"It's not a bad thing finding out that you don't have all the answers. You start asking the right questions." - Dr. Erik Selvig, *Thor*

UNIVERSITY OF
Southampton

Faculty of Natural and Environmental Sciences

Academic Unit of Chemistry

Crystalline Cheminformatics

Big Data Approaches to Crystal Engineering

by

Philip David Felix Adler

Thesis for the degree of Doctor of Philosophy

January 2015

Abstract

Statistical approaches to chemistry, under the umbrella of cheminformatics, are now widespread - in particular as a part of quantitative activity structure relationship and quantitative property structure relationship studies on candidate pharmaceutical studies. Using such approaches on legacy data has widely been termed "taking a big data approach", and finds ready application in cohort medicinal studies and psychological studies. Crystallography is a field ripe for these approaches, owing in no small part to its history as a field which, by necessity, adopted digital technologies relatively early on as a part of X-ray crystallographic techniques.

A discussion of the historical background of crystallography, crystallographic engineering and of the pertinent areas of cheminformatics, which includes programming, databases, file formats, and statistics is given as background to the presented research.

Presented here are a series of applications of Big Data techniques within the field of crystallography.

Firstly, a naïve attempt at descriptor selection was attempted using a family of sulphonamide crystal structures and glycine crystal structures. This proved to be unsuccessful owing to the very large number of available descriptors and the very small number of true glycine polymorphs used in the experiment.

Secondly, an attempt to combine machine learning model building with feature selection was made using co-crystal structures obtained from the Cambridge Structural Database, using partition modelling. This method established sensible sets of descriptors which would act as strong predictors for the formation of co-crystals, however, validation of the models by using them to make predictions demonstrated the poor predictive power of the models, and let to the uncovering of a number of weaknesses therein.

Thirdly, a homologous series of fluorobenzeneanilides were used as a test bed for a novel, invariant topological descriptor. The descriptor itself is based from graph theoretical techniques, and is derived from the patterns of close-contacts within the crystal structure. Fluorobenzeneanilides present an interesting case in this context, because of the historical understanding that fluorine is rarely known to be a component in a hydrogen bonding system. Regardless, the descriptor correlates with the melting point of the fluorobenzeneanilides, with one exception. The reasons for this exception are explored.

In addition, a comparison of categorisations of the crystal structure using more traditional "by-eye" techniques, and groupings of compounds by shared values of the invariant

descriptor were undertaken. It is demonstrated that the novel descriptor does not simply act a proxy for the arrangement of the molecules in the crystal lattice- intuitively similar structures have different values for the descriptor while very different structures can have similar values. This is evidence that the general trend of exploring intermolecular contacts in isolation from other influences over lattice formation. The correlation of the descriptor with melting point in this context suggests that the properties of crystalline material are not only products of their lattice structure.

Also presented as part of all of the case studies is an illustration of some weaknesses of the methodology, and a discussion of how these difficulties can be overcome, both by individual scientists and by necessary alterations to the collective approach to recording crystallographic experiments.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Philip Adler (Student No. 21989575), Declare that the thesis entitled "Statistical Descriptions of Crystalline Compounds" and the work presented in it are my own and have been generated by me as the result of my own original research. I confirm that:

- This work was done wholly while in candidature for a research degree at the University of Southampton.
- Where I have consulted the published work of others this is always clearly attributed.
- Where I have quoted from the work of others, the published source is always given.
- With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- None of this work has been published before submission, except where expressly noted as such.

Acknowledgements

I could write a book as long as War and Peace if I thanked everyone along the way who made this thesis possible. I shall, therefore, stick to the immediate co-conspirators.

First and foremost I should like to thank my Supervisor Simon Coles for securing me the opportunity to pursue my passion in this piece of research, as well as supporting me both Academically and Pastorally throughout, in a fashion that should be exemplar to supervisors everywhere. Furthermore, my Co-Supervisor Dave Woods has been instrumental in my instruction of the rudiments of Statistics and has proven to be invaluable in this research. Both have gone above and beyond the call of duty in terms of their roles as supervisors, in terms of helping with this thesis, and continue to be excellent sources of scientific advice and experience.

If as Müller et al. claim, "The training of a crystallographer is much like the training of a Jedi Knight",[1] then the Obi-Wan to my Luke would assuredly be Graham Tizzard, who has my utmost gratitude in helping me take my first steps into crystallography. By that same token, my heartfelt thanks go to the other members of the National Crystallography Service - namely Mateusz Pitak, Peter Horton and Claire Wilson. There would be no substitute for their expertise and experience as guidance during this Ph. D.

I am indebted to Terry Threlfall who has acted as a voice of timeless experience, and who provided me with the majority of the chemical samples which have sustained not only this Ph. D. work, but also the third year undergraduate project on which it is based. He also assisted in no small way with the proofreading and critique of this thesis.

No small thanks are owed to Sarah Milsted, who helped me navigate the many pitfalls of administration and often navigated them on my behalf. My thanks also go to her for (re)introducing me to the welcome distraction of the Scouting Association to provide me with some perspective when it was sorely needed.

In a similar vein, Sally Dady has been worth her weight in gold in terms of administrating my Postgraduate demonstration duties, among a host of other little odds and ends which make working in the Chemistry Academic Unit practicable.

During this work Andrew Milsted provided much assistance in keeping various computer systems running, fixing forgotten passwords, setting up servers and generally putting up with me breaking things with my various computational experiments.

Eleanor Dodd, an undergraduate student whose work on her third year undergraduate project has supplemented my own, was an invaluable help, and is the kind of project students that Postgraduates dream of having along for the ride. I wish her every success in her future endeavours.

I thank my fellow Ph. D. students in the group; Izzy Kirby, Lucy Mapp and Lisa Blair, for their ongoing companionship, support, proofreading and keeping me from missing a meeting in spite of my subconscious best efforts. Also (now Dr.) Will Peveler, who has been very assistive in stylistic advice for the writing of this document.

I thank the NCS and CCDC both for funding this work, and furthermore for providing invaluable support, insight and resources (which are noted appropriately in this document) throughout the course of this Ph. D.

I should like to note the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work.

Thanks also to Gemma Beale for a lifetime of help with mathematics and for help proofreading this thesis.

For a heroic 11th-hour complete proofreading readthrough, my sincere thanks to my little Sister Jess. CNN!

I should like to note the use of the Diamond Light Source (in partnership with the NCS) for the collection of Synchrotron source X-ray Diffraction Data which was used in this work. Samples where this applies are noted at the appropriate points in this document.

This thesis is Dedicated to the Memories of Irene and Felix Adler, and also Jean and Lloyd Jones, three of whom were laid to rest during the duration of this Ph. D.

Definitions

API Application Programming Interface, not to be confused with the chemistry abbreviation for Active Pharmaceutical Ingredient

CART Classification and Regression Tree

CCDC Cambridge Crystallographic Data Centre

cif *or* .cif crystallographic information framework

cml *or* .cml Chemical Markup Language file

COD Crystallography Open Database

Copyleft A kind of software license, designed to force developers who create derivative works to release their source code for free use

CSD The Cambridge Structural Database

DFT Density Functional Theory

esd *or* e.s.d Estimated standard deviation

GooF Goodness-of-fit as defined in Section 2.3.2

GPL The GNU Public License, a copyleft license used for open source software

LSAM Long-range Synthon Aufbau Module

MOF Metal Organic Framework

mol *or* .mol Molecule file

NCS National Crystallography Service

PDB The Protein Data Bank

R R-factor as defined in Section 2.3.2

refcode The reference codes used by the CSD to identify each crystal structure

STAR Self-defining Text Archive and Retrieval format

string A term for an ordered string of characters used in computer programming

SVM Support Vector Machine

wR Weighted R-factor as defined in Section 2.3.2

Part I

Introductory Material

Chapter 1

General Background

1.1 Overview

The purpose of this section is to introduce the rudiments of the work that has been performed to a sufficiency to explain the initial motivations behind the work, and to lay the ground for later, more in depth chapters.

1.2 Crystallography

Crystallography, in the broadest sense, is the study of crystalline materials. Whilst the term has become synonymous with X-ray Crystallography, formalised observations of crystalline matter date back as far as the middle ages, although a true scientific analysis was lacking until the late 1600s, when Nils Stensen (a.k.a. Nicolaus Steno) demonstrated the First Law of crystallography and ascertained that crystals grow by the progressive acquisition of minute particles.[5]

It was not until the 20th Century however, that the formal analysis of X-rays by Wilhelm Röntgen in 1895[6] permitted the diffraction of X-rays by Max Laue in 1912.[8] Many of the hypotheses of prior scientists received formal validation, with the first successful crystal structure determination[i] by William Lawrence Bragg and William Henry Bragg.[7] Additionally, these experiments confirmed that crystalline matter was constructed of an indefinitely repeating lattice of atoms (as the constituents of molecules or otherwise).

Even with such a capacity to study the three-dimensional arrangement of these atomic lattices, the laws which govern those arrangements as a predictable outcome of a given molecular structure, in particular with respect to organic compounds, remain elusive[9,14,17,18]

[i] Zinc blende, pre-dating the better known work on sodium chloride by about a year.[7]

(though some successes are noted[19–23]). In the 1960s a thesis produced by Michael Hursthouse[ii] contained 4 crystal structures[24] and was typical of the time. By comparison an undergraduate project at Southampton University completed in 2010 contained 14 crystal structures.[25] The increase in the ease and speed of crystal structure determinations is in no small way thanks to the advent of high speed computers and their relative availability.[26] The same ease with which crystal structures can now be determined lends itself in particular to the analysis of bodies of related crystal structures, in order to seek patterns and rules which govern the behaviour of aggregated molecules within the crystal lattice.

1.3 Crystal Engineering

Crystal Engineering stems from the idea that one can create crystal structures with properties known *a priori*, and designed into the system intentionally.[9] Clearly this requires some knowledge of the directing effects of molecular structure upon crystal packing, and this has given rise to the idea of supramolecular synthons;[27] chemical moieties which give predictable crystal structure outcomes. One of the clearest examples of this is the interaction of carboxylic acid groups. The benzylic acid depicted in Figure 1.1 from a paper by Bruno and Randaccio[28] demonstrates this commonly utilised supramolecular synthon. Hydrogen bonding motifs like these are popular as supramolecular synthons owing to their relatively strong directing influence,[10–16] although they do not lead automatically to predictable outcomes.[9] Such synthons are of course, subject to a degree of directional complementarity as well as their nature.[153,154]

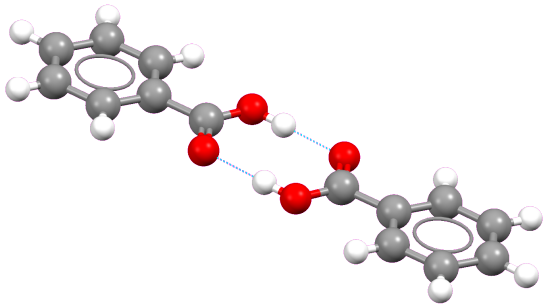


Figure 1.1: The carboxylic acid "supramolecular synthon". The dimer is mediated by two antiparallel hydrogen bonds.

Such an approach generally applies to crystal engineering using solely small organic compounds, and other techniques have found (arguably greater) successes, in particular

[ii]The Author's Academic Grandfather

metallurgy (where the ability to model the materials as collections of hard spheres simplifies the problem greatly[156]), metal organic framework (MOF) design and in the design of coordination polymers.[157,158] These are nevertheless modes of engineering crystal structure. Similarly it can be argued that ab-initio crystal structure prediction techniques and molecular modelling are the reverse-engineering of crystals, and these are discussed in Chapter 2.4.

1.4 Compound Libraries and Crystallographic Databases

There is an analogue for the search for supramolecular synthons in the search for useful interactions with biological molecules (i.e. pharmaceutical research). To aid the latter, those with an interest in researching pharmacologically active compounds would build libraries of compounds to be screened.[159–171] To generate comprehensive combinatorial libraries across chemical space would, of course, be financially ruinous. As such, strategies have been developed for the design of such libraries and, armed with the knowledge that chemicals which are structurally similar ordinarily display similar properties,[172] the norm is to design such libraries to be as diverse as possible with respect to certain heuristic limitations.[172–175]

In a similar fashion, the community of scientists interested in crystal engineering is armed with Crystallographic Databases; for instance the CSD,[177] the PDB[178] and the COD.[179] The disadvantage faced by such scientists as compared to those working with the aforementioned compound libraries is that the crystallographic databases have not, hitherto, been designed as such, but are the result of exhaustive literature surveys and the harvesting of new data from the literature. Even if they were designed, it is difficult to be as readily assured that similar molecules will exist in similar crystal forms- given the phenomenon of polymorphism, which can cause a single compound to exist in many crystal forms.[180] In spite of this, much work has been done in mining the CSD and similar databases for useful information, both in the form of manual surveys and those using more automated methods.[181–184]

1.5 Structural Systematics

In spite of the successes in mining the serendipitously constructed crystallographic databases, the discipline of Structural Systematics exists as a powerful tool. Structural systematics entails the construction of libraries of closely related compounds in order to establish information about a particular kind of interaction and its impact upon the tertiary structure

of a compound. In some senses, the notion pre-dates the existence of such databases as the CSD, the term having been used as early as the 1950s to describe the systematic differences in series of inorganic compounds.[185] Progressing through the 1970s, studies of preselected and synthesised molecules started to be seen in the crystallographic literature,[186] though such studies were not necessarily described with the term "structural systematics". Large scale studies in Crystallography have only relatively recently become manageable with the advent of cheap computer automation and data processing, improved detectors, and increasingly bright laboratory X-ray sources.[26] More recent studies by Hursthouse and Gelbrich in the early 2000s sought to automate this process, seeking out structural motifs and building graphs of similarity within families of compounds.[187]

One could of course make the statement that structural systematics goes back to the very heart of chemical and scientific thinking- to seek for patterns amongst related data in order to establish models for reality- found in Pauling's discussions on ionic radii[188] and Mendeleev's periodic table.[189] It could therefore be said that structural Systematics is the true chemistry of crystalline matter.

1.6 Cheminformatics

As data sets become larger and larger, so then it becomes harder and harder to manually find patterns within the information. In particular, with complex problems with many contributing factors using a simple mechanism. Cheminformatics is the name given to a field that represents a set of tools which coalesce to deal with such problems. Originally the tools stem from separate disciplines; Statistics[iii], Databases, Computer Science and of course, Chemistry. The quantitative tools used in cheminformatics also provide evidence of the importance of a pattern and the likelihood of such occurring by chance.

Cheminformatics has reached its prominence in the search for quantitative structure-activity relationships (QSAR).[190] Such studies are used by the pharmaceutical and agro-chemical industries to establish relationships between structural features of compounds and the level of the intended (and unintended) effect that they have in the target organism. There have, over the time in which studies have been taking place, developed many ways in which to describe the structure of a molecule quantitatively. Often these quantities (herein *descriptors*) will describe different aspects of the compound (electron density, number of atoms, the percentage of the molecule made up of specific atom types,

[iii] Although it is interesting to note that some standard tools for statistics were developed for the purposes of beer brewing

dipole and quadrupole moments, to exemplify a few.[iv]). As such, it is often necessary to try to relate more than one of these to the measured outcome(s) (a *response descriptor*, for instance, duration of effect of a drug). Once this relationship has been established, one has a model of the system, and can begin to start making predictions based upon it. Often one will draw upon databases for known measurements, or programs can be written to calculate descriptors for molecules for use in such models.

For the purposes of the study at hand, descriptor space was found, after an exhaustive and time consuming literature review (detailed more fully in Section 3), to fall into the following broad categories (which are not mutually exclusive).

- Molecular Descriptors: Descriptors of the molecules which make up the crystal structure
- Physical Descriptors: Response descriptors describing physical properties of the crystalline matter
- Topological Descriptors: Response descriptors describing the links between molecules in a crystal structure (e.g. Hydrogen Bonds)
- Spacial Descriptors: Response descriptor indicating the arrangement of atoms and molecules in three-dimensional space within the crystal structure

1.7 Starting out: A Naïve Hypothesis

In a paper published by Zhu et al. it was suggested that, in the case of fluorobenzanilide structures, the electrostatic effects of fluorine were enough to have a directing effect on the overall crystal structure.[46] The paper focused on co-crystals (simplistically: a crystal structure containing 2 compounds) of two pentafluorobenzanilide compounds. Each molecule had one of its phenyl rings completely substituted with fluorine, whilst the other ring remained unsubstituted. Each component of the co-crystal had a different ring substituted. The co-crystal formed a stacking motif such that the substituted phenyl ring for one compound would be positioned above the unsubstituted ring of the other component. This, as per crystalline matter, repeated indefinitely in an alternating pattern.

The naïve hypothesis that underpins the work being presented is that such fluorine directing effects could be applied to the entire homologous series of fluorinated compounds, illustrated by the Markush structure shown in Figure 1.3. The hypothesised outcome is

[iv] In his excellent tome, *Molecular descriptors for chemoinformatics*, Roberto Todeschini gives a close to exhaustive list of molecular descriptors.[191]

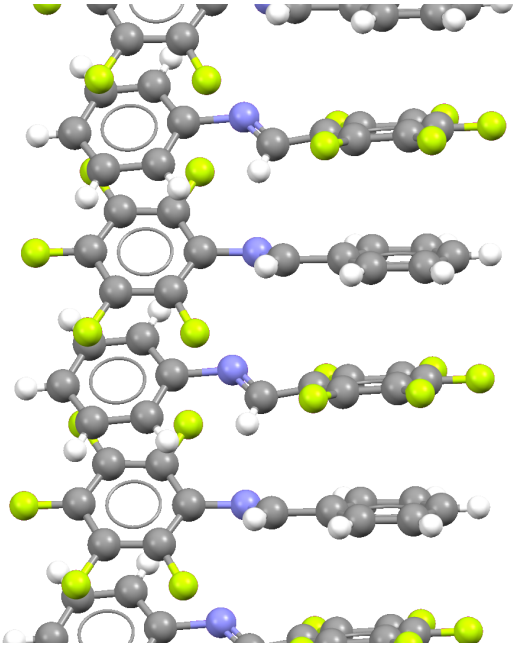


Figure 1.2: The crystal structure presented by Shizheng Zhu et al. Note the stacking motif alternates perfluorophenyl and phenyl ring positions.

that where the fluorine substitutions are complementary (as shown in Figure 1.4) the result will be that the stacking motif will be seen. Where the overlap is deliberately designed to frustrate that stacking motif (Figure 1.5) the motif will not be present.

It is anticipated that, in particular for the non-complementary compounds, that the packing motifs will be necessarily complex, and that the factors which underpin them will be equally intricate. In order to provide a more complete understanding of such outcomes, tools from cheminformatics will for the first time be employed in understanding a particular family of crystalline compounds, with the hope of building a predictive model for some characteristics of the crystal structures and, more broadly, developing a procedure and tools for the application of cheminformatics to this sphere of chemistry. What follows in the rest of this introductory material is a full explanation of the relevant chemical and statistical understanding required to address such a problem.

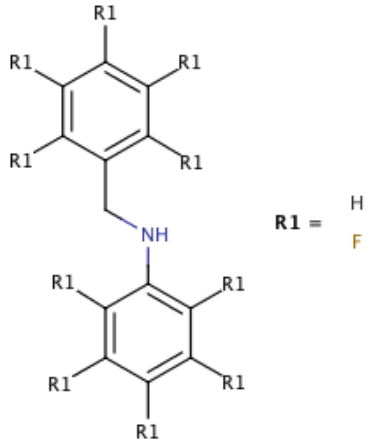


Figure 1.3: The Markush structure defining the homologous series of fluorobenzanilides under examination.

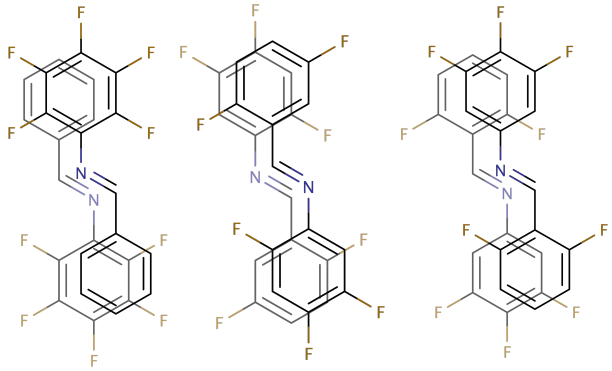


Figure 1.4: Three hypothesised examples of 'complementary overlap' structures.

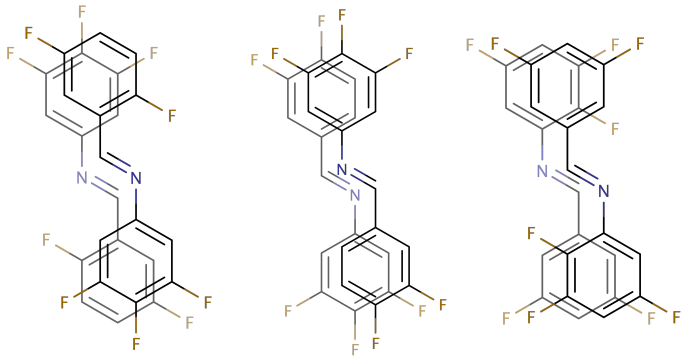


Figure 1.5: Three examples of fluorobenzanilides which, when stacked, display non-complementary overlap

Chapter 2

Crystallography

2.1 Crystalline Matter

2.1.1 Powders, Single Crystals and Twinned Crystals

Crystalline matter is a three-dimensional extensively[i] repeating lattice of a given substance.[7,8]

In general, X-ray crystallography is considered either from the perspective of dealing with powders or single crystals. Powders are simply collections of a very large number of very small crystals. Because these crystals are oriented randomly within the powder, the X-ray diffraction patterns they produce are a set of radial bands about a centre point. Single crystals produce a series of spots as an X-ray diffraction pattern.[192]

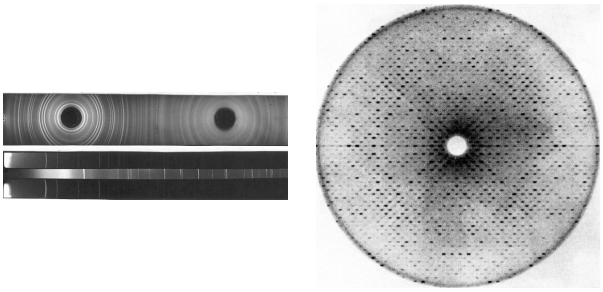


Figure 2.1: A powder diffraction pattern and a single crystal diffraction pattern. Note that the powder diffraction pattern forms rings because of the effectively random orientation of the crystals and their unit cells.[193,194]

Whilst single crystals are, in principle, crystals formed of one infinitely repeating lattice, the truth is that a single crystal contains many subunits which all share the same

[i] although in crystallographic models, it is treated as being infinite

approximate orientation. The extent to which there are different orientations of these subunits is called the mosaicity of the crystal.[192]

Twinning is a complication in crystal structure characterisation. Twinning occurs either when two crystals fuse or grow from the same nucleus in two different directions or otherwise undergo some change of state owing to physical conditions. The result of this is two crystals with a mutually fixed orientation which appear to be a single crystal. The unit cell structure itself does not necessarily change, but the data that is collected requires some disambiguation in order to account for the two different orientations of the crystals when the data is being collected.[192] There are methods for dealing with the data these situations produce,[195,196] though anecdotally, success rates are lower for dealing with such crystals.

2.1.2 The Unit Cell

Defining an entire crystalline entity at the molecular level is impractical (unless one possesses an infinite amount of computational power); thus such substances are described by a subsection of an infinite repeating lattice which, by use of symmetry operators, can be used to reconstruct an arbitrarily large segment of the crystal structure. This subsection which is the smallest repeatable unit which can be used to reconstruct a crystal structure in this way defines the crystal structure and is designated the *unit cell* of a crystal lattice.

The dimensions of this unit cell are expressed in terms of three axes: a, b, and c, and three angles α, β and γ, as shown in Figure 2.2. The unit cell repeats infinitely along the axes. These dimensions and angles are characteristic of a given species - it is a known method to identify an unknown sample by searching databases of unit cell dimensions. It should be noted, however, that the reverse is not true - a given compound does not necessarily always form into the same unit cell. Cases where multiple unit cells are found for a single species are called *polymorphs*.

In the illustrated example, the corners of the unit cell are termed lattice points. A lattice point is merely a relative point in space to the other lattice points of the unit cell, and the absolute position is arbitrary[ii]. Two primary conventions exist for the positioning of lattice points when presenting unit cells. In the majority of inorganic chemistry (with the notable exceptions of organometallic chemistry and complex chemistry among others), lattice points tend to be placed such that they coincide with atomic entities. For most other chemistry molecular moieties tend to be situated such that their centre of mass lies within the unit cell bounds. To illustrate this, observe the two dimensional example

[ii]This isn't strictly true in all cases; as shall be seen in later chapters, symmetry is used to further condense unit cell descriptions, and this introduces constraints.

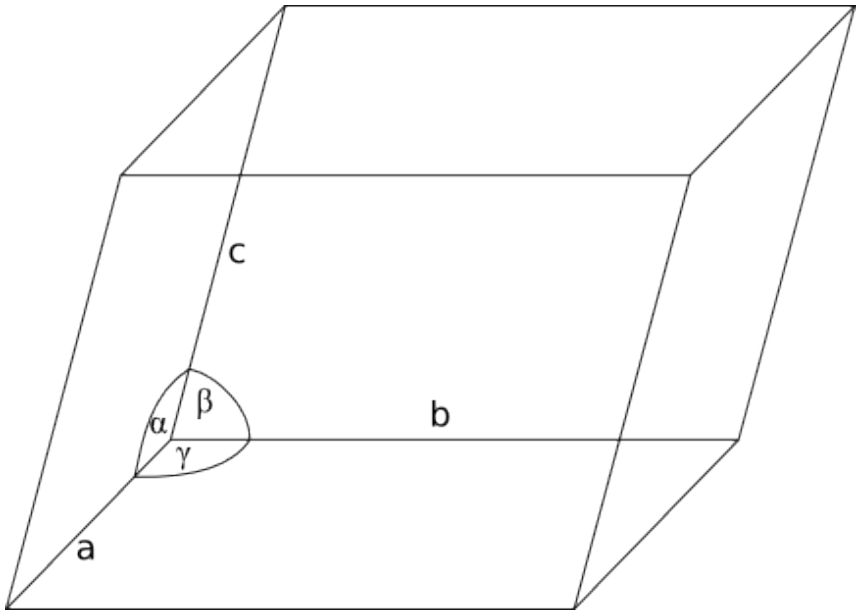


Figure 2.2: A triclinic unit cell. The axes and angles are labeled with the nomenclature which is common in crystallography.

in Figure 2.3. Both unit cells would be considered entirely equivalent in spite of their different positions.

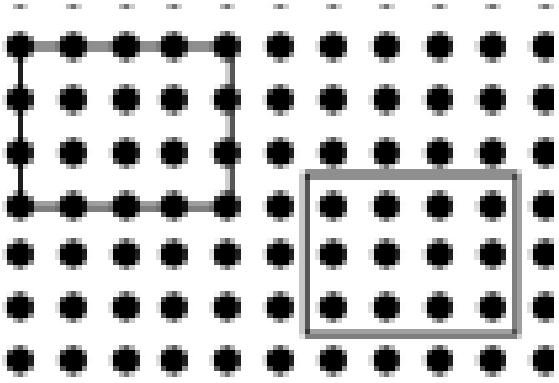


Figure 2.3: Two equivalent unit cells; their independence holds regardless of the location of the contents relative to the unit cells

2.1.3 A Hierarchy of Crystal Structures

Overview

The unit cell description of a crystalline compound can be further minimised by using *point symmetry operations* in conjunction with the unit cell translations (which in this context are referred to as *space symmetry operations*). The combination of these kinds of symmetry operators also yield additional symmetry operations. These are useful during the characterisation process to minimise the amount of data which needs collecting, but also have the useful effect of providing a mathematical description of the symmetry relationships between moieties in a crystal structure.

Notation

Symmetry operations have two main forms of notation in crystallography, throughout this report the notation used is the *Herman-Mauguin* notation.

Point Symmetry Operations

Point symmetry operators (that is, a transformation according to a predetermined set of rules), when applied to an isolated object (for instance, a unit cell), must leave at least one point within the object unmoved.

Point symmetry operations include rotation, mirror reflection and inversion through a point. Rotation operations are described as n-fold rotation axes. Such a definition means that a rotation by $\frac{360^\circ}{n}$ leaves the object or structure appearing identical to the state prior to the operation. Similarly, n-fold rotatory-inversion axes represent a rotation of $\frac{360^\circ}{n}$, immediately followed by an inversion. A 1-fold rotatory-inversion is simply identical to a plain inversion. Normal rotations are normally notated simply by their numeric value (e.g. 1), whilst rotatory inversions are notated with a bar (e.g. $\bar{1}$).

A mirror plane is intuitively defined - it creates a mirror image of the structure. Rigorously, a mirror plane is identical to the symmetry operation $\bar{2}$.

Additional Operations

As mentioned, space group symmetry adds translation operations to the operators available. Their combination with point group symmetry yields additional operations. One such are n-fold screw axes, which are a combination of translation along an axis at a distance of $1/r$ where r/x is an integer fraction of the identity period x along the axis, followed by an n-rotation as defined previously. These are denoted as n_r; for instance 2_1.

The other operation is that of glide planes, which are a combination of a translation and a mirror plane. The translation must be of distance $x/2$ where x is the period along the axis. Glide planes are named after the axis along which they are found. Since in crystallography the three axes of a unit cell are described with the letters a, b and c, this results in the nomenclature c-glide, b-glide and a-glide.

Crystal Families and Crystal Systems

The highest level of the hierarchy of crystal structures is that of crystal families - of which there are six. These are defined by the angles α, β and γ, and the relative sizes of the dimensions a, b and c, as tabulated in Tables 2.1 and 2.2.

As shown, crystal systems are an extension of the crystal families, separating the hexagonal crystal family into the trigonal and hexagonal crystal systems according to the minimum required symmetries, also shown in Table 2.1.

Crystal Family	Crystal System	Symmetry Req.
Triclinic		$\bar{1}$
Monoclinic		$2/m$
Orthorhombic		mmm
Tetragonal		$4/mmm$
Hexagonal	Trigonal	$\bar{3}m$
	Hexagonal	$6/mmm$
Cubic		$m\bar{3}m$

Table 2.1: The crystal systems and families

Bravais Lattices

Lattices, as mentioned, are often represented by their lattice points. When viewed in this context, crystalline compounds are divided up into lattice systems, which are subtly different to crystal systems. Lattice systems, rather than being defined by their symmetry operations, are defined by the dimensions of the unit cell. It should be noted that this results in some relationship between crystal systems and lattice systems, but some differences can be seen by examining Table 2.1 and Table 2.2.

These are known as *primitive lattices*. Other lattice types are observed by the existence of additional lattice points within the lattice under examination. These additional lattice points have an effect upon the observed symmetry elements of a given lattice, which as shall be shown later, is what makes them observable. There are 14 *Bravais Lattices*, which are distributed among the 6 lattice systems as follows:

Lattice	Dimensions	Crystal System(s)
Triclinic	$\alpha, \beta, \gamma \neq 90°$	Triclinic
Monoclinic	$\alpha, \gamma = 90°,\ \beta \neq 90°$	Monoclinic
Orthorhombic	$\alpha, \gamma = 90°,\ \beta \neq 90°$	Orthorhombic
Tetragonal	$a \neq c$	Tetragonal
Rhombohedral	$a = b = c,\ \alpha, \beta, \gamma \neq 90°$	Trigonal
Hexagonal	$b = a,\ \alpha, \beta = 90°$	Hexagonal
Cubic	$a = b = c, \alpha, \beta, \gamma = 90°$	Cubic

Table 2.2: The Primitive Lattices

- **Triclinic** is completely synonymous with the same Bravais lattice.
- The **Monoclinic** crystal family can be divided into *primitive* (no additional lattice point) and *base-centred*, where an additional lattice point exists at the base of the hexahedron.
- The **Tetragonal** family can be *primitive* or *body-centred*, wherein an additional lattice point exists at the centre of the unit cell.
- The **Hexagonal** family is divided into the *rhombohedral* and *hexagonal* lattices, the latter of which may only be *primitive*. Rhombohedral cells may be *primitive* or *rhombohedrally centered*.
- The **Cubic** family has members which are *primitive*, *body-centred* (in the same manner as Tetragonal) or *face-centred*, wherein each face has an additional lattice point at it's centre.
- **Orthorhombic** family members can be *primitive*, *body-centred*, *face-centred* or *base-centred*.

Crystal Classes

The crystal systems are more commonly divided by crystallographers in to the 32 crystal classes, which are more commonly called crystal point groups. As their more commonly used name implies, these originate from their point group symmetry operations, and are labelled according to the primary symmetry elements that they possess. It is this that leads to the term *point group*, as the crystal class is based on the group of point symmetry operations that it possesses.

The point groups are not listed here, but can be found in the IUCr reference "International Tables for Crystallography".[197]

Space Groups

As mentioned in 2.1.3, combining point group and space group symmetry operations gives rise to additional operations. Their inclusion into the model of crystal lattices thus far results in the 230 space groups. For non-chiral species there are only 219, as some space groups become equivalent owing to the additional symmetry innate to the species.

2.2 Principles of X-ray Crystallography

2.2.1 The Diffraction Pattern and the Unit Cell

As stated, crystalline compounds can be considered lattices of atoms. Taking this as a supposition, Laue generated the first diffraction pattern, drawing from an analogy with Young's diffraction slits. He was also able to derive a law that linked the diffraction pattern with the distances between the planes of the repeated molecular lattice; the dimensions of the unit cell.[5]

Bragg later developed a more intuitive method of calculating interplanar spacing in his eponymous equation (equation number 2.1).

$$n\lambda = 2d\sin\theta \tag{2.1}$$

Bragg's model surmises that as X-rays pass into the crystal they are reflected by the repeated, infinite planes of atoms therein. There are, it was deduced, only certain angles at which the reflected rays would be in phase and hence, constructively interfere. Knowing the wavelength of the incident radiation, the angle of incidence, and the angle of reflection (from the diffraction pattern), one can then calculate the interplanar spacing because each 'reflection' (as they have come to be known) must have a reflected angle that satisfies the Bragg equation.[198] It should be noted that this only gives the measurements of the unit cell, however, and not the position of the atoms within it.

In these diffraction patterns, additional information can be derived by the existence of *systematic absences* in the diffraction pattern. These are points in the diffraction pattern which, given a primitive unit cell of specified dimensions, would appear to be missing some fraction of the reflections in the diffraction pattern. Some of these absences are in fact caused by non-primitive Bravais lattices, whilst others are caused by the presence of symmetry groups in the crystal lattice involving some translation element.[199]

2.2.2 Unit Cell Contents

The above analyses, whilst it allows the ready deduction of the unit cell, does not allow the identification of the contents thereof. The contents of a unit cell are deduced from the locations of electron density in the unit cell. This is calculated using the electron density equation shown in Equation 2.2.[200]

$$\rho(xyz) = \frac{1}{V} \sum_{hkl} F(hkl) e^{-2\pi i(hx+ky+lz)} \tag{2.2}$$

Where:

- V is the volume of the unit cell.
- The sum is taken for all values of $F(hkl)$
- $F(hkl)$ are the *structure factors* of the crystal.
- x, y and z are the relative coordinates at which the electron density will be found.
- $\rho(xyz)$ is the electron density function at the point at the xyz coordinates.

The structure factors $F(hkl)$ represent the information contained in the diffracted beams, and are a complex number of the amplitude and phase of the 'reflected' beam in the form $|F(hkl)|e^{(i\phi(hkl))}$.

In point of fact this is a reflection of the fact that the beams are not reflected from the lattice plains at all, but are in fact caused by the scattering influence of electrons in the crystal structure. When the incident X-ray beam interacts with the crystal the electrons oscillate in the beam and then re-emit the X-rays in random directions. This propagates throughout the lattice, and once the waves exit the crystal structure, along with interference patterns, results in the diffraction pattern that is collected. Regardless of this, the term *reflections* has remained in common use to refer to each spot in a diffraction pattern.

This relationship between the scattering of individual atoms in the crystal structure and the structure factor can be expressed by the structure factor equation (Equation 2.3).[200]

$$F(hkl) = \sum_{j=1}^{N} f_j e^{2\pi i(hx_j+ky_x+lz_j)} \tag{2.3}$$

Where:

- $F(hkl)$ is as defined previously.
- f_j is the scattering factor of the jth atom.
- N is the number of atoms in the unit cell.
- x_j, y_j and z_j are the fractional coordinates of the atoms in the cell.

The intensity of a given reflection in a diffraction pattern, (I_{hkl}) is ultimately proportional to the square of the modulus of the structure factor (expression 2.4).

$$I_{hkl} \propto |F(hkl)|^2 \tag{2.4}$$

And it is this that permits the creation of a real-space model of the electron density from the diffraction data, via the Fourier transform in Equation 2.2.

2.2.3 Crystal Structure Solution

Whilst we are able to calculate the modulus of the structure factor by consequence of expression 2.4, in collecting the X-ray diffraction data we only collect the amplitude information of the structure factor, not the phase information (ϕ). This is known as the phase problem.

The three main methods of resolving this issue in small molecule crystallography are the *Patterson*, the direct *Direct* and *Dual Space* methods. The Patterson technique tends to be employed when there exist a small number of heavy atoms present in the sample, whilst Direct methods tend to be better suited towards species which are constructed primarily of atoms with a similar mass. Both are generally performed by computer programs in the current mode of crystallographic practice.

The Patterson technique simply replaces the structure factors in Equation 2.2 with their complex conjugate. This allows the production of a 'Patterson map' which describes vectors between atoms in the unit cell (that is, their relative positions with respect to each other). The largest of the peaks in the Patterson map will in most cases be due to the vectors between whichever heavy atoms exist within the unit cell. The basic principle is to deduce locations in the unit cell which explain all of the large peaks in the Patterson map.[201]

Direct methods, by contrast, use mathematical relationships between the amplitudes and phases of the structure, as well as some 'prior knowledge' about the nature of the crystal structure (for instance; the fact that one cannot have a negative value for electron density, or the nature of the atoms expected in the crystal structure) from this prior

knowledge of the crystal structure, one can form mathematical constraints on the probable electron densities of said crystal structure. This is the more commonly used method, because it is readily amenable to automation (indeed, programs are freely available that are able to perform this kind of analysis). In the most ideal cases, a model for the structure can be derived from the first electron density map that is produced by this method - although this is not always the case. The method is limited to a few hundred atoms, since the statistical relationships it depends on break down for larger structures.[202]

Direct Methods are solved primarily by placing constraints on information in Fourier space, whilst Patterson methods place constraints primarily (and somewhat confusingly) in direct-space.[203] Hence, methods which operate in both spaces are known as Dual-Space methods. Probably the most commonly used in crystallography are the Shake-and-Bake methods (as applied in the crystal structure solution program ShelXD[204]) and Charge-Flipping algorithms (as applied in the crystal structure solution program SuperFlip[207]).

The basic approach to charge flipping methods is to take a random assignment of electron density. This random electron density is then "flipped" subject to an arbitrary threshold value:

$$\tau_i = \begin{cases} \rho_i & \text{if } \rho \geq \delta \\ -\rho_i & \text{if } \rho < \delta \end{cases} \tag{2.5}$$

Where:

- ρ_i is the electron density at a given position i
- δ is the arbitrary threshold value, and can be set according to a variety of factors according to the situation
- τ_i is the new electron density at i

Once this has taken place the trial electron density τ is subjected to the constraints in both direct and Fourier space. The electron density resulting from this is then fed back into the flipping step and so on until the electron density converges on a solution.[203]

Shake-and-Bake methods are superficially similar to the Charge Flipping algorithm, in that a random charge density is used as an initial trial solution which is then subjected to constraints in both the real and Fourier transformed space. However, on each iteration, many solutions are obtained as the results of an adjustment of the phase parameters of the previous solution via an arbitrary function (shaking)[iii]. The phases are then passed

[iii]The term arbitrary means that in theory any function can be used, but of course some functions are more useful than others.

into a minimisation function, the result of which should be at its lowest with the correct phases. When the new phases result in a lower value from this function, the solution with the lowest value is passed through the constraints in real and direct space, and then is passed back through the algorithm. This is repeated until the result of the minimisation function reaches a convergence.[208]

2.2.4 Crystal Structure Refinement

In any case, what results from a crystal structure solution is an electron density map, and for some programs a partial model. The electron density map can be used visually to find regions where atoms may be located, and these atoms are added to the model. Information from the model about the location and nature of the atoms can be added to the Fourier transformations which grants additional phase information, which in turn will yield more information in the next electron density map, and this iterative cycle continues until one has a completed crystal structure. The cycle may also refine co-ordinates of atoms in the model structure.

These cycles are a least-squares refinement. The full mathematical details of least squares solutions are detailed fully in the section on statistics (Section 3.2.4. For present purposes, the refinement algorithms seek to maximise the correlation between the calculated structure factors for the model crystal structure and those from the collected x-ray data; this is, in general measured by minimising the R-factor or correlation coefficient between the two. This is defined explicitly in Section 2.3.2.

It should be noted that the results gained from these analyses are an average, both through space and time. The data is collected from an entire crystal, not one unit cell (the diffraction from such a small source would not generate sufficient data), and so the structures which are gathered represent an average throughout the sample. In addition, the results are an average through time. Often, each diffraction pattern can take many seconds to collect, during which time atoms will move through thermal motion, and these will appear both in models and electron density maps as smears, as 'on average' they will have been spread through space.

Eventually cycles of crystal structure refinement will reach a convergence, wherein changes to the crystal structure model will either not change the electron density map, or will make the model worse; this is measured in a variety of ways, as discussed in the next section.

There are cases where conformations or bond distances may not be sensible in a crystallographic model. Many programs for crystal structure refinement permit the inclusion

of constraints and restraints as a way of fixing this information in the crystal structure model.[205,206]

A constraint forces a piece of information about the crystal structure to be expressed absolutely in terms of a constant or another piece of information about the crystal structure. For instance, the following constraints are available to the program SHELX:

- AFIX: Fixes a given number of atoms to relative positions based on the location of an anchor or 'pivot' atom
- EXYZ: Fixes the location of a number of atoms to be the same
- EADP: Fixes the atomic displacement parameter (thermal motion) of a number of atoms to be the same

A restraint, on the other hand, confines a measurement about the crystal structure to a value with a probability distribution. Some examples (again from SHELX) are:

- DFIX: restrains the distance between two atoms
- SADI: restrains two 3-atom angles to be approximately the same
- FLAT: restrains a set of atoms to share the same plane

Restraints are included as extra data in models for the purposes of validation.

Such constraints and restraints are an important part of refining disordered crystal structures. The most common instances in the crystallography of small organic molecules (especially this thesis) are in the placement of hydrogen atoms as a part of a crystallographic model and in the handling of aromatic rings.

Aromatic rings are a common construct in small organic molecules. By definition they are flat, rigid systems which are normally regular polygons; consider the regular hexagonal shape of the phenyl ring, for instance. For a variety of reasons, diffraction data quality may be poor. This may give rise to the appearance (in this example) of six membered rings in which the atoms are not equally distributed, have differing thermal displacement parameters (contradictory to a rigid system) or are not flat. The chemist will normally have some *a priori* information about what substance is under analysis, and would wish to include such information as data in the model. In such a case, DFIX, SADI and FLAT can be used to restrain the distances and angles of the phenyl ring in a manner which corresponds to the information held by the chemist.

The inclusion of hydrogen atoms using constraints is an extremely common practice in organic crystallography. This is because the normal modelling techniques of crystallography assume that atoms are 'points' within a lattice, and that those points are identified by the location of electron density in the crystal structure solution. In the case of covalent bonds with hydrogen atoms, the electron density associated with the electron will be primarily localised in the bond connecting the hydrogen to the other atom. Placing the hydrogen atom at the site of the electron density would, therefore, give rise to an artificially short bond. In the case of shelx, a special constraint, HFIX, is used to place a hydrogen atom appropriately to the other atom. In truth, this is actually a shorthand for a range of AFIX constraints. The distances used for this come from neutron diffraction experiments. This allows the placement of hydrogen atoms at reasonable distances from atoms to which they are covalently bonded whilst permitting the refinement of the overall structure based on the electron density calculated from the data.

2.2.5 Disorder In Crystal Structures

Disorder is a complication in refining a crystal structure. It is an artefact of the averaged nature of the crystal structure. Broadly speaking, disorder falls into two types: thermal disorder and static disorder.

Thermal disorder results from the thermal motion of atoms in the crystal structure. If the range of motion is large enough for a given section of the crystal structure, then it can seem in the electron density maps to appear in two places at once, for instance. There is an additional complication in that the electron density will appear much smaller in each location. This is (continuing the example of being split equally between two locations) because the atoms are only in each location around half the time.

Static disorder appears superficially similar in the data (although the thermal displacement of the atoms will probably be somewhat smaller), but is a result of the averaging through space, rather than through time. In some instances two arrangements of atoms or molecules in a crystal structure will be energetically similar. Thus, randomly throughout the structure, the species will be found in different positions. This manifests in the averaged unit cell as diminished electron density in those different locations in the unit cell.

In some cases, the arrangement of static disorder is not in fact random, but is periodic. Structures where this is the case are called incommensurate structures, and special methods exist for their solution.[203] This will not be discussed here as it is not pertinent to this study.

2.2.6 Z, Z' and Z"

A number of parameters in crystal structures are related to the number of entities contained within the unit cell:

Z is defined in the original .cif specification as:[210]

> "The number of the formula units in the unit cell as specified by _chemical_formula_structural, _chemical_formula_moiety or _chemical_formula_sum".

These three are in turn defined thusly in the IUCr maintained dictionary:[211]

_cell_formula_structural The chemical structure, as reported, giving as much detail using parentheses about the structure as reported

_chemical_formula_moiety The formula with each discreet bonded residue or ion shown as a separate moiety, showing charges where appropriate

_chemical_formula_sum The sum total of atoms, grouped by element in the unit cell

Z' in turn seems to lack a formal literature definition[iv]. A website maintained by Professor John Steed at Durham University states that Z' is strictly defined as[213]

> "the number of formula units in the unit cell divided by the number of independent general positions"

This definition is not exactly supported by the paper which introduces Z", in which Z' is regarded as $Z' = Z/M$, where M is the multiplicity of the general position, i.e. it is the count of whatever unit was used to define Z in the asymmetric unit.[214] This definition is obviously less clear, since there are three definitions for Z which are not strictly equivalent, particularly in cases of formulae containing more than one species (herein *co-crystals*) or where a molecular entity sits upon the site of a symmetry element within a unit cell (a so-called 'special position' as opposed to a 'general position'). So the term already possesses ambiguity without considering corner cases (is a benzylic acid dimer one or two moieties?).

Z" is in turn defined as[214]

> "The number of crystallographically non-equivalent molecules [in the unit cell]"

[iv] an entry could not be found in the IUCr International Tables for Crystallography, and digital literature searching is complicated as Z' is also used as a term in subatomic particle experiments[212]

Thus, for crystalline matter made up of only one species and where the unit cell is equivalent to the asymmetric unit, Z' and Z" are equivalent, but for instance a monohydrate structure with a Z' value of 1 would have a Z" value of 2.

So it is that analysing crystal structures with more than one molecular unit in the asymmetric unit can become somewhat problematic, both in general, but as shall be seen, especially in statistical models.

2.3 Validation of Crystallographic Models

2.3.1 Chemical Sense

The foremost check for validating crystallographic models is that they make chemical sense. Bond lengths must be of a sensible magnitude, and constrained regions should be of a sensible conformation (phenyl rings, for instance, should be flat). As mentioned in Section 2.2.4, constraints and restraints can be applied to preserve chemical sense. Obviously, only atom identities which were a part of the synthetic method at some stage should be included - atoms which cannot be accounted for could be an error or a problem.

2.3.2 R-factor, weighted R-factor, Goodness-of-Fit

The three factors, R-factor (R), the weighted R-factor (wR) and Goodness-of-Fit ($GooF$), are mathematical expressions of agreement of the model with the X-ray data from which that model was derived. All seek to be minimised to convergence during the refinement sequence, except the $GooF$, which should reach a minimum value of 1. A value smaller than this suggests a problem with the model. Often this will mean that the solution as had improper corrections applied for absorption by the sample, or that the wrong space group has been selected for the refinement.[1] These will alter the observed structure factors or artificially inflate the number of independent reflections respectively. The effect of the artificially large number of independent reflections can be seen from inspection of Equation 2.8. The relationship between the observed structure factors and the weighting scheme is nontrivial, but is the mechanism which gives rise to an erroneous goodness of fit in the case of improper absorption corrections.

The factors are defined as follows:[1]

$$R = \frac{\sum ||F_{obs}| - |F_{calc}||}{\sum F_{obs}} \tag{2.6}$$

$$wR = \left[\frac{\sum w(F_{obs}^2 - F_{calc}^2)}{\sum (F_{obs}^2)} \right]^{\frac{1}{2}} \tag{2.7}$$

$$GooF = \left[\frac{\sum w(F_{obs}^2 - F_{calc}^2)^2}{N_R - N_P} \right]^{\frac{1}{2}} \tag{2.8}$$

Where:

- F_{obs} is the observed structure factor from the data
- F_{calc} is the calculated structure factor from the model
- w is a weighting factor derived from the uncertainty of the reflections
- N_R is the number of crystallographically independent reflections
- N_P is the number of refined parameters in the model

It has been known that nonsense solutions can give very low R-factors for some data.[208] Such solutions highlight the issue of using purely numeric methods to validate models without chemical reasoning. This is related to the chance correlation between two datasets that can be seen in statistical methods. In statistical methods, there are other techniques used to protect against such chance correlations, and this is detailed in Chapter 3.

2.3.3 Thermal Displacement Parameters

Thermal displacement parameters are used in crystal structure models to describe the thermal motion mentioned previously, however, groups which are attached to each other should, in most cases, have similar directions of motion - and instances where this is not the case should have a clear rationale as to why not.[1]

2.3.4 Estimated Standard Deviations

Whilst the R-factor describes the extent of agreement between the model and the data, each individual coordinate, bond length and most other measurements in a crystallographic model will be accompanied by an estimated standard deviation (abbreviated e.s.d.). The physical meaning of this value is somewhat subtle (a full treatment is given in a different context in Section 3.2.4), but it measures the spread of possible values for a parameter based on random error.[1] It is often misinterpreted to be an absolute limit for said values.

2.3.5 Data Quality Measures

Whilst not strictly a part of *model* validation, problems with model statistics do not always arise from problems with the model per-se but from issues with the underlying data. In general, three main quality measures for data are used: redundancy, completeness, and resolution. The three are somewhat interelated, completeness describes to what extent all of the unique reflections in the data set have been collected up to a provided resolution threshold. Resolution is the precision to which features can be detected in the crystal structure, and is generally considered to be the highest angle (as per the Bragg equation, crystal structure resolution and angle of diffraction are related) at which a reasonable completeness of data is maintained.

It can be seen from Equation 2.1 the resolution is related to the wavelength of the radiation being used. Copper diffraction sources therefore have a lower possible resolution than Molybdenum sources.

Redundancy refers to the number of times that each unique reflection in a data set has been collected. Abstractly, higher redundancy increases the certainty in the value of the intensity of any given reflection, which increases the certainty in any model based upon that data. In general, small molecule crystallographers aim for a minimum redundancy of 3.

2.4 Crystallographic Databases

There are two main crystallographic databases which store information about small organic molecules (the focus of this work): the Crystallographic Open Database, and the Cambridge Structural Database. The latter of these is the larger of the two,[215,216] and so this is favoured for the presented work. Furthermore, the software which is used in conjunction with the CSD is more full featured,[217,218] making it less difficult for use in a project involving any degree of data searching. Each crystal structure is given a reference code (*refcode*) which uniquely identifies each structure, and these are used to reference the crystal structures throughout this document.

2.5 Intra- and Inter- Molecular Interactions

2.5.1 Preamble

As alluded to in Section 1.3, Crystal Engineering works on the basis that one can know *a priori*, what crystal structure a given species will form, and furthermore that one can

select a species to give a specific crystal structure, or motifs within said structure. To do this requires an understanding of the interactions of all compounds in general. This section discusses the established models for those interactions from the scientific literature. The vast majority are referred to under the general umbrella term van der Waals forces.[219] Because of the subject matter of this thesis, the discussion is limited to forces which are thought to apply in the solid state.

The majority of techniques which self-declare as being crystal engineering eschew a quantitative understanding of these interactions in favour of retro-justification from crystallographic data. There are a few notable exceptions, such as Gavezzotti, who seek to use crystal structures to quantify the interactions observed in crystal structures energetically.[220] As Gavezzotti reminds us - the following segregation of intermolecular forces is not necessarily the most correct understanding,[221] nevertheless, to quote the late George Box: "All models are wrong, some models are useful".[2]

2.5.2 Dispersion Forces

Dispersion Forces, otherwise known as London Forces, were introduced by their namesake, Fritz London, to describe the attraction between molecules which could not be ascribed to other sources such as ionic effects or permanent multipoles.[222,223] It should be noted that dispersion forces are a purely non-classical effect,[224] and so the classical explanation of instantaneous dipoles[155] bears little relevance to the London equation, which approximate the quantum mechanical treatment:[224]

$$U_{disp} \approx -\frac{3U_A U_B}{2(U_A + U_B)} \frac{\bar{\alpha}^A \bar{\alpha}^B}{(4\pi\epsilon_0)^2 R^6} \tag{2.9}$$

Where:

- U_{disp} is the dispersion energy
- U_A is the average excitation energy (the ionisation energy) of species A
- U_B is the average excitation energy (the ionisation energy) of species B
- $\bar{\alpha}^A$ is the polarisability component of species A
- $\bar{\alpha}^B$ is the polarisability component of species B
- ϵ_0 is the permittivity of free space
- Hence, the term $\frac{\bar{\alpha}^A \bar{\alpha}^B}{(4\pi\epsilon_0)}$ defines the product of the polarisabilities of the two species

- R is the distance between the two species

Note that in this version of the formula, the species under consideration are presumed to be spherical, and the species under consideration are two discreet entities on the atomic/molecular scale.[224] For the purposes of our discussion, the important thing to note is that the dispersion energy has a relatively short spacial reach, decreasing proportionally to the sixth power of the inter-species distance.

2.5.3 Polar Interactions

Monopolar Interactions

The interaction of two monopoles (that is, in this context a coulombic interaction of two ions) is denoted by the equation:[155]

$$U_{mono} = \frac{q_1 q_2}{4\pi\epsilon_0 r} \tag{2.10}$$

Where:

- U_{mono} is the energy of the monopolar interaction
- q_1 and q_2 are the point charges of the two species involved
- ϵ_0 is the permittivity of free space (for species in a medium, this is replaced by the dielectric constant of that medium)
- r is the distance between the two species

Note that the coulombic interaction maintains a much larger range of effect than the London forces.

Dipolar Interactions

Dipoles refer to rigid bodies which maintain a polarisation, a localisation of partial or full charge, across the rigid body. Carbon monoxide would be a good example of such a polarisation. The coulombic interaction can be adapted to describe the interaction of a dipole with a point charge:[155]

$$U_{md} = -\cos(\theta)\frac{\mu_1 q_2}{4\pi\epsilon_0 r^2} \tag{2.11}$$

Where:

- U_{md} is the energy of the monopole-dipole interaction
- q_2 is the point charge
- μ_1 is the dipole moment of the dipole, calculated by $\mu_1 = q_1 l$, q_1 being the polarised charge of the dipole, and l being the length of the dipole
- ϵ_0 is the permittivity of free space (for species in a medium, this is replaced by the dielectric constant of that medium)
- r is the distance between the two species
- θ is the angle between the dipole and the point charge

Note that the range of effect of this interaction is still higher than that for London forces, but smaller than that of the monopolar interactions.

Dipole-dipole interactions are more complicated, since the relative orientation of the dipoles in three dimensions must be considered. It ceases to be useful to attempt to handle the dipole in terms of a dipole moment for these purposes, but instead as a matrix of charge density, and a vector of euclidean positions in three-dimensional space. Working in a two-species system, one can begin by setting up an electric field function for species A:[225]

$$V_A(\mathbf{r}_B) = \int \frac{\phi_A}{|\mathbf{r}_A - \mathbf{r}_B|} d\mathbf{r}_A \tag{2.12}$$

Where:

- $V_A(\mathbf{r}_B)$ describes the electric field effect of species A on locations described by the vector $\mathbf{r}_B$
- ϕ_A is a vector which describes the electronic distribution in species A; this has an involved quantum mechanical derivation[226]
- $\mathbf{r}_A$ and $\mathbf{r}_B$ are vectors describing the relative location of charge in species A and B

$$U_{es} = \int V_A(\mathbf{r}_B)\phi_B d\mathbf{r}_B \tag{2.13}$$

Where U_{es} is the energy of electrostatic interaction and the other terms are the same as in the previous set of definitions. This formula is generally applicable for all multipole moments. Working out the range-of-effect from these equations becomes evidently less

than straightforward, dependent as it is on the orientation and size of the species in question. In the most straightforward instances it is directly proportional to the distance between the species, giving electrostatic interactions the greatest range of effect (although not necessarily the strongest).[155] A broadly similar approach is taken for the atom-atom potential method, although that deals with both attractive and repulsive forces, and can furthermore have different levels of theory applied.[227]

Quadrupolar Interactions

Whilst equations 2.12 and 2.13 apply for all multipole moments, it is useful to illustrate quadrupolar moments as resulting from phenyl rings, because of their prominence within this work. As shown in Figure 2.4, the electron density in a phenyl ring sits above and below the ring as a result of the π orbitals, forming negative regions in that space, and positive regions around the edge of the ring, as shown in Figure 2.5. As illustrated in Figure 2.6, this balance can be altered by introducing different inductive effects to the system.[228]

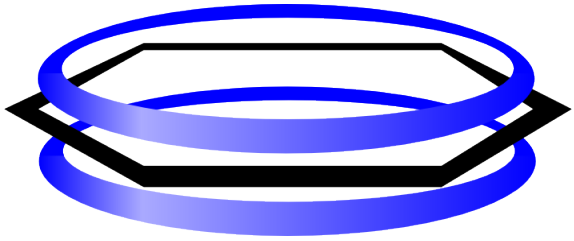


Figure 2.4: An illustration of the π orbitals in benzene

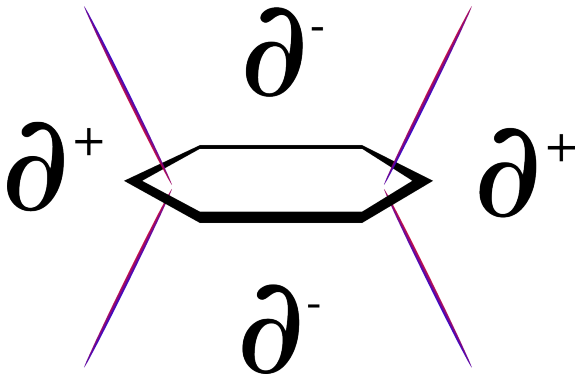


Figure 2.5: An illustration of the quadrupole moment in benzene

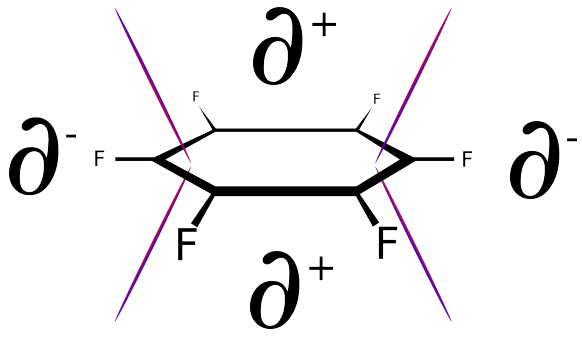


Figure 2.6: An illustration of the quadrupole moment in hexafluorobenzene

2.5.4 Charge Transfer/ Electron Donor-Acceptor Interactions

This kind of interaction is included here for completeness, but is not discussed in detail as it does not pertain to the species that are under examination in this study. Charge Transfer interactions generally occur in metal-ligand complexes,[155,156] but are known in other systems, normally as Lewis acid-base complexes.[156] The interaction involves the transfer of an electron from the highest occupied molecular orbital of one species to the lowest unoccupied molecular orbital of the other; such interactions are stronger than standard electrostatic interactions but not as strong as fully covalent interactions[155] (though they are depicted as such in standard texts[156]).

2.5.5 Hydrogen Bonding

Complicated though multipolar interactions are, Hydrogen bonding is, quantitatively speaking, more complicated again. A hydrogen bond exists between an acceptor atom and a donor atom, mediated by a hydrogen atom, as illustrated in Figure 2.7. The atom which is not covalently bound to the hydrogen atom is generally an electronegative species such as a halogen, oxygen, or nitrogen, and must have a 'lone pair'. Organic synthetic chemists consider hydrogen bonds to have 'covalent character'.

Figure 2.7: A hypothetical hydrogen bond.

In general, the attraction generated by Hydrogen bonds is greater than for other kinds

of interaction, and for this reason they are deliberately avoided for the purposes of the studies discussed in this thesis. A great deal of study has gone into categorising and attempting to develop a quantitative theory of hydrogen bonds, and whilst great strides have been made, a complete theory remains elusive, though there are valuable empirical models available, and much has been learned in recent years with the increasingly available *Density Functional Theory* level quantum calculations.[229] Suffice it to say for the purposes of this text that the Hydrogen bond is considered to be strong enough to be considered capable of directing crystal structure. In short, this means that the arrangement of the crystal structure is in part dictated by hydrogen bonds over and above other interactions, though how this compares to molecular shape considerations is not clear. The notion that hydrogen bonds are structure directing is illustrated by the wealth of literature dependent upon this concept referenced in Section 2.6.

Whilst the quantitative detail of hydrogen bonding is complex in terms of the mechanics, simplifications exist for the purposes of the identification of Hydrogen bonds. The CSD software suite, for example, sets a default cut off that the three point angle of X, H and A (in the hypothetical figure) must be no less than 120 degrees, and further that the distance between the X and the H should be smaller than the sum of the van der Waals radii - and this has become the community standard. Though, as noted in the version changes document,[230] the standard was until quite recently to use 90 degrees as the default.

In general it is considered to be the case that hydrogen bonds are shorter when they are stronger (as is generally true for most interactions). In addition, 'strong' hydrogen bonds are considered to exist where X is an electronegative species such as oxygen or nitrogen, and A is also an electronegative species such as a halogen, oxygen, or nitrogen (though fluorine is excepted from this in organic systems, as will be discussed shortly). Weak hydrogen bonds are considered to exist between species where X is specified to be carbon and A remains a hydrogen atom. This is particularly considered where the hydrogen atom is attached to an aromatic group - where the quadrupole system places the hydrogen in a position of partial positive charge, which is much the same effect as when the hydrogen is attached to an electronegative species.

As can be seen in Section 2.6, there are a large variety of such hydrogen bonding systems in organic molecular species' crystal structures. Fluorine, however, has been excluded from considerations of hydrogen bonding in organic systems since the paper self-explanatorially titled 'Fluorine hardly ever accepts hydrogen bonds' was published in 1997.[231] It was noted therein that where carbon-fluorine groups were present in an organic molecule, they did not find examples of hydrogen bonds between such species

with hydrogen which was bound to oxygen, nitrogen, or carbon. This in spite of the fact that the fluoride ion forms very strong hydrogen bonds. In addition, in the few such species existed that the geometry was correct for a hydrogen bond, other considerations cast doubt as to their veracity. Namely, other incidental factors which could give rise to the same formation or a low-level quantum theoretical calculation.

