

DEVELOPMENT OF THE VARIABLE-CELL POWDER DIFFERENCE
(VC-PWDF) METHOD AND APPLICATIONS IN THE COMPARISON
AND DETERMINATION OF CRYSTAL STRUCTURES

by

R. Alex Mayo

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
August 2023

© Copyright by R. Alex Mayo, 2023

CONTENTS

List of Tables	vii
List of Figures	ix
List of Abbreviations and Symbols Used	xiii
Abstract	xvi
List of Tables	xvi
List of Figures	xvi
Acknowledgements	xvii
Chapter 1 Introduction	1
1.1 Background	1
1.2 Experimental identification and determination of molecular crystal poly- morphs	2
1.3 Crystal structure prediction	3
1.4 Thesis overview	6
Chapter 2 Crystallographic Theory	8
2.1 The unit cell	8
2.2 Lattice planes, <i>d</i> -spacing, and Miller indices	11
2.3 X-Ray diffraction in crystals	13
2.3.1 Calculated powder diffractograms, multiplicity, and systematic absences	14
2.3.2 Friedel's law, crystal classes and Laue classes	16

2.4	Powder X-ray Diffraction (PXRD)	17
2.5	Structure Determination from Powder Data (SDPD)	18
2.5.1	Confirming the solution: Rietveld refinement	19
2.5.2	PXRD indexing	21
2.5.3	Structure generation by direct methods	21
2.5.4	Structure generation by real space methods	22
Chapter 3	Quantitative Methods of Crystal Structure Comparison	25
3.1	Measurement of relative atomic positions	25
3.2	Powder diffractogram comparison	27
3.3	Challenges in powder diffraction-based comparison methods	28
3.4	The VC-PWDF protocol	30
3.5	VC-PWDF example	33
Chapter 4	Improved Quantitative Crystal-Structure Comparison using Powder Diffractograms via Anisotropic Volume Correction	39
4.1	Introduction	39
4.2	Dataset	42
4.3	Methods	44
4.3.1	Mercury CPS tool	44
4.3.2	Newly developed VC-PWDF code	45
4.3.3	Similarity value notation	48
4.4	Results	49
4.4.1	PXRD comparison	49
4.4.2	Analysis of additional VC structure matches	51
4.4.3	COMPACT comparison	52
4.5	Discussion	55
4.5.1	XXII polytypes	55

4.5.2	Extra matches not identified in BT6	57
4.5.3	Grey areas in structure comparison	59
4.5.4	Poor candidate geometries	62
4.6	Conclusions	64
Chapter 5	Development and Assessment of an Improved Powder-Diffraction- Based Method for Molecular Crystal Structure Similarity	67
5.1	Introduction	67
5.2	Methods	71
5.2.1	Mercury's CPS tool	71
5.2.2	Variable-cell powder difference (VC-PWDF)	72
5.3	Dataset	73
5.4	Results	75
5.4.1	Outcomes of structure comparisons	75
5.4.2	Dependence on tolerances and cutoffs	76
5.4.3	COMPACT tolerance behaviour	78
5.5	Discussion	82
5.5.1	COMPACT issues with highly branched molecules	82
5.5.2	VC-PWDF same / COMPACT different	84
5.5.3	COMPACT same / VC-PWDF different	88
5.6	Conclusions	90
Chapter 6	Quantitative Matching of Crystal Structures to Experimental Pow- der Diffractograms	94
6.1	Introduction	94
6.2	Results and discussion	98
6.3	Conclusion	107

Chapter 7	Conclusions and Future Work	109
7.1	Conclusions	109
7.2	Future work	111
Appendix A	Supplementary Information for Chapter 4	113
A.1	Target structures	114
A.2	Dataset	114
A.2.1	Lists removed from all BT6 submissions	114
A.2.2	Data processing	115
A.3	Cell transformation matrices	117
A.4	RMSD drift from BT6 results	119
A.5	Dependence on COMPACK options	120
A.6	Example output tables	121
A.7	Effect of VC-PWDF tolerance	122
A.8	Effect of RMSD tolerance	123
A.9	Effect of volume correction on RMSD(1)	123
A.10	Correlations between RMSD(20) and powder difference values	124
Appendix B	Supplementary Information for Chapter 5	125
B.1	Dataset	126
B.1.1	Edits to the dataset	126
B.1.2	Pruning of the dataset	126
B.2	Comparisons using CSD-housed vs. cif structures	131
B.3	Example of the dependence of RMSD(N) on tolerance	133

Appendix C	Supplementary Information for Chapter 6	134
C.1	Experimental	135
C.1.1	Powder X-ray diffraction	135
C.2	Data	135
C.2.1	Target structures from the CSD	135
C.2.2	CSP landscapes	135
C.3	Computational methods	136
C.3.1	VC-(x)PWDF	136
C.3.2	autoFIDEL	137
C.4	Results	138
C.4.1	Rietveld refinement	149
Bibliography		151

LIST OF TABLES

2.1	Summary of crystal space symmetry elements.	9
2.2	Conventional unit-cell shapes along with their classification of crystal and lattice systems, and number of associated crystal classes and space groups.	9
2.3	The different centring types and the coordinates of the symmetry-unique lattice points.	10
2.4	Systematic absences that result from unit cell centring types.	16
2.5	Relation of the crystal systems, observable Laue classes from PXRD, SC-XRD, and all crystal classes.	17
3.1	Maximum changes in molecular geometries after distortion of selected trial unit cells.	38
4.1	Selected RMSD and powder difference measures for four borderline cases in which the molecules display notable conformational differences relative to the target, or a notable positional shift is observed.	59
4.2	Similarity comparisons for structure XXII-G09-L1-E02, before and after volume correction, and after constant-volume geometry relaxation with rigid molecules.	63
5.1	Confusion matrices for the outcomes of 44,939 structure comparisons conducted with COMPACK and either CPS-PWDF or VC-PWDF using literature tolerance/cutoff values.	76
5.2	Confusion matrices for the outcomes of 44,939 structure comparisons conducted with COMPACK and either CPS-PWDF or VC-PWDF using optimal tolerance/cutoff values.	78
5.3	Number of structure comparisons with specified change in the number of molecule matches predicted by COMPACK, as a function of tolerances.	80
5.4	Some examples of molecules (and associated structure refcode families) that are difficult to compare using COMPACK.	83
6.1	VC-xPWDF scores from comparison of the collected powder diffractograms with the CSD structures.	100

A.1	CCDC identifiers for the BT6 target structures.	114
A.2	Lists of BT6 submissions removed prior to analysis.	114
A.3	Transformation matrices applied to six structures identified as matches in BT6 in order to apply the anisotropic volume correction.	119
A.4	COMPACT results for structures submitted for compound XXIII that had a 180° rotation of the carboxylic acid group relative to the target.	120
A.5	Example <code>vc-pwdf</code> output of structures that pass the unit-cell dimension criteria, when given a 10% deviation allowance from the reference structure.	121
A.6	Example <code>vc-pwdf</code> output.	121
B.1	Refcode changes made since the study done by Sacchi <i>et al.</i>	126
B.2	List of 12 duplicate comparisons in the previous dataset.	126
B.3	List of 30 comparisons eliminated due to the crystal structures involving different molecular species.	127
B.4	List of the 116 structures identified as disordered by ConQuest. . .	127
B.5	List of 78 structures identified by <code>checkcif</code> as having <code>Alert level A</code> flagged voids.	129
B.6	List of 8 structures identified as having missing non-H atoms. . . .	129
B.7	List of 146 refcode families removed from the data set due to excessively long run times.	131
C.1	Summary of the result of duplicate screening on the structure-energy landscapes obtained from the CPOSS database.	137
C.2	Summary of indexed unit cell dimensions from the experimental powder diffractograms collected and comparison to the unit cell dimensions of the matching polymorph in the CSD.	138
C.3	Summary of R_{wp} and χ^2 values from Rietveld refinement of CSD structures with the collected powder diffractograms.	149
C.4	Summary of R_{wp} and χ^2 values from Rietveld refinement of uracil structures with the collected powder diffractograms.	150
C.5	Summary of R_{wp} and χ^2 values from Rietveld refinement of caffeine structures with the collected powder diffractograms.	150
C.6	Summary of R_{wp} and χ^2 values from Rietveld refinement of urea structures with the collected powder diffractograms.	150

LIST OF FIGURES

1.1	Hypothetical polymorphic landscape.	2
1.2	Comparison of the structure-energy landscape of 5-fluorouracil from force field and DFT calculations.	4
1.3	Overlay of simulated powder diffractograms of the crystal structure of $C_{10}H_5NO_2S_2$ collected at ambient conditions, and optimized by DFT.	6
2.1	Unit-cell dimensions of the 7 crystal systems.	10
2.2	Diagrams of the 14 Bravais lattices.	11
2.3	Plot of the atomic scattering factor as a function of the scattering angle.	13
2.4	Schematic aiding in the description of Bragg's law by showing the geometry of radiation diffraction by crystal planes.	14
2.5	Plots that highlight the poor guiding ability of the χ^2 figure of merit in biased structure generation algorithms.	23
3.1	Diagram showing how COMPACK measures interatomic distances and angles.	26
3.2	Flowchart of the steps undertaken by the VC-PWDF protocol when comparing two crystal structures.	32
3.3	Comparable views of the two example crystal structures	33
3.4	Sample critic2 output showing which of the two structures being compared is chosen as the reference structure.	33
3.5	Comparable views of the Niggli cells of the two example crystal structures.	34
3.6	Sample critic2 output showing the Niggli-cell lattice vectors and dimensions of the two structures being compared.	34
3.7	Sample critic2 output showing possible lattice vectors of the candidate structure, along with their assignment to the Niggli-cell vectors of the reference structure.	35
3.8	View of the (100) plane of the Niggli cell of the candidate crystal structure.	35

3.9	Sample critic2 output showing the trial unit cells of the candidate structure.	36
3.10	Overlays of the reference crystal structure with different distorted trial unit cells of the candidate structure.	37
3.11	Overlays of the simulated powder diffractograms of the reference crystal structure with different distorted trial unit cells of the candidate structure.	37
3.12	Molecular overlays and RMSD(1) values for the distorted trial unit cells.	38
4.1	Schematic of the five target compounds in BT6.	43
4.2	Comparison of the unit-cell vectors for a BT6 target structure with a generated structure that requires the application of a transformation matrix.	48
4.3	Histograms showing the distribution of powder difference values obtained by various methods.	50
4.4	Classification of 113 relevant structures in the VC-POWDIF histogram.	52
4.5	Histograms highlighting the effect of the developed volume correction on RMSD(20) values.	53
4.6	Plot of VC-RMSD(20) vs. VC-PWDF values.	54
4.7	Overlay of the XXII target crystal structure with a generated polytype structure.	56
4.8	Histograms of the CPS-PWDF and VC-PWDF values for the 13 matches and 10 polytype structures identified for target XXII.	56
4.9	Overlays of the two missed match crystal structures with their respective target crystal structures.	57
4.10	Two examples of structures with notable conformational differences with their target, but are successfully overlaid in a cluster of 20 molecules with COMPACK.	60
4.11	Overlay of target XXIII form C with a generated structure identified as a conformational phase.	60
4.12	Overlay of the volume-corrected XXV-G15-L1-E24 and XXV target structures, with H-bonds highlighted.	61
4.13	COMPACK overlay, in the <i>ab</i> -plane, of the volume-corrected XXII-G09-L1-E02 structure with the target.	63

5.1	Heat maps representing the percentage of comparisons for which COMPACK and CPS-PWDF or VC-PWDF disagree on the outcome, as a function of the PWDF cutoff and COMPACK tolerances used.	77
5.2	Heat maps of the percentage of comparisons that are considered the same by COMPACK and different by VC-PWDF, and different by COMPACK and the same by VC-PWDF.	79
5.3	Δ RMSD(<i>N</i>) values from changes in COMPACK tolerances.	81
5.4	Best overlay generated by COMPACK for a hypothetical molecule consisting of four tri-tert-butylsilane substituents bonded to a central silicon atom by ethyne linkers.	84
5.5	COMPACK overlay of ZEDCUG and ZEDCUG01.	85
5.6	COMPACK overlay of ZITZUX and ZITZUX01.	87
5.7	COMPACK overlay of JIYKAD and JIYKAD01.	87
5.8	COMPACK overlay of UHIKUR and UHIKUR01.	88
5.9	COMPACK overlay of SILVAL and SILVAL02.	89
5.10	COMPACK overlay of CBMZPN03 and CBMZPN11.	90
6.1	Compounds studied and accompanying CSD refcode family.	97
6.2	Plots showing the computed VC-xPWDF scores resulting from comparison of each input crystal structure to the experimental powder diffractogram collected for that compound.	102
6.3	Overlay of the experimentally collected powder diffractogram and the simulated powder diffractogram for DMANTL07 after the VC-xPWDF protocol.	103
6.4	(VC-x)PWDF scores and powder diffractogram overlays for 1,4-dicyanobenzene computed using VC-xPWDF and autoFIDEL.	106
A.1	Comparison of RMSD values reported in BT6 with those obtained in this work using the current version of Mercury.	119
A.2	Histograms of powder difference values for structures that pass step (3) of our computational algorithm, with different volume and cell-length tolerances selected.	122
A.3	RMSD(20) values (before and after volume correction) as a function of the tolerance required to obtain a 20/20 molecule match with COMPACK.	123

A.4	RMSD(1) values for molecules before and after anisotropic volume correction.	123
A.5	Correlations between various RMSD(20) and powder difference values for the 113-structure dataset.	124
B.1	Structures of LUYKOF and LUYKOF01	128
B.2	Comparison of unit cells of PEZMEM10 and PEZMEM.	130
B.3	Molecular structures for ACLLEU and ACLLEU01.	132
B.4	COMPACK overlays of SUCROS27 and SUCROS33.	133
C.1	Experimental powder diffractograms collecting with the 3hr scan conditions.	139
C.2	Experimental powder diffractograms collecting with the 2min scan conditions.	140
C.3	Overlays of the experimentally collected powder diffractograms with the simulated powder diffractogram of the matching polymorph from the CSD after running the VC-xPWDF protocol. . . .	141
C.4	Overlay of simulated powder diffractograms of urea crystal structures after the VC-xPWDF protocol with the experimental powder diffractogram.	142
C.5	Overlay of simulated powder diffractograms of caffeine crystal structures after the VC-xPWDF protocol with the experimental powder diffractogram.	143
C.6	Overlay of simulated powder diffractograms of uracil crystal structures after the VC-xPWDF protocol with the experimental powder diffractogram.	144
C.7	Images of the packing of two <i>in silico</i> generated crystal structures of uracil, and a COMPACK overlay of the two structures.	145
C.8	Plots showing the VC-xPWDF value from comparison of each input crystal structure to the experimental powder diffractogram collected from a 2 minute scan for that compound.	146
C.9	Plots showing the POWDIFF value after optimization with the autoFIDEL code between the crystal structure and the experimental powder diffractogram collected for that compound.	147
C.10	Overlay of simulated powder diffractograms from crystal structures after optimization with autoFIDEL with the experimental powder diffractogram.	148

LIST OF ABBREVIATIONS AND SYMBOLS USED

Abbreviation	Description
1D	One-Dimensional
2D	Two-Dimensional
3D	Three-Dimensional
API	Active Pharmaceutical Ingredient
BGMN	J. Bergmann's automated Rietveld refinement protocol
BT6	The CCDC's Blind Test 6
CCDC	Cambridge Crystallographic Data Centre
COMPACK	COMparison of PACKing method
CONV	CONstant Volume relaxation
CPS	The CCDC's Crystal Packing Similarity tool
CSD	The CCDC's Crystal Structure Database
CSP	Crystal Structure Prediction
C-CSP	Constrained Crystal Structure Prediction
CPOSS	Control and Prediction of the Organic Solid State
DFT	Density-Functional Theory
DICVOL	Successive dichotomy indexing method
D3	Grimme's dispersion correction
DMACRYS	Distributed Multipole Analysis for CRYStals force field
FF	Force Field
FIDEL	FIIt with DEviating Lattice parameters method
FoM	Figure of Merit
HF-3c	Minimal-basis Hartree-Fock theory with 3 empirical corrections
ITO	A zone-indexing method by J. Visser (indexing method)
KOHL	F. Kohlbeck's heuristic search (indexing method)
MMFF94	Merk Molecular Force Field 1994
PBE	Perdew-Burke-Ernzerhof density functional
POWDIFF	Raw POWder pattern DIFFerence

Abbreviation	Description
PWDF	Dissimilarity measure using the weighted cross-correlation function
PXRD	Powder X-Ray Diffraction
RMSD	Root Mean Squared Deviation
SC-XRD	Single Crystal X-Ray Diffraction
SDPD	Structure Determination from Powder Data
TAUP	D. Taupin's index-permutation search (indexing method)
TPSS	Tao-Perdew-Staroverov-Scuseria density functional
TREOR	P.E. Werner's heuristic search (indexing method)
VC-PWDF	Variable Cell PoWder DiFference (method, score)
VC-xPWDF	Variable Cell experimental PoWder DiFference (method, score)
XDM	eXchange-hole Dipole Moment dispersion model

Symbol	Description
A, B, C	Crystal structures
$A(\theta)$	Absorption factor as a function of diffraction angle
a, b, c	Unit-cell crystallographic axes lengths
b_i	Length of the weighted triangle
c	Cutoff value
c_{fg}	Cross-correlation function of functions f and g
D_{fg}	Difference value for functions f and g
$d(A, B)$	Similarity between two crystal structures, A and B
d_{hkl}	Distance between crystallographic planes of the same Miller indices (hkl)
\mathbf{F}_{hkl}	Structure factor corresponding to lattice plane (hkl)
F_N	Smith-Snyder figure of merit
f_i	Scattering factor for atom i
f, g	functions
(hkl)	Miller indices defining a lattice plane
$[hkl]$	Miller indices defining a crystallographic direction
i, j	Numbering index
I_{hkl}	Peak intensity of the radiation diffracted by crystal planes (hkl)
$K\alpha$	X-ray emission from electron transitions from the $2p$ orbitals to the $1s$ orbital
$K\alpha_1$	X-ray emission from the electron transition from the $2p_{j_s=1/2}$ orbital to the $1s$ orbital

Symbol	Description
$LP(\theta)$	Lorentz-Polarization factor
M_{20}	de Wolff 20-peak figure of merit
M_{hkl}	Reflection multiplicity of the (hkl) plane
M_T	Value being minimized during Rietveld refinement
M, N	Numbers of molecules
n, p	Integer numbers
R_{wp}	Weighted profile residuals from point-wise comparison of powder diffractograms
R -factor	Measure of agreement between modeled and collected structure factor amplitudes
r	Interatomic distance
S_{fg}	Similarity value for functions f and g
s	Scaling factor
$T(\theta, hkl)$	Texture factor as a function of diffraction angle and (hkl) preferred orientation
V	Volume
$w(\delta)$	Weighting function for offset δ
w_i	Weighting parameter in Rietveld refinement
x, y, z	Cartesian directions
y_i^{obs}	Observed intensity of experimental data point i
y_i^{calc}	Calculated intensity of data point i
α, β, γ	Unit-cell angles between axes
Δ	Additional distance traveled by incident X-ray
δ	Offset shift between two functions
θ	Angle
2θ	Bragg's angle
λ	Electromagnetic radiation wavelength
χ^2	Figure of merit from Rietveld refinement

ABSTRACT

The identification and classification of crystal structures is fundamental in materials science, as the crystal structure is an inherent factor of what gives solid materials their properties (conductivity, magnetism, hardness, solubility, etc.). Being able to identify the same crystallographic form from unique origins (e.g. different temperatures, pressures, or *in silico*-generated) is a complex challenge. In particular, the use of simulated powder diffractograms for this purpose has not seen general success due to the intimate relationship between the lattice dimensions and peak positions, which are strongly affected by experimental conditions. Herein is presented the development and application of the VC-PWDF (Variable-Cell PoWder DiFference) method to resolve this scientific problem. This new approach of comparing crystal structures using their powder diffractograms involves an automated series of steps that identifies the lattice distortion necessary to align the two crystal structures being compared, provided one exists (i.e. provided they are the same form). The quantitative value (VC-PWDF score) yielded by the protocol provides a measure of similarity more accurate than other available methods of structure comparison based on powder diffractograms. For a set of nearly 45,000 structure pairs in the Cambridge Structure Database (CSD), the VC-PWDF method is shown to be as successful as the COMPACK method, which compares atomic positions, in distinguishing the same form under disparate conditions from a different polymorph structure. When comparing known polymorphs to *in silico*-generated structures from Crystal Structure Prediction (CSP) studies, the VC-PWDF method is shown to be a valuable complementary method to COMPACK, which is prone to false negatives that VC-PWDF readily identifies as true positives, determining two missed matches from the 6th CSP blind test. Finally, the ability of the VC-PWDF approach to match an experimental powder diffractogram collected on a regular laboratory diffractometer to a crystal structure (from the CSD or CSP) *via* its simulated powder diffractogram is demonstrated (Variable-Cell eXperimental PoWder DiFference, VC-xPWDF), outlining a path for structure determination from powder data. The VC-(x)PWDF methods are anticipated to become commonplace tools for crystal structure comparison and determination in academia and industry alike.

ACKNOWLEDGEMENTS

First, I would like to thank my supervisor, Prof. Erin Johnson, for taking me on as a PhD student and giving me the opportunity to explore my research ideas with freedom and encouragement. I'm so grateful for the opportunities you gave me to attend numerous international conferences to present my work and interact with prominent members of some small and specialized research communities. Although I find myself having performed work in the area of "theoretical/applied crystallography", our many discussions of theoretical and quantum chemistry and sharing of your deep knowledge of electronic structure theory with me has probably also made me an acceptable computational chemist.

Next, I need to thank the Johnson group collaborator, scripting wizard, and crystallography consultant, (and alumnus!) Prof. Alberto Otero de la Roza. If his name isn't in the author list in the papers presented here-in, it's in the acknowledgments because I spent a lot of time discussing (and occasionally arguing) about various crystallographic topics and project ideas with him. The critic2 program that Alberto created and maintains was a fundamental tool for the research presented in this thesis and its importance in making this work possible cannot be overstated.

I would also like to thank my advisory committee members—Prof. Josef Zwanziger, Prof. Peng Zhang and Prof. Stephen Bearne—for their contributions to this work with time they spent providing reviews and edits on written documents, and comments and suggestions made in meetings. In particular, thanks to Prof. Zwanziger for challenging me to continuously improve my depth of knowledge, and my communication of the more complex aspects of my work to non-experts.

I am grateful for the Department of Chemistry at Dalhousie and all of its members; in administration helping to keep things running smoothly and on track, faculty and staff members who were accommodating, and graduate student colleagues. Though I did not spend much time on-campus as a result of the COVID-19 pandemic, my experience with the department and university was extremely positive. A special thank you to all the Johnson group members for their support and contributions, always making time for practice talks and providing me with advice and suggestions whenever I sought it over the group Discord channel.

I was fortunate enough to be awarded a few smaller scholarships during my studies. These helped to provide me with both encouragement and funds and I would like to extend my sincere thanks to the Walter C. Sumner Foundation, Gerry Dauphinee scholarship trust, the ICDD Ludo Frevel award funders and selection committee, the CNCC Larry Calvert travel award, and the Dalhousie Chemistry Department Analytical Chemistry travel award.

I'd like to thank my family for their love and support, even if they don't really understand why I quit my job to go back to school and kind of just think that I like being a career student.

And finally, I must thank my partner, Kate, who has not only been by my side to push me along, keep me focused, and bring me late night snacks while I obsessively tried to fix this or that problem, but also been a collaborator! Chapter 6 is our first paper together and it was very special to publish a scientific article with both of our names on it. I am so excited about her career as a professor at Carleton University and all the scientific conversations we'll continue to have, whether they end up published or not.

CHAPTER 1

INTRODUCTION

1.1 Background

Molecular solids play a vital role in the anthropogenic age and promise a future of rational material design due to their ease of processing and tunability. The development of technological materials (commonly associated with network-inorganics) such as electronics, semiconductors, memory storage, screen displays, and photovoltaics is now being undertaken by novel molecule-based materials, which have the potential to eventually outperform those currently used.¹⁻⁶ Pharmaceuticals in modern medicine are primarily small-molecule drugs and their solid-form properties are critical for both their process engineering and performance, as well as intellectual property protection.^{7,8} The fundamental utility of a molecular material depends on its solid-state properties and, therefore, its structure, as this is what dictates its properties. Whether a new optoelectric device, explosive, dye, nutraceutical, or pesticide, the performance and utility of the material used is dependent on its solid-state structure.⁹⁻¹⁷

An ensemble of molecules that forms a 3-dimensional (3D) repeating array (i.e. possesses translational symmetry) upon solidifying is called a crystal. Often, there are different ways in which the molecules are capable of packing together to form a 3D array, yielding unique crystal polymorphs. The relative stability of molecular crystal polymorphs is assessed at a particular temperature and pressure and may change with a change in temperature (enantiotropic), or remain the same (monotropic). Thermodynamically, the polymorph with the lowest chemical free energy is the most stable form under the conditions considered and, barring kinetic effects, the metastable (less stable) form will eventually transform into the thermodynamically stable form. The identification of polymorphs and elucidation

of their relationships is critical in materials and pharmaceutical development due to their structure-properties relationship; different polymorphs will have different properties.

1.2 Experimental identification and determination of molecular crystal polymorphs

Common techniques for experimental differentiation of crystal polymorphs include thermal analysis and powder X-ray diffraction (PXRD). These two methods can be used in combination with an exploration of crystallization conditions to discover and identify polymorphs, and determine the relationships between the discovered polymorphs of the molecule being studied (a hypothetical example is shown in Figure 1.1). Polymorph screening is a common practice in academia and industry, both for research and discovery, and for quality control.⁸

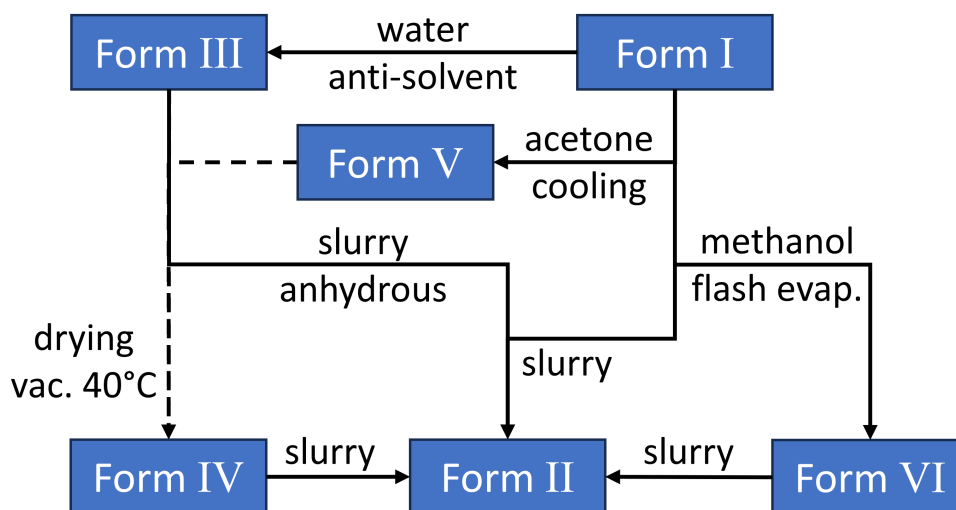


Figure 1.1: Hypothetical polymorphic landscape from experimental polymorph screening including conditions from which the different forms are generated or inter-converted.

