsolid state contributions? *Mol Pharmaceutics*. 2015;12(6):2126–2141.
12. Boobier S, Hose DRJ, Blacker AJ, Nguyen BN. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nat Commun*. 2020;11. Article number: 5753.
13. Berkovitch-Yellin Z. Toward an ab initio derivation of crystal morphology. *J Am Chem Soc*. 1985;107(26):8239–8253.

14. Clydesdale G, Docherty R, Roberts KJ. HABIT - a program for predicting the morphology of molecular crystals. *Comput Phys Commun*. 1991;64(2):311–328.
15. Lovette MA, Doherty MF. Needle-shaped crystals: Causality and solvent selection guidance based on periodic bond chains. *Cryst Growth Des*. 2013;13(8):3341–3352.
16. Winn D, Doherty MF. A new technique for predicting the shape of solution-grown organic crystals. *AlChE J*. 1998;44(11):2501–2514.
17. Nguyen TTH, Rosbottom I, Marziano I, Hammond RB, Roberts KJ. Crystal morphology and interfacial stability of rsibuprofen in relation to its molecular and synthonic structure. *Cryst Growth Des*. 2017;17:3088–3099.
18. Pickering J, Hammond RB, Ramachandran V, Soufian M, Roberts KJ. Synthonic engineering modelling tools for product and process design. In: Roberts KJ, Docherty R, Tamura R, eds. *Engineering Crystallography: From Molecule to Crystal to Functional Form*. Dordrecht: Springer Netherlands; 2017:155–176.
19. Rosbottom I, Roberts KJ, Docherty R. The solid state, surface and morphological properties of p-aminobenzoic acid in terms of the strength and directionality of its intermolecular synthons. *CrystEngComm*. 2015;17(30):5768–5788.
20. Turner TD, Dawson N, Edwards M, Pickering J, Hammond RB, Docherty R, Roberts KJ. A digital mechanistic workflow for predicting solvent-mediated crystal morphology: the α and β forms of L-glutamic acid. *Cryst Growth Des*. 2022;22:3042–3059.
21. Hammond RB, Pencheva K, Roberts KJ. A structural-kinetic approach to model face-specific solution/crystal surface energy associated with the crystallization of acetyl salicylic acid from supersaturated aqueous/ethanol solution. *Cryst Growth Des*. 2006;6(6):1324–1334.
22. Ballard D, Pickering J, Rosbottom I, Tangparitkul S, Roberts KJ, Rae R, Dowding P, Hammond RB, Harbottle D. Molecular survey of strongly and weakly interfacially active asphaltenes: an intermolecular force field approach. *Energy Fuel*. 2021;35:17424–17433.
23. Wang C, Rosbottom I, Turner TD, Laing S, Maloney A, Sheikh AY, Docherty R, Yin Q, Roberts KJ. Molecular, solid-state and surface structures of the conformational polymorphic forms of ritonavir in relation to their physical chemical properties. *Pharm Res*. 2021;38:971–990.
24. Kaskiewicz PL, Rosbottom I, Hammond RB, Warren NJ, Morton C, Dowding PJ, George N, Roberts KJ. Understanding and designing tailor-made additives for controlling nucleation: case study of p-aminobenzoic acid crystallising from ethanolic solutions. *Cryst Growth Des*. 2021;21:1946–1958.
25. Rosbottom I, Pickering J, Hammond RB, Roberts KJ. A digital workflow supporting the selection of solvents for optimizing the crystallizability of p-aminobenzoic acid. *Org Process Res Dev*. 2020;24:500–507.
26. Rosbottom I, Turner TD, Ma CY, Hammond RB, Roberts KJ, Yong CW, Todorov IT. The structural pathway from its solvated molecular state to the solution crystallisation of the α- and β-polymorphic forms of para amino benzoic acid. *Faraday Discuss*. 2022;235:467–489.
27. Rosbottom I, Yong CW, Geatches D, Hammond RB, Todorov IT, Roberts KJ. DL_POLY/DL_FIELD/DL_ANALYSER – an integrated software platform for molecular simulations to explore the synthonic interactions in benzoic acid/hexane solutions. *Mol Simul*. 2021;47:257–272.
28. Ramachandran V, Murnane D, Hammond RB, Pickering J, Roberts KJ, Soufian M, Forbes B, Jaffari S, Martin GP, Collins E, Pencheva K. Formulation pre-screening of inhalation powders using computational atom–atom systematic search method. *Mol Pharmaceutics*. 2015;12(1):18–33.
29. Nguyen TTH, Hammond RB, Styliari ID, Murnane D, Roberts KJ. A digital workflow from crystallographic structure to single crystal particle attributes for predicting the formulation properties of terbutaline sulphate. *CrystEngComm*. 2020;22:3347–3360.
30. Roberts KJ, Hammond RB, Ramachandran V, Docherty R. Synthonic engineering: from molecular and crystallographic structure to the rational design of pharmaceutical solid dosage forms. editor. In: Abramov YA, ed. *Computational Approaches in Pharmaceutical Solid State Chemistry*. John Wiley & Sons, Ltd; 2015:175–210.