2.5.6 Repulsion Forces

Repulsion forces come into effect when species are placed in such close proximity that nuclear and electronic repulsions and kinetic energies begin to dominate over attractive forces between the species. As such, they are modelled as being incredibly short range interactions. The *hard sphere potential*, for instance, postulates that the potential energy of the two species becomes infinite as soon as the species are closer than a specific separation, mathematically:[155]

$$U_{rep} = \begin{cases} \infty & \text{if } r \leq d \\ 0 & \text{if } r > d \end{cases} \tag{2.14}$$

Where:

- U_{rep} is the repulsion energy
- r is the distance between the two species
- d is the threshold distance at which the potential becomes infinity

Another way in which the repulsion energies can be modelled is the Mie potential, presented here in a simplified form from the standard textbook[155v]

$$U_{tot} = \frac{C_n}{r^n}\frac{C_m}{r^n} \tag{2.15}$$

Where:

- U_{tot} is the total (non hydrogen bonded) interaction energy
- C_n and C_m are repulsive and attractive coefficients respectively, according to the identity of the species concerned
- r is the distance between the two species

[v] Though an expression resembling this cannot be found in the original publication.[232]

- n and m are exponents which dictate the distance over which the repulsive and attractive interactions persist. $n > m$ must be true.
- Hence, the repulsion is quantified by the term $\frac{C_n}{r^n}$

A special case of the Mie potential is the Lennard-Jones potential, which is often used in textbooks because of its clarity, but has a number of weaknesses for practical application.[155]

$$U_{tot} = 4\epsilon \left\{ \left(\frac{r_0}{r} \right)^{12} - \left(\frac{r_0}{r} \right)^{6} \right\} \tag{2.16}$$

Where:

- U_{tot} is the total (non hydrogen bonded) interaction energy
- ϵ is the magnitude of the minimum energy of the interaction
- r_0 is the distance between the species at which U_{tot} is 0
- r is the distance between the two species
- Hence, the repulsion is quantified by the term $\left(\frac{r_0}{r}\right)^{12}$

It should be noted that in all of these examples, the range of effect for the repulsive forces is always much shorter than for the attractive forces.

2.6 The Aufbau Approach

2.6.1 Overview

In general within the scientific literature, crystal engineering refers to the attempt to design crystal structures based on knowledge of probable motifs that will arise from given supramolecular synthons, to the exclusion of crystal structure prediction using *ab initio* methods or a discussion of a full theoretical treatment of the interactions. Here, however, the findings of such crystal engineering studies are presented only as a part of a discussion on crystal engineering, alongside *ab initio* crystal structure prediction techniques and a theoretical discussion of intermolecular interactions. This is because of an analogy with traditional engineering, where known mechanisms are used to construct predictable outcomes, and other machinery can be reverse-engineered. The analogy here should be apparent; the aufbau approach is the construction, whilst *ab initio* studies and theoretical

discussions on intramolecular interactions are the analogue to reverse-engineering - but engineering nevertheless!

'Aufbau' in this context derives from Kitaigorodskii's aufbau principle, which poses that molecular entities in crystals will coalesce into smaller constructs, which will in turn coalesce into the body of the crystal structure.[233]

The term supramolecular synthon has arisen by analogy with the term synthon from retrosynthetic analysis within organic synthetic chemistry. In retrosynthetic analysis, a synthon is described as an 'idealised reagent'; a (often charged) fragment which stands in place of a reagent, from which one can deduce a reagent for use in the forward reaction. This is illustrated by the retrosynthetic analysis in figs 2.8 and 2.9.[238] However, when first introducing the notion of the synthon, Corey used the term to refer to the chemical moieties within a molecule, which are the result of the chemical reactions implied by the retrosynthetic analysis.[239] In spite of the fact that Corey's definition has broadly been dropped in the scientific literature, to the point where the IUPAC gold book uses the term in the manner described previously[219] (although it does not formally define the term), when coining the term supramolecular synthon Desiraju opted to mimic Corey's definition stating that

> "supramolecular synthons are structural units within supermolecules which can be formed and/or assembled by known or conceivable synthetic operations involving intermolecular interactions."[240]

Somewhat confusingly he later revisited the term, implying two other definitions, only one of which is partially compatible with the original.[38,104] Multiple definitions coupled with the already broad remit of the term supramolecular synthon have given rise to the term being used to describe a broad range of supramolecular constructs. In a review of more than 100 papers[29–151] it can be seen to refer to supramolecular entities which can more precisely be described as interactions,[29–85,152] motifs,[87,112,113,115,116,118–124] bonding networks,[87–111] and supramolecular reagents.[126–133] In addition, several papers misappropriate or misuse terminology which has been in use for considerably longer, with some papers misusing chemical notations,[36–40,98–100,116,118] and one misuses the concept of denticity.[60] This implies that only some errors that are seen with the notion of supramolecular synthons are entirely the responsibility of confusion surrounding the terminology.

Some supramolecular reagents have been extended into the more generalised concept of tectons.[49,243] Tecton is a general term for a structure directing moiety of one or more molecules within a crystal structure which links together to create a supramolecular pattern. This encodes both shape and intermolecular interactions.[244] This is much more akin

to the synthesists' use of the term synthon.

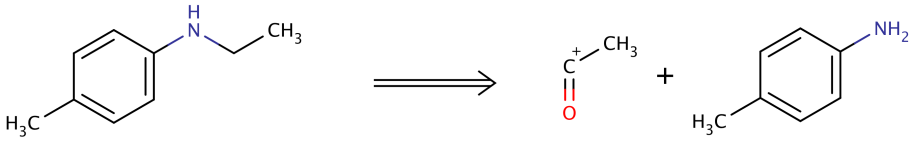


Figure 2.8: A basic retrosynthetic analysis of Paracetamol resulting in a synthon.

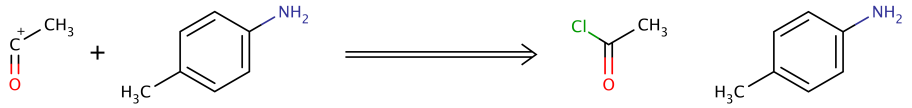


Figure 2.9: An illustration of the deduction of the acyl chloride reagent from its respective synthon.

There are a couple of unspoken assumptions within retrosynthetic analysis; firstly, that the reagents are either naturally or commercially available, and secondly, that the end product has a known or hypothesised useful property. In supramolecular chemistry, the establishment of relationships between crystalline structure and macro-scale properties is relatively recent.[121,242] The former consideration that reagents are commonly available has not necessarily been applied in the context of supramolecular synthons. Furthermore, the focus of crystal engineers on this approach seems to be predicated on the notion that crystal structure is primarily dependent upon the directing effects of functional groups, by way of the analogy with synthetic chemistry. This assumption manifests by the absence of directly comparable evidence of the alternative hypothesis, that crystal structure is directed primarily by shape (in the aforementioned review, only five attempted this comparison[46,59,100,123,142]); this being difficult to test for - changing the constituents in a molecule will change its sterics as well as its electrostatics. This is despite the fact that, thus far, shape has the greatest demonstrable effects on crystal structure to date,[233–237] and this clearly has an impact on the outcomes of otherwise similar intermolecular interactions demonstrated in the referenced papers throughout this section.

As a further divergence from traditional retrosynthetic analysis, there currently exists no established notation for a synthon as opposed to a full reagent. A cursory examination of the referenced papers in this section will reveal that the synthon segment of the molecules illustrated - that is the portion which is responsible for the crystallographic structure motifs observed is often not marked, illustrating further the complicated interaction of molecular structure and molecular shape in influencing the crystal structure. Moreover, the majority of supramolecular synthons that have been identified are in fact retro-justified

from crystallographic data, and do not yet represent the capacity of design that we see in traditional synthesis.

Hitherto, there does not exist in the literature a formal classification for supramolecular synthons. Many entities have been identified as synthons without an attempt to clarify their robustness and/or without an attempt to relate them to a specific motif which the synthon directs[vi]. This often stems from an *ex-post-facto* justification of crystal structure.[76]

More practical challenges to this approach are illustrated where compounds with otherwise similar arrangements of functional groups show different packing arrangements,[245,246] demonstrating that even apparently equivalent functional group arrangements do not completely determine crystal structure.

2.6.2 Interactions of Interest

Given the apparent controversy surrounding the supramolecular synthon nomenclature, this section will simply describe some examples of the interactions which are thought to be important in influencing crystal structure, giving reference to the literature purporting to demonstrate the fact. The papers referenced are the results of an extensive search of the available literature that purports to pertain to supramolecular synthons.

Hydrogen Bonding

The discussion of hydrogen bonds and hydrogen bonded formations makes up the majority of the literature discussing supramolecular synthons. The following exemplify the types of interaction seen which are considered to be 'structure directing'.

[vi] See citations: 32, 33, 35, 37, 48, 50, 60, 70, 75, 76, 80, 81, 83, 85, 86, 93, 94, 98, 103, 106, 108, 109, 114, 121, 124, 133–136, 143, 144

- Two hydroxyl groups[143]
- A hydroxyl group (or water) and any halogen as an acceptor, this interaction may or may not be charge assisted.[81]
- A hydroxyl group and a tertiary amine[54,121]
- A hydroxyl group and a carbonyl group[121]
- A hydroxyl group and a primary amine[68]
- A hydroxyl group and a phenyl π system[32]
- A hydroxyl group and a chlorine ion[129]
- Two carboxylic acid groups[147]
- A carboxylic acid group and a carboxylate group[47,147]
- A carboxylic acid group and a tertiary amine[48,53,54,64,65]
- A carboxylic acid group and a chloride ion[131]
- A primary amine and a carboxylate group[84]
- A primary amine and an ether[33]
- A primary amine and a hydroxyl group[68]
- A primary amine and a sulphonate group[39]
- A primary amine and a cyano group[101]
- A protonated primary amine group and a chloride group[83]
- A protonated primary amine group and a carboxylate group[67,79]
- A secondary amine and a carbonyl group[71,89,121]
- A secondary amine and a carboxylate group[123]
- A secondary amine and a hydroxyl group or water[121,122]
- A secondary amine and a tertiary amine[60,121]
- A secondary amine and a thioamide-group sulphur atom[71]
- A secondary amine and a fluorine ion[107]
- A secondary amine and a chlorine ion[45,107]
- A protonated secondary amine and a carboxylate group[64,66,67,145]
- A protonated tertiary amine and a tertiary amine group[70]
- An amide nitrogen and an amide carbonyl group[48]
- An aminosulphonyl nitrogen and an amide carbonyl group[48]
- An aminosulphonyl nitrogen and an aminosulphonyl SO group[48]

- A triazole nitrogen atom and a different triazole nitrogen atom[91]
- A methyl group and a copper atom or a nickel atom[85]
- An N-methyl group and a carbonyl group[113]
- A methine group and a copper atom or a nickel atom[85]
- A terminal alkyne group and a hydroxyl group[104]
- A terminal alkyne group and a cyano group[43]
- A terminal alkyne group and an arene π system[32,33]
- A terminal alkyne group and an alkyne π system[78,104]
- A phenyl ring hydrogen and a hydroxylate group
- A phenyl ring hydrogen and a cyano group[38]
- A phenyl ring hydrogen and a chlorine atom[33]
- A phenyl ring hydrogen and a chlorine ion[45,128,129]
- A phenyl ring hydrogen and a fluorine atom[76]
- A phenyl ring hydrogen and a carbon-carbon triple bond[33]
- A phenyl ring hydrogen and a phenyl π system[140,145]
- A phenyl ring hydrogen and a carboxylic acid group[131]

It should be noted that the 'singular' interactions are often notionally used to construct more complicated nets and frameworks, however, the robustness of these is rarely demonstrated and it is self-evident in most cases that the shape of the rest of the molecule cannot be ignored, though this is never quantified or discussed.

Carboxylic Acid Like Constructs

As noted in Fig. 1.1, carboxylic acid dimers have proven to be prevalent in the scientific literature on the matter of supramolecular interactions. A great many other interactions exist by analogy with the carboxylic acid dimer as well, and actually are exemplars of a very limited manipulation of molecular shape in conjunction with intermolecular interactions. These include:

- Two carboxylic acid groups[28,47–52,93–95,114–117]
- Carboxylic acids with primary amides[38,49,54,62,63,116]
- A carboxylic acid group paired with the nitrogen and a CH group of a pyridyl ring.[41,50]
- A carboxylic acid group and a quinolene-like configuration of Nitrogen and CH[41,50,121]
- A carboxylic acid group and a pyridone oxygen and CH[41]
- A pair of conformationally locked ureas[88,89]
- Carboxylate groups with guanidinium compounds[135]
- Two amidine groups[60,61,113,115,116,137,143]
- Carboxylate groups and ureas[57,70,110,116,131]
- Two amides[48,49,53–57,93,101,102]
- Two tautomerised amides[66,146]
- A cyanide group and an amine group arranged in a bracket[112,146]
- A pair of pyridine rings[50,53]
- A pair of acetylhydrazine groups[80]

Some interactions are reminiscent of the shape of the interactions listed above, but have a slightly different arrangement of protons.

- Two Phenyl Hydrogens and a 1,8-Naphthyridine[61]
- Two oximes[92,95,127]
- Two hydrazones[91,92]
- An amide and a primary amine[40]
- A metal centre, bonded to two amines and another metal centre bonded to fluorine ions[142]
- A metal centre, bonded to two amines and a boron centre attached to two fluorine atoms[96]
- A diamine and a nitrate ion[96]
- A two amine groups arranged in a bracket and a carboxylate group[60,63,109,119,120]

Bracket-Shape Interactions

Other interactions still, mimic the principle of the bracket-arrangement of the carboxylic acid, but feature more atoms in one or both of the brackets.

- An amide and an amine group linked by a short alkyl chain to a carboxylate group.[123]
- A pair of interactions between a protonated and deprotonated amine[58]
- A carboxylic acid group and an amine group linked by a short alkyl chain to a carbonyl group[58]
- A carboxylate ion and a diol[82]
- A thiane group and a carboxylic acid group[41]
- A carboxylate group and a pair of linked amines[108]
- A pair of carbonyl groups and a pair of linked amines[108]
- Two amine groups and two carbonyl groups[107]

Polyfurcated hydrogen bonds and over-coordinated hydrogen bonds

Some hydrogen bonds are bifurcated or even trifurcated to two receptors. Conversely, more than one hydrogen bond may form with a receptor atom which has more than one available lone pair, such as oxygen. Such have been published as structure directing interactions:

- A bifurcated hydrogen bond from an alkyne hydrogen with a nitro group[42,43]
- Bi- and tri- furcated hydrogen bonds from amines to metal-connected chlorine[44,45]
- A bifurcated hydrogen bond from a phenyl hydrogen and two metal-connected chlorine atoms[45]
- An amine and a carbonyl linked to a hydroxy group[34,35]
- An over-coordinated hydrogen bond from urea NH_2 groups to urea carbonyl groups[70,72,73]
- An over-coordinated hydrogen bond from a methyl group and a phenyl group to an urea carbonyl group[72]
- An over-coordinated hydrogen bond from two amine groups to an urea carbonyl group[120]
- A bifurcated hydrogen bond from two carbonyl groups and a pyridyl group[74]
- An over-coordinated hydrogen bond from two amine groups and a chloride ion[75,149,150]
- An over-coordinated hydrogen bond from two amine groups and a bromide ion[75,96,149,150]
- An over-coordinated hydrogen bond from two amine groups and an oxygen atom[75,96,149]
- An over-coordinated hydrogen bond between a fluorine attached to phosphorus and a pair of amines[75,142]
- An over-coordinated hydrogen bond between a fluorine attached to boron and a pair of amines[96]
- An over-coordinated hydrogen bond between a pair of amines and a cyano group[97]

Complementary Hydrogen Bonded Shapes

By extension of the above, many occurrences exist in the literature of creating molecules shaped that hydrogen bond donor/acceptor pairs can potentially align, and thus form multiple hydrogen bonds between species in a crystalline structure.[69,106,107,119,136–139]

Hydrogen Bonded Networks

Other interactions which are published as robustly structure directing require formations which are more complex than the two or three bodied interactions listed so far.

- A ring formation of water and carboxyl groups[87,141]
- A ring formed from six water molecules or hydroxy groups[36,68,98]
- A ring formed from 3 hydrazone groups[92]
- A ring formed from 3 oximes[92]
- A 42 membered ring involving hydrogen bonds between carboxyl, azide, hydroxyl and amine groups[91]
- A chain of oxime groups[92,95]
- A chain of amine groups[92]
- A tetrahedron of hydroxy groups[100]
- Cubic formations involving hydrogen bonds between carboxyl group and amine groups[151]
- A ring involving a hydroxy group and two carboxylic acid groups[50]
- A three membered ring composed of three hydroxy groups[37,41]
- A complex interaction between a hydrazone group and some iodine groups[107]
- A square of hydroxy groups[68,98,106,148]
- A chain of hydroxy groups[68]
- A square of hydroxy and methyne triple bonds[104]
- Several formations involving water as a bridging species[47,93,124]

Halogen Bonds

Halogen bonds have been defined by analogy with hydrogen bonds. They are an attractive interaction between two halogen atoms, or a halogen atom and a heteroatom - and electron density evidence has been found for their nature as bonds (rather than simple electrostatic interactions[241]). It should be noted that in general, fluorine is not considered among the halogens to form this particular interaction, and that the property is linked to the polarisability of the halogen atom.[38] Furthermore, the interaction is asymmetric, forming attractive interactions reliably requires a specific geometric orientation of the involved atoms, as it is induced by electron-rich and electron-poor regions on the halogen atom.[38,241]

Several examples of this exist in the literature, many of which are analogous with the hydrogen bonding interactions listed previously.

- bifurcated halogen bonds[38,57,112]
- One-to-one interactions involving amine groups[38,49]
- One-to-one interactions involving cyano groups[38]
- One-to-one interactions between two halogens[35,37,41]
- Network formed from suitably arranged halogen atoms[37,41,118]
- Between the oxygen in water and an iodine atom[81]
- Between the oxygen in a nitro group and an iodine atom[57]
- Between a halogen and a carbonyl group[59]
- Metal halides and amines[126]
- Metal halides and other halogens[77]
- Halogen atoms and halide ions[81]

Other Supramolecular Interactions

Other supramolecular interactions have also been proposed as structure directing such as π-π interactions.[46,55,58,87,112,138,140] However, a great deal of literature has been written on the subject, and such interactions are somewhat problematic. Whilst they may exist, the nature of the interaction is hard to characterise, and the term π-π may imply a false specificity,[247] with much literature on the subject now holding the opinion that such stacks,

where they are genuinely energetically favourable interactions, are in fact quadrupolar in origin.[248]

Pnicogen and Chalcogen bonds have also been suggested as structure directing interactions. Evidence for their existence is presently drawn from theoretical and geometric considerations.[251–253]

General Remarks

It can be noted from the above that these interactions, even traditionally 'robust' ones, do not have a guaranteed outcome. Furthermore, in the references for each interaction 'class' that has been listed here, it can be readily seen that each interaction can produce a very different effect depending on the shape of the molecules to which the interacting moieties are attached. Even where shapes in molecules are not significantly different, instances of the above can demonstrate the ability for tectons to form more than one supramolecular pattern. As such, they are something of a blunt instrument, albeit an apparently useful one. This underpins the need for a more incisive approach.

2.6.3 Interactions Involving Fluorine Atoms

In the species being used as an exemplar in this thesis, it can be noted that there is a great potential for hydrogen-fluorine interactions and fluorine-fluorine interactions. Organic fluorine is particularly interesting because it very rarely forms hydrogen bonds.[231] That there is an attractive interaction is evident, but such evidence also indicates that these interactions are primarily dispersive in character.[254] Hydrogen bonds using fluorine have been reported by utilising boron as a substituent of an organic molecule.[255] That true organic fluorine hydrogen bonds exist is still the subject of some debate in the literature, with the case in favour primarily resting on theoretical and geometric considerations - though some charge density results do exist for a few cases.[248]

Of relevance to this thesis in particular is a publication by Kaur and Choudhury which claimed, via the use of geometric considerations, that fluorine based hydrogen bonds exist consistently in fluorobenzanilides.[256] Similarly, Kaur *et al.* saw fit to claim similarly for a different group of fluorobenzanilides.[252]

In a similar vein, the interaction of organic fluorines with each other has become the subject of some considerable controversy. Intuitively, such an interaction would be highly unfavourable, fluorine being a strongly electronegative atom with a very low polarisability. Pauling's principles therefore dictate that this would make interactions of the type seen with more polarisable halogens such as iodine unlikely. Nevertheless, some research

groups still consider the fluorine-fluorine interactions to be likely and stabilising considerations.[248–250] Where such justifications emerge, they are often in tightly constrained subsets which therefore fail to address the question as to whether there is some secondary factor that forces fluorine atoms into close proximity, one which is specific to the test cases observed,[249,250] whilst others do not show consistent behaviour even within such a subset.[257]

2.7 XPac2 and Crystallographic Construct Analysis

With one of the chief criticisms of the aufbau principle being that it does not directly take into account the effects of chemical shape, and furthermore that the approach as presented in the scientific literature does not take into account the robustness of the different interactions. An approach was developed by Thomas Gelbrich which approaches the issue in an agnostic fashion with respect to intermolecular interactions, instead treating molecules as sets of corresponding ordered points in space. This was implemented in the program XPac2.[3]

The details of the algorithms are not stated explicitly in the original publication, and the source code for this program has not been made available. The program analyses sets of crystal structures by comparing the relative positions of a selected set of ordered crystallographic points. Where commonalities between different crystal structures are identified, these are reported as constructs. These constructs fall into four categories: zero, one, two, or three dimensional. At present, the program only performs the analysis in a pairwise fashion among all of the selected crystal structures, which makes some aspects of the analysis problematic for large datasets, as will be seen later in this text.

Zero-dimensional constructs are those that do not repeat in the same way between different crystal structures, but are found throughout compared crystal structures. In principle, all crystal structures with corresponding ordered sets of points share at least one zero-dimensional construct – those of the points themselves. However, these are assumed and hence ignored for the purposes of the XPac2 program. Whilst the approach is agnostic to intermolecular interactions, a good exemplar would be the example of a hydrogen bonded dimer using the carboxylic acid paired interaction. The atoms involved in the pairing will be arranged similarly with respect to each other, but the location and orientation of each pairing throughout the crystal structure need not be the same.

By contrast, one-dimensional constructs repeat infinitely along a given propogation vector, in a manner which is common to the crystal structures being compared. Two, and three dimensional constructs are an extension of this principle, with the two-dimensional

constructs having two directions of propogation, and so on. It should be noted that compounds which share a three-dimensional construct are by definition, isostructural in the traditional crystallographic sense. Such a statement does not include incommensurate structures, which in principle can possess constructs in the fourth dimension; but these are not presently accounted for in XPac2.

By assessing the regularity with which given constructs arise, one can build crystal engineering suppositions which take into account not only intermolecular interactions but also molecular shape, making this approach much more generalisable.

2.8 Crystal Structure Prediction

Ab-initio crystal structure prediction is the 'reverse engineering' of crystal engineering. Instead of seeking to know how to create a given crystal structure or interaction, it seeks to know what interactions will form based on molecular structure. Regardless of the particulars, methods which seek to perform this task follow the same general pattern. This pattern is ultimately grounded in the field of thermodynamics - and seeks to generate the most stable crystal structures - that is, those systems with the lowest internal energy. This, of course, cannot be measured directly.

It is known that it is not only the most thermodynamically stable crystal structures are the ones that form; the existence of polymorphism in crystal structures demonstrates that meta-stable crystal structures must also be able to form. This gives rise to the following general procedure:

1. Randomly or pseudorandomly generate a series of potential crystal structures for the species under examination.
2. Calculate energies for those crystal structures
3. Generate more crystal structures, using the lower energy structures from the starting group
4. Calculate energies for the new crystal structures
5. Repeat as needed

Such a procedure allows the generation of an energy 'surface', from which one can calculate not only the most thermodynamically stable compounds but also contemplate regions which might be considered local minima, in which a crystal structure would require

a significant amount of energy to destabilise, and thus may also indicate meta-stable forms and polymorphism.

Whilst successes have been noted with this technique,[19–23] such methods are not without their drawbacks. Because of their reliance on quantum mechanical calculations (either via molecular mechanics, DFT, or other related methods), such calculations are computationally demanding, requiring long stretches of time on powerful supercomputers - often longer than it would take to simply run the crystal structure experiment. Furthermore, the predictions of such calculations tend to produce multiple suggestions, one of which may be the actual crystal structure.[19–23] Additionally, the predictions give no indication of the conditions under which a given crystal structure may be viable; which means that one cannot be certain if a predicted crystal structure has not been seen because of the invalidity of the prediction, or simply because a suitable experimental space has not yet been explored.

Chapter 3

Cheminformatics

3.1 Descriptors

3.1.1 Overview

When describing a system statistically, one does so in terms of descriptors. A descriptor can be qualitative (e.g. crystal quality measured in terms of "good" or "bad") or quantitative (e.g. Number of Molecules in the Unit Cell).

A descriptor's only real required quality is that it discriminates between members of a population of objects. Thus, a so-called descriptor which possesses the same value for all members of the population cannot be said to be a descriptor at all. Furthermore, a boundless descriptor which possesses a many-to-many mapping with a population cannot successfully discriminate between members of a population. The key point here is 'boundless'. Space groups from crystallography, for instance, are mathematically bounded - a finite number of them exist. This means that even with a many-to-many mapping, there are a finite number of groups in which a molecular species can exist (if it is highly polymorphic). Thus, there are a finite set of combinations of space groups in which a molecular species may exist and, that being the case, the set of combinations can be used as a descriptor to differentiate instead.

Imagine now an extremely hypothetical universe where space groups were unbounded in number. Any compound could exist in any number of space groups which has not been a part of our previously collected data, and the descriptor is categorical rather than being ordered. It becomes logically impossible to make any predictions about these other-universe space groups which have not been present in the training data.

In the broadest sense, descriptors can be collated into four categories: numerical, ordinal, categorical and boolean. Numerical descriptors are those which exist on natural

numerical scales which progress uniformly if not linearly (e.g. linear or logarithmic scales). Ordinal descriptors can also be ordered, but are generally qualitative categories like large, small, and medium. Categorical descriptors, by contrast, have no sense of ordering. Both ordinal and categorical descriptors are subject to the many-to-many mapping corollary. Boolean descriptors are a further special case of Ordinal descriptor, being always bounded at 0 and 1.

Another differentiation exists between ordinary descriptors and spectral descriptors, which is illustrated below.

3.1.2 Molecular Descriptors

The popularity of QSAR experiments has led to the development of a great many molecular descriptors.[190,191] In one sense, this is greatly beneficial, since it means little work in devising novel methods to describe molecular entities. On the other hand, one off the shelf package can calculate many thousands of descriptors, and it is not immediately obvious which ones are important or relevant to any given study. This problem is addressed for the present study in Section 5.

What follows are examples which serve to illustrate the distinction between ordinary and spectral descriptors.

Example 1: An ordinary descriptor, Total Polar Surface Area

The term ordinary, when applied to descriptors in this thesis, refers to a descriptor which produces a single value for a molecular structure. A good example of such a descriptor is the topological polar surface area (TPSA[i]). The total polar surface area describes the area which is accessible to solvent molecules. There are a variety of algorithms for calculating this.[191] For the course of this Ph. D. the algorithm used is that of Ertl, Rohde, and Selzer.[280] The algorithm is based on predefined fragments of molecular structures. Molecular structures are broken down into these fragments, each of which possess a value of surface area assigned to them which defines their contribution to the Polar Surface Area.

The values assigned to each fragment are derived from the world drug index set of molecules.[281] For these drug molecules the calculation of their TPSA was calculated using a (slower) method put forward by David Clark,[282] wherein the total solvent accessible surface area is calculated according to the methods of Dodd and Theodorou.[283] The geometric

[i]Dragon Descriptor Code: TPSA(tot)

details of this method are now well established in terms of computational problems, and are quite tedious, so the following summary will suffice:

1. The van der Waals surface areas of each atom in the molecule are calculated.

2. The areas of these spheres where they overlap is subtracted from each sphere.

3. The area contributed to the total surface area of each sphere is output.

From this, it can be deduced to what extent certain elements, counted as polar (normally nitrogen and oxygen, but often extended to include sulphur and phosphorous) contribute to the surface area, and this area is considered the polar surface area.

This algorithm has the obvious drawback that it does not distinguish between accessible and inaccessible surface areas. For instance, large molecules can have regions of 'surface' that would not be exposed to either themselves or small solvent molecules (a commonly used solvent molecule for the purposes of such a deduction is water[191]). For some species, therefore, the TPSA calculated for this algorithm may be less relevant.

In any case, the method used by Ertl, Rohde, and Selzer optimises the above calculation by removing the need for a three-dimensionally arranged molecule by pre-assigning TPSA contribution values to fragments of the molecule. These values are derived using a least-squares fit such that the calculated TPSA using the fragments correlates strongly with that of those calculated by the traditional method.[280]

Example 2: A 'Spectral' Descriptor

A spectral descriptor, for the purposes of this thesis, is the result of a calculation, which is some arbitrary function the nature of which cannot be known *a-priori* and cannot necessarily be compared directly between different molecules. The result is, therefore, a series of descriptors which are values taken at arbitrary but consistent points along this spectrum.

The molecular walk count of order n (where n is an arbitrary integer) would be just such a descriptor[ii]. The standard molecular drawings used by (particularly organic) chemists can also be thought of as a graph, in the mathematical sense of that word, where each atom is a *vertex*, and each bond is an *edge* on that graph. In the context of graph theory, a walk is a group of edges and vertices such that one can move from each vertex along an edge within the group. Repetitions of any edge or vertex are permitted. The walk count

[ii] Dragon Descriptor Code: MWC01 through MWC10

of order n, is the number of these walks that can be found for n moves within a graph (molecule).[191]

In the course of this thesis, a Spectral Descriptor will be said to have a spectral value. For the example of a molecular walk count, this is the value of n.

3.1.3 Crystallographic Descriptors

Whilst molecular descriptors are well represented in the literature, descriptors for crystal structure are harder to obtain. This is complicated further if one intends to use the descriptors as response descriptors, as this adds the additional constraint that the descriptors should be invariant (i.e. do not directly depend upon) with respect to the molecular descriptors.

For instance, one could conceive of using the dimensions of the three axes of the unit cell as a descriptor. The problem being is that this is obviously dependent upon the size and symmetry of the molecular species involved - hence it is not invariant.

The lack of pre-derived crystallographic descriptors has lead to the development of new crystallographic descriptors, which is discussed in Section 6. What follows is a discussion of valid descriptors which already exist (although there may or may not exist software which readily derives them from crystal structures), and a handful of mathematical constructs which have been presented in the literature as descriptors but either do not function for the purposes of this study, or are not descriptors at all.

Ab-initio Energy Calculations

Ab-initio energy calculations are an ideal descriptor from a statistical standpoint; they are numeric, and they are calculable for any crystal structure - without suffering some of the headaches posed by crystal structures that have more than one entity within the unit cell. The key drawback with any *ab-initio* calculation in crystallography is that they are computationally extremely expensive to reproduce.

Specific Geometric Descriptions

The standard approach when analysing crystal structures is presently to assess specific geometric features and trends thereof within a crystal structure or within a set of crystal structures belonging to a related series of compounds. Geometric features might include things such as a specific bond length, the presence of absence of an intermolecular or intramolecular contact, void spacing, inter-planar angles, torsion angles or 3 bodied angles. Such descriptors are valid for a given subset of compounds, however, they are impossible

to extend to all crystalline compounds, by virtue of the fact that the ability to measure them is inherently dependent upon the species under examination.

Graph Sets

Graph Sets in their current incarnation were first formalised by L. N. Kuleshova and P. M. Zorky[284] and a symbol was devised for their display: $G_m^n(k)$. G is the formalism for the character of the graph under examination. In the original scheme proposed by Zorky and Kuleshova, it could take the values I (islands), C (chains), L (layers) or F (frameworks). This was later modified by Etter to the set C (infinite chains), R (rings), D (dimers or other finite group) or S (selfs, for intramolecular interactions), along with a more rigorous definition for classification.[285]

k is the symbol for the degree of the system. Again, the utility differs between Zorky and Etter. Zorky defined this symbol as being the dimension (number of members) of the rings present in a graph.[284] In Etter's more precisely defined definition, the definition of this term is dependent upon G. If G is R or S, then the value of k is the number of atoms in the ring. For the case of G being set to D, k is the number of atoms involved in the entire hydrogen bonding system, along the shortest path. If G is C, then the value of k is the number of atoms between the first donor and the last acceptor in the system, along the shortest path.[285] the symbol n refers to the number of hydrogen bond acceptors in the graph, whilst m refers to the number of hydrogen bond donors in the graph.

It should be noted, that whilst the formalisms of n and m are described using hydrogen bonds as the topology determining interaction, the software Rpluto,[286] and recently Mercury[230] allowed the use of any given set of atom-pair interactions to be used to determine the graph sets.

Despite the utility of graph sets as a topological descriptor, there are issues with using them in a statistical model. The only means of using this descriptor would be to treat it as a binary descriptor (contrary to its description as a *quantitative descriptor* by Bernstein et al.[287]). One might have thought to use it as a categorical descriptor, however, the mapping between crystal structures is a many to many mapping. Many crystal structures can 'belong to' one graph set, whilst many graph sets can apply to any one crystal structure - ergo it would have to be used as a binary descriptor. The problem with using the graph sets as a binary descriptor is that the set of graph sets is infinite, and so whichever constraint was placed on the set of graph sets to be used would be arbitrary, and thus any model would not have any predictive power beyond the sets specified beyond this arbitrary constraint - although it may yet emerge that whilst the set

is not mathematically constrained, it might be practically constrained, such as has been found in some organometallic materials.[288]

In addition, the use of graphs in this way creates additional problems in the form of subgraphs, which will always be associated with their respective supergraphs. This means that it is impossible to orthogonalise graph set descriptors when characterised in this way, which would make statistical models which use them as binary descriptors challenging to interpret.

Space Groups and Symmetry Elements

On the surface, a space group appears to be an ideal descriptor. There is a mathematically constrained set of 230 space groups,[197] with the mapping between these and a crystal structure being that a crystal structure cannot belong to more than one space group. However, a few problems do arise in trying to implement space groups as a descriptor. First and foremost, space groups do not readily differentiate between different species, especially for organic compounds. A survey of the CSD indicates that over a third of crystal structures contained in that database are of space group $P2_1/c$.[289] With such an uneven distribution, use of space groups are not going to be entirely helpful either.

The Unit Cell

The use of the six measured dimensions (the unit cell lengths a, b, c and their corresponding angles α, β, γ) of the unit cell also seems like an obvious choice for describing a crystal structure. Indeed, it is considered to be a fingerprint by which one can identify a crystal structure. However, these descriptors are not orthogonal to the size of the species under examination, and so one would expect to see a very strong correlation between molecular size and unit cell volume for organic compounds.

Attempts have been made to normalise the unit cell. One such attempt is the *packing coefficient*. This very simple normalisation scheme is the ratio of the volume of the unit cell contents to the total volume of the unit cell.

A more subtle and complex approach is seen in the derivation of the box model. The box model was devised by Elna Pidcock and Sam Motherwell at CCDC, and describes crystal unit cells in terms of constrained arrangements of molecular boxes, and provides a powerful descriptor for use in statistical models, when applied in conjunction with values of Z and Z'. This descriptor is named the pattern coefficient, and measures the ratio in size of a molecular box to the unit cell.[234–237]

For instance, a Z=4 structure (a crystal lattice with 4 molecules in the unit cell) will

have each unit cell divided into 4 corresponding boxes, one for each molecule. These boxes (equal in size) will have three unequal dimensions which are perpendicular to each other; L, M and S. The assignment of dimensions L and M are made using the principal axes of inertia of the molecule (that is, its longest and next longest dimensions). The dimension S is then assigned as being the remaining dimension perpendicular to both L and M.[234]

Each of the axes of the box is then assigned to the axes of the unit cell with which the axes of the unit cell most closely align. The ratio of the lengths of each pair of axes provides the pattern coefficient.[234] Population graphs of these coefficients mined from data in the CSD have demonstrated distinct population profiles for these ratios at various Z and Z' values. There is also some data that suggests an asymptotic, non-trivial relationship with the orientation of the molecule within the unit cell - and whilst this relationship is not transparent, it does suggest another descriptor for use in statistical models.

As well as the numerical descriptor described, the arrangement of the boxes within a unit cell also proves to be a descriptor of sorts, although that is categorical rather than numeric. For instance, for Z=4 structures, there are three different pattern structures.[237]

This descriptor also seems to have some correlation with the location of molecular entities within the unit cell. Once again, this relationship is not as transparent as the pattern coefficient, but still provides another potential crystallographic descriptor. It also implies the necessity of including the relevant crystallographic descriptors that are already available, namely those which describe the unit cell, and the Z and Z' values thereof.

3.2 Statistics

3.2.1 Variance, Covariance and Degrees of Freedom

Variance is a measure of how much a variable deviates from the average. In the common use case of n equally likely values of a random variable, the variance is calculated thus:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n} \tag{3.1}$$

Where:

- σ is the standard deviation of the population in consideration
- σ^2 is the variance of the population in consideration
- n is the number of members in the population

- x_i is the variable value for the ith member of the population
- μ is the mean value of the population

Notice that this is the variance of the population. It is unusual in any statistical analysis to have access to the full population, and so a calculation has to be done which estimates the population variance based upon a random sample of that population.

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - m)^2}{n - 1} \tag{3.2}$$

Where:

- s is the (biased) estimated standard deviation of the population
- s^2 is the estimated variance of the population
- n is the number of members in the sample
- x_i is the ith member of the sample
- m is the mean value of the sample

There are two key differences between 3.1 and 3.2. The first is that Greek letters have been re-rendered as Latin characters, and this is simply a matter of convention when discussing estimated sample scalar values as opposed to population scalar values. The second is that the denominator has changed from n in 3.1 to $n - 1$ in 3.2.

This denominator is referred to as the degrees of freedom, and elsewhere is often denoted ν. In his online book on statistical practice, Gerard Dallal writes of this quantity:

> "One of the questions an instructor dreads most from a mathematically unsophisticated audience is, 'What exactly is degrees of freedom?'"[290]

The fact being that many of the justifications for 'degrees of freedom' rarely hold. The mathematics that gives rise to the "degrees of freedom" adjustments is somewhat complicated to justify in full.[291] Suffice it to say, for this Thesis, that it is a mathematical consequence of estimating parameters from sampled data. It occurs throughout several statistical analyses.

In context of the given equations, it can be interpreted as the number of independent pieces of data which can vary independently in order to estimate a quantity. For instance,

the estimated variance is considered a fixed parameter; only $n-1$ data points are free to vary, since the last data point is calculable from the (fixed) variance and the other values.

In any event, the calculation of the variance can also be described using vectors.

$$\mathrm{Var}(X) = \frac{(X - \bar{x}\mathbf{1})^{\mathbf{T}}(X - \bar{x}\mathbf{1})}{n-1} \tag{3.3}$$

Where:

- X is a vector containing the data from the sample
- $\mathbf{1}$ is a vector of 1s the of the same dimension as X
- $\bar{x}$ is the average of the sample
- n is the number of members of the sample
- Superscripted $\mathbf{T}$ indicates the transpose of a vector or matrix

The variance is actually simply a special case of the covariance, in which one measures the variation in one variable with respect to another:

$$\mathrm{Cov}(X, Y) = \frac{(X - \bar{x}\mathbf{1})^{\mathbf{T}}(Y - \bar{y}\mathbf{1})}{n-1} \tag{3.4}$$

Where X and $\bar{X}$ and n are as defined previously, and Y and $\bar{Y}$ are similarly defined for a separate variable. The variance calculation for X in Equation 3.3 simply sets $Y = X$, and $\mathbf{T}$ is defined as before.

For completeness, if one wishes to calculate the variances of p variables taken from the same sample, one can calculate the variance-covariance matrix:

$$S = \frac{(\mathbf{X} - \bar{X}\mathbf{1})(\mathbf{X} - \bar{X}\mathbf{1})^{\mathbf{T}}}{n-1} \tag{3.5}$$

Where:

- S is $p \times p$ variance-covariance matrix
- n is defined as before
- p is the number of variables in X
- X is an $n \times p$ matrix of the sample data
- $\mathbf{1}$ is a p long column matrix of 1s

- $\bar{X}$ is a vector containing the p averages of the sample data
- Superscripted **T** indicates the transpose of a vector or matrix

The diagonal values of S are the variances, and the off diagonal variables are the covariances, such that the first row, second column item in the variance-covariance matrix is the value of the covariance between the first and second column variables of **X**. S is a symmetric matrix.

3.2.2 Correlation

Correlation coefficients are the means by which the strength of a relationship between two variables can be measured. It should be noted that correlation does not necessarily imply that two variables are causally related - only that the data demonstrates a numerical relationship. Equally, given a suitable experimental design correlation *can* imply causation (see Section 3.2.5).

There are several different correlation coefficients, and the choice of which one is appropriate depends on the data types under examination.

The Pearson Product-moment Correlation Coefficient

Most often referenced as just r or Pearson's r, this is the most familiar correlation coefficient. It measures the level of linear correlation, and direction thereof, between two variables. A value of 1 indicates a perfect linear correlation with both values increasing. A value of -1 implies perfect correlation with one variable decreasing with respect to the other. Normally, this is the dependent variable decreasing with respect to the independent variable. A value of 0 implies no relationship between the two variables. Other levels of correlation are more frequently seen, and the interpretations of the strength of these relationships is open to some interpretation - a correlation of 0.7 might be considered very poor in some fields, whilst it might be considered quite strong in others.

The calculation for this correlation coefficient is:

$$r = \frac{\mathrm{Cov}(X, Y)}{s_x \cdot s_y} \tag{3.6}$$

Where:

- r is the correlation coefficient for a sample
- $\mathrm{Cov}(X, Y)$ is defined as in Equation 3.4

- s_x is the estimated standard deviation of x, and similarly for s_y

Pearson's r is best suited for measuring linear relationships between continuous variables. If dealing with nonlinear relationships, one can in principle linearise variables, but this can affect the statistical inferences which can be made using Pearson's r. For instance, measuring r between x^2 and y rather than x and y in order to linearise the relationship would also affect the values of the variances of x, which would affect the certainty measures outlined in Section 3.2.3.

Spearman's Rank Correlation

For dealing with nonlinear relationships it can be preferable to use the Spearman's Rank Correlation, also called Spearman's ρ, which is nonparametric. Spearman's ρ, as the name implies, is to be calculated for ranked variables. This necessitates that any continuous data is taken and placed into ranks. Where the data would tie, the rank assigned should be the mean of the tied rank and the next rank. The score following this should be assigned the next rank again, and so on.

Spearman's ρ is formulated for a sample as follows:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} \left(R(x_i) - R(y_i)\right)^2}{n(n^2 - 1)} \tag{3.7}$$

Where:

- n is the number of members in the sample
- $R(x_i)$ is the rank of the ith score of x and similarly for y

If two variables have a monotonic relationship, then ρ will be equal to ± 1, with the sign indicating the direction of the relationship.

3.2.3 Statistical Inference

Calculating correlations is useful, but there is always a chance that such correlations are 'chance correlations'. Statistics has largely been concerned, therefore, with calculating certainties that this mis-assignment is not happening.

Frequentist statistical experiments tend to be arranged with two hypotheses; a null hypothesis, and an alternative hypothesis.[iii] In general, when speaking about tests of

[iii] Other set ups exist, but are not pertinent to the work presented.

correlation, one considers the null hypothesis to be that there is no correlation between two variables. The alternative hypothesis is simply that there is correlation – note that the alternative hypothesis is *not* that correlation is observed at the level found in our sample data.

In the most general sense, the scientist performing the statistical experiment should set an α value. The α can be considered as a satisfactory risk that the scientist will reject the null hypothesis incorrectly. As such, it must be set *before* any statistical analysis is performed. To test against this alpha level, a p value is calculated for the data under examination. The method for calculating a p value varies from test to test, but if it is lower than the α level, this is considered to be a statistically significant result, and reject the null hypothesis.

The p value in the general case is informative of the likelihood of seeing a value of some statistic (for instance, correlation) that would meet or exceed the value of that statistic if the null hypothesis is true - i.e. if a random set of data were taken from the null distribution, how likely would it be that the statistic value would be as extreme as that seen from the data. This is often misinterpreted as the likelihood of falsely rejecting the null hypothesis. The alpha value that is set, therefore, should be cognisant of the relative size of the test sample as to the population; if the population is orders of magnitude larger than the test set, then it becomes more feasible for an independent test set with a statistically unlikely characteristic to arise.

It should be also noted that if many correlation analyses are made on the same data, that the likelihood of seeing a value of the equal or higher test statistic goes up. The simplest and most conservative protection against this is the Bonferroni method, which is to set the alpha level at a level proportionately smaller level to the number of tests being performed. That is, if would set $\alpha = 0.05$ for a single test, then one would in fact set $\alpha = 0.05/9$ for 9 tests.

The p value for Pearson's r

The p value for testing Pearson's r relies upon the fact that the underlying variables are normally distributed. Given this assumption, it is a well established result that uncorrelated random pairs from a bivariate normal distribution follow Student's[iv] t-distribution. It follows therefore that the test statistic:

$$t = r\sqrt{\frac{n-2}{1-r^2}} \tag{3.8}$$

[iv] The *nom de plume* of William Sealy Gosset, a chemist at the Guinness brewery company.

With r being the correlation coefficient and n being the number of members of the sample, will provide a p value from Student's t-distribution.

The p value for Spearman's ρ

Spearman's ρ can also be tested for certainty. In addition, it is a non-parametric statistic, meaning that it will apply for cases where variables do not follow a normal distribution. The trade off for this is that non-parametric tests have less 'statistical power' than parametric tests, that is, they are less likely to detect an effect that is really present in the system (they cause the false rejection of the alternative hypothesis).

To do this, one utilises a calculation called the Fisher transformation:

$$F(\rho) = \frac{1}{2} \cdot \ln \frac{1+\rho}{1-\rho} \tag{3.9}$$

Where ρ is the Spearman's ρ.

The transformed statistic can then be used to calculate another statistic:

$$z = F(r) \cdot \sqrt{\frac{n-3}{1.06}} \tag{3.10}$$

p values calculated this way for Spearman's Rank Correlation can be unreliable when there are many 'ties' in the ranking score. It is for this reason that the programming language R does not permit the direct calculation of the p values for Spearman's ρ under these circumstances.[269] Some implementations of testing for Spearman's ρ, notably that used in the programming language R's default libraries, simply prohibit tests in this instance. An alternative method of calculating the p value is therefore the permutation method, in which one approximates the distribution under the null hypothesis by simulation.

In brief, one calculates the ρ value, which shall be called ρ_0. One then randomises the pairing of the paired values in the data used to calculate ρ_0. Then one recalculates ρ, and one repeats this numerous times. The proportion of times that ρ_0 is exceeded is the p-value. An implementation of this algorithm can be seen in Section 5.3.2.

3.2.4 Least Squares Regression

Simple Least Squares Regression

A linear equation is always (written in some variant) of the form $y = mx + b$. When working with two variables (x and y), Least Squares Regression attempts to calculate the terms in the linear equation based on the data presented. It does this by attempting to

minimise the total offset between the points of data and the linear equation estimate. Such minimisation operations are not possible analytically on functions which are not differentiable, and so rather than simply using the magnitudes of the offsets it is easier to use their squares. As a result of using the square values, outlying values will have a disproportionate effect on any fit, therefore it is important to remove any unreasonable outliers before the fitting process. Hence, the operation is to minimise the squares of the offsets, which gives the name least squares regression.

$$Y = X\beta \tag{3.11}$$

Where:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \tag{3.12}$$

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \tag{3.13}$$

- Y is the vector of response variable values
- β_0 is the Y intercept
- β_1 is the gradient
- n is the number of members in the sample
- x_i is the value of the independent variable of the ith member of the sample.

Data for use in a least squares analysis is considered as the form:

$$Y = X\beta + \epsilon \tag{3.14}$$

Where Y and X are defined as in Equation 3.11, and ϵ is a vector containing the normally distributed random error (residuals) between the line of the equation and the data.

The β parameters can be estimated by the following formulation:

$$\hat{\beta} = (X^{\mathbf{T}}X)^{-1} \cdot X^{\mathbf{T}}Y \tag{3.15}$$

Where:

- $\hat{\beta}$ is the estimate of the beta parameters as described in Equation 3.12
- X is as defined in Equation 3.13
- Y is the vector of response variable values

Hypothesis testing for a univariate linear model is the test of how well the calculated model matches the data. The hypothesis test is calculated using Pearson's r, and with it the appropriate test given in Equation 3.8.

$$t = r\sqrt{\frac{n-2}{1-r^2}} \tag{3.8 restated}$$

r^2 also has the useful property of being the ratio of the sum of squares accounted for by the model to the sum of squares in the data, making it a useful measure to assess the quality of the model.[v]

Firstly, the following are defined:

$$ss_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3.16}$$

$$ss_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{3.17}$$

$$ss_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{3.18}$$

$$ss_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3.19}$$

Where:

- ss_{xx} is the sum of the squares in the variable x
- ss_{yy} is the sum of the squares in the variable y

[v]This is also mathematically equal to the correlation between the estimates of the response variable from the model and the data. This is in fact used in crystallographic models and is how the R-factor is calculated for use as mentioned in Section 2.3.2

- ss_{res} is the sum of the squares of the residuals between the model and the data
- x_i is the ith value of the variable x, and similarly for y
- $\hat{y_i}$ is the model estimated value of y_i

$\hat{y_i}$ is the model estimated value of y_i, and is thus:

$$\hat{y_i} = \hat{\beta}_0 + x_i \cdot \hat{\beta}_1 \tag{3.20}$$

It follows from the matrix calculations in Equation 3.15 that:

$$\hat{\beta}_1 = \frac{ss_{xy}}{ss_{xx}} \tag{3.21}$$

It also follows from Equation 3.6 that:

$$r^2 = \frac{ss_{xy}^2}{ss_{xx}ss_{yy}} \tag{3.22}$$

We are able to deduce from Equation 3.20 and the definition of ss_{res} that:

$$ss_{res} = \sum_{i=0}^{n}(y_i - \frac{\sum_{i=0}^{n} y}{n} + \hat{\beta}_1 \frac{\sum_{i=0}^{n} x_i}{n} - \hat{\beta}_1 x_i)^2 \tag{3.23}$$

Which multiplies out to become:

$$ss_{res} = ss_{yy} + \hat{\beta_1}^2 ss_{xx} - 2\hat{\beta}_1 ss_{xy} \tag{3.24}$$

Substituting in Equation 3.21:

$$ss_{res} = ss_{yy} + \frac{ss_{xy}^2}{ss_{xx}^2} ss_{xx} - 2\frac{ss_{xy}}{ss_{xx}} ss_{xy} \tag{3.25}$$

$$ss_{res} = ss_{yy} - \frac{ss_{xy}^2}{ss_{xx}} \tag{3.26}$$

$$\frac{ss_{res}}{ss_{yy}} = \frac{ss_{xy}^2}{ss_{xx}ss_{yy}} = r^2 \tag{3.27}$$

Multivariable Least Squares Regression

The multivariable case is accounted for by extending out the dimensions of X in Equation 3.15, adding extra columns for each predictor variable in the model. The dimensions

of $\hat{\beta}$ necessarily increase, adding an extra row for each estimated relationship. r^2 can be calculated via a correlation of the estimates of y, $\hat{y}$, with the observations from data.[292,293]

3.2.5 Experimental Design

Warnings about the fact that "correlation does not imply causation" have become so commonplace that the fallacy to which they pertain has become a part of modern popular culture:

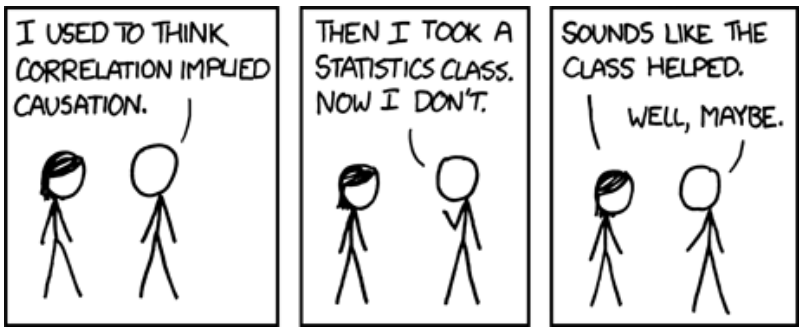


Figure 3.1: The text accompanying this comic read "Correlation does not imply causation, but it does waggle its eyebrows furtively and gesture suggestively while mouthing 'look over there'"[294]

The statement requires adjustment for accuracy; correlation does not *intrinsically* imply causation. However, given a sufficiently designed and controlled experiment, it can. As a further restriction, such an experiment should also check for interacting features of the experiment. A further and more pertinent fallacy, therefore, is to what extent such a conclusion is valid.

One can consider a hypothetical chemical reaction. A student runs the reaction at high and low temperatures, whilst controlling for the amount of reactants and any other factor that they can expect to have an impact upon the reaction. At the higher temperature, they consistently record a higher yield, and so they conclude that a higher temperature induces a greater amount of reaction product.

The student then elects to test for variations of product yield with respect to time. Again, the student controls for all the other factors, and records a higher yield at higher temperature consistently. Thus, they conclude that at longer reaction times a higher yield will be returned.

The student goes on to conclude that the best conditions for that particular reaction would be at high temperature for a long time. Whilst the two conclusions about temperature and time are separately valid - the final conclusion is not. When asked by their

professor to run a reaction under the recommended reaction, they find that the reaction records a poor yield because remaining hot for a long time decomposes the product.

In this particular instance, temperature and time make up the "descriptor space" being explored. In this particular experiment, the descriptor space at high temperature and high time had not been populated, and so the conclusions drawn from the original experiment were not valid in this descriptor space.

A method of experimental design to avoid such problems is known as a 2-level full factorial design. Each descriptor in the space is set at a high and low level, and every possible combination thereof is tested. Two level factorial designs have two problematic corollaries:

1. To run such a design for any significant number of factors requires many procedures to be run; 2^n for n factors.

2. Such an experiment may still fail to spot effects if the reaction to a given factor is nonmonotonic.

A method to avoid the first of these problems exists, and depends on the fact that interactions between more than a certain number of variables may be considered unlikely by the domain expert. Such situations can be exploited to confound multiple factor effects with the single factor effects. A full discussion of this is not pertinent for the presented work.

The latter problem can only normally be identified by running at multiple levels, and this necessarily complicates the experimental design and increases the number of procedures that need to be run. The necessity of such a design is normally left to the judgement of a domain expert.

The reliance of cheminformaticians and statisticians on domain experts is one of the largest caveats of the entire procedure - and relates to the caveat in Section 2.3 on validating crystallographic models. All statistical models and experimental designs *must* make sense within the field they are applied, in this case, they must make chemical sense.

3.2.6 Classification and Regression Trees

Overview

Classification and regression trees, or CARTs, are a means of modelling data which are amenable to nominal (categorical) response descriptors. They work particularly well, therefore, for experiments with a binary response descriptor. They are also useful as a

means of 'feature selection' - the process by which one determines the important governing factors in a system. The generation of such a model results in a decision tree structure, by which one can use the descriptors of a new system to make a prediction about the outcome descriptor for that system.

Many implementations of regression trees exist.[295,296] The implementation discussed here is the one found in the commercially available software JMP 11,[297] as this is the software chosen to work with for this particular type of analysis in the presented work.

The objective of a CART algorithm is to split the data into n classes (defined by the response variable) based on independent variables in the data. This makes a number of splits in the data. Each split will be based upon one independent variable. This means that each split is one dimensional. Interactions between descriptors are represented by subsequent splits on other independent variables.

Construction

In practice, the construction of a CART is automated software. Different algorithms exist for constructing different models. Because JMP 11 is the software in use for the analysis performed in this thesis and is not discussed in the traditional academic literature, we will present the mathematical details of that algorithm here.

One might imagine that the same measures of accuracy which apply to the model overall (i.e. how well the data is divided up) would apply to each split made in the tree. However, as in chess, where one might imagine the best move is to take the piece with the highest point value at each opportunity, one will find oneself defeated when met with a more sophisticated approach which adopts a wider view.

The JMP package utilises the *LogWorth* function to determine which of the many potential splits is most preferable for a given system at each given node on a tree.

$$L_W = -\log(p) \tag{3.28}$$

Where L_W is the value of *LogWorth*. Ordinarily, the value of p would be calculated by using Pearson's Chi-Squared Test:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{3.29}$$

Where:

- n is the number of 'levels' in the model (2 for the binary case)

- E_i is the expected number of items at level i assuming the proportions in the population were true for this node in the tree
- O_i is the number of observed elements at level i in this split

A p value can be taken from the appropriate χ^2 distribution. In some instances, particularly in models with large numbers of factors, this is considered ill suited, and so the Bonferroni adjustment is applied to the p value with respect to the number of splits thus far in the model.

In JMP, the p value is calculated using Monte Carlo methods. The JMP documentation reasons that the calculation it uses is fairer than the unadjusted p and Bonferroni adjusted p.[297]

Utilising the Model to Make Predictions

Predictions of an outcome i from the decision tree can be made by calculating the p_i statistic for each endpoint on the tree:

$$p_i = \frac{n_i + d_i}{\sum_{i=0}^{j} (n_i + d_i)} \tag{3.30}$$

Where:

- p_i is the probability of the ith level at a given node.
- n_i is the number of members of the ith level at a given node
- j is the number of levels

d_i is the prior probability of the ith level, and is calculated:

$$d_i = \lambda d_{i(p)} + (1 - \lambda) p_{i(p)} \tag{3.31}$$

Where:

- $d_{i(p)}$ is the prior from the parent node for the ith level
- λ is a weighting constant
- $p_{i(p)}$ is the probability of the ith level from the parent node

This recursive definition results in a situation where an explicit parent node no longer exists, in which situation the prior is equal to the ratio of the proportion of the data belonging to the ith level.

In practice, these probabilities are calculated and presented automatically by the program.

Validation

One must define a measure of the overall quality of the CART. The obvious criteria is that the splits must, over the entirety of the CART, result in the fewest misclassification errors. That is to say that, on our test data, the model makes the fewest mistakes in assigning the class to which a given datum belongs. This is called the misclassification rate - and is used in many kinds of regression tree.[296] JMP eschews this in favour of what is termed the entropy R^2 in their documentation[vi], which serves the same purpose as the r^2 seen in regression models, but has a very different calculation method.

$$R^2 = -2 \log \left(\frac{L_m}{L_0} \right) \tag{3.32}$$

Where:

- L_m is the likelihood of the model
- L_0 is the likelihood of the constant probability model

The likelihood of a partition model is calculated in the case of JMP by:

$$L = \sum_{i=0}^{m} - \log \left(p_{corr} \right) \tag{3.33}$$

Where m is the count of final nodes in the decision tree, and p_{corr} is the probability of correct response in that node of the tree. A correct response, in JMP, is considered to be the response which makes up the majority of a population at a final node in a tree.[298]

The R^2 value must also be validated. There are several approaches for this, one of which is called k-fold validation. This entails dividing the data up into k sets. The model is constructed using the same splitting criteria, but using only $k - 1$ of the sets of split data. The model is then tested using the remaining set of data - and the p_{corr} values are calculated based upon the misclassification of the test set of data.

[vi] but is elsewhere described as McFadden's R^2.

This is then repeated k times, each time leaving a different set of data aside for to be used as the test. The k-fold R^2 value is then:

$$R^2_{kfold} = \frac{\sum_{i=0}^{k} R_i^2}{k} \tag{3.34}$$

Where R_i^2 is the ith R^2 value.

When R^2 and R^2_{kfold} diverge, this is normally a symptom of over-fitting, and indicates that there are too many splits in the model to be justified by the data.

3.3 Previous Work on Feature Selection

Feature selection refers to the process of deciding which aspects of the population under experimentation are relevant to the response variable under examination. In general, statisticians rely very heavily on domain experts in order to establish what factors are thought to be important by established practice and theory. Crystallography represents an interesting conundrum in this sphere since there are very few easily calculable molecular properties which have been established by long-standing practice or theory. Work by Kitaigorodskii did suggest very strongly that molecular size and shape were strong influencing factors,[233] but very little work has been done since seeking a governing factor from a collection of work, in spite of the existence of databases such as the CSD.

Work done in the past by Elna Pidcock of the CCDC has attempted to address this. The shape of a molecule was described by three principal, orthogonal axes - long, medium and short. A descriptor was calculated by dividing the volume of this box by the volume of the unit cell of the compound. This was termed the pattern coefficient. This was then related to various packing arrangements. Clear patterns did emerge in the data, backing up the notion that shape and molecular volume are governing factors in the formation of crystalline materials. In particular, strong relationships could be identified between the packing coefficient and the location of the molecular centres within the unit cell.[234–237]

Terence Threlfall also conducted (and at present still conducts) a two-and-a-half day course in which he identified no less than 30 factors which had been described by various scientists as having an impact upon crystallisation processes - including factors such as molecular shape and molecular polarity, but also very external factors such as the phase of the moon.[299] Of course, the observations to which he gives reference are not conducted in properly controlled studies, nor were they given a statistically rigorous analysis *post-hoc*, and should be interpreted in that light.

More recently, work by Lazlo Fabian set out to establish the important governing factors in the formation of co-crystals. Such work is hindered by the absence of large bodies of 'failure data' - data pertaining to procedures and experiments which fail to give rise to co-crystals. Instead, Fabian sought examples of complementarity between the member molecules of co-crystalline species. To do this, correlations were drawn between descriptors for each molecular pair in each co-crystal which were obtained from the CSD. Correlations were observed for the various dimensions of the 'box' of the molecules as described in the box model work by Pidcock et al. Correlations were also identified for fractional polar volume (a descriptor which loosely describes the volume in a molecule which can be attributed to polar atoms) and the heavy atom count. Fabian therefore concluded that these aspects of the molecular system were likely important in the formation of co-crystals. Such conclusions carry the caveat of the absence of counterexamples - without examining the cases that did not form co-crystals, one cannot be certain that these correlations are unique features of the co-crystalline systems. Furthermore, all such trials necessarily assume that data from the CSD is a true representation of crystalline descriptor space, which may or may not be true.[153]

More recently still, Richard Cooper et al. published work with the titular question "Will it Crystallise?".[300] The work elaborated on the use of Support Vector Machines (SVMs) to predict whether molecular species would crystallise, and under what conditions. As stated in that paper, SVMs are opaque, and difficult to interpret. As a pragmatic solution, statistical tests were performed on individual descriptors to assess their significance. However, such a method makes the assumption that interaction effects which are important will only arise from descriptors which are themselves important in a one-dimensional sense. This, again, may or may not hold as an assumption - although it can be validated by comparative assessment of models with and without additional descriptors, as was performed in the work. Two key descriptors were found to be important in the study: namely rotatable bond count, and a value called ${}^{0}\chi^{v}$, which is a measure of the valencies of atoms in a network description of the molecules with the hydrogen atoms omitted. It correlates very strongly with molecular volume.

Part II

Laboratory Procedures

3.4 Synthesis

Synthesis of the various fluorobenzanilides was performed by Terence Threlfall, and followed some analogue of the reaction scheme depicted in Figure 3.2

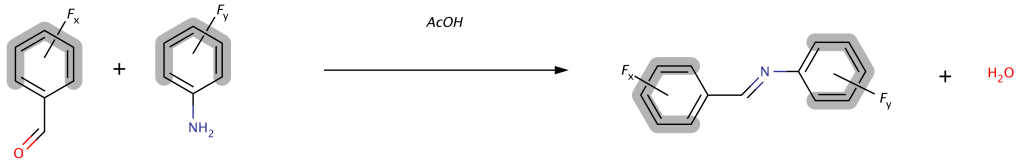


Figure 3.2: An approximate schema of the synthesis of the fluorobenzanilides. Crystals were retrieved from a variety of solvents under varying conditions. Some required significantly more effort to recrystallise than others.