While a polymorphic landscape can be generated through polymorph screening with only PXRD and thermal analysis data (and sometimes a few additional experiments and analyses), an elucidation of the structure-properties relationship of the different polymorphs cannot be accomplished without the crystal structure (the 3D representation of the crystal lattice and atomic positions). A crystal structure is commonly obtained through single-crystal X-ray diffraction (SC-XRD) studies. This method of determination requires a single crystal of sufficient quality to be grown in the lab for analysis on a single-crystal

X-ray diffractometer. The growth of a quality single crystal is often much more difficult in practice than on paper, and single-crystal X-ray diffractometers and crystallographers are less accessible than powder diffractometers. A powder diffractogram contains a useful amount of structural information about the corresponding crystalline material, but this information is compressed since it is a 1D projection of the 3D diffraction pattern that would be generated by SC-XRD. Therefore, crystal structure determination from powder data (SDPD) is much less common than determination from single-crystal diffraction data and, for molecular crystals, is a considerably more challenging endeavour.^{18–21}

Once a crystal structure has been determined for a polymorph, it is desirable to use this structure as a reference for comparison. Comparison to other solved crystal structures can be accomplished by various methods,^{22–28} including atomic position-based,²² or simulated PXRD-based methods.²³ The obvious benefit of a PXRD-based comparison method is that it can, in principle, be used for comparison with experimentally collected powder diffractograms as well as other crystal structures. The simulation of a powder diffraction pattern from a crystal structure is trivial using modern computers and free software. However, the disparate conditions under which SC-XRD is commonly performed (ca. 100 K) and PXRD is routinely done (ambient conditions, 298 K) causes difficulties in the quantitative comparison of powder diffractograms as the peak positions shift markedly with minor changes in the lattice dimensions. A thorough discussion of this phenomenon and developments intended to account for it is presented in Chapter 3. Thus, in practice, quantitative comparisons are performed using atomic position-based methods as they are intrinsically less sensitive to minor changes in lattice parameters, but require the structure solution from the diffraction data.

1.3 Crystal structure prediction

Crystal structure prediction (CSP) has developed considerably over the past 25 years to the point where various commercial companies offer CSP services, primarily to pharmaceutical developers. The interest in CSP is multifold; however, current commercial application is generally in assessing polymorphic risk for new active pharmaceutical ingredients (APIs). Experimental polymorph screening can only explore a finite number of crystallization conditions, so some polymorphs of the studied molecule may be missed. These are unlikely to pose any serious issues if they are metastable forms, but a missed polymorph

that turns out to be a thermodynamically stable form around ambient conditions can be disastrous.^{29,30} CSP has also been identified within academia as an aid in the development of porous materials³¹ and organic semiconductors,³² among other materials with desirable properties,³³ and its use in the development of any material with targeted properties is increasingly likely to be adopted as methods of CSP and property simulation continue to improve.

A CSP study begins with a structure generation step, which uses a computer program to create tens to hundreds of thousands of hypothetical crystal structures for the molecule of interest. These hypothetical crystal structures are optimized, often using classical-mechanics force fields, to yield a chemically reasonable structure and a corresponding energy for each crystallographic arrangement. The energy and density of these optimized hypothetical structures are then used to plot each crystal structure as a data point on a crystal energy landscape (energy on the y-axis and density on the x-axis), as illustrated in Figure 1.2. Low-energy structures towards the bottom of the landscape are considered more likely to be observed experimentally.

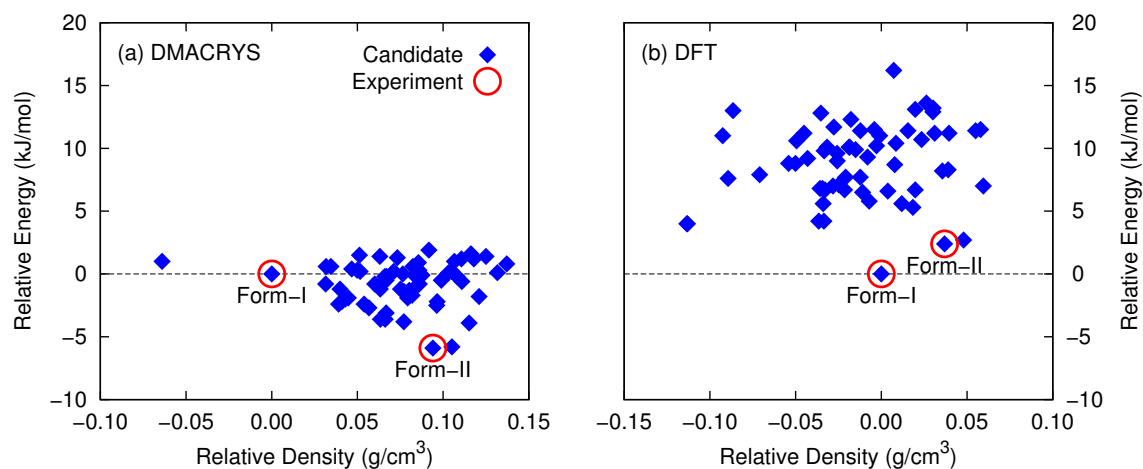


Figure 1.2: Two structure–energy landscapes of 5-fluorouracil produced using data from ref. 34. The landscape on the left was generated with a classical force field (DMACRYS), and that on the right was generated with periodic-boundary DFT using a plane-wave basis set.

The level of theory used to optimize the hypothetical crystal structure and calculate its corresponding energy is found to have a dramatic effect on the crystal energy landscape, with higher levels of theory (e.g. density-functional theory, DFT) able to better reproduce experimental results of polymorph energy ranking.^{34,35} Unfortunately, the use of DFT

requires orders of magnitude greater computational resources than classical mechanics models, so optimization and energy evaluation of many thousands of hypothetical (mostly unobservable) crystal structures would be an excessive waste of resources. Therefore, it is desirable to identify methods for reducing the number of structures that need to be re-optimized with a higher level of theory. The computer program that generates the hypothetical crystal structures may generate the same crystal structure in various, slightly different, descriptions that are effectively identical. A rapid, accurate, quantitative method of comparison is valuable in eliminating these duplicate structures so that the same calculation is not performed multiple times.

In order to assess the performance of a CSP method, benchmarking tests are done whereby a crystal structure-energy landscape is produced for a molecule that has undergone polymorph screening and crystal structure determination of the polymorphs. The best test cases are those where the researchers producing the CSP structure-energy landscapes do not have access to the known crystal structure(s) of the molecule until after their submission, i.e. the Cambridge Crystallographic Data Centre (CCDC) blind tests.³⁶⁻⁴¹ It is necessary, however, to ensure that one is able to identify the matching crystal structure to the target experimental crystal structure(s) of the molecule from the many thousands of hypothetical crystal structures generated by the CSP study.

The hypothetical crystal structures generated are static, without any consideration of the thermal vibrations that contribute to the averaged position of the atoms in an experimentally determined crystal structure. This results in quite notable distortions in the crystal structure, approximated to the comparison of a “0 K” crystal structure to the target crystal structure (likely determined at 100 K). Therefore, once again, atomic position-based methods of structure comparison have gained popularity in the field of CSP.

The requirement of two crystal structures for comparison restricts the potential of CSP to help with structure determination from powder data, whereby an experimental powder diffractogram (collected at ambient conditions, 298 K) could be matched to a hypothetical crystal structure generated by CSP using its simulated powder diffractogram. The difference in conditions between the hypothetical crystal structure from CSP and the PXRD data is even more extreme than comparison with a solved SC-XRD structure, adding to the challenge of even a visual assessment of similarity due to the magnitude of the (non-linear) disparity in peak positions (e.g. Figure 1.3).

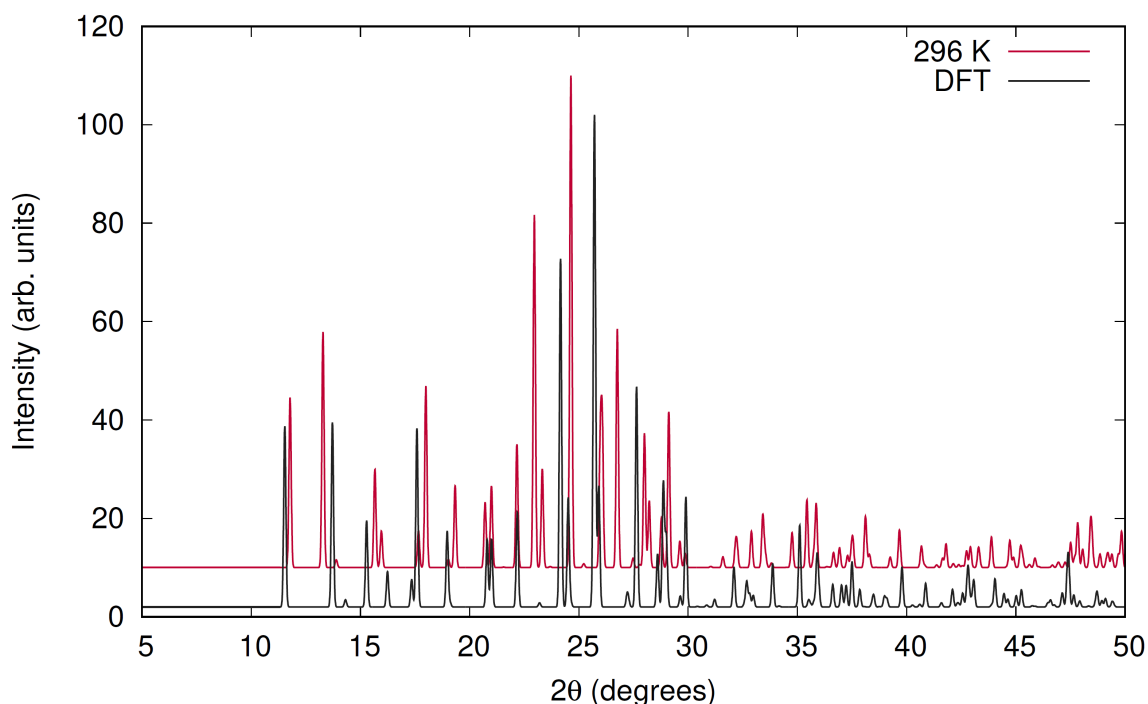


Figure 1.3: Overlay of the simulated powder diffractograms of the crystal structure of $C_{10}H_5NO_2S_2$ (Ref. 42) collected at 296 K (red) and after optimization by DFT (black) using the FHI-aims program,⁴³ (B86bPBE-XDM, light basis sets, and dense integration grids³⁵) to simulate a structure generated by CSP. The dramatic shifting in the peak positions make a quantitative determination of similarity useless, and visual comparison challenging.

1.4 Thesis overview

This thesis explores the importance of accurate quantitative crystal structure comparison, and the development and applications of a new method of comparison using powder diffractograms to determine whether the compared diffractograms represent the same crystal structure, or two distinct crystal structures. The relevance of this research is clear in the context of the preceding discussion of material properties of solids and polymorphism. A thorough background on crystallographic theory is presented in Chapter 2. Chapter 3 provides more details on the crystal structure comparison methods used in this work, and a detailed description of the newly developed variable-cell powder difference (VC-PWDF) method, including an example case walk-through.

Chapters 4, 5, and 6 are (reproduced with permission) published peer-reviewed articles outlining the research performed and developments made in accurate and efficient crystal structure comparison using the new VC-PWDF approach; their accompanying (curtailed)

supplementary information documents are provided in the Appendices. The application of the VC-PWDF method to identify matching target structures in CSP landscapes is highlighted, and its utility as a complementary method to atomic position-based comparison methods demonstrated (Chapter 4). The VC-PWDF method is shown to dramatically improve agreement between an atomic position-based method and a powder diffractogram-based method of crystal structure comparison in Chapter 5, wherein ca. 45,000 crystal structure pairs from the Crystal Structure Database (CSD) are compared, and advantages and disadvantages of the different methods are explored. Having unambiguously demonstrated the robustness of the VC-PWDF method in comparing two simulated powder diffractograms, Chapter 6 extends the methodology to compare simulated powder diffractograms with experimentally collected ones. The result is a method capable of matching an experimental powder diffractogram to a crystal structure, i.e. structure determination. Conclusions and future work for the VC-PWDF approach are summarized in Chapter 7.

CHAPTER 2

CRYSTALLOGRAPHIC THEORY

This chapter has been written with reference to *The International Tables for Crystallography*,^{44–50} *Basic Solid State Chemistry* by Anthony R. West,⁵¹ *Fundamentals of Powder Diffraction and Structural Characterization of Materials* by Vitalij Pecharsky and Peter Zavalij,⁵² and *Structure Determination from Powder Diffraction Data* by William David *et al.*⁵³ The bulk of the content is well established and can be found in the above-mentioned references. Any non-established ideas are referenced in the text.

2.1 The unit cell

In the context of chemistry, a crystalline solid is one composed of atoms and possessing translational symmetry in 3D (under the approximation of a perfect, infinite system). In addition to translational symmetry, a crystal may possess other space symmetry elements (Table 2.1). Using these symmetry elements, points (called *lattice points*) can be chosen within the crystal that are indistinguishable from one-another, forming a 3D array of points called a *crystal lattice*. By joining a single, chosen lattice point (i.e. the origin) to three other lattice points using non-colinear vectors, one can produce a *unit cell* that, through application of the symmetry operations that exist for the crystal, will generate the entire crystal from the finite parallelepiped that is defined by these three basis vectors.

The general dimensions of the possible parallelepipeds that may result from the choice of *lattice vectors* (connecting the origin lattice point with another lattice point) and their corresponding crystal system and lattice system are summarized in Table 2.2 and Figure 2.1. The criteria for choosing a *conventional unit cell*, defined by the International Union for Crystallography (IUCr), requires that it a) has a right-handed axial setting, b) has its edges

Table 2.1: The space symmetry elements that may be observed in crystals and a brief description of their operation. Hermann-Mauguin symbols are used in crystallography instead of the Schönflies notation used in spectroscopic/molecular point groups.

Type	Element(s) Symbol(s)	Description
Translation	(none)	-
Rotation	2, 3, 4, 6	rotation by $360^\circ/N$
Reflection	m	reflection
Rotoinversion	$\bar{1}, \bar{3}, \bar{4}, \bar{6}$	combination of rotation by $360^\circ/N$ and inversion through the central point on the rotation axis
Screw	$2_1; 3_1, 3_2; 4_1, 4_2, 4_3$ $6_1, 6_2, 6_3, 6_4, 6_5$	for N_x , combination of translation of x/N and rotation by $360^\circ/N$
Glide	a, b, c, n, d, e	combination of translation and reflection

along symmetry directions of the lattice, and c) is the smallest cell that meets the above criteria, while able to portray all symmetry elements of the crystal. Conventional unit cells that are generated in the *standard settings* aim to order the axes lengths from smallest to largest where possible, and keep all angles as close to 90° as possible or, in the case of monoclinic crystal systems, make β the unique angle.

Table 2.2: Conventional unit-cell shapes along with their classification of crystal and lattice systems, and number of associated crystal classes and space groups.

Crystal System	Lattice System	Parallelepiped Dimensions		Crystal Classes	Space Groups
		Lengths	Angles		
Triclinic	Triclinic	$a \neq b \neq c$	$\alpha \neq \beta \neq \gamma \neq 90^\circ$	2	2
Monoclinic	Monoclinic	$a \neq b \neq c$	$\alpha = \gamma = 90^\circ, \beta \neq 90^\circ$	3	13
Orthorhombic	Orthorhombic	$a \neq b \neq c$	$\alpha = \beta = \gamma = 90^\circ$	3	59
Tetragonal	Tetragonal	$a = b \neq c$	$\alpha = \beta = \gamma = 90^\circ$	7	68
Trigonal	Rhombohedral	$a = b = c$	$\alpha = \beta = \gamma \neq 90^\circ$	5	7
	Hexagonal	$a = b \neq c$	$\alpha = \beta = 90^\circ, \gamma = 120^\circ$		18
Hexagonal	Hexagonal	$a = b \neq c$	$\alpha = \beta = 90^\circ, \gamma = 120^\circ$	7	27
Cubic	Cubic	$a = b = c$	$\alpha = \beta = \gamma = 90^\circ$	5	36

After the assignment of the lattice vectors to the parallelepiped axes, the resulting unit cell may contain a varied number of lattice points. The unit-cell *centring types* provide the classification for the possible outcomes and are summarized in Table 2.3. From the combination of the seven lattice systems and 4 centring types, only 14 unique conventional unit-cell lattices are possible, called the 14 Bravais lattices shown in Figure 2.2. There is a finite number of mutually compatible combinations of the space symmetry elements with

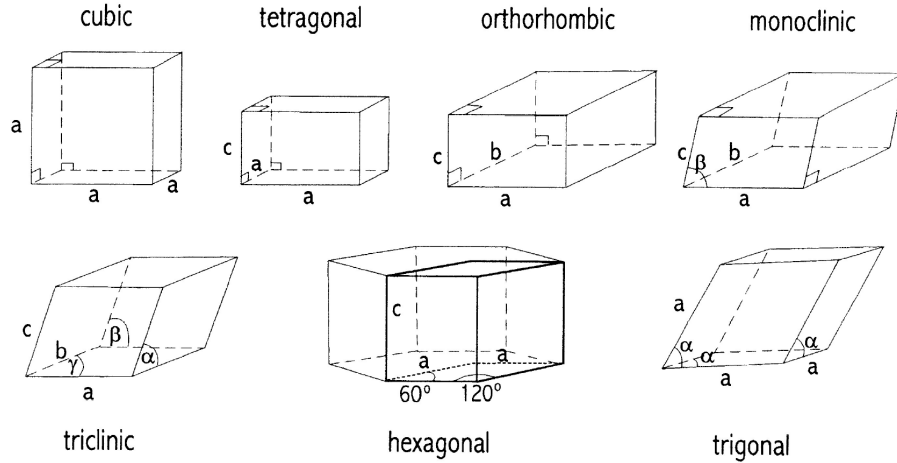


Figure 2.1: Unit-cell dimensions of the 8 crystal systems (reproduced with permission from ref. 51).

Table 2.3: The different centring types and the coordinates of the symmetry-unique lattice points.

Centering	Symbol(s)	# of lattice points	Coordinate(s) of lattice point(s)
Primitive	P	1	0,0,0
Base-centred	C	2	0, 0, 0 ; 1/2, 1/2, 0
	B		0, 0, 0 ; 1/2, 0, 1/2
	A		0, 0, 0 ; 0, 1/2, 1/2
Body-centred	I	2	0, 0, 0 ; 1/2, 1/2, 1/2
Face-centred	F	4	0, 0, 0 ; 1/2, 1/2, 0 ; 1/2, 0, 1/2 ; 0, 1/2, 1/2

the 14 Bravais lattices, leading to the 230 space groups that describe the unit-cell shape and symmetry of a crystal.

Primitive conventional unit cells are in their reduced form when they are also in their standard setting. Unit cells of the other centring types are converted to a reduced cell that is primitive through transformation matrices. A reduced cell is by definition primitive.

The Niggli reduced cell is that recommended and outlined in the International Tables for Crystallography, Volume A. The Niggli reduced cell is unique for a given crystal lattice, is always primitive, and comprised of the three smallest non-coplanar vectors in the standard setting. There are two “Types” of reduced cell based primarily on the possible resultant angles between the vectors. The Type I cells may be triclinic or rhombohedral, as a condition requires that all vector dot-products be greater than 0 (making the angle $< 90^\circ$). The Type II cells may describe any lattice system, as all vector dot products must be less than or equal to 0 (making the angle $\geq 90^\circ$). Additional conditions relevant to special

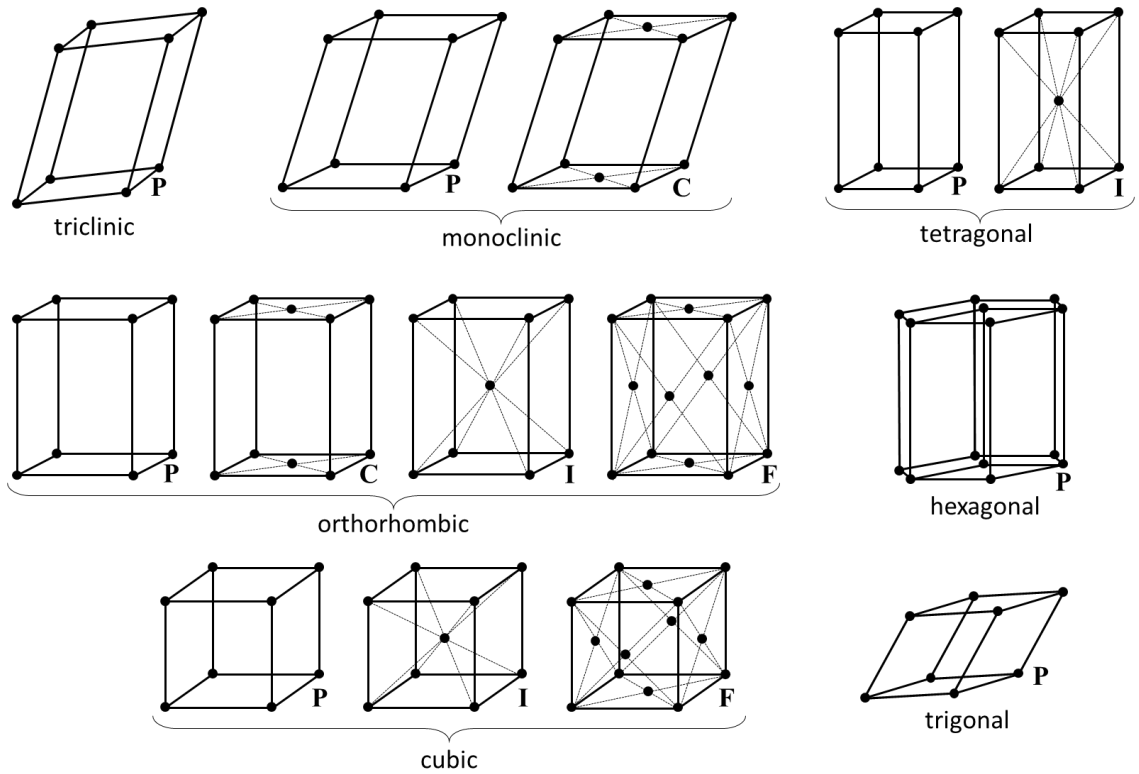


Figure 2.2: Diagrams of the 14 Bravais lattices.

cases of Type I and II cells can be found in the International Tables of Crystallography, Volume A.

2.2 Lattice planes, d -spacing, and Miller indices

Crystal lattice planes may be defined based on connection of the lattice points, which exist parallel to one another *ad infinitum* due to the treatment of a crystal as an infinite 3D system. The naming of these planes is done in a standard format, using Miller indices. An index variable is given to each unit-cell axis, with h corresponding to the a -axis, k to the b -axis, and l to the c -axis. First, one identifies the lattice plane passing through the origin of the unit cell that is symmetry equivalent to the plane of interest, then finds the intersection of this lattice plane with the unit-cell axes. The inverse is taken as the index value (i.e. intersection at $a/2$ would yield $h = 2$), and the indices are given as (hkl) .

These Miller indices are used to refer to a particular lattice plane; however, the planes exist regardless of the chosen unit cell basis. Therefore, the indices of a particular plane will change when the choice of lattice vectors that define the unit cell are changed. Nothing

about the crystal has changed—the distance between the planes (referred to as the d -spacing) is completely unaffected—but the indices used to name the plane have simply changed. This subtle point is important as confusion about which plane is being referred to may arise if one is not careful to consider the unit cell used to describe the crystal.

While the distance between planes is unchanged, the formula used to calculate the d -spacing is also dependent on the unit cell chosen, as it is necessary to reference to the cell axes lengths (a , b , c) and angles (α , β , γ). The general formula for determining the d -spacing is

$$\begin{aligned} \frac{1}{d^2} = & \frac{1}{V^2} [h^2 b^2 c^2 \sin^2 \alpha + k^2 a^2 c^2 \sin^2 \beta + l^2 a^2 b^2 \sin^2 \gamma \\ & + 2hkabc^2 (\cos \alpha \cos \beta - \cos \gamma) + 2kla^2 bc (\cos \beta \cos \gamma - \cos \alpha) \\ & + 2hlab^2 c (\cos \alpha \cos \gamma - \cos \beta)]. \end{aligned} \quad (2.1)$$

This simplifies readily for monoclinic (unique β angle),

$$\frac{1}{d^2} = \frac{1}{\sin^2 \beta} \left(\frac{h^2}{a^2} + \frac{k^2 \sin^2 \beta}{b^2} + \frac{l^2}{c^2} - \frac{2hl \cos \beta}{ac} \right), \quad (2.2)$$

hexagonal,

$$\frac{1}{d^2} = \frac{4}{3} \left(\frac{h^2 + hk + k^2}{a^2} \right) + \frac{l^2}{c^2}, \quad (2.3)$$

and orthogonal ($\alpha = \beta = \gamma = 90^\circ$) cells,

$$\frac{1}{d^2} = \frac{h^2}{a^2} + \frac{k^2}{b^2} + \frac{l^2}{c^2}. \quad (2.4)$$

It is worth noting that Miller index notation is also commonly used to describe directions in a crystal, which are given by indices in square brackets, $[hkl]$. The values are determined by drawing a line (that is parallel to the direction of interest) through the unit cell origin and a point in the unit cell with known fractional coordinates, then multiplying all indices by a common factor to yield the smallest integer values.