31. Roberts KJ, Docherty R, Taylor S. Material science: solid form design and crystallisation process development. In: Blacker J, Williams MT, eds. *Pharmaceutical Process Development: Current Chemical and Engineering Challenges*. Cambridge: The Royal Society of Chemistry; 2011:286–313. RSC Drug Discovery Series No 9.
32. Bryant MJ, Black SN, Blade H, Docherty R, Maloney AGP, Taylor SC. The CSD drug subset: the changing chemistry and crystallography of small molecule pharmaceuticals. *J Pharm Sci*. 2019;108(5):1655–1662.
33. Bhutani P, Joshi G, Raja N, Bachhav N, Rajanna PK, Bhutani H, Paul AT, Kumar R. U.S. FDA approved drugs from 2015-June 2020: a perspective. *J Med Chem*. 2021;64:2339–2381.
34. Brown DG, Wobst HJ. A decade of FDA-approved drugs (2010-2019): trends and future directions. *J Med Chem*. 2021;64:2312–2338.
35. Sheikh AY, Mattei A, Bhardwaj RM, Hong RS, Abraham NS, Schneider-Rauber G, Engstrom KM, Diwan M, Henry RF, Gao Y, Juarez V, Jordan E, DeGoey DA, Hutchins CW. Implications of the conformationally flexible, macrocyclic structure of the first-generation, direct-acting anti-viral paritaprevir on its solid form complexity and chameleonic behavior. *J Am Chem Soc*. 2021;143(42):17479–17491.
36. Hammond RB, Pencheva K, Roberts KJ. Structural variability within, and polymorphic stability of, nano-crystalline molecular clusters of L-glutamic acid and D-mannitol, modelled with respect to their size, shape and 'crystallisability. *CrystEngComm*. 2012;14:1069–1082.
37. Hammond RB, Pencheva K, Roberts KJ. An examination of the polymorphic stability and molecular conformational flexibility as a function of crystal size associated with the nucleation and growth of benzophenone. *Faraday Discuss*. 2007;136:87–102.
38. Roberts KJ, Docherty RPB, Jetten LAMJ. The importance of considering growth-induced conformational change in predicting the morphology of benzophenone. 1993 *J Phys D: Appl Phys*. 1993;26B:7–21. 26B.
39. Barbour LJ. Crystal porosity and the burden of proof. *Chem Commun*. 2006:(11):1163–1168.
40. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. *Nucleic Acids Res*. 2000;28:235–242.
41. Bruno IJ, Cole JC, Lommerse JPM, Rowland RS, Taylor R, Verdonk ML. IsoStar: a library of information about non-bonded interactions. *J Comput -Aided Mol Des*. 1997;11:525–537.
42. Mayo SL, Olafson BD, Goddard WA. DREIDING: a generic force field for molecular simulations. *J Phys Chem*. 1990;94(26):8897–8909.
43. Stewart JJP MOPAC 6.0,(CQCPE program# 455). Quantum chemistry program exchange, creative arts building 181, Indiana University, Bloomington, IN 47405 USA.
44. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP. Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. *J Am Chem Soc*. 1985;107(13):3902–3909.
45. Clydesdale G, Roberts KJ, Docherty R. HABIT95 - a program for predicting the morphology of molecular crystals as a function of the growth environment. *J Cryst Growth*. 1996;166:78–83.
46. Hammond RB, Ma C, Roberts KJ, Ghi PY, Harris RK. Application of systematic search methods to studies of the structures of urea-dihydroxy benzene cocrystals. *J Phys Chem B*. 2003;107(42):11820–11826.
47. Hammond RB, Pencheva K, Ramachandran V, Roberts KJ. Application of grid-based molecular methods for modeling solvent-dependent crystal growth morphology: aspirin crystallized from aqueous ethanolic solution. *Cryst Grow Des*. 2007;7(9):1571–1574.
48. Reback J, Jbrockmendel, McKinney W, Bossche VdJ, Augspurger T, Cloud P, Hawkins S, Gfyoung, Roeschke M, Sinhrks, Klein A, Petersen T, Tratner J, She C, Ayd W, Hoefler P, Naveh S, Garcia M, Schendel J, Hayden A, Saxton D, Darbyshire JHM, Shadrach R, Gorelli ME, Li F, Zeitlin M, Jancauskas V, McMaster A, Battiston P, Seabold S 2021. pandas-dev/pandas: Pandas 1.3.4.
49. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng*. 2007;9(3):90–95.
50. Waskom ML. seaborn: statistical data visualization. *J Open Source Software*. 2021;6(60):3021.
51. Flack HD. Chiral and achiral crystal structures. *Helv Chim Acta*. 2003;86(4):905–921.
52. Dunitz JD. Organic fluorine: odd man out. *ChemBioChem*. 2004;5(5):614–621.
53. Bauer J, Spanton S, Henry R, Quick J, Dziki W, Porter W, Morris J. Ritonavir: an extraordinary example of conformational polymorphism. *Pharm Res*. 2001;18(6):859–866.
54. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res*. 2018;46(D1):D1074–D1082.