3.5 X-ray Procedural Configurations

There are 7 procedural configurations which were used for X-ray analysis. Six of these were given a moniker: Dot, Ros, Kat[vii], Del Boy, Damien[viii], and Spider[ix].

The seventh procedural arrangement is that of the Diamond Light Source I19 beamline.

The differences between the procedural arrangements are detailed below. In general, the procedure for running an X-ray experiment is the same.

A crystal is selected under a microscope. Crystals which display faults or flaws should not be selected - these can produce multiple diffraction patterns which are then difficult to disambiguate akin to twinned crystals. In addition crystals (with the notable exception of high symmetry unit cells) have the property of extinguishing plane-polarised visible light when appropriately orientated. This can be used as an examination tool - a crystal that is truly singular will uniformly extinguish the polarised light as it is rotated in the light. This also makes it easier to examine for passenger crystals which may otherwise become attached to the crystals under examination.

Crystals should be no larger in any dimension than the X-ray beam, although in practice this does not always prove to be problematic for low molecular weight compounds.

Once a crystal is selected, they are mounted on a 'pip' made from glass fibre, polyimide, or human hair (materials which are not ordered and so only produce uniform scattering of

[vii] Named for Dorothy Hodgkin, Rosalind Franklini and Kathleen Londsdale
[viii] Named for the father and son in the T.V. series *Only Fools and Horses*
[ix] Named for the model of the diffractometer. This was at one point named Peg after Margeret Thatcher, but the name didn't take.

X-ray radiation which does not interfere with the diffraction pattern). This is then placed in the X-ray beam, and a diffraction pattern is collected.

In general, procedures are carried out under a nitrogen stream. This benefits the experiment by decreasing the amount of thermal motion in the crystalline lattice, improving the quality of the models. Furthermore, the temperature of the crystals is prevented from increasing too drastically as a result of X-ray absorption - although in practice this is less of a problem for organic crystals as the amount of X-ray absorption is related to the number of electrons in the species. Some species prove to be thermally sensitive, and so the temperature must be increased to prevent thermal shock, which causes splintering of the crystals. Cases where this has taken place are detailed in Section 3.5.2.

3.5.1 Fixed Arrangements

Dot

A 007-HF High Flux Copper rotating anode source. A Saturn 944+ CCD detector was utilised. An Oxford Cryosystems Cobra device permits a cold nitrogen stream to a minimum temperature of 80K.

Ros

An FR-E+ SuperBright Molybdenum rotating anode X-Ray generator by Rigaku, utilising VariMax VHF optics. The detector is a Saturn 724+ CCD detector. The cold nitrogen stream is provided by an Oxford Crysosystems Cobra device to a minimum temperature of 80K.

Kat

An FR-E+ SuperBright Molybdenum rotating anode X-ray generator equipped with VariMax HF optics. A Saturn 724+ CCD detector is utilised. The cold nitrogen stream is provided by an Oxford Cryosystems Cobra device to a minimum temperature of 80K.

Spider

A Molybdenum sealed tube X-ray generator and a RAPID image plate detector system. An Oxford Cryosystems Cobra device provide the cold nitrogen stream, although in this instance the minimum temperature is 80K

Damien

A Molybdenum rotating anode source with a Bruker Nonius APEXII area detector. The incident beam was focused using 10cm confocal mirrors. The cold nitrogen stream is provided by an Oxford Cryostreams Cobra device to a minimum of 80K

Del Boy

A Molybdenum rotating anode source with a KappaCCD Roper area detector. A graphite monochromator was employed, and the nitrogen cold stream was provided by a Oxford Cryosystems Cobra device to a minimum of 80K.

Diamond Beamline I19

Diamond is a third generation Synchrotron light source. Radiation is generated by accelerating electrons around a ring at relativistic speeds. The acceleration, aided by special magnetic arrangements around the ring, permits the generation of a wide spectrum of radiation types. Several enclosures called beamlines are placed around the ring, of which I19 is one. I19 is configured to be a small-molecule X-ray crystallography procedural arrangement, using a Saturn 724+ diffractometer, and a CryoStream unit very similar in most regards to the units listed previously - with a minimum temperature of 80K.

3.5.2 Specific Procedural Details

Data was collected for this project as an ongoing exercise over many years. Some data collections were performed by undergraduate students, and this necessitated varying amounts of involvement from supervising staff and Ph. D. students. This information is given in Appendix B.

The grid in Figure 3.4 gives an illustration of the chemical space covered by the compounds that can hypothetically exist within the homologous series. The numbers across the top of the image indicate positions around the “aniline end” of the molecule which are filled by fluorine atoms. The numbers down the left side similarly, reflect the substituions for the “benzyl end” of the molecule. The numbers for the positions are shown in Figure 3.3. Numbers displayed in the grid reflect specific crystal structures, and are used through the rest of this thesis. Where more than one number appears in a square of the grid, this indicates a hydrate or a polymorphic structure.

A full list of compound structures which have been analysed in this thesis can be found in Appendix B.

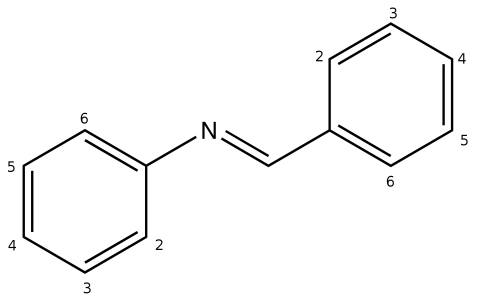


Figure 3.3: The fluorobenzanilide core, with positions on the phenyl rings numbered; the numbers correspond to the patterns indicated in Figure 3.4.

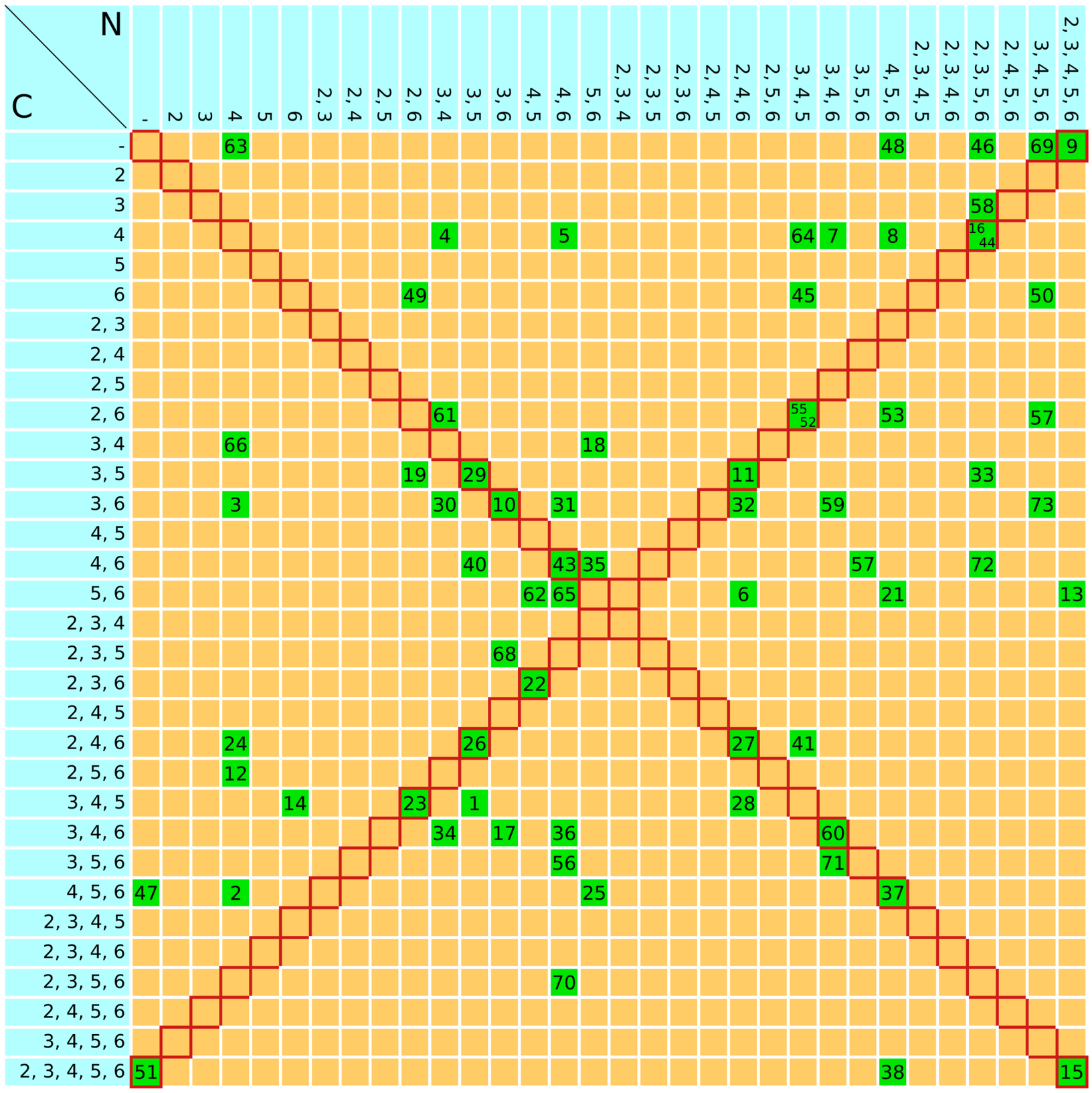


Figure 3.4: An overview of the possible fluorobenzanilide substitutions in a grid format
An illustration of the chemical space covered by the compounds that can hypothetically exist within the homologous series. The numbers across the top of the image indicate positions around the “aniline end” of the molecule which are filled by fluorine atoms. The numbers down the left side similarly, reflect the substituions for the “benzyl end” of the molecule. The numbers for the positions are shown in Figure 3.3. Numbers displayed in the grid reflect specific crystal structures, and are used through the rest of this thesis. Where more than one number appears in a square of the grid, this indicates a hydrate or a polymorphic structure.

3.6 Structure Solution and Refinement Details

The individual structure solution and refinement details of each procedure can be found in the .cif files of each crystal structure, stored in Appendix F, since in principle they do not alter the validity of the model. However, an overview of some pertinent features are given here.

Structure solution utilised one of a handful of programs: Olex2, ShelXS, ShelXD, ShelXT, Superflip or SIR92.

The solution methods in Olex2 and Superflip utilise Dual-Space charge flipping techniques.[206,207] ShelXD uses dual space according to a shake-and-bake algorithm.[204] SIR92, by contrast, uses primarily direct methods.[301] All of these programs provide the end user with an attempt at a full structure solution. ShelXS, by comparison, stops at the end of the phasing step, providing only a three dimensional electron density map. A number of ShelX versions have been used over the course of this project, utilising its implementation of direct methods.[205]

The reason for having such a large number of tools for performing what is essentially the same job is that each of the techniques implemented by the programs anecdotally perform better for different data sets. Such evidence has been matched by experience in this project, although no systematic review has taken place, nor has one been found in the literature.

For structure refinement, there are markedly fewer tools available, and in this project only ShelXL and Olex2 have been employed. Although each uses a different mathematical interpretation of the problem, both employ minima-seeking techniques to refine the crystal structure to a minimum for the R^2 value of the crystallographic model against the data by adjusting atomic locations and displacement parameters.[205,206]

Part III

Descriptor Libraries and Statistical Modelling

Chapter 4

Descriptor Libraries

4.1 Software Review

4.1.1 Overview

In both the cases of crystallographic and molecular descriptors it was necessary (for different reasons in each case) to conduct a thorough review of the available software which could calculate descriptors for crystalline and molecular systems respectively. In both cases the criteria for usefulness were strikingly similar:

Descriptor(s)

The problem of obtaining descriptors is frequently a two-part problem. In some software, potentially useful descriptors are calculated incidentally to the functioning of the software. The second part of the problem is to have the program output the data in a usable, computer processable format. The format is not entirely relevant, but standards such as xml, or even *de facto* standards such as csv are preferable.

Access to Source Code

In some cases software may calculate descriptors, but may not output the information, or it may calculate some description of a system which provides a shortcut to calculating a descriptor. In such cases, obtaining the underlying source code is required so that such tweaks or feature additions can be made.

Coding Language

Many older languages such as COBOL and tcl/tk have fallen out of fashion[i]. Other languages such as C and FORTRAN, seem to spring eternal (though thankfully it seems no more new code is produced in FORTRAN 77). The upshot of this is that some software may no longer be compilable by virtue of being written in a language for which interpreters or compilers no longer exist, or which was only available on a specific hardware system. In other cases, some languages do not scale well for large volumes of calculations, and this must be weighted against the costs of re-implementing the algorithms in a more suitable language.

System Requirements

Whilst, in theory, most code in languages such as C can be compiled for any system, some environmental considerations give way to system-specific code. Frequently, commercial software vendors are forced to consider the costs associated with generating code compatible with minority operating systems against their commercial benefit. Conversely, there are frequently costs associated with developing with popular commercial operating systems which open source software vendors are not always willing to undertake. This can lead to 'incompatibility by environment', whereby different software suites cannot be used in conjunction because they cannot be placed on the same system.

Maintainance

Crystallography, in particular, has had a rich history of software development (as evidenced in Appendix E). However, much of this software has fallen into disuse, either because research groups have moved on, or the software has been superseded. Often, the software may have been ahead of its time, and while the description may seem promising on first inspection, the reality can prove to be outdated code, no longer compatible with modern computing environments. Moreover, if code has become unmaintained - a phenomenon known as abandonware - it can be extremely hard to understand the methodology of the software, especially in cases of poorly documented code.

[i]COBOL actually is still very prevalent in the financial sector, where it makes up such a large volume of code that it has become difficult to replace - it was in fact part of the origin of the millennium bug[302]

Licensing

Much software that is produced is considered 'proprietary', in that someone owns the intellectual property embodied in the source code[ii]. In such cases, the source code is frequently unavailable for inspection as an added barrier to reverse-engineering of code. Even where source code is available, the licence conditions for the use thereof may still be severely restrictive (as is the case with Gaussian), and such matters must be checked before making use of source code. That being the case, certain facets of these project aim or aimed to be integrated with software produced by CCDC. Whilst the open source software movement has made much software available, it is often available under *copyleft* licences such as the GNU Public License. Such licences make software and source code freely available for re-use and distribution on the proviso that software which integrates functions, or works with the libraries which have been made available must be licenced either under the same licence or one which has been designated as compatible. Such viral licensing logically prohibits the use of such libraries that would integrate them with proprietary software such as the CCDCs code libraries. This is at least as restrictive as for other, proprietary software.

4.2 Molecular Descriptors

4.2.1 Available Software

The legacy of QSAR means that there are off-the-shelf packages available which can calculate a wealth of descriptors. Many of these are in constant development, and at the time of writing for this document, the current versions of the software represent very different entities to their incarnations when the assessment of this software was originally made. Only a handful of the currently maintained items of software are discussed here.

JChem

JChem is a commercial, closed source, commercial piece of software by ChemAxon. It is written in Java, and so can be made to run on most software environments, including a wide variety of open source distributions, OS X, and various versions of Microsoft Windows. The fact that it is written in Java and packaged appropriately also provides it with a relatively convenient API for use by external software. Comparatively, it offers a fair variety of molecular descriptors numbering in the hundreds, although this is not as many

[ii] Recognised as a work of literature in the United Kingdom

as some items of software for the price.[303]

RDKit

RDKit is a completely open-source software library specifically for cheminformatics use.[304] At the time this assessment as to a preferred item of software was being made, it did not produce many descriptors, although its Python and C++ bindings and extremely permissive software licence make it very appealing to develop with. More recently the software has made large amounts of progress with including increasing numbers of descriptors in the library, although still not as many as other software packages.

ARIANA.code

Superseded by Corina Symphony, this closed-source software was C++ coded and available for windows systems, calculating over 1500 different molecular descriptors.

Dragon Software

Dragon by Talete is concieved by Roberto Todeschini, who quite literally wrote the book on Molecular Descriptors.[191] It calculates circa 5000 different descriptors. It is closed-source, coded in an unknown language, but provides a very powerful command line API for use in scripting environments, making it very useful as a component in a cheminformatics pipeline. In addition, it is available on both Microsoft Windows and OS X, as well as a variety of open source systems, in particular Linux.

Dragon has within its framework a code which identifies each descriptor. Inside the results that the program generates, it does not specify the full names of the descriptor and so it becomes more convenient to notate the descriptors in this way throughout the report. The naming scheme is detailed extremely briefly here, but a full list of descriptor code names is made available by Talete.[305]

Whilst Dragon does suffer from some issues with respect to some missing descriptors - see the following note - it calculates by far the largest assortment of descriptors with a minimum of difficulty, and so was selected for use in this project.

Notable Omissions

In spite of the large variety of descriptors available in Dragon, there are some general omissions worth noting. The key fact that some descriptors in the collection are generally correlated with each other - for instance, total polar surface area is likely, at least in

some regard, to be correlated with the general size of a molecule. No software package surveyed offered a manner to normalise these against each other in a straightforward way, or even provided normalised values for obvious cases (TPSA normalised with respect to total surface area, for instance). This means that some information may be lost from models which might otherwise have been preserved.

4.3 Crystallographic Descriptors

4.3.1 Overview

As noted previously, Crystallography has had a rich history of software development borne primarily from necessity - the increase in productivity that the software has helped cultivate is evidenced simply by the number of crystallographic procedures that have been performed as a part of this work, whereas a mere half-century ago such a proposition would have been unthinkable.

Much of the software is poorly documented, and only briefly mentioned in the literature. The significant undertaking of examining each literature reference and often source code (spread across many languages) has been undertaken. The primary result of this is Appendix E, which is possibly the most up to date list of crystallographic software in (and out of) existence.

In detail, however, are presented works of specific interest which may prove useful either in this work or in the future for the calculation of crystallographic descriptors.

4.3.2 Property Calculators

VIBRATE! [sic] is abandonware which does not appear to be currently licenced, with no known access to source code.[306] XANADU is not known to be maintained, but is listed as having an unspecified open-source licence, and the code, written in FORTRAN77, is available.[307] Both pieces of software calculate Vibrational modes, which offer an interesting mechanism by which to capture information pertaining to the relationship between the symmetry of the molecular structure and the crystal structure. This was not ultimately used in the presented work, since XANADU's code is written in an old dialect of FORTRAN and is somewhat obfuscated. Nevertheless, this idea may prove to be of interest should a more readily implementable piece of software become available.

PIXEL is a piece of software currently maintained and developed by Angelo Gavezzotti. It is issued under a proprietary licence, though access to the FORTRAN code is granted

under licence.[220] The software aims to make the calculation of intermolecular interaction energies practicable on standard hardware (as opposed to a supercomputer).

The methodology of PIXEL is to take valence-only electron densities of a crystal system, calculated using an external package such as NWChem or Gaussian. The electron density values are treated as rigid. The physical space under consideration is partitioned into "pixels"[iii]. The interaction energies of the moieties in the system are calculated as the sum of the interaction between pixel-pixel nucleus-pixel and nucleus-nucleus pairs, which are 'owned' by each moiety in the system. Thus, the interaction between each moiety is the sum of the shared interactions between the spaces owned by the moieties under examination.

This method is reported to give a reasonable approximation of the actual lattice enthalpies for a relatively lightweight computation - it has also been used in crystal structure prediction.[19–22]

Whilst the ability to include these interaction energies in statistical models is of interest, it cannot be said that this is readily implementable. Firstly, it is not trivial to render the descriptor invariant. In addition, the source code is complex, and the licence is not amenable to software modification. As such, this software was not used in the course of this work.

4.3.3 Crystallographic Toolkits

PLATON is a very large crystallographic toolkit[308] currently maintained by Ton Spek. The source code is not available and the software is provided under a proprietary licence. It has a full suite of geometric calculations as a part of its framework. However, the output from this program is not highly parsable, being designed for human consumption rather than automated consumption by a computer. This a symptom of the time at which the software was produced - often the tables it generates were (and still are) included in crystal structure papers.

It also has a mechanism to calculate non-covalent contacts. Although such information is not easily digestable into an invariant descriptor this could prove useful if the difficulties with parsing the input are overcome. This information was found more readily from other sources, and so PLATON was not used.

The CCDC Toolbox is a vast array of code libraries which underpin the CCDC's crystallographic software suite. Until quite recently this was only available in C++. The library is extremely large and highly complex and developing with it, particularly as an

[iii] Although, strictly speaking, voxels would probably be a more appropriate term to capture the volumetric nature of the units

individual developer, is complicated and time consuming. Furthermore, it is only available as decompiled code under very specific agreements with CCDC.

CCDC have made the libraries accessible using a python API (application programming interface). This makes the library much easier to code with, owing to Python's deliberately flexible nature, and much faster to develop with, since one does not require the compilation of large segments of code repeatedly. It is also much simpler to use Python to interact with the external environment than C++, enabling more free interaction with external software components. However, the available functionality of the Python API is necessarily a subset, and so there is a trade off in terms of what can be achieved with this library.

4.3.4 Visual Crystal Structure Examination Software

TOPOS is a program whose sole focus is to examine the topology of crystal structures. It is currently maintained but the source code is unavailable. The software is provided under proprietary licence terms. Until quite recently it was primarily focused on inorganic compounds. Its primary use is not automated calculation as much as it is visualisation, though it does have some capacity to calculate graph sets in an automated fashion. Ultimately, whilst TOPOS is powerful, its current feature set does not exceed those of other available software packages outside of the functions it has available for inorganic systems.

CrystalExplorer is a piece of software maintained by the research group of Mark Spackman at the University of Western Australia. It is coded in C++ and maintained under a proprietary licence. CrystalExplorer primarily serves as a front-end to the open source software Tonto, although increasing amounts of calculation are done in CrystalExplorer itself. Tonto is coded in the language 'Foo', and is maintained by Dylan Jayatilaka under the LGPL licence - this permits the use of the code as a library but not for direct inclusion in other software.

One of the more unique features of CrystalExplorer is the ability to calculate values of distance between molecules, and generate 'fingerprints' of these interactions. Some development work was undertaken to develop code as a component of Tonto to output this information as raw data rather than a graphic, but this went unused in the presented research owing to the complexity involved in translating this information into an invariant descriptor.

Xpac is a piece of software which has been used in studies of crystallographic compounds previously.[187] The approach taken is that, given a crystallographic moiety constructed from an ordered set of points, it can find corresponding ordered sets of points in

several structures. When rendered, these give evidence for motifs which bear similarity in the group of crystal structures under inspection.

These similarities between pairs of systems are grouped into 0-dimensional (molecule only), 1-dimensional, 2-dimensional and 3-dimensional motifs. XPac also calculates a dissimilarity measure. However, the formula by which this is calculated has not been found in the literature. This similarity information can be used to generate a graphical representation of related motifs. Called Hasse diagrams, these representations cannot currently be automatically created and are time consuming to produce. In principle, the creation of an algorithm to generate graphs such as these is not trivial as such, but should be feasible. However, the closed source nature of Xpac renders this challenging.

Chapter 5

Feature Selection

5.1 Overview

As discussed in Section 3.1.2, there is a wealth of molecular descriptors available in off-the-shelf packages. In fact, for modelling the fluorobenzanilide collection of structures, there proves to be too many for sensible modelling of the sample size available.

As discussed previously, attempts to discern important descriptors for the purposes of crystallisation have lacked statistical rigour. As such, some statistical experiments were put together to attempt to restrict the descriptor space and illuminate descriptors which are important in governing crystal structure.

$Z' > 1$, and other structures with more than one species or otherwise distinct unit per unit cell complicate descriptor based modelling processes. In the simplest case of more than one molecule in the asymmetric unit, this is simply because the different conformations of the molecule resulting in different descriptor values for some descriptors. Multiple species systems will have radically different descriptor values.

There are many methods available for handling such systems- various averages (simple mean, geometric averages), and maximum and minimum values offer avenues for modelling such systems statistically. Where possible for such an exploratory study it was deemed best policy to attempt to avoid them where possible.

Disordered structures also present complications for the same reason as $Z' > 1$ structures, but present additional complications in terms of describing crystal structures. They are avoided completely for the purposes of the statistical models.

5.2 Correlation Analysis

5.2.1 Overview

Initially, full factorial designs were the method of statistical analysis being explored. With nearly 5000 descriptors available in Dragon, this brings to the fore that one would have to have 2^{5000} x-ray crystallographic results, selected at appropriate points in what is assumed to be an orthogonal descriptor space. This is impractical in terms of time constraints, even assuming that components could be found which crystallised at adequate points in the descriptor space. In addition, it is extremely unlikely that the descriptor space would prove to be orthogonal in 5000 dimensions.

Initial methods to try to lower the dimensions of the descriptor space focused on identifying descriptors which, for these purposes, were not orthogonal. Statistical methods rely on 'expert intuition' to identify likely descriptors for use in a model. Hitherto, most crystal structure prediction techniques have relied upon electron density calculations.[19–23] One can also consider the notion that molecular shape will also have some power to affect crystal structure- from which the idea of tectons appears in the literature.[49]

5.2.2 Sulphonamides

To that end, two families of compounds were examined. Firstly a group of sulphonamides, the data for which were originally collected by Susanne Huth.[309] These structures are all isostructural- and a large proportion of the compounds differ only in terms of one atomic position.

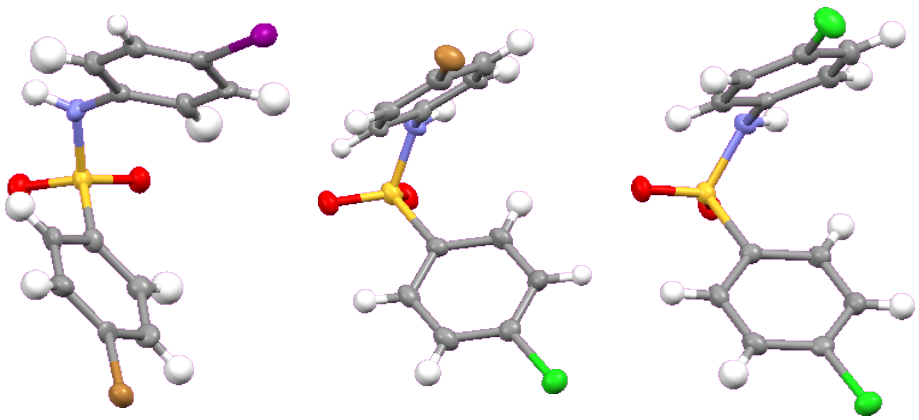


Figure 5.1: Examples of the three of the Sulphonamides examined by Susanne Huth

Intuitively, therefore, one might surmise that descriptors which change their values

across this series are less connected with geometry, given that the geometry is consistent amongst all species, and more connected with electron density distribution. In addition, descriptors which change in step may indicate themselves as being correlated, and thus non-orthogonal descriptors could be removed from the space. It was thought that this might be utilised to select important descriptors around which hypotheses for testing could be based.

The correlation tables resulting from the sulphonamides can be found in the Digital Appendix F. 600 Orthogonal descriptors could be found amongst the dataset, but there are caveats which undermined the reliability of these results. First and foremost, certainty values for the dataset were untenable because of the comparatively small sample with the large descriptor space. For each correlation comparison for each descriptor, the certainty value goes down because the probability of finding a correlation by chance goes up. With such a large number of descriptors, this renders the correlation results meaningless.

In addition, deriving meaning from even 600 descriptors would still prove to be a prohibitively time consuming process, and in any event, 600 descriptors is still too large a space to explore systematically given the time constraints. The experiment therefore fails to save us time even if we were to accept the results without the certainty values.

5.2.3 Glycine

At the same time and separately, Glycine was also considered. Glycine is ostensibly the most polymorphic crystal structure contained in the CSD, having 20 apparently distinct crystal structures in the CSD (it emerged that not all of these are true polymorphs, and some of the structures differ by only small distances and/or angles). Therefore, changes in descriptor values could be considered associated with changes in geometry rather than electron density, and are almost certainly more related to the variations in pressure used in the many studies of glycine than any other reaction condition. However, such a small number of true polymorphs would rapidly encounter the same issues discussed for the sulphonamides.

5.3 Co-Crystalline Experiment

5.3.1 Overview

A fellow Ph. D. student, Lucy Mapp, has been working on co-crystallisation experiments and wished to know what factors were important when forming co-crystals. In addition, she wished to examine systems which did not contain hydrogen bonds. It is reasonable to

pursue the hypothesis that factors important in forming crystalline systems with more than one molecular species present may also be important in governing the crystal structures of single-species crystals. Furthermore, there was overlap in the study owing to the search for non-hydrogen bonding systems, whilst the fluorobenzanilides had been selected as a test family for much the same reason. Therefore, a statistical experiment was set up in order to attempt to determine the governing factors for co-crystallisation, and hence a testing set of descriptors for single-species crystalline models where hydrogen bonding was not a governing factor.

Obviously, examining co-crystals necessitates handling more than one species per asymmetric unit. To maintain simplicity the co-crystals under examination were restricted to systems involving only two species. Any 'ordering' of the molecules in these structures would be arbitrary, so the descriptors for co-crystalline systems were re-defined as the maximum descriptor value from the values for each molecule in the co-crystal, and the difference between the values for each species in the co-crystal. This ensures that the values for both members of a co-crystalline system are included in the models, albeit obliquely, whilst removing any artefacts in the resultant models resulting from arbitrary ordering.

5.3.2 Method

Co-crystal Data

An initial subset of the CSD data set was found by Lucy Mapp by using ConQuest.[176,177] The search results were limited to structures which matched the following criteria:

- An R-factor of less than 5%- granting only the highest quality data set for the purposes of statistical study
- No disorder present in the structure- disordered structures further complicate the process of calculating descriptor values
- No chemical errors - only accurate and complete structures can be used for the statistical analysis at this time
- Not polymeric - the data of interest to the study is small molecule studies
- Not ionic - van der Waals forces are of primary interest, so ionic re-enforcement would serve to complicate the picture
- Not powder structures

- Organic compounds only

The first search limited the data set to compounds containing at least one carbon atom, in which the heaviest element was iodine[i]. This was then supplemented by another search for compounds which did not contain silicon. The final data set was the union of these two search results.

Unfortunately, technical constraints in ConQuest prevent searching directly for co-crystals and the elimination of solvent molecules, as well as the elimination of structures containing hydrogen bonds[ii]. Therefore, Lucy Mapp searched 'by hand' through the subset of data to pick out an arbitrary set of data which matched the criteria of being co-crystals without evident hydrogen bonds. This set of data is referenced in Appendix A.

The data for these crystals was then translated into the .cml format using the program OpenBabel.[310] The semantic structure of .cml, which includes the notion of 'molecules', permitted the removal of systems which had more than two species present, which was done using functions written in python[259] (See Appendix F).

The .cml format also retained all the information necessary to perform calculations in Dragon. By-hand searching still had to be done in order to remove hydrates and crystals containing solvent molecules. Ultimately, an arbitrary subset was found from those which matched the criteria. Ideally a random subset would have been used but this was not feasible given technological and time constraints. Once the searching was complete the .cml files were translated back into a .mol (.sybyl) file format, which Dragon requires for descriptor calculation.[311]

Dragon also requires the creation of an instruction file to calculate descriptors,[311] this was managed with a short php script.

Failed co-crystallisation Data

In order to prevent biasing of our findings, data had to be collected on experiments which failed to form co-crystals between two compounds. This information was found in a handful of papers in the relevant literature,[312–316] and the molecular structures were inputted into a computer by hand by Lucy Mapp. These structures were then energy minimised using the algorithms provided in ChemAxon's Marvin package to provide a three-dimensional conformation, necessary for the calculation of certain descriptors, and then descriptors were calculated for these molecular structures.

[i] This constraint was necessary as a result of a bug in the ConQuest software which gave rise to some metals appearing in spite of the organic-only constraint in earlier searches. In newer versions of the software this bug appears to be fixed

[ii] Or geometric constraints that resemble them

Problem of Experimental Design

The problem with a study of data gathered over several studies like this is that it is difficult to eliminate biases in the harvested data. For instance: if when graphed, the compound pairs which did not form co-crystals cluster together, it is difficult to ascertain whether that is because that region of that descriptor space precludes the formation of co-crystals, or if it is a coincidence that the data harvested focuses upon that region of descriptor space. Ultimately, one has to make the assumption that this latter scenario is not the case, as there is simply no way to tell without performing a fully designed experiment. Furthermore, technical limitation have prevented the removal of polymorphic co-crystalline compounds from the dataset. This means that there could remain biasing introduced by effective duplicates of those compounds in the dataset.

Processing of Data

Initial correlation analysis was done by eye, by plotting the various descriptors on multidimensional plots, whereby the descriptor value for each compound in the co-crystallisation attempt was placed along the x and y axes, and the colour of the plots indicated whether the attempt succeeded in creating a co-crystalline compound. The raw graphs are contained in the Appendix F. Unfortunately, such information proved difficult to transform directly into an intuitive model. However, the graphs did demonstrate patterns of data which lend themselves to Characterisation and Regression Trees (CARTs).

After creating these CARTs, it was noted that some descriptors may well be closely related, as such a correlation analysis was performed using the R programming language to assess how closely related the molecular descriptors were to each other. The Spearman correlation using pairwise complete observations was used to measure the degree of correlation between descriptors. The purpose of this excercise was not to remove descriptors before rebuilding the CARTs but to augment the cart creation algorithm. To use the information in this way would be invalid- examining such a large number of correlations would mean that little certainty could be understood of the correlations on such a small data set, as per the bonferonni approximation described earlier. Instead, descriptors which were closely correlated would not be permitted to form nodes which descend from each other in the decision tree. This assists in the prevention of silent overfitting by using closely related descriptors successively. This does not, therefore, decrease the number of descriptors supplied to the CART algorithm, nor does it prevent any number of descriptors being used in the ultimate CART.

This functionality was provided by the standard correlation libraries in R. Whilst

the P values can be calculated using the standard functionality in R, they cannot be calculated for anything other than fully complete data, which this sample was not, owing to the incalculability of some descriptors for some molecules. An implementation of the permutation algorithm was used to calculate the p values for the Spearman correlation:

```
1  permute <-function(dataMatrix, dataMatrix2, correlationMatrix, ncores) {
2          corrReps = replicate(
3            (10^3)/ncores,
4              cor(dataMatrix,
5                  dataMatrix2[sample(
6                                     1:nrow(dataMatrix2),
7                                     replace=T),
8                             ],
9                  use="pairwise.complete.obs",
10                 method="spearman"), simplify=F ); #create the resampled
                       correlation matrices
11      corrSums = matrix(data = 0, nrow=nrow(correlationMatrix), ncol=ncol(
            correlationMatrix)); #initialise a zero matrix the same size as the
            correlation matrix
12      #count the exceedences by correlation coefficients in the resampling
            matrices vs. the original correlation matrix
13      for(m in corrReps) {
14          corrSums = corrSums + (abs(m) > abs(correlationMatrix));
15      }
16      return(corrSums);
17 }
```

Listing 5.1: The permutation algorithm used in the calculation of p values for the Spearman Correlation

The key feature of note in this implementation of the algorithm are that it was designed to be run in a highly parallel environment - hence the variable 'ncores' which is used on line 3. R provides mechanisms to run tasks in this fashion using the 'parallel' package. This algorithm belongs to a class of problems described colloquially as 'embarrassingly parallel'; the replicate command in line 2 repeats the command provided as its second argument (line 4, 'cor', the correlation command in R) a number of times equal to its first argument (line 3). These runs are independent; the results of the first run do not depend on the results of the second, and so on. These can be done in parallel.

Therefore, this code is run in parallel on a number of cores equal to 'ncores' in order to speed execution - else running this many correlation calculations on such large matrices would prove to be prohibitively time consuming, and the apparatus for running the code in parallel is outside the scope of the snippet provided.

The correlations are drawn between the variables 'dataMatrix', and 'dataMatrix2'; in the case of self-correlating variables, these would be identical matrices. However, in the case of the co-crystalline systems, it proved to be algorithmically easier to calculate two matrices - one containing the differences between each molecular species in the co-crystal, the other the maximum. Hence, this algorithm is run three times: once for the self-correlations of the maximum values, once for the self-correlation of the difference values and once for the correlations between the two.

The sample function on line 6 randomises the order of the values in the second data matrix. The actual values of the correlations are stored in the 'corellationMatrix' variable. The result of the code in lines 2 to 10 is a list of correlation matrices which have been drawn between random arrangements of the data. Then, in lines 13 to 15, the number of times that these correlations are higher than the 'true' correlation value.

Once these values have been returned from across the different processors- the proportion of times that the randomised correlations exceed the value of the 'true' correlations corresponds to the probability of a Type I error, in short, the matrix of p values, the algorithm for which is illustrated below:

```
1  pValue <- function(dataMatrix, correlationMatrix, cluster, ncores,
       dataMatrix2=NULL) {
2      if(is.null(dataMatrix2)) {
3          dataMatrix2 = dataMatrix;
4      }
5      corrSumList = clusterCall(cluster, permute, dataMatrix, dataMatrix2,
           correlationMatrix, ncores);
6      corrSums = Reduce("+", corrSumList);
7      pValues = corrSums/nrow(correlationMatrix); #the proportion of
           exceedences gives us the matrix of pValues
8      return(pValues);
9  }
```

Listing 5.2: The function to calculate p values from the correlation matrix exceedence counts

Chapter 6

A New Crystallographic Descriptor

6.1 A New Graphical Descriptor

6.1.1 The Spectral Radius

As mentioned previously, descriptors for crystalline systems are few and far between. One of the descriptors that does exist - graph set descriptors - is not ideal for use in statistical models because of its many-to-many relationship with crystal structures, as discussed in Section 3.1.3. Nevertheless, the notion that the arrangement of the intermolecular interactions in a crystalline lattice govern the material properties and formation is a key theory which has yet to be unequivocably proven.

Graphs can be expressed in a format which does not rely on labels for their description. Connectivity matrices are frequently used to describe molecular systems in terms of their number of bonds. In general, such representations are made ignoring hydrogen atoms, which for most organic molecules can be considered implicit.

For instance, one of the pentafluorobenzanilide compounds could be expressed thus:

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	N_1	F_1	F_2	F_3	F_4	F_5
C_1	0	1.5	0	0	0	1.5	0	0	0	0	0	1.5	0	1	0	0	0	0
C_2	1.5	0	1.5	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
C_3	0	1.5	0	1.5	0	0	0	0	0	0	0	0	0	0	0	1	0	0
C_4	0	0	1.5	0	1.5	0	0	0	0	0	0	0	0	0	0	0	1	0
C_5	0	0	0	1.5	0	1.5	0	0	0	0	0	0	0	0	0	0	0	1
C_6	1.5	0	0	0	1.5	0	1	0	0	0	0	0	0	0	0	0	0	0
C_7	0	0	0	0	0	1	0	0	0	0	0	0	2	0	0	0	0	0
C_8	0	0	0	0	0	0	0	0	1.5	0	0	1.5	1	0	0	0	0	0
C_9	0	0	0	0	0	0	0	1.5	0	1.5	0	0	0	0	0	0	0	0
C_{10}	0	0	0	0	0	0	0	0	1.5	0	1.5	0	0	0	0	0	0	0
C_{11}	0	0	0	0	0	0	0	0	0	1.5	0	1.5	0	0	0	0	0	0
C_{12}	1.5	0	0	0	0	0	0	0	0	0	1.5	0	0	0	0	0	0	0
N_1	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0
F_1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F_2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F_3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F_4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F_5	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

(6.1)

Note that double bonds are counted as a value of two in the matrix.

Note, however, that such a representation is not invariant to the order of the atoms examined. One can come up with canonical orderings of atoms such as the Cahn-Ingold-Prelog rules, however, when examining crystalline systems where the repeating units are frequently identical, this does not apply as a method. In either case, the matrices are still not invariant to the number of nodes (atoms in the case of a single compound) in the graph.

However, graphs have a property known as the *spectral radius* which can reduce the character of a graph to a single variable, and is calculated by solving for lambda in the equation:

$$det[\mathbf{A} - \lambda I] = 0 \tag{6.2}$$

Where I is the identity matrix and $\mathbf{A}$ is the connectivity matrix under examination.

λ can take many values for any given graph. The spectral radius of the graph is the

largest of these values. This is more recognisable to physical scientists in the context of a molecule as being the principal eigenvalue.

The spectral radius describes the nature of the graph, but it is a lossy reduction; one cannot reconstruct the graph given the principal eigenvalue alone, but two graphs which are similar should have similar spectral radii.

Whilst the spectral radius is invariant to the order in which the nodes which build up the graph are chosen, it is not invariant with respect to the number of nodes in the graph. Therefore, if one compares spectral radii, it is important to ensure that the graphs being analysed have the same number of nodes.

6.1.2 Connectivity Graphs in Crystalline Systems

While the bonding in molecular entities is well established as a graphical depiction, in crystalline systems, there is no established method of drawing a graph from a crystal structure. There are recognised intermolecular interactions as mentioned previously. Nevertheless, most crystallographic papers retro-justify the existence of such interactions based upon close contacts in a crystal lattice. Such close contacts between two entities are defined in the software Mercury as being points at which the centres of two atoms are closer than the sum of the van-der-Waals radii of the two atoms.

In the present system, this permits us to use the molecular entities in a crystal structure as being our nodes in the graph, and the connections are the short contacts. What is needed is a method to select the entities in a crystal structure to use to build up a matrix. Whichever method is chosen, it must be able to apply to all systems, and must do so in a way that is consistent.

One way is to choose the n-closest molecular entities to some kernel. For instance the molecular entities in the asymmetric unit. Such a collection of n-closest entities is generally known as a *packing shell*. n is some semi-arbitrary integer. It can be considered semi-arbitrary, because the value of n is not itself important and is subject only to the constraint that there must be the same n for any set of crystal structures being compared. However, if one is comparing a crystal structure whose asymmetric unit contains three molecules, and another whose asymmetric unit contains two molecules, a graph built from only three molecules will actually lose information in the case of the two molecule graph-some connectivity will be missing. As a result, the value of n, when dealing with many crystal structures, should be the lowest common multiple of the number of entities in the asymmetric unit.

A problem with choosing the n closest entities is that the concept of closeness requires

quantification. For anisotropic entities, such as flat molecules like the fluorobenzanilides, measuring whether something is close is not trivial. At least four definitions for this concept can be deduced:

1. The distance between the central positions in the molecules
2. The distance between the centre of gravities of the molecules
3. The distance between the van der Waals surfaces of the molecules
4. The lowest potential energy, as used in the Pixel method by Gavezzotti et al.[220]

The first of these is the easiest to implement, but by no means the most accurate. The third is conceptually easy to approximate by using the centre points of the atoms in the molecules, and calculating the nearest of these. However, as shall be shown in Subsection 6.1.3, this could not be implemented straightforwardly at this time.

6.1.3 Concrete Implementation

A concrete implementation of the above descriptor was instantiated using the python version of the CCDC toolkit. Python solutions are frequently more straightforward to implement than solutions in C++, owing to the flexible type system and interpreted nature of the language. However, in this instance this necessitated some trade-off in terms of conceptual precision. Of the four concepts of closeness discussed previously, the python library has only the capacity to generate the first. The approximation to the third method of closeness calculation could be implemented by the libraries involved in the C++ libraries, but the interfaces have not yet been made compatible with the Python libraries.

On the other hand, this does have the advantage of being the same calculation which takes place internally inside the visualisation program Mercury,[230] which permits a visual analysis of the descriptor's interpretation- in particular for any outliers that may arise.

```
1       crystals = []
2       for f in glob(sys.argv[1]):
3           reader = io.CrystalReader(f, 'cif')
4           crystals = crystals + [c for c in reader]
5       zPrimes = [c.z_value for c in crystals]
6       packingLcm = lcm(*zPrimes)
7       if packingLcm < 1:
8           raise Exception('A structure has a z value of <1, and this is not
                valid with this program')
```

Listing 6.1: The initial phase of the graph descriptor algorithm

The first phase of the algorithm reads the crystal structures in from the cif file using the reader implementation provided by the CCDC toolbox library. It then collects the Z' prime values of the crystal structures into a list. It should be noted that this is problematic for systems where the Z' is not defined in the .cif file; the library defaults this value to zero rather than calculating it as a property. It should be also noted that because this algorithm depends on Z', it only applies to single-species crystal structures.

The lowest common multiple of the Z' values is calculated using a recursive algorithm:

```
1  def lcm(arg1, arg2, *args): #calculates the lowest common multiple of an
       arbitrarily long series of numbers.
2      baseLcm = (arg1*arg2)/gcd(arg1, arg2)
3      if len(args) < 1:
4          return baseLcm
5      else:
6          return lcm(baseLcm, args[0], *args[1:])
```

Listing 6.2: The calculation of the lowest common multiple; this will be the number of molecules in the minimum common packing shell (provided it is greater than 16)

Wherein, the 'gcd' function is a standard library function in python which calculates the greatest common denominator.

```
1      while packingLcm < 16: #16 is a widely accepted value for packing shells
2          packingLcm = packingLcm * i
3          i = i + 1
```

Listing 6.3: The actual number of molecules to pack is calculated by multiplying the lowest common multiple by incrementing integers

In order to generate a true *packing shell*, it is necessary to somehow scale the value of the lowest common multiple to a value that will include sufficient crystallographic entities to produce a three-dimensional shell. In general, 16 closest molecules is the default for Mercury to generate packing shells for a system, and so this has been used as a first-draft approximation for this code as a threshold value.

```
1      shells = []
2      for crystal in crystals:
3          shells.append(crystal.packing_shell(int(packingLcm)))
```

Listing 6.4: The packing shells for calculating the novel descriptor are generated

The packing shells are then generated by making a call to a function for that purpose found in the CCDC toolkit. Note that the number of molecules is coerced to an integer value - effectively rounding it down. This prevents difficulties in comprehension of partial molecules.

This action is performed once per crystalline species provided, yielding a list of packing shells.

```
1       for shell in shells:
2           components = [comp for comp in shell.components]
3           matrix = []
4           for i in range(0, len(components)):
5               matrixRow = []
6               for j in range(0, len(components)):
7                   matrixRow.append(0)
8                   if i != j:
9                       for atom1 in components[i].atoms:
10                          for atom2 in components[j].atoms:
11                              if iad(atom1, atom2) < (vdwRadii[atom1.
                                    atomic_symbol] + vdwRadii[atom2.atomic_symbol
                                    ]):
12                                  matrixRow[j] = matrixRow[j] + 1
13              matrix.append(matrixRow)
14          contactMatrices.append(matrix)
```

Listing 6.5: The close-contact matrices are calculated

The close-contact matrices are then calculated. Note that for the moment this has had to be performed using code which is not based in the CCDC toolkit as, although this functionality is in some sense present, it was not readily amenable to this task. Therefore this code checks the inter-atomic distances between the centres of every pair of atoms in every pair of molecules possible in each packing shell. The inter-atomic distance for each pair of atoms is compared, and if it is smaller than the sum of the van-der-Waals radii of the atomic pair, then it is counted as a close contact. The corresponding value of the matrix is incremented for the molecular pair, and so on.

The inter-atomic spacing is calculated by the function, and yields a measurement in picometers:

```
1   def iad(atom1, atom2): #calculates the distance between two atoms in
        picometers (assuming 3d coords are done in angstroms)
2       dx = abs(atom1.coordinates.x - atom2.coordinates.x)
3       dy = abs(atom1.coordinates.y - atom2.coordinates.y)
4       dz = abs(atom1.coordinates.z - atom2.coordinates.z)
5       return 100*sqrt(pow(dx, 2) + pow(dy, 2) + pow(dz, 2))
```

Listing 6.6: The inter-atomic distance calculation

And the atomic radii are stored in a very sparse python dictionary- note that the radii present in this dictionary are only those which are relevant to the presented work. They

too, are measured in picometers, and are taken from work by Bondi[317]

```
1  def iad(atom1, atom2): #calculates the distance between two atoms in
       picometers (assuming 3d coords are done in angstroms)
2      dx = abs(atom1.coordinates.x - atom2.coordinates.x)
3      dy = abs(atom1.coordinates.y - atom2.coordinates.y)
4      dz = abs(atom1.coordinates.z - atom2.coordinates.z)
5      return 100*sqrt(pow(dx, 2) + pow(dy, 2) + pow(dz, 2))
```

Listing 6.7: The relevant van-der-Waals Radii

The eigenvalues are then calculated and output, using functions from the numpy linear algebra module.[318]

```
1      for m in contactMatrices:
2          characteristicValues.append(max(eig(mat(m))[0]))
```

Listing 6.8: The calculation of the Eigenvalues- these are the graph descriptor values

6.2 Melting points of Fluorobenzanilides

During the course of the presented work Liam Oliver, a project student working in the Coles group at Southampton University, collected melting points of a subset of the Fluorobenzanilides. During the course of his project he selected the fluorobenzanilides which, if overlayed, would have a fully complementary overlap (see Figure 6.1). Furthermore, the subset was restricted to systems with R-factor was lower than 10%, and were not disordered, in order to ease analysis.

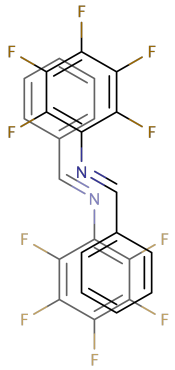


Figure 6.1: A hypothesised examples of 'complementary overlap' structure re-illustrated

The melting points were collected using a Mettler Toledo FP82 hot stage and an FP90 controller. The observation of the melting points were specifically made by capturing the

images of the melting process using a camera attached to a computer running software called studio capture. The melting points used in the analysis are the midpoint of the melting range.

In order to assess the utility of the packing network descriptor described in the previous section, correlations were drawn between the values obtained for the network descriptor and the values established for the melting points, using the in-built functions in the programming language R.

Part IV

Results and Discussion

Chapter 7

Fluorobenzanilide Crystal Structure Results

7.1 Overview

As described in Section 1.7, a hypothesis was drawn up that considered the interaction of hydrogen and fluorine to be a favourable interaction, even perpendicular to the rings in the fluorobenzanilides. Therefore, the fluorobenzanilides are arranged in three categories; those whose structures when lain in a stack displayed complementary overlap between fluorine and hydrogen atom positions, those which consistently displayed like-to-like positioning of such atoms when lain in a stack, and those which had varying levels of complementary and clashing positions. In that naïve hypothesis, it was considered that those structures with complementary layers would be more likely to form such stacks, whilst those that did not, would not.

As the naïve hypothesis is based on the stacking structure, only those structures, and structures which are otherwise of more general crystallographic interest (polymorphs, isostructures, hydrates, and disordered structures) will be discussed in detail. Other structures will be summarised briefly.

In previous presentations of a subset of this data, some structures were described in terms of a ribbon motif;[319] these being constructed from the side-to-side interactions of the molecular species. Whilst such a broad categorisation does have some merit, the notion of a ribbon actually covered too much variety to be useful, and so this construct is not presented directly here.

7.2 Xpac Analysis

The analysis of the crystal structures using X-Pac yielded more information than could reasonably handled; a large number of 1- and 2- dimensional constructs were observed, and the relationships between all of the crystal structures via these constructs can be found in a human and machine readable format in the digital appendix, though interpretation of the large volume of data is elusive.

The problem of interpreting such a large volume of crystal structures in this way stems from the software itself. XPac is capable of locating the constructs in a pairwise fashion between crystal structures, but cannot as yet follow up with the meta-analysis of which constructs are shared between more than one pair of structures. Neither the source code nor a specification for the files created by the program have been released, and so adding this capacity would require a complete reimplementation of the original program.

This being the case, it is still possible to use the program to glean information about isostructural lattices, as these are few enough to be extracted from the data relatively straightforwardly.

7.3 Structures of Special Interest

7.3.1 Polymorphs

There is only one polymorphic compound in this data set, which is compounds 16 and 44.

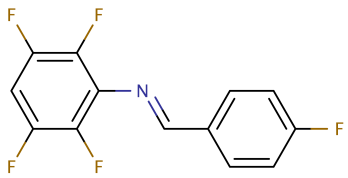
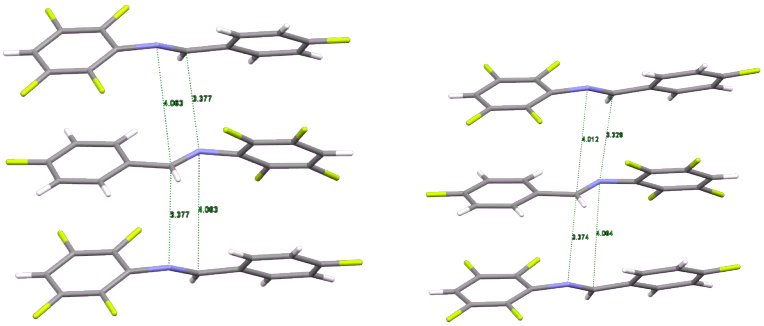


Figure 7.1: A reiteration of the molecular structure of compounds 16 and 44

A cursory examination of the crystal structures might give the impression of two structures the same, if one did not have ready access to the lattice perameters. This is not least because the two structures still feature the same base construct; that of the head-to-tail stack (which is discussed in more depth later in this chapter). In each structure, members of the stack are also similarly spaced.

However, the first dissimilarity can be noted with the torsion angles between the rings. In compound 16 we find only one measure, 49.77°. In compound 44, with less symmetry present, we find three different torsion angles; 41.94°, 39.28°, and 37.97°.



(a) A stack in compound 16 (b) A stack in compound 44

Figure 7.2: A comparison of distances a common motif in compounds 16 and 44. The distances displayed are in Ångströms, and seem to relate to the symmetry of the fluorine substitution pattern.

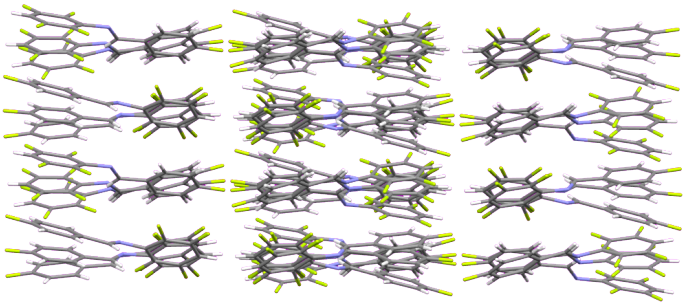


Figure 7.3: A transverse view of the stacks in compound 44, demonstrating the angulation of the molecules in the stacks

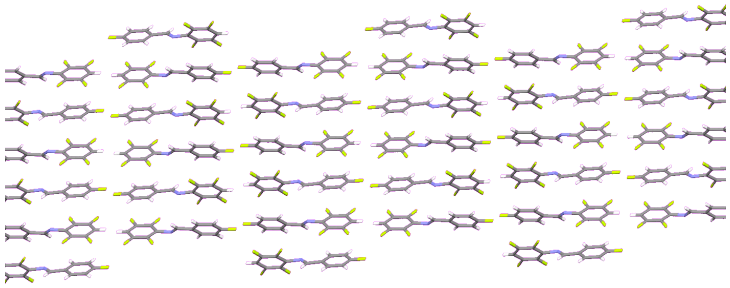


Figure 7.4: A transverse view of the stacks in compound 16, illustrating the lack of angle between molecules in adjacent stacks

The key difference in the crystal structures however, is in how these constructs are arranged in the larger structure. A lateral viewing of the crystal structures, as in figures 7.4 and 7.3, reveals that the stacks in compound 44 are tilted with respect to each other.

To see two crystal structures of the same compound which seem to be formed of different arrangements of similar intermolecular constructs lends support to the hypothesis of Kitaigorodskii that crystal structures build in this way. It also credence to the approach taken in the XPac software of looking for supramolecular constructs of different dimensionalities.

7.3.2 Isostructures

There are several families of isostructure in this series of compounds. Some are very straightforward cases, and others less so. In all cases, the existence of the isostructures creates difficulties for the underlying assumption at the outset of this thesis, that is that the atomic constituents of the fluoraniline compounds have some directing effect on the structures, as will be illustrated by re-iterating the compound structures as a part of the analysis.

Isostructures in homologous series are also an area in which XPac excels. Of the six isostructural systems presented, only two were detected using a by-eye inspection. Xpac also managed to detect one visually-self evident isostructure which has a radically different unit cell, which gives a good rationalisation for the symmetry independent approach that Xpac takes.

Compounds 36 and 43

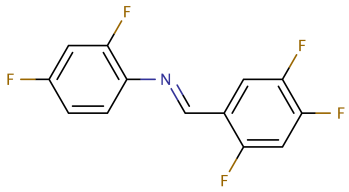


Figure 7.5: The molecular structure of compound 36

This pair of crystal structures constitute arguably the simplest case of isostructurality in this dataset. The two compounds have extremely similar unit cells (differences are all less than 0.2Å and 3° at worst), and distances between common points between molecules in the unit cell differ by less than 0.1Å. In addition, the difference in torsion angles between the rings is less than 2°. The substitution patterns on the molecules are different in only

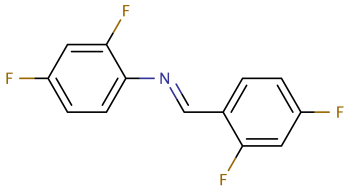


Figure 7.6: The molecular structure of compound 43

one posision, and this does not challenge the underlying hypothesis of QSAR that similar molecular structures will form similar crystal structures.

Compounds 44 and 57

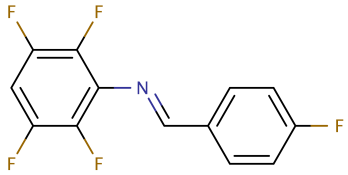


Figure 7.7: The molecular structure of compound 44

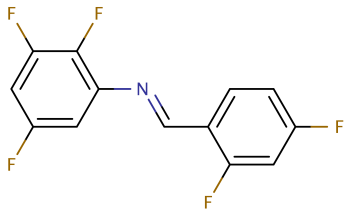


Figure 7.8: The molecular structure of compound 57

This is another textbook isostructure. The unit cells and symmetry are the same (<0.5Ådifference for any dimension, <1°for any angle), and the distances between corresponding pairs of atoms in close molecules tend to be different by less than 0.1Å. The torsion angles between corresponding molecules in the lattice differ by less than 1°. The molecules themselves differ only in one position of substitution per ring, whilst themselves being substantially different from the molecules in the previous subset which are isostructural to one another.

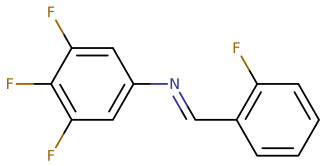


Figure 7.9: The molecular structure of compound 14

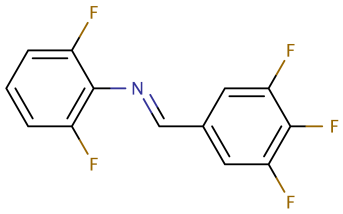


Figure 7.10: The molecular structure of compound 23

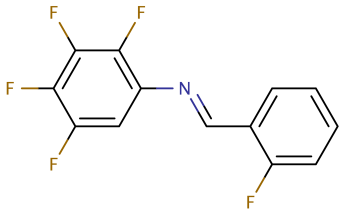


Figure 7.11: The molecular structure of compound 50

Compounds 14, 23, and 50

This subset represent a simple case of isostructurality in this dataset. That aside, the structures do illustrate the difficulties in asigning a quantitative meaning to the term isostructural. Although in each case, the unit cell measurements are very close (within a precision of about 0.2Å and 0.5°, and the central nitrogen/carbon linkers are spaced similarly, the torsion angles between the rings in the molecules vary (max 44.44°, min 33.19°). Whether this is of import owes itself to the purposes for which one is using the data. If one is interested in mechanical properties, then it only counts to the extent that these differences manifest in those outcomes. To the purpose in this thesis, identifying directing effects on crystal structure, it can be argued that it is of more significance that the molecules as a whole arrange into extremely similar three dimensional structures.

In this case it comes as no surprise that the structures are similar, since the molecular substitution patterns are similar. However, the differences in the torsion angles do not seem to be well explained by the substitution patterns in a manner which can be identified

without a much deeper electron density analysis.

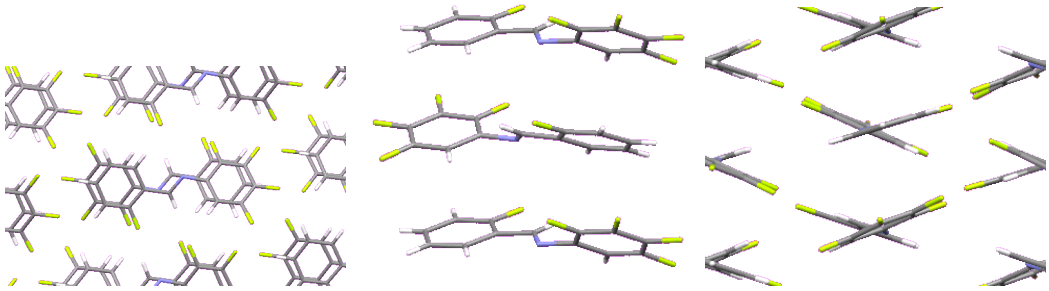


Figure 7.12: An illustration of the packing structure in compound 50

Compounds 11, 16 and 46

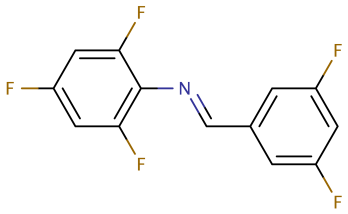


Figure 7.13: The molecular structure of compound 11

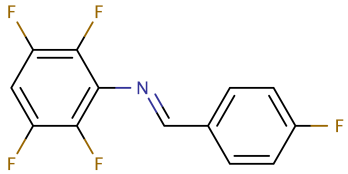


Figure 7.14: The molecular structure of compound 16

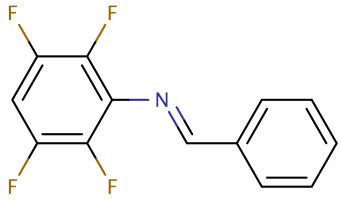


Figure 7.15: The molecular structure of compound 46

These three structures have very similar unit cells, and visually similar arrangements. However, the substitution patterns are quite different. A detailed look at the geometric measurements of the crystal structures however, reveals that there are subtle differences in these crystal structures. The differences in the structures fit with the underlying hypothesis of this thesis, that fluorine has weak interactions which can direct crystal structure formation, albeit in a minor fashion. The structures nevertheless possess the same characteristic arrangements, and are similar enough to be considered isostructral.

Figure 7.16 illustrates the structure of compound 16, and hence, the basic formation of the structure for all of the compounds in this group. In compound 46 we see that the c-axis becomes nearly a full Ångström shorter than in compound 16. Meanwhile, in compound 11 the β angle is 3 degrees wider.

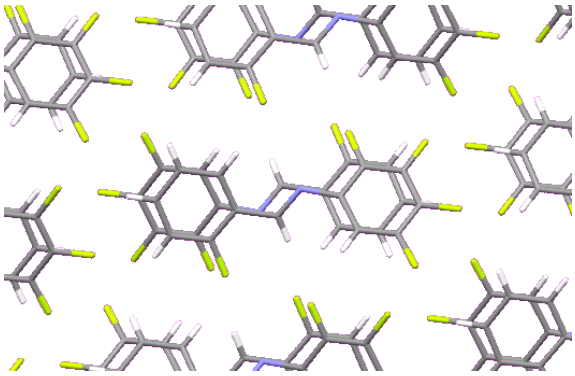


Figure 7.16: An illustration of the packing structure in compounds 11, 16 and 46

Examining the molecular structure of compound 11, we can see that when arranged in this way, adjacent molecules will have alternating hydrogen-fluorine interactions along their long edge. Neither of the other two molecules have this. Compound 16, by contrast, does have two hydrogen-fluorine interactions for each ring, but they are not alternating. If one considers the electron withdrawing effect of fluorine on the ring, one can see that the two fluorine atoms on adjacent positions of the ring in compound 16 'compete' for electron density. Hence, they do not generate the same attractive effect as seen in compound 11, where we observe the wider angle, and hence closer molecules.