2.3 X-Ray diffraction in crystals

The atoms that occupy lattice planes, denoted using Miller indices, are responsible for the diffraction of electromagnetic radiation. Radiation in the X-ray region of the electromagnetic spectrum is most commonly used in diffraction studies of crystals as the wavelength is on the correct scale for atomic resolution, while being low enough in energy that it is non-destructive to the sample being analyzed (usually). The ability of an atom to diffract the radiation is described by its scattering factor, f , and is dependent on the radiation wavelength, incident angle, and its number of electrons. (Figure 2.3). Atomic scattering factors are compiled in the International Tables for Crystallography.

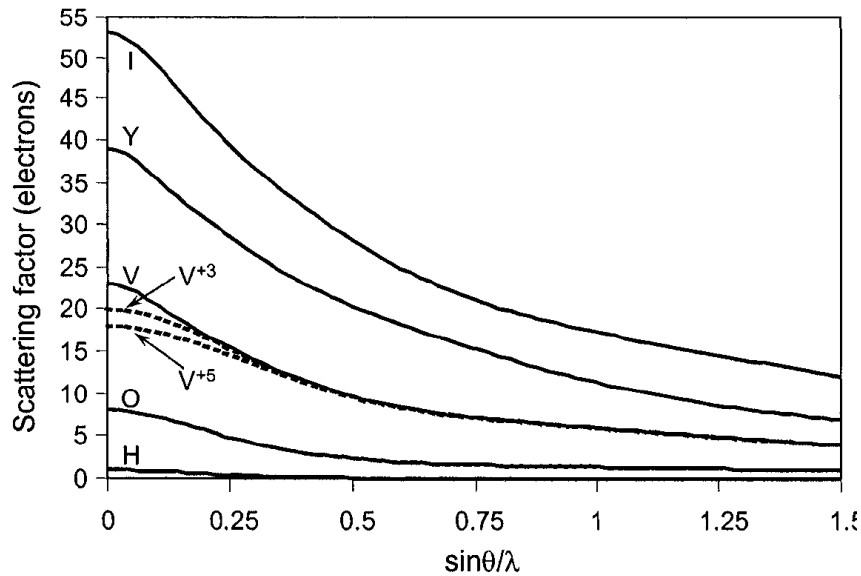


Figure 2.3: Plot of the atomic scattering factor (f) of iodine, yttrium, vanadium (neutral and oxidized states), oxygen and hydrogen as a function of the scattering angle (wavelength-normalized). The scattering factor is equal to the number of electrons in the element at a scattering angle of 0 (y-intercept). (reproduced with permission from ref. 52.)

The structure factor associated with a particular lattice plane, F_{hkl} , is a combination of the atomic scattering factors of the atoms that occupy that plane:

$$\mathbf{F}_{hkl} = \sum_j f_j e^{2\pi i(hx_j + ky_j + lz_j)} \quad (2.5)$$

Here, f_j is the scattering factor of atom j in the plane (hkl); h , k , and l are the Miller indices of the plane; and x_j , y_j , and z_j are the fractional coordinates of the atom in the unit cell.

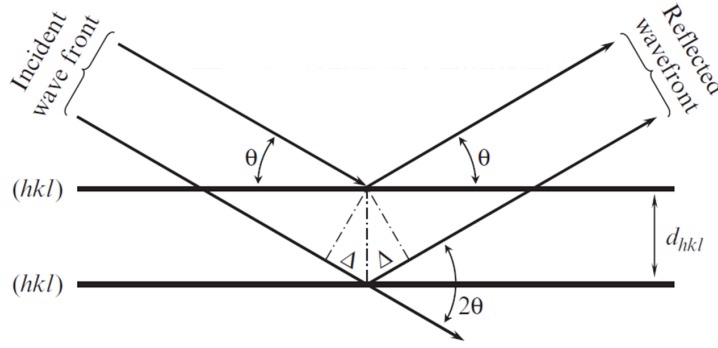


Figure 2.4: Schematic showing the geometry of radiation diffraction by crystal planes. When Δ is equal to an integer multiple of the wavelength of the X-ray radiation used, Bragg's law is satisfied and yields constructive interference. (Reproduced with permission from ref. 52).

The angle at which the radiation is diffracted to generate constructive interference naturally depends on the d -spacing, as shown in Figure 2.4. These angles are defined by Bragg's law:

$$n\lambda = 2d_{hkl} \sin \theta, \quad (2.6)$$

where n is an integer, λ is the wavelength of the electromagnetic (X-ray) radiation, d is the distance between crystallographic planes, and θ is the angle that the X-ray beam makes with the sample interface. The X-rays only reach the detector in appreciable intensity when Bragg's condition is satisfied, that is when the extra distance traveled by the transmitted X-ray is equal to an integer multiple of $2d_{hkl} \sin \theta$. Thus, the peaks observed in diffraction experiments are at angles of 2θ where this condition is satisfied, which is related to the d -spacing.

2.3.1 Calculated powder diffractograms, multiplicity, and systematic absences

The simulation of a powder diffractogram from a known crystal structure can be done using the following equation:

$$I_{hkl} = sLP(\theta)A(\theta)T(\theta, hkl)M_{hkl}|F_{hkl}|^2 \quad (2.7)$$

where s is a scaling factor, $LP(\theta)$ is the Lorentz-polarization function, which affects peak shape, A is the absorption factor, T is a factor accounting for preferred-orientation

effects (a.k.a. “texture”), M_{hkl} is the reflection multiplicity, and F_{hkl} is the structure factor. Absorption and preferred orientation are assumed to be absent when generating a theoretical powder diffractogram. The multiplicity and structure factors must be calculated based on the crystal structure.

In a crystal that belongs to the $P1$ space group (translational symmetry only), each lattice plane is unique. In a single-crystal diffraction experiment, the reflection from each plane is identified and the intensity integrated. Coincidence may result in overlapping peaks when this diffraction pattern is projected onto the 1D powder diffractogram, making (hkl) identification and intensity measurements difficult and of poorer quality. However, with crystals of higher-symmetry space groups, certain lattice planes become equivalent, and/or their reflection is annihilated due to the presence of a particular symmetry element or another lattice plane. The former is referred to as the *multiplicity* of a particular reflection, while the latter is referred to as a *systematic absence* of a particular reflection.

As an example of multiplicity, the (100) , (010) and (001) planes of a crystal belonging to the cubic crystal system will all be equivalent, and so the measurement of each of these reflections should yield the same intensity at the same Bragg’s angle; they are indistinguishable from one another. In powder diffraction studies, the peaks will overlap perfectly and the intensity of the reflection will relate to its multiplicity by being $3\times$ the calculated intensity of the diffraction from one of the planes.

The absence of the diffraction of a particular lattice plane may arise from one of two features: a non-primitive lattice centring type or space symmetry elements. An example of the effect of centring type is the body-centred cell, where the (100) reflection is absent and the (200) reflection will have a multiplicity of 2. The (100) plane is symmetry equivalent to the (200) plane and, thus, the interference will be destructive at the angle where the (100) plane would diffract due to Bragg’s law. Table 2.4 shows the general rules for the observed reflections for the conventional centring types.

As an example for a systematic absence due to the presence of a space symmetry element, consider the 2_1 screw axis. If the screw axis is in the $[100]$ direction, then only reflections with $(h00)$ will be observed, where $h = 2n$. Similarly, when the screw axis is along $[010]$, only $(0k0)$ reflections are observed, where $k = 2n$. These absences arise in an analogous way to the cell centring, where there is an equivalence in the magnitude of the diffraction of the lattice planes and yield complete destructive interference. Systematic absence

Table 2.4: The systematic absences that accompany the centring types (additional rules may apply given space symmetry elements present).

Centring	Symbol(s)	Rules for observed reflections
Primitive	P	None
Base-centred	C	$h + k = 2n$
	B	$h + l = 2n$
	A	$k + l = 2n$
Body-centred	I	$h + k + l = 2n$
Face-centred	F	h, k, l either all even or all odd

conditions exist for all space symmetry elements that include a translational component (screw axes and glide planes) and are direction-dependent (as they depend on definition by cell axes, which are user-defined). The reader is referred to the International Tables for Crystallography, Volume A⁴⁴ for the complete list of conditions.

2.3.2 Friedel's law, crystal classes and Laue classes

The point group of a crystal works on the same principles as molecular point groups (exclusion of space symmetry elements). There are 32 crystal point groups based on mutually compatible combinations of symmetry elements, which are called the *crystal classes*. Practically, the differentiation of all 32 crystal classes may not be possible, and only a sub-set, called the Laue groups or Laue classes can be identified in diffraction experiments due to the centrosymmetric nature of diffraction. In single-crystal diffraction experiments, all non-centrosymmetric crystal classes collapse into the 11 distinguishable centrosymmetric Laue classes when there is negligible adsorption (resonant scattering) of the diffracting radiation, and the reflection intensity from the hkl planes are equal in intensity to those of the $\bar{h}\bar{k}\bar{l}$ planes (ie. Friedel's law is obeyed).

The determination of the crystal class experimentally by comparing the intensity of the hkl planes with the $\bar{h}\bar{k}\bar{l}$ planes can be regularly done in single crystal diffraction when there is significant resonant scattering. Due to the overlap of the hkl and $\bar{h}\bar{k}\bar{l}$ reflections in PXRD and much lower accuracy in the intensity measurements, these reflections cannot be resolved in 1D powder diffractograms. In fact, the number of observable Laue classes decreases in PXRD compared to SC-XRD, with only 6 differentiable Laue classes. The peaks that would be unique in SC-XRD to differentiate these Laue classes overlap at the same Bragg's angle in PXRD since the diffraction pattern has been compressed from 3D into only 1D. Thus, determination of a space group using indexed reflection intensities

Table 2.5: Relation of the crystal systems, observable Laue classes from PXRD, SC-XRD, and all crystal classes.

Crystal system	PXRD Laue class	SC-XRD Laue class(es)	All Crystal classes
Triclinic	$\bar{1}$	$\bar{1}$	$\bar{1}, 1$
Monoclinic	$2/m$	$2/m$	$2/m, 2, m$
Orthorhombic	mmm	mmm	$mmm, 222, 2mm$
Tetragonal	$4/mmm$	$4/m$ $4/mmm$	$4/m, 4, \bar{4}$ $4/mmm, 422, \bar{4}2m, 4mm$
Trigonal	$6/mmm$	$\bar{3}$ $\bar{3}m$	$\bar{3}, 3$ $\bar{3}m, 32, 3m$
Hexagonal		$6/m$ $6/mmm$	$6/m, 6, \bar{6}$ $6/mmm, 622, \bar{6}2m, 6mm$
Cubic	$m\bar{3}m$	$m\bar{3}$ $m\bar{3}m$	$m\bar{3}, 23$ $m\bar{3}m, 432, \bar{4}3m$

and systematic absences once the unit cell is identified does not benefit from reduction of options using the Laue classes when solving structures from powder data. Simple extraction of the unit-cell parameters based exclusively on the peak positions provides the maximum amount of information for the least amount of effort when analyzing powder diffractograms.

2.4 Powder X-ray Diffraction (PXRD)

Powder X-ray Diffraction (PXRD) is a routine method of characterizing and/or identifying a crystalline material, indispensable in high-throughput screening. A polycrystalline sample is prepared and irradiated with an X-ray beam, scanning from small to large 2θ angles in 1D, and the detector monitors the intensity of the reflected X-rays at the opposite angle with respect to the sample. Peaks appear in the powder diffractogram at angles where Bragg's law is satisfied. Ideally, the orientation of the crystallites in the sample are completely random to obtain an average sampling of each of the lattice planes and yield an accurate relative intensity of the reflections from the planes. In practice this is seldom the case, and obtaining sufficiently accurate peak intensities is further complicated by preferred orientation alongside coincident overlap of various independent reflections.

While intensity data from powder diffraction is of low-quality, positional data (2θ , and

thus d -spacing) can be very accurate with modern diffractometers without resorting to long, time-consuming scan conditions. Positional data may be skewed if the sample is displaced with respect to the expected position in the diffractometer, but this may be corrected by either diligent sample preparation or post-data collection.

Peak overlap is commonplace in PXRD, which can make utilization of the information contained in a powder diffractogram additionally challenging. Overlap may result for various reasons; however, the resolution of overlapping peaks is further complicated when peaks are of low intensity and broadened. Peaks become broader as the temperature at which the data is collected increases. Additionally, peaks both broaden and are reduced in intensity as the diffraction angle 2θ increases. In crystals formed by drug-like organic molecules, the larger size of the unit cell and light-element composition results in a tendency for low-intensity, broad, and overlapping peaks to be highly prevalent beyond $30^\circ 2\theta$. These conditions make deconvolution of the intensity data of a powder diffractogram into individual peaks very challenging (if not impossible) to do accurately for the vast majority of drug-like molecular crystals.

2.5 Structure Determination from Powder Data (SDPD)

There are a variety of reasons why one may seek to solve a crystal structure from a powder diffractogram. It is possible that growing a single crystal of sufficient quality is costly, inconvenient, been unsuccessful, or that reproducing the crystallization of the desired polymorph is a significant challenge and the only data available is the powder pattern (so-called “disappearing polymorphs”^{54,55}). The collection of powder diffraction data is routine, while equipment and expertise for single-crystal diffraction experiments is less common and more expensive. The endeavour of Structure Determination from Powder Data (SDPD) is not new, having been developed nearly a century ago. However, the latest (2022) statistics from the CSD (a database focused on molecular crystals) shows 1.23 million crystal structures in the database, with only 4,773 of these solved by SDPD ($\ll 1\%$) showcasing the disproportionate success of SC-XRD over SDPD. The protocol followed for most SDPD solutions is outlined in the sub-sections below.

2.5.1 Confirming the solution: Rietveld refinement

Some SDPD endeavours start and end with Rietveld refinement. The refinement procedure takes the calculated powder diffractogram from a trial crystal structure and attempts to match it to the experimentally collected diffractogram using point-wise comparison. If the trial crystal structure is a good representation of the true crystal structure, the refinement procedure will confirm the match. For inorganic crystals, where derivatives may be formed by substitution of central transition metals in (pseudo-)tetrahedral or (pseudo-)octahedral environments, while keeping the network assembly of the atoms in the unit cell relatively unchanged, it is easy to substitute the metal from the original structure with that of the derivative and have a very good representation of the true crystal structure of the derivative.⁵⁶

Rietveld refinement was a major development at its inception due to its ability to model a powder diffractogram as-is rather than requiring the deconvolution of overlapping peaks. It uses a least-squares approach to fit a theoretical powder diffractogram to a measured experimental diffractogram.⁵⁷ Minimization of the function:

$$M_T = \sum_i w_i [y_i^{\text{obs}} - sy_i^{\text{calc}}]^2 \quad (2.8)$$

is performed iteratively, where w_i is a weighting factor, y_i^{obs} is the experimental data point i , and y_i^{calc} is the corresponding i data point from the calculated profile (based on the trial structure) scaled by a phase scaling factor s . The list of possible refinable parameters that affect the y_i^{calc} values include the background, sample displacement (sample transparency and zero-shift), peak shape (Caglioti parameter and asymmetry), unit-cell dimensions (within about 1%), preferred orientation, adsorption, porosity, extinction coefficients, scale factor, atomic positions, occupancy factors (disorder), and thermal atomic displacement parameters. Not all of these parameters will be implemented in every program that runs a Rietveld refinement, and the default refinement protocol may vary between programs and users as well.

The success of a Rietveld refinement is commonly given by the value of one or more figures of merit. The weighted profile residual R_{wp} is one of the most utilized figures of

merit, defined by

$$R_{\text{wp}} = \left(\frac{\sum_i^n w_i (y_i^{\text{obs}} - y_i^{\text{calc}})^2}{\sum_i^n w_i (y_i^{\text{obs}})^2} \right)^{\frac{1}{2}} \times 100\%. \quad (2.9)$$

Often given in addition is a “goodness of fit parameter”, χ^2 :

$$\chi^2 = \sum_i^n \frac{(y_i^{\text{obs}} - y_i^{\text{calc}})^2}{n - p}, \quad (2.10)$$

where n is the number of data points and p is the number of refined parameters (note that formally, the statistical goodness of fit value is actually $\sqrt{\chi^2}$). However, there is no agreement on the maximum values of these figures of merit for a matching fit, and their values are highly dependent on the quality of the experimental data. Often, a visual assessment of the fit is done for confirmation of a match.⁵⁸

The generation of a suitable trial crystal structure to use for Rietveld refinement in the case of molecular crystals is very different from the case of inorganic crystals — a molecular derivative packing the same way as the original molecule is the exception rather than the norm. In these cases, using the crystal structure of the original as a guess for the crystal structure of the derivative will result in a failed Rietveld refinement.

The influence of some of the initial parameters given for a Rietveld refinement cannot be overstated. In particular, if the atomic positions and unit-cell dimensions are not very close to their experimental positions (i.e. unless you have a good initial candidate crystal structure), the refinement will prove unsuccessful as no refinement-level modification to the structure will be able to generate a calculated powder pattern that will match the experimental one. For molecular crystals, it is even possible for crystal structures collected at different temperatures to fail a Rietveld refinement since anisotropic changes in unit-cell dimensions have a profound effect on the observed powder diffractogram. Additionally, impurities in the experimental pattern will make the refinement mostly unsuccessful unless the theoretical patterns of both (all) phases are used in the refinement.

The following subsections describe the procedure undertaken in order to generate a sufficiently accurate initial guess at the crystal structure for a molecular crystal such that a successful Rietveld refinement can be done to confirm the trial crystal structure matches the experimental powder diffractogram.

2.5.2 PXRD indexing

Crystal structure solutions from powder data begin with the determination of the unit cell from the peak positions. The problem is 6-dimensional, with three axes and three angles to determine for the cell that must match the peak positions in the diffractogram. The term *indexing* comes from the assignment of Miller indices to the peaks that correspond to viable unit cell dimensions by solving numerous Bragg's law problems. The success of these algorithms depends on phase purity of the sample (peaks may be present from impurities as low as 1-2% by weight in PXRD), peak position accuracy, and a reasonable number of peaks (usually ≥ 20).

There are many different algorithms that undertake the task of determining the unit cell dimensions from the powder pattern peak positions. Common programs include zone-indexing (ITO⁵⁹), index-heuristics (TREOR⁶⁰), and successive dichotomy (DICVOL⁶¹) search methods. As each method employs a different approach, it is common to use more than one in an attempt to index a powder pattern; one method may miss the correct solution, or they may agree on the best solution and the redundant identification improves confidence in the solution found.

The confidence in the (many) unit cells proposed by an indexing program is also given by the program as a figure of merit. The two most commonly used figures of merit are the de Wolff M_{20} ⁶² value and the Smith-Snyder F_N ⁶³ value. The general rule is that a larger value indicates a better fit of the unit cell to the experimental data. It is always possible that the unit-cell dimensions that give the best figure of merit are incompatible with the compound of interest, however, and checks to ensure the validity of the cell (for example, a reasonable Z and Z' are obtained using a rough molecular volume²¹ in the proposed cell) should always be done.

2.5.3 Structure generation by direct methods

Direct methods of SDPD are derived from SC-XRD-based solutions, where individual peaks are identified and, using their position and intensity, an electron density map is generated. As mentioned previously, the generation of an electron density map is severely complicated when using a powder diffractogram vs. the diffraction pattern from SC-XRD due to peak overlap. While this method can be used successfully when the powder pattern has many sharp, well-resolved peaks, and their intensities can be determined with high accuracy, this effectively restricts direct methods to inorganic solids and simple organic

molecules with small unit cells and high symmetry. No further discussion of direct methods is warranted in this thesis; however, the interested reader may refer to the recent review of the topic in the International Tables for Crystallography, Volume H.¹⁹

2.5.4 Structure generation by real space methods

The basis of real space methods of SDPD is an iterative-improvement, trial-and-error approach. Fundamentally, only the composition of the material is required. However, in general, the unit cell and space group are used to constrain the search space. Additional constraints and restraints can be applied, such as bond lengths, angles, dihedrals, etc. between different atoms. An algorithm then generates structures and compares the calculated powder diffractogram of the trial structure to the experimental diffractogram. Even with all the constraints and restraints that can be applied, often the search space is still enormous. Therefore, simple stochastic algorithms that randomly generate trial crystal structures are far from the most effective use of computational resources. In order to generate the matching structure more quickly and with fewer trial structures, algorithms incorporate feedback based on how well the trial structures already generated match the target diffractogram, which requires some kind of figure of merit.

Figures of merit used vary by program, but commonly take a whole-profile approach, using the χ^2 or R_{wp} values from Rietveld refinement (see Section 2.5.1). Sometimes the residuals on the extracted intensities or structure factors from a LeBail refinement are used; however, these introduce additional sources of error based on peak deconvolution and should be avoided in assessing the fit for molecular crystal structures. This lack of confidence in the peak intensities alludes to the issues inherent in using profile-matching as the exclusive guiding function for these algorithms; it may be that certain features of the diffractogram drive the structure generation but are not chemically sensible. Additionally, the hypersurface generated by the χ^2 or R_{wp} value is known to be quite flat in regions where the structure is very far from matching, up to nearly matching, with steep wells to local and global minima (see Fig. 2.5), making it a rather poor director of the algorithm.²⁰

The figure of merit is used in different ways by the different algorithms in order to direct the generation of the next trial structure. A thorough discussion of these global optimization methods and references to the development and parameters pertinent to their usage is presented in David's 2002 book on *Structure Determination from Powder Data*.⁵³

In simulated annealing, if the figure of merit value of the most recently generated trial

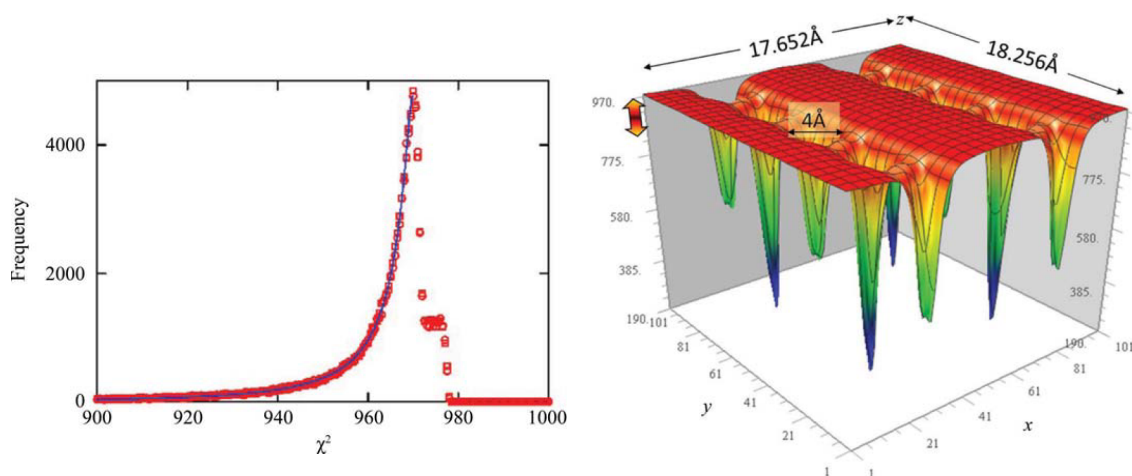


Figure 2.5: The crystal structure of famotidine ($C_8H_{15}N_7O_2S_3$) can be reduced to 13 degrees of freedom (7 internal torsion angles, and 6 molecular position/orientation parameters) after applying constraints on the atomic bond lengths and angles within the molecule, and defining the units cell from indexing. LEFT: Distribution of χ^2 values from randomly selecting values for the 13 degrees of freedom. RIGHT: The “slice” of the χ^2 hypersurface that corresponds to the variation of selected 2D molecular position parameters in the unit cell with the 11 other parameters held constant at their optimized values. (Reproduced with permission from ref. 20.)

structure is smaller than the previous trial structure, it is accepted and the next structure is generated by applying some random modification to the most recent trial structure. If the figure of merit of the new structure is larger than that of the previous trial structure, the structure is only accepted with a Boltzmann probability that is related to a combination of the difference between the figures of merit of the two structures, and a parameter that changes over the course of the simulation (reducing the chances of up-hill changes as it progresses). If the new structure is rejected, a new change to the initial structure is made and the algorithm continues.

Parallel tempering performs multiple simulated annealing runs in parallel, then at chosen (possibly random) intervals, allows for change-over and mixing of the structures between the parallel runs, also with a particular probability. Since the success of identifying the global minimum from a single simulated annealing run decreases as the complexity of the system increases, multiple simulated annealing runs are commonly done for completeness. Parallel tempering improves on multiple serial simulated annealing runs by allowing cross-over between runs. The structure with the best figure of merit at the end of a defined number of steps is taken to be the solution.

Genetic or evolutionary algorithms draw inspiration from biology, and start by generating a batch (population) of trial structures. The best structures according to the figure of merit are then combined (mating), generating a new batch (child population) with a random mix of features belonging to the fittest structures from the previous generation. Random changes (mutations) may also be imparted on the child structures of a subsequent generation. This continues for a set number of generations.

As the χ^2 and R_{wp} hypersurfaces have no direct consideration of chemical feasibility, structures may be generated that have overlapping atoms/molecules. In order to reduce the probability of such structures being carried through the algorithms, the addition of penalty functions that consider atomic overlap have been combined with the χ^2 and R_{wp} figures of merit to generate hybrid cost functions that are determined for each structure.⁶⁴ This is the simplest implementation of an “energy” parameter, effectively using overlap as a crude measure of repulsion. The use of two-body repulsive potentials ($\propto 1/r^{12}$) have also been implemented to impart more physically realistic repulsion energies as a penalty.⁶⁵⁻⁶⁷

CHAPTER 3

QUANTITATIVE METHODS OF CRYSTAL STRUCTURE COMPARISON

Rietveld refinement changed the way that structure models were refined to prove a match between the trial structure and the experimentally observed powder diffractogram. However, the comparison of crystal structures in general is of substantial interest as well, not in order to refine a match, but to identify whether two crystal structures are the same form or different forms. Our ability to compare crystal structures visually is often used to qualitatively compare similarities and differences to determine whether the crystal structures being compared are polymorphs. While predominantly correct, this method of comparison is neither automatic nor quantitative, making the question “which is *more* similar to the target crystal structure?” often challenging to answer. While developed quantitative methods may also vary in their answer to this question of greater similarity, they can be automated to compare vast numbers of crystal structures to one another without requiring human attention. The two most common methods of quantitative crystal structure comparison are described in this section using two specific programmed algorithms.