The argument for the rather smaller difference in intermolecular distances is much less apparent. However, it may be to do with the fact that it has one ring which remains completely unsubstituted, and therefore has a very slightly smaller size when these two tail-ends face each other in the three dimensional arrangement.

Compounds 2, 24 and 47

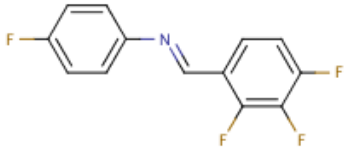


Figure 7.17: The molecular structure of compound 2

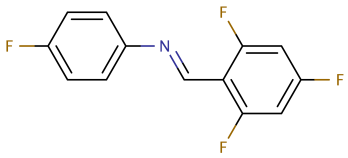


Figure 7.18: The molecular structure of compound 24

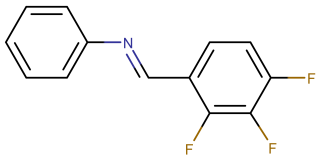


Figure 7.19: The molecular structure of compound 47

This triad of molecules represents a similar and complementary case; two of the crystal structures are a very straightforward case of polymorphism. Compounds 2 and 47 share unit cell measures (within 0.3Å and 2° per dimension). The torsion angle between the rings only varies by 1.8°.

Compound 24 does not differ in terms of the torsion angle between the rings, but its unit cell dimensions are a full Ångström shorter in the a and b axes, and a full angstrom longer in the c axis. It is interesting to note that for the second time in two sets of related crystal structures, this dimensional change is associated with a molecule which has a 2,4,6 substitution pattern on the ring. It does not appear self-evident that the difference in substitution patterns here change the number of favourable interactions possible in the demonstrated crystal structures. It seems logical therefore that the different substitution patterns change something of the character of the interactions. For instance, the alternating hydrogen/fluorine substition pattern may give a distribution of electrons which render the area of the molecule near the hydrogen atom a greater $\delta+$ charge, whilst the fluorines

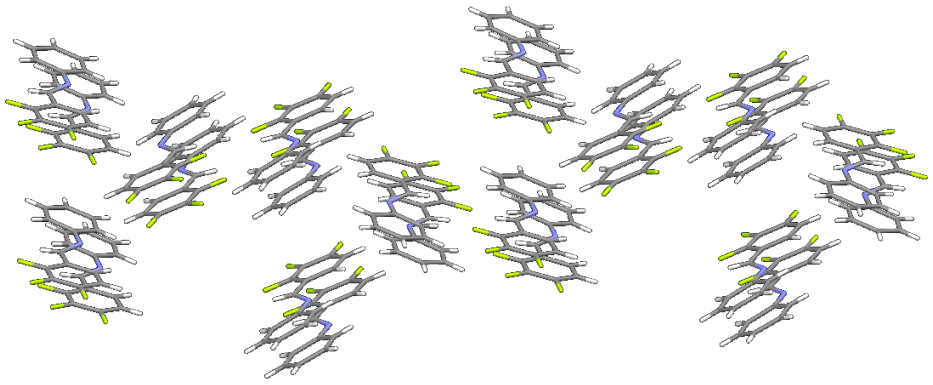


Figure 7.20: An illustration of the packing structure in compounds 2, 24 and 47

in turn are able to withdraw more electron density from the ring when not placed adjacent to each other, becoming more δ-. This would create stronger interactions between adjacent molecules and explain the shortened unit cell lengths (and associated changes in intermolecular distances).

Compounds 4, 5, 7, 8 and 48

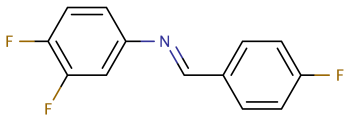


Figure 7.21: The molecular structure of compound 4

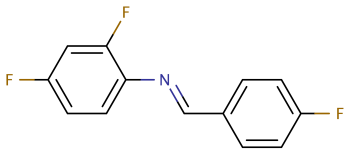


Figure 7.22: The molecular structure of compound 5

This group of compounds is simultaneously a simple and yet complex case of isostructurality. It is simple in that all of the unit cells are extremely similar (<0.5Å difference across the set, <0.01° difference in angle[i]. It is complex because the only similarity in

[i]Though this is of little relevance since all but one were solved in an orthorhombic space group.

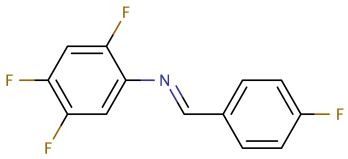


Figure 7.23: The molecular structure of compound 7

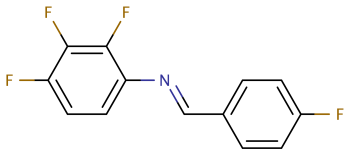


Figure 7.24: The molecular structure of compound 8

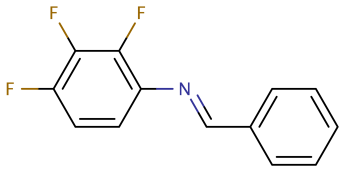


Figure 7.25: The molecular structure of compound 48

fluorine substitution between the molecules is in the 4 position on the 'benzyl' end of the molecule. In addition, the close contacts in each system seem to be different owing to this differing arrangement of atoms. It is tempting to argue that this is a 'default' structure which arises from the shape of the compound, but this is difficult to justify as it is different from the structures seen in either the fully substituted or non-substituted members of this homologous series.

One point of interest with this set of structures is that compound 4 was not selected by XPac as being a member. This is likely owed to the fact that the crystal structure as represented in the data file is enantiomeric to the others in the series. However, the data contains only light elements and was collected using a molybdenum source, and therefore it is not possible to determine the absolute structure. XPac does not seem to check for this information however, and so fails to correct for this information.

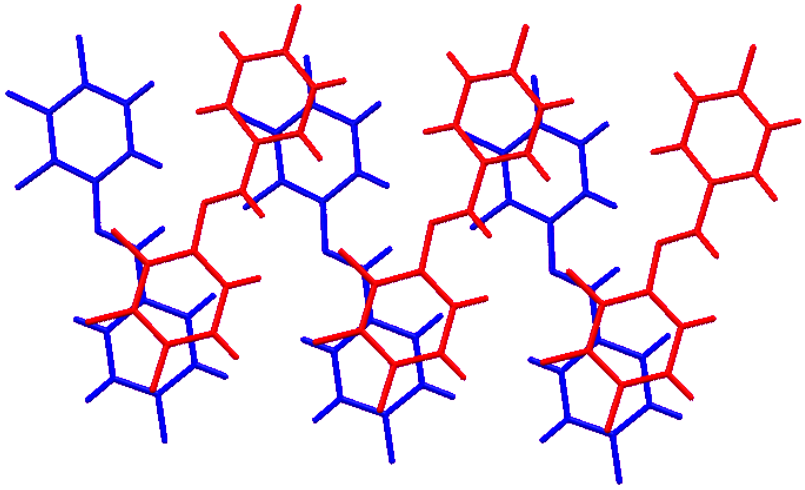


Figure 7.26: A depiction of the layers in compound 8, with the layers coloured red and blue rather than by atom to contrast the depths

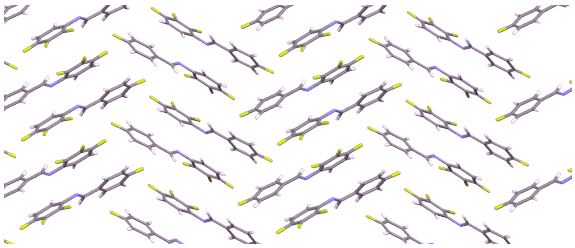


Figure 7.27: A transverse view of the layers in compound 8

Compounds 22, 45, 51, 52, 54, 61, and 71

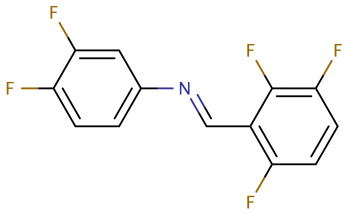


Figure 7.28: The molecular structure of compound 22

This large group of compounds is the most complex set of isostructural compounds in the homologous series.

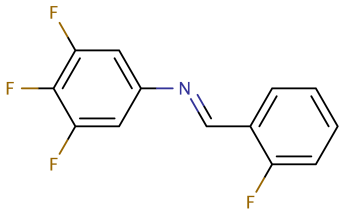


Figure 7.29: The molecular structure of compound 45

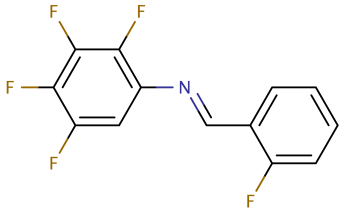


Figure 7.30: The molecular structure of compound 51

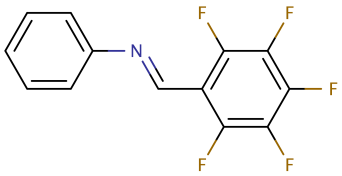


Figure 7.31: The molecular structure of compound 52

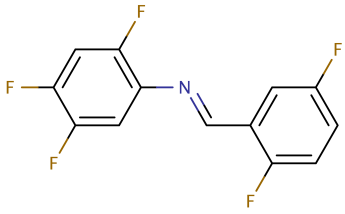


Figure 7.32: The molecular structure of compound 54

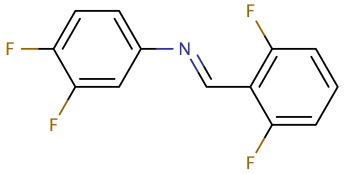


Figure 7.33: The molecular structure of compound 61

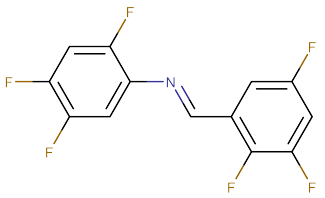


Figure 7.34: The molecular structure of compound 71

Compound 54 has a completely different unit cell compared to the others in this subset, though this is more to do with data quality than some absolute truth of the crystal structure, and this structure should be recollected and solved.

In order to successfully explain the differences in the unit cell measurements in the other species here, it is necessary to change our frame of reference. It is common practice in crystallography to make the a axis the shortest axis by definition. The problem with this in this context is that for compound 45, this results in a different orientation of the unit cell relative to the molecules it contains.

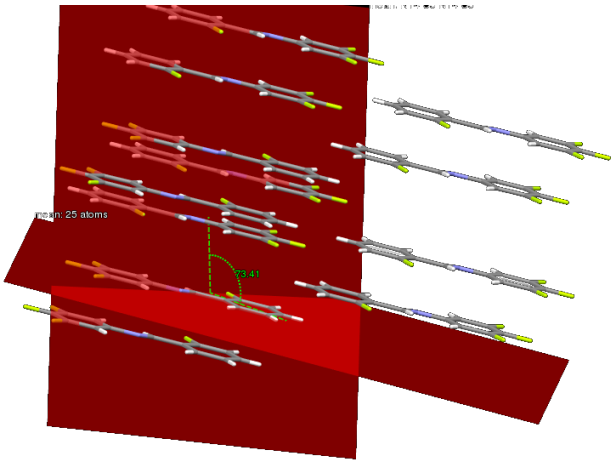


Figure 7.35: An illustration of the offset measurement in compound 45; one plane is the average plane of hte molecule, the other of the stack centers measured using the nitrogen and carbon linker. The angle between the planes indicates the extent of the offset.

Therefore, for the remainder of this subset, the cell axes will be discussed in terms of their relation to the stacked molecules in this group - the axes which is paralell with the stack, the axis which is paralell to the long edge of the molecule and the axis which is paralell to the short edge of the molecule. This will allow an unambiguous discussion.

Compound 51, completely fluorinated at one end and unfluorinated at the other, has

the shortest axis parallel to the stack. By contrast, the offset between molecules vertically is much lower. Consider 7.35, where we measure the angle between the mean plane of the molecule and the plane created between the 4 linking atoms in the center of two molecules[ii]. The angle seen here is 73.41°. The angle in compound 51, by contrast is 82.62, much closer to a perpendicular arrangement. Compound 61, which also has a shortened stack-parallel axis, has an angle of 76.62°. The other systems, which do not have shortened axes generally shallower angles, with the exception of compound 54, whose angle is 76.51, whose axes are not aligned with the stack in the same way as other systems.

This is strongly implicative of the quadrupolar interaction taking place between the stacks. While it is often (but not universally) taken to be the case that π-π interactions perpendicular to the molecule are favourable, the results here imply that is the case only for alternating stacks of electron rich and electron poor moieties.

7.3.3 Hydrates

Hydrates are not unusual in crystalline systems. Some compounds do not form except for their hydrates. They are however, numerically unusual in this homologous series. The incorporation of water into Compounds 41 and 55 is not surprising, given that the nitrogen atom central to the series is an obvious candidate for hydrogen bonding with the small water molecules. There is nothing specific about these two molecules which lends itself to incorporation of water however, and indeed compound 55 is the hydrated form of Compound 52. Whether these systems preferentially form hydrates or their pure crystal is a matter left for further investigation. If other compounds in this system can be induced to form hydrates, it may be possible to build hypotheses based on the non-hydrates and then test them in the hydrated system.

[ii] This is an alternative measure to overcoming the challenge of defining the same absolute planes in two different crystal structures which are defined using different unit cells.

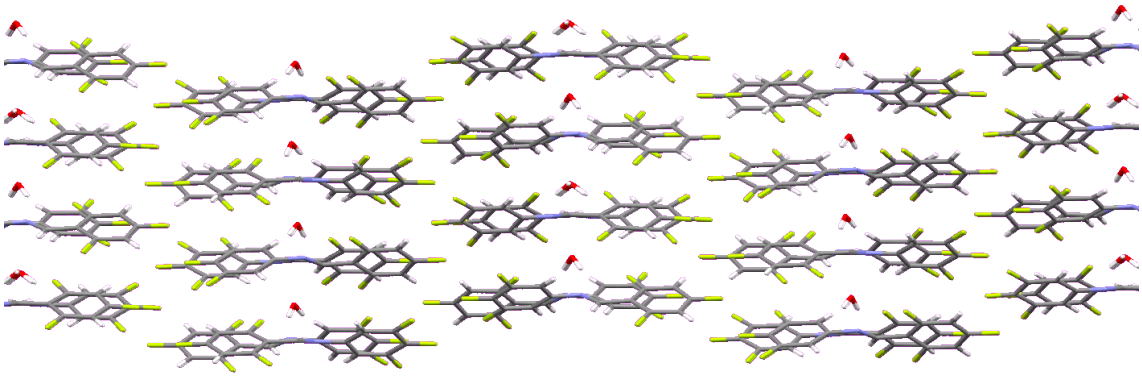


Figure 7.36: A transverse view of the stacks in compound 41

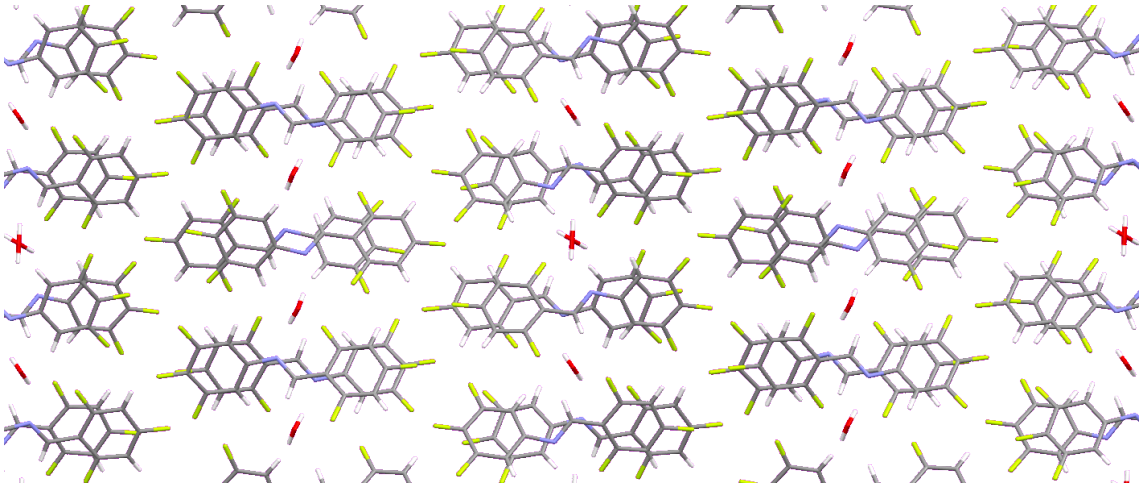


Figure 7.37: A top-down view of the stacks in compound 41

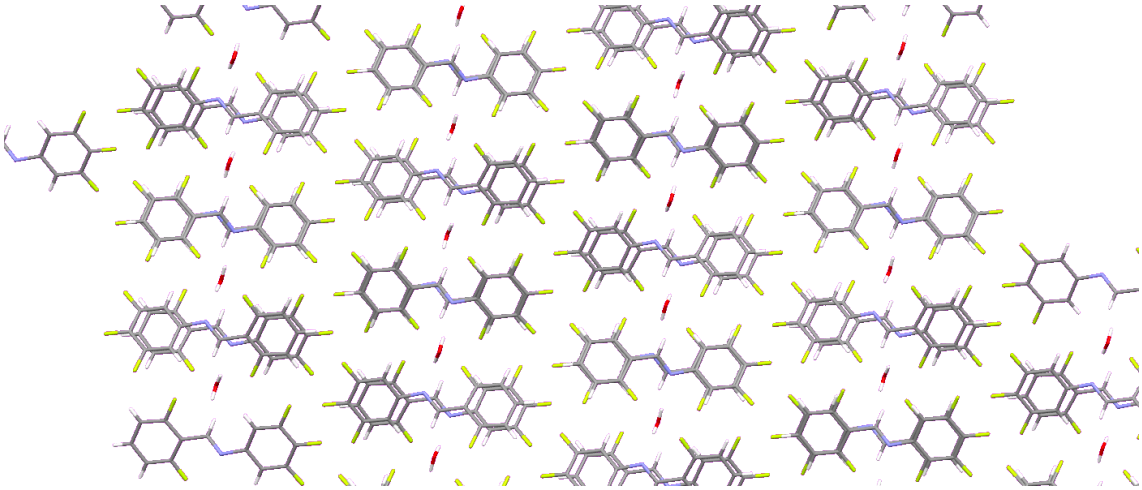


Figure 7.38: A top-down view onto the stacks from Compound 55

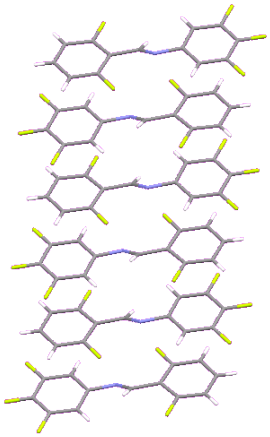


Figure 7.39: A single stack in isolation from Compound 55

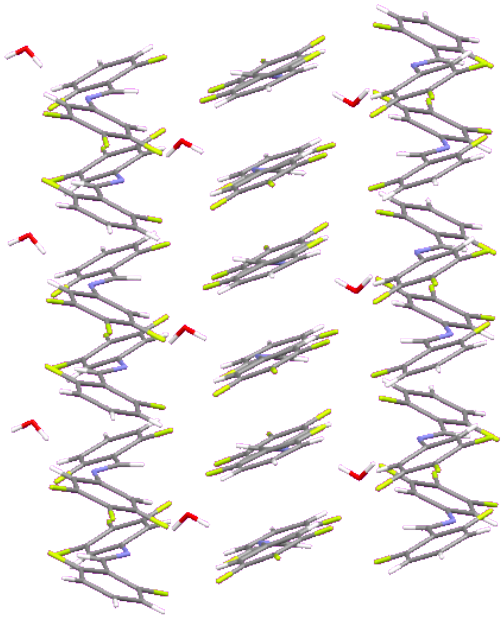


Figure 7.40: Demonstration of the two different stacking motifs in Compound 55

7.3.4 Disordered Structures

There are five compounds which show varying degrees of disorder in the data set. Disorder arises in systems where multiple different arrangements are present irregularly throughout the extended crystal lattice. Because of the fact that X-ray diffraction solutions are averages both through space and time, it is difficult to assert thermal or static disorder unambiguously. owing to the nature of the system under examination (a small, rigid molecule) and the fact that there are not large spaces within which the molecules in these lattices can move, it is reasonable to assume in this case that the disoder arises from an irregular static arrangement throughout the lattice.

The disorder that is seen in four of the five the demonstrated cases can be easily rationalised by observing the near-symmetry of the molecules about which they are disordered. In compounds 15 and 27 we see that both halves of the molecule have similar substitution patterns, so that when the molecules are flipped, any properties dictated by the ring systems will remain unchanged. This is similar for compounds 34 and 68, though the substitution pattern varies in one position of that ring.

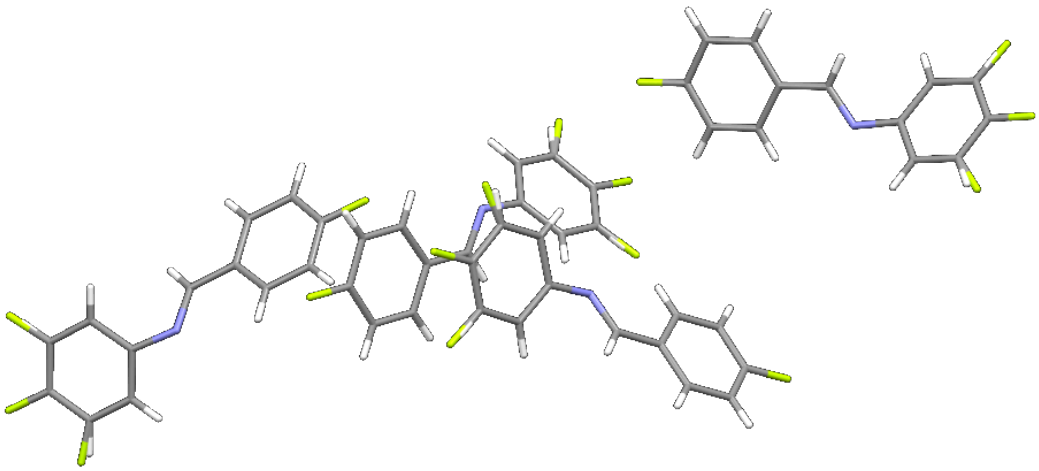


Figure 7.41: An illustration of compound 4

For compound 4, we see a case of isomerism in the molecule; the aniline ring has formed a racemate around the bond with restricted rotation in the center, and the two isomers have formed a disordered co-crystal. This, incidentally, is why compound 4 has two locations in the grid in Section 3.5.2.

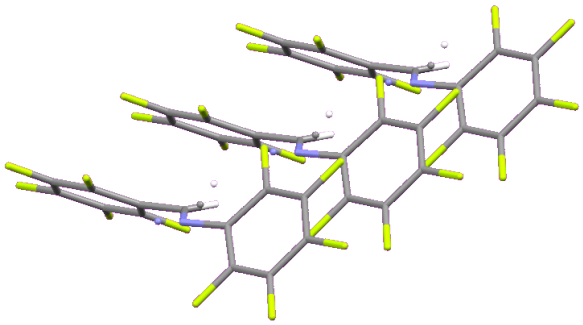


Figure 7.42: An illustration of compound 15

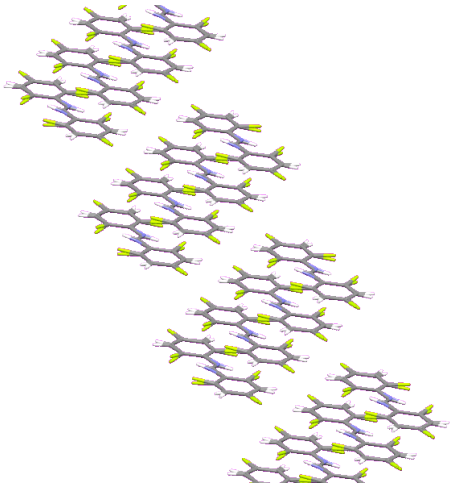


Figure 7.43: A view of the structure of compound 27

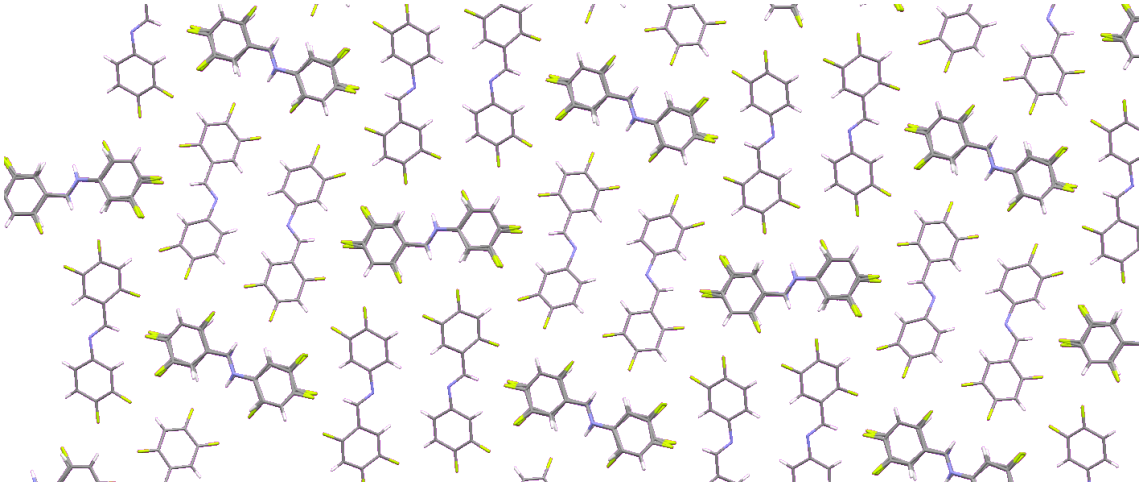


Figure 7.44: A top-down view of the stacks in compound 34

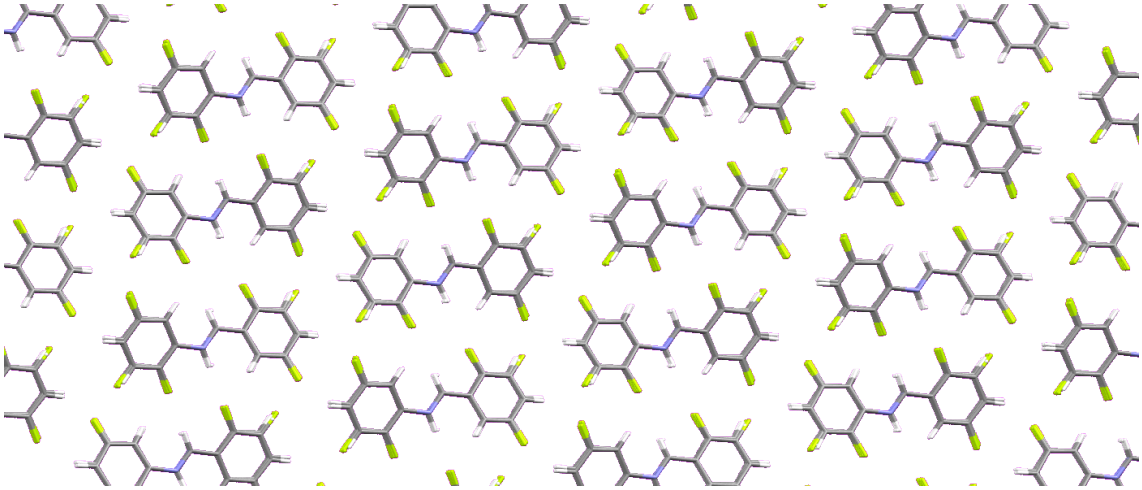


Figure 7.45: A top-down view of the stacks in Compound 68

7.4 Stacked Crystal Structures

The stacked crystal structures were the first observed in this homologous series of compounds. As such, the naïve hypothesis was created with these structures in mind. Many compounds in the homologous series form a variety of stacked structures. Broadly, there are two classes of stacked structure.

7.4.1 Head-To-Tail Stacked Structures

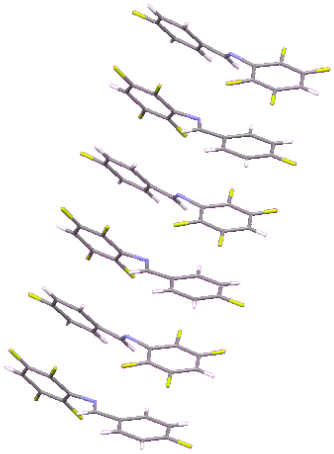


Figure 7.46: A visualisation of the head-to-tail stack, in this case from compound 16

The stacks which are described as head to tail are as depicted in figure 7.46. Alternating molecules are rotated 180 degrees from each other.

Of the 22 compounds which form this type of stack, 20 form visually similar structures[iii] an exemplar of which is depicted in figure 7.46. Although they are visually similar, there are geometric alterations which can be as simple as a changing in the offset of the stack (as discussed in section 7.3.2. It becomes quite impossible to intuit justifications for the variety of differences in structure manually - though there are options for automating this in the future (see further work, section 9.

An example which illustrates some of the difficulty in this task however, is the pair of compounds 9 and 51. Compound 51 has been described earlier in this document, and a number of other structures in this dataset were isostructural to it. Curiously, one which was not among that group was compound 9.

[iii] For reference, compounds 9, 11, 13, 16, 19, 22, 23, 26, 44, 45, 46, 50, 51, 52, 54, 56, 61, 68, 71, 73

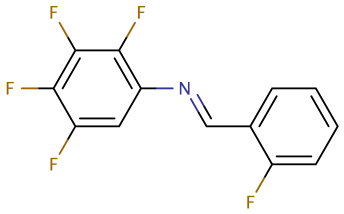


Figure 7.47: The molecular structure of compound 50

Electronically speaking the two molecules are identical save for the respective location of a lone pair, which is not delocalised into the pi system, in relation to the pentafluorinated ring. And yet, within the two compounds, despite the same symmetry and same unit cell lengths, we see different unit cell angles (by 8° in the worst case). Further, the torsion in the rings are very different- 44.78° in compound 9 and only 3.20°in compound 51.

One trend emerges, however: in the complimentary stacks, the molecules in adjacent stacks tend to sit in a plane The exceptions are compounds 73 and 44, which express a very slight torsion angle of 9.94° and 11.01° respectively.

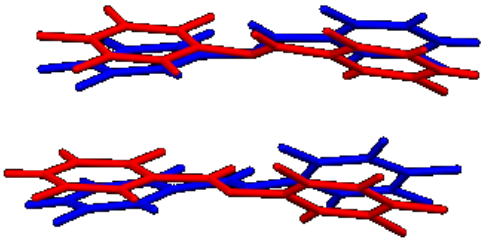


Figure 7.48: An illustration of the angling between molecules in adjacent stacks.

There exists three other structure types which only have one exemplar molecule each.

Compound 1

The structure of compound 1 differs from that in the others in that the stacks themselves are organised in a staggered fasion, as in figure 7.49. There is a plausible hypothesis for this which relates back to the hydrogen-fluorine paired interactions seen in compound 11.

This construct is observable again in this compound structure between the adjacent, staggered, stacks. Moreover, this is the only compound in the collected data set which is

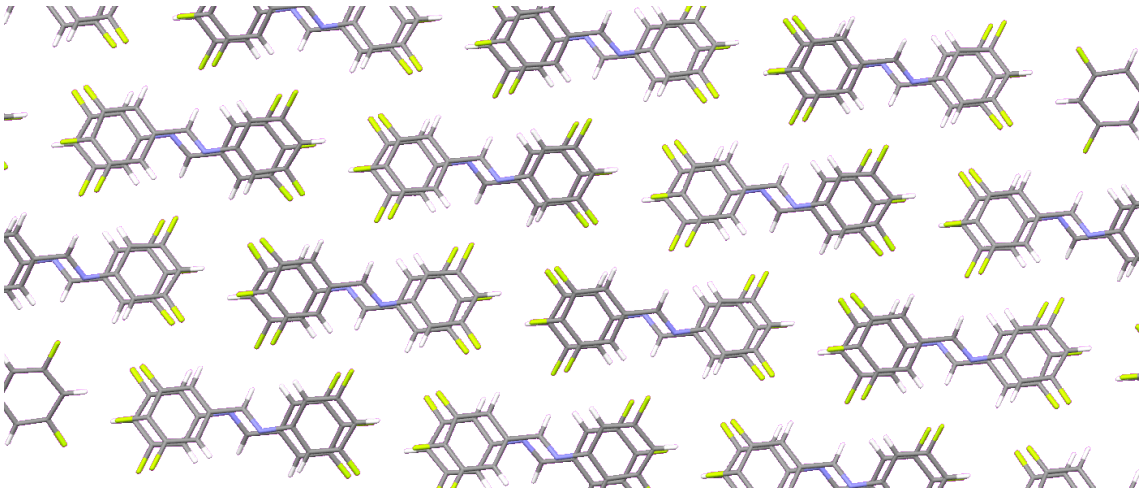


Figure 7.49: The staggered arrangement of stacks in compound 1

capable of arranging in this particular way to produce this favourable arrangement. That there is a causal arrangement is not proven by this, but it is strongly indicative that this may be a genuinely structure directing construct, especially when coupled with the details of compound 11.

Compound 13

Compound 13 is unusual in that the compound notionally stacks in that the molecules form isolated columns - but each column contains a large lateral offset between alternating molecules. Examining the arrangement of this offset suggests that this may be an arrangement designed to best fit quadrupolar moments which have been distorted by the substitution patterns - but without proper electron density experiments or quantum mechanical examination this remains speculative.

Compound 53

Compound 53 forms a unique crystal structure among the whole dataset. In principle, it forms head-to-tail stacks. These are organised into head to tail rows, and then between those rows is a completely different crystal motif, which is much more closely resemblant of the brick-wall pattern that will be discussed later.

Summary

It is interesting to note in terms of the naïve hypothesis, that all of the structures with complementary overlap can be found in this group of compounds, while none of those with perfect clashing arose. This result can be analysed statistically, and that will be shown in section 8

7.4.2 Head-To-Head Stacked Crystal Structures

In this set of compounds we see a far wider variety of structure types. Arguably of most interest in relation to the naïve hypothesis is the fact that these compounds exist in a system where fluorine atoms stack above fluorine atoms. It does not necessarily follow from this that the naïve hypothesis will produce incorrect predictions - a more rigorous assessment of which is given later. It does however tend to indicate that the mechanistic understanding of those predictions is incorrect, or else that there is some other component to these crystal structures that allows them to overcome energy costs that the mechanistic understanding implies.

Compounds 30, 49, and 60

The first type of structure seen in this subset is similar to the modal secondary arrangement of stacks seen in the head-to-tail stacking formations. Stacks are arranged in regular rows which extend parallel ad. infinitum. Each has a very distinct unit cell, although two of the three compounds that comprise this group share a space group. Unlike the structures which the top-down view gives the impression of being similar, the arrangement of the molecules within the stacks is angled with respect to one another.

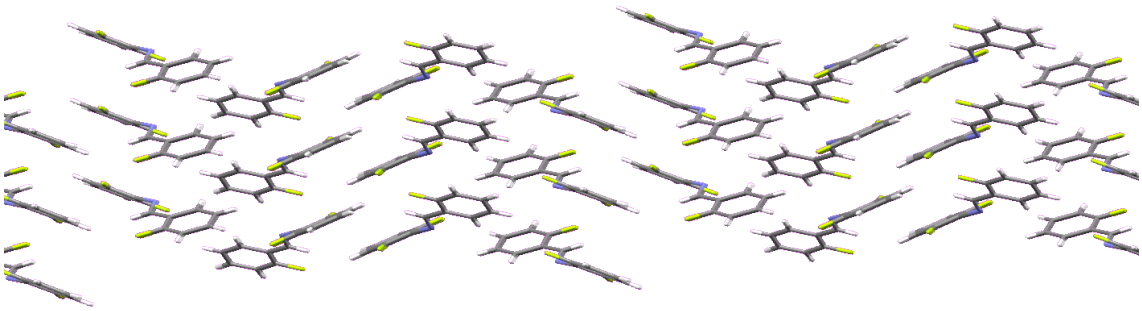


Figure 7.50: The angled stacks of compound 49

Compounds 31, 36, 43, and 65

We also see a subset where the rows of stacks are staggered. Still, we see the same slanting of the molecules in the crystal structure. The only common interactions between adjacent stacks in this subset is that there is some short contact between a hydrogen and fluorine atom. But each is located differently, although this gives rise to a similar motif in terms of spacial arrangement.

Compounds 6, 21, 29, 32, and 72

In these structures, the stacks arrange head to tail, but with an angle between each stack. These angles vary from 156.85° in Compound 21 to 133° for compound 6. The commonality seems to be that this arrangement maximises, H-F contacts, but this is difficult to empirically confirm, as it is difficult to demonstrate that there are any other alternative arrangements which are physically viable.

Compounds 40, 64 and 66

These three compounds have their stacks paired, and then arranged in a brickwork pattern. The pairing of the stacks is actually inverted for the two structures, however, note that compound 40 is paired such that the formation of the spiral of in figure 7.51

Compound 64, by contrast, is seemingly mediated by a longer range connectivity graph, that is nevertheless composed of hydrogen and nitrogen contacts.

It is informative to view compound 34 alongside the other structures presented here.

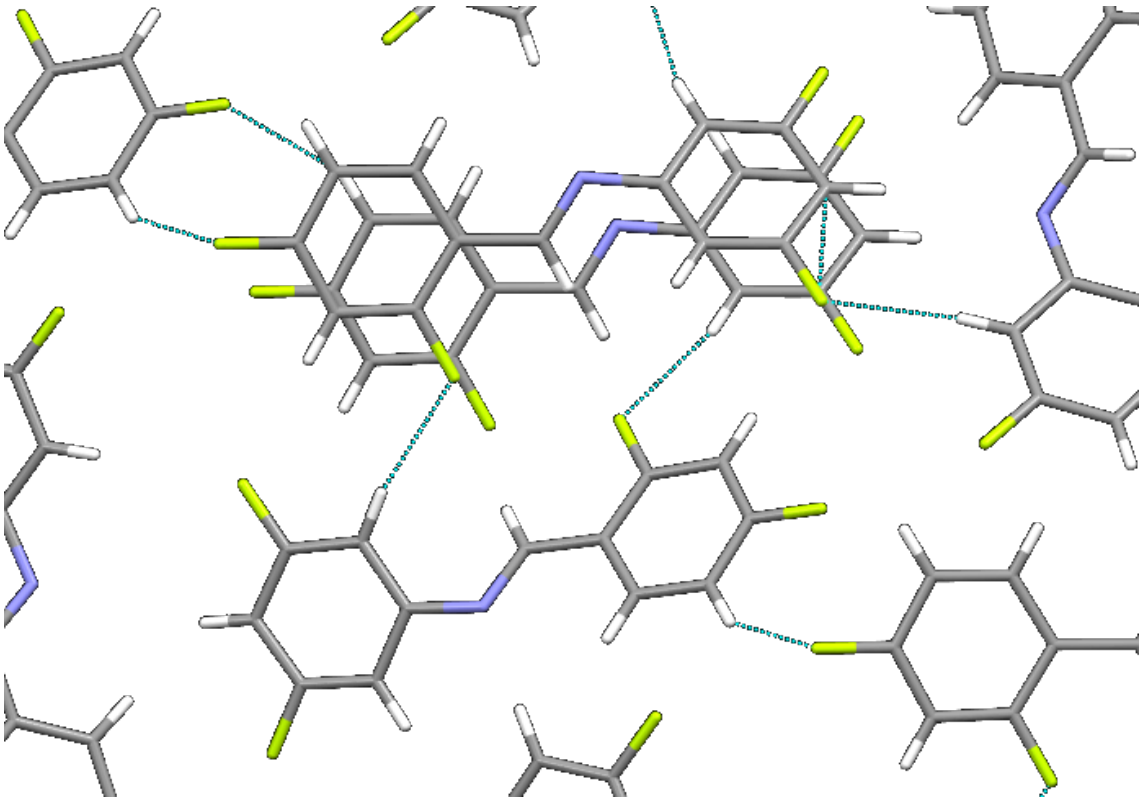


Figure 7.51: An illustration of the close contacts in compound 40

Although the structure is very different overall, note that the paired stacks reappear, and are not mediated by connectivity between the pairs. It may be that the relative increase in fluorine-hydrogen contacts from the head of the disordered molecule compensates for this.

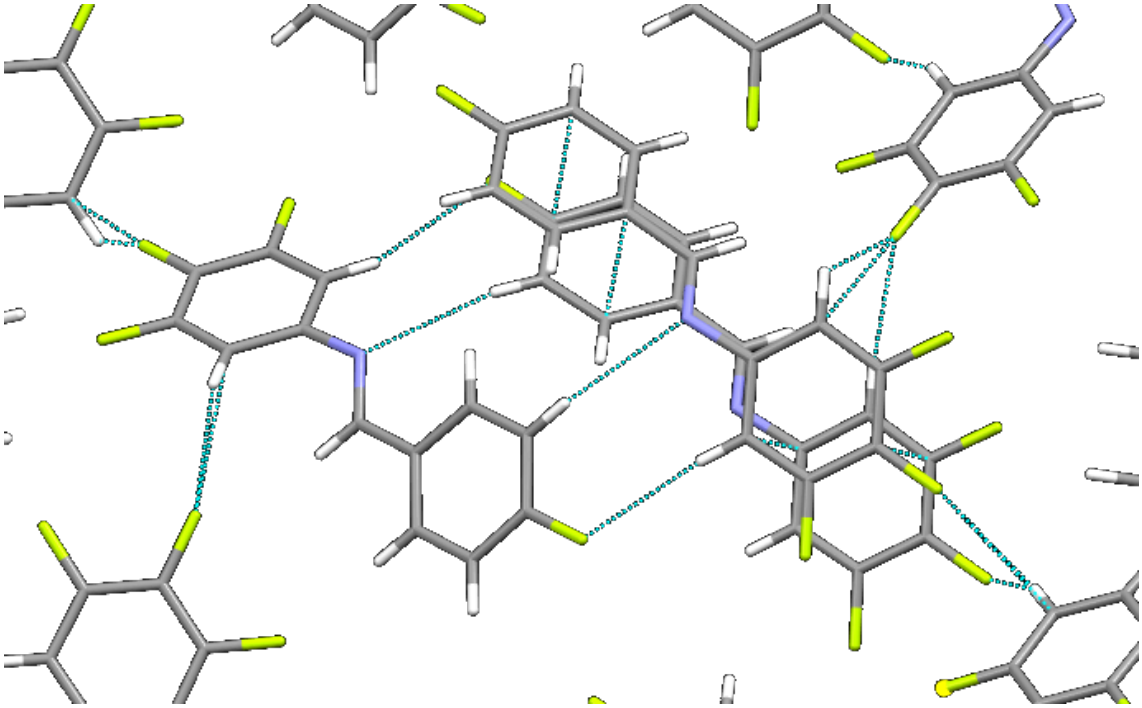


Figure 7.52: An illustration of the close contacts in Compound 64

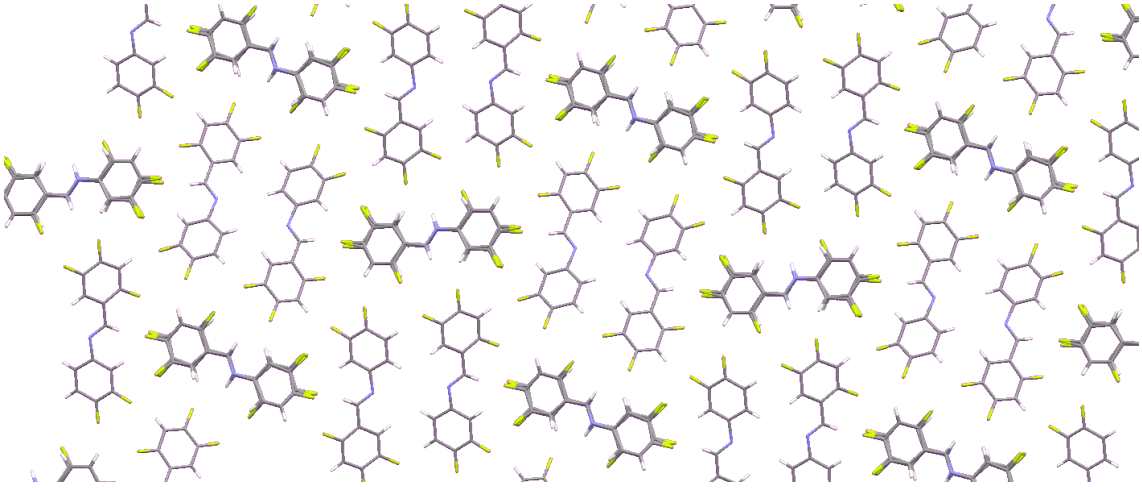


Figure 7.53: An illustration of the stacking motif in Compound 34

Compound 28

Compound 28 possesses a unique arrangement in that it is head-to-head stacked, but the molecules are also flipped alternately about their central axis.

Summary

In terms of the original hypothesis, this set of structures contain three (compound 60, 29, 43) of the "fully clashing" compounds. In point of fact, though, all of these compounds form perfectly clashed structures owing to their head-to-head arrangement. All do so, however, with a slant in the stacks- which is consistent with the quadrupolar moment model of the π-stacking systems.

In addition, we again see a handful of potentially structure facilitating formations which occur only in subsets of molecules. Tempting though it is to simply state that these are structure directing constructs, predictions must be made and tested based upon their presence to establish this.

7.5 Other Crystal Structures

A good number of crystal structures in the homologous series lack commonalities with others in the same series when inspected by eye. These are detailed presently.

7.5.1 Compound 15

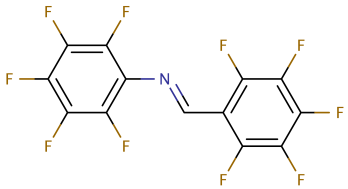


Figure 7.54: A reiteration of the molecular structure of compound 15

Compound 15 is isolated among the crystal structures. This structure forms in a grid-like system. The extreme torsion angle, the largest of any compound in this group, prevents this from being described as a 'stack', since the molecules in this species no longer lie flat. This torsion angle, the whole-molecule disorder that is observed in the crystal structure, and the generally low quality of the crystals, serve to underline the notion that fluorine-fluorine interactions are in fact, unfavourable. This, in spite of the fact that great lengths were gone to by Terry Threlfall to obtain crystals of this particular compound, whereas most species considered in this thesis crystallised without additional measures. The R-factor is notably higher for this crystal structure than others presented in this paper; so specific measurement comparisons should not be drawn, however, this too, underlines the lack of crystallisability of this particular small molecule.

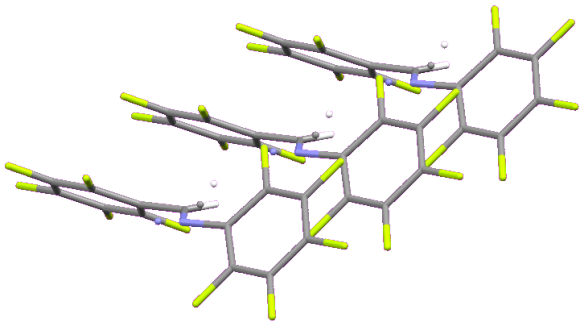


Figure 7.55: An illustration of adjacent molecules in the grid structure in compound 15

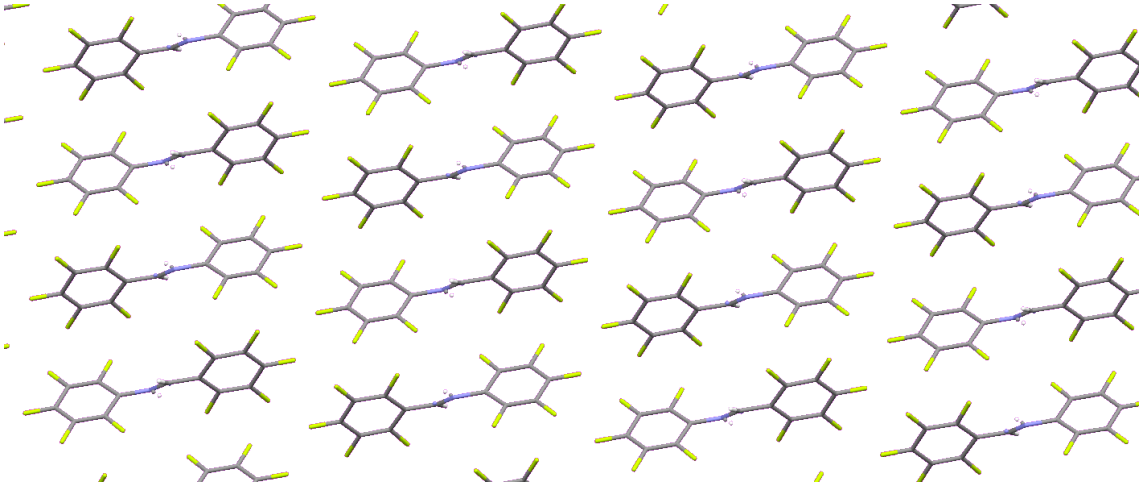


Figure 7.56: A side-on view of the grid in compound 15

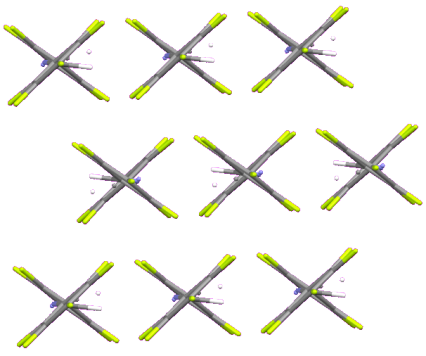


Figure 7.57: An end on view of the grid in compound 15

Compound 38

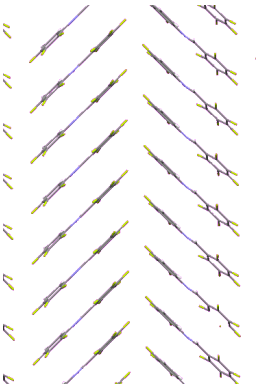


Figure 7.58: The structure of compound 38

Compound 38 forms a unique crystal structure in this dataset. Given that both rings are heavily fluorinated, it is curious that it does not form the same structure as compound 15. However, the two unsubstituted points on the aniline end of the molecule enable a continuous network of hydrogen-fluorine contacts throughout the system. In addition, the angles between pseudo stacks (see figure 7.58 and the offset within provide an arrangement which is consistent with the recieved behaviour of quadrupolar systems with electronegative substitutents described in our introduction.

7.5.2 Compound 10

Compound 10 displays a unique motif. Molecules arrange in rows, with hydrogen atoms and fluorine atoms in complementary positions transversely across the row as in Figure 7.59. These rows are paired, and become arranged in a brickwork pattern as seen in Figure 7.61.

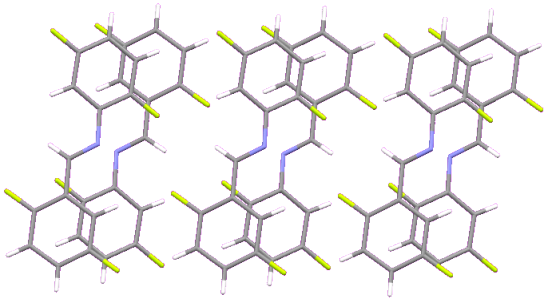


Figure 7.59: Two rows in atom colour

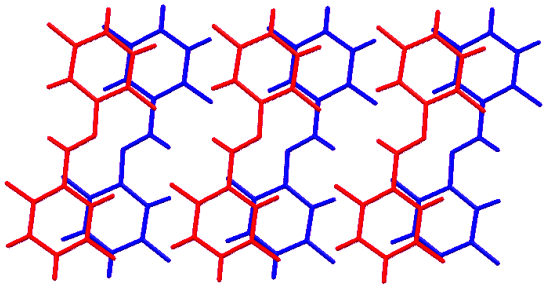


Figure 7.60: A false-colour view of two rows

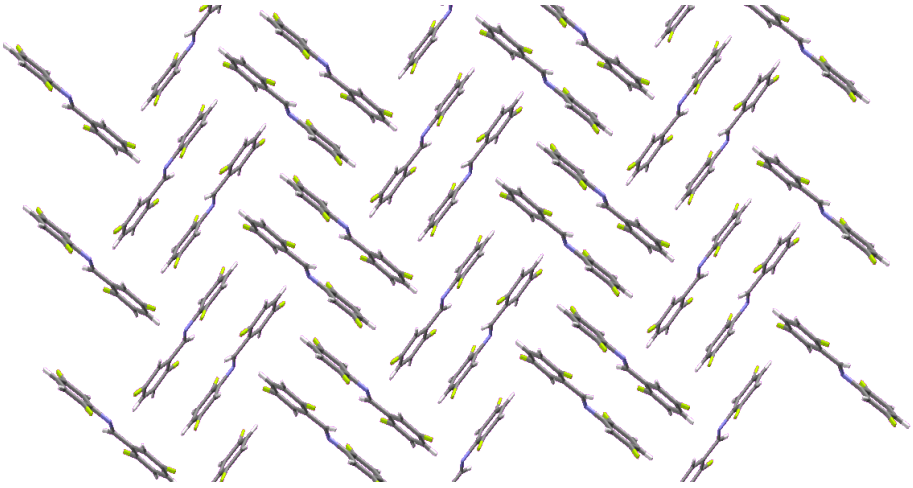


Figure 7.61: The brickwork pattern of compound 10

Whilst the side-to-side arrangement of the molecules in this structure makes intuitive sense, why the structure should limit the stacking arrangement to two molecules is rather

less intuitive. The brickwork arrangement may be a result of the quadrupole moments in the molecular structure, but such interactions are known to be weak.

7.5.3 Angled Layers

This set of compounds includes the set of isostructures with compound 4 discussed previously. What alters for the other 5 compounds in this group[iv] is the angle between the layered molecules.

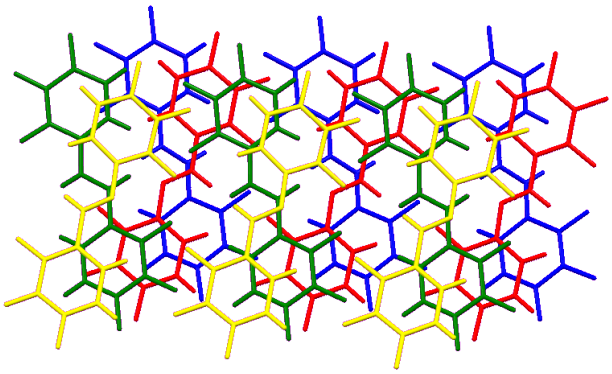


Figure 7.62: The angled layers structure, here exemplified by compound 37. The colours are non-atomic to better illustrate the layered nature of the structure.

These have extremes, for compound 63, this is about 1.5°, while for compound 70 it is around 55°. The angle directly corresponds with the distribution of the fluorine atoms, those molecules which have the fluorine atoms distributed symmetrically about the ring, as in compound 60, have the steeper angles, while those in this group with less substitution, or asymmetric substitution, see shallower angles.

Compound 37 differs from the other structures in that it has more layers before the original formation 'repeats'. That is to say, one axis of transformational symmetry is longer.

Compound 63 has by far a narrower angle, and could almost be considered not to be a member of this group. It bears a close resemblance to the unsubsituted structure in the CSD (BENZON11).[4] Should an adequate descriptor of structure be derived in the future, it may be that this allows us to see a continum of the structures.

[iv] compounds 16, 63, 69, 70 and 37

Compounds 12, 27 and 62

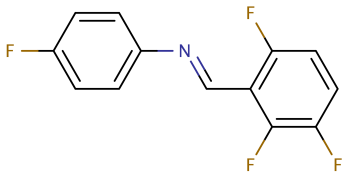


Figure 7.63: The molecular structure of compound 12

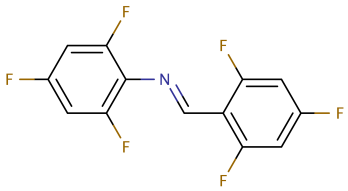


Figure 7.64: The molecular structure of compound 27

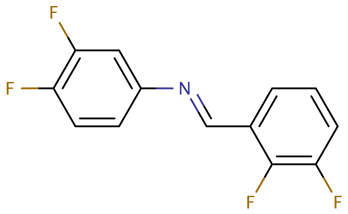


Figure 7.65: The molecular structure of compound 62

Compounds 27 and 62 share a motif of threads which are layered offset from one another. These layers form themselves flat constructs which are layered. These secondday layers are perpendicular to one another, alternating through the three dimensional structure.

Both structures have the molecules in the threads offset such that a molecule in one layer aligns its 'head' with the nitrogen-carbon linker in the center of the next molecule. However, in compound 27 the heads of the molecules are disaligned from the tail of the next molecule by about 17°. Compare with compound 62 where this directional offset is only 6°. This difference seems to be a compensation for the symmetrical distribution of fluorine atoms in compound 27, which would give rise to fluorine-fluorine interactions.

Compound 12 shares the overal motif, but the threads offset is greater than in the other two species, so that the head now sits above the tail in the molecule below. Note that this

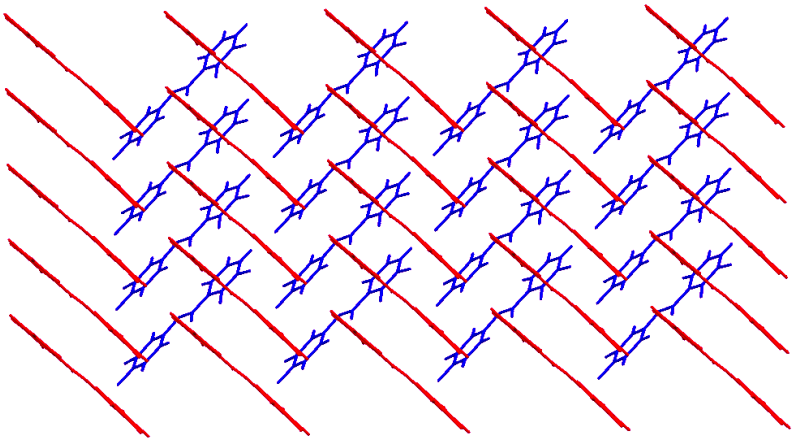


Figure 7.66: An illustration of the overlayed threads, taken from compound 62.

compound has much greater difference in substitution between the rings, possibly meaning that the quadrupolar arrangement to have rings above one another is more favourable. Given that information, however, it becomes much harder to substantiate why compound 12 did not form a simple, stacked structure.

Compounds 24, 35, 47 and 53

Compound 53, it should be noted, has already been discussed in the stacked structures. It appears here because although it possesses the stacked crystalline motif, it also shares another motif with the other structures here.

The primary motif for discussion here is not unlike that seen in the previous subset, save for the fact that the three dimensional arrangement is to have all of the threads pointing in the same direction, rather than alternating.

Again, it is true to observe that compounds with greater substitution have a greater angular deviation between the molecules and the overall 'direction' of the motif. Further, molecules with greater differentials between the rings in terms of the degree of substitution also sit with the rings as the overlap point, while those with similar degress of substitution sit with the ring above the linker in the next row.

Compound 53

Compound 53 is an interesting special case which warrants some further discussion. In it, we see two constructs which are found in other materials in this homologous series,

but combined together. This provides direct evidence for the the Aufbau principle of Kitaigorodskii. The principle holds molecules cohese to form constructs, which then go on to form crystal structures as we understand them in the three dimensional sense. That being the case, an observable proof of that hypothesis would be a crystal structure which combines two or more constructs seen in crystal structures built of similar molecules. This compound, and its hybrid crystal structure, provide a demonstrable proof of that idea.

7.5.4 Overall Summative Remarks

The Naïve Hypothesis

The existence of head-to-head stacking in the listed structures, alongside the existence of stacked structures in the completely clashing overlap structures, tends to indicate that the prevalence of stacking is not directly connected to the complementary overlap of the fluorine and hydrogen atoms in a molecular species. That being the case, it is interesting to note that the completely complementary arrangement of fluorine and hydrogen yields exclusively stacked compounds, whilst stacks are much less prevalent in the clashing structures. The net effect of this is that while the naïve hypothesis gives a good prediction in terms of the clashing and complementary stacks (a full statistical analysis of which is given in Section 8.

Crystal Engineering and F-F interactions

The importance of the crystal structure shown in Compound 53 as a part of two families of crystal structures cannot be overstated. To the best knowlege of the author this represents a first genuinely direct proof of an implicit prediction by the aufbau principle first put forward by Kitaigorodskii. Further work should be performed on the energetics of these crystal structures, but this will require correlated methods (in the quantum mechanics sense of these words). This would be a work unto itself.

The disordered structure of compound 15, the low quality of the crystals that are capable of being produced and the difficulty of producing them relative to those of the other compounds should illustrate firmly that fluorine-fluorine interactions are uniformly repulsive. This is further implied by the steep torsion angles between the rings in that structure which permit a sane quadrupolar interaction between the partial charges on the rings. Again, energetic studies can be performed to follow up on this, and this should be done as a completion of this analysis.

Hydrogen-Fluorine Motifs

The repeated occurrence of proximity of hydrogen fluorine pairs in all the structural motifs, illustrated in Figure 7.67, bear a marked resemblance to groups which have been described as synthons in other literature, and which were enumerated in the opening chapters of this thesis. In structures such as asymmetrically populated phenyl groups shown in the complementary structures, such interactions, if they exist, would be exacerbated by the quadrupolar moments present in the rings as was described in Section 2.5.3, which are also a factor in stacking systems. Those structures which have complementary overlap also, by dint of the substitution pattern, have the capacity for these complementary pairings of hydrogen and fluorine. In particular, the structures of compound 64, 40 and 66 show a preference to the centre of the neighbouring molecule, which contains a hydrogen near to a nitrogen atom.

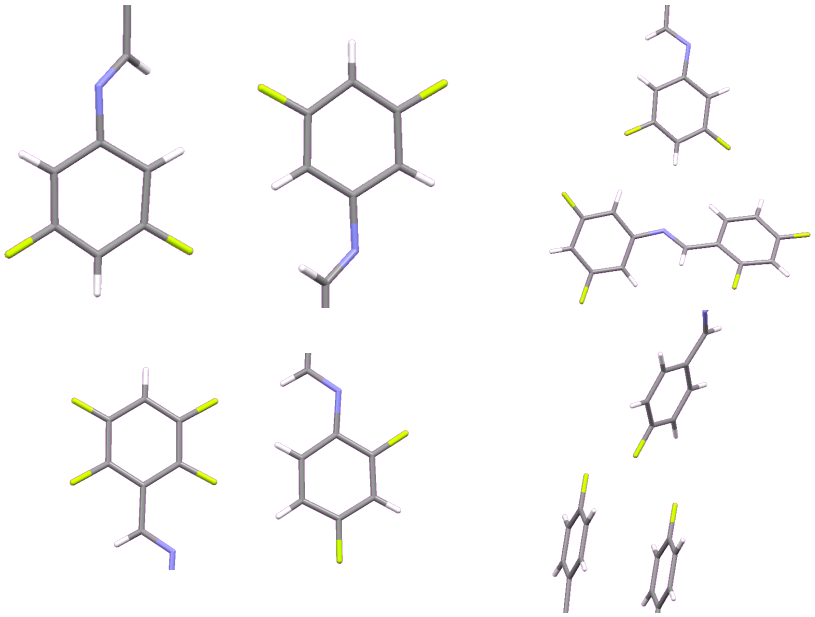


Figure 7.67: Views of recurring proximity motifs in various compounds.

The difficulty with the hypothesis of such an interaction is that the inter-atomic distances being described, as shown in Table 7.1, are of an order more commonly associated with weak hydrogen bonds. But it has long been an established notion that organic fluorine does not form hydrogen bonds - indeed it is ironically the very reason for which these compounds were selected. In the oft-cited paper which established, titularly, the fact that *Organic Fluorine Hardly Ever Accepts Hydrogen Bonds*[231], the evidence clearly supported the title statement - that at the time it was not seen that, by the definitions of

that time, Fluorine accepted hydrogen bonds in the context of organic chemistry. This was surprising given Fluorine's strong electronegativity, a fact which has largely been explained by the fact that fluorine's strong electronegativity also gives it a low polarisability, and so the donation of electrons to assist with hydrogen bonding is less likely than for Oxygen, Nitrogen, or the other Halogens. It is relevant therefore, that at the time of publication of the aforementioned paper - weak hydrogen bonds, such as were exemplified as structure directing motifs from the literature in the introductory part of this thesis - had not long been accepted or recognised as hydrogen bonds. As such, the search for fluorine-based hydrogen bonds in the paper focused on systems which may form strong hydrogen bonds: those which contain OH or NH groups, for instance, and these are, for reasons already described, much more likely to form hydrogen bonds with each other than with fluorine. This is compounded by the fact that distances seen in the structures presented in this paper would not necessarily have been reported as weak hydrogen bonds even if they had been of oxygen or nitrogen based groups. There is a difference of course, in the reported literature stating that an interaction is not observed, and stating that it *cannot* happen. Nevertheless, the notion that organic fluorine cannot form hydrogen bonds has become established theory to the point that intermolecular distances which might otherwise be noted as weak hydrogen bonds by software commonly used for passive searching of intermolecular interactions such as Mercury (whom shares in Jack Dunitz a progenitor with the paper being discussed), are specifically ignored for the case of fluorine.

While on the basis of the results presented it is very tempting to present them as a stand-alone counter-argument. A more full investigation of the stabilities of these crystal structures, and their relation to alternatives needs to be assessed in a rigorous way, lest the counter argument be hoisted upon its own rhetorical petard.

Analysis using XPac

Lastly, the too XPac has been utilised in this section to give insight into the isostructurality within these systems. It has been found that the method is flawed with respect to two aspects of this work. The least detrimental is arguably that of the failure to spot an isostructure which has an enantiomeric relationship to other structures. It could be argued that this is a feature of the software produced, rather than a bug. But it is an inaccurate result for cases where the absolute configuration cannot be known, for instance as in the case presented in this work, where light atoms are present and molybdenum radiation is used. Moreover, even were it not the case that this is incorrect, such results are valuable to our understanding of crystallisation.

Structure No.	Distance (Å)	Angle (°)
49	2.543	133.98
49	2.374	108.74
49	2.781	140.59
49	2.917	104.35
41	2.560	159.24
41	2.577	158.30
41	2.582	98.82
41	2.582	143.94
41	2.595	149.95
5	2.780	110.98
5	2.523	146.88
5	2.804	143.16
5	2.523	158.20
5	2.616	150.31
55	2.493	115.14
55	2.580	123.63
55	2.688	125.97
55	2.810	175.20
66	2.502	132.59
66	2.605	144.21
66	2.624	122.83

Table 7.1: Example H/F distances from 5 randomly selected fluorobenzanilides, angles and distances are as reported from Mercury[230]

Secondly, the large volume of related crystal structures caused XPac to give rise to a plethora of possible constructs or motifs which were common to pairs of structures. What the software lacks however, is the ability to perform the necessary meta-analsys to see which of these motifs are shared by multiple pairs of crystal structures. The result is that this was effectively unusable for this case, as the number of pair-wise constructs became impractical to manage 'manually'.

Unfortunately, there is no recent news of development work being done with XPac, and the publicly funded source code remained closed and is unavailable for third party development. One of two things therefore is required for the above analyses to take place in a reasonable way, either the source code must be released, or funding must be procured to repeat the work that has already been performed as a platform for new work to take place. Unfortunately, neither of these seem likely outcomes in the near future.

Chapter 8

Statistical Models

8.1 Naïve Hypothesis

A simple test can be used to demonstrate the effectiveness of the naïve hypothesis with respect to the stacking or non-stacking of the clashing and non-clashing molecular substitution arrangements. Fisher's exact test provides with a way of testing the valid outcomes of such a true/false hypothesis.[i] In this case, a true positive is where the naïve hypothesis states that a stack would form, and a true-negative would be where a clash was predicted not to form a stack. While the head-to-head stacks were an unanticipated structure type, we shall include them as stacks for the purpose of testing the hypothesis. This results in 8 true positives, no false positives, 5 true negatives and 3 false negatives. Therefore, the naïve hypothesis predicts accurately for this population with a p-value ¡ 0.05. However, the mechanism underpinning this seems to be clearly erroneous- the existence of such a large body of head-to-head structures clearly indicates that any vertical interactions between fluorines in the stack formation are not guaranteed to be structure selecting.

8.2 Novel Descriptor Analysis

8.2.1 Relation to Physical Property

Melting points and Invariant Graph Descriptor values were collected and calculated as previously detailed. The results have been plotted blow. The bars represent the start and end of the melt points, and the circles the mean value of the start and end points. Raw data for the graph below can be found in the digital appendix of this document.

[i] It is apocryphally said that the test was first used to test whether or not a woman was able to tell correctly whether milk had been first added to tea under blind control conditions.

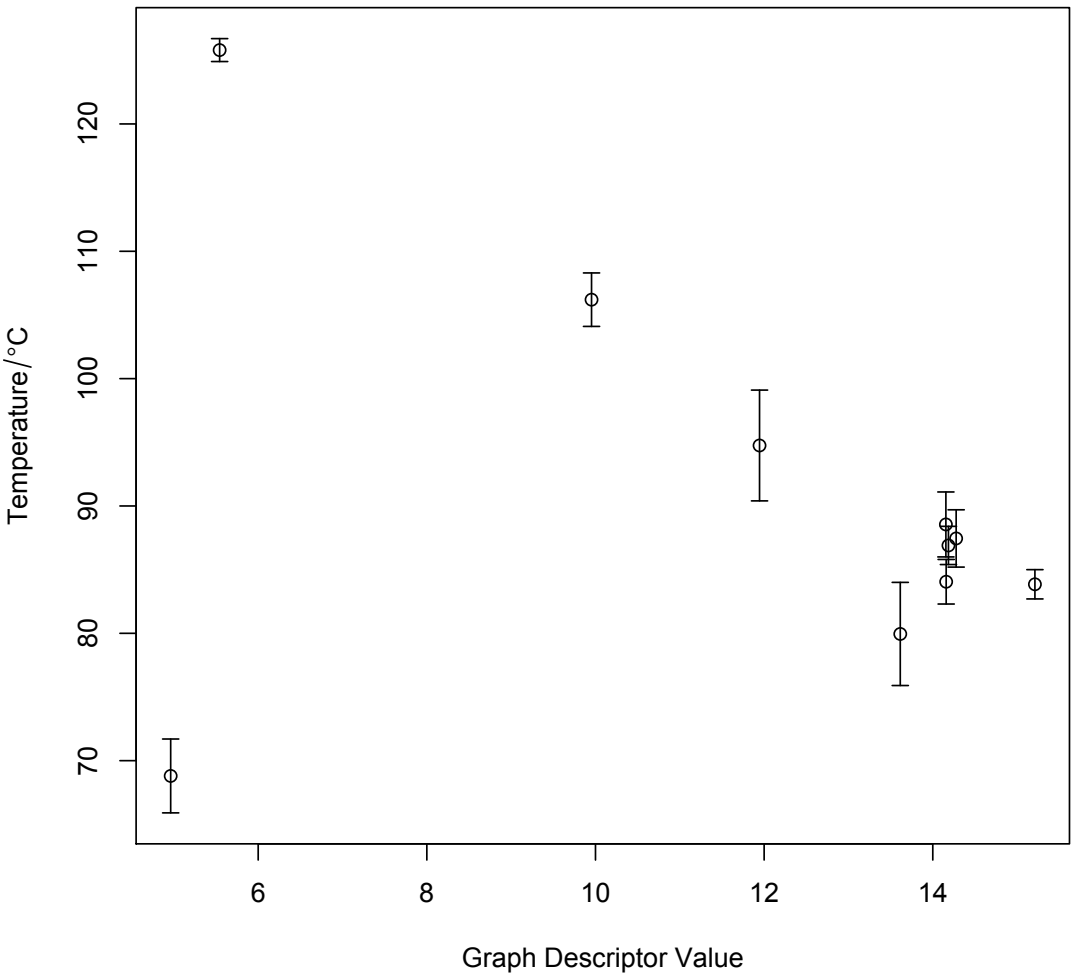


Figure 8.1: Melting Point plotted against the Novel Invariant Graph Descriptor

It is evident from visual inspection that there is an outlying value.

Note the markedly different shape of the packing structure involved in the compound, that gives rise to a very different connectivity graph, and this may render it an outlier. However, reconstructing this crystal structure in such way that it more closely resembled the packing shells did not change the value of the graph descriptor. This is, in the broadest sense, a good thing, as this indicates that the descriptor itself is relatively immune to arbitrary changes is reference point. However, it does raise the question as to why this

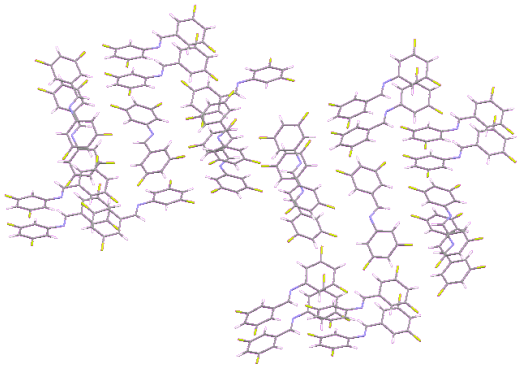


Figure 8.2: The packing shell of compound 29, which is the outlying value in Figure 8.1

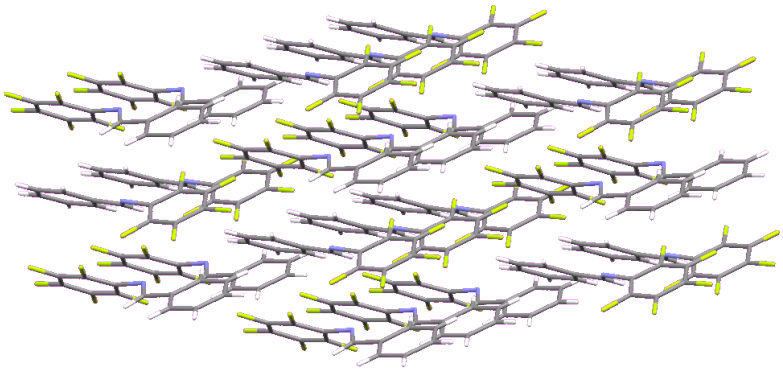


Figure 8.3: The packing shell of compound 9, in the top left of Figure 8.1

particular crystal structure gives rise to such an outlier.

We see a correlation coefficient of -0.69. This therefore demonstrates an extremely moderate correlation between a novel, quantitative descriptor of a crystal structure and a measured property. The descriptor has the added benefit of being invariant to the number of molecules within the unit cell of a crystal structure, a problem which has plagued the comparative quantification of crystal structures.

It should be noted that although a correlation value has been calculated, a p-value has not - this is because the subset of fluorobenzanilides cannot be considered a random sample, though it is representative of the set. No meaning can truly be ascribed to such

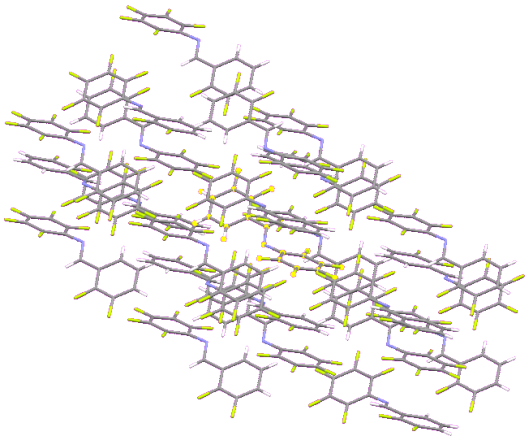


Figure 8.4: The packing shell of compound 42, which lies in the cluster of values in the bottom right of 8.1

a p value calculated for such data, and p-hacking has become a problem in the scientific literature;[320] the task of obtaining a truly random sample, therefore, remains for further work.

8.2.2 Interpreting the Novel Descriptor

Eigenvalues are inherently lossy descriptors. The condensation of matrices into vectors and scalars inevitably removes some data; this is why the graph descriptor as presented does not necessarily code graph identity. The upshot of this is that it is difficult, if not impossible, to intuit meaning directly from the results.

There are clear groupings from the descriptor. However, it has not been possible to relate these to specific features of the graphs manually. With the advent of new versions of the CCDC's python library, this may recently have become possible to correlate with the graph sets of the molecules, and a method for this is described in further work.

Nevertheless, an interpretation may be acquired by the analogy with graphs used to describe molecules. In those cases, the principle eigenvalue of the molecular graph corresponds to an energy level which describes the relative stability of that molecule. To follow this analogy through to its conclusion, if the short contacts represent, as they are often assumed to, electron dense regions in the crystal structure (as per hydrogen bonding and other interactions), then this would also follow for crystal structures too. This would be much more generalisable, and would also go some way to explaining the correlation with melting point of the descriptor. However, again, this should be followed up with

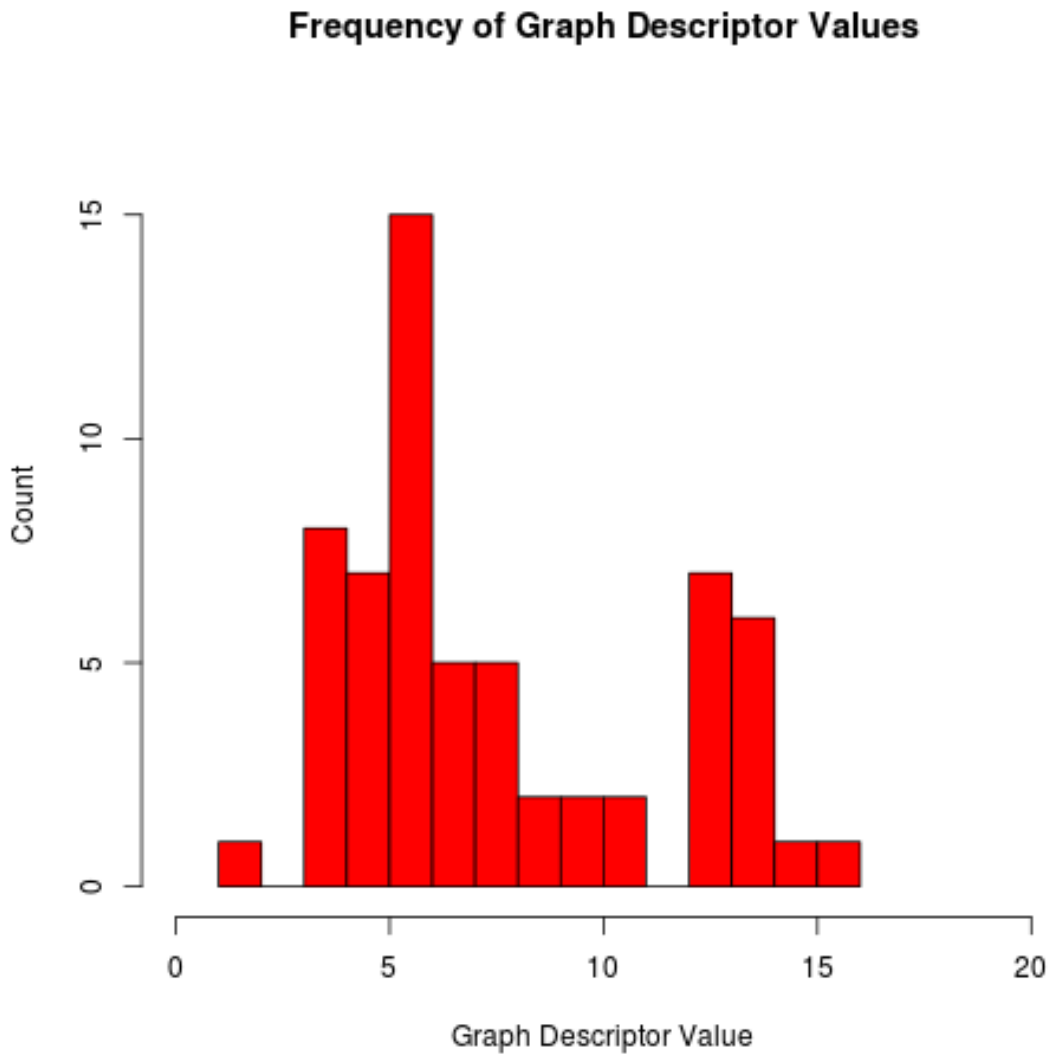


Figure 8.5: The population density of different values of the Graph Descriptor

quantum mechanical simulations or charge density studies.

Unfortunately, this does not assist with the decoding of the outlying melting point/-graph descriptor value. The structure that gives rise to it shows nothing characteristically unusual in terms of the pattern of close contacts as compared to other structures in the experiment. It is true to say that it is the only compound with its packing structure in the examined data set. It may be that examining the other compounds with similar packing structures may yield information as to why, systematically, this structure is an outlier.

8.3 Co-Crystallisation Experiment

A Characterisation And Regression Tree was constructed as already described using JMP 11 software, on approximately 600 instances of co-crystal experimental data taken from the literature, a full listing of which can be found in Appendix A. The initial decision tree is shown in Figure 8.6.

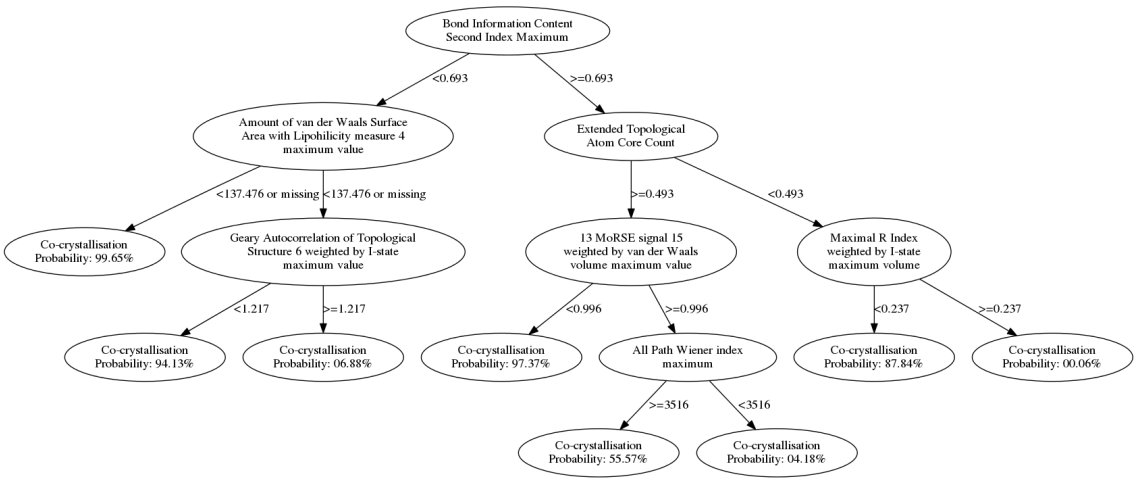


Figure 8.6: A representation of the decision tree generated for the co-crystallisation. Each node details the descriptor. When one calculates for a new co-crystal pair, one follows the appropriate values for the descriptors based on the value description on the edges. (arrows). At the base of the decision tree, the probabilities are given for the likelihood of co-crystallisation. These are derived from but not identical to the proportions of co-crystals in the learning data set at that point in the tree.

The descriptors which emerge in this decision tree are chemically relevant, and the shortness of the decision tree coupled with a high R^2 value of 0.98 indicated that this decision tree was potentially useful.

When assigning the probable result a pair of molecules which one intends to co-crystallise, one calculates the descriptor values for that molecular pairing and then traverses the tree, following the branch which matches the value. The following description of the relevant descriptors attempts to mirror that approach.

One problematic aspect of the decision tree is that in general, the only matter being described is the 'maximum' value; that is, of the two components of the co-crystal, the only value which appears to matter as far as prediction of the co-crystallisation outcome is concerned, is the higher of the two, regardless of what the other might be. This does place constraint on the secondary (effectively silent) value, since it cannot be higher than

the maximal value, but is not necessarily a tight constraint in all cases.

The bond information content second index maximum (BIC2) describes the bonding pattern in a molecule. It can be calculated by virtue of the fact that molecules can be described as a set of vertices and edges - such a set is described as a graph by mathematicians, and the study of such graphs is known as graph theory. Such graphs can be used to represent molecules in computer programs, and can also be used to calculate descriptors by representing them as matrices (as discussed in 6).

It is calculated by separating the vertices (atoms) into equivalency classes (mathematically, a set), each class belonging to a corresponding topological distance m. Vertices are considered equivalent (are in the same equivalence) class if they meet the following conditions:

- The vertices represent the same chemical element
- The vertices have the same *order* (They have the same total number of connections)
- There exists a distinct path of length m for each vertex such that the vertex orders, number of edges, and chemical element types is the same in each path

From the equivalency classes, the neighbourhood information content index IC_m is defined by:

$$IC_m = -\sum_{g=1}^{G} \frac{A_g}{A} \cdot log_2 \frac{A_g}{A} \tag{8.1}$$

Where:

- G is the count of the equivalence classes
- A_g is the cardinality of equivalence class g
- A is the total number of vertices

This is then normalised to the number of edges and their multiplicity:

$$BIC_m = \frac{IC_m}{log_2 \left(\sum_{b=1}^{B} \pi_b^* \right)} \tag{8.2}$$

Where:

- B is the number of edges in the graph representation

- π_b^* is the conventional bond order of the edge b (a double bond gives this a value of 2, for instance)

BIC2, therefore, is a characterisation of the bonding patterns in a given compound. The larger the value, the more 'complex' the connectivity graph is considered to be. The inclusion of this descriptor in the decision tree is on one hand, intuitive, in that less complex molecular systems clearly have a greater probability of forming co-crystals. The difficulty with such a descriptor is that the mathematical abstraction required for its calculation means that it is difficult to intuitively interpret what the descriptor 'means' in terms of the connectivity of the molecule.

A Amount of van der Waals Surface Area with Lipophicity Measure 4 is defined as the amount of the van der Waals surface area (VSA) with a given property of a given value (4).

The VSA of each atom is estimated using the following formula:

$$VSA_i = 4\pi R_i^2 - \pi R_i \cdot \sum_{j=1} Aa_{ij} \left(\frac{R_j^2 - (R_i - g_{ij})^2}{g_{ij}} \right) \tag{8.3}$$

Where:

- R_i is the radius of the radius of atom i, and similarly for j
- A is the count of atoms in the molecule
- $g_{ij} = min\left\{max\left\{\left|R_i - R_j\right|, b_{ij}\right\}, (R_i + R_j)\right\}$
- $b_ij = r_i^* j - c_{ij}$
- r is an idealised bond distance between two atoms
- c is an adjustment parameter based on the valency of the atoms

Many descriptors can be mapped into P_VSA like descriptors, in this case, the logP value is the relevant descriptor.

The P (and by extension logP) are measures of lipophilicity. There exist a number of ways to estimate this; One of the most intuitive of these is the water-octanol partition coefficient:

$$logP = log[C]_{octanol} - log[C]_{water} \tag{8.4}$$

Where $[C]$ is the concentration of chemical C in a given solvent (water or octanol).

Thus, this descriptor describes the behaviour of a compound with respect to solvents. The fact that one examines the amount of surface area with a given logP value rather than an overall logP value for the molecule rather obfuscates the intuitive understanding of the value of the descriptor. A logP of 4-5 (represented by the descriptor class mentioned in the decision tree) would be the same for the compound Phenanthrene. Dragon is not well documented as to which units it uses for these calculations, but by implication that the Handbook of Chemical Descriptors and Dragon were authored by the same person (Todeschini), it can be assumed that the correct unit is square Angstroms.

This descriptor proves therefore, to be reasonable in terms of our overall understanding of chemistry, in that solvent interaction is important in crystallisation, but proves to be non-intuitive in that it requires no more than a specified amount of surface area to fall within a specified value of a solubility ratio for crystallisation to occur. It is a component of a spectral descriptor, and demonstrates a difficulty with that type of descriptor. Furthermore, the split is not large - only 14 molecular pairs fall one side of the split. Lastly, this descriptor is not invariant to the overall size of a molecule, which further hinders interpretation.

At the same level of the decision tree, but on the other side of the BIC2 split, Extended Topological Atom Core Count is selected as a splitting factor. More fully identified as the Extended Topochemical Atom Average Core Count, it is derived as follows.

The core count for the ith atom, α_i is calculated as:

$$\alpha_i = \frac{Z_i - Z_i^v}{Z_i^v} \cdot \frac{1}{L_i - 1} \tag{8.5}$$

Where:

- Z_i is the proton number of the ith atom
- Z_i^v is the valence electron count of the ith atom
- L_i is the principle quantum number of the ith atom

The Average Core Count is the arithmetic mean of these values over the atoms in the molecule, and acts as an invariant description of the atomic makeup of the molecule. Molecules with high values for this descriptor tend to imply large molecules, though this rule of thumb is not always adequate. Atomic carbon has an α value of 0.5; nitrogen a value of about 0.29. Given the organic character of the species under consideration, the decision tree therefore implies that molecules with a large number of atoms higher on the periodic table than carbon are more likely to form co-crystals. Intuitively, this makes

sense; species with these more electronegative constituent atoms are more likely to have polar regions (they will contain nitrogen, oxygen, etc), and these will provide additional electrostatic impetus for crystallisation.

The GATS6s descriptor, fully the Geary Autocorrelation of Topological Structure weighted by I-state, belongs to another group of descriptors like the VSA descriptors, to which can be mapped another quality of the makeup of the molecule, in this case the I-state.

The Geary coefficient is calculated thusly:

$$c_k = \frac{\frac{1}{2\Delta_k} \cdot \sum_{i=1}^{A} \sum_{j=1}^{A} (w_i - w_j)^2 \cdot \delta\left(d_{ij}; k\right)}{\frac{1}{A-1} \cdot \sum_{i=1}^{A} (w_i - \bar{w})^2} \tag{8.6}$$

Where:

- w_i is a property calculated for the ith atom, and similarly for j
- $\bar{w}$ is the average of that property on the molecule
- δ is the Kronecker delta which is set to 1 if the topological distance between atoms i and j (d_{ij}) is equal to k
- k is the topological distance under consideration; the original statement of this descriptor calls this the *lag*
- Δ_k is the number of atoms at topological distance k
- A is the number of atoms in the molecule

The I-state, or intrinsic state of an atom, by which this and other descriptors may be weighted, is a means to encode both the electronic and topological properties of an atom within a molecule. It is mathematically defined thus:

$$I = \frac{v + 1}{n} \tag{8.7}$$

Where v is the number of valence electrons in the atom, and n is the number of "nearest neighbours" - the number of atoms to which the atom under consideration is bonded to.[321]

This descriptor is another description of the complexity of bonding within a molecule, which is not in this case invariant to the species under consdieration; it should also be noted that it does not correlate strongly with the BIC2 descriptor. Broadly speaking,

the larger the value of the GATS6s descriptor, the greater the difference in connectivity environment between atoms at a distance of 6 bonds. Thus, the group of molecules which have the lower values for this may either be those which are very simple molecules, or those which are very intra-connected.

Mor15v is fully known as 3D-MoRSE descriptor signal 15, is an involved descriptor which is based on the calculation of electron scattering patterns, and weighting by a quality (in this case, van der Waals volume) of the molecule under consideration. The set of this spectral descriptor as a whole describe the shape of the molecule, but what the meaning of signal 15 in isolation is remains unclear.

R5s. is the Maximal R Index Weighted by I-State The R index is defined using the off-diagonal elements of what is termed the influence/distance matrix **R**, which has the following elemental definition:

$$R_k^+(w) = \max_{ij} \left(\sum_{i=1}^{A-1} \sum_{j=i+1}^{A} \frac{\sqrt{h_i \cdot h_j}}{r_i j} \cdot w_i \cdot w_j \cdot \delta\left(d_{ij}; k\right) \right) \tag{8.8}$$

Where:

- r_{ij} is the geometric distance between atoms i and j
- m_i is the molecular mass of the atom, and similarly for j
- h_{ii} is the diagonal element from the influence matrix
- A is the number of Atoms
- k is the topological distance for which the R index is being calculated
- δ is a Kronecker delta function equal to 1 when the topological distance between atoms i and j (d_{ij}) is equal to k
- $i \neq j$

The influence matrix is defined as:

$$\mathbf{H} = \mathbf{M} \cdot (\mathbf{M^T} \cdot \mathbf{M})^{-1} \cdot \mathbf{M^T} \tag{8.9}$$

Where M is an $A \times 3$ matrix of the Cartesian coordinates of the A atoms in the molecule. Its diagonal elements h_{ii} in this case modelling the influence of each atom on the shape of the molecule as a whole, and are valued between 0 and 1.

The descriptor therefore denotes relative bonding complexity at a bonding distance of 5 bonds, in a fashion similar to the GATS6s descriptor. Again, this descriptor does not represent a large split of the training data, and its meaning is not very intuitive.

Wap is the 'all path Weiner index', the formal mathematical definition of which is omitted here for clarity and brevity; it is another characterisation of the bonding patterns within the molecule, but its relationship to an intuitive understanding of bonding complexity is not as straightforward as other descriptors mentioned previously.

Testing of the Model

To test the predictive power of the model, Lucy Mapp provided an example list of co-formers and example target compounds with which she was intending to perform co-crystallisation experiments, and these were run through the decision tree to provide a set of predictions of outcome as per the decision tree. The results of these experiments were universally that, even where the reactions were expected to generate co-crystals, this did not occur - no reaction in this set generated a co-crystalline species. On inspection of the target compounds it became obvious as to the cause of this outcome; the model had been deliberately optimised for molecules without hydrogen bonding capacity, whilst the target compounds (Artemisenin, Griseofulvin and Fenofibrate) are candidates for hydrogen bonded co-crystals - ones which, furthermore, have proven problematic for generating co-crystals of previously. An exact list of the reactions cannot be provided doing to intellectual property concerns of third parties. The reactions were performed under a variety of different conditions, including melts and liquid assisted grinding experiments.

The lack of hydrogen bonding co-crystals in the training data set would likely lead to two sets of outcomes; either the model simply doesn't contain descriptors which describe the involvement of hydrogen bonds. This would yield predictions which would be meaningless, both false positives and false negatives. An alternative case, the descriptors may bias hydrogen bonded systems to 'failure' predictions, as the training set contained some molecular structures with known hydrogen bonding donors and acceptors in the failure data. However, the majority of the predictions which did not transpire to be accurate were false positives, and so the former scenario is the most likely.

In light of this, a new decision tree was drawn up including an additional 400 results detailing hydrogen-bonded co-crystal systems. The results of the decision tree building process for that are included in the digital appendix, and are (as expected) far more complex and much more challenging to interpret in the direct fashion that was possible for the non-hydrogen bonded co-crystal model. It also proved to be impractical to draw

predictions from in a sensible time frame[ii]. Part of the reason that this decision tree may have been so complex[iii] is that Dragon actually contains very few descriptors well suited to explicitly describing hydrogen bonding. For instance, it does not include any counts of hydrogen bond donors or acceptors. This oversight may prevent any decision tree being produced using those descriptor calculations for hydrogen bonded systems, and alternative descriptor calculator should probably be selected for future work. Other work performed has also utilised a different classifier type, the Support Vector Machine, with more successful results,,[300] svms being better suited to problems containing necessarily correlated descriptors, for instance molecular bonding patterns and hydrogen bonding donors and acceptors.

Furthermore, the fact that many of the experimental details which clearly effect the outcome of the experiments beyond the reactants involved, meant that the ongoing practical experiments no longer utilised the model - the observation being that the model might be a useful filter for what *can* form co-crystalline systems, but not under which conditions, and hence it would not be a useful predictor of what *will* produce co-crystalline systems.

Nevertheless, it is important to note that there are some similarities in the more complex combined model and the simpler model for non-hydrogen co-crystalline systems which bear some examination, in particular, the retained presence of the bond information content descriptor and the extended topochemical atom index as the two main factors in the decision tree. Given that these emerge as descriptors regardless of the subsets being observed, it seems reasonable that these descriptors confirm what has long been held to be true, that the shape and bonding of organic molecules is paramount in the formation of crystal structures before other considerations come into play.

Whilst the decision tree models here have failed to produce accurate predictions on new data, or have simply not yet been tested owing to external factors, it should be remembered that co-crystals are relatively rare among crystal structures and are considered among the more difficult challenges in crystal engineering. Furthermore, the lack of availability of both positive and negative data in a truly representative fashion of co-crystallisation experiments is a challenge that must be met by the crystal engineering community - the continued absence of this 'failure' data from the literature discussion is probably the largest hindrance to the creation of robust statistical models.

In addition, the approach of being able to generate hypotheses and lend numerical strength to chemical intuitions as to the governing factors of crystallisation in systems has

[ii] It emerged that JMP has no automated facility to do this

[iii] And hence, why the raw data rather than an image representation of the tree is included in the digital appendix.

proven to be useful in and of itself for more tractable systems than organic co-crystals, and work based on this method has recently been accepted for publication in Polyhedron.[322]

Chapter 9

Conclusions and Further Work

9.1 Conclusions

In this thesis has been examined a homologous series of related fluorobenzilidine crystal structures. The structures were hypothesised to form stacks which had the substitution patterns in a complementary H/F pattern vertically within the stack. For those structures with perfect complementary overlap and perfect clashing, this proved to be a strong hypothesis. However, the existence of head-to-head stacked structures, which necessitate clashing overlap, indicate that the hypothesis success is mediated by some other mechanism than the vertical interaction between substituted atoms in the stacks, be that hydrogen-fluorine interactions between stacks or quadrupolar moments or some combination of the two. Indications of both occur frequently throughout the family of crystal structures.

The crysal structures themselves have been grouped into related compounds, and some hypotheses have been drawn as to the origins of the patterns. In particular, one compound gives rise to implications that Kitaigorodskii's aufbau principle can be demonstrated to be the case without the aid of mechanistic evidence, which is very difficult to obtain for crystallisation processes.

A common set of hydrogen-fluorine interactions were identified within the various subsets of the fluorobenzilidine crystal structures. A rigorous test for their structure directing nature is proposed in the further work section. Their distanced would, but for the involved atom types, be characteristic of hydrogen bonding, suggesting a need for the review of the commonly accepted position within organic crystal structures that organic fluorine does not form hydrogen bonds. There were also a large number of indicators that for this family of structures at least, fluorine-fluorine interactions can only be considered to be a repulsive interaction.

A new descriptor has been developed for the interaction networks in crystalline compounds. The fluorobenzilidine crystal structures were used as a test set for the descriptor. Rather than attempting to relate the graph descriptor to the molecular structure which proves difficult to do in a generalised way, the descriptor was related to the melting points of the crystals. This proved to have a moderate correlation with one unexplained outlier.

Two attempts have been made at feature selection operations for crystallisation. The initial attempts using sulphonamides and glycine were eliminated early since the data sets could not be large enough to support true correlations on such a large number of potential descriptors. Another attempt was made on co-crystalline systems. This yielded a sensible model with good validation statistics, but when applied, failed to generate good predictions. This was laid down to the lack of hydrogen-bonded crystal systems in the training data as compared to the trial data. In addition, it remains the case that failure data is hard to obtain from the literature, and much of the failure data revolved around a small number of closely related systems. Finally, the chosen descriptor calculator does not contain direct hydrogen bonding descriptors for molecular systems. It seems self-evident that for that biases surrounding these systems in terms of both descriptors and data resulted in models which whilst apparently valid would not be successful. This also explains why a new model that was generated on a new data subset proved to be complicated beyond utility.

9.2 Further Work

Much further work arises from the work presented in this thesis. Approaches have been developed with varying degrees of success, and these now need to be applied across greater ranges of fields or with additional refinements.

The graph descriptor value still lacks a good intuitive meaning. There is a reasonable supposition that it might imply stability of the crystal structure. This should be assessed next to stability calculations performed using quantum mechanical methods, or by further testing against melting points using a greater range of structures. In addition, although it was not possible at the time of this work, a new release of the CCDC python API should make it possible to correlate the graph sets (which are non-quantitative) with the descriptor value. This would not necessarily extend beyond any one compound family, but may be able to provide a more intuitive understanding of the novel descriptor.

Additionally, correlation studies should be performed between the molecular structures involved and the graph descriptor values. Even if the study can only show results within the homologous series, it may give important insight into the key substitution locations

in the molecules and their effect on the overall crystal structure.

It was also noted in the crystal structure analysis section of this thesis that a number of hydrogen-fluorine constructs arise commonly throughout the series. With the data already available, it would be pertinent do perform a rigorous statistical assessment of whether these interactions arise purely by co-incidence or whether they are actually anomalously common. This would require an assessment as to which molecules *can* find an orientation which produces these interactions within the homologous series, and which ones actually *do*. Then a fisher's exact test can be performed once more, as it was for the naïve hypothesis in this thesis.

Further to the naïve hypothesis, it would be ideal to attempt to correlate the existing groups of structures with the degree of potential vertical overlap in a vertically stacked structure. This would require the development of a descriptor which adequately captured the overlap extent information. This would allow the determination of whether the hypothesis extends to the full range of data produced by this work.

To supplement the further categorisation of work of this kind, it would be useful to have a more advanced toolset within the program XPac. For such large datasets it proves to be invaluable for spotting isostructures, even if there are flaws in this method which have already been discussed. However, it lacks the capacity to meaningfully gather information across the large dataset about one- and two-dimensional constructs, simply because the volume of data becomes to great. This will likely require either the release, or the reproduction of the source code of that program.

In order to redress the descriptor selection work which proved largely unsuccessful in this thesis, one might better look towards design of experiment methods from statistics to actively construct a well-designed data set from scratch, including the absent information on synthesis methods. By definition, this would also provide the much needed systematically related failure data which is so badly needed for creating sound and predictively relevant models. Until such data becomes widely available for historical data, such rigorous investigations will have to remain the norm. However, such work can contribute to the change in culture by actively releasing the failure data.

As with much science, some answers and aims have been met during this research, however, some challenges remain, and some new ones have arisen. First and foremost it has become self evident as a result of performing this research that much data is left unreported, most key being that which represents experiments with unanticipated results. Mechanisms need to be developed for recording, quantifying, accessing *en masse*, and modelling this information in a manner which is convenient to practising chemists. One of the key components in the disproving of the theory of phlogiston was access by Antoine

Levoisier to data well recorded by Joseph Priestly. This necessity of access to well recorded data is no less true now than it was in those early days of chemical understanding - in fact, with the rise in the complexity of the problems under examination, it could and should be argued that it is in fact *more* so. It is, not coincidentally, to a project with this in mind which the Author is currently attached.

The fluorobenzanilide experiments provide two obvious avenues of further research. Firstly, the collection of a truly random sample to confirm the invariant descriptor as being consistent, and secondly that some more quantitative explanation be gained for the recurring hydrogen-fluorine motif that was seen throughout the set. In particular, work using Pixel, and NCI plots[323] to elucidate whether there is in fact, a true interaction taking place would be of great benefit, particularly given that because of the isomerism issue these molecules are unlikely to be suitable for a statistical approach to this problem.

The other avenue of research would be a more automated way of identifying common motifs. The program XPac provides some assistance in this regard, but does not perform any kind of clustering among crystal structures based on the motifs or constructs present within the crystal structures, and this assignment of groups would be necessary to develop a model for predicting such motifs.

Taken together, it would also be a worthy exercise to measure the invariant descriptor among small organic molecular crystal structures, and correlated this with the shape descriptors of the molecule. Further, it would also be a boon to have some link between the quantitative measures of the molecular structure shape descriptors and the intuitive understanding of molecular structures that exists.

Part V

Appendices

Appendix A

Co-Crystalline Modelling RefCodes

The following is a list of refcodes, which identify crystal structures in the CSD, in line with the suggested policy on referencing structures from the CSD by CCDC.[324] These crystal structures were used in the analysis of co-crystalline systems presented in Section 5.3. Items where a literature reference could not be confirmed have been left uncited.

ABUNAM[325]	ASAVIZ[336]	BEZSIJ[350]	CAPTOC[362]	COKNUM[376]
ABUNAM01[325]	ASIBAG[337]	BICVUE01[351]	CAPTOC01[363]	CRMESF[377]
ABUNOA[325]	ASIHUF[338]	BILXUP[352]	CAZLAR[364]	CRPACX10[378]
AJUMOI[326]	ASIWIJ[339]	BIRDIP[353]	CAZLAR01[364]	CUKXIP[379]
ALUMOJ[327]	ASIZEI[340]	BIRQUO[354]	CAZLAR02[364]	CUPJAZ[380]
AMILUD[328]	AYEBAH[341]	BITROL[355]	CDSCDS[365]	CUPJED[380]
ANTCYB13[329]	BALVOA[342]	BORNUS[356]	CEHPUC[366]	CUPJIH[380]
ANTCYB14[330]	BAPMAH[343]	BORPAA[356]	CEJTAM[367]	CUPJON[380]
ANTPML01[331]	BAZDAH[344]	BOVQUY[357]	CEKBUP[368]	CUPJUT[380]
ANUPIJ[332]	BDTNBB[345]	BUVKOS01	CEKYUM[369]	CUPKAA[380]
ANUPUV[332]	BECNUS02[346]	BUWCIF	CENHAE[370]	CUPKEE[380]
ANUQAC[332]	BEFGIC[347]	BZANTC10[358]	CENHAE01[371]	CUPKII[380]
APANBZ[333]	BERMOB[348]	BZAPMA10[358]	CEWYOT[372]	DARZOM[381]
ARIWAA[334]	BERZED[349]	BZATNB20[358]	CIFWUJ[373]	DATCEI[382]
ARIWAA01[334]	BEZRUU[350]	BZQTCQ10[359]	CLAHMB01[374]	DATCIM[382]
ARIWAA02[334]	BEZSAB[350]	CAFWAH[360]	CLAHMB02[375]	DBTTNB
ASAKOU[335]	BEZSEF[350]	CAMBAU[361]	COKNOG[376]	DEBVAI[383]

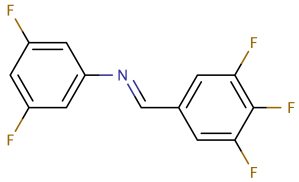
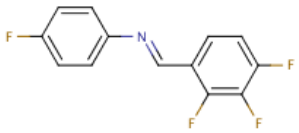
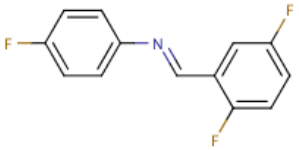
DENFUX[384]
DENMIT[385]
DESDIO[386]
DESDIO01[386]
DEXWAE[386]
DIVDUI[387]
DNPCPH[388]
DOCLIQ[389]
DUJTAD[390]
DUJTEH[390]
DURYUK[391]
DURYUK01[391]
DURZAR[391]
DURZAR01[391]
DURZAR02[391]
EBIHIH[392]
EBIHUT[392]
EBIJAB[392]
EBIJEF[392]
EBIJIJ[392]
EBIJOP[392]
ECUTUR[393]
ECUVIH[393]
ECUVON[393]
EDAGIZ[394]
EDAGUL[394]
EDAWAH[395]
EKIGEK[396]
EPAQES
EQOPAB[397]
ERAFAE[398]
ERAFEI[398]
ETELOF[399]
EXIFAT[400]
EXIFEX[400]
GIDMAI[401]
HACYER[402]
HAFVEQ[403]
HAVQUQ[404]
HAYCOZ[405]
HAYCOZ01[406]
HAYYOW[407]
HEBHOM[408]
HELGUC[409]
HELHEN[409]
HELHIR[409]
HETMEZ[410]
HEVXIP[411]
HIZBEY[412]
HMTFCQ[413]
HMTNTI[414]
HORVOA[415]
HORVUG[415]
HORXOB[416]
HORXOB01[416]
HORXUH[416]
HOVDAY[417]
HUKPIM[418]
HUMLOQ[419]
HURYIC[420]
IKUHUR[421]
IKUHUR01[422]
IKUJAZ[421]
IKUJIH[421]
IKUJON[421]
IKUJUT[421]
IQEWIK[423]
ISIJEZ[177]
JAQMEU[424]
KABLAC[425]
LEZPIQ[426]
LEZPOW[426]
LEZPUC[426]
LEZQAJ[426]
LUKMIN[427]
MALGUB[428]
MASVUZ[429]
MASWAG[429]
MASWEK[429]
MAXBET[430]
MAXBIX[430]
MECXAT[431]
MEGXOL[432]
MIYKOU[433]
MOCCEM[434]
MODVEG[435]
MOFFUI[436]
MORIPA01[437]
MOZNEU[438]
MUGLAB[439]
MULYAU[440]
MULYOI[440]
MUQKOZ[441]
MXTTCQ01
NEBXUP[442]
NEDROF[443]
NUGPOV[444]
NUGPUB[444]
OCOMUO[445]
PASLAY[446]
PIFVIK[447]
QABZUQ[448]
QACBAZ[448]
QIHBAK[449]
QIHBEO[449]
QIHBEO01[445]
QIHCAL[449]
QIHCAL01[450]
REZDEG[451]
RIWZIH[452]
RUYHID[453]
RUYHOJ[453]
RUYHUP[453]
RUYJAX[453]
SAKLAR[454]
SAMTEF
SEHJAS[455]
SENKIG[456]
SETWOD[457]
SETWOD10[458]
SEVWIA[459]
SIBDUC[460]
SIBFAK[460]
SIBFEO[460]
SIBGUF
SIKCIY[461]
SIMHAY[462]
SIVBAA[463]
SOWLOF[177]
SOXJIY
SUBQAH
SUBQIP
SUWYIT[464]
TIWNET[465]
TIWNIX[465]
TOJBOK[466]
TOJBUQ[466]
TOJCEB[466]
UWOFES[467]
VABNUJ[468]
WEXVUR[469]
WEXWAY[469]
WEXWEC[469]
WUZMUZ[470]
YAMZOC[471]
YANGAX[472]
YANGEB[472]
YARNAG[473]
YATJUA[474]
YAVXEA[475]
YAZTIC[476]
YESZAX
YESZEB
YIPCEG[477]
YIPCIK[477]
YISDIN[478]
YISPAS[479]
YOLJIS[480]
YUHLAP[481]
YUHLET[481]
YUHLIX[481]
YUHLOD[481]
YURPAD[482]
YUWNEJ[483]
ZAGKOH[484]
ZAJDIX[485]
ZAPNAF
ZARFUV[486]
ZARQEO[487]
ZAYQEV[488]
ZEBLAV[489]
ZEFKOM[490]
ZEFWAK[491]
ZEKQAJ[492]
ZEKQEN[492]
ZELZOF[493]
ZIHVIV[494]
ZONYOQ[495]
ZPHCYQ
ZPHCYQ10[496]

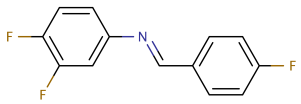
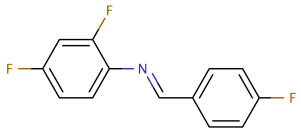
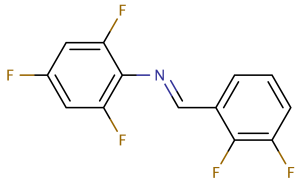
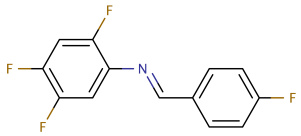
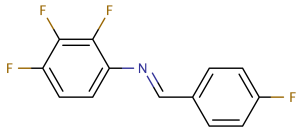
ZUGRUO[497]	ZUPJID[498]	ZUPKUQ[498]	ZZZGKE01[500]
ZUPJEZ[498]	ZUPJOJ[498]	ZUZDUT[499]	ZZZOZY01[501]

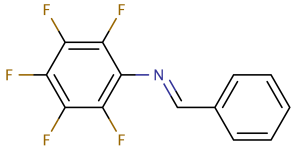
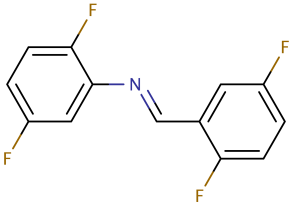
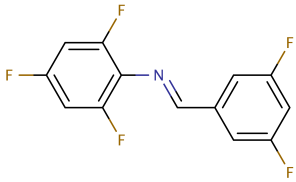
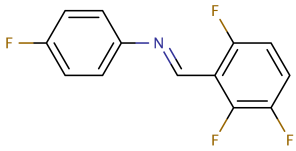
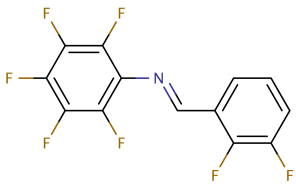
Appendix B

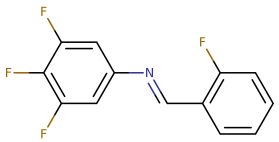
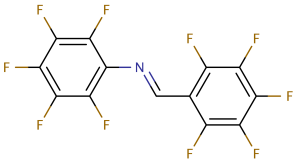
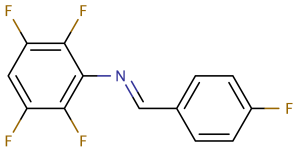
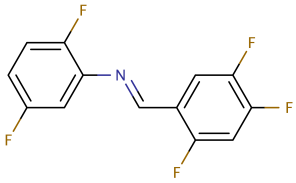
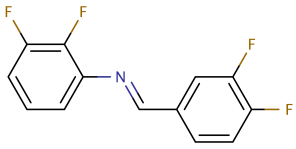
X-Ray Experimental Specifics

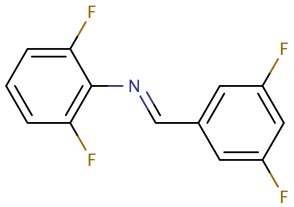
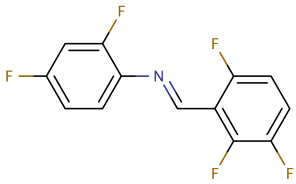
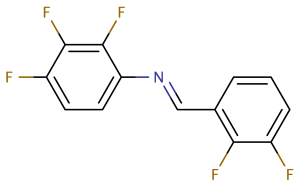
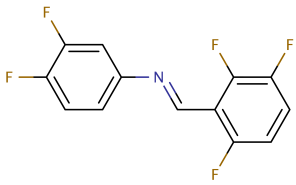
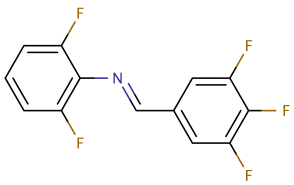
Table B.1: X-ray Experimental Details for Fluorobenzanilide Compounds

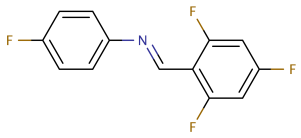
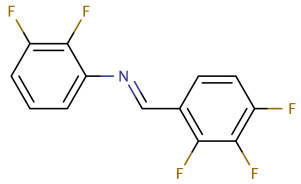
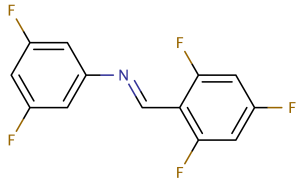
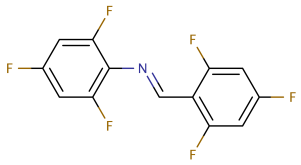
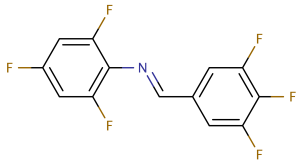
ID	Compound Structure	Setup	T (K)	Collector	Supervisor	CSD Refcode
1		Del Boy	120	Samuel O. Ling	Graham Tizzard	KOMZAP[177]
2		Del Boy	120	Samuel O. Ling	Graham Tizzard	KOMZIX[177]
3		Del Boy	120	Samuel O. Ling	Graham Tizzard	KOMZOD[177]

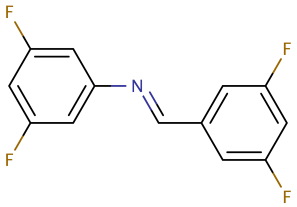
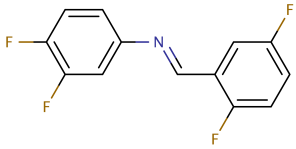
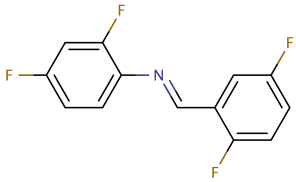
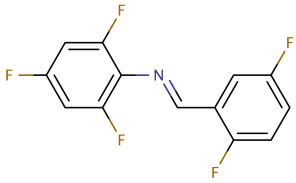
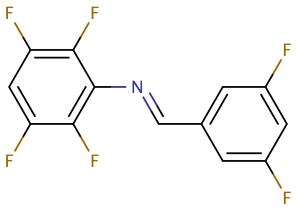
ID	Compound Structure	Setup	T (K)	Collector	Supervisor	CSD Refcode
4		Del Boy	120	Samuel O. Ling	Graham Tizzard	KOMZET[177]
5		Del Boy	120	Samuel O. Ling	Graham Tizzard	KOMYUI[177]
6		Del Boy	120	Samuel O. Ling	Graham Tizzard	KOMYOC[177]
7		Del Boy	120	Samuel O. Ling	Graham Tizzard	KONBAS[177]
8		Del Boy	120	Samuel O. Ling	Graham Tizzard	KOMYIW[177]

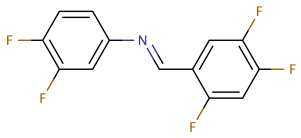
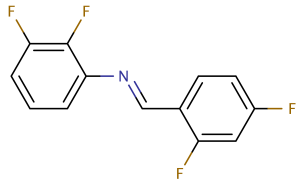
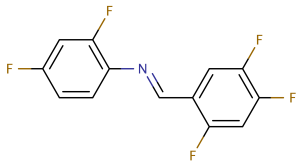
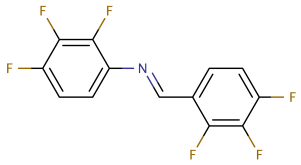
ID	Compound Structure	Setup	T (K)	Collector	Supervisor	CSD Refcode
9[ii]		Damien	120	Philip Adler	Graham Tizzard	N/A
10[ii]		Del Boy	120	Philip Adler	Graham Tizzard	BUCLUI[256]
11		Spider	120	Eleanor Dodd	Philip Adler	N/A
12		Spider	120	Eleanor Dodd	Philip Adler	N/A
13		Spider	120	Eleanor Dodd	Philip Adler	N/A

ID	Compound Structure	Setup	T (K)	Collector	Supervisor	CSD Refcode
14		Spider	120	Eleanor Dodd	Philip Adler	N/A
15		I19	100	Eleanor Dodd	Philip Adler, Graham Tizzard	N/A
16[ii]		Del Boy	120	Philip Adler	Graham Tizzard	N/A
17[ii]		Del Boy	120	Philip Adler	Graham Tizzard	N/A
18[ii]		Del Boy	120	Philip Adler	Graham Tizzard	N/A

ID	Compound Structure	Setup	T (K)	Collector	Supervisor	CSD Refcode
19[ii]		Damien	120	Philip Adler	Graham Tizzard	N/A
20[ii]		Damien	120	Philip Adler	Graham Tizzard	N/A
21[ii]		Del Boy	120	Philip Adler	Graham Tizzard	N/A
22[ii]		Del Boy	120	Philip Adler	Graham Tizzard	N/A
23[ii]		Del Boy	120	Philip Adler	Graham Tizzard	N/A

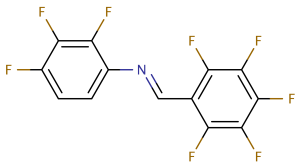
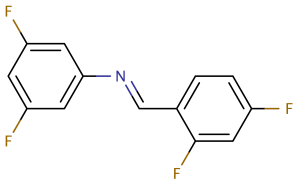
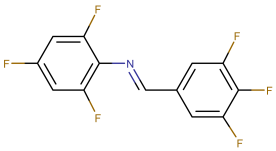
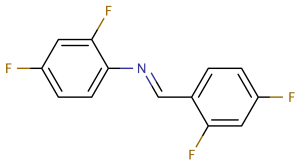
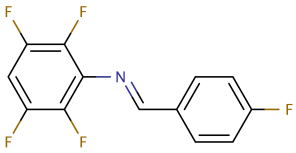
ID	Compound Structure	Setup	T (K)	Collector	Supervisor	CSD Refcode
24[ii]		Del Boy	120	Philip Adler	Graham Tizzard	N/A
25[ii]		Del Boy	120	Philip Adler	Graham Tizzard	N/A
26[ii]		Del Boy	120	Philip Adler	Graham Tizzard	N/A
27[ii]		Del Boy	120	Philip Adler	Graham Tizzard	N/A
28[ii]		Del Boy	120	Philip Adler	Graham Tizzard	N/A

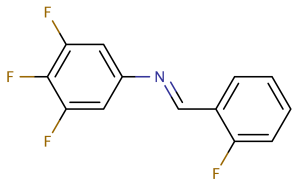
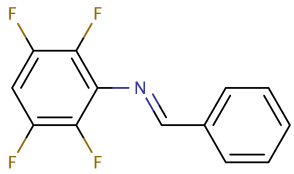
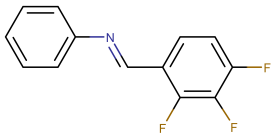
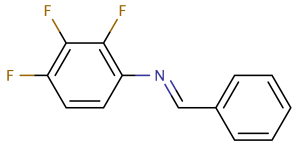
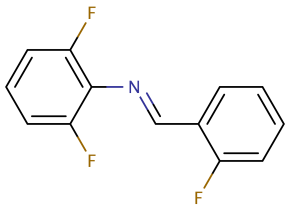
ID	Compound Structure	Setup	T (K)	Collector	Supervisor	CSD Refcode
29[ii]		Del Boy	120	Philip Adler	Graham Tizzard	N/A
30[ii]		Del Boy	120	Philip Adler	Graham Tizzard	N/A
31[ii]		Del Boy	120	Philip Adler	Graham Tizzard	N/A
32[ii]		Del Boy	120	Philip Adler	Graham Tizzard	N/A
33[ii]		I19	100	Philip Adler	Graham Tizzard	N/A

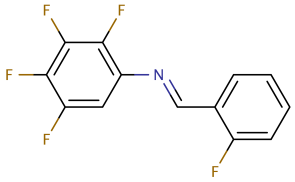
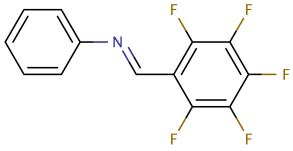
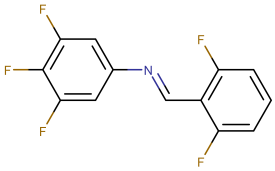
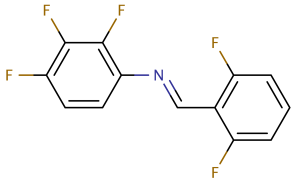
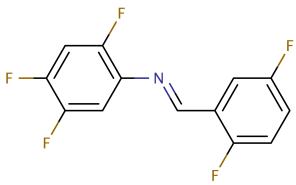
ID	Compound Structure	Setup	T (K)	Collector	Supervisor	CSD Refcode
34[ii]		I19	100	Philip Adler	Graham Tizzard	N/A
35[i]		Del Boy	120	Philip Adler	Graham Tizzard	N/A
36[ii]		Del Boy	120	Philip Adler	Graham Tizzard	N/A
37		Dot	100	Philip Adler	N/A	N/A

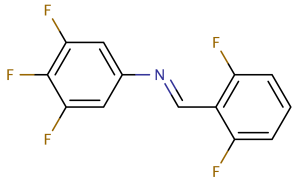
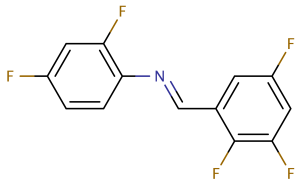
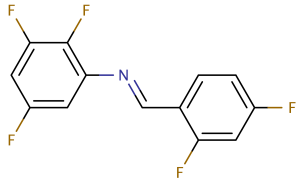
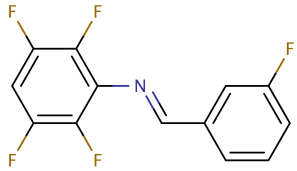
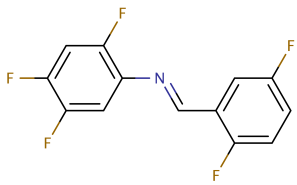
[i] The data for this sample was collected before the formal start of the Author's Ph. D. course, but was refined during said course.

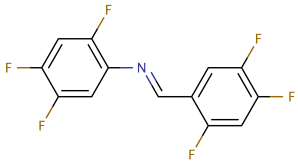
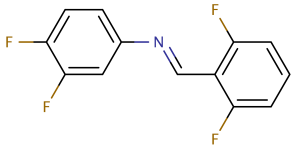
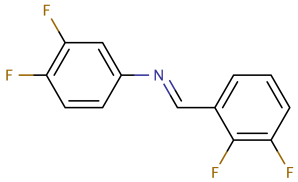
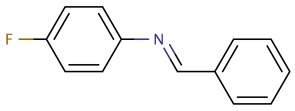
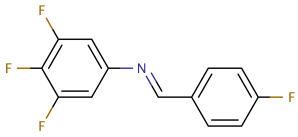
[ii] It should be noted for the purposes of the regulations of the University of Southampton, that this sample's data was collected and processed prior to the formal start of the Author's Ph. D. course.