3.1 Measurement of relative atomic positions

If one chooses the “same” atom as the origin in two different crystal structures, measurements from this atom to other atoms in the two structures should yield the same distances to the same atoms if, indeed, the two structures are the same. By making a set of measurements between the “origin” atom and other atoms in the two crystal structures, and comparing these distances and the angles between different measurements (Figure

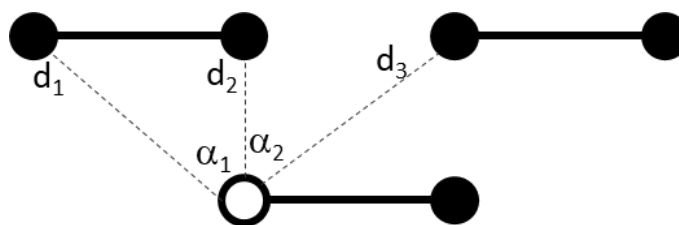


Figure 3.1: Diagram showing how COMPACK measures distances (d_1, d_2, d_3) and angles (α_1, α_2) from the chosen origin atom, identified as the hollowed circle, to atoms of the surrounding molecules. In this example, the cluster size would be 3.

3.1), the COMPACK algorithm²² provides quantitative values of similarity between crystal structures.

The algorithm aims to determine whether the two crystal structures are the same within a given tolerance of difference in the measured interatomic distances and angles. COMPACK is primarily designed for molecular crystals and includes a check to ensure that the crystal structures being compared contain the same molecular species. Accordingly, a cluster size (how many molecules) is also required for the comparison. A cluster size of 1 compares the molecular conformation (or determines whether the molecules match, if one were to compare different molecules). A cluster size larger than one begins to compare intermolecular positions, i.e. packing. For crystal structures, a representative comparison of packing is generally accepted to be achieved at a cluster size of 20 molecules (see Chapter 5 for counter examples).⁶⁸

When a match of 20/20 molecules in the cluster is made, an optimal overlay of the two structures is generated and the root-mean-squared-deviation (RMSD) of the atomic positions of the structure overlay is calculated^{69,70} and reported. Generally, matching 20/20 molecules in the cluster allows one to conclude that the two crystal structures being compared are the same, so the RMSD(20) (the number in parentheses indicates the size of the molecular cluster compared) value is related to the degree of similarity, with smaller values indicating higher similarity (RMSD(20) = 0 Å is an exact match). The RMSD values are given in Å as the inter-atomic distance measurements are made in Å.

If the two crystal structures being compared fail to match for the particular cluster size within the given tolerances, the algorithm reports the total number of molecules that could be matched for the cluster (e.g. 3/20 indicates that three molecules of the cluster of 20 match within the given tolerances). In these cases, there is little value in reporting the RMSD if one is solely concerned with whether the two structures are the same or not—the

calculated RMSD (continuing with the example, this would be RMSD(3)) is unrelated to the cluster size of interest. The fact that the inter-atomic measurements do not agree within the cluster size of interest is more representative of the degree of similarity (i.e. they are completely different since only 3 of the 20 molecules match) than an RMSD value.

Some interesting properties and caveats to the COMPACK algorithm are discussed in Chapter 5.

3.2 Powder diffractogram comparison

As shown in Section 2.3.1, the calculation of a 1D powder diffractogram from a crystal structure is facile using modern computers. Thus, the comparison of the calculated powder diffractograms from two crystal structures is a viable method for determining their similarity. The figure of merit values from Rietveld refinement are such a measure of similarity, yielding values that relate to the disagreement between the model crystal structure and collected data. Analogously, a simple point-wise difference calculation performed on the normalized simulated powder diffractograms of two crystal structures can be used as a quantitative measure of (dis)similarity between the two crystal structures. This approach is found to perform generally poorly for powder diffractograms due to the effect of even minor deviations in the lattice dimensions on the peak maxima of the pattern.²³

A metric that is general in its application, but has been identified as particularly useful in the comparison of powder diffractograms, is the similarity value yielded by a triangle-weighted cross-correlation function developed by de Gelder and colleagues in 2001.²³ The method treats the two 1D diffractograms as functions (e.g. $f(x)$ and $g(x)$, where x is 2θ), and establishes similarity from measurement of the overlap area of the two functions:

$$c_{fg}(\delta) = \int f(x)g(x + \delta)dx. \quad (3.1)$$

The patterns are normalized such that an absolute scale of similarity can be established despite different patterns having different absolute integral values.

The comparison differs from a simple point-wise comparison by taking into account the “surroundings” of every point on the plot (the off-set shift between the two functions, δ).

This is done using a triangle-weighting function, $w(\delta)$, where

$$w(\delta) = \begin{cases} 1 - |\delta|/b_t, & |\delta| < b_t \\ 0, & |\delta| \geq b_t. \end{cases} \quad (3.2)$$

with the length of the “base” of the triangle defined by b_t .

The final expression of similarity is

$$S_{fg} = \frac{\int w(\delta)c_{fg}(\delta)d\delta}{\left(\int w(\delta)c_{ff}(\delta)d\delta \int w(\delta)c_{gg}(\delta)d\delta \right)^{\frac{1}{2}}}, \quad (3.3)$$

and the expression for dissimilarity is

$$D_{fg} = S_{ff} + S_{gg} - 2S_{fg}, \quad (3.4)$$

which is simply equal to $1 - S_{fg}$ when normalized to 1. This algorithm works very well when using calculated powder diffractograms since both will be simulated with the same peak shape parameters, and only their positions and intensities will change based on the crystal structure. Thus, with accounting of the “surroundings”, this method of comparison using a weighted cross-correlation function is able to handle some minor shifting in the peak positions and still yield a low dissimilarity value for cases where two crystal structures of the same form are used. The range of 2θ shift that can be accommodated by the weighted cross-correlation function, however, is completely insufficient to yield useful dissimilarity values for molecular crystals collected at different temperatures or pressures.

3.3 Challenges in powder diffraction-based comparison methods

Generally, powder diffraction-based quantitative comparison methods are only useful in identifying identical structures collected at the same set of conditions, severely restricting their utility, because shifting of diffraction peaks due to changes in temperature or pressure will yield a value inseparable from one yielded by the comparison of two polymorphic structures. In order to account for condition-induced peak shifting, modifications to the crystal structure(s) may be applied prior to simulation and comparison of the powder

diffraction patterns.

The approach taken by van de Streek and Motherwell in 2005⁷¹ was to isotropically correct both crystal structures being compared to a calculated unit-cell volume based on empirical molecular volumes as tabulated by Hofmann.⁷² Following the volume adjustment, the powder diffraction patterns were simulated and compared using the weighted cross-correlation function method. While some improvement was observed with this approach, it was also noted that cases of significant anisotropic changes in the crystal lattice were poorly handled. A survey of the CSD in 2021 determined that anisotropy in thermal expansion is more the norm than the exception when considering molecular crystals.⁷³ Thus, a modification that is solely isotropic is unlikely to be generally effective, particularly when comparing crystal structures from more extreme conditions (e.g. “0 K” for *in silico* generated structures).

The FIDEL (Fit with DEviating Lattice) method was specifically developed for the comparison of unindexed experimental powder diffraction patterns to the simulated powder diffraction pattern of a model crystal structure.⁷⁴ A hill climber’s algorithm is used to make modifications to the unit cell dimensions (and other structural parameters) of the model crystal structure in order to maximize the agreement between the two powder diffraction patterns (measured using the weighted cross-correlation function method). Ideally, this results in alignment of the Bragg’s peaks; however, the method is susceptible to false minima where proper alignment of all peaks is not achieved (see Chapter 6). The application of this method to the comparison of two crystal structures/simulated powder diffraction patterns has remained absent from the literature at the time of writing, however it is assumed to be used.

The development of the VC-PWDF method stemmed from the assumption that the unit cell of both structures being compared was known and – provided the two crystal structures were the same form – a deformation could be applied to distort one structure to match the other and yield perfect overlap of the Bragg’s peaks to yield a quantitative value that would clearly classify the two structures as “matching”. Accordingly, non-matching structures would yield values that indicated their dissimilarity, as there is no deformation that yields a coincident overlap of the crystal structures and, therefore, the Bragg’s peaks in the powder diffraction patterns.

3.4 The VC-PWDF protocol

The development of the Variable Cell PoWder DiFference (VC-PWDF) method was the result of a need for a robust powder diffraction-based method of crystal structure comparison. The initial version of the method was conceived, developed, and coded by the author (R. Alex Mayo), presented in Chapter 4. Collaboration with Dr. Alberto Otero de la Roza resulted in the integration of the method into the *critic2* program,⁷⁵ which was coded by Dr. Otero de la Roza. During the merger, the finite number of variable cells explored by the initial version of the program was expanded in order to provide a general solution. This change, presented in Chapter 5, was a result of contributions from both the author and Dr. Otero de la Roza. Thus, the inception of the methodology was a result of ideas and work done by the author, while the current version of the method and its greater accessibility through the *critic2* program is a result of contributions from both the author and Dr. Otero de la Roza.

Given two crystal structures, **xtal1** and **xtal2**, the first step of the VC-PWDF protocol is to convert both structures to their Niggli reduced cell. The Niggli reduction algorithm yields a unique unit cell that is the smallest primitive unit cell for that crystal structure, with axes assigned in order of increasing length (a is the shortest, b is second shortest, and c is longest). The protocol then chooses a reference structure and candidate structure from analysis of the Niggli cells of **xtal1** and **xtal2**. The reference structure is chosen to have the higher number of atoms in its Niggli cell, or the crystal structure entered first if both contain the same number of atoms.

In the majority of cases, the distortion that results from matching the dimensions of the Niggli cell of the candidate with the Niggli cell of the reference (i.e. $a_1 \rightarrow a_2$, $b_1 \rightarrow b_2$, $c_1 \rightarrow c_2$) yields the coincident overlay. However, anisotropic thermal expansion can cause a switch in the lattice vectors used for the Niggli cell (i.e. their rank in terms of shortest length) depending on the temperature and pressure conditions. Thus, the direct deformation from the candidate's Niggli cell to the dimensions of reference's Niggli cell is not a general solution to the problem of aligning two crystal structures collected under different conditions.

The first iteration of the VC-PWDF method looked to resolve this issue by applying a finite number of transformation matrices (maximum of 24) to the Niggli-cell basis of the candidate structure, followed by a deformation of each resulting trial unit cell to match

the dimensions of the reference structure's Niggli cell. The deformed trial structures were all compared to the reference structure using the cross-correlation function and the lowest value (measure of dissimilarity, so smaller indicates more similar) was taken as the representative value, or "VC-PWDF score". While still an incomplete solution, the exploration of "variable cells" made considerable strides towards a general solution for this problem, and was a notable improvement on the other available methods at the time.

The second iteration of the VC-PWDF method saw its integration into the critic2 program. Rather than being run as a bash script that utilized many of the critic2 functions (generating inputs and reading outputs) to produce the results, the protocol was now included as a critic2 subroutine. With this integration, some of the other critic2 subroutines became more accessible and a general solution to the cases of inequivalent Niggli cells of matching structures from different conditions was developed.

Rather than applying transformation matrices to the Niggli cell of the candidate structure, a list of the lattice vectors of the candidate structure was compiled and each lattice vector matched to the lattice vector(s) forming the Niggli-cell basis of the reference structure. From this list of lattice vectors, trial unit cells were generated for the candidate structure, deformed to match the Niggli-cell dimensions of the reference structure, and compared with the reference structure using a cross-correlation function. Again, the lowest dissimilarity value obtained is taken as the VC-PWDF score.

The following additional criteria were applied in order to set reasonable boundaries on the protocol while being overly generous in allowing dramatic distortions:

- lattice vectors of the candidate structure are assigned to the lattice vectors of the Niggli-cell basis of the reference structure only if they are within 30% of the length (only these assigned lattice vectors are used to generate trial unit cells of the candidate structure), and
- trial unit cells are discarded if:
 - two (or more) lattice vectors used are colinear,
 - the resultant angles differ by $\pm 20^\circ$ from the reference Niggli cell angles,
 - the total volume of the unit cell exceeds a 50% change in volume, and/or
 - it does not contain the same number of atoms as the reference Niggli cell.

A flowchart outlining the main steps of the VC-PWDF protocol is given in Figure 3.2.

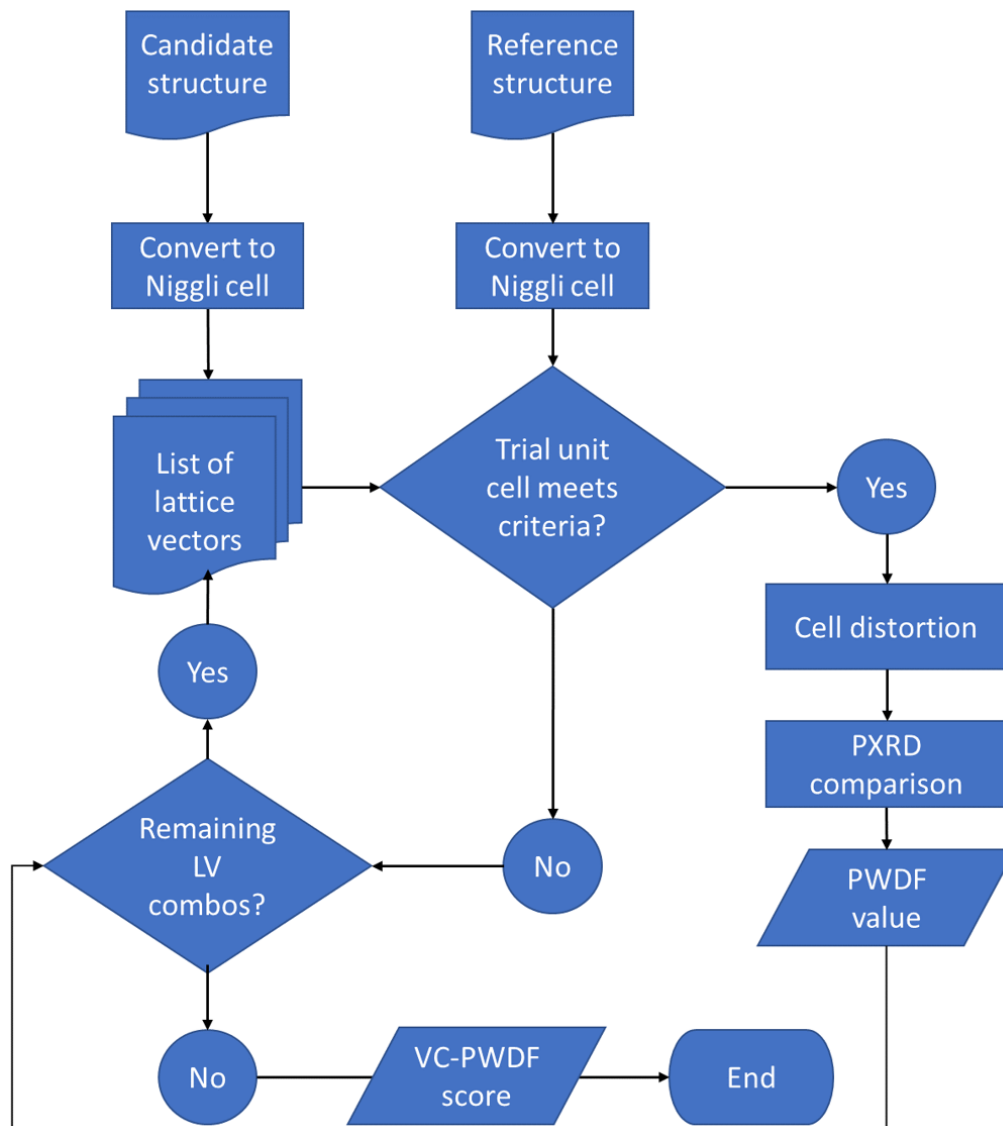


Figure 3.2: Flowchart of the steps undertaken by the VC-PWDF protocol when comparing two crystal structures.

3.5 VC-PWDF example

As an example case, the experimental crystal structure of form D of 2-((4-(3,4-dichlorophenethyl)phenyl)amino)benzoic acid (XXIII_D.cif) and the structure ranked 11th by energy from Group 14 of the 6th CSP blind test⁴¹ (G14_Erank11.cif) will be compared (Figure 3.3).

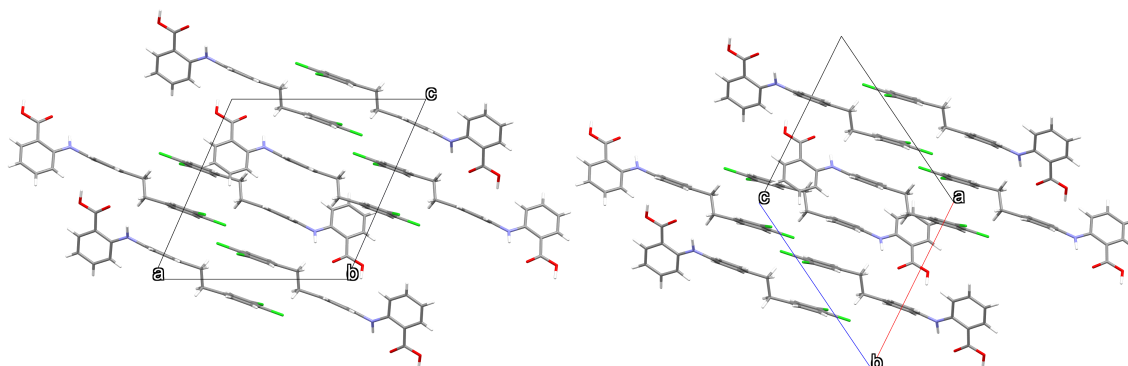


Figure 3.3: Comparable views of the two example crystal structures, XXIII_D (left) and G14_Erank11 (right).

The first step is to load the crystal structures into `critic2` (Figure 3.4). They are immediately reduced to their Niggli cells (Figure 3.5), and the reference structure (XXIII_D.cif) is selected. The lattice vectors of the Niggli reduced cells and Niggli-cell dimensions for both structures are also printed (Figure 3.6).

```
critic2:1> %% comparevc XXIII_D.cif G14_Erank11.cif
* COMPARE, allowing for deformed cells
+ Reading the structure from: XXIII_D.cif
+ Reading the structure from: G14_Erank11.cif
+ Using as reference: XXIII_D.cif
  The other crystal will be transformed to match the reference.
```

Figure 3.4: Sample `critic2` output showing which of the two structures being compared (XXIII_D.cif, vs. G14_Erank11.cif) is chosen as the reference structure.

Linear combinations of lattice vectors of the candidate structure (with respect to the Niggli-cell basis) are then listed from shortest to longest, along with their assignment to the unit-cell dimensions of the reference structure's Niggli cell (Figure 3.7). The numbers in brackets (used-by column) are the corresponding lattice vector assignments with that of the Niggli-cell basis of the reference structure. The last vector (#21) is unassigned as it is outside the range of viable lattice vector lengths.

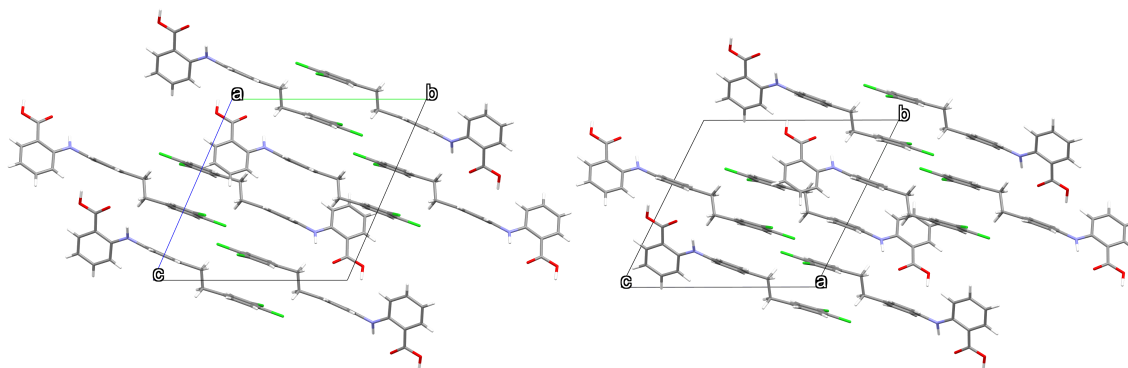


Figure 3.5: Comparable views of the Niggli cells of XXIII_D (left) and G14_Erank11 (right).

```

+ Niggli lattice vectors (bohr)
# Structure 1 (XXIII_D.cif):
  a:  -0.00000000000  -20.2729819536   0.00000000000  length =  20.2729819536
  b:  -26.2407370813   0.00000000000   0.00000000000  length =  26.2407370813
  c:   10.6643251791  -0.00000000000  -24.3725626989  length =  26.6035644988
Lengths (ang): 10.72800 13.88600 14.07800
Angles (deg): 113.632  90.000  90.000
# Structure 2 (G14_Erank11.cif):
  a:  -0.00000000000  -20.5040071643   0.00000000000  length =  20.5040071643
  b:  -25.6309445870   0.00000000000   0.00000000000  length =  25.6309445870
  c:   12.0023583778  -0.00000000000  -24.0097470966  length =  26.8425885911
Lengths (ang): 10.85025 13.56331 14.20449
Angles (deg): 116.560  90.000  90.000

```

Figure 3.6: Sample critic2 output showing the Niggli-cell lattice vectors and dimensions of the two structures being compared (XXIII_D.cif, vs. G14_Erank11.cif).

The next step is to form trial unit cells for the candidate structure using three of the lattice vectors listed in Figure 3.7. Some examples of trial unit cells are shown in Figure 3.8. When a viable unit cell is formed, it is distorted to match the reference structure's Niggli cell and the PWDF value is calculated, as shown in Figure 3.9. The three lattice vectors of the candidate structure used are listed (e.g. the first viable trial cell used lattice vectors 1 as axis *a*, 3 as axis *b*, and 5 as axis *c*), along with the maximum change in cell lengths and angles, and the powder difference (PWDF) value post-distortion. While the length of candidate lattice vector 1 is within 30% of all three Niggli cell lattice vectors of the reference structure (Figure 3.7, ID #1 has (123) in the “used-by” column), there is no PWDF value associated with all three unit cell vectors assigned to lattice vector #1 as this would be an invalid unit cell.

```
+ Candidate lattice vectors for structure 2 (referred to the Niggli basis):
```

#Id	x	y	z	length	used-by
1	1.00	0.00	0.00	20.504007	(123)
2	-1.00	0.00	0.00	20.504007	(123)
3	0.00	-1.00	0.00	25.630945	(123)
4	0.00	1.00	0.00	25.630945	(123)
5	0.00	0.00	-1.00	26.842589	(23)
6	0.00	0.00	1.00	26.842589	(23)
7	0.00	1.00	1.00	27.608084	(23)
8	0.00	-1.00	-1.00	27.608084	(23)
9	1.00	1.00	0.00	32.823157	(23)
10	-1.00	-1.00	0.00	32.823157	(23)
11	-1.00	1.00	0.00	32.823157	(23)
12	1.00	-1.00	0.00	32.823157	(23)
13	1.00	0.00	-1.00	33.777787	(23)
14	-1.00	0.00	1.00	33.777787	(23)
15	-1.00	0.00	-1.00	33.777787	(23)
16	1.00	0.00	1.00	33.777787	(23)
17	-1.00	-1.00	-1.00	34.389252	(3)
18	-1.00	1.00	1.00	34.389252	(3)
19	1.00	1.00	1.00	34.389252	(3)
20	1.00	-1.00	-1.00	34.389252	(3)
21	-2.00	0.00	0.00	41.008014	

Figure 3.7: Sample critic2 output showing possible lattice vectors of the candidate structure (with respect to the Niggli-cell basis), along with their assignment (number in brackets) to the Niggli-cell vectors of the reference structure.

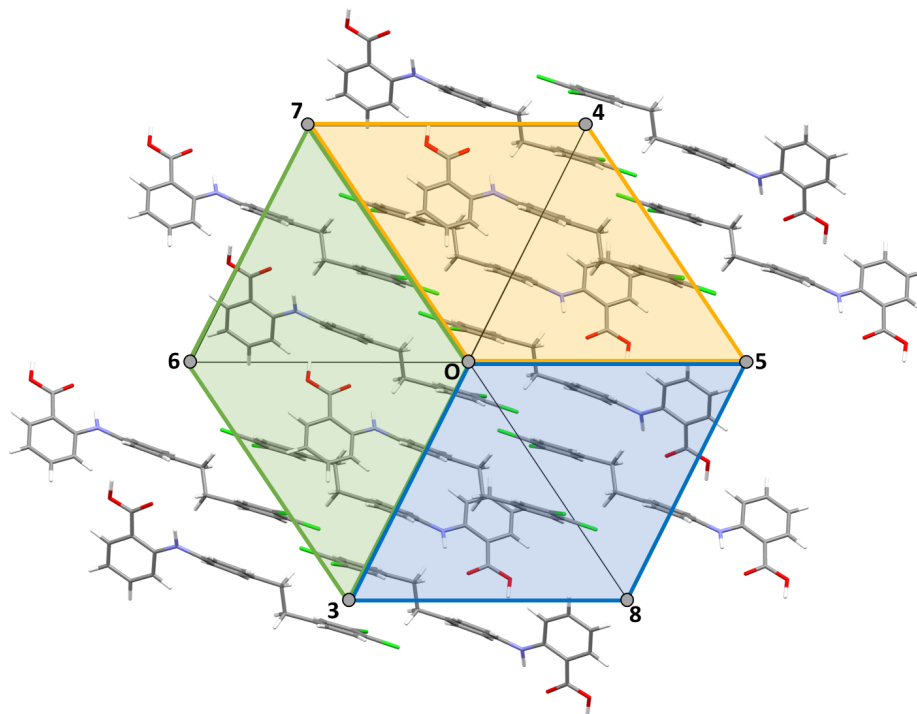


Figure 3.8: View of the (100) plane of the Niggli cell of the candidate crystal structure. Lattice points are shown as grey circles, and numbers identify the lattice vector (according to Figure 3.7) formed by tracing from the origin to that lattice point. Coloured parallelograms provide examples of possible unit cell descriptions of the structure (constant a -axis).

```

+ Structural comparison of candidate structures
# Reference structure is 1.
# Structure 2 takes lattice vectors (a,b,c) from the list above.
# max-dlen = maximum difference in cell lengths (bohr).
# max-dang = maximum difference in angles (degree).
#a b c max-dlen max-dang powdiff
+ INITIAL DIFF = 0.087223955
 1 3 5 0.6098 2.928 0.0872240
 1 3 7 1.0045 5.948 0.2032314
 1 4 6 0.6098 2.928 0.0872240
 1 4 8 1.0045 5.948 0.2032314
 1 5 3 0.9726 2.928 0.0057937
 1 5 7 1.0045 10.227 0.4841775
 1 6 4 0.9726 2.928 0.0057937
 1 6 8 1.0045 10.227 0.4841775
 1 7 3 1.3673 5.948 0.1132656
 1 7 5 1.3673 10.227 0.4918237
 1 8 4 1.3673 5.948 0.1132656
 1 8 6 1.3673 10.227 0.4918237
 2 3 5 0.6098 2.928 0.0872240
 2 3 7 1.0045 5.948 0.2032314
 2 4 6 0.6098 2.928 0.0872240
 2 4 8 1.0045 5.948 0.2032314
 2 5 3 0.9726 2.928 0.0057937
 2 5 7 1.0045 10.227 0.4841775
 2 6 4 0.9726 2.928 0.0057937
 2 6 8 1.0045 10.227 0.4841775
 2 7 3 1.3673 5.948 0.1132656
 2 7 5 1.3673 10.227 0.4918237
 2 8 4 1.3673 5.948 0.1132656
 2 8 6 1.3673 10.227 0.4918237
+ FINAL DIFF = 0.005793737

```

Figure 3.9: Sample critic2 output showing the trial unit cells of the candidate structure (using the numbered lattice vectors shown in Figure 3.7), the maximum difference with respect to the original candidate structure in terms of lattice vector length and unit-cell angle differences, and the resultant PWDF value after deformation of the trial unit cell to match the reference structure.

Once all valid trial unit cells comprised of unique combinations of lattice vectors of the candidate structure have been explored, the last line of the output gives the VC-PWDF score (+ FINAL DIFF). The score is simply the smallest PWDF value calculated throughout the protocol. Removal of hydrogen atoms prior to the powder diffractogram simulation and comparison can be done by using the NOH keyword. This has a minimal effect on the final VC-PWDF score due to the very small atomic scattering factor of hydrogen; in this case, the score changes to 0.0053.

Various crystal structure overlays post-distortions are shown in Figure 3.10, and their overlaid simulated powder diffractograms in Figure 3.11. The overlay shown in the bottom right of each of these figures corresponds to the coincident unit-cell description.

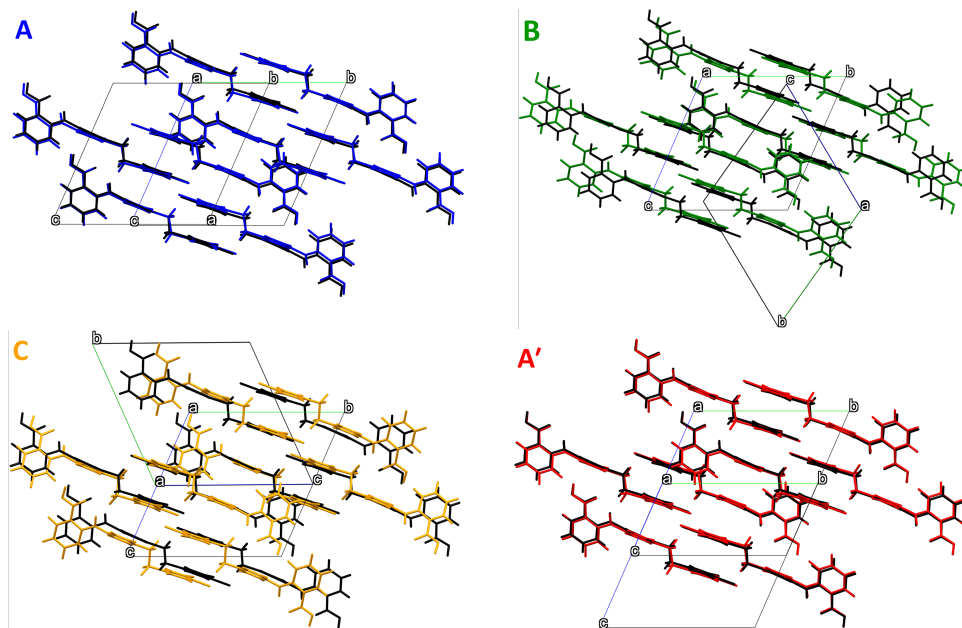


Figure 3.10: Overlays of the reference crystal structure XXIII_D (black) with different distorted trial unit cells of the candidate structure G14_Erank11 (different colours/letters corresponding to distortions from different unit-cell description). A' is the coincident unit-cell description and correct overlay.

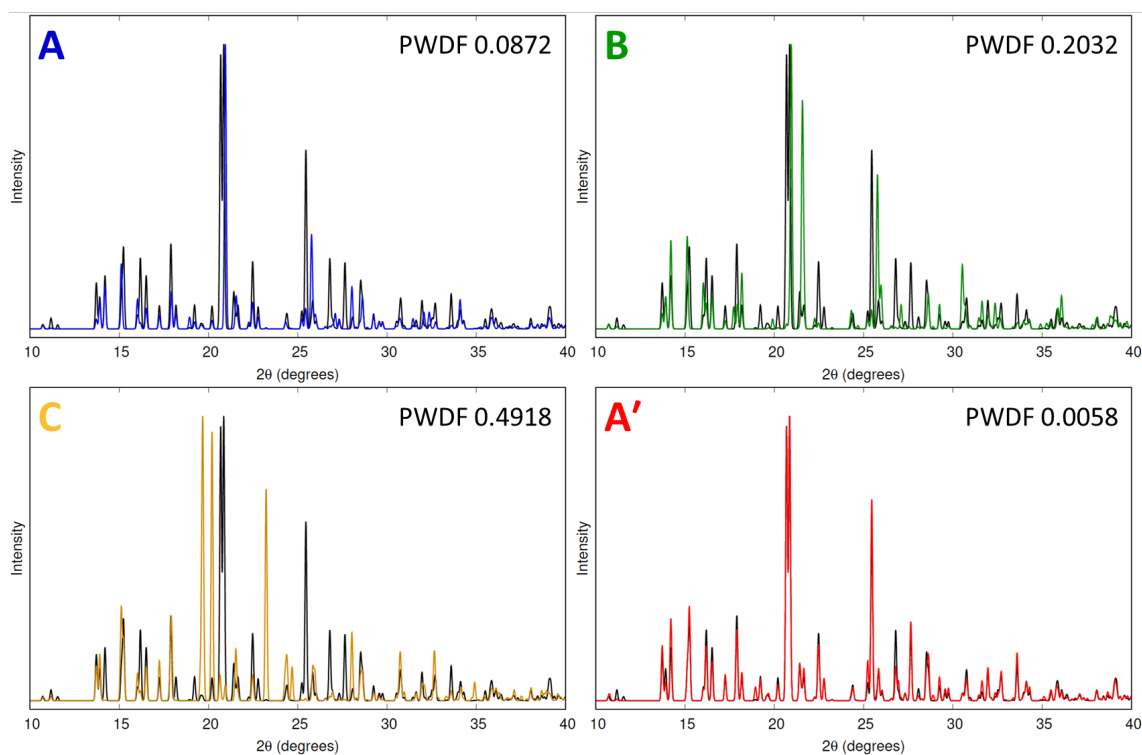


Figure 3.11: Overlays of the simulated powder diffractograms of the reference crystal structure XXIII_D (black) with different distorted trial unit cells of the candidate structure G14_Erank11 (different colours/letters corresponding to distortions from different unit-cell description). A' is the coincident unit-cell description and correct overlay.

As the atomic positions in the unit cell are given in fractional coordinates of the basis (formed by the lattice vectors comprising the unit cell), the distortion of the unit cell affects the molecular geometry, but minimally. A measure of the RMSD(1) and corresponding visual overlay for each of the selected example trial unit cells, before and after the distortion, is shown in Figure 3.12. Table 3.1 summarizes the maximum change in axis length, molecular bond length, angle, and dihedral, RMSD(1) and PWDF score. The distortion of the correct corresponding trial unit cell results in the smallest change in overall molecular geometry measured by RMSD (A', bottom-right).

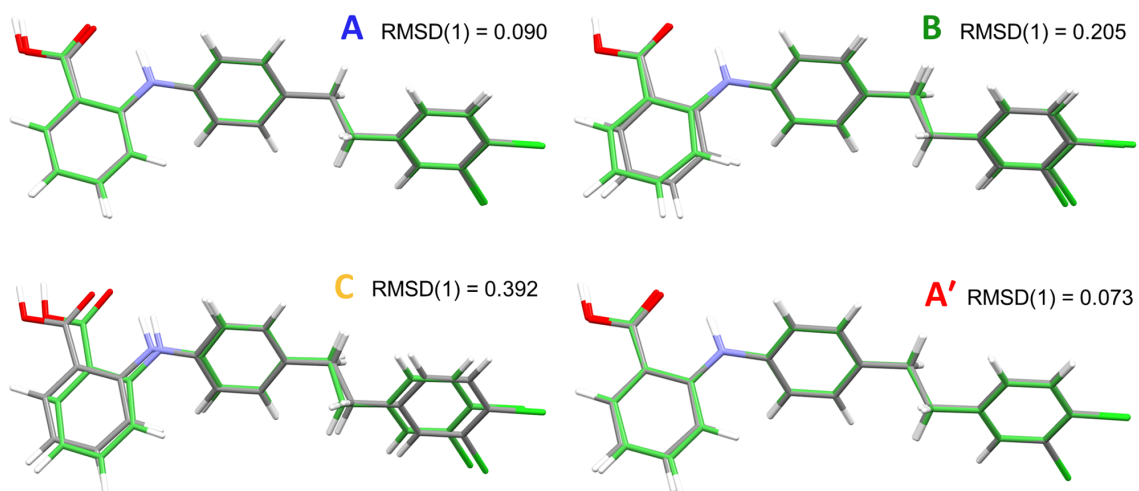


Figure 3.12: Molecular overlays and RMSD(1) values comparing the initial conformation prior to distortion (G14_Erank11) and after distortion (green-coloured carbon atoms) of the selected example trial unit cells. The smallest RMSD(1) is associated with distortion of the coincident unit cell description A'.

Table 3.1: Maximum change in axis length (\AA), bond length (\AA), bond angle ($^\circ$), dihedral angle ($^\circ$), and RMSD(1) (\AA) after application of the distortion for the four example trial unit cells.

trial unit cell	axis	bond length	bond angle	dihedral	RMSD(1)	PWDF
A	0.323	0.07	2.8	1.3	0.090	0.0872
B	0.532	0.10	5.4	0.4	0.205	0.2032
C	0.532	0.12	9.2	3.2	0.392	0.4918
A'	0.515	0.09	3.2	1.7	0.073	0.0058

CHAPTER 4

IMPROVED QUANTITATIVE CRYSTAL-STRUCTURE COMPARISON USING POWDER DIFFRACTOGRAMS VIA ANISOTROPIC VOLUME CORRECTION

Reprinted with permission from **R. Alex Mayo** and Erin R. Johnson, “Improved Quantitative Crystal-Structure Comparison using Powder Diffractograms *via* Anisotropic Volume Correction”, *CrystEngComm*, **23**, 7118-7131 (2021), DOI: 10.1039/D1CE01058A, Copyright 2021 Royal Society of Chemistry.

R. Alex Mayo wrote the VC-PWDF code, performed all exploratory and benchmarking calculations and crystal structure comparisons, analyzed the data, made figures, and wrote the first draft of the manuscript. ERJ supervised the project and edited the manuscript.

4.1 Introduction

The phenomenon of polymorphism is inextricably bound to the fields of materials science and pharmaceuticals, where the determination of the crystal structure is a critical step in compound discovery and characterization. If polymorphs exist, it is important to distinguish between them since even subtle changes between crystal structures can cause dramatic changes in their bulk properties.^{76–79} The discovery of polymorphs for a compound of interest has the potential to realize the desired properties of a material,^{80–83} or to severely complicate its production.^{29,84}

The ideals of first-principles crystal structure prediction (CSP)^{85–89} are to provide a means to screen molecules (before they are synthesized in the laboratory) to predict whether they will yield materials with desired properties, and to assess polymorphism risk for new pharmaceuticals.^{31,33,84,90–92} In practice, the crystal structure-energy landscapes generated by CSP do not usually provide a definitive structure, or list of polymorphs, that will be observed experimentally for the molecule of interest. Rather, hundreds of thousands of trial structures are generated in the first step of the CSP protocol, which are ranked energetically to identify the most likely candidates. The choice of theoretical method can have a profound influence on the resultant structure-energy landscape, so that energy re-ranking with higher-level theoretical methods is often performed at later stages of the CSP protocol.^{30,91,93–97}

CSP methods are commonly benchmarked by performing studies on previously characterized molecules, in order to determine how the method ranks the experimentally observed crystal structure(s). This approach forms the basis of the CSP blind tests coordinated by the Cambridge Crystallographic Data Centre (CCDC).^{36–41} Identifying whether any of the candidate structures generated in the CSP study match the experimental structure(s) is, therefore, key in assessing the relative abilities of various CSP protocols.

Two commonly employed quantitative methods of crystal structure comparison are the measurement and comparison of inter-atomic distances for a defined cluster size, and comparison of calculated powder X-ray diffractograms (PXRD). The COMPACK algorithm,²² implemented in the CCDC's Mercury software Crystal Packing Similarity (CPS) tool,⁹⁸ is a common example of the former. This approach provides a simple pass/fail metric to identify structure matches within a certain user-defined tolerance for a cluster of M molecules. For matching structures, a quantitative comparison is also provided in the form of a root-mean-square deviation (RMSD) in the atomic positions for the given molecular cluster size. To directly compare $\text{RMSD}(M)$ values with this method, a pass must be achieved (with a consistent cluster size of M molecules) for all structures being compared. Effectively, a smaller RMSD value indicates greater similarity with the reference crystal structure.

Alternatively, the algorithm developed by de Gelder²³ has become popular for comparison of powder diffractograms calculated from crystal structure data. This algorithm uses the normalized integral of a weighted correlation function to give a result between 0 and

1 that quantifies the similarity of the two diffractograms. Two implementations of this scale have been adopted: (i) powder pattern similarity, where larger values (approaching 1) indicate more similar structures,⁷¹ and (ii) powder pattern difference, where smaller values (approaching 0) indicate more similar structures.⁷⁵ Due to the powder difference values being analogous to the RMSD values from the COMPACK algorithm, this metric will be used for PXRD comparison through the remainder of this work.

Developing quantitative comparison methods for crystal similarity is complicated by the innate differences between *in silico* generated structures and real experimental X-ray structures. Structures generated by CSP predominantly correspond to a “static lattice”, neglecting both zero-point lattice vibrations and thermal effects; this is referred to as zeroth-order CSP.⁹⁹ While thermal effects on the lattice can be modeled via molecular dynamics simulations,^{100–102} or through use of the quasi-harmonic approximation,^{103–106} these approaches are extremely expensive computationally and can not be broadly applied across all generated structures in a CSP landscape. One should expect static-lattice structures to have more compact unit cells compared to experimental structures solved from the collection of X-ray diffraction data.¹⁰⁵ They may also potentially exhibit unphysical conformational differences in cases with highly flexible molecules due to neglect of thermal entropy.¹⁰⁷

Use of static-lattice structures will adversely affect structure comparisons using both RMSD and PXRD metrics, although apparent differences are magnified for PXRD as the peak positions are quite sensitive to the changes in cell volume that result from thermal expansion. To address such differences in peak positions, an isotropic volume correction was developed by van de Streek and Motherwell.⁷¹ It uniformly scales the unit cell axes lengths in order to achieve a particular cell volume, which is obtained by summation of the calculated atomic volumes¹⁰⁸ for all atoms in the unit cell. This appears to be the methodology employed to yield the “PXRD similarity” metric in Mercury’s CPS tool.⁹⁸ However, when applied to distinguishing distinct polymorphs from structural re-determinations at differing temperatures in the early 2004 CSD (Cambridge Structural Database), this volume correction was not particularly effective for materials with significant anisotropy in their thermal expansion, which has been noted to be rather common in molecular crystals.^{71,109}

A recent study by Bernstein and co-workers⁶⁸ compared the ability of the COMPACK and PXRD methods implemented in Mercury to differentiate polymorphs from structural

re-determinations in the July 2018 CSD. The two methods were found to be in agreement for 89% of 47,422 pairwise comparisons of structures extracted from the database. The majority of the cases where the methods disagreed arose when the PXRD comparison erroneously indicated differing structures (commonly due to substantial differences in the conditions under which the re-determined data was collected) and the COMPACK method correctly identified a structural match. This implies that PXRD will be less successful than COMPACK when comparing static-lattice structures from zeroth-order CSP to experiment. However, the conformational differences observed in some static-lattice structures of flexible molecules, despite effectively identical packing arrangements, may conversely pose an issue for COMPACK comparisons. For a number of flexible molecules, PXRD comparisons matched structures collected at different temperatures, but COMPACK did not, unless the tolerances on the interatomic distances were increased from their default value of $\pm 20\%$ to $\pm 50\%$.⁶⁸ Overall, the study concluded that relying exclusively on one method has the potential to yield both false positives and false negatives, depending on the structures and the nature of the difference between them.

In this work, we present a simple approach to improve the reliability of PXRD comparisons using an anisotropic volume correction scheme. Our method is targeted to comparisons between zeroth-order CSP candidates and a reference, finite-temperature experimental crystal structure, although it can also be utilized to compare experimental structures obtained at several temperatures. Our method is applied to identify the matching structures from the structure-energy landscape lists submitted to the 6th CSP blind test,⁴¹ and reveals two uncredited matches from that work. The results highlight the improved ability of PXRD comparisons using the anisotropic volume correction to identify structural matches, compared to the isotropic volume correction implemented in the Mercury CPS tool. Anisotropic volume correction is also found to improve RMSD-based comparisons of *in silico* generated structures and experimental structures using COMPACK.

4.2 Dataset

All crystal structures were gathered from the supporting information accompanying the CCDC's 6th blind test (BT6) of CSP methods.⁴¹ Contributors were allowed to submit two lists of up to 100 structures for each of 5 target compounds, labeled XXII-XXVI and shown in Figure 4.1, with 5 target polymorphs (A-E) for compound XXIII. A list of the

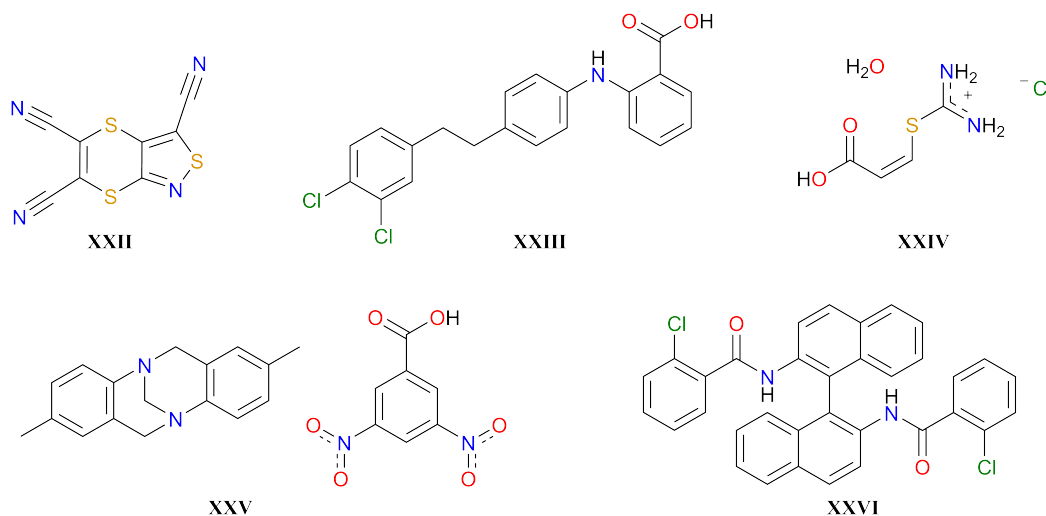


Figure 4.1: The five target compounds used in the CCDC's 6th blind test. Note that there were five target polymorphs (A-E) for compound XXIII, two of which (C,E) have $Z' = 2$.

CCDC identifiers for the target structures is provided in Table A.1.

A total of 115 lists, containing a varied number of structures, were submitted. In the test, 62 structures were identified within these 115 lists that match the corresponding target structure. However, a number of the secondary lists submitted were not different in structure, but simply re-ranked energetically (i.e. from single-point energy calculations with a different method, inclusion of free-energy approximations, etc.). Since the objective of this study is comparison of the generated structures, not their energetic rank, these secondary lists are effectively duplicates and 11 of the secondary lists were removed from the dataset (details can be found in Section A.2). Of the 11 lists removed, 10 contained a matching structure, so the number of “unique” matches was reduced to 52. Throughout this work, references to specific structures will make use of the following notation: *[Target]-[Group]-[List]-[Energy rank]*. As an example, XXII-G18-L2-E5 would be the structure ranked 5th by energy in the second list submitted by Group 18 for target XXII.

We note that the list submitted by Group 12 for target XXII, which did not contain a match to the target, was also omitted. This was due to a number of issues concerning unit-cell dimensions and corresponding crystal system and space group assignments, as well as complete connectivity breakdown of the molecular structure for a number of the candidates contained in the list. A single additional occurrence of a complete molecular difference was identified in list 2 submitted by Group 21 (also for XXII) and this structure was excluded, but the remainder of the list kept. Thus, 103 lists containing a total of 9,104

structures were searched, making 16,532 comparisons, with the expectation of identifying the same 52 hits identified in the original BT6 study.

4.3 Methods

4.3.1 Mercury CPS tool

To compare the developed method to standard alternatives, we performed crystal structure comparisons using the Crystal Packing Similarity (CPS) tool in Mercury⁹⁸ (v2020.1). Results were obtained from the CPS implementations of both (i) the COMPACK²² algorithm (i.e. the number of molecules matched and RMSD(M)) and (ii) PXR similarity.

A cluster size of 20 molecules was used in the COMPACK comparison to identify matching crystal structures. Initial comparisons were made with a tolerance of $\pm 20\%$ on the distances and $\pm 20^\circ$ on the angles. If these tolerances were too strict to obtain a match of 20/20 molecules, the tolerances were increased in increments of 5% and 5° for the distances and angles, respectively, until such a match was achieved, provided the structures continued to overlay in reasonable visual agreement. If an increase in the tolerance was accompanied by a dramatic change in the structural overlay, then the loosening of tolerances ceased and it was concluded that obtaining a representative RMSD(20) value was not possible for that structure. Notably the RMSD(20) values between submitted and target structures were found to differ moderately from previously reported values in BT6⁴¹ (Figure A.1). The RMSD values calculated in the current version of Mercury are those reported throughout this study.

For COMPACK comparison, hydrogen-atom counts and bond counts for each atom were ignored. These optional selections were important for comparison of structures of compound XXIII [2-((4-(3,4-dichlorophenethyl)phenyl)amino)benzoic acid]. Here, the carboxylic acid moiety can be rotated by a full 180° to yield a different conformer, without otherwise affecting the crystal packing as all COOH groups form two strong hydrogen bonds with the COOH of a neighbouring molecule in the lattice. While H atom count (bonded to an atom) is considered by default, the H atom positions are not considered as they are regularly refined by applying constraints instead of being solved from the electron density; thus, in general, both possible proton orderings should be counted as structural matches. A comparison of the number of molecules in common (#/20) and RMSD (1)/(20) values obtained with and without the selection of these options is given in Table A.4 for

structures where this had an effect.

There is little documentation regarding how PXRD similarity values are determined from the CPS tool, although it appears that an isotropic volume correction procedure, similar to that outlined by van de Streek and Motherwell,⁷¹ is used. In the work done by Bernstein and coworkers,⁶⁸ they report that the powder diffractograms were calculated using ideal Cu $K\alpha_1$ radiation (1.54056 Å) and Pseudo-Voigt peak shapes from $0 - 50^\circ 2\theta$. The diffractograms were then compared with de Gelder's cross-correlation function to yield a similarity value (1 being identical). The resulting PXRD similarity values are subtracted from 1 to convert them into powder pattern difference values, facilitating comparison with results from our `critic2` program.⁷⁵

4.3.2 Newly developed VC-PWDF code

To implement an anisotropic volume correction, we have developed a `bash` script to be run from the command line by Linux OS. The `vc-pwdf` code is available from github¹¹⁰ and interfaces with the latest version of `critic2`.⁷⁵ It automates a protocol of unit-cell reduction, screening by unit-cell parameters, and performing the volume correction, followed by powder diffractogram comparison. The code has been designed for application to a set of candidate structures resulting from CSP; it currently accepts as input CSP structure lists (e.g. submission to BT6) and a target reference structure, both in `.cif` format. However, it should be noted that the code is also applicable to pair-wise comparison of only two given structures.

The required inputs are:

- A single file that contains geometries of all the candidate structures to be compared to the reference target structure
- A reference target structure

The algorithm undertaken by the code is as follows:

1. Split the catenated `.cif` into separate files for each candidate structure.
2. Convert each structure, including the reference, to its Niggli reduced cell^{111,112} (using `NEWCELL PRIMITIVE` and `NEWCELL NIGGLI` sequentially in `critic2`).

3. Compare unit-cell dimensions of each candidate structure to those of the reference structure to identify which are potential matches.
 - (a) Eliminate candidate structures where the volume is not within a given threshold (default 20%) of the reference structure.
 - (b) Eliminate structures where each axis length is not within a given threshold (default 20%) of one of the three axis lengths of the reference structure.
 - (c) Eliminate structures where the crystal system (triclinic, monoclinic, orthorhombic, tetragonal, hexagonal, or cubic) does not match that of the reference structure.
 - (d) Eliminate structures that do not possess the same space-group symmetry as the reference. This screening criterion can be toggled off by the user if, upon review of the log file, there is concern that one or more structure(s) have similar unit-cell dimensions as the target but failed the space-group match. All data shown here were obtained using space-group screening. If this screening step is omitted, no additional matching structures were found, although one additional XXII polytype structure was identified.
 - (e) Each candidate structure that makes it to this stage undergoes a number of transformations in order to account for possible inconsistencies in the unit-cell description with respect to the target structure. The transformation matrices are applied via `critic2` with `NEWCELL [matrix]`. Each transformation generates a new structure file that is carried through the remainder of the protocol. Only the structure file with the smallest VC-PWDF value out of all the variations generated for that candidate structure is kept at the end.
 - i. A check of the unit-cell axes is performed. If any of the candidate structures have two axes within 1 Å of each other, it is deemed possible that the axes may be swapped relative to the reference structure (e.g. the *a*-axis vector of the candidate structure's unit cell matches the *b*-axis vector of the reference structure's unit cell). The transformation matrix that interchanges the axes of interest is then applied to the candidate structure, generating an additional structure with these axes interchanged that is carried through to the next steps of the algorithm. Interchanging axes was

necessary to identify 6/52 of the original BT6 matches.

ii. Additional structure files are generated using linear combinations of the unit-cell vectors, and combined linear combination and axes swaps. This compensates for cases where a candidate structure and the target will have incompatible lattice-parameter definitions, even after Niggli reduction (see Figure 4.2 for an example). Three sets of transformation matrices are used depending on the case. One set of 24 matrices is used for triclinic unit cells with acute angles. Another set of 24 matrices is used for the obtuse-angle triclinic unit cells, and a subset of 12 of these 24 matrices is used for monoclinic cells (which, by definition, must have an obtuse non-right angle). These additional structures are carried through to the next steps of the algorithm. Details regarding the sets of transformation matrices are available in Section A.3. Applying transformation matrices was necessary to identify 6/52 of the original BT6 matches.