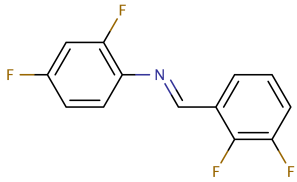
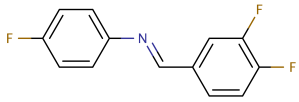
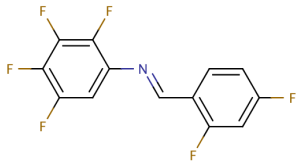
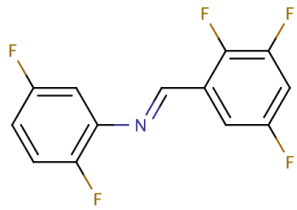
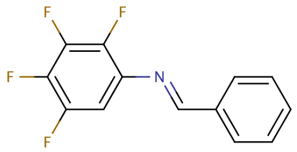
ID	Compound Structure	Setup	T (K)	Collector	Supervisor	CSD Refcode
38		Spider	120	Philip Adler	N/A	N/A
40		Dot	100	Philip Adler	N/A	N/A
41		Ros	100	Philip Adler	N/A	N/A
43		Dot	120	Philip Adler	N/A	N/A
44		I19	100	Philip Adler	N/A	N/A

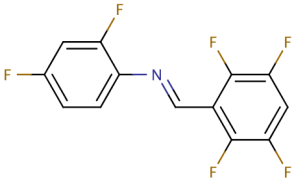
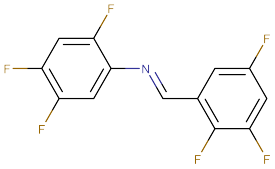
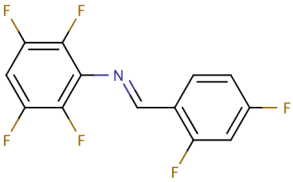
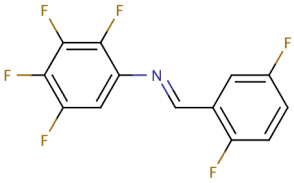
ID	Compound Structure	Setup	T (K)	Collector	Supervisor	CSD Refcode
45		Spider	120	Philip Adler	N/A	N/A
46		Spider	120	Philip Adler	N/A	N/A
47		Dot	100	Philip Adler	N/A	N/A
48		Spider	120	Philip Adler	N/A	N/A
49		Kat	100	Philip Adler	N/A	N/A

ID	Compound Structure	Setup	T (K)	Collector	Supervisor	CSD Refcode
50		Spider	120	Philip Adler	N/A	N/A
51		Ros	100	Philip Adler	N/A	N/A
52		Spider	120	Philip Adler	N/A	N/A
53		Ros	100	Philip Adler	N/A	N/A
54		Kat	100	Philip Adler	N/A	N/A

ID	Compound Structure	Setup	T (K)	Collector	Supervisor	CSD Refcode
55		Ros	100	Philip Adler	N/A	N/A
56		Kat	100	Philip Adler	N/A	N/A
57		Kat	100	Philip Adler	N/A	N/A
58		Dot	100	Philip Adler	N/A	N/A
59		Dot	100	Philip Adler	N/A	N/A

ID	Compound Structure	Setup	T (K)	Collector	Supervisor	CSD Refcode
60		Ros	100	Philip Adler	N/A	N/A
61		Spider	120	Philip Adler	N/A	N/A
62		Ros	100	Philip Adler	N/A	N/A
63		Ros	100	Philip Adler	N/A	N/A
64		Ros	100	Philip Adler	N/A	N/A

ID	Compound Structure	Setup	T (K)	Collector	Supervisor	CSD Refcode
65		Kat	100	Philip Adler	N/A	N/A
66		Ros	100	Philip Adler	N/A	N/A
67		Ros	100	Liam Oliver	Graham Tizzard	N/A
68		Ros	100	Liam Oliver	Graham Tizzard	N/A
69		Ros	100	Liam Oliver	Graham Tizzard	N/A

ID	Compound Structure	Setup	T (K)	Collector	Supervisor	CSD Refcode
70		Ros	100	Liam Oliver	Graham Tizzard	N/A
71		Ros	100	Liam Oliver	Graham Tizzard	N/A
72		Ros	100	Liam Oliver	Graham Tizzard	N/A
73		Ros	100	Liam Oliver	Graham Tizzard	N/A

Appendix C

Programming Languages

C.1 Preamble

A wide variety of programming languages were encountered and used during the creation of this document. One or two of these were quite esoteric and, as such, a very brief overview of the languages used in this work, and their features, is given here. It should be noted that this is not a technical specification of the languages so much as an overview of the languages in relation to the context of the project.

C.2 Language Features

C.2.1 Design Purpose

The design purpose of a language is the use for which a language was either primarily created, or for which it has become primarily used.

C.2.2 Primary Modus Operandi

Many languages each support many modes of use. That said, each language tends to lend itself most effectively to a particular mode of use. The languages seen here tend to lend themselves to one of the following modes:

Imperative Languages that follow an imperative paradigm translate the code literally into a set of instructions which are processed in the order provided.

Declarative Declarative paradigm languages are less concerned with the how of data manipulation, but instead simply on the right result - therefore the code which is

written may not represent an accurate description of the operations that are actually performed on the data.

Functional Functional languages represent the manipulation of data through functions, which obey the rules that apply to mathematical functions. In general, Declarative languages tend to lend themselves to functional language use.

Procedural Procedural programming represents operations through the use of procedures or subroutines (often mislabelled as being functions). In general it is distinguished from functional programming by the notion of 'state', wherein a procedure can produce different effects for the same output, for instance, by keeping track of information like how many times the procedure has been used. In general, Procedural style implies an Imperative language.

Object Oriented Object Oriented languages represent data as objects. Objects have properties which represent data associated with that object, and methods, which perform calculations and algorithms based on the object with which they are associated and other objects which are given to the method as inputs. The concept is somewhat orthogonal to the other notions listed here, but is most common in Imperative paradigms. In general the data and methods associated with an object are determined by the *class* of an object. A full exploration of this topic can be found in many standard programming texts.[258,260]

C.2.3 Compilation

Languages are often divided into two categories: compiled and interpreted.

Compiled languages are passed though a program called a compiler, which reads the written code and translates it into a bytecode which is directly interpretable by a computer processor. This compilation phase can often be time consuming, but compiled code tends to run more quickly than interpreted code - and so the consideration is a trade off in terms of time saved for calculations against the time required for compiling. Compilers will often also analyse algorithms in order to optimise them for speed.

The notion of interpreted languages is something of a misnomer, since many 'interpreted languages' are actually compiled just before the program is run, in a method aptly called just-in-time compilation. The net result however is the same, insofar as a special compilation phase need not be carried out, saving time. However, algorithms may run more slowly in interpreted languages. The trade off considerations are therefore the inverse as for compiled programming languages.

C.2.4 First Class Citizen

The first class citizen of a language is a data type which supports all of the operations generally available to other entities in the language. More intuitively; it is the data type which is easiest to use within the language to describe information. Sometimes these notions of data type are very concrete (e.g. Integer, Matrix), others are more abstract (e.g. object, as per Object Orientated programming).

C.2.5 Libraries

A programming language, either over time or by design, can feature libraries, which are segments of code that are pre-written and can be included in programs by a programmer. Some libraries make a language more useful than others for a particular purpose.

C.2.6 Parallel Programming

Relatively recently, consumer grade computers have started including multiple processors. Furthermore, institutions will often possess large supercomputers. In either case, the use of multiple processing units is known as parallel computing. Some languages possess the capacity to deal with this innately, and others require third party libraries. It is rare that a language will possess no parallel computing capacity, but the ease of implementation varies from language to language.

C.3 Languages

C.3.1 C++

C++ is an Imperative, Object Oriented, Compiled language[260] which lends itself to a procedural style of programming. It is a compiled language, and its first class citizens are 'objects'. The language is a general purpose language, and is used in many systems. It is standardised according to a specification[262] first generated in 1985 by Bjarne Stroustrup.[261] C++ comes with a very large feature set provided by standard libraries.[260,262] Parallel computing is granted to this language by a variety of third party libraries.[260,261]

C.3.2 FORTRAN

FORTRAN is an older language, having been first specified in 1956 by personnel working at IBM.[263] It is a compiled language, and is equally suited to Procedural and Functional programming, having separate semantics for each style and, as such, can intermix

styles.[209,264] It is an imperative language, and newer editions permit an Object Oriented style.[264] Its design pattern is optimised for use in highly optimised calculation, and as such, input/output methods for this language can be somewhat cumbersome, whilst the language itself contains additional features for use in numerical calculation, including the innate ability to handle complex numbers.[264] The feature set of FORTRAN is smaller than C++, even disregarding libraries. However, such a lightweight feature set has until recently allowed the compiler to make assumptions about data handling which could not be made for C++, rendering some mathematical algorithms to be faster in FORTRAN. Parallel computing is granted to FORTRAN by third party libraries.

C.3.3 PHP

The name PHP is a recursive acronym for "PHP hypertext preprocessor",[265] and refers to the languages original and primary role in the serving of web-pages which required some level of algorithmically calculated output. In relatively recent versions Object Oriented patterns have been included in the language,[266] but the primary mode of operation remains purely Procedural and Imperative, ignoring Object Oriented concepts.[267] The purpose of the language does not really require parallel processing, and so implementing this for this language can be cumbersome, although it is possible. It has a rich set of tools for handling markup languages such as XML because of its origins as a web language. It is an interpreted language, and a string of characters (herein *string*) could be regarded as its first class citizen. There is hitherto no formal specification for PHP.

C.3.4 Python

Python is an interpreted, object oriented, imperative language which is generally used in a Procedural style. Like C++, it is designed to be a general purpose language and so has found a wide range of applications. Parallel processing is provided by a variety of third party libraries. In particular, Python is designed for very rapid development with a very clear, maintainable syntax, and so lends itself to the creation of utility scripts which process lightweight information for the handling by more heavyweight programs. Python is an interpreted language and its first class citizen is the Object.[258,259]

C.3.5 Foo

Foo is a language developed by Prof. Dylan Jayatilaka, and is primarily used in the program TONTO. The language is designed to have a Python-like syntax, and is a compiled language. It compiles to FORTRAN90 (which in turn is compiled into bytecode), and so

is effectively a very terse, Object Oriented dialect of that language. It is an imperative language, and parallel libraries are provided by the implementations of FORTRAN on the system. The word Foo comes from the term Foobar[i], a term commonly used in programming examples, though the documentation maintains that the language is named after the Children's Book "Little Bunny Foo Foo".[268]

C.3.6 R

R is a reimplementation of the statistical language 'S'. It is designed primarily for data handling and statistical analysis, and has a wealth of libraries to support that end. Parallel processing is easily implementable with libraries provided by the language. The language lends itself to a functional style. Imperative elements are available, but tend to result in slow code, and are frowned upon in the communities which use this language. Its First Class Citizens are the 'data frame' (analogous to a data table in most respects) and the Matrix. It is an interpreted language and some Object Oriented features are now beginning to be implemented in the language.[269–271]

[i] A cacography of the term FUBAR; an acronym which loosely means 'broken beyond repair'

Appendix D

Data Formats

D.1 Preamble

A large part of cheminformatics is in fact purely data handling.[190] This necessitates the translation of data into both human and machine readable formats. Here follows a brief discussion about those which are relevant to the study at hand.

D.2 .cif Files

D.2.1 Origin, syntax and content

The .cif file was formally defined in 1991,[210] and is the most commonly used standard for the transfer of crystallographic information (cif in fact stands for crystallographic information file). The general syntax is one of tags and values separated by whitespace characters (tabs, spaces, carriage returns etc.):

```
data_signifies_a_data_block
_this_is_a_tag      'this is a string value'
_this_is_a_second_tag 0.2532 #this is a numerical value.
#The area after the hash is a comment.
#comments are ignored by the language interpreters
#but must still obey line lengths.
_this_is_another_tag
;
This is a long body of text with some verbose detail. Such lines
   become
longer than was commonly permitted on machines during the time when
```

```
.cif was developed so a special syntax was designed so that multi-
    line
pieces of text could be represented
;
```

There are also tags to abbreviate repeated uses of tags, among other functions. A full exposition of all tags which exist for the .cif format is contained in the International Tables for Crystallography Volume G.[211]

As such, it fits the definition of being a markup language, however, its inception predates that of XML (Extensible Markup Language), and although SGML was present at the time, SGML is highly complex,[272] and required more computing resource to implement than was generally available at the time.[211] As such it is not an SGML language, and it is not practicable nor necessary to create such a schema retroactively.

Like SGML languages, .cif files are self-defining - the dictionary which defines .cif files is also itself a .cif file.[273]

There are a number of weaknesses in the .cif file format. In general, it requires a specific parser, not being in a common markup scheme (such as XML). Furthermore, a legacy from its time is that it is constrained to 80 characters per line[210i]. Such limitations added necessary complications to the general syntax such that data longer than 80 characters could still be retained. Although the 80 character limit has been removed,[211] many commonly used, if ancient, programs retain the limit, thus preserving the status quo. This problem is compounded by the fact that researchers are rarely rewarded for creating or updating such tools, a problem which was noted throughout this study, and will be observed again in Section 4.3.

Furthermore, the .cif dictionaries are not updated often, which in the current time of rapidly moving technologies can present difficulties when trying to include additional functionality into the .cif files, although there are facilities to include custom tags into a .cif file.[211]

Such practical considerations have largely been addressed however, since software exists already to cope with such matters. Nevertheless, other problems remain; a key one for the purposes of this study being a lack of semantics within the data. For instance, .cif files possess no abstract concept of a molecule. This makes sense given the origin of the file format - it must be remembered that not all crystal structures are molecular,[274] Nevertheless, it can make processing crystal structures in terms of molecules problematic,

[i]This was a common constraint - FORTRAN77, a commonly used language of the period had a similar constraint of 72 characters per line, as do the instruction files for the commonly used crystallographic program SHELX.[205,209]

as shall be seen later.

Another problem which has not been rectified is the fact that the STAR (Self-defining Text Archive and Retrieval) syntax on which the cif file is based does not permit non-ASCII (American Standard Code for Information Interchange) characters to be included within the file. This creates large problems for internationalisation of the standard which modern text encodings that are compatible with ASCII (for instance Unicode) can overcome, although this does not appear to have yet been implemented.

D.3 .mol and related files

.mol files are generally a catch-all term used by various programs for storing data about molecules.[275,276] The abbreviation is generally short for 'molecule file', although often they can contain reactions schemes or multiple molecules.[275] Some have alternative file extensions that can be used, for instance TRIPOS .mol files can have the file extension .sybyl. These files are normally described by proprietary specifications, the licensing for which is frequently unclear. Some specifications are released and intended for wider use.[275] As a result of this design philosophy, different types of .mol file are better equipped to contain certain data, and so where they are utilised by third party programs some styles are favoured over others for specific purposes. For instance, a sybyl molfile could be preferred where charge calculations are concerned because the specification permits detailed charge data to be stored as a part of the file. Their specific nature also tends to mean that for their use a parser needs to be written, since libraries either do not exist or are not well documented.

D.4 .cml files

By contrast to mol files, .cml (chemical markup language) files are clearly intended to be used to transfer data between users and programs. Chemical markup language is an application of XML,[277] and this general standard means that parsers are more widely available/more readily implemented. It is an open standard although it is not maintained by a standards organisation.[278] As a result of being an XML application, it inherits the extensibility of that format,[279] but it cannot in its default state contain reaction schemes, it is able to replicate most, if not all, of the capacities of the .mol file formats.

The following example cml file describes methane:

```
<?xml version="1.0" encoding="MacRoman"?>
```

```
<cml xmlns="http://www.xml-cml.org/schema" xmlns:convention="http:
   //www.xml-cml.org/convention" convention="convention:molecular"
   xmlns:marvin="http://www.chemaxon.com/marvin/marvinDictRef"
   version="ChemAxon file format v6.2.0, generated by v6.2.0">
<molecule id="m1">
        <atomArray>
                <atom id="a1" elementType="C" x2="
                   1.1549999713897705" y2="7.204999923706055"></
                   atom>
                <atom id="a2" elementType="H" x2="
                   1.1549999713897705" y2="8.744999923706054"></
                   atom>
                <atom id="a3" elementType="H" x2="
                   2.6949999713897705" y2="7.204999923706055"></
                   atom>
                <atom id="a4" elementType="H" x2="
                   1.1549999713897705" y2="5.664999923706055"></
                   atom>
                <atom id="a5" elementType="H" x2="
                   -0.3850000286102295" y2="7.204999923706055"></
                   atom>
        </atomArray>
        <bondArray>
                <bond id="b1" atomRefs2="a1 a2" order="1"></bond>
                <bond id="b2" atomRefs2="a1 a3" order="1"></bond>
                <bond id="b3" atomRefs2="a1 a4" order="1"></bond>
                <bond id="b4" atomRefs2="a1 a5" order="1"></bond>
        </bondArray>
</molecule>
</cml>
```

As can be seen, the bracketed "tags" contain either further tags, or data, and indicate information about the data they contain. Such tags are defined explicitly in the cml schema.[277]

Chemical markup language files are not without their weaknesses however. Whilst, unlike .mol files, it has the semantic notion of a molecule, this semantic notion of a 'molecule' this notion was erroneously implemented within the specification to refer to any chemical entity (an ion, for instance, or a group of molecules in a substance).[277] To generalised to these cases, 'moiety' would have been a much more appropriate, general term. In general, the philosophical viewpoint of the file format appears to stem from the

common organic drawing of a molecule, and this limits it when attempting more nuanced descriptions of molecules than the bonds-between atoms perspective. By way of example, delocalised systems struggle for a description in cml.

The significance of this is manifest when one needs to separate the molecular entities in a cml file for use by programs which require files which only contain single molecules (for instance, many descriptor calculation programs). The algorithms to do this necessarily become much more complex and commensurately slower as a result of this seemingly philosophical oversight.

Appendix E

Crystallographic Software

E.1 Abandonware

- *ABSCYL* - A program for absorption correction of needle crystal forms.[502]
- *ABSEN* - "For the study and display of crystal structures".[508]
- *Altwyk* - Produces general and Wyckoff positions for many space groups[502]
- *BAXMAP* - Allows non-crystallographic transformations of SHELX output files[509]
- *CAF2* - Refinement of Harmonic Approximation Parameters.[502]
- *CUBINDEX* - Indexing software for Cubic Systems. Claimed to perform 'additional tasks'.
- *CVIS* - Crystal visualisation software.[502]
- *Chekcel* - Software for finding alternative space group settings using an alternative measure to FOM.[502]
- *Chem-Ray* - Chemical graphics software for windows 95.[502]
- *CifSieve* - .cif file parsing library.[510]
- *DATARED* - Data reduction package[502]
- *DATCOR* - Semi empirical absorption correction package[502]
- *DEF4* - Plane Wave Topography Program[502]

- *DEFW* - X-ray Topography Simulation Program[502]
- *DIMS* - Direct methods solution program for incommensurate crystal structures.[502]
- *DPLOT* - 2d Plotting Program.[502]
- *DREMABLP* - Single Crystal Data Reduction Program.[502]
- *DataTheif* - "Program to reverse engineer scanned graphs to datapoints"[502]
- *EIKONA 3D* - 3d visualisation software.[502]
- *FOCUS* - Model Depiction and Electron Density Map Program[502]
- *Fhkl* - Calculation of structure factors from .hkl file.[502]
- *GRASP* - Visualisation program.[502]
- *GULP* - X-ray simulation software.[502]
- *GraphEnt* - Crystal Structure Determination program using the Maximum Entropy Axiom[511]
- *INTLDM* - X-Ray raw data indexing and integration software.[512]
- *ISODISPLACE* - web interface for ISOTROPY. Discontinued.[513]
- *Jas* - Image Contrast Enhancement software for diffraction images.[502]
- *Java Stereograms* - "plots stereographic projections of poles onto a Wulff net"[502]
- *LCC Cell* - Crystal Structure Refinement Program.[502]
- *MAINDEX* - Manual indexation of diffraction data.[514]
- *MIMS* - Software for modelling incommensurate structures.[502]
- *MODPLT* - Software for modelling modulated structures.[502]
- *MOLGEN* - Automatic Structure Elucidation.[502]
- *MOMO* - Modelling software for organic structures.[502]
- *MOPRO* - Structure and Charge Density refinement.[502]
- *MULTAN88* - Structure resolution program.[502]

- *Molecular Studio* - Functionality not described, home page gone.[502]
- *Oscail* - Shell Structure in the fashion of Wingx.[502]
- *Quantum Image* - Image Processing program.[502]
- *Quasitiler* - Draws penrose tilings (quasicrystal patterns)[515]
- *RES2INS* - Converts .res files to .ins files for use with SHELX[516]
- *SAPI* - Structure determination package.[502]
- *SDP for Windows* - Structure determination package.[502]
- *SDS* - Structure determination package.[502]
- *SFAC331* - Calculates structure factors for an X-ray structure.[502]
- *STOE IPDS* - Calculates intensity of Twinned or grown-together crystals.[502]
- *ShakePSD* - Structure solution and refinement. Former competitor to SHELX.[517]
- *VOID* - Searches and displays voids in crystal structures.[502]
- *WYCKSPLIT* - Determination of Wyckoff positions for a group-subgroup pair.[502]
- *XABS2* - Empirical absorption correction program.[502]
- *XAct* - Database for storing crystallographic experiment data.[502]
- *XITE* - X-based diffraction image processing program.[502]
- *XMol* - Molecular structure viewer and format converter.[502]
- *XPMA* - Mouse driven menu based graphical program for the manipulation of crystal structures.[502]
- *XRDA* - Complete X-ray data handling program.[518]
- *XTAL4POV* - Crystal shape drawing program.[502]
- *Xtal-3d* - 3d visualisation of crystal structures.[502]
- *XtalView* - "Package for fitting electron density maps and solution of structures by MIR and MAD"[502]

- *Zldb* - Data framework for crystallographic results.[502]
- *ALCHEMY II* - Visualisation program.[520]
- *ATOMS* - Visualisation program.[520]
- *BALL & STICK* - Visualisation program.[520]
- *CHEMMOD II* - Visualisation program.[520]
- *CRYSTAL STRUCTURE and LATTICE ENERGY* - software for BBC microcomputers, aimed at old-style 'A'-level students.
- *MOLDRAW* - Visualisation software.[520]
- *NEMESIS* - Visualisation software.[520]
- *PCPDFWIN* - Search program for the old International Centre for Diffraction Data database.[520]
- *aixCCAD* - Molecular dynamics calculator specifically for ionic structures.[504]
- *XSEED* - Overlay for SHELX.[504]
- *DIMS* - Incommensurate structure solution program using *ab-initio* methods[504]
- *OASIS* - Direct method phasing software.[504]
- *DIRAX* - Indexing software.[504,521]
- *Queen of Spades* - 'A stochastic approach to molecular replacement'[504]
- *LinGX* - Linux equivalent to WinGX.[504]
- *asf88* - Calculates atomic/ionic scattering factors.[523]
- *CSDSHL* - Converts old CSD atom coordinate files into SHELX files.[524]
- *CIFtbx2* - Fortran library for manipulation of .cif files.[525]
- *IVTON* - Program for calculation of geometric aspects of inorganic crystal structures.[526]
- *CELLTR/HKLTR/COORDTR* - Transformation program for cell data, Miller indices and atomic coordinates.[527]

- *ATOMCHAR* - Calculates the atomic charges in a molecule. Abandonware.[504]
- *XANADU* - An open source Fortran program which calculates vibrational modes, torsion angles and least squares planes, among other descriptors.[307]
- *CRYC3D* - A program that allowed geometric parameters of crystal structures to be calculated, including vector operations; this may have been similar to the internal representations of XPac.
- *VIBRATE!* - Identifies and calculates irreducible representations of vibration modes in a crystal lattice.[306]
- *WinXPRO* - Program to calculate electronic properties of a crystalline system.[306]
- *RELEXPL* - A Program compatible with X-PLOR for calculating electron density maps.[520]
- *VOLCAL* - Calculates polyhedron volumes, with the implicit intention of calculation of molecular volumes.[520]
- *UNISOFT* - advertised as being able to calculate 'lattice dynamical calculations'.[528]
- *TOPXD* - Claimed to examine topology based upon electron density considerations.[529]
- *SADIAN91* - a program which can calculate distances and angles in crystal structures.[520]
- *STRUCTURE TIDY* - Places inorganic crystal structures into a standardised space group for comparison using atomic coordinates (which are also standardised).[530]
- *SEXIE* - This program, it was claimed, calculated coordination shells and geometries.[520]
- *Tessel* - A '3D compiler' to produce crystal and molecular models, parametric surfaces and several forms of sphere tessellations.[520]
- *PRO-CHEMIST* - Modelling program with additional functionality for PCA based on molecular descriptors and dynamic energy minimisation.[531]
- *BALSAC* - Program to generate lattices, surfaces and clusters for analysis[520]

- *CALCRYS* - a piece of software specifically for the calculation of distance vectors within molecules.[532]
- *crystana* - Calculates some graph theoretic descriptors for silicates via a web interface.[533]
- *HYPERCHEM* - Molecular dynamics simulation software.[520]

E.2 Commercially Available

- *ATOMS* - Atomic visualisation program.[534]
- *SHAPE* - Tool for drawing crystal models.[534]
- *CRYSCON* - Converts between 'popular' file formats.[534]
- *BREADTH* - Calculates line broadening in diffraction patterns[502]
- *BUNYIP* - Detects additional symmetry elements in crystal structures[502]
- *CSD* - Crystal Structure Determination Package for DOS. Possibly abandonware. Not related to the Cambridge Structural Database.[535]
- *POLYSNAP* - Spectroscopic data matching program, with cluster analysis functionality.[536]
- *JCrystal* - Computer program for modelling crystal shapes.[537]
- *Krystal Shaper* - Computer program for modelling crystal shapes.[537]
- *Win-Wulff* - "...a program for plotting stereographic projections of (hkl) and [uvw] onto a Wulff-net or polar net."[537]
- *Kossel/Kikuchi* - Program for calculating K-Patterns for Periodic Crystals.[537]
- *QuaRef* - Program for calculating lists of reflections for quasi-crystals[537]
- *SPEC/C-PLOT* - programs for diffractometer control, data collection and refinement.[538]
- *PROW* - Program for the integration of weak or overlapped data.[539]
- *GAUSSIAN* - Quantum theory calculation program.[540]

- *Diamond* - Crystal structure visualisation and animation package[541]
- *CRYSCOMP-CRYSDRAW* - "Basic computation and drawing" package for MS-DOS[542]
- *Carine* - Crystallographic calculation, visualisation and instruction tool.[543]

E.3 Free to Academic Software

- *INDEX* - Indexing program for output files of EFLECH.[544]
- *PDFFIT* - Refinement Program for Pair Distribution Function.[545]
- *KUPLOT* - Plotting program for output of DISCUS and PDFFIT.[546]
- *HEAVY* - Solution and Refinement by Heavy Atom search. Possible abandonware.[502]
- *PATGEN* - Manual implementation of Patterson methods. Possible abandonware.[547]
- *LCells* - Unit Cell Search Engine and database.[548]
- *PSILAM* - For the calculation and graphical display of 'multiple diffraction patterns'.[549]
- *ROD* - Refinement of surface Structures from X-ray synchrotron data.[550]
- *SPACER* - "A program to display space group information for a conventional and non-conventional coordinate system. "[551]
- *SPGR4D* - A program for the derivation of (3+1) dimensional symmetry operations (refinement of incommensurate crystal structures).[552]
- *TRY* - A program for the automatic solution and refinement of hard crystallographic problems (large incompleteness of data).[553]
- *TWIN3.0* - A program for testing for merohedrally twinned crystals. Possible abandonware.[554]
- *UMWEG* - A program for calculating and displaying multiple diffraction patterns.[555]
- *WinXMorph* - Crystal morphology visualisation software.[556] Claims to be able to make rough prediction of crystal morphology from .cif files,[557] but this could not be made to work by the author of this report.

- *AnoDe* - Structure solution program.[558]
- *PLATON* - Crystallographic calculation toolkit.[308]
- *TOPOS* - Crystallographic visualisation and analysis program.[559]
- *rPLUTO* - Crystallographic calculation package.[286]
- *Mercury* - Crystallographic visualisation and analysis program.[230]
- *Crystal Explorer* - Crystallographic property visualisation software, with particular capacity for Hirshfield surfaces.[560]
- *SHELX* - Structure solution program.[205]
- *PIXEL* - Program for calculating lattice energies.[220]

E.4 Web Interface Software (Free to Access)

- *COPL* - Finds complete lists of order parameters for a phase transition.[561]
- *INVARIANTS* -"Generate invariant polynomials of the components of order parameters"[561]
- *SMODES* - "Find the displacement modes in a crystal which brings the dynamical matrix to block-diagonal form, with the smallest possible blocks."[561]
- *BRL* - Multiple Bragg Diffraction calculator.[562]
- *GID_sl/TER_sl/TDRS_sl* A group of CGI based programs for calculating reflections from known crystal structures[563]
- *x0h* - Program to calculate crystal susceptibilities to X-rays.[564]
- *FROZSL* - Performs lattice dynamical calculations on a provided lattice.[520]
- *VIBRATZ* - Calculates of vibration modes (Raman, IR) for crystalline compounds.[520]

E.5 Free or Open Source Software

It should be noted that simply because source code is available does not mean that it is currently maintained, or indeed, that it is functional.

- *Crystals* - Resurrected, open source, structure solution and refinement program.[519,520i]
- *cctbx* - A toolbox for crystallographic refinement subroutines in Python.[565]
- *DRAWxtl* - A 3 dimensional display tool for crystal structures[566]
- *EUHEDRAL* (formerly *f*) - refinement of crystal description from reflection intensity.[567]
- *GSAS* - Structure solution software for x-ray and neutron diffraction data.[568]
- *HARDPACK* - Structure Prediction by energy minimisation for the use of poor diffraction data.[569]
- *ISOTROPY* - Software for exploration of space groups, irreducible representations and phase changes.[561]
- *ISODISTORT* - software for exploration of incommensurate and distorted crystal structures.[561]
- *ISOCIF* - Modification of cif files for ISOTROPY suite of programs[561]
- *COMSUBS* - "Find common subgroups of two structures in a re-constructive phase transition."[561]
- *JSV* - Structure Viewer.[570]
- *Jana* - Structure Determination Package. Particularly useful for incommensurate structures.
- *JMap3D* - Display of 3D electron density maps.[571]
- *Fourier Transform Lab* - Program for 2 dimensional FFT calculations common in X-Ray Diffraction.[537]
- *KOQUA2* - Program for calculating lists of reflections for quasi-crystals[572]
- *LAC* - Linear Absorption Coefficient Java Applet[573]
- *LAPODS* - Refinement of lattice parameters using optimal regression.[574]
- *Lauept* - Laue pattern simulation.[502]

[i]The second reference here, whilst unorthodox, is a nice illustration of the age of this software, which dates back to before 1993.

- *LaueX* - Laue Simulation and Calculation Program.[575]
- *MCE* - Electron density visualisation.[576]
- *Mollso* - Electron density visualisation.[577]
- *ORTEP* - Thermal Ellipsoid Plotting Program.[578]
- *Orientation Library* - A generic library for rotating coordinates.[502]
- *PARST* - Calculation of molecular parameters from crystallographic results.[579]
- *RMERGE* - Calculation of R merge factors to assess quality of X-ray data.[502]
- *SAS-OMEGA* - Calculates Hauptman's three-phase structure invariants estimate.[580]
- *SIR2011* - The latest in the SIR family of structure solution and refinement programs using *ab-initio methods.*[581]
- *SUPERFLIP* - Solves small molecule, macromolecular and incommensurate structures, and makes an automated structure refinement attempt.[207]
- *CCSL* - A large library of crystallographic and mathematical subroutines, which can be compiled to a suite of programs.[582]
- *Voxel* - A small program which represents 'sliced data' in three dimensions, with the implicit use for electron density display[502]
- *WinGX* - An interface program to many other crystallographic programs, like SUPERFLIP, SIR, and SHELX.[583]
- *XR-shape* - MSDOS program for drawing crystal habit.[573]
- *XR95* - MSDOS program for calculating X-ray diffraction patterns and viewing crystal structures.[573]
- *XRSV* - MSDOS program for viewing of crystal structures.[573]
- *LMCTEP* - Program for the space-filling representation of atomic crystal structures.[584]
- *Xtal* - Open Source Crystal Structure solution and refinement program.[585]
- *patmat-67* - Fortran source code for subroutines which allow Patterson method structure solution.[507]

- *cryls-68* - Structure factor determination routines in Fortran[507]
- *datap-68* - Absorption correction code in Fortran.[507]
- *Lsqpl-68* - Calculation for molecular planes in Fortran.[507]
- *orfls-69* - Crystallographic least squares refinement calculations in Fortran.[507]
- *weight-69* - Automated weighting scheme routine for crystallographic data in Fortran.[507]
- *fordap-70/fordap-79* - Fourier transform routine for diffraction data.[507]
- *REDUCE-79* - Data reduction program for single crystal diffraction.[507]
- *AGNOST-74* - Crystal orientation code in Fortran.[507]
- *ICON-74* - Fortran implementation of the assembler program ICON8. No explanation is given of the functionality of the code.[507]
- *LINEX-74* - Fortran code library which appears to be for structure refinement. No commenting nor explanation is given to allow deeper interpretation of the code.[507]
- *CAMEL JOCKEY* - Fortran implementation of absorption correction. Uses specific binary format files.[586]
- *xfls-77* - Structure factor refinement by least squares method in Fortran.[507]
- *Exfft* - Fast Fourier transform program for output of the MULTAN program.[507]
- *MULTAN-80* - 1980 version of the MULTAN88 software package.[507]
- *NORMAL* - Fortran program written in the 1980s for the calculation of normalised structure factors.[507]
- *SEARCH* - Electron density peak finding and interpretation program in Fortran.[507]
- *Struplo* - Early Fortran program for creating crystal structure illustrations.[507]
- *block-85* - Least squares refinement program.[507]
- *getpec* - Calculates the space group from the Hall symbol and the symmetry setting.[507]
- *geom* - Some form of crystallographic geometry program.[507]

- *gx* - A package of compatible programs for complete structure determination and refinement, comprising block, cad4, absorb, calcomp, checklist, difabs, fft, ftab, geom, gx, ortep, rbls, refil, scfs, search, sort, stand, wtanal, and xyz.[507]
- *lsq* - A least squares refinement program.[507]
- *rbls* - Rigid body least squares refinement program.[507]
- *sort* - reads output from CAD4, sorts reflections and merges them, rejecting systematic absences.[507]
- *wtanal* - weighting analysis program.[507]
- *hole* - 'Calculates Holes in structures.' Possibly an early void calculation program in Fortran.[507]
- *SHADOW* - Appears to be a diffraction pattern indexing and integration program. No formal documentation.[507]
- *PATSEE - Structure solution program, with Patterson, packing and direct methods.*[587]
- *XLAT* - Program for the refinement of cell constants.[502,507ii]
- *crym* - structure solution and refinement package in Fortran comprised of many smaller programs.[507]
- *rmca* - Program reporting to be for 'the fitting of diffraction data', without specification to powder or single crystal data in the documentation.[507]
- *xyz* - Various manipulations of a crystallographic model.[507]
- *absorb* - Absorption correction program.[507]
- *strumo* - Program for modelling inorganic crystal structures.[507]
- *cascade* -A shell, that allows conversion and visualization of outputs from semi-empirical calculations.[507]
- *DIFABS* - Absorption correction program[507]

[ii] In the software listing where this was found, a reference was given to "B.Rupp, Scripta Metallurgica 22, 1 (1988)", however, the paper in question could not apparently be found in the given issue of that Journal.

- *GTSYM* - Space group information calculated from symbols or number.[507]
- *lhpm* - Code with little documentation. Appears to be another structure solution and refinement package.[507]
- *STRUVIR* - Patched version of STRUPLO.[507]
- *caos* - Crystal structure solution and refinement package.[507]
- *DIRDIF* - Crystal Structure Solution Program.[588]
- *laue* - Examination of Laue symmetry from Shelx programs.[507]
- *ICURVAL* - Precursor to the cifcheck web interface used for checking the validity of crystallographic results.[507]
- *THMA* - Thermal motion analysis program[507]
- *hydrogen* - Program for modelling hydroxyl groups and water molecules.[507]
- *promet* - Very early days crystal structure prediction based on packing energies.[507]
- *Babel* - Translates different crystallographic formats.[507]
- *AtomInfo* - Scattering factors calculated in ANSI C.[507]
- *CRYSTAL* - Visualisation program.[507]
- *drawxtl* - Visualisation program.[507]
- *Space Group Information* - Presumably self explanatory program. No documentation or commented code.[507]
- *ESPOIR* - Translates from French to English as 'hope'. Uses Monte Carlo methods to solve and refine crystal structures as a last ditch effort. Most sources, and the in-code documentation only state that this is used for powder structure determination,[502] however, Other locations also state that this can be used for single crystal data as well. It is the only open source software of it's type.[507]
- *'alpha'* - Thermal expansion tensor Fortran code.[504]
- *ANHARM* - Anharmonic Thermal motion refinement software.[504]
- *Drear* - Absorption correction program.[504]

- *COSET* - Derives potential merohedral and pseudomerohedral twin laws.[504]
- *Crunch* - Crystal Structure Solution Program[504]
- *ABSORB* - Brennan-Cowan X-ray absorption, reflection and dispersion calculation. It is not clear that this is not a different piece of software to 'absorb', also listed.[504,522]
- *DISCUS* - Diffraction Simulation Program.[589]
- *layer* - Reads ASCII formatted reflection data and renders precession-style bitmap.[504]
- *XFIT* - Peak fitting program.[504]
- *ZORTEP* - ORTEP-like crystal structure viewing program.[504]
- *DIFFax* - Structure Determination Program for Faulted and Twinned Crystals.[502,504iii]
- *DS*SYSTEM* - Structure solution and refinement package made from conjoining other individual programs into one executable.[504–506]
- *GAMATCH* - Genetic algorithm based program for face-indexing.[504]
- *Gzwilling* - Integration of single crystal area detector data.[504]
- *Kohl* - Indexing program.[504]
- *STRATEGY* - Aids in calculation of data collection strategy.[503,504]
- *LaueCell* - A program for indexing diffraction data with no a-priori information about the unit cell.[504]
- *CHANGEDAT* - Fortran source code to rotate a crystal structure in PARST format using a rotation and translation matrix provided by the user.[504]
- *CIFPARST* - Translates a .cif file into suitable input for PARST-97[504]
- *CYLABS* - Absorption correction for cylindrical crystals (needles)[504]
- *CSDPARS* - Generates PARST input from a CSD entry of FDAT format.[504]
- *DSTANTAB* - Generates a table of bond distances and angles from slightly adjusted output of PARST.

[iii] According to the cited source - this program has a traditional literature reference: Proc. R. Soc. A (1991) 433, 499-520, however, this could not be accessed for verification.

- *MORPHO* - Generates descriptors of crystal morphology (as opposed to crystal structure morphology)[504]
- *ORDRIFL* - hkl data parsing for the purpose of detecting systematic absences by visual inspection of data.[504]
- *PARS9396* - Creates PARST-97 input from PARST-93 input.[504]
- *PARSTCIF* - Translates a PARST-97 output into a .cif file.[504]
- *PARSTINS* - Translates a PARSt-97 input file into a SHELX-93 input file.[504]
- *PREP97* - Creates a PARST97 or THMV7 input file from the .lst file from SHELX-93 or SHELX-97[504]
- *ROTENER* - Calculates the difference in molecular non bonded potential energies when a subgroup has been rotated about an axis.[504]
- *SPHERABS* - Absorption correction for spherical crystals.[504]
- *STATRIFL* - "Considers the distribution of the observed and unobserved reflections."[504]
- *TORSTAB* - Produces a table of the torsion angles in a molecule.[504]
- *PATE* - Takes input from GSAS and outputs ASCII formatted data suitable for plotting.[504]
- *FOUE* - Reads GSAS binary map and outputs to a common format such as WinGX mapview.[504]
- *Equiv* - Analyses equivalent reflections from Single Crystal Data, before the refinement stage.[590]
- *Prometheus* - Crystal structure refinement program.[504]
- *RASMOL* - Structure visualisation program.[591]
- *remos* - A package for the refinement of modulated crystal structures.[504]
- *QUASI06* - Structure refinement package for quasi-crystals.[504]
- *RMCX* - Reverse Monte Carlo modelling for disordered structures.[504]

- *ROTAX* - determines the twin matrix from the F-obs and F-calc.[504]
- *SCHAKAL* - Visualisation program.[592]
- *SYSTER* - Analysis systematic errors in crystal structures.[593]
- *WINCELL* - Structure determination program.[504]
- *Xtaldraw* - Program for viewing crystal and molecular structures.[504]
- *XY2GSAS* - Program for converting XY format crystal data in to GSAS compatible data.[504]
- *CIF2CELL* - "Generates the geometrical setup of a crystallographic cell for a number of electronic structure programs from data contained in a .cif file"
- *CIFLIB* - Library providing ready access to CIF dictionaries and read write operations on .cif files.[594]
- *enCIFer* - Program for editing .cif files.[595]
- *EXPGUI* - A GUI for the program GSAS[596]
- *JMap3D* - Renders electron density on to isosurfaces.[571]
- *FINDSYM* - Finds the space group of a crystal, given the position of atoms within a unit cell.[520]
- *dSNAP* - Compares intramolecular similarity for crystalline compounds.[597]
- *Tonto* - Generates crystallographic information for use by CrystalExplorer

Appendix F

Digital Appendix file Descriptions

At the end of this appendix chapter will be a blank page with a compact disc attached. This contains a set of files, listed here, which contain data related to the presented work that could not readily be represented in a text format. Here follows a list of the file names, and a brief description of what the file contains, and the format in which it is stored.

fluoroanil_crystal_structures A folder with all of the crystal structure files reported in this thesis. Each folder therein is named with a number corresponding to the crystal structure IDs in Chapter 3.5.2, and each contains the following:

****.cif** The crystallographic information framework file for this compound; '**' is replaced by two numerical digits

****.hkl** The text-format data file containing the reflection information from the diffraction procedure

****.fcf** The structure factor file for the diffraction procedure

Other files may be present in these folders, and arise from the processing of the data. They can generally be disregarded but are included, where available, for completeness.

cocrystals A folder containing the following files:

cmlsep.py A program used for separating individual molecules contained in cml files generated from co-crystal structures

maxCorrs_c.tsv A tab separated value file containing the correlations between the maximal values of molecular descriptors of the co-crystal pairs from Chapter 5.3

diffCorrs_c.tsv A tab separated value file containing the correlations between the difference values of molecular descriptors of the co-crystal pairs from Chapter 5.3

interCorrs_c.tsv A tab separated value file containing the correlations between the difference and maximal values of molecular descriptors of the co-crystal pairs from Chapter 5.3

descriptors.txt A tab separated value file containing the raw values of the descriptors for each component of each co-crystal potential pair

maxs.txt A tab separated value file containing the maximal values for each descriptor from each co-crystal

diffs.txt A tab separated value file containing the difference values for each descriptor from each co-crystal

diffPValues_c.tsv A tab separated value file containing the p values for the descriptor difference correlations

maxPValues_c.tsv A tab separated value file containing the p values for the descriptor maximal value correlations

intPValues_c.tsv A tab separated value file containing the p values for the descriptor difference/maximal value correlations

descriptor_graphs A folder containing .png images of graphs of descriptor pairs. Each axis represents a value for the descriptor for each molecule in the co-crystal, and the colouration of the region represents the outcome as to whether a co-crystal was formed (blue=false, red=true). The names of the files correspond to the labels assigned by the dragon program to the descriptors.

sulphonamides A folder containing two subfolders:

cifs The cif files of the selected sulphonamides

dragonresults.txt The descriptor values of the selected sulphonamides calculated by Dragon

Bibliography

[1] Peter Müller, Regine Herbst-Irmer, and Anthoney Spek. *Crystal Structure Refinement: A crystallographer's Guide to SHELXL*. Oxford Science Publications, 2006.

[2] George E. P. Box, William G. Hunter, and J. Stuart Hunter. *Statistics for Experimenters*. John Wiley and Sons, 1978.

[3] Thomas Gelbrich, Terence L. Threlfall, and Michael B. Hursthouse. "XPac dissimilarity parameters as quantitative descriptors of isostructurality: the case of fourteen 4,5[prime or minute]-substituted benzenesulfonamido-2-pyridines obtained by substituent interchange involving CF3/I/Br/Cl/F/Me/H". *CrystEngComm* 14 (17) 2012, pp. 5454–5464. DOI: 10.1039/C2CE25508A.

[4] Jun Harada, Mayuko Harakawa, and Keiichiro Ogawa. "Torsional vibration and central bond length of *N*-benzylideneanilines". *Acta Crystallographica Section B* 60 (5) 2004, pp. 578–588. DOI: 10.1107/S0108768104016532.

[5] Kresĭmir Molcănov and Vladimir Stilinović. "Chemical Crystallography before X-ray Diffraction". *Angewandte Chemie International Edition* (53) 2014, pp. 638–652.

[6] Otto Glasser. *Dr. W.C. Röntgen*. 1st ed. Charles C Thomas, 1945.

[7] S. W. Wilkins. "Celebrating 100 years of X-ray crystallography". *Acta Crystallographica Section A* 69 (1) 2013, pp. 1–4. DOI: 10.1107/S0108767312048490.

[8] Michael Eckert. "Max von Laue and the discovery of X-ray diffraction in 1912". *Annalen der physik* 524 (5) 2012, A83–A85.

[9] Gautam R. Desiraju. "Cryptic Crystallography". *Nature Materials* 1 2002, pp. 77–79.

[10] Frank H. Allen et al. "Systematic analysis of the probabilities of formation of bimolecular hydrogen-bonded ring motifs in organic crystal structures". *New J. Chem.* 23 (1) 1999, pp. 25–34. DOI: 10.1039/A807212D.

[11] Gautam R Desiraju. "Crystal engineering. From molecules to materials". *Journal of Molecular Structure* 656 (1–3) 2003. Studies in Supramolecular Chemistry and Molecular Structure, pp. 5–15. DOI: 10.1016/S0022-2860(03)00354-5.

[12] A. Anthony et al. "CRYSTAL ENGINEERING: SOME FURTHER STRATEGIES". *Crystal Engineering,* 1 (1) 1998, pp. 1–18.

[13] Ranjit Thakuria et al. "Pharmaceutical cocrystals and poorly soluble drugs". *International Journal of Pharmaceutics* (453) 2013, pp. 101–125.

[14] Brian Moulton and Michael J. Zaworotko. "From Molecules to Crystal Engineering: Supramolecular Isomerism and Polymorphism in Network Solids". *Chemical Reviews* (101) 2001, pp. 1629–1658.

[15] Detlef W. M. Hofmann, Ludmila N. Kuleshova, and Mikhail Yu. Antipin. "Supramolecular Synthons and Crystal Structure Prediction of Organic Compounds". *Crystal Growth & Design* 4 (6) 2004, pp. 1395–1402. DOI: 10.1021/cg049969f.

[16] Jack D. Dunitz and Angelo Gavezzotti. "Molecular Recognition in Organic Crystals: Directed Intermolecular Bonds or Nonlocalized Bonding?" *Angewandte Chemie International Edition* 44 (12) 2005, pp. 1766–1787. DOI: 10.1002/anie.200460157.

[17] Scott M. Woodley and Richard Catlow. "Crystal Structure prediction from first principles". *Nature Materials* 7 2008, pp. 937–946.

[18] Heinrich R. Karfunkel, Frank J.J. Leusen, and Robert J. Gdanitz. "The ab initio prediction of yet unknown molecular crystal structures by solving the crystal packing problem". *Journal of Computer Aided Materials Design* 1 1993, pp. 177–185.

[19] J. P. M. Lommerse et al. "A test of crystal structure prediction of small organic molecules". *Acta Crystallographica Section B* 56 (4) 2000, pp. 697–714.

[20] W. D. S. Motherwell et al. "Crystal structure prediction of small organic molecules: a second blind test". *Acta Crystallographica Section B* 58 (4) 2002, pp. 647–661.

[21] G. M. Day et al. "A third blind test of crystal structure prediction". *Acta Crystallographica Section B* 61 (5) 2005, pp. 611–527.

[22] Graeme M. Day et al. "Significant progress in predicting the crystal structures of small organic molecules - a report on the fourth blind test". *Acta Crystallographica Section B* 65 (2) 2009, pp. 107–175.

[23] David A. Bardwell et al. "Towards crystal structure prediction of complex organic compounds – a report on the fifth blind test". *Acta Crystallographica Section B* (67) 2011, pp. 535–551.

[24] Michael B. Hursthouse. "Packing in Molecular Complexes - Antimony Trichloride with some Polynuclear Aromatics". PhD thesis. University of Southampton: Chemistry, 1965.

[25] Philip Adler. "CHEM3012 Final Report: Smart Labs". Year 3 Project Final Report.

[26] Michael B. Hursthouse and Simon J. Coles. "The UK National Crystallography Service; its origins, methods and science". *Crystallography Reviews* 20 (2) 2014, pp. 117–154. DOI: 10.1080/0889311X.2014.884565.

[27] Prof. Gautam R. Desiraju. "Supramolecular Synthons in Crystal Engineering—A New Organic Synthesis". *Angewandte Chemie International Edition in English* 34 (21) 2003, pp. 2311–2327.

[28] G. Bruno and L. Randaccio. "A refinement of the benzoic acid structure at room temperature". *Acta Crystallographica Section B* 36 (7) 1980, pp. 1711–1712. DOI: 10.1107/S0567740880007030.

[29] Srinu Tothadi, Palash Sanphui, and Gautam R. Desiraju. "Obtaining Synthon Modularity in Ternary Cocrystals with Hydrogen Bonds and Halogen Bonds". *Crystal Growth & Design* 14 (10) 2014, pp. 5293–5302. DOI: 10.1021/cg501115k.

[30] Shaunak Chakraborty, Somnath Ganguly, and Gautam R. Desiraju. "Synthon transferability probed with IR spectroscopy: cytosine salts as models for salts of lamivudine". *CrystEngComm* 16 (22) 2014, pp. 4732–4741. DOI: 10.1039/C3CE42156B.

[31] Joanna A. Bis et al. "Hierarchy of Supramolecular Synthons: Persistent Hydroxyl···Pyridine Hydrogen Bonds in Cocrystals That Contain a Cyano Acceptor". *Molecular Pharmaceutics* 4 (3) 2007, pp. 401–416. DOI: 10.1021/mp070012s.

[32] Clair Bilton et al. "Crystal engineering in the *gem*-alkynol family; synthon repetitivity and topological similarity in diphenylethynylmethanols: structures that lack O—H···O hydrogen bonds". *Acta Crystallographica Section B* 56 (6) 2000, pp. 1071–1079. DOI: 10.1107/S010876810001154X.

[33] Archan Dey and Gautam R. Desiraju. "Supramolecular equivalence of ethynyl, chloro, bromo and iodo groups. A comparison of the crystal structures of some 4-phenoxyanilines". *CrystEngComm* 6 (104) 2004, pp. 642–646. DOI: 10.1039/B416962J.

[34] Md. Badruz Zaman, Masaaki Tomura, and Yoshiro Yamashita. "Crystal Engineering Using Anilic Acids and Dipyridyl Compounds through a New Supramolecular Synthon". *The Journal of Organic Chemistry* 66 (18) 2001, pp. 5987–5995. DOI: 10.1021/j001746i.

[35] Md. Badruz Zaman, Masaaki Tomura, and Yoshiro Yamashita. "New Hydrogen-Bonded Donor-Acceptor Pairs between Dipyridylacetylenes and 2,5-Dichloro-3,6-dihydroxy-1,4-benzoquinone". *Organic Letters* 2 (3) 2000, pp. 273–275. DOI: 10.1021/ol991229q.

[36] Roger Bishop et al. "Role of the $(O–H)_6$ Synthon in the Construction of Organic Inclusion Compounds". *Molecular Crystals and Liquid Crystals Science and Technology. Section A. Molecular Crystals and Liquid Crystals* 356 (1) 2001, pp. 289–297. DOI: 10.1080/10587250108023708.

[37] Srinivasulu Aitipamula et al. "Topological Equivalences between Organic and Coordination Polymer Crystal Structures: An Organic Ladder Formed with Three-Connected Molecular and Supramolecular Synthons". *Organic Letters* 4 (6) 2002, pp. 921–924. DOI: 10.1021/ol017284s.

[38] Arijit Mukherjee, Srinu Tothadi, and Gautam R. Desiraju. "Halogen Bonds in Crystal Engineering: Like Hydrogen Bonds yet Different". *Accounts of Chemical Research* 47 (8) 2014, pp. 2514–2524. DOI: 10.1021/ar5001555.

[39] Jiwen Cai et al. "A novel supramolecular synthon for H-bonded coordination networks: syntheses and structures of extended 2-dimensional cadmium(II) arenedisulfonates". *J. Chem. Soc., Dalton Trans.* (16) 2001, pp. 2370–2375. DOI: 10.1039/B102729H.

[40] Peddy Vishweshwar et al. "Supramolecular synthons based on N-HN and C-HO hydrogen bonds. Crystal engineering of a helical structure with 5,5-diethylbarbituric acid". *Chem. Commun.* (17) 2002, pp. 1830–1831. DOI: 10.1039/B204388B.

[41] V. S. Senthil Kumar et al. "Molecular Complexes of Some Mono- and Dicarboxylic Acids with trans-1,4-Dithiane-1,4-dioxide". *Crystal Growth & Design* 2 (4) 2002, pp. 313–318. DOI: 10.1021/cg025523s.

[42] James M. A. Robinson et al. "Predictable solid state structures incorporating the $C{\equiv}C\text{-}H\cdots O_2N$ supramolecular synthon". *Chem. Commun.* (4) 1999, pp. 329–330. DOI: 10.1039/A809115C.

[43] Philip J. Langley et al. "Supramolecular synthons mediated by weak hydrogen bonding: forming linear molecular arrays via C≡C-H· · ·N≡C and C≡C-H· · ·O_2N recognition". *New J. Chem.* 22 (12) 1998, pp. 1307–1309. DOI: 10.1039/A807552B.

[44] Juan C. Mareque Rivas and Lee Brammer. "Self-Assembly of 1-D Chains of Different Topologies Using the Hydrogen-Bonded Inorganic Supramolecular Synthons N—H···Cl_2M or N—H···Cl_3M". *Inorganic Chemistry* 37 (19) 1998, pp. 4756–4757. DOI: 10.1021/ic9805269.

[45] Marina Felloni et al. "Conserved hydrogen-bonded supramolecular synthons in pyridinium tetrachlorometallates". *CrystEngComm* 6 (19) 2004, pp. 87–95. DOI: 10.1039/B401129E.

[46] Shizheng Zhu et al. "Strong phenyl–perfluorophenyl π–π stacking and C–H· · ·F–C hydrogen bonding interactions in the crystals of the corresponding aromatic aldimines". *Tetrahedron Letters* 46 (15) 2005, pp. 2713–2716. DOI: 10.1016/j.tetlet.2005.01.183.

[47] Alankriti Bajpai, Paloth Venugopalan, and Jarugu Narasimha Moorthy. "Self-Assem-bly of Rigid Three-Connecting Mesitylenetribenzoic Acid: Multifarious Supramolecular Synthons and Solvent-Induced Supramolecular Isomerism". *Crystal Growth & Design* 13 (11) 2013, pp. 4721–4729. DOI: 10.1021/cg400805c.

[48] Bethany I. Harriss et al. "Furosemide Cocrystals: Structures, Hydrogen Bonding, and Implications for Properties". *Crystal Growth & Design* 14 (2) 2014, pp. 783–791. DOI: 10.1021/cg401662d.

[49] Christer B. Aakeröy, Prashant D. Chopade, and John Desper. "Avoiding "Synthon Crossover" in Crystal Engineering with Halogen Bonds and Hydrogen Bonds". *Crystal Growth & Design* 11 (12) 2011, pp. 5333–5336. DOI: 10.1021/cg2009013.

[50] Peddy Vishweshwar, Ashwini Nangia, and Vincent M. Lynch. "Recurrence of Carboxylic Acid—Pyridine Supramolecular Synthon in the Crystal Structures of Some Pyrazinecarboxylic Acids". *The Journal of Organic Chemistry* 67 (2) 2002, pp. 556–565. DOI: 10.1021/j0162484.

[51] Maria Gdaniec et al. "Supramolecular Assemblies of Hydrogen-Bonded Carboxylic Acid Dimers Mediated by Phenyl–Pentafluorophenyl Stacking Interactions". *Angewandte Chemie International Edition* 42 (33) 2003, pp. 3903–3906. DOI: 10.1002/anie.200351432.

[52] Uttam Kumar Das et al. "Supramolecular Synthons in Noncovalent Synthesis of a Class of Gelators Derived from Simple Organic Salts: Instant Gelation of Organic Fluids at Room Temperature via in Situ Synthesis of the Gelators". *The Journal of Organic Chemistry* 74 (18) 2009, pp. 7111–7121. DOI: 10.1021/j901463k.

[53] Christer B. Aakeröy, Alicia M. Beatty, and Brian A. Helfrich. "A High-Yielding Supramolecular Reaction". *Journal of the American Chemical Society* 124 (48) 2002, pp. 14425–14432. DOI: 10.1021/ja027845q.

[54] Peddy Vishweshwar, Ashwini Nangia, and Vincent M. Lynch. "Supramolecular synthons in phenol-isonicotinamide adducts". *CrystEngComm* 5 (31) 2003, pp. 164–168. DOI: 10.1039/B304078J.

[55] L. Sreenivas Reddy, Ashwini Nangia, and Vincent M. Lynch. "Phenyl-Perfluorophenyl Synthon Mediated Cocrystallization of Carboxylic Acids and Amides". *Crystal Growth & Design* 4 (1) 2004, pp. 89–94. DOI: 10.1021/cg034106s.

[56] Christer B. Aakeröy et al. "Establishing Amide· · ·Amide Reliability and Synthon Transferability in the Supramolecular Assembly of Metal-Containing One-Dimensional Architectures". *Inorganic Chemistry* 48 (9) 2009, pp. 4052–4061. DOI: 10.1021/ic801992t.

[57] Srinu Tothadi and Gautam R. Desiraju. "Designing ternary cocrystals with hydrogen bonds and halogen bonds". *Chem. Commun.* 49 (71) 2013, pp. 7791–7793. DOI: 10.1039/C3CC43822H.

[58] Shailesh Upreti and Arunachalam Ramanan. "Structure-Directing Role of Hydrogen-Bonded Dimers of Phenylenediammonium Cations: Supramolecular Assemblies of Octamolybdate-Based Organic—Inorganic Hybrids". *Crystal Growth & Design* 5 (5) 2005, pp. 1837–1843. DOI: 10.1021/cg050100m.

[59] J. Narasimha Moorthy et al. "Crystal Engineering: Identification of a Unique Supramolecular Synthon Based on CO···X Interaction in Halogen-Substituted Aromatic Carboxaldehydes". *Crystal Growth & Design* 3 (4) 2003, pp. 581–585. DOI: 10.1021/cg034001p.

[60] Laura J. Thompson et al. "Supramolecular Behavior of Adenine with Succinic, Fumaric, and Maleic Acids: Tautomerism, Cocrystallization, Salt Formation, and Solvation". *Crystal Growth & Design* 13 (4) 2013, pp. 1464–1472. DOI: 10.1021/cg301561j.

[61] Christopher E. Marjo et al. "Crystal Engineering Involving C—H···N Weak Hydrogen Bonds: A Diquinoxaline Lattice Inclusion Host with a Preference for Polychlorocarbon Guests". *European Journal of Organic Chemistry* 2001 (5) 2001, pp. 863–873. DOI: 10.1002/1099-0690(200103)2001:5<863::AID-EJOC863>3.0.CO;2-2.

[62] V. S. Senthil Kumar et al. "Supramolecular synthesis of brick wall and honeycomb networks from the T-shaped molecule 5-nitrosalicylic acid". *New J. Chem.* 27 (2) 2003, pp. 224–226. DOI: 10.1039/B209350B.

[63] Valeria Ferretti, Valerio Bertolasi, and Loretta Pretto. "Supramolecular aggregation by means of charge-assisted hydrogen bonds in acid-base adducts containing amidinium cations". *New J. Chem.* 28 (5) 2004, pp. 646–651. DOI: 10.1039/B314143H.

[64] Xuefeng Mei and Christian Wolf. "Neutral and Ionic Supramolecular Structures of Unsaturated Dicarboxylic Acids and Acridine: Significance of Molecular Geometry and Proton Transfer". *European Journal of Organic Chemistry* 2004 (21) 2004, pp. 4340–4347. DOI: 10.1002/ejoc.200400396.

[65] Hadi D. Arman et al. "Crystal and Molecular Structures of the 2:1 Cocrystal of 4-Nitrophenylacetic acid and N,N'-bis(pyridin-3-ylmethyl)oxalamide, and with the Thioxalamide Analogue". English. *Journal of Chemical Crystallography* 42 (7) 2012, pp. 673–679. DOI: 10.1007/s10870-012-0298-5.

[66] Maholy Linares and Alexander Briceno. "Solid-state synthesis of head-to-tail photodimers from supramolecular assemblies directed by charge-assisted hydrogen bonds". *New J. Chem.* 34 (4) 2010, pp. 587–590. DOI: 10.1039/B9NJ00608G.

[67] Uttam Kumar Das, Vedavati G. Puranik, and Parthasarathi Dastidar. "Supramolecular Synthon Transferability and Gelation by Diprimary Ammonium Monocarboxylate Salts". *Crystal Growth & Design* 12 (12) 2012, pp. 5864–5868. DOI: 10.1021/cg301242p.

[68] Venu R. Vangala et al. "Dianiline-Diphenol Molecular Complexes Based on Supra -minol Recognition". *Crystal Growth & Design* 5 (1) 2005, pp. 99–104. DOI: 10.1021/cg049967v.

[69] Kumar Biradha et al. "C-HLO Hydrogen bonded multi-point recognition in molecular assemblies of dibenzylidene ketones and 1,3,5-trinitrobenzenes". *J. Mater. Chem.* 7 (7) 1997, pp. 1111–1122. DOI: 10.1039/A607106F.

[70] Kaiqiang Liu and Jonathan W. Steed. "Triggered formation of thixotropic hydrogels by balancing competitive supramolecular synthons". *Soft Matter* 9 (48) 2013, pp. 11699–11705. DOI: 10.1039/C3SM51949J.

[71] G. Pavlović et al. "Supramolecular amide and thioamide synthons in hydrogen bonding patterns of N-aryl-furamides and N-aryl-thiofuramides". English. *Structural Chemistry* 17 (3) 2006, pp. 275–285. DOI: 10.1007/s11224-006-9029-x.

[72] Sumod George et al. "Crystal engineering of neutral N-arylpyrimidinones and their HCl and HNO3 adducts with a C-HO supramolecular synthon. Implications for non-linear optics". *New J. Chem.* 25 (12) 2001, pp. 1520–1527. DOI: 10.1039/B104646M.

[73] Benjamin Isare et al. "Conformational Control of Hydrogen-Bonded Aromatic Bis-Ureas". *Langmuir* 28 (19) 2012, pp. 7535–7541. DOI: 10.1021/la300887p.

[74] Thomas J. Podesta and A. Guy Orpen. "Use of the Ni(dithiooxalate)22- unit as a molecular tecton in crystal engineering". *CrystEngComm* 4 (60) 2002, pp. 336–342. DOI: 10.1039/B203138H.

[75] Unchulee Suksangpanya et al. "Hydrogen-bonded supramolecular architectures in (N-(methylpyridin-2-yl)-amidino-O-alkylurea)copper(ii) halides". *CrystEngComm* 6 (28) 2004, pp. 159–167. DOI: 10.1039/B405875E.

[76] P. Ganguly and G. R. Desiraju. "Long-range synthon Aufbau modules (LSAM) in crystal structures: systematic changes in C6H6-nFn ($0 <= n <= 6$) fluorobenzenes". *Crystal Engineering Communications* 12 (3) 2010, pp. 81–833.

[77] Firas F. Awwadi et al. "The Electrostatic Nature of Aryl—Bromine—Halide Synthons: The Role of Aryl—Bromine—Halide Synthons in the Crystal Structures of the trans-Bis(2-bromopyridine)dihalocopper(II) and trans-Bis(3-bromo
-py
-ridine)
-di
-halo
-cop
-per(II) Complexes". *Crystal Growth & Design* 6 (8) 2006, pp. 1833–1838. DOI: 10.1021/cg060154b.

[78] Binoy K. Saha and Ashwini Nangia. "Ethynyl Group as a Supramolecular Halogen and C≡C—H· · ·C≡C Trimer Synthon in 2,4,6-Tris(4-ethynylphenoxy)-1,3,5-

triazine". *Crystal Growth & Design* 7 (2) 2007, pp. 393–401. DOI: 10.1021/cg060744+.

[79] Andreas Lemmerer, Susan A. Bourne, and Manuel A. Fernandes. "Robust Supramolecular Heterosynthons in Chiral Ammonium Carboxylate Salts". *Crystal Growth & Design* 8 (4) 2008, pp. 1106–1109. DOI: 10.1021/cg701020s.