(f) A check of the angles is performed, comparing those of the candidate structures (and additional transformed structure files) to the reference structure. If an angle is 90° in the reference structure, but not also 90° for the candidate structure, the structure is eliminated (most relevant for monoclinic structures).

4. Apply the anisotropic volume correction. This is done by replacing the unit-cell dimensions (cell lengths and angles) of the candidate structure with those of the reference cell. This replacement of the unit-cell vectors is done within a .res file format, where the atomic positions are given in fractional coordinates. Thus, the volume correction will cause a distortion of the molecular geometries, but only marginally (*vide infra*).

5. Compare computed powder diffraction patterns of the candidate structures with the reference structure using the COMPARE keyword in `critic2`. Powder diffractograms are generated from $5 - 50^\circ 2\theta$ and compared with de Gelder's cross-correlation function to yield the dissimilarity value (with a value of 0 indicating identical structures). The output consists of two ranked lists of powder difference values from comparison of the candidate structures, before and after volume correction, with the reference (examples are shown for both in Section A.6).

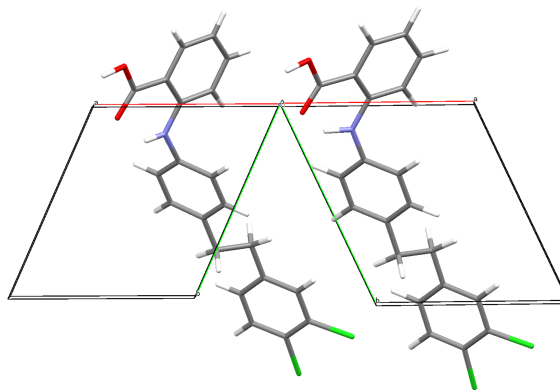


Figure 4.2: Comparison of the unit cells of (left) Group 09's matching structure (XXIIIB-G09-L1-E13) and (right) the experimental structure of target XXIII form B, viewed in the *ab* plane. Application of the $[-1\ 0\ 0]$, $[-1\ 1\ 0]$, $[0\ 0\ -1]$ transformation matrix to the G09 structure is required for its *b*-axis vector to align with the *b*-axis vector of the target, allowing the volume correction to be properly applied.

For all candidate structures, the results of the protocol are output to a log file to explain if/how the structures were modified and why structures were eliminated. The ordering of the screening steps is meant to run from least to most stringent, permitting the greatest number of structures to be carried forward at every step. This allows the user to track a structure through the screening. Eliminated structure files are removed from the working directory, leaving only matching structure files (the parent file containing all the CSP-generated structures remains unedited).

4.3.3 Similarity value notation

The following terminology will be used to discuss the different values generated by the different comparison methods:

A **raw-POWDIFF** value is the result of PXRD comparison between the reference structure and a candidate structure without any volume correction, using the COMPARE functionality in `critic2`.

A **VC-PWDF** value is the result of PXRD comparison between the reference structure and a candidate structure after anisotropic volume correction, using the algorithm described in Section 4.3.2.

A **CPS-PWDF** value is the result of PXRD comparison between the reference structure and a candidate structure after isotropic volume correction, using the Mercury CPS tool. The PXRD similarity value yielded by Mercury is converted to CPS-PWDF by subtracting the result from 1.

A **raw-RMSD(1)** value is the result of COMPACK comparison between the reference structure and a candidate structure without any volume correction for a cluster size of one molecule.

A **VC-RMSD(1)** value is analogous to the above, but using the candidate structure after application of the developed anisotropic volume correction.

A **raw-RMSD(20)** value is the result of COMPACK comparison between the reference structure and a candidate structure without any volume correction for a cluster size of 20 molecules.

A **VC-RMSD(20)** value is analogous to the above, but using the candidate structure after application of the developed anisotropic volume correction.

4.4 Results

4.4.1 PXRD comparison

Powder difference values for the full dataset were obtained using the COMPARE keyword in `critic2`. The full histogram of raw-POWDIFF values in Figure 4.3(a) displays a normal Gaussian distribution. Figure 4.3(b) shows an expanded view of this histogram, highlighting the BT6 matches, which have raw-POWDIFF values ranging from 0.03 – 0.54. When isotropic volume correction⁷¹ is applied to the dataset, a skewed distribution of the resulting CPS-PWDF values is observed in Figure 4.3(c). A histogram of the 0 – 0.05 CPS-PWDF range (considered to be a relatively small value^{68,71}) is shown in Figure 4.3(d). The distribution of the 52 matches identified in BT6 is again quite broad, spanning this range and beyond, with 6 matching structures having CPS-PWDF values > 0.05.

The newly developed code for anisotropic volume correction was also applied to the dataset. The distribution of VC-PWDF values, for the structures that pass the unit-cell screening (Step 3 described in Section 4.3.2), is shown in Figure 4.3(e). The VC-PWDF values for the 52 BT6 matches are reduced by roughly an order of magnitude, compared to the raw-POWDIFF values. They now fall into the 0 – 0.05 range for all but two cases: XXII-G09-L1-E02 (raw-POWDIFF of 0.5363 and VC-PWDF of 0.1120) and XXIIIB-G13-L1-E88 (raw-POWDIFF of 0.2783 and VC-PWDF of 0.0546). These two structures will be discussed in more detail in Sections 4.5.3 and 4.5.4.

The distribution of the VC-PWDF values for the BT6 matches in Figure 4.3 sharply contrasts with the CPS-PWDF results. As shown in Figure 4.3(f), there is a decay in

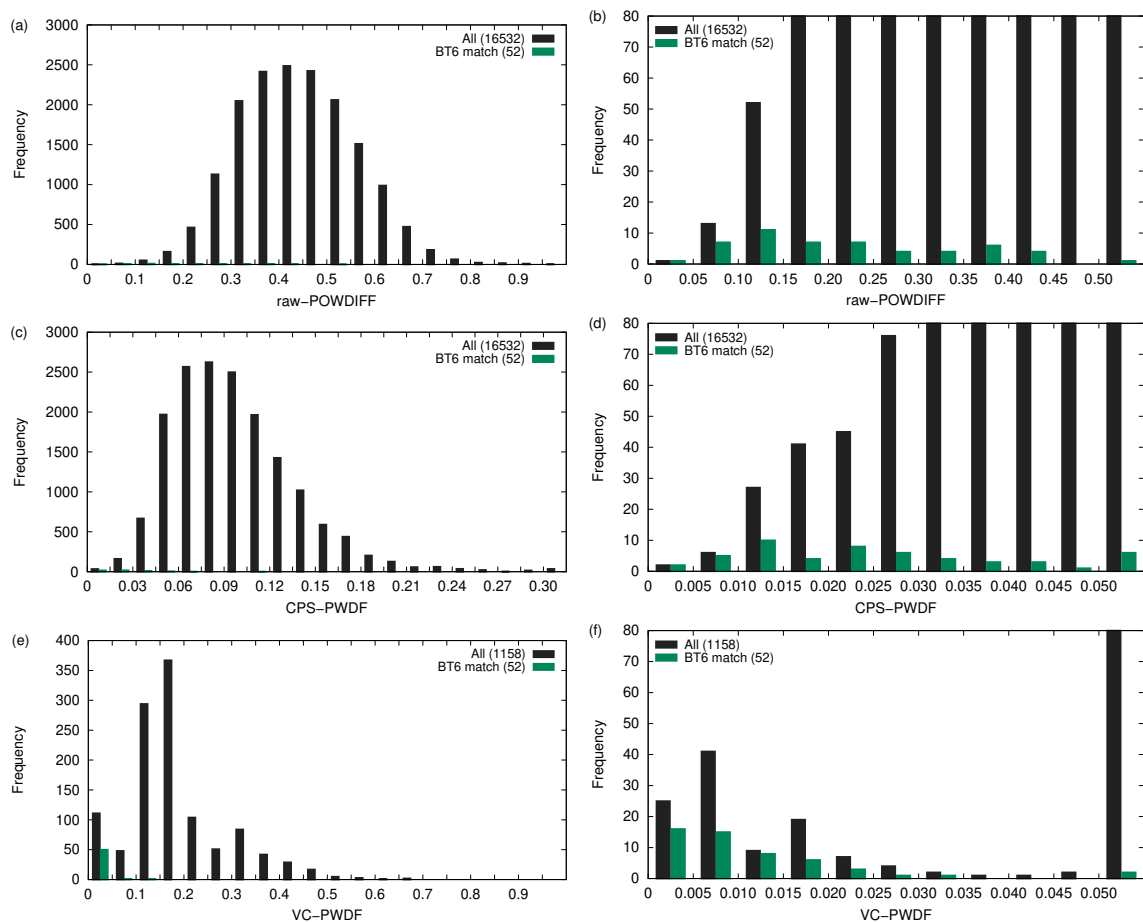


Figure 4.3: Histograms showing the distribution of powder difference values obtained using the various comparison methods for the full dataset (black bars), and matches identified in the 6th blind test (green bars), within relevant ranges. Shown are raw-POWDIFF values for the unmodified structures from critic2 (a,b), CPS-PWDF values after isotropic volume correction from Mercury (c,d), and VC-PWDF values after anisotropic volume correction from critic2 (e,f). Note the differences in x -axis scale. Powder difference values range from 0–1; any data points with values surpassing the x -axis range are included in the final bin.

the number of structures as the VC-PWDF values increase from 0, to 0.035, and only a couple matching structures with VC-PWDFs between 0.035 – 0.05. Thus, the volume correction provides a natural segregation between possible structure matches and other candidate structures. A detailed breakdown of all structures identified by our algorithm to have VC-PWDFs of < 0.05 , and thus be likely structure matches, will be presented in Section 4.4.2.

4.4.2 Analysis of additional VC structure matches

None of the three PXRD comparison methods clusters only the 52 matches within the lowest powder difference bins, segregated from all of the other structures (Figure 4.3). However, the VC-PWDF histogram clearly stands out in having some ability to group the BT6 matches, with a total of only 111 structures having VC-PWDF values less than 0.05 (although this range misses 2 of the 52 matches identified in BT6). The reasonable number of candidates in the VC-PWDF 0 – 0.05 range makes it possible to analyze all of these structures to determine if additional matches were found.

The majority of the additional structures (47/61) are duplicate matching structures. These are structures that match the target structure, but were included in a list that already contained one of the 52 identified matches. It was noted in the BT6 competition that, if there were duplicates within a list, the matching structure with the lowest energy would be chosen for the energy ranking in the results table. The bulk of the duplicate structures are part of the two lists submitted by group 23 for target XXIII, and match form B (29/47). This is interesting as Group 23 re-optimized a sub-set of the force-field¹¹³ structures generated by Group 18 using either HF-3c¹¹⁴ (list 1) or TPSS-D3^{115,116} (list 2). Thus, unique structures generated by the force field converged to the same structure when optimized with the quantum-mechanical methods, since this duplication is not observed in the lists provided by Group 18.

Figure 4.4 shows the distribution of the full set of 113 structures either yielding VC-PWDF values less than 0.05, or identified as a match in BT6. The majority (98/113) of the structures are classified as matches, including duplicate matches. Notably, two of these (non-duplicate) matching structures were missed in BT6 (see Section 4.5.2). A further 10 structures were identified as polytypes of compound XXII. While they are not proper matches, they possess a fairly similar packing to the reference compound and will be discussed further in Section 4.5.1. Three structures were found to have significant conformational differences from the reference, but would be expected to be close matches upon geometry relaxation with a quantum-mechanical method, such as dispersion-corrected density-functional theory. One of these structures was identified as a BT6 match (XXIIIB-G13-L1-E88), while the other two were part of a list that already contained a match identified in BT6 (see Section 4.5.3). One structure (XXV-G15-L1-E24) was found to have a slightly differing packing than the target due to a 180° rotation of

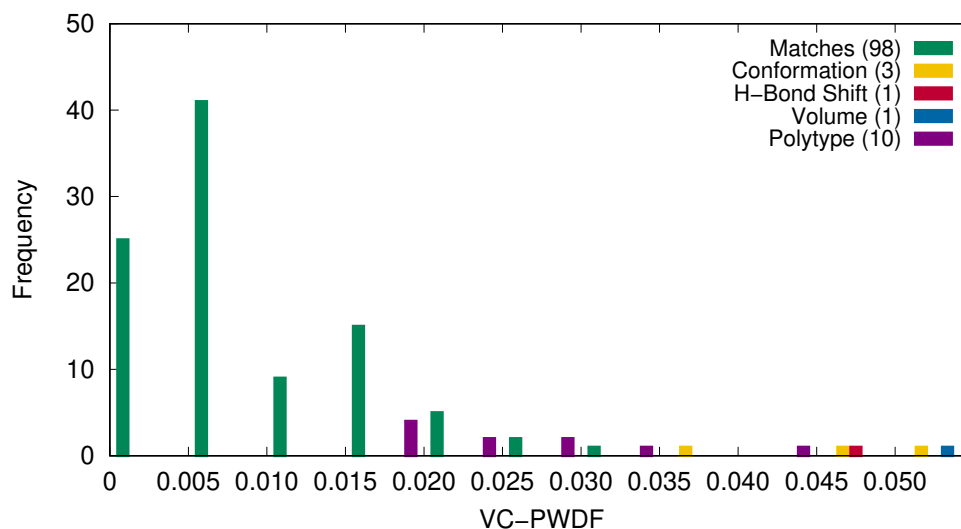


Figure 4.4: Classification of the 113 structures with VC-PWDF values less than 0.05, or identified as a match in the 6th blind test. Powder difference values range from 0–1; any data points with values surpassing the x -axis range are included in the final bin. The XXII polytypes, as well as cases with significant differences in conformation/H-bond alignment or volume, are discussed in Sections 4.5.1, 4.5.3, and 4.5.4, respectively.

the 3,5-dinitrobenzoic acid COOH group (see Section 4.5.3). Finally, one BT6 match (XXII-G09-L1-E02, see Section 4.5.4) has a volume that is anomalously large compared to the reference structure, and is displayed separately on the histogram (and discussed further in Section 4.5.4).

4.4.3 COMPACK comparison

Mercury’s CPS tool was used for COMPACK comparison of all 113 structures with a VC-PWDF less than 0.05, or identified as a match in BT6 (i.e. 111 structures with VC-PWDF values < 0.05 , plus XXII-G09-L1-E02 and XXIIIB-G13-L1-E88). RMSD(1) values were computed for the 98 $Z' = 1$ structures within this set. On average, the volume correction resulted in a negligible difference in the VC-RMSD(1) values compared to the raw-RMSD(1) values of the unmodified structures (see Figure A.4). In 55/98 cases, there was actually a slight improvement in the RMSD(1) value with the application of the volume correction.

The distribution of the raw-RMSD(20) values for the 113-structure dataset is shown in Figure 4.5(a). The classified matches from Figure 4.4 are now subdivided into two groups: those that are COMPACK matches with the default tolerances ($\pm 20\%$ and $\pm 20^\circ$) and those that required looser tolerances. This latter group is labeled as “CPS-tolerance” in

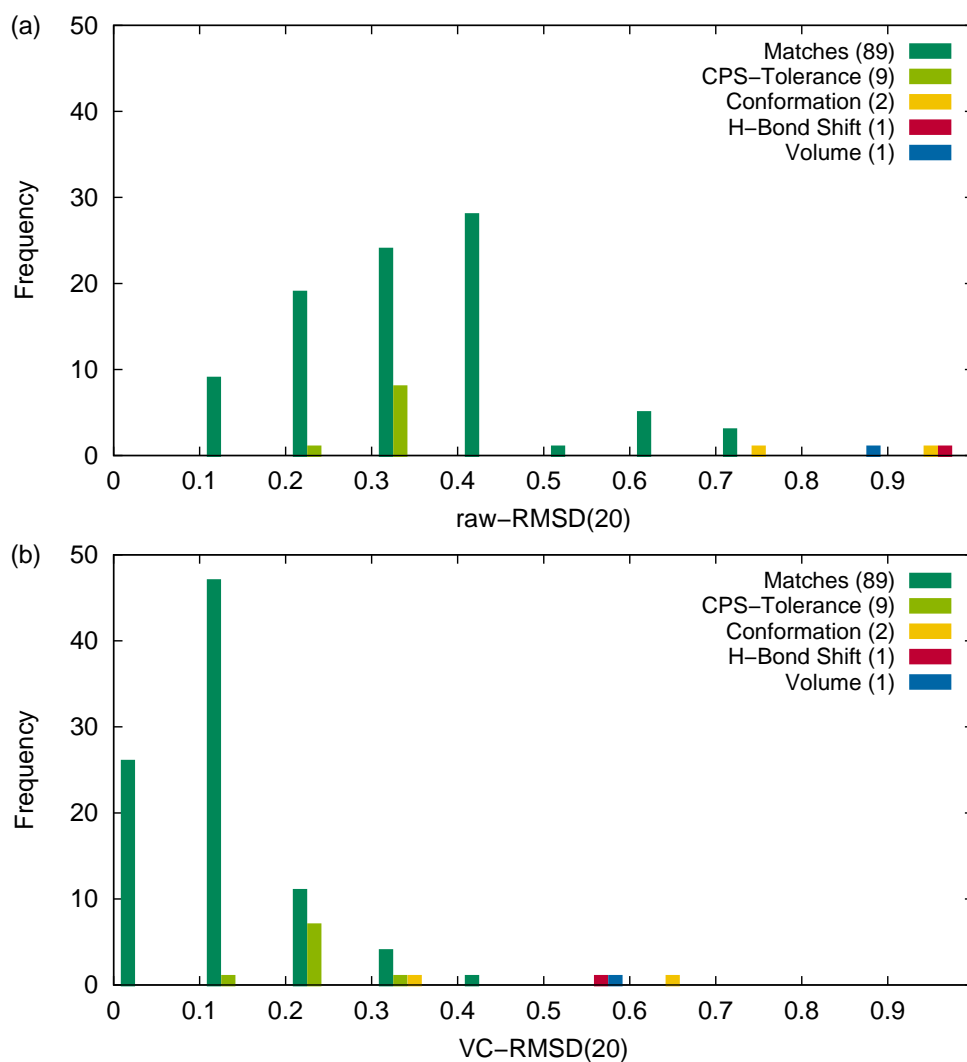


Figure 4.5: Histograms showing the distribution of RMSD(20) values (in Å) for the 102/113 structures plotted in Fig. 4.4 that yield a viable 20/20 molecule match. Shown are raw-RMSD(20) values (a) and VC-RMSD(20) values (b). The final far-right bin in the raw-RMSD(20) distribution (a) includes all values larger than 0.9 Å. “CPS-tolerance” indicates structures where the COMPACK tolerances had to be loosened from their default values. The 10 XXII polytypes and one “conformational” structure are excluded as no valid 20/20 molecule match is possible for any tolerance.

Figure 4.5. Overall, the tolerances had to be increased for a total of 13 structures, including four BT6 matches. Two of the BT6 matches required tolerances looser than the $\pm 25\%$ and $\pm 25^\circ$ used in that work.⁴¹ Two of the structures indicated in Figure 4.4 as exhibiting significant conformation differences, and the structure exhibiting a significant volume difference, from the target also yielded viable 20/20 molecule matches once the tolerances were loosened. However, for the other “conformational” structure, a reasonably overlaid

20/20 molecule match could not be achieved at any tolerance.

VC-RMSD(20) values (calculated using the structures output from the anisotropic volume correction) were also determined and the distribution of these values is shown in Figure 4.5(b). The same 13 structures still required loosening of the tolerances to match all 20 molecules of the cluster, and there appears to be no correlation between the required tolerance and the resulting RMSD(20) values (Figure A.3). As shown in Figure 4.5, the range of RMSD(20) values is nearly halved upon volume correction, compared to the results for the unmodified structures. All but three structures have a VC-RMSD(20) less than 0.5 Å. Because volume difference is no longer a contribution to the calculated VC-RMSD(20), a much tighter grouping of the matching structures is observed at lower values. This demonstrates the developed volume correction's improvement of COMPACK, as well as PXR, structure comparison.

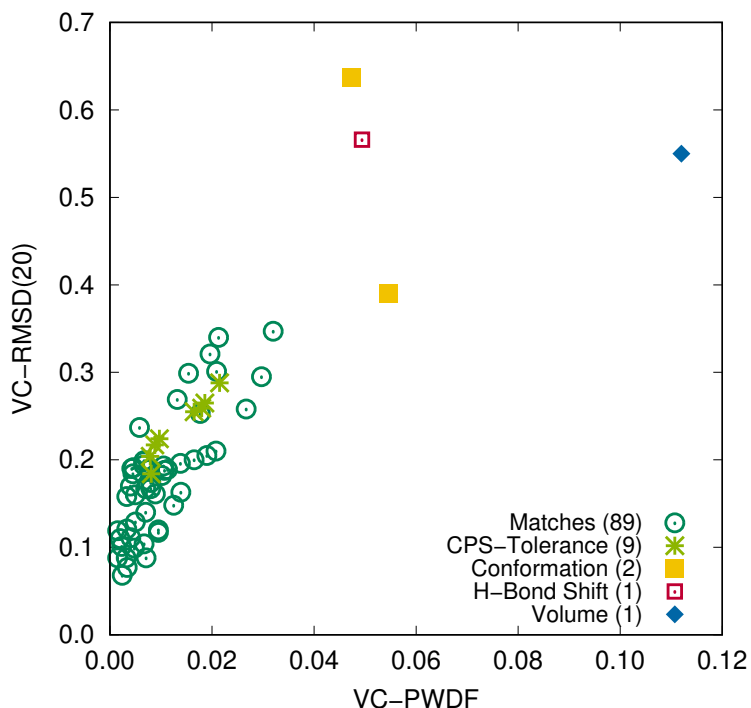


Figure 4.6: Plot of VC-RMSD(20) versus VC-PWDF for all BT6 matches and other structures with VC-PWDFs <0.05. Structures that do not yield RMSD(20) values with reasonable structure overlap are not included (the 10 XXII polytypes, as well as XXIIIC-G14-E25).

Finally, Figure 4.6 shows a good correlation between the VC-PWDF and VC-RMSD(20) values. This scatter plot clearly distinguishes closely matched structures from the two “conformational” structures and the shifted structure with a different H-bond alignment.

Analogous plots involving the raw-POWDIFF or CPS-PWDF values show considerably worse correlations and lose the distinct groupings of structure types (Figure A.5). We find that VC-PWDF values are arguably as useful as RMSD(20) in providing a quantitative similarity comparison of two crystal structures. The VC-PWDF value can even represent an improvement over RMSD(20) in some cases, as it does not require varying tolerances to identify matching structures. We view these as complementary metrics that can be used most effectively in combination to prevent omission of any matching structures generated by CSP.

4.5 Discussion

4.5.1 XXII polytypes

Ten structures were submitted in lists from different groups for compound XXII that are not a match to the target and do not yield a viable RMSD(20) value, but all match each other. This common structure can be viewed as a polytype of the target. Figure 4.7 shows the experimental XXII structure overlaid with the polytype (XXII-G03-L1-E56 is used as a representative example) to highlight the considerable similarity in the packing. When viewed in the *bc* plane (left), the molecules appear to align perfectly. However, when rotated and in the *ac* or *ab* planes (center and right, respectively), the difference in the packing of the two structures is revealed. If one considers there to be two rows of molecules in the *ab* plane, then the bottom row of molecules match perfectly in both structures; however, the top row is translated by half of the *b*-axis length. Similarly, viewing the unit cell in the *ac* plane, the left column of molecules is not properly overlaid and is instead translated by half the *c*-axis length. This considerable packing similarity is identified by the powder difference methods, but not by the COMPACK algorithm, which will fail at 9/20 molecules matched for most tolerances.

Figure 4.8 shows histograms of CPS- and VC-PWDF values for the 13 matches and 10 polytype structures identified for target XXII. CPS-PWDF is unable to distinguish the matches from the polytypes and the histogram shows a complete intermingling of the two categories. In contrast, the VC-PWDF results show a segregation of the matches from the polytypes, with a single bin (0.015–0.020) occupied by one matching structure and 4 of the polytype structures. Thus, the VC-PWDF method clearly does much better than the CPS-PWDF method at separating these two classes of structures, despite their very strong

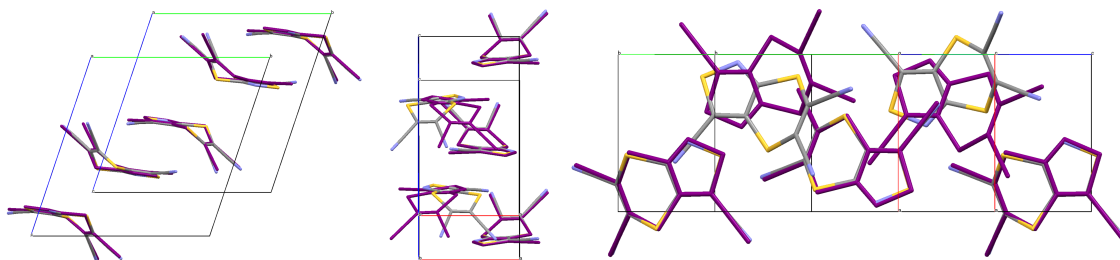


Figure 4.7: Overlay comparing the packing of the polytype (XXII-G03-L1-E56, shown in purple) with the target structure of compound XXII in the bc plane (left), the ac plane (center), and the ab plane (right).

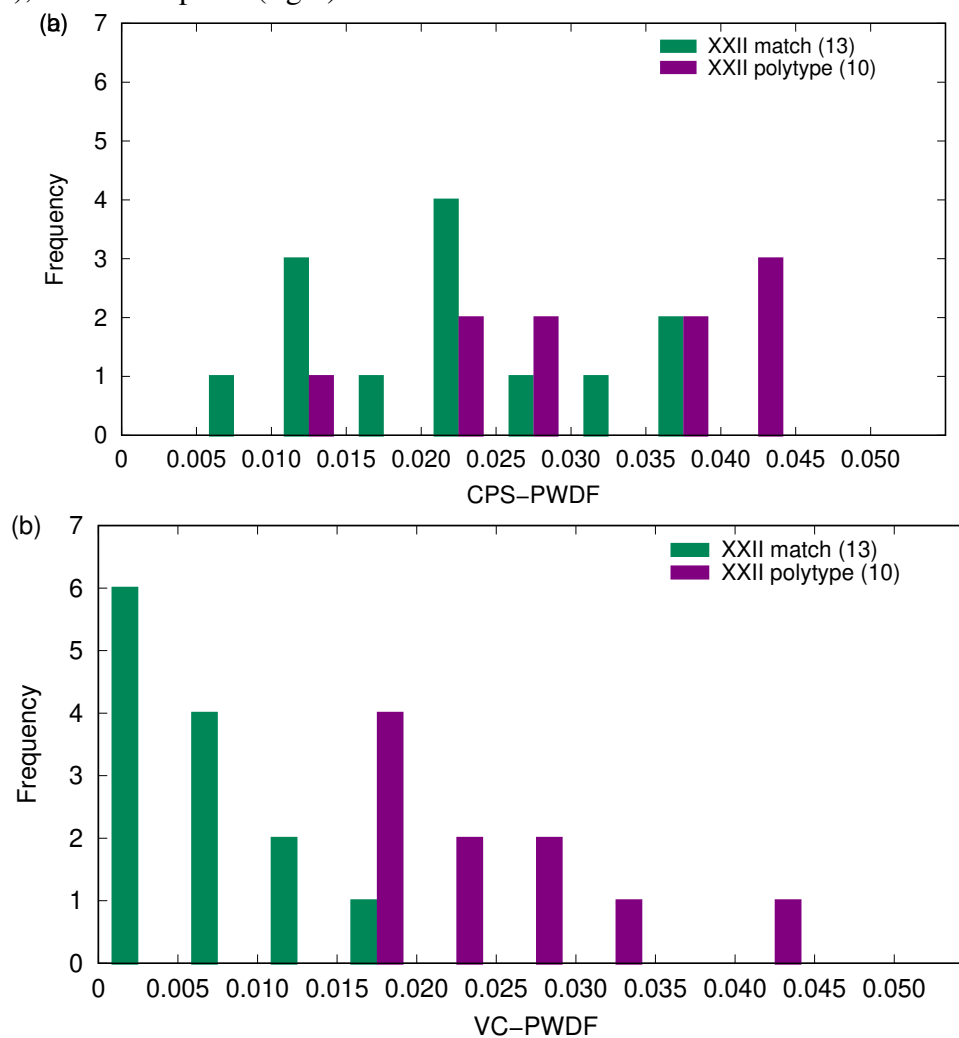


Figure 4.8: Histogram of the CPS-PWDF (a) and VC-PWDF (b) values for the 13 matches and 10 polytype structures identified for target XXII.

similarity.

This example showcases the requirement for flexible cutoffs depending on the system

in question. While a VC-PWDF threshold of 0.035 is needed to include most matches identified for the full set of BT6 compounds (Figure 4.4), a tighter threshold is clearly needed for the rigid compound XXII. Here, an optimal choice of 0.017 for a VC-PWDF threshold would actually result in complete separation of the polytypes from the structure matches (although this is clearly specific to compound XXII).

4.5.2 Extra matches not identified in BT6

With our anisotropic volume correction, two structural matches are identified that were missed in BT6: XXIII A-G09-L2-E19 and XXV-G06-L1-E08. Overlays of these two structures with their respective targets are shown in Figure 4.9. Both are classified as matching structures in Figure 4.4 and XXV-G06-L1-E08 falls into the “CPS-tolerance” group in Figures 4.5 and 4.6.

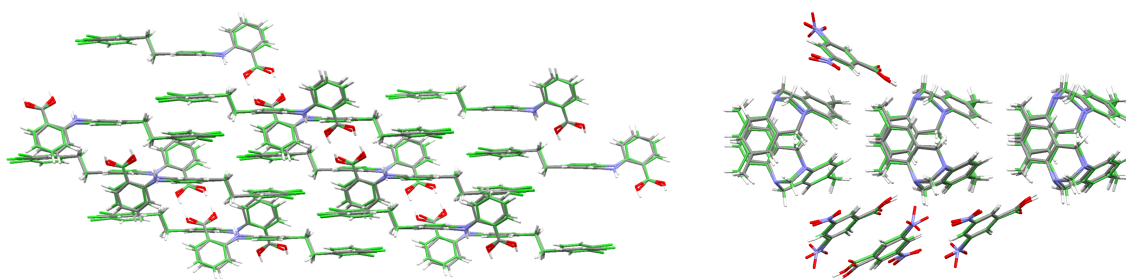


Figure 4.9: Overlays of the packing of XXIII A-G09-L2-E19 (left) and XXV-G06-L1-E08 (right), after anisotropic volume correction, with their respective target structures.

First, XXIII A-G09-L2-E19 was identified as a likely match to the target, with a VC-PWDF of 0.0297. As shown in the left panel of Figure 4.9, there is a 180° difference in the orientation of the COOH group between this candidate structure and the target. However, proton exchange across this COOH–HOOC doubly hydrogen-bonded dimer does not otherwise alter the crystal packing. Assignment of H atom positions from residual electron density is challenging, requiring high quality data, and is therefore regularly done by applying constraints instead. As such, the two structures should be considered equivalent (despite the obvious minimum-energy conformer). A 20/20 molecule match with a raw-RMSD(20) of 0.551 \AA (VC-RMSD(20) of 0.295 \AA) can be obtained with the default COMPACT tolerances if H-atom and bond counts are ignored, confirming this as a structural match. However, if these factors are considered in determination of a structural match, the tolerances must be increased to 30% and 30° , which are higher than

the thresholds used in BT6. While two structures with this same COOH rotation were identified in BT6 (XXIIID-G06-L1-E73 and XXIIID-G09-L1-E66), indicated in the results table with their energy rank in brackets to denote this difference, XXIIIA-G09-L2-E19 was not.

Next, XXV-G06-L1-E08 was also identified as a likely match to the target structure, with a VC-PWDF value of 0.0213. As shown in the right panel of Figure 4.9, the main difference between XXV-G06-L1-E08 and the target structure is a rotation of one of the 3,5-dinitrobenzoic acid nitro groups. In the XXV-G06-L1-E08 structure, one of the nitro groups is rotated 60° out the plane of the benzene ring, while both nitro groups lie in plane in the experimental structure to maximize conjugation. The deviation from planarity is likely the result of Group 06's use of the MMFF94 force field¹¹⁷ for the intramolecular degrees of freedom during geometry optimization.⁴¹ Refinement with a quantum-mechanical method would be expected to restore the planarity of the 3,5-dinitrobenzoic acid. Unfortunately, this structure was not in the top 50 chosen for subsequent reoptimization with PBE-XDM when Group 06 generated their second list, according to the SI provided with BT6.⁴¹

XXV-G06-L1-E08 was missed in the original publication of BT6 results due to the inherent functionality of the COMPACK algorithm to weigh heavily on conformation. Were this an amine rather than a nitro group, the COMPACK algorithm would likely have classified this as a match (by default, H atoms are excluded in the inter-atomic measurements of distances and angles). As the unphysical conformation change in the candidate structure does involve a nitro group, the large deviations in oxygen positions result in failure to achieve a 20/20 molecule match until the tolerances are increased to $\pm 60\%$ and $\pm 60^\circ$. At this point, a perfectly agreeable raw-RMSD(20) value of 0.363 Å is obtained.

These two examples illustrate the danger inherent in using COMPACK alone to determine structure matches when evaluating CSP methods. Pairing COMPACK with PXRD methods is necessary to avoid missing structural matches with differing proton assignments, or with conformational differences that may result from using low levels of theory for geometry relaxation in the early steps of CSP.

4.5.3 Grey areas in structure comparison

We now consider the outliers shown in Figures 4.4 and 4.6, with VC-PWDF values above 0.035. In this section, we focus on the three “conformational” cases (XXIIIB-G13-L1-E88, XXIIIC-G14-L2-E25, and XXVI-G14-L1-E25) and the one “H-bond shifted” case (XXV-G15-L1-E24). Metrics quantifying the similarity of each of these structures with their respective targets are collected in Table 4.1. The two $Z' = 1$ entries have the two largest RMSD(1) values seen for the entire dataset of 113 structures, indicating the greatest conformational differences from the target. All other matches have RMSD(1) values well under 0.3 Å.

Table 4.1: Selected RMSD and powder difference comparison measures for four borderline cases in which the molecules display notable conformational differences relative to the target, or a notable positional shift is observed. RMSD values are given in Å. The cases in which no RMSD(1) value is reported have $Z' = 2$.

Structure	RMSD(1)	RMSD(20)	VC-RMSD(20)	raw-POWDIFF	CPS-PWDF	VC-PWDF
XXIIIB-G13-L1-E88	0.339	0.758	0.390	0.2783	0.0593	0.0546
XXIIIC-G14-L2-E25	N/A	N/A	N/A	0.2305	0.0197	0.0357
XXV-G15-L1-E24	N/A	0.949	0.566	0.3314	0.0563	0.0494
XXVI-G14-L1-E06	0.561	1.522	0.637	0.3037	0.0444	0.0473

As shown in the left panel of Figure 4.10, XXIIIB-G13-L1-E88 has a significant conformational difference with the target. However, the overlap of the molecular positions in the packing remains nearly the same and this structure was identified as a match in BT6. A reasonable overlay of all 20 molecules can be made once the tolerances are loosened to ± 30 (%) and $^\circ$) for the raw structure, or ± 25 (%) and $^\circ$) for the volume-corrected structure. Despite this, the PXRD methods give fairly large CPS-PWDF and VC-PWDF values (see Table 4.1), indicating a less similar structure than most of the BT6 matches.

XXVI-G14-L1-E06, shown in the right panel of Figure 4.10, is a duplicate match on the list submitted by Group 14 (the BT6 match is XXVI-G14-L1-E01). This structure also shows considerable conformational differences compared to the target. COMPACK only identifies a match when the tolerances are loosened to ± 45 (%) and $^\circ$) for the raw structure, or ± 30 (%) and $^\circ$) for the volume-corrected structure. Here, the PXRD methods once again give relatively large CPS-PWDF and VC-PWDF values, as listed in Table 4.1.

Structure XXIIIC-G14-L2-E25 is another duplicate match submitted by G14 (the BT6

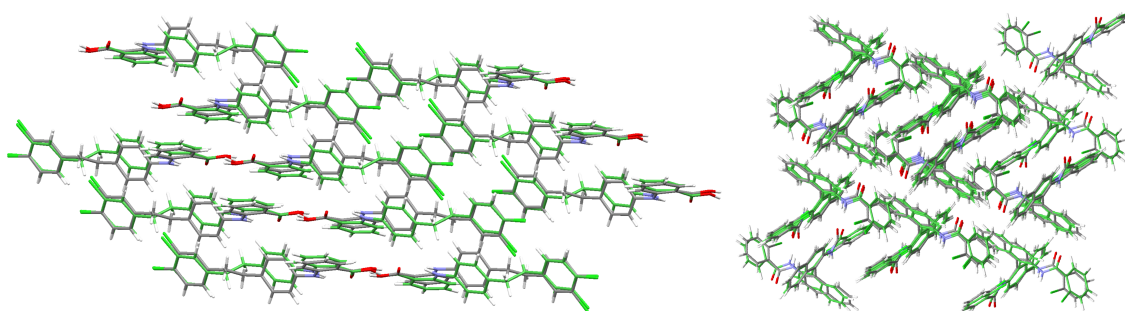


Figure 4.10: Two examples of structures with notable conformational differences with their target, but are successfully overlaid in a cluster of 20 molecules with COMPACK. Left: BT6 match for XXIII B from Group 13 (XXIII B-G13-L1-E88). Right: A duplicate on the list submitted by Group 14 for target XXVI (XXVI-G14-L1-E06).

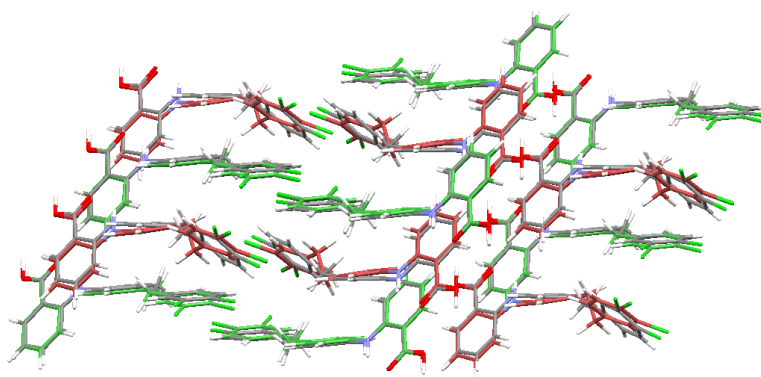


Figure 4.11: Overlay of target XXIII form C with XXIIC-G14-L2-E25, with default tolerances, showing matching molecules in green. Non-matching molecules are shown in red, and possess a conformational change in the terminal dichloro-phenyl moiety.

match is XXIIC-G14-L2-E06). The volume-corrected PXRD methods predict this structure to have strong packing similarities to the target, with CPS-PWDF and VC-PWDF values that are notably lower than the two examples showcased above (see Table 4.1). Despite this, half of the molecules have a visually distinguishable difference in conformation from the target, as shown in Figure 4.11, that prohibits determination of a viable RMSD(20) value. As noted previously,⁶⁸ and showcased by the missed BT6 match for compound XXV, COMPACK weighs heavily on molecular conformation when assessing crystal-structure matches. In some cases, structures with clear conformational differences fail to achieve a “pass” from the COMPACK algorithm, despite having the same packing as the target. At looser tolerances, a reasonable 20/20 molecule overlay can be made; however, not all molecules pass according to the tolerance given. If tolerances are loosened further, then the overlay is distorted and becomes unreasonable. XXIIC-G14-L2-E25 is

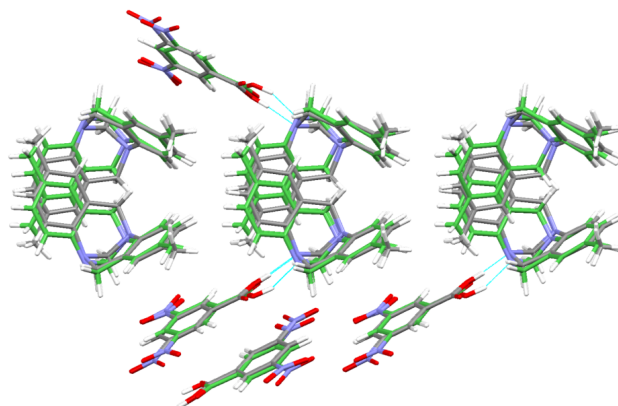


Figure 4.12: Overlay of the volume-corrected XXV-G15-L1-E24 and XXV target structures, with H-bonds highlighted.

an example where a 20/20 match cannot be made up to a tolerance of 75 (% and °). Once the tolerance is loosened to 80 (% and °), the overlap becomes unreasonable, the number of matching molecules in the cluster decreases, and the RMSD value jumps.

Finally, XXV-G15-L1-E24 shows a fairly poor overlay with target XXV in Figure 4.12, commensurate with the large RMSD(20) values in Table 4.1. XXV-G15-L1-E01 is the BT6 match from this list, while XXV-G15-L1-E24 differs in the COOH proton position (i.e. 180° rotation of the COOH group relative to the target). This leads to a visible rotation of the 3,5-dinitrobenzoic acid molecules and some shifting of the Tröger's base to accommodate the intermolecular H-bonding. This structure is not a proper match to the target, as the 180° rotation of the COOH group would prohibit optimization to the same energy minimum as the target with either force fields or quantum-mechanical methods. A tolerance of ± 35 (% and °) was required to obtain a 20/20 match for XXV-G15-L1-E24, ignoring H-atom and bond counts. COMPACK is able to achieve a match for XXV-G15-L1-E24 at a tolerance nearly twice as strict as that required to match XXV-G06-L1-E08 (the missed match for target XXV), even though the missed match has virtually identical packing to the target. Conversely, the VC-PWDF value for XXV-G15-L1-E24 is more than twice as large as XXV-G06-L1-E08, reflecting the significant difference in packing between XXV-G15-L1-E24 and the target.

These four examples showcase the grey areas of crystal structure comparison. The differences in molecular conformation result in larger RMSD(20) or powder difference values than seen for other matches. The three conformational structures for compounds XXIII and XXVI could still be considered matches since they would be expected to relax to

the same energy minimum as the experimental structure upon full geometry optimization. However, the “H-bond shifted” structure for XXV would not optimize to an identical structure as the target, due to the 180° rotation of the COOH group causing a difference in 3,5-dinitrobenzoic acid orientation to maintain intermolecular H-bonding. These examples also illustrate that larger cutoffs for VC-PWDFs may be used for flexible molecules, such as XXIII and XXVI, where intramolecular conformation differences are common. For rigid molecules, such as the components of the XXV co-crystal, a smaller VC-PWDF cutoff is likely required to weed out non-matching structures with similar packing.

4.5.4 Poor candidate geometries

A final interesting case is that of XXII-G09-L1-E02, which has by far the greatest volume difference, at 14.7% *larger* than the target. The next greatest volume difference occurs for the XXIIIB-G13-L1-E88 “conformational” match, which has a volume 9.2% larger than the target. For comparison, the mean absolute percent volume difference is 4.0% for the 113 structure subset, while the mean percent volume difference is -1.6%. It is more common (72/113) to see candidate structures from zeroth-order CSP studies that are more compact than the target structures due to neglect of thermal expansion^{41,105} (and only 2/113 structures considered here were generated with thermal effects included⁴¹). In the case of XXII-G09-L1-E02, the large volume mismatch is likely a result of the MMFF94 force field used for geometry optimization by Group 09.^{41,117}

Table 4.2 shows the similarity metrics comparing XXII-G09-L1-E02 to the XXII target structure, before and after volume correction. As expected, this structure has the largest increase in RMSD(1) of the entire data set upon volume correction. While the VC-RMSD(20) and VC-PWDF values are significantly reduced by the volume correction, both remain much higher than for the other BT6 matches, as seen from the scatter plot in Fig. 4.6. This occurs because the unit-cell volume of the submitted structure is so large that a shift in the packing relative to the target can be observed. As shown by the structural overlay in Figure 4.13, this shift is retained after volume correction. The planes formed by the molecules in the candidate structure are angled considerably with respect to the *a*-axis, whereas the planes formed by the molecules in the target structure are essentially parallel to the *a*-axis. This shift in packing is a result of the expanded volume of the original unit cell, analogous to the temperature-dependent shifts in molecular packing that may occur experimentally.

Table 4.2: Similarity comparisons for structure XXII-G09-L1-E02, before and after volume correction, and after constant-volume (CONV) geometry relaxation with rigid molecules. RMSD values are given in Å.

Structure	RMSD(1)	RMSD(20)	raw-POWDIFF or VC-PWDF
Raw	0.049	0.833	0.5363
VC	0.180	0.550	0.1120
CONV	0.180	0.277	0.0231

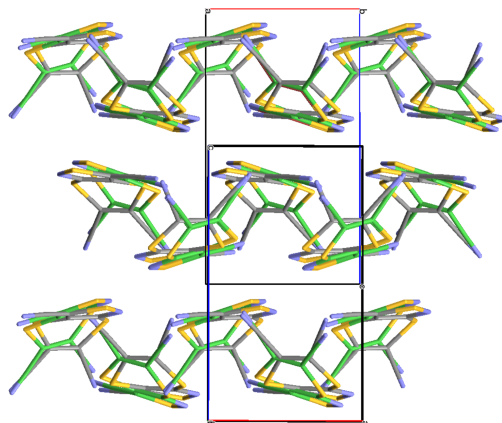


Figure 4.13: COMPACK overlay, in the ab -plane, of the volume-corrected XXII-G09-L1-E02 structure with the target.

The CONV (constant-volume relaxation) functionality of DMACRYS,¹¹³ which holds the unit-cell dimensions constant, was applied to the submitted structure after volume correction. Relaxation of the molecular positions visually corrected the angling of the molecules and the resulting structural overlap is now quite good, as quantified by the VC-RMSD(20) and VC-PWDF values in the final row of Table 4.2. After the CONV relaxation, these metrics correlate with the true similarity to the target as well as for any of the other matches in the dataset (see Fig. 4.6).

We recommend CONV optimization as a final step before quantitative comparison of volume-corrected structures in cases with exceptionally high volume differences ($>10\%$) to compensate for the use of crude force fields in the geometry optimization steps of CSP. However, with this secondary manipulation of the structure, a discussion as to whether the candidate structure should be considered a match to the target is warranted. As mentioned above, changes in molecular packing and conformation may occur between experimental structures collected at different temperatures. Unless these shifts result in changes in properties or symmetry, the structures are considered to be the same, rather than

distinct polymorphs. Use of a force-field method to relax a structure that has undergone a substantial volume change mimics the relaxation that would occur experimentally, when a corresponding volume difference results from a change in temperature. The inclusion of a method for eliminating intramolecular distortions would also be beneficial in cases with dramatic volume differences to improve further the final similarity value for the candidate structure. The intramolecular distortion imparted by our volume correction is not physical and is not corrected by the DMACRYS CONV relaxation, which assumes rigid molecules. That being said, only one structure out of the 52 unique matches from the original study (<2%) has such a dramatic difference in volume relative to its corresponding target.

4.6 Conclusions

This work presents a tailored anisotropic volume correction to improve PXRD comparison of crystal structures. The approach's ability to identify all candidate crystal structures submitted during the 6th CSP blind test⁴¹ that match the target, experimental structures was assessed. In contrast to existing PXRD comparisons, which either involve no volume correction or only an isotropic volume correction⁷¹ (using the CPS tool in Mercury⁹⁸), our approach is capable of segregating the BT6 matches from the remaining candidate structures. All but two of the BT6 matches were found to have variable-cell powder pattern differences (VC-PWDF) of < 0.035. Considering all candidate structures having VC-PWDF values within this threshold, we were also able to identify two matching structures that went uncredited in BT6. These were a match to target XXIII, form A, submitted by Group 09 and a match to target XXV submitted by Group 06.

A limitation of the method is cases where there is an extremely large volume difference between the target structure and a candidate match. Rigid-cell relaxation of the volume-corrected candidate structure with a distributed-multipole force field,¹¹³ or better yet a low-cost quantum-mechanical method such as HF-3c,¹¹⁴ will improve identification of matching crystal structures. However, this would greatly increase the computational cost of our algorithm and is not generally practical.

The optimum VC-PWDF threshold needed to indicate a structural match is highly dependent on the target molecule in question. For the rigid compound XXII, a relatively small VC-PWDF threshold of 0.017 was required to distinguish 10 instances of a polytype structure from matches to the experimental target. In contrast, a threshold of 0.035 is

needed to identify the majority of the BT6 matches. For the flexible molecules XXIII and XXVI, several structures were found to have similar packing, but visible differences in conformation from the target, leading to larger VC-PWDF values in the range 0.035-0.055. While these fall into more of a grey area, they would be expected to give identical structures to the target upon relaxation of the atomic positions, and can therefore be deemed matches. Thus, larger VC-PWDF thresholds must be used to identify structural matches for flexible molecules, compared to rigid molecules.

This work also illustrated some disadvantages of the COMPACK algorithm in cases of flexible molecules with minor conformational differences. Thirteen matching structures, including one of the two missed BT6 matches, required larger tolerances than the COMPACK defaults of ± 20 (% and $^{\circ}$) to achieve a 20/20 molecule match. Tolerances of up to ± 60 (% and $^{\circ}$) were needed, which meet or exceed those reported by Bernstein and coworkers,⁶⁸ who noted similar issues with COMPACK for flexible molecules. However, setting too large of a tolerance can lead to unreasonable cluster alignments and large jumps in RMSD values. While COMPACK has long been the default method for identifying matching crystal structures, the sensitivity of the alignment and RMSD values to the choice of tolerance emphasizes the need to be diligent when using this comparison method.

Overall, the VC-PWDF measure was able to provide as much information as the raw-RMSD(20) with respect to quantifying the true similarity of the compared structure to the target. Anisotropic volume correction was also found to significantly reduce RMSD(20) values obtained from comparison of matching crystal structures, and a strong correlation between VC-PWDF and VC-RMSD(20) values was identified. We recommend utilization of both the VC-PWDF and COMPACK methods in concert to ensure that all matching structures are identified, and that false positives can be readily removed. Pairing with COMPACK is particularly important as a structure that is similar to the target, but presented in a different crystal system, will not be identified as a match by the current version of `vc-pwdf`. Decoupling the anisotropic volume correction from the unit-cell parameters and crystal system presents an opportunity for further development.

The comparison of crystal structures is critical in the analysis of structure-energy landscapes and assessing the ability of CSP methods to reproduce experimentally known structures. The use of the developed VC-PWDF method, in conjunction with COMPACK, is proposed as an improved tool for such analysis. Anisotropic volume correction may also

aid in the use of CSP to match a generated structure to experimental powder diffractograms. This would be of significant interest, particularly in the pharmaceutical industry where solid-form screening is routinely undertaken, where PXRD is common but obtaining a single crystal for every polymorph found can be a daunting endeavour, if not impossible.

CHAPTER 5

DEVELOPMENT AND ASSESSMENT OF AN IMPROVED POWDER-DIFFRACTION-BASED METHOD FOR MOLECULAR CRYSTAL STRUCTURE SIMILARITY

Reprinted with permission from **R. Alex Mayo**, Alberto Otero de la Roza, and Erin R. Johnson, Development and assessment of an improved powder-diffraction-based method for molecular crystal structure similarity, *CrystEngComm*, **24**, 8326–8338 (2022), DOI: 10.1039/D2CE01080A, Copyright 2022 Royal Society of Chemistry.