[80] Yong Yang et al. "Noncovalent Synthesis of Shape-Persistent Cyclic Hexamers from Ditopic Hydrazide-Based Supramolecular Synthons and Asymmetric Induction of Supramolecular Chirality". *Journal of the American Chemical Society* 131 (35) 2009, pp. 12657–12663. DOI: 10.1021/ja9029335.

[81] Marcos D. García et al. "Interplay between Halogen/Hydrogen Bonding and Electrostatic Interactions in 1,1'-Bis(4-iodobenzyl)-4,4'-bipyridine-1,1'-diium Salts". *Crystal Growth & Design* 9 (12) 2009, pp. 5009–5013. DOI: 10.1021/cg901175e.

[82] Padmini Kavuru et al. "Hierarchy of Supramolecular Synthons: Persistent Hydrogen Bonds Between Carboxylates and Weakly Acidic Hydroxyl Moieties in Cocrystals of Zwitterions". *Crystal Growth & Design* 10 (8) 2010, pp. 3568–3584. DOI: 10.1021/cg100484a.

[83] Shahedeh Tayamon et al. "2,2-Dibenzylhydrazin-1-ium chloride". *Acta Crystallographica Section E* 69 (3) 2013, p. 382. DOI: 10.1107/S1600536813003966.

[84] Uttam Kumar Das and Parthasarathi Dastidar. "Extending Primary Ammonium Dicarboxylate (PAD) to Diprimary Ammonium Dicarboxylate (DPAD) Synthon and Its Implication in Supramolecular Gelation". *Crystal Growth & Design* 13 (10) 2013, pp. 4559–4570. DOI: 10.1021/cg401052a.

[85] Kafeel Ahmad Siddiqui and Edward R. T. Tiekink. "A supramolecular synthon approach to aid the discovery of architectures sustained by C-HM hydrogen bonds". *Chem. Commun.* 49 (76) 2013, pp. 8501–8503. DOI: 10.1039/C3CC44808H.

[86] Jonathan Starbuck, Nicholas C. Norman, and A. Guy Orpen. "Secondary bonding as a potential design element for crystal engineering". *New J. Chem.* 23 (10) 1999, pp. 969–972. DOI: 10.1039/A906352H.

[87] Arijit Mukherjee et al. "Aniline–phenol recognition: from solution through supramolecular synthons to cocrystals". *IUCrJ* 1 (4) 2014, pp. 228–239. DOI: 10.1107/S2052252514012081.

[88] Li-Ping Cao et al. "The R 2 2 (8) Hydrogen-Bonded Supramolecular Synthon in Two Novel Glycoluril Derivatives". English. *Journal of Chemical Crystallography* 41 (3) 2011, pp. 425–429. DOI: 10.1007/s10870-010-9963-8.

[89] Basab Chattopadhyay et al. "Supramolecular Architectures in 5,5'-Substituted Hydantoins: Crystal Structures and Hirshfeld Surface Analyses". *Crystal Growth & Design* 10 (10) 2010, pp. 4476–4484. DOI: 10.1021/cg100706n.

[90] Sadhana Singh et al. "Synthesis, Characterization and Photoluminescent Property of a Trinucleated Cadmium (II) Coordination Polymer Involving In Situ Ligand Reaction". English. *Journal of Chemical Crystallography* 43 (2) 2013, pp. 82–90. DOI: 10.1007/s10870-012-0388-4.

[91] Lei Wang et al. "Construction of interesting organic supramolecular structures with synthons cooperation in the cocrystals of 1H-benzotriazole and hydroxybenzoic acids". English. *Science China Chemistry* 55 (12) 2012, pp. 2515–2522. DOI: 10.1007/s11426-012-4652-4.

[92] Tetsuya Nyui, Takashi Nogami, and Takayuki Ishida. "Organic kagome lattice consisting of trimeric pyrazole and oxime supramolecular synthons from 3-pyrazolecar
-boxal
-doxime". *CrystEngComm* 7 (101) 2005, pp. 612–615. DOI: 10.1039/B512977J.

[93] Alajos Kálmán et al. "Basic forms of supramolecular self-assembly organized by parallel and antiparallel hydrogen bonds in the racemic crystal structures of six disubstituted and trisubstituted cyclopentane derivatives". *Acta Crystallographica Section B* 57 (4) 2001, pp. 539–550. DOI: 10.1107/S0108768101006723.

[94] Miao Du, Zhi-Hui Zhang, and Xiao-Jun Zhao. "Cocrystallization of Bent Dipyridyl Type Compounds with Aromatic Dicarboxylic Acids: Effect of the Geometries of Building Blocks on Hydrogen-Bonding Supramolecular Patterns". *Crystal Growth & Design* 5 (3) 2005, pp. 1199–1208. DOI: 10.1021/cg049595q.

[95] Eric A. Bruton et al. "Hydrogen bond patterns in aromatic and aliphatic dioximes". *New J. Chem.* 27 (7) 2003, pp. 1084–1094. DOI: 10.1039/B301045G.

[96] Unchulee Suksangpanya et al. "Complementarity of anion-mediated hydrogen-bonding and alkyl substitution in the construction of two-dimensional rhombic (4,4) grids by bis(N-alkylamidino-O-alkylurea)copper(ii) nitrates and tetrafluoroborates". *CrystEngComm* 5 (4) 2003, pp. 23–33. DOI: 10.1039/B210627B.

[97] Atish Dipankar Jana et al. "Towards rational design of supramolecular helices using linear pseudohalides in Cd(ii) - 2′,-biimidazole system". *CrystEngComm* 9 (4) 2007, pp. 304–312. DOI: 10.1039/B614979K.

[98] Vi T. Nguyen et al. "Crystallisation of C2-symmetric endo,endo-bicyclo[3.3.1]nonane-2-6-diols: supramolecular synthons and concomitant degrees of enantiomer separation". *New J. Chem.* 33 (8) 2009, pp. 1736–1741. DOI: 10.1039/B900463G.

[99] Lian-Cheng Wang et al. "Observation of a Persistent Supramolecular Synthon Involving Carboxyl Groups and H_2O That Guides the Formation of Polycatenated Co-crystals of a Tritopic Carboxylic Acid and Bis(pyridyls)". *Crystal Growth & Design* 13 (1) 2013, pp. 1–5. DOI: 10.1021/cg301015a.

[100] Alessia Bacchi, Elsa Bosetti, and Mauro Carcelli. "Unusual hydrogen bonded $(OH)_4$ tetrahedral nests organize zinc(ii) coordination complexes in a non covalent diamondoid network". *CrystEngComm* 9 (4) 2007, pp. 313–318. DOI: 10.1039/B616955D.

[101] David R. Turner et al. "Ammonium salts of carbamoyldicyanomethanide, $C(CN)_2(CONH_2)$-: Effects of hydrogen-bonding cations on anionic networks". *CrystEngComm* 11 (2) 2009, pp. 298–305. DOI: 10.1039/B815926B.

[102] Jungang Wang et al. "Robust R_{22}^8 hydrogen bonded dimer for crystal engineering of glycoluril derivatives". *CrystEngComm* 15 (46) 2013, pp. 10079–10085. DOI: 10.1039/C3CE41702F.

[103] Venu R. Vangala et al. "Correspondence between Molecular Functionality and Crystal Structures. Supramolecular Chemistry of a Family of Homologated Aminophenols". *Journal of the American Chemical Society* 125 (47) 2003, pp. 14495–14509. DOI: 10.1021/ja037227p.

[104] Rahul Banerjee et al. "Synthon Robustness and Solid-State Architecture in Substituted gem-Alkynols". *Crystal Growth & Design* 6 (4) 2006, pp. 999–1009. DOI: 10.1021/cg050598s.

[105] Man-Li Cao et al. "Assembly of a Cubic Nanocage Co_8L_{12} and a Hydrogen-Bonded 3D NbO Net Based on the $[(HCO_3)_2]_2$— Synthon and Water". *Inorganic Chemistry* 47 (18) 2008, pp. 8126–8133. DOI: 10.1021/ic800585d.

[106] Somnath Ray Choudhury et al. "Robust recognition of malonate and 2-amino-4-picolinium in conjunction with M(ii) as a triad (M = Ni/Co/Mn): role of this highly stable hydrogen-bonded motif in driving supramolecular self-assembly". *Dalton Trans.* (37) 2009, pp. 7617–7624. DOI: 10.1039/B905127A.

[107] Ishtvan Boldog et al. "Hydrogen Bonding Patterns and Supramolecular Structure of 4,4'-Bipyrazolium Salts". *Crystal Growth & Design* 9 (6) 2009, pp. 2895–2905. DOI: 10.1021/cg9002109.

[108] Cédric Borel et al. "Oxalate- and Squarate-Biimidazole Supramolecular Synthons: Hydrogen-Bonded Networks Based on $[Co(H_2biimidazole)_3]^{3+}$". *Crystal Growth & Design* 9 (6) 2009, pp. 2821–2827. DOI: 10.1021/cg900075j.

[109] M. Mirzaei et al. "Comprehensive studies of non-covalent interactions within four new Cu(ii) supramolecules". *CrystEngComm* 14 (24) 2012, pp. 8468–8484. DOI: 10.1039/C2CE26442K.

[110] Inese Sarcevica et al. "Crystal and Molecular Structure and Stability of Isoniazid Cocrystals with Selected Carboxylic Acids". *Crystal Growth & Design* 13 (3) 2013, pp. 1082–1090. DOI: 10.1021/cg301356h.

[111] Anna Portell, Mercè Font-Bardia, and Rafel Prohens. "Self-Assembling of Zwitterionic Squaramides through Electrostatically Compressed Face-to-Face π-Stacking: A New Supramolecular Synthon". *Crystal Growth & Design* 13 (10) 2013, pp. 4200–4203. DOI: 10.1021/cg401161d.

[112] Gautam R. Desiraju. "Designer crystals: intermolecular interactions, network structures and supramolecular synthons". *Chem. Commun.* (16) 1997, pp. 1475–1482. DOI: 10.1039/A607149J.

[113] Valerio Bertolasi et al. "Competition between hydrogen bonding and donor-acceptor interactions in co-crystals of 1,3-dimethylbarbituric acid with aromatic amines". *New J. Chem.* 25 (3) 2001, pp. 408–415. DOI: 10.1039/B008262G.

[114] Adina Lazar et al. "Assembly of a novel supramolecular synthon of calix[4]arene presenting four carboxylic acids". *Chem. Commun.* (8) 2006, pp. 903–905. DOI: 10.1039/B516065K.

[115] Tejender S. Thakur and Gautam R. Desiraju. "Crystal Structure Prediction of a Co-Crystal Using a Supramolecular Synthon Approach: 2-Methylbenzoic Acid—2-Amino-4-methylpyrimidine". *Crystal Growth & Design* 8 (11) 2008, pp. 4031–4044. DOI: 10.1021/cg800371j.

[116] Keisuke Maruyoshi et al. "Identifying the intermolecular hydrogen-bonding supra -molecular synthons in an indomethacin-nicotinamide cocrystal by solid-state NMR". *Chem. Commun.* 48 (88) 2012, pp. 10844–10846. DOI: 10.1039/C2CC36094B.

[117] Peter R. Ashton et al. "Combining Different Hydrogen-Bonding Motifs To Self-Assemble Interwoven Superstructures". *Chemistry – A European Journal* 4 (4) 1998, pp. 577–589. DOI: 10.1002/(SICI)1521-3765(19980416)4:4<577::AID-CHEM577>3.0.CO;2-T.

[118] Mariya E. Brezgunova et al. "Charge Density Analysis and Topological Properties of Hal_3-Synthons and Their Comparison with Competing Hydrogen Bonds". *Crystal Growth & Design* 12 (11) 2012, pp. 5373–5386. DOI: 10.1021/cg300978x.

[119] N. Stanley et al. "Crystal Engineering of Organic Salts: Hydrogen-Bonded Supramolecular Motifs in Pyrimethamine Hydrogen Glutarate and Pyrimethamine Formate". *Crystal Growth & Design* 2 (6) 2002, pp. 631–635. DOI: 10.1021/cg020027p.

[120] Gustavo Portalone. "Supramolecular association in proton-transfer adducts containing benzamidinium cations. I. Four molecular salts with uracil derivatives". *Acta Crystallographica Section C* 66 (6) 2010, pp. 295–301. DOI: 10.1107/S0108270110016252.

[121] Bipul Sarma and Basanta Saikia. "Hydrogen bond synthon competition in the stabilization of theophylline cocrystals". *CrystEngComm* 16 (22) 2014, pp. 4753–4765. DOI: 10.1039/C3CE42332H.

[122] Concepción Foces-Foces et al. "Role of the Molecular Conformation in the Two- and Three-Dimensional Supramolecular Structure of 10 Hydroxyl-N-alkylamides". *Crystal Growth & Design* 7 (5) 2007, pp. 905–911. DOI: 10.1021/cg060671u.

[123] Agata Bialonska and Zbigniew Ciunik. "When a host becomes a guest-competition between decreasing hydrophobic spaces and supramolecular synthon propagation". *CrystEngComm* 13 (3) 2011, pp. 967–972. DOI: 10.1039/C0CE00388C.

[124] Ivan V. Fedyanin and Konstantin A. Lyssenko. "New hydrogen-bond-aided supra -molecular synthon: a case study of 2,4,6-trinitroaniline". *CrystEngComm* 15 (46) 2013, pp. 10086–10093. DOI: 10.1039/C3CE41227J.

[125] Paul V. Bernhardt. "A Supramolecular Synthon for H-Bonded Transition Metal Arrays". *Inorganic Chemistry* 38 (15) 1999, pp. 3481–3483. DOI: 10.1021/ic990074f.

[126] Basem Fares Ali and Rawhi Al-Far. "A novel framework of N—H$\cdots$Br hydrogen bonds forming Br(4,4′-bipyridinium)$_4$ supramolecular synthons: bis(4,4′-bipyridinium) tris[tetrabromidoferrate(III)] bromide". *Acta Crystallographica Section C* 63 (10) 2007, pp. m451–m453. DOI: 10.1107/S0108270107039856.

[127] Christer B. Aakeroy, Alicia M. Beatty, and Destin S. Leinen. "Supramolecular assembly of low-dimensional silver(i) architectures: testing the reliability of the self-complementary oximeoxime hydrogen-bond interaction". *CrystEngComm* 4 (55) 2002, pp. 310–314. DOI: 10.1039/B202303M.

[128] Megha S. Deshpande et al. "Supramolecular Self-Assembled Ruthenium—Polypyridyl Framework Encapsulating Discrete Water Cluster". *Crystal Growth & Design* 6 (3) 2006, pp. 743–748. DOI: 10.1021/cg0505719.

[129] V. Balamurugan et al. "Designing neutral coordination networks using inorganic supramolecular synthons: Combination of coordination chemistry and C-HCl hydrogen bonding". *CrystEngComm* 6 (66) 2004, pp. 396–400. DOI: 10.1039/B406744B.

[130] Deyuan Kong, Abraham Clearfield, and Jerzy Zoń. "Crystal-Engineered Three-Dimensional Hydrogen-Bonding Networks Built with 1,3,5-Benzenetri(phosphonic acid) and Bipyridine Synthons". *Crystal Growth & Design* 5 (5) 2005, pp. 1767–1773. DOI: 10.1021/cg050033w.

[131] Alexander Briceño et al. "Hydrogen-bonded networks in *trans*-3-(3-pyridyl)acrylic acid and *rctt*-3,3′-(3,4-dicarboxycyclobutane-1,2-diyl)dipyridinium dichloride". *Acta Crystallographica Section C* 63 (8) 2007, pp. 441–444. DOI: 10.1107/S0108270107027497.

[132] Shuang-Quan Zang, Ping-Shing Cheng, and Thomas C. W. Mak. "SilverX-aryl (X = I and Br) interaction in a network assembly with a flexible polynuclear silver-ethynide supramolecular synthon". *CrystEngComm* 11 (6) 2009, pp. 1061–1067. DOI: 10.1039/B900308H.

[133] Siegfried M. J. Wang, Liang Zhao, and Thomas C. W. Mak. "Coordination networks constructed with the multinuclear silver-ethynide supramolecular synthon 4-nitrophenyl-C≡C⊃Ag_n (n = 3, 4, 5)". *Dalton Trans.* 39 (8) 2010, pp. 2108–2121. DOI: 10.1039/B917789B.

[134] Daniel E. Lynch et al. "A New Supramolecular Synthon Using N-Methylaniline. The Crystal Structure of the 1 : 1 Adduct of N-Methylaniline with 5-Nitrofuran-2-carboxylic Acid". *Australian Journal of Chemistry* 51 (9) 1998, pp. 867–870. URL: http://www.publish.csiro.au/paper/C98072.

[135] Muhammad Irfan Ashiq et al. "Dimeric supramolecular motifs of two carboxylate–guanidinium compounds". *Acta Crystallographica Section C* 66 (9) 2010, pp. 455–458. DOI: 10.1107/S0108270110029252.

[136] Shiki Yagai et al. "Synthesis and noncovalent polymerization of self-complementary hydrogen-bonding supramolecular synthons: N,N′-disubstituted 4,6-diamino-pyri -midin-2(1H)-ones". *Chem. Commun.* (9) 2004, pp. 1114–1115. DOI: 10.1039/B401132E.

[137] Patrick Gamez and Jan Reedijk. "1,3,5-Triazine-Based Synthons in Supramolecular Chemistry". *European Journal of Inorganic Chemistry* 2006 (1) 2006, pp. 29–42. DOI: 10.1002/ejic.200500672.

[138] JolantaNatalia Latosińska et al. "Supramolecular synthon pattern in solid clioquinol and cloxiquine (APIs of antibacterial, antifungal, antiaging and antituberculosis drugs) studied by 35Cl NQR, 1H-17O and 1H-14N NQDR and DFT/Q-TAIM". English. *Journal of Molecular Modeling* 17 (7) 2011, pp. 1781–1800. DOI: 10.1007/s00894-010-0876-4.

[139] Borys Ośmiałowski et al. "2-Acylamino- and 2,4-Bis(acylamino)pyrimidines as Supra -molecular Synthons Analyzed by Multiple Noncovalent Interactions. DFT, X-ray Diffraction, and NMR Spectral Studies". *The Journal of Organic Chemistry* 77 (21) 2012, pp. 9609–9619. DOI: 10.1021/j301643z.

[140] Ernest Meštrović and Branko Kaitner. "A supramolecular structure of bis(1,3-diphenyl-1,3-propanedionato-O,O') (1,10-phenanthroline-N,N')cobalt(II) based on C–H...O, C–H...π and π...π interactions". English. *Journal of Chemical Crystallography* 36 (9) 2006, pp. 599–603. DOI: 10.1007/s10870-006-9105-5.

[141] Feng-Ling CUI et al. "Investigation of the Interaction between Adenosine and Human Serum Albumin by Fluorescent Spectroscopy and Molecular Modeling". *Chinese Journal of Chemistry* 26 (4) 2008, pp. 661–665. DOI: 10.1002/cjoc.200890125.

[142] Unchulee Suksangpanya et al. "Generation and structural characterisation of a difluorodimethoxyborate-mediated hydrogen-bonded supramolecular synthon". *CrystEngComm* 4 (108) 2002, pp. 638–643. DOI: 10.1039/B210393C.

[143] Hiroyuki Tanaka et al. "Diaminotriazine-substituted nitronyl nitroxide: a novel building block for organic magnets having multiple hydrogen bonding substituents as structure-determining supramolecular synthons". *CrystEngComm* 12 (2) 2010, pp. 526–531. DOI: 10.1039/B909757K.

[144] Jack D. Dunitz. "A Supramolecular Three-Dimensional Hydrogen-Bonded Network with Potential Application in Crystal Engineering Paradigms". *Chemistry – A European Journal* 4 (4) 1998, pp. 745–746. DOI: 10.1002/(SICI)1521-3765(19980416)4:4<745::AID-CHEM745>3.0.CO;2-4.

[145] Darshak R. Trivedi and Parthasarathi Dastidar. "Cation-Induced Supramolecular Isomerism in the Hydrogen-Bonded Network of Secondary Ammonium Monocar-

boxylate Salts: A New Class of Organo Gelator and Their Structures". *Crystal Growth & Design* 6 (9) 2006, pp. 2114–2121. DOI: 10.1021/cg060325c.

[146] David R. Turner, Sze Nee Pek, and Stuart R. Batten. "Heterotapes: A Persistent, Dual-Synthon Hydrogen-Bonding Motif". *Chemistry – An Asian Journal* 2 (12) 2007, pp. 1534–1539. DOI: 10.1002/asia.200700208.

[147] Hua-Kui Wu, Yan-Qing Ji, and Yu Liu. "Crystal Engineering of Supramolecular Interaction Based on Different Molecular Synthons". *Synthesis and Reactivity in Inorganic, Metal-Organic, and Nano-Metal Chemistry* 41 (10) 2011, pp. 1293–1298. DOI: 10.1080/15533174.2011.594843.

[148] Qing-ling Ni et al. "4,4′-Dithiodiphenol: a chain of rings generated by two pairs of O—H$\cdots$O hydrogen bonds". *Acta Crystallographica Section E* 60 (12) 2004, pp. 2318–2320. DOI: 10.1107/S1600536804028636.

[149] Unchulee Suksangpanya et al. "Complementarity of halide-mediated hydrogen-bonding and alkyl substitution in the construction of two-dimensional rhombic (4,4) grids using bis(N-alkylamidino-O-alkylurea)copper(ii) halides". *CrystEngComm* 5 (3) 2003, pp. 10–22. DOI: 10.1039/B210185H.

[150] Unchulee Suksangpanya et al. "Hydrogen-bonded supramolecular synthons in complexes of copper(ii) halides with polymethylene-linked bis(amidino-O-alkylurea) ligands". *CrystEngComm* 4 (92) 2002, pp. 552–563. DOI: 10.1039/B208617D.

[151] Tetsuharu Yuge et al. "Guest-induced topological polymorphism of pseudo-cubic hydrogen bond networks-robust and adaptable supramolecular synthon". *CrystEngComm* 10 (3) 2008, pp. 263–266. DOI: 10.1039/B715926A.

[152] Christer B. Aakeröy et al. "Supramolecular Synthesis Based on a Combination of Hydrogen and Halogen Bonds". *Crystal Growth & Design* 9 (1) 2009, pp. 432–441. DOI: 10.1021/cg8006712.

[153] László Fábián. "Cambridge Structural Database Analysis of Molecular Complementarity in Cocrystals". *Crystal Growth & Design* 9 (3) 2009, pp. 1436–1443. DOI: 10.1021/cg800861m.

[154] Jose Fayos. "Molecular Crystal Prediction Approach by Molecular Similarity: Data Mining on Molecular Aggregation Predictors and Crystal Descriptors". *Crystal Growth & Design* 9 (7) 2009, pp. 3142–3153. DOI: 10.1021/cg801122m.

[155] Peter Atkins and Julio de Paula. *Atkins' Physical Chemistry*. 8th Ed. Oxford University Press, 2006.

[156] Peter Atkins et al. *Inorganic Chemistry*. Oxford University Press, 2014.

[157] Makoto Fujita et al. *Metal-Organic Frameworks: Design and Application*. Ed. by Leonard R. MacGillivray. 1st ed. Wiley and Sons, 2010.

[158] Edurne S. Larrea et al. *Coordination Polymers and Metal organic Frameworks*. Ed. by Oscar L. Ortiz and Luis d. Ramírez. Chemical Engineering Methods and Technology. Nova Science Publishers, 2012.

[159] Roland E. Dolle. "Comprehensive survey of chemical libraries yielding enzyme inhibitors, receptor agonists and antagonists, and other biologically active agents: 1992 through 1997". English. *Molecular Diversity* 3 (4) 1997, pp. 199–233. DOI: 10.1023/A:1009699413828.

[160] Roland E Dolle and Kingsley H Nelson. "Comprehensive survey of combinatorial library synthesis: 1998". *Journal of combinatorial chemistry* 1 (4) 1999, pp. 235–282.

[161] Roland E Dolle. "Comprehensive survey of combinatorial library synthesis: 1999". *Journal of combinatorial chemistry* 2 (5) 2000, pp. 383–433.

[162] Roland E Dolle. "Comprehensive survey of combinatorial library synthesis: 2000". *Journal of combinatorial chemistry* 3 (6) 2001, pp. 477–517.

[163] Roland E Dolle. "Comprehensive survey of combinatorial library synthesis: 2001". *Journal of combinatorial chemistry* 4 (5) 2002, pp. 369–418.

[164] Roland E Dolle. "Comprehensive survey of combinatorial library synthesis: 2002". *Journal of combinatorial chemistry* 5 (6) 2003, pp. 693–753.

[165] Roland E Dolle. "Comprehensive survey of combinatorial library synthesis: 2003". *Journal of combinatorial chemistry* 6 (5) 2004, pp. 623–679.

[166] Roland E Dolle. "Comprehensive survey of combinatorial library synthesis: 2004". *Journal of combinatorial chemistry* 7 (6) 2005, pp. 739–798.

[167] Roland E Dolle et al. "Comprehensive survey of combinatorial library synthesis: 2005". *Journal of combinatorial chemistry* 8 (5) 2006, pp. 597–635.

[168] Roland E Dolle et al. "Comprehensive survey of chemical libraries for drug discovery and chemical biology: 2006". *Journal of combinatorial chemistry* 9 (6) 2007, pp. 855–902.

[169] Roland E Dolle et al. "Comprehensive survey of chemical libraries for drug discovery and chemical biology: 2007". *Journal of combinatorial chemistry* 10 (6) 2008, pp. 753–802.

[170] Roland E Dolle et al. "Comprehensive survey of chemical libraries for drug discovery and chemical biology: 2008". *Journal of combinatorial chemistry* 11 (5) 2009, pp. 739–790.

[171] Roland E Dolle et al. "Comprehensive survey of chemical libraries for drug discovery and chemical biology: 2009". *Journal of combinatorial chemistry* 12 (6) 2010, pp. 765–806.

[172] Peter Willett. "Chemoinformatics–similarity and diversity in chemical libraries". *Current opinion in biotechnology* 11 (1) 2000, pp. 85–88.

[173] Jens Sadowski. "Optimization of chemical libraries by neural networks". *Current Opinion in Chemical Biology* 4 (3) 2000, pp. 280–282. DOI: 10.1016/S1367-5931(00)00089-2.

[174] Konstantin V. Balakin et al. "Rational Design Approaches to Chemical Libraries for Hit Identification". *Current Drug Discovery Technologies* 3 (1) 2006, pp. 49–65. URL: http://www.ingentaconnect.com/content/ben/cddt/2006/00000003/00000001/art00002.

[175] Hugo O Villar and Mark R Hansen. "Design of chemical libraries for screening". *Expert opinion on drug discovery* 4 (12) 2009, pp. 1215–1220.

[176] Ian J. Bruno et al. "New software for searching the Cambridge Structural Database and visualizing crystal structures". *Acta Crystallographica Section B* 58 (3 Part 1) 2002, pp. 389–397. DOI: 10.1107/S0108768102003324.

[177] Frank H. Allen. "The Cambridge Structural Database: a quarter of a million crystal structures and rising". *Acta Crystallographica Section B* 58 (3 Part 1) 2002, pp. 380–388. DOI: 10.1107/S0108768102003890.

[178] Frances C. Bernstein et al. "The protein data bank: A computer-based archival file for macromolecular structures". *Archives of Biochemistry and Biophysics* 185 (2) 1978, pp. 584–591. DOI: 10.1016/0003-9861(78)90204-7.

[179] Saulius Gražulis et al. "Crystallography Open Database – an open-access collection of crystal structures". *Journal of Applied Crystallography* 42 (4) 2009, pp. 726–729. DOI: 10.1107/S0021889809016690.

[180] Joel Bernstein. *Polymorphism in Molecular Crystals*. IUCr Monographs on Crystallography. Oxford University Press, 2002.

[181] Ashwini Nangia. "Database research in crystal engineering". *CrystEngComm* 4 (17) 2002, pp. 93–101. DOI: 10.1039/B201206E.

[182] Delia A. Haynes et al. "Supramolecular synthon competition in organic sulfonates: A CSD survey". *CrystEngComm* 6 (95) 2004, pp. 584–588. DOI: 10.1039/B413797C.

[183] Jack D. Dunitz, Giuseppe Filippini, and Angelo Gavezzotti. "A Statistical Study of Density and Packing Variations among Crystalline Isomers". *Tetrahedron* 56 (36) 2000, pp. 6595–6601. DOI: 10.1016/S0040-4020(00)00460-9.

[184] Detlef W.M. Hofmann and Joannis Apostolakis. "Crystal structure prediction by data mining". *Journal of Molecular Structure* 647 (1–3) 2003. Polymorphism, prediction, transformations and activity in organic crystal chemistry., pp. 17–39. DOI: 10.1016/S0022-2860(02)00519-7.

[185] Rudolf Hoppe. "Neue Untersuchungen an Komplexen Fluoriden". *Journal of Inorganic and Nuclear Chemistry* 8 1958, pp. 437–440.

[186] H.B. Bürgi and J. D. Dunitz. "Crystal and Molecular Structures of Benzylideneaniline, Benzylideneaniline-p-carboxylic acid and p-methylbenzylidene-p-nitroaniline". *Helvetica Acta* 7 (53) 1970, pp. 1747–1764.

[187] Thomas Gelbrich and Michael B. Hursthouse. "A versatile procedure for the identification, description and quantification of structural similarity in molecular crystal". *Crystal Engineering Communications* 53 (7) 2005, pp. 324–336.

[188] Linus Pauling. *The Nature of the Chemical Bond and the Structure of Molecules and Crystals*. 3rd ed. Oxford University Press, 1960.

[189] "Ueber die Beziehungen der Eigenschaften zu den Atomgewichten der Elemente". *Zeitschrift für chemie* 12 1869, pp. 405–406.

[190] A.R. Leach and V.J. Gillet. *An Introduction to Cheminformatics*. Revised Ed. Springer, 2007.

[191] Roberto Todeschini. *Molecular descriptors for chemoinformatics*. 2nd rev. and enl. ed. Vol. 1. Methods and principles in medicinal chemistry. Wiley and Sons, 2009.

[192] H. L. Monaco C. Giacovazzo et al. *Fundamentals of Crystallography*. IUCr Texts On Crystallography. Oxford Science Publications, 1992.

[193] Unknown. *Pulveraufnahmen nach Debye-Scherrer (oben) und Guinier (unten) für ein Kaliumthioplatinat*. Public Domain Image.

[194] Jane Shelby Richardson. *Photographic X-ray diffraction image from a bovine Cu,Zn superoxide dismutase crystal, taken on a precession camera around 1971. Space group C2, resolution out to 2Å*. Creative Commons Attribution 3.0 Unported.

[195] *CrystalClear 3.1*. Software Released by Rigaku. 2013.

[196] *CrystalClear 2.1*. Software Released by Rigaku. 2012.

[197] H. Arnold et al. *Space-group symmetry*. Ed. by Th. Hahn and M. I. Aroyo. 5th ed. Vol. A. International Tables for Crystallography. Wiley and Sons, 2006.

[198] *The Reflection of X-rays by Crystals*. Royal Society. Royal Society, 1913.

[199] Jenny Pickworth Glusker and Kenneth N. Trueblood. *Crystal Structure Analysis A Primer*. 2nd ed. Oxford University Press, 1985.

[200] Alexander J. Blake et al. *Crystal Structure Analysis*. Ed. by William Clegg. 2nd ed. IUCr Texts On Crystallography. Oxford Science Publications, 2009.

[201] A. L. Patterson. "A Direct Method for the Determination of the Componenets of Interatomic Distances in Crystals". *Z. Kristallogr.* 90 1935, pp. 517–542.

[202] Herbert A. Hauptman. "The phase problem of x-ray crystallography". *Reports on Progress in Physics* 1991, pp. 1427–1454.

[203] Lukáš Palatinus. "The charge-flipping algorithm in crystallography". *Acta Crystallographica Section B* (69) 2013, pp. 1–16.

[204] Thomas R. Schneider and George M. Sheldrick. "Substructure solution with *SHELXD*". *Acta Crystallographica Section D* 58 (10 Part 2) 2002, pp. 1772–1779. DOI: 10.1107/S0907444902011678.

[205] George M. Sheldrick. "A short history of *SHELX*". *Acta Crystallographica Section A* 64 (1) 2008, pp. 112–122. DOI: 10.1107/S0108767307043930.

[206] Oleg V Dolomanov et al. "OLEX2: a complete structure solution, refinement and analysis program". *Journal of Applied Crystallography* 42 (2) 2009, pp. 339–341.

[207] Lukáš Palatinus and Gervais Chapuis. "*SUPERFLIP* – a computer program for the solution of crystal structures by charge flipping in arbitrary dimensions". *Journal of Applied Crystallography* 40 (4) 2007, pp. 786–790. DOI: 10.1107/S0021889807029238.

[208] Charles M. Weeks et al. "Ab initio phasing by dual-space direct methods". In: *Proceedings of the 18th European Crystallographic Association Meeting: Advances in Structure Analysis*. Ed. by R. Kužel and J. šek. Obtained from. Czeck & Slovak Crystallographic Association. 2001, pp. 37–64.

[209] Delores M. Etter. *Structured FORTRAN 77 for engineers and scientists*. 4th ed. Benjamin/Cummings, 1993.

[210] S. R. Hall, F. H. Allen, and I. D. Brown. "The crystallographic information file (CIF): a new standard archive file for crystallography". *Acta Crystallographica Section A* 47 (6) 1991, pp. 655–685. DOI: 10.1107/S010876739101067X.

[211] F. H. Allen et al. *Definition and exchange of crystallographic data*. Ed. by S. R. Hall and B. McMahon. First online edition. Vol. G. Wiley and Sons, 2006.

[212] Michael Dittmar, Anne-Sylvie Nicollerat, and Abdelhak Djouadi. "Z' studies at the LHC: an update". *Physics Letters B* 583 (1–2) 2004, pp. 111–120. DOI: 10.1016/j.physletb.2003.09.103.

[213] *Z'*. URL: https://community.dur.ac.uk/zprime/.

[214] Bouke P. van Eijck and Jan Kroon. "Structure predictions allowing more than one molecule in the asymmetric unit". *Acta Crystallographica Section B* 56 (3) 2000, pp. 535–542. DOI: 10.1107/S0108768100000276.

[215] *CSD Entries: Summary Statistics*. 2014. URL: http://www.ccdc.cam.ac.uk/SupportandResources/Resources/pages/Resources.aspx?mc=-1%5C&sc=-1%5C&rt=9%5C&p=-1.

[216] *Crystallography Open Database*. 2014. URL: http://www.crystallography.net.

[217] Unknown. *CSD 5.35 Release & Installation Notes - v5.35*. Online. 2014. URL: http://www.ccdc.cam.ac.uk/Lists/DocumentationList/release%5C_install%5C_notes535.pdf.

[218] *Crystallographic Open Database Features*. 2014. URL: http://www.crystallography.net/search.html.

[219] A. D. McNaught et al. *IUPAC Compendium of Chemical Terminology (the "Gold Book")*. Online Corrected Version (2.3.3). Blackwell Scientific Publications, 2014. URL: http://goldbook.iupac.org/V06597.html.

[220] Angelo Gavezzotti. "Calculation of lattice energies of organic crystals: the PIXEL integration method in comparison with more traditional methods". *Z. Kristallogr.* (220) 2005, pp. 499–510.

[221] Angelo Gavezzotti. "Crystal formation and stability: Physical principles and molecular simulation". *Crystal Research and Technology* 48 (10) 2013, pp. 793–810. DOI: 10.1002/crat.201200706.

[222] Fritz London. "Properties and applications of molecular forces". *Zeitschrift für physikalische Chemie B* 11 1930, pp. 222–251.

[223] Fritz London. "Zur Theorie und Systematik der Molekularkräfte". *Zeitscrift Für Physik* 63 (3-4) 1930, pp. 245–279.

[224] Anthony Stone. *The Theory of Intermolecular Forces*. 2nd ed. Oxford Scholarship Online, 2013.

[225] H. Ratajczak and W. J. Orville-Thomas. *Molecular Interactions*. Ed. by Mavis Redshaw. Vol. 3. Wiley and Sons, 1982.

[226] Carlo Gatti and Piero Macchi, eds. *Modern Charge-Density Analysis*. Springer, 2012.

[227] Alexander J. Pertsin and A.I. Kitaigorodskii. *The atom-atom potential method : applications to organic molecular solids*. Vol. 43. Springer series in Chemical Physics. Springer-Verlag, 1987.

[228] Julian Vrbancich and Geoffrey L. D. Ritchie. "Quadrupole moments of benzene, hexafluorobenzene and other non-dipolar aromatic molecules". *J. Chem. Soc., Faraday Trans. 2* 76 1980, pp. 648–659. DOI: 10.1039/F29807600648.

[229] Gastone Gill and Paola Gilli. *The Nature of the Hydrogen Bond*. Vol. 23. IUCr Monographs on Crystallography. Oxford Science Publications, 2009.

[230] Clare F. Macrae et al. "*Mercury CSD 2.0* – new features for the visualization and investigation of crystal structures". *Journal of Applied Crystallography* 41 (2) 2008, pp. 466–470. DOI: 10.1107/S0021889807067908.

[231] Jack D. Dunitz and Robin Taylor. "Organic Fluorine Hardly Ever Accepts Hydrogen Bonds". *Chem. Eur. J.* 3 (1) 1997, pp. 89–98.

[232] Gustav Mie. "Zur kinetischen Theorie der einatomigen Körper". *Annalen der physik* 316 (8) 1903, pp. 657–697.

[233] A.I. Kitaigorodskii. *Organic Crystallography*. Authorised Translation. Heywood, 1961.

[234] Elna Pidcock and W. D. Sam Motherwell. "A Novel Description of the Crystal Packing of Molecules". *Crystal Growth & Design* 4 (3) 2004, pp. 611–610.

[235] Elna Pidcock. "Spatial arrangement of molecules in homomolecular Z' = 2 structures". *Acta Crystallographica Section B* 62 2006, pp. 268–279.

[236] Elna Pidcock and W. D. Sam Motherwell. "Parameterization of the close packing of molecules in the unit cell". *Acta Crystallographica Section B* 60 2004, pp. 725–733.

[237] Elna Pidcock and W. D. Sam Motherwell. "Distribution of molecular centres in unit cells with respect to packing patterns". *Acta Crystallographica Section B* 60 2004, pp. 539–546.

[238] Jonathan Clayden et al. *Organic Chemistry*. Oxford University Press, 2001.

[239] E. J. Corey. "General Methods For The Construction Of Complex Molecules". *Pure Appl. Chem.* 14 (1) 1967, pp. 19–38.

[240] G. R. Desiraju. "Supramolecular Synthons in Crystal Engineering—A New Organic Synthesis." *Angewandte Chemie International Edition in English* 34 (21) 1995, pp. 2311–2327.

[241] Thai Thanh Thu Bui et al. "The Nature of Halogen···Halogen Interactions: A Model Derived from Experimental Charge-Density Analysis". *Angewandte Chemie International Edition* 48 (21) 2009, pp. 3838–3841. DOI: 10.1002/anie.200805739.

[242] A. van de Walle. "A complete representation of structure–property relationships in crystals". *Nature Materials* 7 2008, pp. 455–458.

[243] James D. Wuest. "Engineering crystals by the strategy of molecular tectonics". *Chem. Commun.* (47) 2005, pp. 5830–5837. DOI: 10.1039/B512641J.

[244] Jonathan W. Steed. *Encyclopedia of Supramolecular Chemistry*. Ed. by Jerry L. Atwood. 4th ed. Vol. 2. CRC Press LLC, 2004.

[245] Elisa Nauha et al. "Comparison of the polymorphs and solvates of two analogous fungicides-a case study of the applicability of a supramolecular synthon approach in crystal engineering". *CrystEngComm* 13 (15) 2011, pp. 4956–4964. DOI: 10.1039/C1CE05077J.

[246] Sihui Long et al. "From Competition to Commensuration by Two Major Hydrogen-Bonding Motifs". *Crystal Growth & Design* 14 (1) 2014, pp. 27–31. DOI: 10.1021/cg401532j.

[247] Chelsea R. Martinez and Brent L. Iverson. "Rethinking the term "pi-stacking"". *Chem. Sci.* 3 (7) 2012, pp. 2191–2201. DOI: 10.1039/C2SC20045G.

[248] Klaus Merz and Vera Vasylyeva. "Development and boundaries in the field of supramolecular synthons". *CrystEngComm* 12 (12) 2010, pp. 3989–4002. DOI: 10.1039/C0CE00237B.

[249] William H. Ojala et al. "Molecular symmetry and fluorine-containing supramolecular synthons as structure-differentiating agents in some "bridge-flipped" isomeric bis-benzylideneanilines". *CrystEngComm* 16 (31) 2014, pp. 7226–7235. DOI: 10.1039/C3CE42540A.

[250] Harrison Omorodion et al. "Further Evidence on the Importance of Fluorous–Fluorous Interactions in Supramolecular Chemistry: A Combined Structural and Computational Study". *Crystal Growth & Design* 16 2014, pp. 7226–7235. DOI: 10.1021/acs.cgd.5b00254.

[251] Steve Scheiner. "The Pnicogen Bond: Its Relation to Hydrogen, Halogen, and Other Noncovalent Bonds". *Accounts of Chemical Research* 46 (2) 2013, pp. 280–288. DOI: 10.1021/ar3001316.

[252] K. Eskandari and N. Mahmoodabadi. "Pnicogen Bonds: A Theoretical Study Based on the Laplacian of Electron Density". *The Journal of Physical Chemistry A* 117 (48) 2013, pp. 13018–13024. DOI: 10.1021/jp4098974.

[253] Weizhou Wang, Baoming Ji, and Yu Zhang. "Chalcogen Bond: A Sister Noncovalent Bond to Halogen Bond". *The Journal of Physical Chemistry A* 113 (28) 2009, pp. 8132–8135. DOI: 10.1021/jp904128b.

[254] Rahul Shukla and Deepak Chopra. "Crystallographic and computational investigation of intermolecular interactions involving organic fluorine with relevance to the hybridization of the carbon atom". *CrystEngComm* 17 (19) 2015, pp. 3596–3609. DOI: 10.1039/C4CE02391A.

[255] Per Restorp et al. "A synthetic receptor for hydrogen-bonding to fluorines of trifluoroborates". *Chem. Commun.* (38) 2009, pp. 5692–5694. DOI: 10.1039/B914171E.

[256] Gurpreet Kaur and Angshuman Roy Choudhury. "A comprehensive understanding of the synthons involving C-HF-C hydrogen bond(s) from structural and computational analyses". *CrystEngComm* 17 (15) 2015, pp. 2949–2963. DOI: 10.1039/C5CE00215J.

[257] Piyush Panini and Deepak Chopra. "Role of intermolecular interactions involving organic fluorine in trifluoromethylated benzanilides". *CrystEngComm* 14 (6) 2012, pp. 1972–1989. DOI: 10.1039/C2CE06254B.

[258] Mark Lutz. *Learning Python*. 3rd ed. O'Reilly Media, 2007.

[259] Anon. *The Python Language Reference*. Online. 2014. URL: https://docs.python.org/3.3/reference/.

[260] Ivor Horton. *Ivor Horton's Beginning Visual C++ 2010*. Wiley and Sons, 2010.

[261] Bjarne Stroustrup. *The C++ Programming Languages*. 1st ed. Addison-Wesley, 1985.

[262] *Information Technology — Programming Languages — C++*. The British Standards Institute. 2011.

[263] J.W. Backus et al. *Fortran Automatic Coding System For The IBM 704*. 1st ed. IBM. 1956.

[264] Stephen J. Chapman. *Fortran 95/2003 for Scientists and Engineers*. 3rd ed. McGraw-Hill, 2008.

[265] *What is php?* URL: `php.net/manual/en/intro-whatis.php`.

[266] *History of PHP*. Online. URL: `http://php.net/manual/en/history.php.php`.

[267] *Internal (built in) Functions*. Online. URL: `http://php.net/manual/en/functions.internal.php`.

[268] Dylan Jayatilaka. URL: `http://ra.bcs.uwa.edu.au/Tonto/wiki/index.php/Welcome_to_Tonto!`.

[269] *R Language Definition*. 3.1.1. R Core Team. 2014.

[270] *Package 'parallel'*. R Core Team. 2013.

[271] Norman Matloff. *The Art of R Programming: A Tour of Statistical Software Design*. No Starch Press, 2011.

[272] *Standard generalized markup language (SGML) for text and office systems*. Revised Ed. The British Standards Institute. 1991.

[273] *CIF Core Definition*. 2012. URL: `ftp://ftp.iucr.org/pub/cif_core.dic`.

[274] W.L.Bragg. "The Structure of Some Crystals as Indicated by Their Diffraction of X-rays". *Proceedings of the Royal Society* 89 (610) 1913, pp. 248–277.

[275] *Tripos Mol2 File Format*. Tripos. 2005.

[276] *CTFile Formats*. MDL and Elsevier. 2005.

[277] Peter Murray-Rust and Henry S. Rzepa. "Chemical Markup, XML, and the World Wide Web. 4. CML Schema". *Journal of Chemical Information and Computer Sciences* 43 (3) 2003, pp. 757–772. DOI: `10.1021/ci0256541`.

[278] *Chemical Markup Language*. 2014. URL: `http://xml-cml.org/`.

[279] Nico Adams et al. "Chemical Markup, XML and the World-Wide Web. 8. Polymer Markup Language". *Journal of Chemical Information and Modeling* 48 (11) 2008, pp. 2118–2128. DOI: 10.1021/ci8002123.

[280] Peter Ertl, Bernhard Rohde, and Paul Selzer. "Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties". *J. Med. Chem.* 43 2000, pp. 3714–3717.

[281] Derwent Publications Ltd. *World Drug Index Database WDI97*. Distributed by Daylight Chemical Information Systems, Inc.

[282] David E. Clark. "Rapid Calculation of Polar Molecular Surface Area and Its Application to the Prediction of Transport Phenomena. 1. Prediction of Intestinal Absorption". *Journal of Pharmaceutical Sciences* 88 (8) 1999, pp. 807–814.

[283] Lawrence R. Dodd and Doros N. Theodorou. "Analytical treatment of the volume and surface area of molecules formed by an arbitrary collection of unequal spheres intersected by planes". *Molecular Physics* 72 (6) 1991, pp. 1313–1345. DOI: 10.1080/00268979100100941.

[284] L. N. Kuleshova and P. M. Zorky. "Graphical Enumeration of Hydrogen-Bonded Structures". *Acta Crystallographica Section B* 36 1980, pp. 2113–2115.

[285] Margaret C. Etter and John C. MacDonald. "Graph-Set Analysis of Hydrogen-Bond Patterns in Organic Crystals". *Acta Crystallographica Section B* 46 1990, pp. 256–262.

[286] CCDC. *R-Pluto*. Software.

[287] J. Grell, J. Bernstein, and G. Tinhofer. "Graph-set analysis of hydrogen-bond patterns: some mathematical concepts". *Acta Crystallographica Section B* 55 1999, pp. 1030–1043.

[288] Zhenjie Zhang and Michael J. Zaworotko. "Template-directed synthesis of metal-organic materials". *Chem. Soc. Rev.* 43 (16) 2014, pp. 5444–5455. DOI: 10.1039/C4CS00075G.

[289] *CSD Space Group Statistics - Space Group Frequency Ordering*. Online. 2014. URL: http://www.ccdc.cam.ac.uk/Lists/ResourceFileList/2014%5C_stats%5C_sgrankorder.pdf.

[290] Gerard E. Dallal. *The Little Handbook of Statistical Practice*. Online. 2012. URL: http://www.jerrydallal.com/LHSP/LHSP.HTM.

[291] Michael J. Wichura. *The Coordinate-Free Approach to Linear Models*. ebook. Cambridge Series in Statistical and Probabalistic Mathematics. Cambridge University Press, 2007.

[292] B. S. Everett and A. Skrondal. *The Cambridge Dictionary of Statistics*. 4th ed. Cambridge University Press, 2010.

[293] Richard Anderson-Sprecher. "Model Comparisons and R^2". *The American Statistician* 48 (2) 1994, pp. 113–117. DOI: 10.1080/00031305.1994.10476036.

[294] Randall Munroe. *Correlation*. Web: http://www.xkcd.com/552/. URL: http://www.xkcd.com/552/.

[295] Mark Hall et al. "The WEKA Data Mining Software: An Update". *SIGKDD Explorations* 11.1 2009.

[296] Leo Breiman et al. *Classification and Regression Trees*. CRC Press LLC, 1984.

[297] SAS Institute Inc. *JMP Version 11.0*. Software. 1989-2007.

[298] *Email Exchange with SAS Institute*. Private Communication. 2015.

[299] Derek Robinson Terry Threlfall. *Understanding Polymorphism & Crystallisation Issues in the Pharmaceutical Industry*. training course presented at the Radisson SAS Hotel, Dublin, Ireland. 2009.

[300] Jerome G. P. Wicker and Richard I. Cooper. "Will it crystallise? Predicting crystallinity of molecular materials". *CrystEngComm* 17 (9) 2015, pp. 1927–1934. DOI: 10.1039/C4CE01912A.

[301] A. Altomare et al. "*SIR*92 – a program for automatic solution of crystal structures by direct methods". *Journal of Applied Crystallography* 27 (3) 1994, p. 435. DOI: 10.1107/S002188989400021X.

[302] *Cobol*. Vol. 2. Codes That Changed The World. BBC Radio 4, 2015.

[303] *JChem*. Software Released by ChemAxon. URL: https://www.chemaxon.com/download/jchem-suite/.

[304] Greg Landrum. *RDKit*. Software published online. URL: http://www.rdkit.org/.

[305] *List of molecular descriptors calculated by Dragon*. Online PDF. 2012. URL: http://www.talete.mi.it/products/dragon_molecular_descriptor_list.pdf.

[306] A. M. Glazer. "*VIBRATE!* A program to compute irreducible representations for atomic vibrations in crystals". *Journal of Applied Crystallography* 42 (6) 2009, pp. 1194–1196. DOI: 10.1107/S0021889809040424.

[307] *Crystallographic Software Museum Hosted by the IUCr*. URL: http://www.iucr.org/resources/commissions/crystallographic-computing/software-museum.

[308] Anthony L. Spek. "Structure validation in chemical crystallography". *Acta Crystallographica Section D* 65 (2) 2009, pp. 148–155. DOI: 10.1107/S090744490804362X.

[309] L. S. Coles. "The Acetanilide Crystal Structures: Packing and Conformational Similarities". PhD thesis. Chemistry, 2011.

[310] Noel M O'Boyle et al. "*Open Babel: An open Chemical Toolbox*". *Journal of Cheminformatics* 3 (33) 2011.

[311] *Dragon 6 user's manual*. Online. 2010. URL: http://www.talete.mi.it/help/dragon%5C_help/.

[312] Srinivasulu Aitipamula et al. "Cocrystal Hydrate of an Antifungal Drug, Griseofulvin, with Promising Physicochemical Properties". *Crystal Growth & Design* 12 (12) 2012, pp. 5858–5863. DOI: 10.1021/cg3012124.

[313] Nizar Issa et al. "Screening for cocrystals of succinic acid and 4-aminobenzoic acid". *CrystEngComm* 14 (7) 2012, pp. 2454–2464. DOI: 10.1039/C2CE06325E.

[314] Shyam Karki et al. "New solid forms of artemisinin obtained through cocrystallisation". *CrystEngComm* 12 (12) 2010, pp. 4038–4041. DOI: 10.1039/C0CE00428F.

[315] Tudor Grecu et al. "Validation of a Computational Cocrystal Prediction Tool: Comparison of Virtual and Experimental Cocrystal Screening Results". *Crystal Growth & Design* 14 (1) 2014, pp. 165–171. DOI: 10.1021/cg401339v.

[316] Brenda C. Félix-Sonda et al. "Nitazoxanide Cocrystals in Combination with Succinic, Glutaric, and 2,5-Dihydroxybenzoic Acid". *Crystal Growth & Design* 14 (3) 2014, pp. 1086–1102. DOI: 10.1021/cg4015916.

[317] A. Bondi. "van der Waals Volumes and Radii". *The Journal of Physical Chemistry* 68 (3) 1964, pp. 441–451. DOI: 10.1021/j100785a001.

[318] S. van der Walt, S.C. Colbert, and G. Varoquaux. "The NumPy Array: A Structure for Efficient Numerical Computation". *Computing in Science Engineering* 13.2 2011, pp. 22–30. DOI: 10.1109/MCSE.2011.37.

[319] Simon Coles. "Are H-F interactions strong enough to drive crystal structure formation?" In: *HALCHEM V Conference*. Halchem.

[320] Megan L. Head et al. "The Extent and Consequences of P-Hacking in Science". *PLoS Biol* 13 (3) 2015, e1002106. DOI: 10.1371/journal.pbio.1002106.

[321] Lowell H. Hall, Brian. Mohney, and Lemont B. Kier. "The electrotopological state: structure information at the atomic level for molecular graphs". *Journal of Chemical Information and Computer Sciences* 31 (1) 1991, pp. 76–82. DOI: 10.1021/ci00001a012.

[322] Philip D.F. Adler et al. "Probing structural adaptability in templated vanadium selenites". *Polyhedron* 2015. Volume and Page not yet assigned. DOI: http://dx.doi.org/10.1016/j.poly.2015.11.038.

[323] Erin R. Johnson et al. "Revealing Noncovalent Interactions". *Journal of the American Chemical Society* 132 (18) 2010, pp. 6498–6506. DOI: 10.1021/ja100936w.

[324] *How should I reference published CSD data in my paper?* 2013. URL: http://www.ccdc.cam.ac.uk/SupportandResources/Support/pages/SupportSolution.aspx?supportsolutionid=287.

[325] Otto Ermer and Jörg Neudörfl. "Comparative Supramolecular Chemistry of Coro-nene and Hexahelicene: Helix Alignment in Crystalline Complexes with Trimesic Acid (=Benzene-1,3,5-tricarboxylic Acid) and π-Acceptor Compounds". *Helvetica Chimica Acta* 84 (6) 2001, pp. 1268–1313. DOI: 10.1002/1522-2675(20010613)84:6<1268::AID-HLCA1268>3.0.CO;2-Z.

[326] Y.Imai et al. "Crystal Structure of 9.10-Dipentafluorophenylanthracene Host System". *Letters in Organic Chemistry* 6 2009, p. 588.

[327] Katsuyuki Ogura et al. "Formation of gold-like metal-lustrous inclusion crystals from 1-phenyl-2,5-bis[5-(tricyanoethenyl)-2-thienyl]pyrrole host and an electron-donating aromatic guest". *Org. Biomol. Chem.* 1 (21) 2003, pp. 3845–3850. DOI: 10.1039/B302689B.

[328] P.S.Donnelly, B.W.Skelton, and A.H.White. "'Neutralmolekülcomplexe'—Struc-tural Characterization of Some Adducts of Urea and Thiourea with N,N'-Bidentate Aromatic Bases". *Australian Journal of Chemistry* 56 (12) 2003, pp. 1249–1253.

[329] J. Lefebvre et al. "Characterization of an orientational disorder in two charge-transfer complexes: anthracene–tetracyanobenzene (A–TCNB) and naphthalene–tetracyanobenzene (N–TCNB)". *Acta Crystallographica Section B* 45 (3) 1989, pp. 323–336. DOI: 10.1107/S0108768189001400.

[330] John J. Stezowski. "Phase transition effects: A crystallographic characterization of the temperature dependency of the crystal structure of the 1:1 charge transfer complex between anthracene and tetracyanobenzene in the temperature range 297

to 119 K". *The Journal of Chemical Physics* 73 (1) 1980, pp. 538–547. DOI: http://dx.doi.org/10.1063/1.439851.

[331] B. E. Robertson and J. J. Stezowski. "The crystal structure of the π-molecular complex of anthracene with pyromellitic dianhydride at –120°C". *Acta Crystallographica Section B* 34 (10) 1978, pp. 3005–3011. DOI: 10.1107/S0567740878009929.

[332] Dominik Cincic, Tomislav Friscic, and William Jones. "Experimental and database studies of three-centered halogen bonds with bifurcated acceptors present in molecular crystals, cocrystals and salts". *CrystEngComm* 13 (9) 2011, pp. 3224–3231. DOI: 10.1039/C0CE00699H.

[333] A. W. Hanson. "The crystal structure of the acepleiadylene, *s*-trinitrobenzene complex". *Acta Crystallographica* 21 (1) 1966, pp. 97–102. DOI: 10.1107/S0365110X66600238X.

[334] Pilar García-Orduña, Slimane Dahaoui, and Claude Lecomte. "Intermolecular interactions and charge transfer in the 2:1 tetrathiafulvalene bromanil complex, $(TTF)_2$-BA". *Acta Crystallographica Section B* 67 (3) 2011, pp. 244–249. DOI: 10.1107/S0108768111015801.

[335] Jonathan C. Collings et al. "Arene-perfluoroarene interactions in crystal engineering. Part 10. Crystal structures of 1:1 complexes of octafluoronaphthalene with biphenyl and biphenylene". *CrystEngComm* 6 (6) 2004, pp. 25–28. DOI: 10.1039/B316169B.

[336] T. Hasegawa et al. "Mixed-stack organic charge-transfer complexes with intercolumnar networks". *Phys. Rev. B* 62 (15) 2000, pp. 10059–10066. DOI: 10.1103/PhysRevB.62.10059.

[337] Ali Saad, Oliver Jeannin, and Marc Fourmigué. "Chiral, flexible binaphthol-subs -tituted tetrathiafulvalenes". *Tetrahedron* 67 (21) 2011, pp. 3820–3829. DOI: http://dx.doi.org/10.1016/j.tet.2011.03.103.

[338] Lakhemici Kaboub et al. "Structural study and electrical conductivity of salts based on functionalized TTF containing peripheral selenium atoms". *J. Mater. Chem.* 14 (3) 2004, pp. 351–356. DOI: 10.1039/B308514G.

[339] Laetitia J. Martin et al. "Safe and Reliable Synthesis of Diazoketones and Quinoxalines in a Continuous Flow Reactor". *Organic Letters* 13 (2) 2011, pp. 320–323. DOI: 10.1021/ol1027927.

[340] Jaroslaw Saczewski et al. "Synthesis of aza-aromatic hydroxylamine-O-sulfonates and their application to tandem nucleophilic addition–electrophilic 5-endo-trig cyclization". *Tetrahedron* 67 (20) 2011, pp. 3612–3618. DOI: http://dx.doi.org/10.1016/j.tet.2011.03.091.

[341] Daniel B. Werz, Rolf Gleiter, and Frank Rominger. "Cyclic Tetra- and Hexaynes Containing 1,4-Donor-Substituted Butadiyne Units: Synthesis and Supramolecular Organization". *The Journal of Organic Chemistry* 69 (9) 2004, pp. 2945–2952. DOI: 10.1021/jo035882+.

[342] Howard M. Colquhoun, David J. Williams, and Zhixue Zhu. "Macrocyclic Aromatic Ether-Imide-Sulfones: Versatile Supramolecular Receptors with Extreme Thermochemical and Oxidative Stability". *Journal of the American Chemical Society* 124 (45) 2002, pp. 13346–13347. DOI: 10.1021/ja027851m.

[343] N. B. Singh, Aradhana Pathak, and Roland Fröhlich. "Solid-State Complexation Reaction between p-Benzoquinone and p-Nitroaniline". *Australian Journal of Chemistry* 56 (4) 2003, pp. 329–333.

[344] T.M. Shchegoleva et al. "Charge-transfer complexes based on diphenoquinone and its derivatives". English. *Journal of Structural Chemistry* 22 (4) 1981, pp. 553–557. DOI: 10.1007/BF00784090.

[345] K. Yakushi et al. "The crystal structure of benzidene–*s*-trinitrobenzene 1:1 molecular complex benzene solvate". *Acta Crystallographica Section B* 31 (3) 1975, pp. 738–742. DOI: 10.1107/S0567740875003718.

[346] N. Karl, W. Ketterer, and J. J. Stezowski. "1:1 donor–acceptor complex between phenazine and pyromellitic dianhydride at approx. 120 K". *Acta Crystallographica Section B* 38 (11) 1982, pp. 2917–2919. DOI: 10.1107/S0567740882010334.

[347] R. Doherty et al. "The structure of the 1:1 molecular complex of pyrene and dicyanomethylenecroconate". *Acta Crystallographica Section B* 38 (3) 1982, pp. 859–863. DOI: 10.1107/S0567740882004245.

[348] C.-H. Chong, H. Yamochi, and G. Saito. "Property of intramolecular charge transfer compounds $D^{\delta+}$-π-$A^{\delta-}$ derived from {TCNQ} derivatives". *Synthetic Metals* 135-136 2003. Proceedings of the International Conference on Science and Technology of Synthetic Metals, pp. 603–604. DOI: http://dx.doi.org/10.1016/S0379-6779(02)00787-7.

[349] T. Dahl. "N^6,N^6-Dimethyladenine–tetraiodoethene (2/1)". *Acta Crystallographica Section C* 55 (9) 1999, pp. 1568–1570. DOI: 10.1107/S0108270199006472.

[350] Carsten Knapp et al. "Inclusion Complexes of the Bicyclic Aryl-Substituted Sulfur-Nitrogen Compounds Ar—CN_5S_3 with Fluorocarbon and Hydrocarbon Aromatics". *European Journal of Inorganic Chemistry* 2004 (12) 2004, pp. 2446–2451. DOI: 10.1002/ejic.200300951.

[351] Jeffrey Huw Williams, Jeremy Karl Cockcroft, and Andrew Nicholas Fitch. "Structure of the Lowest Temperature Phase of the Solid Benzene–Hexafluorobenzene Adduct". *Angewandte Chemie International Edition in English* 31 (12) 1992, pp. 1655–1657. DOI: 10.1002/anie.199216551.

[352] J.E. Mulvaney et al. "The crystal structure of 2,2'-bis-1,3-dithiole and dimethyl dicyanofumarate CT complex". English. *Journal of Crystallographic and Spectroscopic Research* 12 (4) 1982, pp. 331–342. DOI: 10.1007/BF01159049.

[353] Martin R. Bryce et al. "Preparation, properties, and X-ray crystal structure of a 1:1 complex of tetrathiafulvalene and p-dinitrobenzene". *Canadian Journal of Chemistry* 60 (16) 1982, pp. 2057–2061. DOI: 10.1139/v82-291.

[354] Veejendra K. Yadav et al. "7-Hydroxymethyl-6-methyl-1-oxa-4-thiaspiro[4.5]dec-6-ene. Exceptionally High Anti to Sulfur Diastereoselectivity under Conditions of Johnson Ortho Ester Claisen Rearrangement Favors Cieplak Mode of Diastereoselec -tion". *The Journal of Organic Chemistry* 64 (8) 1999, pp. 2928–2932. DOI: 10.1021/jo981634f.

[355] Thomas J. Emge et al. "Crystal Structures for the Electron Donor Dibenzotetrathiafulavalene, DBTTF, and Its Mixed-stack Charge-transfer Salts with the Electron Acceptors 7,7,8,8-tetracyano-p-quinodimethane, TCNQ, and 2,5-difluoro-7,7,8,8-tetracyano-p-quinodimethane, 2,5-$TCNQF_2$". *Molecular Crystals and Liquid Crystals* 87 (1-2) 1982, pp. 137–161. DOI: 10.1080/00268948208083778.

[356] Sergiy V. Rosokha et al. "Unusual structural effects of intermolecular π-bonding in the tetracyanopyrazine (ion-radical) dimer". *New J. Chem.* 33 (3) 2009, pp. 545–553. DOI: 10.1039/B812829D.

[357] Michaele J. Hardie and Colin L. Raston. "Crystalline hydrogen bonded complexes of o-carborane". *CrystEngComm* 3 (39) 2001, pp. 162–164. DOI: 10.1039/B107198J.

[358] D. E. Zacharias. "Structures of three electron donor–acceptor complexes of dibenz -[*a*,*h*]
-anthracene (DBA)". *Acta Crystallographica Section C* 49 (6) 1993, pp. 1082–1087. DOI: 10.1107/S0108270192013325.

[359] B. Shaanan and U. Shmueli. "Structure and packing arrangement of molecular compounds. IX. 7,7,8,8-Tetracyanoquinodimethane-7,8-benzoquinoline (1:1)". *Acta Crystallographica Section B* 36 (9) 1980, pp. 2076–2082. DOI: 10.1107/S0567740880007960.

[360] K. Kozawa and T. Uchida. "Structure of perylene–chloranil molecular complex (1:1), $C_{20}H_{12}.C_6Cl_4O_2$". *Acta Crystallographica Section C* 39 (9) 1983, pp. 1233–1235. DOI: 10.1107/S0108270183008045.

[361] Mikio Yasutake et al. "An Alternative Synthetic Route of [35](1,2,3,4,5)Cyclophane, and Structural Properties of Multibridged [3n]Cyclophanes and Their Charge-Transfer Complexes in the Solid State1". *Journal of the American Chemical Society* 124 (34) 2002, pp. 10136–10145. DOI: 10.1021/ja012363k.

[362] D. Zobel and G. Ruban. "The structures of some charge-transfer complexes containing TCNQ as acceptor and their electrical anisotropy". *Acta Crystallographica Section B* 39 (5) 1983, pp. 638–645. DOI: 10.1107/S0108768183003092.

[363] F. Bertinelli et al. "Optical Properties, Conductivity and Structure of the DTT-TCNQ CT Complex". *Molecular Crystals and Liquid Crystals* 109 (2-4) 1984, pp. 289–302. DOI: 10.1080/00268948408078714.

[364] Pilar García et al. "Crystallographic investigation of temperature-induced phase transition of the tetrathiafulvalene-*p-bromanil, TTF-BA charge transfer complex*". *Phys. Rev. B* 72 (10) 2005, p. 104115. DOI: 10.1103/PhysRevB.72.104115.

[365] Jürgen Steidel, Ralf Steudel, and Ali Kutoglu. "Röntgenstrukturanalysen von Cyclododekaschwefel (S_{12}) und Cyclododekaschwefel-1-Kohlendisulfid ($S_{12} \cdot CS_2$) [1]". *Zeitschrift für anorganische und allgemeine Chemie* 476 (5) 1981, pp. 171–178. DOI: 10.1002/zaac.19814760520.

[366] G. J. Corban et al. "Reactivity of di-iodine toward thiol: Desulfuration reaction of 5-nitro-2-mercapto-benzimidazole upon reaction with di-iodine". *Heteroatom Chemistry* 23 (5) 2012, pp. 498–511. DOI: 10.1002/hc.21042.

[367] Y. Yamaguchi and I. Ueda. "Structure of the 1:1 complex of 1,4-dithiintetracarboxylic *N*,*N*'-dimethyldiimide and acridine, $[C_{10}H_6N_2O_4S_2][C_{13}H_9N]$". *Acta Crystallographica Section C* 40 (1) 1984, pp. 113–115. DOI: 10.1107/S0108270184003218.

[368] F. H. Herbstein and G. M. Reisner. "Molecular compounds and complexes. XV. Pyrene–1,4,5,8-naphthalenetetrone (1:1), $C_{16}H_{10}.C_{10}H_4O_4$". *Acta Crystallographica Section C* 40 (1) 1984, pp. 202–204. DOI: 10.1107/S0108270184003590.

[369] L. I. Foss et al. "Banana". *Acta Crystallographica Section C* 40 (2) 1984, pp. 272–274. DOI: 10.1107/S0108270184003796.

[370] A. Elbasyouny et al. "Host-guest complexes of 18-crown-6 with neutral molecules possessing the structure element XH_2 (X = oxygen, nitrogen, or carbon)". *Journal of the American Chemical Society* 105 (22) 1983, pp. 6568–6577. DOI: 10.1021/ja00360a006.

[371] T. Koritsanszky et al. "Low-temperature x-ray and neutron diffraction studies on 18-crown·2 cyanamide including electron density determination". *Journal of the American Chemical Society* 113 (22) 1991, pp. 8388–8398. DOI: 10.1021/ja00022a028.

[372] Tobias Wedel and Joachim Podlech. "Alkylidene[1,3]dithiolane-1,3-dioxides as Potent Michael-Type Acceptors". *Synlett* 2006 (13) 2006, pp. 2043–2046. DOI: 10.1055/s-2006-948202.

[373] Luigi Pasimeni, Carlo Corvaja, and Dore Augusto Clemente. "Crystal Structure of Diphenylacetylene-Tetracyanobenzene 1:1 CT Complex and Spin Polarization of Triplet Exciton and of Trans-Stilbene Triplet Trap". *Molecular Crystals and Liquid Crystals* 104 (3-4) 1984, pp. 231–247. DOI: 10.1080/00268948408070426.

[374] Pierre Le Magueres, Sergey V. Lindeman, and Jay K. Kochi. "Direct relationship between intermolecular charge-transfer and charge-resonance complexes via structural changes in the arene donor with various π-acceptors". *J. Chem. Soc., Perkin Trans. 2* (7) 2001, pp. 1180–1185. DOI: 10.1039/B009543P.

[375] Stephan M Hubig, Sergey V Lindeman, and Jay K Kochi. "Charge-transfer bonding in metal–arene coordination". *Coordination Chemistry Reviews* 200-202 2000, pp. 831–873. DOI: http://dx.doi.org/10.1016/S0010-8545(00)00322-2.

[376] David Fox et al. "Site-selective supramolecular synthesis of halogen-bonded cocrystals incorporating the photoactive azo group". *CrystEngComm* 10 (9) 2008, pp. 1132–1136. DOI: 10.1039/B806911E.

[377] J. A. Bandy, M. R. Truter, and F. Vögtle. "The structure of the 1,4,7,10,13,16-hexaoxacyclooctadecane (18-crown-6) bis(dimethyl sulphone) complex". *Acta Crystallographica Section B* 37 (8) 1981, pp. 1568–1571. DOI: 10.1107/S0567740881006596.

[378] I. Goldberg. "Structure and binding in molecular complexes of cyclic polyethers. I. 1,4,7,10,13,16-Hexaoxacyclooctadecane (18-crown-6)–dimethyl acetylenedicarboxylate at –160°C". *Acta Crystallographica Section B* 31 (3) 1975, pp. 754–762. DOI: 10.1107/S0567740875003755.

[379] Edzard M. Geertsema, Auke Meetsma, and Ben L. Feringa. "Asymmetric Synthesis of Overcrowded Alkenes by Transfer of Axial Single Bond Chirality to Axial Double Bond Chirality". *Angewandte Chemie International Edition* 38 (18) 1999, pp. 2738–2741. DOI: 10.1002/(SICI)1521-3773(19990917)38:18<2738::AID-ANIE2738>3.0.CO;2-Z.

[380] Carsten Präsang, Adrian C. Whitwood, and Duncan W. Bruce. "Halogen-Bonded Cocrystals of 4-(N,N-Dimethylamino)pyridine with Fluorinated Iodobenzenes". *Crystal Growth & Design* 9 (12) 2009, pp. 5319–5326. DOI: 10.1021/cg900823d.

[381] Clifford W. Padgett et al. "New Conformations and Binding Modes in Halogen-Bonded and Ionic Complexes of 2,3,5,6-Tetra(2′-pyridyl)pyrazine". *Crystal Growth & Design* 5 (2) 2005, pp. 745–753. DOI: 10.1021/cg049730z.

[382] Michael C. Pfrunder et al. "Exploitation of the Menshutkin Reaction for the Controlled Assembly of Halogen Bonded Architectures Incorporating 1,2-Diiodotetra
-fluoro
-ben
-zene and 1,3,5-Tri
-iodotri
-fluoro
-benzene". *Crystal Growth & Design* 12 (2) 2012, pp. 714–724. DOI: 10.1021/cg201017r.

[383] Bernhard Neumüller and Kurt Dehnicke. "Die Reaktion von SePh2 mit N-Chlor
-succin
-imid. Kristallstrukturen von $[SeCl_2Ph_2]$ und $[SeCl_2Ph_2(Succinimid)_2]$". *Zeit
-schrift für anorganische und allgemeine Chemie* 632 (1) 2006, pp. 31–33. DOI: 10.1002/zaac.200500364.

[384] Claus Krieger and François Diederich. "Structure of host-guest complexes of 1′,1″-dimethyl-dispiro[1,6,20,25-tetraoxa[6.1.6.1]paracyclophane-13,4′:32,4″-bispiperidine] with benzene and p-xylene". *Chemische Berichte* 118 (9) 1985, pp. 3620–3631. DOI: 10.1002/cber.19851180916.

[385] Pierangelo Metrangolo et al. "Fluorinated liquid crystals formed by halogen bonding". *Chem. Commun.* (31) 2006, pp. 3290–3292. DOI: 10.1039/B605101D.

[386] James H. Gall et al. "Discovery and crystal structure of a novel chlorocarbon host: perchlorofluorene-9-spirocyclohexa-2′,5′-diene". *Tetrahedron Letters* 26 (33) 1985, pp. 4005–4008. DOI: http://dx.doi.org/10.1016/S0040-4039(00)98710-0.

[387] Dominik Cinčić, Tomislav Friščić, and William Jones. "Isostructural Materials Achieved by Using Structurally Equivalent Donors and Acceptors in Halogen-Bonded Cocrystals". *Chemistry – A European Journal* 14 (2) 2008, pp. 747–753. DOI: 10.1002/chem.200701184.

[388] A. K. Wilkerson, J. B. Chodak, and C. E. Strouse. "X-ray structure determination of the 1:1 charge transfer complex of naphthalene and tetrachlorophthalic anhydride at -153ř". *Journal of the American Chemical Society* 97 (11) 1975, pp. 3000–3004. DOI: 10.1021/ja00844a015.

[389] Kenji Kobayashi et al. "Two-dimensional hexagonal hydrogen-bonded network with triangle-like large cavities: hexakis(4-carboxyphenyl)benzene". *Tetrahedron Letters* 41 (1) 2000, pp. 89–93. DOI: http://dx.doi.org/10.1016/S0040-4039(99)02008-0.

[390] Fumio Toda et al. "INCLUSION COMPLEXES OF 1,1,2,2-TETRAPHENYL -ETHANE-1,2-DIOL AND 1,1,2,2-TETRAPHENYLETHANE". *Chemistry Letters* 15 (1) 1986, pp. 109–112. DOI: 10.1246/cl.1986.109.

[391] John J. Stezowski, Rolf-Dietrich Stigler, and Norbert Karl. "Crystal structure and charge transfer energies of complexes of the donor biphenylene with the acceptors TCNB and PMDA". *The Journal of Chemical Physics* 84 (9) 1986, pp. 5162–5170. DOI: http://dx.doi.org/10.1063/1.450669.

[392] Pierangelo Metrangolo et al. "Metric engineering of perfluorocarbon-hydrocarbon layered solids driven by the halogen bonding". *Chem. Commun.* (13) 2004, pp. 1492–1493. DOI: 10.1039/B402305F.

[393] Jonathan C. Collings et al. "Arene-perfluoroarene interactions in crystal engineering. Part 3. Single-crystal structures of 1 : 1 complexes of octafluoronaphthalene with fused-ring polyaromatic hydrocarbons". *New J. Chem.* 25 (11) 2001, pp. 1410–1417. DOI: 10.1039/B105502J.

[394] Julie I. Jay et al. "Noncovalent Interactions in 2-Mercapto-1-methylimidazole Complexes with Organic Iodides". *Crystal Growth & Design* 1 (6) 2001, pp. 501–507. DOI: 10.1021/cg015538a.

[395] Andrei S. Batsanov et al. "Octafluoronaphthalene–1,8-diaminonaphthalene (1/1)". *Acta Crystallographica Section E* 57 (10) 2001, pp. 950–952. DOI: 10.1107/S1600536801015100.

[396] Paul S. Sidhu et al. "Polymorphism, Structure, Guest Conformation, and Dynamics in the Inclusion Compound of 1,2-Dichloroethane with Tris(5-acetyl-3-thienyl) Methane: a Combined Single Crystal and Powder X-ray Diffraction, ^{13}C CP/MAS, and ^{2}H NMR Study". *The Journal of Physical Chemistry B* 106 (34) 2002, pp. 8569–8581. DOI: 10.1021/jp013983a.

[397] Koichi Tanaka, Yohei Yamamoto, and Mino R. Caira. "Guest-dependent photochromism of 3,3′-bis-(4-fluoro-phenyl)-3H,3′H- [2,2′]biindenylidene-1,1′-dione in its inclusion crystals". *CrystEngComm* 6 (2) 2004, pp. 1–4. DOI: 10.1039/B314140C.

[398] Kapildev K. Arora and V.R. Pedireddi. "Host–guest complexes of 3,5-dinitrobenzo -nitrile: channels and sandwich supramolecular architectures". *Tetrahedron* 60 (4) 2004, pp. 919–925. DOI: http://dx.doi.org/10.1016/j.tet.2003.11.028.

[399] Karl J. P. Davy et al. "Vapour phase assembly of a halogen bonded complex of an isoindoline nitroxide and 1,2-diiodotetrafluorobenzene". *CrystEngComm* 13 (16) 2011, pp. 5062–5070. DOI: 10.1039/C1CE05344B.

[400] Baoming Ji et al. "Symmetrical Bifurcated Halogen Bond: Design and Synthesis". *Crystal Growth & Design* 11 (8) 2011, pp. 3622–3628. DOI: 10.1021/cg200603z.

[401] Luca Russo et al. "Solution stoichiometry determines crystal stoichiometry in halogen-bonded supramolecular complexes". *CrystEngComm* 9 (5) 2007, pp. 341–344. DOI: 10.1039/B702512M.

[402] Sanchao Liu et al. "Polymer charge-transfer complexes for opto-electronic applications". *Synthetic Metals* 159 (14) 2009, pp. 1438–1442. DOI: http://dx.doi.org/10.1016/j.synthmet.2009.03.030.

[403] Rico E. Del Sesto et al. "Chemical Reduction of 2,4,6-Tricyano-1,3,5-triazine and 1,3,5-Tricyanobenzene. Formation of Novel 4,4′,6,6′-Tetracyano-2,2′-bitriazine and Its Radical Anion†". *The Journal of Organic Chemistry* 68 (9) 2003, pp. 3367–3379. DOI: 10.1021/jo025833h.

[404] Hans Bock, Norbert Nagel, and Christian Näther. "Wechselwirkungenin
-Molekül
-kristallen, 147[1,2]. Isostrukturelle Wirt/Gast-Einschluß-Verbindungen von N,N'-Ditosyl- p-phenylendiamin mit Aceton, Cyclopentanon, Cyclopent-2-en-1-on, 1,3-

Dioxolan, Tetrahydrofuran und 2,5-Dihydrofuran". *Zeitschrift für Naturforschung B* 53b 1998, pp. 1401–1412.

[405] Louis A. Carpino et al. "Advantageous applications of azabenzotriazole (triazolopyr -id
-ine)-based coupling reagents to solid-phase peptide synthesis". *J. Chem. Soc., Chem. Commun.* (2) 1994, pp. 201–203. DOI: 10.1039/C39940000201.

[406] Kohji Suda and Toshikatsu Takanami. "A Novel Electrochemical Oxidation Reactions Utilizing Cyclodextrins. Anodic Oxidation of Indole – Cyclodextrin – Alcohol System". *Chemistry Letters* 23 (10) 1994, pp. 1915–1916. DOI: 10.1246/cl.1994.1915.

[407] Doyle Britton. "A 1:1 complex of 2,4,5,6-tetrachloro-1,3-dicyanobenzene with pyrene". *Acta Crystallographica Section E* 61 (12) 2005, pp. 4188–4189. DOI: 10.1107/S1600536805037499.

[408] Ayesha Jacobs, Luigi R. Nassimbeni, and Jana H. Taljaard. "Inclusion compounds of isomeric xanthenol hosts with aniline". *CrystEngComm* 7 (120) 2005, pp. 731–734. DOI: 10.1039/B515436G.

[409] Julien Lieffrig et al. "Competition between the C–H···N Hydrogen Bond and C–I···N Halogen Bond in TCNQFn (n = 0, 2, 4) Salts with Variable Charge Transfer". *Crystal Growth & Design* 12 (8) 2012, pp. 4248–4257. DOI: 10.1021/cg3007519.

[410] Roberta Bertani et al. "Supramolecular rods via halogen bonding-based self-assembly of fluorinated phosphazene nanopillars". *Inorganica Chimica Acta* 360 (3) 2007. Protagonists in Chemistry: Vincenzo Balzani, pp. 1191–1199. DOI: http://dx.doi.org/10.1016/j.ica.2006.10.013.

[411] Marisa Belicchi Ferrari et al. "Synthesis, spectroscopic and structural characterization of methylpyruvate- and pyridoxal-hydrazinopyruvoylthiosemicarbazones". *Inorganica Chimica Acta* 223 (1–2) 1994, pp. 77–86. DOI: http://dx.doi.org/10.1016/0020-1693(94)03991-7.

[412] Carsten Prasang, Adrian C. Whitwood, and Duncan W. Bruce. "Spontaneous symmetry-breaking in halogen-bonded, bent-core liquid crystals: observation of a chemically driven Iso-N-N* phase sequence". *Chem. Commun.* (18) 2008, pp. 2137–2139. DOI: 10.1039/B719555A.

[413] D. Chasseau et al. "Réexamen de la structure du complexe hexaméthylène-tétrathia -fulvalène–tétracyanoquinodiméthane". *Acta Crystallographica Section B* 34 (2) 1978, pp. 689–691. DOI: 10.1107/S0567740878003830.

[414] H. Pritzkow. "Die Kristallstruktur von Stickstofftrijodid-1-Pyridin NJ3
cdot J2
cdot C6H12N4". *Zeitschrift für anorganische und allgemeine Chemie* 409 (2) 1974, pp. 237–247. DOI: 10.1002/zaac.19744090213.

[415] Ion Stoll et al. "Controlling the self assembly of arene functionalised 2-aminopyrimid -ines by arene-perfluoroarene interaction and by silver(I) complex formation". *CrystEngComm* 11 (2) 2009, pp. 306–317. DOI: 10.1039/B811297E.

[416] Todd B. Marder et al. "Crystalline TCNQ and TCNE adducts of the diborane(4) compounds $B_2(1,2\text{-}E_2C_6H_4)_2$ (E = O or S)". *J. Chem. Soc., Dalton Trans.* (13) 1999, pp. 2127–2132. DOI: 10.1039/A901169B.

[417] Jin-Ming Gao et al. "Preparation and structural elucidation of (—)-tetrahydrober -ber
-ine-(+)-2,3-di(p-toluyl) tartaric acid complex". *Journal of Molecular Structure* 892 (1–3) 2008, pp. 466–469. DOI: http://dx.doi.org/10.1016/j.molstruc.2008.06.013.

[418] A. Noman M. M. Rahman et al. "Pi-halogen dimers and V-shaped tetrahalo aryl inclusion hosts". *CrystEngComm* 4 (84) 2002, pp. 510–513. DOI: 10.1039/B207305F.

[419] Doyle Britton and William B. Gleason. "Dicyanodurene–*p*-tetrafluorodiiodobenzene (1/1)". *Acta Crystallographica Section E* 58 (12) 2002, pp. 1375–1377. DOI: 10.1107/S1600536802019943.

[420] Andrei S. Batsanov and Svetlana B. Lyubchik. "4,4′,5,5′-Tetramethyltetrathiafulvalene–1,4-dinitrobenzene (1/1)". *Acta Crystallographica Section E* 59 (2) 2003, pp. 155–157. DOI: 10.1107/S1600536803000114.

[421] Alessandra De Santis et al. "N···Br Halogen Bonding: One-Dimensional Infinite Chains through the Self-Assembly of Dibromotetrafluorobenzenes with Dipyridyl Derivatives". *Chemistry – A European Journal* 9 (16) 2003, pp. 3974–3983. DOI: 10.1002/chem.200204655.

[422] Alessandra Forni et al. "Halogen Bond Distance as a Function of Temperature". *Crystal Growth & Design* 4 (2) 2004, pp. 291–295. DOI: 10.1021/cg034138f.

[423] H. Loc Nguyen et al. "Halogen Bonding: A New Interaction for Liquid Crystal Formation". *Journal of the American Chemical Society* 126 (1) 2004, pp. 16–17. DOI: 10.1021/ja0369941.

[424] Jean-Laurent Syssa-Magale et al. "The tailoring of crystal structures via the self-assembly of organic coordination compounds by NI non-covalent halogen bonds: co-crystals of sterically hindered N-heterocycles and 1,4-diiodo-tetrafluorobenzene". *CrystEngComm* 7 (50) 2005, pp. 302–308. DOI: 10.1039/B500009B.

[425] Hadi D. Arman, Trupta Kaulgud, and Edward R. T. Tiekink. "1,2,4,5-Tetrafluoro-3,6-diiodobenzene–4-(pyridin-4-ylsulfanyl)pyridine (1/1)". *Acta Crystallographica Section E* 66 (10) 2010, p. 2683. DOI: 10.1107/S1600536810038316.

[426] Aleksandra Wasilewska, Maria Gdaniec, and Tadeusz Polonski. "Co-crystals of iodopentafluorobenzene with nitrogen donors: 2-D molecular assemblies through halogen bonding and aryl-perfluoroaryl interactions". *CrystEngComm* 9 (3) 2007, pp. 203–206. DOI: 10.1039/B617929K.

[427] Rosalba Liantonio et al. "2,2′:6′,2″-Terpyridine as Monodentate Ligand: Halogen Bonding Driven Formation of Discrete 2 : 1 Aggregates with 1,2,4,5-Tetrafluoro-3,6-diiodobenzene". *Collection of Czechoslovak Chemical Communications* 67 2002, pp. 1373–1382.

[428] Koichi Tanaka et al. "Chiral inclusion crystallization of tetra(-bromophenyl)ethylene by exposure to the vapor of achiral guest molecules: a novel racemic-to-chiral transformation through gas-solid reaction". *Chem. Commun.* (5) 2000, pp. 413–414. DOI: 10.1039/A909782A.

[429] Catherine Perkins et al. "Diiodoacetylene: compact, strong ditopic halogen bond donor". *CrystEngComm* 14 (9) 2012, pp. 3033–3038. DOI: 10.1039/C2CE00029F.

[430] Teresa Olszewska et al. "Thioamides and selenoamides with chirality solely due to hindered rotation about the C–N bond: enantioselective complexation with optically active hosts". *Tetrahedron: Asymmetry* 16 (22) 2005, pp. 3711–3717. DOI: http://dx.doi.org/10.1016/j.tetasy.2005.09.017.

[431] James M. Blackwell, Warren E. Piers, and Masood Parvez. "Mechanistic Studies on Selectivity in the $B(C_6F_5)_3$-Catalyzed Allylstannation of Aldehydes: Is Hypercoordination at Boron Responsible?" *Organic Letters* 2 (5) 2000, pp. 695–698. DOI: 10.1021/ol0000105.

[432] Gerardus J. Kemperman et al. "Induced fit phenomena in clathrate structures of cephalosporins". *J. Chem. Soc., Perkin Trans. 2* (7) 2000, pp. 1425–1429. DOI: 10.1039/B001692F.

[433] Rosalba Liantonio et al. "4,4′-Bipyridine 1,2-diiodo-3,4,5,6-tetrafluorobenzene". *Acta Crystallographica Section E* 58 (5) 2002, pp. 575–577. DOI: 10.1107/S1600536802007201.

[434] Doyle Britton. "Planar packing of tetrachlorodicyanobenzene isomers". *Acta Crystallographica Section B* 58 (3 Part 2) 2002, pp. 553–563. DOI: 10.1107/S0108768102003348.

[435] John A. Cowan et al. "On the Interaction between N-Heterocyclic Carbenes and Organic Acids: Structural Authentication of the First N—H·C Hydrogen Bond and Remarkably Short C—H·O Interactions". *Angewandte Chemie International Edition* 41 (8) 2002, pp. 1432–1434. DOI: 10.1002/1521-3773(20020415)41:8<1432::AID-ANIE1432>3.0.CO;2-M.

[436] Rosalba Liantonio et al. "Perfluorocarbon–hydrocarbon self-assembly. Part 16: Anilines as new electron donor modules for halogen bonded infinite chain formation". *Tetrahedron* 58 (20) 2002. Fluorous Chemistry, pp. 4023–4029. DOI: http://dx.doi.org/10.1016/S0040-4020(02)00264-8.

[437] Andrei S. Batsanov and Judith A. K. Howard. "Morpholine–β-iodophenylacetylene (1/1) revisited: an exceptionally short I$\cdots$N contact". *Acta Crystallographica Section C* 56 (2) 2000, pp. 252–253. DOI: 10.1107/S0108270199014705.

[438] Sylke Apel et al. "Weak Hydrogen Bonding as a Basis for Concentration-Dependent Guest Selectivity by a Cyclophane Host". *Chemistry – A European Journal* 8 (16) 2002, pp. 3678–3686. DOI: 10.1002/1521-3765(20020816)8:16<3678::AID-CHEM3678>3.0.CO;2-4.

[439] Katharine F. Bowes et al. "*N*,*N*′-Dithiobisphthalimide–nitrobenzene (2/1): a *Pn* solvate with localized solvent molecules ordered head-to-tail in channels". *Acta Crystallographica Section C* 58 (9) 2002, pp. 551–554. DOI: 10.1107/S0108270102012830.

[440] Jan Janczak, Ryszard Kubiak, and Ewa Bukowska. "Isomorphic complexes formed by recrystallisation of M(II)Pc (M(II) = Mg, Mn and Zn) in liquid 2-amino-3-picoline". *Journal of Molecular Structure* 937 (1–3) 2009, pp. 25–33. DOI: http://dx.doi.org/10.1016/j.molstruc.2009.08.007.

[441] EricW. Reinheimer, Hanhua Zhao, and KimR. Dunbar. "Structural Studies of the 1:1 Complex of o-3,4-Dimethyltetrathiafulvalene (o-Me2TTF) and 1,2,4,5-Tetra -cyano

-benzene (TCNB)". English. *Journal of Chemical Crystallography* 40 (6) 2010, pp. 514–519. DOI: 10.1007/s10870-010-9688-8.

[442] Samuel Ebenezer and P. Thomas Muthiah. "Design of Co-crystals/Salts of Aminopyrimidines and Carboxylic Acids through Recurrently Occurring Synthons". *Crystal Growth & Design* 12 (7) 2012, pp. 3766–3785. DOI: 10.1021/cg3005954.

[443] Shohei Saito, Kyohei Matsuo, and Shigehiro Yamaguchi. "Polycyclic π-Electron System with Boron at Its Center". *Journal of the American Chemical Society* 134 (22) 2012, pp. 9130–9133. DOI: 10.1021/ja3036042.

[444] Akiko Hori et al. "Luminescence from π-π stacked bipyridines through arene-perfluoroarene interactions". *CrystEngComm* 11 (4) 2009, pp. 567–569. DOI: 10.1039/B822007G.

[445] Maria T Messina et al. "Intermolecular recognition between hydrocarbon oxygen-donors and perfluorocarbon iodine-acceptors: the shortest O∘I non-covalent bond". *Tetrahedron* 57 (40) 2001, pp. 8543–8550. DOI: http://dx.doi.org/10.1016/S0040-4020(01)00794-3.

[446] Susanta K. Nayak et al. "C-BrO supramolecular synthon: in situ cryocrystallography of low melting halogen-bonded complexes". *CrystEngComm* 14 (13) 2012, pp. 4259–4261. DOI: 10.1039/C2CE25403D.

[447] Pierangelo Metrangolo et al. "Highly Interpenetrated Supramolecular Networks Supported by N···I Halogen Bonding". *Chemistry – A European Journal* 13 (20) 2007, pp. 5765–5772. DOI: 10.1002/chem.200601653.

[448] Michele Baldrighi et al. "Halogen-bonded and interpenetrated networks through the self-assembly of diiodoperfluoroarene and tetrapyridyl tectons". *Journal of Fluorine Chemistry* 131 (11) 2010. 2010 {ACS} Award Issue "For Creative Work in Fluorine Chemistry" Russell P. Hughes, pp. 1218–1224. DOI: http://dx.doi.org/10.1016/j.jfluchem.2010.06.001.

[449] Rosa Bailey Walsh et al. "Crystal Engineering through Halogen Bonding: Complexes of Nitrogen Heterocycles with Organic Iodides". *Crystal Growth & Design* 1 (2) 2001, pp. 165–175. DOI: 10.1021/cg005540m.

[450] Riccardo Bianchi, Alessandra Forni, and Tullio Pilati. "The Experimental Electron Density Distribution in the Complex of (E)-1,2-Bis(4-pyridyl)ethylene with 1,4-Diiodotetrafluorobenzene at 90 K". *Chemistry – A European Journal* 9 (7) 2003, pp. 1631–1638. DOI: 10.1002/chem.200390187.

[451] Andrei S. Batsanov, Ibraheem A. I. Mkhalid, and Todd B. Marder. "4,4′-Di-*tert*-butyl-2,2′-bipyridine–hexafluorobenzene (1/1)". *Acta Crystallographica Section E* 63 (3) 2007, pp. 1196–1198. DOI: 10.1107/S160053680700445X.

[452] Pierangelo Metrangolo et al. "Dendrimeric Tectons in Halogen Bonding-Based Crystal Engineering". *Crystal Growth & Design* 8 (2) 2008, pp. 654–659. DOI: 10.1021/cg700870t.

[453] Laila C. Roper et al. "Experimental and Theoretical Study of Halogen-Bonded Complexes of DMAP with Di- and Triiodofluorobenzenes. A Complex with a Very Short N···I Halogen Bond". *Crystal Growth & Design* 10 (8) 2010, pp. 3710–3720. DOI: 10.1021/cg100549u.

[454] Yoshio Aso et al. "Dichalcogen-Bridged Acenaphthenes as New Electron Donors". *Bulletin of the Chemical Society of Japan* 61 (6) 1988, pp. 2013–2018. DOI: 10.1246/bcsj.61.2013.

[455] Dmitry S. Yufit et al. "Low-melting molecular complexes. Halogen bonds in molecular complexes of bromoform". *CrystEngComm* 14 (23) 2012, pp. 8222–8227. DOI: 10.1039/C2CE26191J.

[456] Mustafa Arslan et al. "A 2:1 complex of 1,3-bis(9H-carbazol-9-yl)propane and tetrachloro-*p*-benzoquinone (*p*-chloranil)". *Acta Crystallographica Section E* 62 (9) 2006, pp. 4055–4057. DOI: 10.1107/S1600536806031886.

[457] Erhard Günther et al. "Novel S-Heteroquinoid Electron Acceptors. Synthesis of 2,5-Bis(cyanoimino)-2,5-dihydrothieno-[3,2-b]thiophenes (DCNTTs) and Conductive Charge Transfer Complexes". *Angewandte Chemie International Edition in English* 29 (2) 1990, pp. 204–205. DOI: 10.1002/anie.199002041.

[458] Erhard Günther et al. "Mehrstufige reversible Redoxsysteme, LIX. Leitfähige Charge-Transfer-Komplexe von 2,5-Bis(cyanimino)-2,5-dihydrothieno[3,2-b]thiophenen mit organischen Donoren". *Chemische Berichte* 125 (8) 1992, pp. 1919–1926. DOI: 10.1002/cber.19921250820.

[459] Anand Sundararaman et al. "Electronic Communication and Negative Binding Cooperativity in Diborylated Bithiophenes". *Journal of the American Chemical Society* 128 (51) 2006, pp. 16554–16565. DOI: 10.1021/ja064396b.

[460] A. Takai S. Nakatsuji A. Kitamura et al. "CT Complexes Derived from Verdazyl Radicals". *Zeitschrift für Naturforschung B* 53 (4) 1998, pp. 495–502.

[461] Roland Köster, Wilhelm Schüßler, and Roland Boese. "9-Fluor-9-borabicyclo -[3.3.1] -nonan - Charakterisierung in Lösung und im Kristall". *Chemische Berichte* 123 (10) 1990, pp. 1945–1952. DOI: 10.1002/cber.19901231004.

[462] Pierangelo Metrangolo et al. "4,4′-Bipyridine–2,4,5,6-tetrafluoro-1,3-diiodobenzene (1/1)". *Acta Crystallographica Section E* 63 (11) 2007, p. 4243. DOI: 10.1107/S1600536807047630.

[463] G. M. Frankenbach, M. A. Beno, and J. M. Williams. "Charge-transfer complex of 2,2',5,5'-tetrathiafulvalene and *p*-benzoquinone". *Acta Crystallographica Section C* 47 (4) 1991, pp. 762–764. DOI: 10.1107/S0108270190007958.

[464] Duncan W. Bruce et al. "Structure–Function Relationships in Liquid-Crystalline Halogen-Bonded Complexes". *Chemistry – A European Journal* 16 (31) 2010, pp. 9511–9524. DOI: 10.1002/chem.201000717.

[465] Duncan W. Bruce et al. "Mesogenic, trimeric, halogen-bonded complexes from alkoxystilbazoles and 1,4-diiodotetrafluorobenzene". *New J. Chem.* 32 (3) 2008, pp. 477–482. DOI: 10.1039/B709107A.

[466] Christer B. Aakeroy et al. "Ten years of co-crystal synthesis; the good, the bad, and the ugly". *CrystEngComm* 10 (12) 2008, pp. 1816–1821. DOI: 10.1039/B811809D.

[467] Burcak Icli et al. "Dative boron-nitrogen bonds in structural supramolecular chemistry: multicomponent assembly of prismatic organic cages". *Chem. Sci.* 2 (9) 2011, pp. 1719–1721. DOI: 10.1039/C1SC00320H.

[468] Hadi D. Arman, Trupta Kaulgud, and Edward R. T. Tiekink. "1,2,4,5-Tetrafluoro-3,6-diiodobenzene–2,3-bis(pyridin-2-yl)pyrazine (1/1)". *Acta Crystallographica Section E* 66 (11) 2010, p. 2885. DOI: 10.1107/S1600536810041668.

[469] André C. B. Lucassen et al. "Co-Crystallization of Sym-Triiodo-Trifluorobenzene with Bipyridyl Donors: Consistent Formation of Two Instead of Anticipated Three N···I Halogen Bonds". *Crystal Growth & Design* 7 (2) 2007, pp. 386–392. DOI: 10.1021/cg0607250.

[470] Claudio Guardigli et al. "Design and Synthesis of New Tectons for Halogen Bonding-driven Crystal Engineering". *Supramolecular Chemistry* 15 (3) 2003, pp. 177–188. DOI: 10.1080/1061027031000078248.

[471] Peter G. Jones et al. "9-Methylene-10,10-bis(phenylethynyl)fluorene–tetracyano-*para*-quinodimethane (2/3)". *Acta Crystallographica Section E* 61 (4) 2005, pp. 957–959. DOI: 10.1107/S1600536805007142.

[472] Kari Raatikainen and Kari Rissanen. "Breathing molecular crystals: halogen- and hydrogen-bonded porous molecular crystals with solvent induced adaptation of the nanosized channels". *Chem. Sci.* 3 (4) 2012, pp. 1235–1239. DOI: 10.1039/C2SC00997H.

[473] Ricardo Diaz Calleja et al. "Structure, electrical conductivity and dielectric relaxation of charge-transfer complexes with 2,4,6,8-tetramethoxydibenzoselenophene as a donor". *J. Mater. Chem.* 3 (5) 1993, pp. 489–497. DOI: 10.1039/JM9930300489.

[474] Dongpeng Yan et al. "A Cocrystal Strategy to Tune the Luminescent Properties of Stilbene-Type Organic Solid-State Materials". *Angewandte Chemie International Edition* 50 (52) 2011, pp. 12483–12486. DOI: 10.1002/anie.201106391.

[475] Amparo Salmerón-Valverde and Sylvain Bernès. "9-(Dicyanomethylidene)fluorene–tetrathiafulvalene (1/1)". *Acta Crystallographica Section E* 68 (4) 2012, p. 932. DOI: 10.1107/S1600536812008124.

[476] Durre Shahwar et al. "3-Acetyl-1-(3-chlorophenyl)thiourea". *Acta Crystallographica Section E* 68 (4) 2012, p. 1189. DOI: 10.1107/S1600536812012147.

[477] Christer B. Aakeroy et al. "Combining halogen bonds and hydrogen bonds in the modular assembly of heteromeric infinite 1-D chains". *Chem. Commun.* (41) 2007, pp. 4236–4238. DOI: 10.1039/B707458A.

[478] Andrei S. Batsanov et al. "X-Ray crystal structure and solid-state properties of a 1:1 complex of tetrathiafulvalene (TTF) and 1-oxo-2,6-dimethyl-4-dicyano
-meth
-yl
-ene
-cyclo
-hexa-2,5-diene". *J. Mater. Chem.* 4 (11) 1994, pp. 1719–1722. DOI: 10.1039/JM9940401719.

[479] Prakash C. Srivastava et al. "X-ray characterization of $Te(S_2CNC_5H_{10})_2$ and $TeI_2[(C_{13}H_{10}N_2S)_2] \cdot 4C_4H_8TeI_2$; the first Te–C bond cleaved products obtained in the substitution reactions of organo(heterocyclic)tellurium(IV) derivatives". *Polyhedron* 27 (2) 2008, pp. 835–848. DOI: http://dx.doi.org/10.1016/j.poly.2007.11.007.

[480] Colm Crean, John F. Gallagher, and Albert C. Pratt. "2,2′-[2,3-Dihydro-2-(prop-2-enyl)-1*H*-isoindole-1,3-diylidene]bis(propanedinitrile)–tetrathiafulvalene (1/1), TCPI–

TTF". *Acta Crystallographica Section C* 58 (1) 2002, pp. 36–38. DOI: 10.1107/S0108270101017632.

[481] Hermann-Josef Frohn et al. "First Isolation and Structural Characterization of N-base Adducts of Mono- and Perfluoroaryliodine(III) Compounds, $ArfIX_2$ (X = F, CN)". *Zeitschrift für anorganische und allgemeine Chemie* 635 (13-14) 2009, pp. 2249–2257. DOI: 10.1002/zaac.200900169.

[482] Masahide Tominaga, Kosuke Katagiri, and Isao Azumaya. "Hydrogen-bonded networks formed from tri- and tetra-substituted adamantanes bearing dimethoxyphenol moieties and their 1,3,5-trinitrobenzene complexes via charge-transfer interactions". *CrystEngComm* 12 (4) 2010, pp. 1164–1170. DOI: 10.1039/B917654C.

[483] Ricardo Diaz Calleja et al. "Structure, electrical conductivity and dielectric relaxation of the 1,2-bis(methylthio)benzene-2,3-dichloro-5,6-dicyano-1,4-benzoquinone 1 : 1 charge-transfer complex". *J. Mater. Chem.* 5 (3) 1995, pp. 389–394. DOI: 10.1039/JM9950500389.

[484] Silas C. Blackstock, Kathy Poehling, and Melinda L. Greer. "Pericyclic Arrays in the Solid State. Azooxide and Azodioxide Donor-Acceptor Complexes with Tetracyanoethylene". *Journal of the American Chemical Society* 117 (24) 1995, pp. 6617–6618. DOI: 10.1021/ja00129a036.

[485] Ken S. Feldman and Robert F. Campbell. "Efficient Stereo- and Regiocontrolled Alkene Photodimerization through Hydrogen Bond Enforced Preorganization in the Solid State". *The Journal of Organic Chemistry* 60 (7) 1995, pp. 1924–1925. DOI: 10.1021/jo00112a007.

[486] Peng Liu et al. "Cyclohexa-2,5-diene-1,4-dione–1,2,4,5-tetrafluoro-3,6-diiodobenzene (1/1)". *Acta Crystallographica Section E* 68 (5) 2012, p. 1431. DOI: 10.1107/S1600536812015930.

[487] Axel Michalides et al. "Supramoleküle aus Kronenethern und geminalen Sulfonen: Synthese von vier binären Komplexen und Kristallstruktur von $(CH_2CH_20)_4$-$2H_2C(S0_2C_2H_5)_2$". *Zeitschrift für Naturforschung B* 50b 1995, pp. 1018–1024.

[488] B. Viossat et al. "12-Imino-12*H*-benzimidazo[2,1-*b*][1,3]benzothiazine–7,7,8,8-Tétra-cyano-*p*-quinodiméthane, IBB–TCNQ". *Acta Crystallographica Section C* 51 (9) 1995, pp. 1896–1898. DOI: 10.1107/S0108270195002472.

[489] Hai Yue Gao et al. "Phosphorescent Cocrystals Assembled by 1,4-Di
-iodo
-tetra
-fluoro
-benz
-ene and Fluorene and Its Heterocyclic Analogues Based on C–I···π Halogen Bonding". *Crystal Growth & Design* 12 (9) 2012, pp. 4377–4387. DOI: 10.1021/cg300515a.

[490] Guillermo Minguez Espallargas et al. "One-dimensional organization of free radicals via halogen bonding". *CrystEngComm* 14 (20) 2012, pp. 6381–6383. DOI: 10.1039/C2CE26131F.

[491] Tsuyoshi Murata et al. "Modulation of charge-transfer complexes assisted by complementary hydrogen bonds of nucleobases: TCNQ complexes of a uracil-substituted EDO-TTF". *CrystEngComm* 14 (20) 2012, pp. 6881–6887. DOI: 10.1039/C2CE25889G.

[492] Louis J. Farrugia, James H. Gall, and David D. MacNicol. "1,2,4,5-Tetrakis(phenyl
-sulfonyl)benzene: a novel quadruped host with D2 symmetry having ordered sulfolane and cycloheptanone guests". *Chem. Commun.* 48 (86) 2012, pp. 10600–10602. DOI: 10.1039/C2CC35647C.

[493] Markus Pridöhl et al. "Preparation and X-Ray structural analysis of the remarkably stable Bis(triphenylmethyl)trisulfane-2-oxide R2S3O". *Zeitschrift für anorganische und allgemeine Chemie* 621 (10) 1995, pp. 1672–1676. DOI: 10.1002/zaac.19956211010.

[494] Stefanie Wilker and Gerhard Erker. "Stereochemistry of the [4 + 2] Cycloaddition of Diarylselenoketones with Conjugated Dienes". *Journal of the American Chemical Society* 117 (44) 1995, pp. 10922–10930. DOI: 10.1021/ja00149a015.

[495] Venkat R. Thalladi et al. "Supramolecular synthons in crystal engineering. Structure simplification, synthon robustness and supramolecular retrosynthesis". *Chem. Commun.* (3) 1996, pp. 401–402. DOI: 10.1039/CC9960000401.

[496] B. Shaanan, U. Shmueli, and M. Colapietro. "Structure and packing arrangement of molecular compounds. X. 9,10-Diazaphenanthrene–2,3-dichloro-5,6-dicyano-1,4-benzoquinone (2:1)". *Acta Crystallographica Section B* 38 (3) 1982, pp. 818–824. DOI: 10.1107/S0567740882004142.

[497] P. W. R. Corfield and S. J. La Placa. "Structure of the charge-transfer salt 2,2',5,5'-tetraselenafulvalene–7,7,8,8-tetracyano-*p*-quinodimethane (TSeF–TCNQ)". *Acta Crystallographica Section B* 52 (2) 1996, pp. 384–387. DOI: 10.1107/S0108768195012286.

[498] V. R. Pedireddi et al. "Creation of crystalline supramolecular arrays: a comparison of co-crystal formation from solution and by solid-state grinding". *Chem. Commun.* (8) 1996, pp. 987–988. DOI: 10.1039/CC9960000987.

[499] Motonari Shibakami, Masanori Tamura, and Akira Sekiya. "Crystal structures of inclusion complexes of cholic acid with aniline, mono- and difluorinated anilines: Guest information expressed by molecular assembly". English. *Journal of inclusion phenomena and molecular recognition in chemistry* 22 (4) 1995, pp. 299–311. DOI: 10.1007/BF00707782.

[500] Jonathan C. Collings et al. "Arene-perfluoroarene interactions in crystal engineering 8: structures of 1:1 complexes of hexafluorobenzene with fused-ring polyaromatic hydrocarbons". *New J. Chem.* 26 (12) 2002, pp. 1740–1746. DOI: 10.1039/B207102A.

[501] Hans Bock et al. "Wechselwirkungen in Kristallen, 88 [1 - 3] Donator/Akzeptor - Komplexe von Alkylbenzolen, Pyren oder Perylen mit Tetrahalogen-/?-benzochinonen: Strukturen und Eigenschaften". *Zeitschrift für Naturforschung B* 51b 1996, pp. 1521–1537.

[502] *Sincris Software List*. Online. URL: http://ww1.iucr.org/sincris-top/logiciel/abc.html.

[503] R. B. G. Ravelli et al. "*STRATEGY*: a program to optimize the starting spindle angle and scan range for X-ray data collection". *Journal of Applied Crystallography* 30 (5 Part 1) 1997, pp. 551–554. DOI: 10.1107/S0021889897003543.

[504] *CCP14 Repository*. Online. URL: http://www.ccp14.ac.uk/mirror/mirror.htm.

[505] S Okada and Kenji Okada. "A DIRECT-SEARCHER Automatic System (Version 4) Running on Windows Personal Computers". *Z. Kristallogr.* 215 2000, pp. 131–143.

[506] Sachiko Okada and Kenji Okada. "Detailed Differential Coefficients of Anomalous Dispersion Terms to Derivatives in Least-Squares Refinement for X-ray Crystallography". *Comput. Chem.* 24 (2) 2000, pp. 143–158.

[507] Armel Le Bail. *Open Source Crystallography Software lists by*. Online. 1995-2012. URL: http://sdpd.univ-lemans.fr/.

[508] P. McArdle. "*ABSEN* – a PC computer program for listing systematic absences and space-group determination". *Journal of Applied Crystallography* 29 (3) 1996, p. 306. DOI: 10.1107/S0021889895016311.

[509] Ian Baxter. Online. URL: http://ww1.iucr.org/sincris-top/logiciel/baxmap/index.html.

[510] J. R. Hester and F. P. Okamura. "CIF Applications. X. Automatic Construction of CIF Input Functions: *CifSieve*". *Journal of Applied Crystallography* 31 (6) 1998, pp. 965–968. DOI: 10.1107/S0021889898008978.

[511] Nicholas M. Glykos and Michael Kokkinidis. "*GraphEnt*: a maximum-entropy program with graphics capabilities". *Journal of Applied Crystallography* 33 (3 Part 2) 2000, pp. 982–985. DOI: 10.1107/S0021889800004246.

[512] J. W. Campbell et al. "*LAUEGEN* version 6.0 and *INTLDM*". *Journal of Applied Crystallography* 31 (3) 1998, pp. 496–502. DOI: 10.1107/S0021889897016683.

[513] Branton J. Campbell et al. "*ISODISPLACE*: a web-based tool for exploring structural distortions". *Journal of Applied Crystallography* 39 (4) 2006, pp. 607–614. DOI: 10.1107/S0021889806014075.

[514] S. Penel and P. Legrand. "*MAINDEX* – manual indexation for area-detector crystallographic data". *Journal of Applied Crystallography* 30 (2) 1997, p. 206. DOI: 10.1107/S0021889896015129.

[515] *Quasitiler*. Online. URL: http://www.geom.uiuc.edu/apps/quasitiler/.

[516] Leonard J. Barbour and Jerry L. Atwood. "*RES2INS*: a graphical interface for the *SHELX* program suite". *Journal of Applied Crystallography* 31 (6) 1998, pp. 963–964. DOI: 10.1107/S0021889898008395.

[517] Sachiko Okada and Kenji Okada. "*ShakePSD*: automatic phase solver covering heavy-atom methods and direct methods for Windows PCs". *Journal of Applied Crystallography* 33 (2) 2000, pp. 406–414. DOI: 10.1107/S0021889899015058.

[518] S. Desgreniers and K. Lagarec. "*XRDA*3.1 - a program for X-ray diffraction analysis on a PC". *Journal of Applied Crystallography* 31 (1) 1998, pp. 109–110. DOI: 10.1107/S0021889897007322.

[519] Paul W. Betteridge et al. "*CRYSTALS* version 12: software for guided crystal structure analysis". *Journal of Applied Crystallography* 36 (6) 2003, p. 1487. DOI: 10.1107/S0021889803021800.

[520] Kate Krennel. *Sincris additional list*. Online. 1993. URL: http://ww1.iucr.org/sincris-top/logiciel/prg-micro.html.

[521] A. J. M. Duisenberg. "Indexing in single-crystal diffractometry with an obstinate list of reflections". *Journal of Applied Crystallography* 25 (2) 1992, pp. 92–96. DOI: 10.1107/S0021889891010634.

[522] R. J. Angel. "Absorption corrections for diamond-anvil pressure cells implemented in the software package *Absorb6.0*". *Journal of Applied Crystallography* 37 (3) 2004, pp. 486–492. DOI: 10.1107/S0021889804005229.

[523] *IUCr Software List*. http://www.iucr.org/resources/other-directories/software.

[524] D. R. Powell. "*CSDSHL* - a utility for converting Cambridge Structural Database atom coordinate files to *SHELX* format". *Journal of Applied Crystallography* 25 (5) 1992, p. 663. DOI: 10.1107/S0021889892002474.

[525] S. R. Hall and H. J. Bernstein. "CIF Applications. V. *CIFtbx*2: extended tool box for manipulating CIFs". *Journal of Applied Crystallography* 29 (5) 1996, pp. 598–603. DOI: 10.1107/S0021889896006371.

[526] T. Balić Žunić and I. Vicković. "*IVTON* – a program for the calculation of geometrical aspects of crystal structures and some crystal chemical applications". *Journal of Applied Crystallography* 29 (3) 1996, pp. 305–306. DOI: 10.1107/S0021889895015081.

[527] I.-H. Suh, M.-R. Oh, and K.-H. Kim. "*CELLTR*, *HKLTR* and *COORDTR*: transformations of cell data, Miller indices and atomic coordinates and their estimated standard deviations". *Journal of Applied Crystallography* 31 (1) 1998, pp. 110–111. DOI: 10.1107/S0021889897008984.

[528] G. Eckold, M. Stein-Arsic, and H. J. Weber. "*UNISOFT* – a program package for lattice-dynamical calculations". *Journal of Applied Crystallography* 20 (2) 1987, pp. 134–139. DOI: 10.1107/S0021889887086977.

[529] Carlo Gatti and Anotoliy Volkov. *TOPXD*. Online. 2001. URL: http://harker.chem.buffalo.edu/public/topxd/index.html.

[530] L. M. Gelato and E. Parthé. "*STRUCTURE TIDY* – a computer program to standardize crystal structure data". *Journal of Applied Crystallography* 20 (2) 1987, pp. 139–143. DOI: 10.1107/S0021889887086965.

[531] *PRO-CHEMIST*. Online. URL: http://pro.chemist.online.fr/.

[532] L. M. Urzhumtseva and A. G. Urzhumtsev. "Tcl/Tk-based programs. IV. *CALCRYS*: crystallographic calculator". *Journal of Applied Crystallography* 33 (3 Part 2) 2000, p. 992. DOI: 10.1107/S0021889800003794.

[533] Hans-Joachim Klein. Online. URL: http://www.is.informatik.uni-kiel.de/~hjk/crystana.html.

[534] Ross J Angel. Online. URL: http://www.rossangel.com/home.htm.

[535] L. G. Akselrud, P. Yu. Zavalij Yu. N. Grin, and Pecharsky V. Online. URL: http://imr.chem.binghamton.edu/zavalij/CSD.html.

[536] Gordon Barr, Wei Dong, and Christopher J. Gilmore. "*PolySNAP3*: a computer program for analysing and visualizing high-throughput data from diffraction and spectroscopic sources". *Journal of Applied Crystallography* 42 (5) 2009, pp. 965–974. DOI: 10.1107/S0021889809025746.

[537] *JCrystal Software*. Online. 2012. URL: http://www.jcrystal.com/.

[538] *SPEC and C-PLOT*. Online. URL: http://www.certif.com/.

[539] D. Bourgeois et al. "An Integration Routine Based on Profile Fitting with Optimized Fitting Area for the Evaluation of Weak and/or Overlapped Two-Dimensional Laue or Monochromatic Patterns". *Journal of Applied Crystallography* 31 (1) 1998, pp. 22–35. DOI: 10.1107/S0021889897006730.

[540] M. J. Frisch et al. *Gaussian 09*. Online. 2009. URL: http://www.gaussian.com/.

[541] *Diamond*. Online. 2012. URL: http://www.crystalimpact.com/diamond/Default.htm.

[542] Yves Epelboin. *CRYSCOMP and CRYSDRAW*. Online. 1996. URL: http://ww1.iucr.org/sincris-top/logiciel/prg-cryscomp.html.

[543] *Carine*. Online. URL: http://carine.crystallography.pagespro-orange.fr/WWW%5C_FULL%5C_DESCRIPT%5C_CRYST%5C_SIM.html.

[544] J. Bergmann. *BGMN Homepage*. Online. URL: http://www.bgmn.de/index.html.

[545] Th. Proffen and S. J. L. Billinge. "*PDFFIT*, a program for full profile structural refinement of the atomic pair distribution function". *Journal of Applied Crystallography* 32 (3) 1999, pp. 572–575. DOI: 10.1107/S0021889899003532.

[546] *KUPLOT*. Online. 2003. URL: http://www.pa.msu.edu/cmp/billinge-group/discus/kuplot.html.

[547] B. Chevrier. "*PATGEN* – an automatic program to generate theoretical Patterson peaks and to compare them with experimental Patterson peaks". *Journal of Applied Crystallography* 27 (5) 1994, pp. 860–861. DOI: 10.1107/S0021889894006898.

[548] Oleg V. Dolomanov et al. "*LCELLS*: an efficient search engine for laboratory unit cells". *Journal of Applied Crystallography* 36 (3 Part 2) 2003, p. 955. DOI: 10.1107/S0021889803003704.

[549] Elisabeth Rossmanith. "*PSILAM*: a program for the calculation and graphical representation of multiple-diffraction peak location plots". *Journal of Applied Crystallography* 36 (4) 2003, pp. 1098–1100. DOI: 10.1107/S0021889803010756.

[550] Elias Vlieg. "*ROD*: a program for surface X-ray crystallography". *Journal of Applied Crystallography* 33 (2) 2000, pp. 401–405. DOI: 10.1107/S0021889899013655.

[551] K. Stróż. "*SPACER*: a program to display space-group information for a conventional and nonconventional coordinate system". *Journal of Applied Crystallography* 30 (2) 1997, pp. 178–181. DOI: 10.1107/S0021889896013453.

[552] Z.-Q. Fu and H.-F. Fan. "A computer program to derive (3+1)-dimensional symmetry operations from two-line symbols". *Journal of Applied Crystallography* 30 (1) 1997, pp. 73–78. DOI: 10.1107/S0021889896006711.

[553] Prof. A. Immirzi. *TRY*. Online. URL: http://www.theochem.unisa.it/try.html.

[554] "Guidelines for Computer Program Abstracts". *Journal of Applied Crystallography* 18 (3) 1985, pp. 189–190. DOI: 10.1107/S0021889885010111.

[555] Elisabeth Rossmanith. "*UMWEG*: a program for the calculation and graphical representation of multiple-diffraction patterns". *Journal of Applied Crystallography* 36 (6) 2003, pp. 1467–1474. DOI: 10.1107/S002188980301851X.

[556] Werner Kaminsky. "*WinXMorph*: a computer program to draw crystal morphology, growth sectors and cross sections with export files in VRML V2.0 utf8-virtual reality format". *Journal of Applied Crystallography* 38 (3) 2005, pp. 566–567. DOI: 10.1107/S0021889805012148.

[557] Werner Kaminsky. "From CIF to virtual morphology using the *WinXMorph* program". *Journal of Applied Crystallography* 40 (2) 2007, pp. 382–385. DOI: 10.1107/S0021889807003986.

[558] Andrea Thorn and George M. Sheldrick. "*ANODE*: anomalous and heavy-atom density calculation". *Journal of Applied Crystallography* 44 (6) 2011, pp. 1285–1287. DOI: 10.1107/S0021889811041768.

[559] Vladislav A. Blatov. "A method for hierarchical comparative analysis of crystal structures". *Acta Crystallographica Section A* 62 (5) 2006, pp. 356–364. DOI: 10.1107/S0108767306025591.

[560] *Crystal Explorer User Manual*. Online. URL: http://ra.bcs.uwa.edu.au/CrystalExplorer/wiki/index.php/Manual/.

[561] Harold T. Stokes, Dorian M. Hatch, and Branton J. Campbell. *Isotropy*. Online. 2007. URL: http://stokes.byu.edu/iso/isotropy.html.

[562] Sergey Stepanov. *BRL*. Online. URL: http://sergey.gmca.aps.anl.gov/brl.html.

[563] *GID_sl/TER_sl/TDRS_sl*. Online. URL: http://sergey.gmca.aps.anl.gov/.

[564] S. A. Stepanov. *x0h*. Online. URL: http://sergey.gmca.aps.anl.gov/x0h.html.

[565] *CCTBX*. Online. URL: http://cctbx.sourceforge.net/.

[566] Larry W. Finger, Martin Kroeker, and Brian H. Toby. "*DRAWxtl*, an open-source computer program to produce crystal structure drawings". *Journal of Applied Crystallography* 40 (1) 2007, pp. 188–192. DOI: 10.1107/S0021889806051557.

[567] M. Lutz and A. M. M. Schreurs. *euhedral*. Online. URL: http://www.crystal.chem.uu.nl/distr/euhedral/index.html.

[568] Allen C. Larson and Robert B. Von Dreele. *GSAS*. Online. 2011. URL: http://www.ncnr.nist.gov/programs/crystallography/software/gsas.html.

[569] R. Rudert. "Program Hardpack: Prediction and Optimisation of Crystal Structures using Atom-Atom Portentials and Point Charges". *Acta Crystallographica Section A, Supplemental* 52 (C94) 1996.

[570] S. Weber. "*JSV*1.07 – a Java structure viewer". *Journal of Applied Crystallography* 32 (5) 1999, pp. 1027–1028. DOI: 10.1107/S0021889899009140.

[571] S. Weber. "*JMap3D* – VRML isosurface generator". *Journal of Applied Crystallography* 32 (5) 1999, p. 1028. DOI: 10.1107/S0021889899009152.

[572] S. Weber. "*KOQUA*2.0: a program for simulating divergent-beam diffraction patterns for crystals and quasicrystals". *Journal of Applied Crystallography* 30 (1) 1997, pp. 85–86. DOI: 10.1107/S0021889896012459.

[573] Steffen Weber. *SteffenWeber Misc. Software*. Online. URL: http://www.jcrystal.com/steffenweber/.

[574] Cheng Dong and J. I. Langford. "*LAPODS*: a computer program for refinement of lattice parameters using optimal regression". *Journal of Applied Crystallography* 33 (4) 2000, pp. 1177–1179. DOI: 10.1107/S0021889800006622.

[575] A. Soyer. "*LAUEX*: a user-friendly program for the simulation and indexation of Laue diagrams on UNIX systems". *Journal of Applied Crystallography* 29 (4) 1996, p. 509. DOI: 10.1107/S0021889896004736.

[576] Jan Rohlíček and Michal Hušák. "*MCE2005* – a new version of a program for fast interactive visualization of electron and similar density maps optimized for small molecules". *Journal of Applied Crystallography* 40 (3) 2007, pp. 600–601. DOI: 10.1107/S0021889807018894.

[577] C. B. Hübschle and P. Luger. "*MolIso* – a program for colour-mapped iso-surfaces". *Journal of Applied Crystallography* 39 (6) 2006, pp. 901–904. DOI: 10.1107/S0021889806041859.

[578] Michael. N. Burnett and Carrol K. Johnson. *ORTEP*. Online. URL: http://www.ornl.gov/sci/ortep/ortep.html.

[579] M. Nardelli. "*PARST*95 – an update to *PARST*: a system of Fortran routines for calculating molecular structure parameters from the results of crystal structure analyses". *Journal of Applied Crystallography* 28 (5) 1995, p. 659. DOI: 10.1107/S0021889895007138.

[580] *SAS-OMEGA*. Online. URL: ftp://ftp.lmcp.jussieu.fr/pub/sincris/software/structure/.

[581] Maria Cristina Burla et al. "*SIR2011*: a new package for crystal structure determination and refinement". *Journal of Applied Crystallography* 45 (2) 2012, pp. 357–361. DOI: 10.1107/S0021889812001124.

[582] J. C. Matthewman, P. Thompson, and P. J. Brown. "The Cambridge Crystallography Subroutine Library". *Journal of Applied Crystallography* 15 (2) 1982, pp. 167–173. DOI: 10.1107/S0021889882011728.

[583] Louis J. Farrugia. "*WinGX* suite for small-molecule single-crystal crystallography". *Journal of Applied Crystallography* 32 (4) 1999, pp. 837–838. DOI: 10.1107/S0021889899006020.

[584] A. Soyer. "*LMCTEP*: software for crystal-structure representation". *Journal of Applied Crystallography* 26 (3) 1993, p. 495. DOI: 10.1107/S0021889892010343.

[585] S.R. Hall, D.J. du Boulay, and R. Olthof-Hazekamp. *Xtal*. Online. 2000. URL: http://xtal.sourceforge.net/.

[586] H. D. Flack. "*CAMEL JOCKEY*, an absorption correction program". *Journal of Applied Crystallography* 8 (5) 1975, pp. 520–521. DOI: 10.1107/S0021889875011168.

[587] E. Egert and G. M. Sheldrick. "Search for a fragment of known geometry by integrated Patterson and direct methods". *Acta Crystallographica Section A* 41 (3) 1985, pp. 262–268. DOI: 10.1107/S0108767385000551.

[588] P.T. Beurskens et al. *dirdif*. Online. 2008. URL: http://www.xtal.science.ru.nl/documents/software/dirdif.html.

[589] Th. Proffen and R. B. Neder. "*DISCUS*, a program for diffuse scattering and defect structure simulations – update". *Journal of Applied Crystallography* 32 (4) 1999, pp. 838–839. DOI: 10.1107/S0021889899004860.

[590] M. Prencipe. "*EQUIV*: a program for the analysis of equivalent reflections from single-crystal data collection". *Journal of Applied Crystallography* 31 (1) 1998, p. 109. DOI: 10.1107/S0021889897007085.

[591] *RASMOL*. Online. URL: http://rasmol.org/.

[592] E. Keller. *SCHAKAL*. Online. URL: http://www.krist.uni-freiburg.de/ki/Mitarbeiter/Keller/download.html.

[593] J.M.M. Smits and R. de Gelder. *SYSTER*. Online. 2000. URL: http://www.xtal.science.ru.nl/documents/software/syster.html.

[594] J. D. Westbrook, S.-H. Hsieh, and P. M. D. Fitzgerald. "CIF Applications. VI. *CIFLIB*: an application program interface to CIF dictionaries and data files". *Journal of Applied Crystallography* 30 (1) 1997, pp. 79–83. DOI: 10.1107/S0021889896008643.

[595] Frank H. Allen et al. "CIF applications. XV. *enCIFer*: a program for viewing, editing and visualizing CIFs". *Journal of Applied Crystallography* 37 (2) 2004, pp. 335–338. DOI: 10.1107/S0021889804003528.

[596] Brian H. Toby. "*EXPGUI*, a graphical user interface for *GSAS*". *Journal of Applied Crystallography* 34 (2) 2001, pp. 210–213. DOI: 10.1107/S0021889801002242.

[597] Gordon Barr et al. "*dSNAP*: a computer program to cluster and classify Cambridge Structural Database searches". *Journal of Applied Crystallography* 38 (5) 2005, pp. 833–841. DOI: 10.1107/S0021889805021308.