R. Alex Mayo performed all calculations and crystal structure comparisons, data analysis, and wrote the first draft of the manuscript. RAM and AOR both contributed to improvements in the generalization of the VC-PWDF method. AOR did all coding to implement the VC-PWDF code into the critic2 program. All authors contributed to editing and input on the final form of the manuscript. ERJ and AOR supervised the project.

5.1 Introduction

The physical properties of solid-state molecular materials are dictated by their composition and, critically, the three dimensional arrangement of the component molecules within the solid. Polymorphism arises from the ability of the same compound to form different crystal structures with different properties. These varying properties may make a particular polymorph valuable or cause significant complications for the intended use of

a compound.^{3,29,118–120} The molecular packing is a critical part of what determines the luminescent,¹²¹ optoelectric,^{9–11} and magnetic^{3,4} properties of materials, and the efficacy of pesticides,^{16,17} nutraceuticals,¹⁵ and drugs⁷. Polymorphism is a particularly challenging phenomenon in the pharmaceutical industry, since it affects patentability and intellectual property claims.

Consequently, when the crystal structure of a novel solid is determined, we must have reliable methods to determine whether it is a new or a known polymorph, taking into account that experimental conditions may be different from previous structural determinations, and therefore may result in slightly distorted structures. Comparing molecular crystal structures visually is a highly taxing endeavour and prone to error since molecular crystals are quite complex and there are infinitely many cells with which they may be represented. Thus, an automated and quantitative method of comparing crystal structures is required. This is particularly important for practitioners in the field of molecular Crystal Structure Prediction (CSP). Many computer programs and algorithms have been proposed to quantify structural similarity.^{22–28}

In general, we expect that any quantitative similarity index $d(A, B)$, where A , B , and C are arbitrary crystal structures, has the mathematical properties of a metric:

- $d(A, B) = 0$ if and only if $A = B$;
- $d(A, B) = d(B, A)$; and
- $d(A, C) + d(C, B) \geq d(A, B)$ for any C , the triangle inequality.

A similarity index does not, on its own, distinguish whether A and B are the same structure, or correspond to redeterminations of the same polymorph. A similarity index can be made into such a comparison method by choosing a cutoff value, c , used to classify all possible pairs of crystal structures. If $d(A, B) \leq c$ the two structures are considered the same, and if $d(A, B) > c$ they are considered to be different. We expect that a well-behaved comparison method adheres to the following “cutoff principle”: If two structures A and B are classified as equal for a given cutoff c , any cutoff higher (i.e. more lenient) than c also classifies A and B as equal. Conversely, if A and B are different for cutoff c , any cutoff lower (i.e. more strict) than c also classifies A and B as different. If the comparison method is derived from a similarity index that fulfills the properties of a metric, the cutoff principle is met.

A common comparison method employed for molecular crystals is the COMPACK algorithm.²² COMPACK uses two values to decide whether a pair of structures is equal. COMPACK matches molecules within a given cluster size, M , (commonly 20) from two given crystal structures (A and B) based on the inter-atomic distances and angles in each structure, and generates an optimal overlay of the two structures. The output values include both the number of matching molecules, N , in the cluster, and the root-mean-square-deviation, $\text{RMSD}(N)$, of the atomic positions (in Å) calculated from the optimal overlay of the cluster of N matching molecules. The N value is commonly used to determine the outcome, with $N = M$ indicating a match between the two structures. The $\text{RMSD}(N)$ value may be used to discuss the degree of similarity between two matching structures. COMPACK does, however, require a specified tolerance of how much the inter-atomic distances and angles are allowed to differ for the molecules in the cluster to be considered matching. An alternative to COMPACK is to compare crystal structures based on their simulated powder X-ray diffraction (PXRD) patterns. A similarity index is obtained by comparing the two powder diffractograms using de Gelder's triangle-weighted cross-correlation function, such that a value of 0 corresponds to identical structures, while 1 indicates maximum dissimilarity.²³ This similarity index has the properties of a metric (although the claim that two different crystal structures always generate different diffractograms has not been proven rigorously²⁸).

Another desirable feature of a comparison method is that redeterminations of the same polymorph are classified as equal, even if the two structures differ somewhat due to changes induced by temperature, pressure, or other experimental conditions. This feature is also important in CSP, where calculated and experimental structures are compared, even though the effect of thermal expansion is usually not included in the former. PXRD-based similarity indices and comparison methods are particularly sensitive to changes in unit-cell dimensions: unless two crystal structures were determined at exactly the same conditions, their peak positions will be shifted, potentially resulting in a large dissimilarity measure. To account for this, volume corrections (isotropic⁷¹ or anisotropic¹²²) can be applied to account for cell distortions prior to the generation and comparison of the simulated powder diffractograms.

An alternative approach to account for peak-shifting in the comparison of powder diffractograms is the FIDEL (FIt with DEviating Lattice parameters) method.⁷⁴ It uses an

optimization procedure to maximize the overlap of the two diffractograms by adjustment of numerous structural parameters (molecular conformation, position, and orientation, as well as lattice parameters), using de Gelder's cross-correlation function as the figure of merit. The FIDEL method is commonly applied to cases where an experimental powder diffractogram cannot be indexed and, thus, the unit cell of the experimental structure is unknown. In this work, we focus on cases where the cell parameters of both structures are known (i.e. comparing two solved crystal structures). Here, a correction using only the lattice parameters is proven to be effective (*vide infra*). An optimization strategy, such as the one undertaken by FIDEL, is unnecessary and may be prone to local maxima when significant differences in cell dimensions exist since a cross-over of peak positions may occur.

A 2020 report by Sacchi *et al.*,⁶⁸ assessed the two comparison methods available within the Cambridge Crystallographic Data Centre's (CCDC's) software suite:⁹⁸ COMPACK and a PXRD similarity measure. The outcome of that study highlighted the poor performance of the PXRD comparison tool, which failed to identify many pairs of structures as being redeterminations of the same polymorph due to temperature- and pressure-induced changes in unit-cell dimensions. While details of the PXRD similarity measure used in the CCDC software are lacking, the isotropic volume correction developed by van de Streek and Motherwell is straightforward but insufficient to consistently detect polymorph redeterminations obtained under disparate conditions.⁷¹ The anisotropic nature of thermal expansion in molecular crystals has been discussed in recent studies and, indeed, is more commonly the norm than the exception.^{73,109}

We recently developed a new approach to improve PXRD-based comparison methods using anisotropic volume corrections.¹²² The method was applied to identify candidate structures generated from first-principles crystal structure prediction (CSP) that match known experimental structures and, in the process, identified two uncredited matches from the 6th CSP blind test.⁴¹ However, because the proposed method relied on the transformation to the Niggli reduced cell, which does not depend continuously on the cell parameters, it was susceptible to yielding incorrect results in some cases.

In this work, we present an updated version of the variable-cell powder difference (VC-PWDF) method that performs an exhaustive search over candidate cells. The new VC-PWDF method has been incorporated into the `critic2` program.⁷⁵ We apply VC-PWDF

to compare pairs of experimental structures hosted in the CCDC’s crystal structure database (CSD), specifically the same dataset considered by Sacchi *et al.*⁶⁸ The use of VC-PWDF is found to dramatically improve the results yielded by a PXRD-based comparison method. In addition, we perform a systematic analysis of the effects of changing cutoffs/tolerances on the outcomes and the agreement between VC-PWDF and the CCDC crystal packing similarity (CPS) tool’s implementation of COMPACK. Certain counter-intuitive behaviours of the COMPACK method that violate the cutoff principle are identified and discussed, along with classes of molecular structures that prove problematic for the method due to highly-branched functional groups and/or conformational chirality. Structure pairs that cannot be agreed upon in terms of classification by the two methods are analysed in detail.

5.2 Methods

5.2.1 Mercury’s CPS tool

The CPS tool provided with the CCDC’s Mercury program⁹⁸ was executed through the CSD Python application programming interface. All structures were accessed from the CSD using their refcodes (see Section B.2 for details on some anomalies between the use of local, downloaded cifs and the CSD-housed structures). The CPS tool includes two comparison methods that are both applied automatically. In this work, we considered only the CPS implementation of the COMPACK algorithm, while results from the simulated powder diffractogram comparison were not recorded. COMPACK was used to obtain the number of matching molecules, N , out of a cluster size of $M = 20$ ($N/20$), and accompanying $\text{RMSD}(N)$ values. A variety of user-defined search options in addition to the default parameters are available. Unless otherwise specified, only the following default parameters were modified:

- The cluster size was changed to 20 molecules (default is 15 molecules).
- Each atom’s hydrogen count was ignored (default is to be considered).
- Each atom’s bond count was ignored (default is to be considered).

In addition, COMPACK defines two tolerances: a percentage tolerance for the interatomic distances and an angular tolerance. In the following, we combine both in a single value, so a COMPACK tolerance of 10 signifies $\pm 10\%$ in the distances and $\pm 10^\circ$ in the angles. If

these tolerances are exceeded, two molecules are not considered a match by COMPACK. The COMPACK tolerances were systematically varied from 10 to 60 in increments of 10. The particular tolerance used is specified in the discussion of the results (the default tolerance is 20). When compared using COMPACK, two structures are considered equal if there is a 20/20 match, regardless of RMSD.

5.2.2 Variable-cell powder difference (VC-PWDF)

The method described herein is the same as presented in Chapter 3, and is an improvement of the previous version from Chapter 4. The dependence that the previous version had on the cell description is resolved as described below, and the code integrated into the `critic2` program.⁷⁵ In order to calculate the variable-cell powder-diffraction pattern difference (VC-PWDF) between two crystal structures, the following steps are carried out:

1. Both structures are transformed to their Niggli reduced cell.¹¹¹
2. The structure with more atoms in the Niggli cell is chosen as the reference. (If both structures have the same number of atoms, the choice is arbitrary, so the first structure is the reference.) The objective is to find the cell transformation that brings the other structure (the “candidate” structure) into closest agreement with the reference, as measured by the powder diffraction similarity index.²³
3. Maximum elongations and angle differences relative to the reference cell are defined. By default, these are $\pm 30\%$ in the cell lengths and $\pm 20^\circ$ in the cell angles. Only transformations of the candidate cell that bring it into agreement with the reference cell within these tolerances will be considered.
4. Lattice vectors of the candidate structure are listed in order of increasing length, up to 30% longer than the longest basis vector in the reference cell. This is a finite list and, if the two structures are equal, it contains the three lattice vectors that transform the candidate cell into the reference cell.
5. The basis vectors of the reference structure are each associated with the subset of the candidate structure lattice vectors whose lengths are within $\pm 30\%$ of the reference.
6. All possible triplets of lattice vectors from the candidate structure are considered as a potential new basis that matches the reference basis. Triplets whose vectors are

collinear, or whose angles differ from the candidate structure cell angles by $\pm 20^\circ$ are discarded. Also, the transformed cell must have the same number of atoms as the reference cell.

7. For the surviving triplets, the change of basis is carried out. Then, the basis vectors of the transformed candidate structure are replaced by those of the reference structure, in the spirit of our previous work.¹²² Finally, the powder diffraction similarity index is calculated. The final VC-PWDF is the lowest of all these calculated values.

The simulated powder diffractograms are calculated using Cu $K\alpha_1$ radiation ($\lambda = 1.54036$ Å) from 5 – 50° 2θ and compared with a triangle base-length of $b_t = 1$ in the weighted cross-correlation function.²³

There are some important observations about this algorithm. First, no symmetry information about the crystal is used. The problem caused by the discontinuity in the Niggli cell when the cell is continuously distorted that plagued our previous method¹²² no longer exists. The search over candidate vectors is exhaustive, so the best matching transformation is found within the distance and angle cutoffs set by the user. The computational cost of the method increases with increasing cutoffs, but we have found that the quite generous 30% distance and 20° angle tolerances are a reasonable and efficient choice, with a comparison run time of a few seconds on average.

VC-PWDF identifies two structures as equal if the similarity index is lower than a given value (the PWDF cutoff, see below). In this work, the search over candidate bases is stopped if a comparison yields a similarity index lower than 0.001 (which is lower than any PWDF cutoff we consider). This reduces the computational cost. We also removed all hydrogens prior to the comparison, given that they are often auto-generated and have a negligible effect on the simulated powder diffractograms.

5.3 Dataset

The set of structures used in this work is the same as in Sacchi *et al.*⁶⁸ To make the CPS powder pattern comparison results directly comparable to our VC-PWDF, we consider the powder pattern difference (PWDF), which is one minus the similarity—this method is denoted CPS-PWDF in the remainder of this work. The COMPACK results from the 2020 study are also used for comparison here. As the CSD has been updated since the list was

generated by Sacchi *et al.*, some refcode changes had been made and these are listed in Table B.1.

The data set used by Sacchi *et al.* contains 47,422 individual comparisons between pairs of crystal structures. A single structure may be present in more than one pair. While processing this list, the data set was reduced to a total of 44,939 pairs as follows:

- 12 duplicate pairs were removed.
- 30 pairs were removed due to the crystal structures involving different molecular species. These cases were identified because COMPACK was unable to provide even a single-molecule match.
- 685 pairs, involving 116 disordered structures, were removed after using ConQuest to search for structures with disorder. Neither COMPACK nor VC-PWDF can handle disorder correctly at present.
- 124 pairs were removed after PLATON's¹²³ checkcif identified Alert Level A flagged voids in one of the structures of the pair. See Section B.1.2.3 for an illustrative example and the list of the 78 structures with voids.
- 87 pairs were removed due to 8 problematic structures (see Section B.1.2.4), in which there were missing non-hydrogen atoms in the cif, such that the given structure did not match the correct stoichiometry of the compound. These structures are incompatible with the COMPACK algorithm.
- A final 1,545 pairs, involving 146 refcode families (see Section B.1.2.5), were removed because COMPACK took (what we considered) an unreasonably long time to compare some of the structures in these refcode families. Specifically, if any pair took longer than 1 hour to complete, the whole refcode family of comparisons was eliminated from the dataset. The removed structures are generally, although not always, characterised by highly branched substituents; additional discussion of this issue is presented in Section 5.5.1.

The outcomes of the remaining 44,939 comparisons form the basis of the results and discussion in the rest of this chapter.

5.4 Results

5.4.1 Outcomes of structure comparisons

A confusion matrix is a concise way of comparing the outcomes from two different methods. The rows and columns in a confusion matrix correspond to all possible outcomes of the two methods, and each cell displays the fraction of points in the data set that had a particular outcome from both methods. In our case, we evaluate the COMPACK and VC-PWDF (or COMPACK and CPS-PWDF) comparison methods regarding their ability to evaluate whether a given pair of structures correspond to the same or a different polymorph. For simplicity, in the rest of the article we use the shorthand notation “structure A is equal to B” to mean that structures A and B correspond to the same polymorph, even though one may be a significant distortion of the other.

Disagreements between COMPACK and VC-PWDF (or CPS-PWDF) are reflected in the off-diagonal cells of the confusion matrix, and must correspond to a misassignment by either of the two methods. Although it is possible there are cases in which both methods agree but misassign, examination of the off-diagonal cases in the confusion matrix is likely to reveal problems inherent to each method. This analysis is carried out in Sections 5.5.2 and 5.5.3.

Using the same cut-off of 0.035 for the powder diffractogram comparison (in the following, the PWDF cutoff) and the same CPS results generated in the 2020 study,⁶⁸ the confusion matrices comparing COMPACK with the two different PXRD-based methods (VC-PWDF and CPS-PWDF) are shown in Table 5.1. We note that the small reduction in data-set size has only a minor effect on the results compared to previous work. Using the two CPS methods, 15.91% of structure comparisons yield different outcomes, which is similar to the 16.3% figure obtained by Sacchi *et al.*⁶⁸

Replacing CPS-PWDF with VC-PWDF yields a dramatic improvement in the agreement with COMPACK results, with a total of only 2.89% disagreements—a 5-fold reduction compared to CPS-PWDF. By far the largest change seen by switching to VC-PWDF is the increase in cases where both methods identify a structural match, and a concomitant reduction in cases where COMPACK yields a match and CPS-PWDF indicates different structures. This is explained by the ability of VC-PWDF to account for anisotropic changes in cell dimensions caused by redetermination of the same polymorph under different experimental conditions. As mentioned above, powder diffractogram differences

Table 5.1: Confusion matrices for the outcomes of 44,939 structure comparisons conducted with COMPACK and either CPS-PWDF (top) or VC-PWDF (bottom). In both cases, a PWDF cutoff of 0.035 was used to differentiate “same” and “different” structures. The COMPACK results reported in the literature⁶⁸ were used.

Literature data ⁶⁸ using CPS-PWDF		
CPS-PWDF	COMPACK	
	same	different
same	47.79%	2.05%
different	13.87%	35.88%

Current data using VC-PWDF		
VC-PWDF	COMPACK	
	same	different
same	61.04%	1.96%
different	0.93%	36.06%

are particularly sensitive to changes in cell dimensions and, therefore, PXRD-based methods used without volume correction tend to reject matching (but significantly distorted) structures.

5.4.2 Dependence on tolerances and cutoffs

The PWDF cutoff of 0.035 used by Sacchi *et al.* in 2020⁶⁸ was selected based on the initial survey of the CSD by van de Streek and Motherwell.⁷¹ In the 2020 study, no analysis regarding the effect of changing the PWDF cutoff or the COMPACK tolerances was performed. Instead, the COMPACK tolerances were loosened to 50 only in cases where the CPS-PWDF value was below 0.05 and the COMPACK result using a 20 tolerance indicated non-matching structures. We now evaluate systematically the fraction of comparisons for which the COMPACK and PXRD-based methods disagree, as a function of both PWDF cutoff and COMPACK tolerances. Figure 5.1 presents these results in the form of a heat map, where either CPS-PWDF (top) or VC-PWDF (bottom) is used.

The minimum percentage of comparisons in disagreement between COMPACK and CPS-PWDF is 12.35%, obtained with a PWDF cutoff of 0.05 and a COMPACK tolerance of 10. In contrast, the minimum disagreement between COMPACK and VC-PWDF is 2.84%—a 4-fold decrease from the CPS-PWDF minimum. Interestingly, this minimum occurs at the intersection of a PWDF cutoff of 0.03 and a COMPACK tolerance of 20, which are commonly taken to be the default cutoff and tolerances for these methods.^{22,71}

		COMPACT tolerance					
		10	20	30	40	50	60
CPS-PWDF cutoff	0.005	29.10%	34.30%	36.17%	37.29%	38.33%	40.11%
	0.010	22.72%	27.41%	29.19%	30.26%	31.30%	33.05%
	0.020	16.07%	20.04%	21.66%	22.68%	23.69%	25.41%
	0.030	13.64%	16.88%	18.41%	19.40%	20.32%	21.99%
	0.040	12.76%	15.56%	16.91%	17.78%	18.57%	20.22%
	0.050	12.35%	14.67%	15.85%	16.62%	17.35%	18.99%
	0.060	12.56%	14.40%	15.45%	16.10%	16.76%	18.40%
	0.070	13.72%	15.12%	15.87%	16.44%	16.98%	18.58%
	0.080	15.10%	16.20%	16.69%	17.18%	17.57%	19.13%
	0.090	16.48%	17.14%	17.50%	17.94%	18.19%	19.73%
	0.100	18.18%	18.36%	18.62%	18.92%	19.07%	20.58%

		COMPACT tolerance					
		10	20	30	40	50	60
VC-PWDF cutoff	0.005	4.77%	7.75%	9.39%	10.36%	11.45%	13.24%
	0.010	4.07%	4.97%	6.46%	7.40%	8.47%	10.27%
	0.020	5.69%	3.02%	3.94%	4.65%	5.60%	7.36%
	0.030	7.28%	2.84%	3.33%	3.92%	4.80%	6.36%
	0.040	8.49%	3.20%	3.15%	3.39%	3.93%	5.46%
	0.050	9.36%	3.87%	3.32%	3.44%	3.72%	5.22%
	0.060	10.01%	4.32%	3.44%	3.48%	3.61%	5.05%
	0.070	11.38%	5.59%	4.53%	4.39%	4.40%	5.68%
	0.080	12.55%	6.73%	5.56%	5.20%	5.10%	6.32%
	0.090	12.98%	7.14%	5.88%	5.45%	5.31%	6.41%
	0.100	13.46%	7.59%	6.22%	5.72%	5.55%	6.54%

Figure 5.1: Heat maps representing the percentage of comparisons for which COMPACT and CPS-PWDF (top) or VC-PWDF (bottom) disagree on the outcome, as a function of the PWDF cutoff and COMPACT tolerances used.

As shown by the confusion matrix in Table 5.2, this choice results in the instances of disagreement where COMPACT predicts different polymorphs but VC-PWDF does not being more prevalent (by a factor of 2). This is the opposite behaviour to that seen previously with CPS-PWDF.⁶⁸

In addition to the difference between the minimum disagreement values, the difference in the topography of the two heat maps in Figure 5.1 is dramatic. The expected correlation for two well-behaved comparison methods (i.e. the minimum following the diagonal) is only observed in the VC-PWDF case. Breakdowns of the total disagreement into the cases where COMPACT considers pairs the same and VC-PWDF considers them different, and vice-versa, are shown in Figure 5.2 (top and bottom, respectively). Some anomalies are

Table 5.2: Confusion matrices for the outcomes of 44,939 structure comparisons conducted with COMPACK and either CPS-PWDF (top) or VC-PWDF (bottom). In both cases, the optimal COMPACK tolerance and PWDF cutoff identified for each method was used to differentiate “same” and “different” structures. These values are 10 and 0.05 for COMPACK/CPS-PWDF and 20 and 0.03 for COMPACK/VC-PWDF.

Literature data ⁶⁸ using CPS-PWDF		
CPS-PWDF	COMPACK	
	same	different
same	48.74%	6.01%
different	6.35%	38.61%

Current data using VC-PWDF		
VC-PWDF	COMPACK	
	same	different
same	60.30%	1.94%
different	0.90%	36.86%

revealed in the data, particularly in the bottom panel, corresponding to the cases where VC-PWDF predicts equal and COMPACK predicts unequal structures for PWDF cutoffs between 0.005 and 0.02. In this region, for each choice of PWDF cutoff, the frequency that COMPACK identifies different structures increases with looser tolerances from 40 to 50 and from 50 to 60. This is due to comparisons that yielded a 20/20 match at the tighter tolerance, but a lower number of matching molecules at the looser tolerance. This behaviour violates the cutoff principle posited above. These cases are considered in more detail in Section 5.4.3.

5.4.3 COMPACK tolerance behaviour

5.4.3.1 Cluster matches

We now take a closer look at the number of molecules ($N/20$) matched by COMPACK as well as the $\text{RMSD}(N)$ obtained from the comparison. We first assessed the changes in the $N/20$ matches given by COMPACK for the full set of 44,939 structure comparisons as a function of the tolerance used. The results are summarized in Table 5.3 and are grouped according to the change in tolerance in increments of 10. More than half (55.83%) of the total number of comparisons do not change N over the full range of tolerances. $\Delta N > 0$ indicates more matching molecules at the looser tolerance, which occurs for 19,816 unique comparisons (44.10% of the data set) for at least one change in tolerance. $\Delta N \geq 0$ is the expected behaviour with increasing tolerance based on the cutoff principle.

		COMPACK tolerance					
		10	20	30	40	50	60
VC-PWDF cutoff	0.005	2.95%	7.41%	9.18%	10.23%	11.30%	13.06%
	0.010	1.05%	4.47%	6.16%	7.20%	8.26%	10.02%
	0.020	0.25%	1.89%	3.30%	4.22%	5.22%	6.96%
	0.030	0.14%	0.90%	2.09%	2.95%	3.92%	5.55%
	0.040	0.12%	0.45%	1.37%	2.06%	2.86%	4.48%
	0.050	0.11%	0.33%	1.00%	1.64%	2.30%	3.91%
	0.060	0.07%	0.20%	0.71%	1.29%	1.89%	3.46%
	0.070	0.06%	0.14%	0.56%	1.06%	1.59%	3.09%
	0.080	0.06%	0.12%	0.49%	0.87%	1.36%	2.82%
	0.090	0.06%	0.11%	0.42%	0.78%	1.24%	2.65%
	0.100	0.05%	0.09%	0.35%	0.67%	1.11%	2.47%

		COMPACK tolerance					
		10	20	30	40	50	60
VC-PWDF cutoff	0.005	1.82%	0.34%	0.22%	0.13%	0.14%	0.18%
	0.010	3.03%	0.50%	0.30%	0.20%	0.21%	0.25%
	0.020	5.44%	1.13%	0.64%	0.43%	0.38%	0.40%
	0.030	7.14%	1.95%	1.24%	0.97%	0.88%	0.80%
	0.040	8.36%	2.75%	1.78%	1.33%	1.07%	0.97%
	0.050	9.25%	3.54%	2.31%	1.81%	1.42%	1.31%
	0.060	9.94%	4.12%	2.74%	2.19%	1.72%	1.58%
	0.070	11.32%	5.45%	3.98%	3.34%	2.81%	2.60%
	0.080	12.49%	6.61%	5.08%	4.32%	3.75%	3.50%
	0.090	12.93%	7.03%	5.45%	4.67%	4.07%	3.77%
	0.100	13.41%	7.50%	5.87%	5.05%	4.43%	4.08%

Figure 5.2: Heat maps of the percentage of comparisons that are considered the same by COMPACK and different by VC-PWDF (top), or that are considered different by COMPACK and the same by VC-PWDF (bottom), as a function of the PWDF cutoff and COMPACK tolerances used.

The number of structure pairs that change from $N < 20$ to $N = 20$ (i.e. a change from being considered different to equal) is considerable when the tolerance increases from 10 to 20. This implies that a tolerance of 10 is insufficient to provide accurate classification of many structure pairs with COMPACK. This interpretation is supported by the dramatic reduction in the number of cases identified as a match by VC-PWDF, but as different by COMPACK, with increased tolerance from 10 to 20 in the lower panel of Figure 5.2. Our previous study¹²² showed that loosening the tolerances up to 60 can be necessary in order to achieve a 20/20 match for some structures with modest RMSD(20) values of ca. 0.36 Å. Similarly, Table 5.3 shows that loosening the tolerances beyond 40 identifies a further

Table 5.3: Number of structure comparisons with specified change in the number of molecule matches ($N/20$) predicted by COMPACK, as a function of changes in the COMPACK tolerances.

Cases of:	Change in COMPACK tolerance				
	10→20	20→30	30→40	40→50	50→60
$\Delta N \neq 0$	11,388	10,527	11,538	11,380	10,536
$N < 20 \rightarrow N = 20$	2,671	851	513	489	827
$\Delta N < 0$	0	0	8	28	134
$N = 20 \rightarrow N < 20$	0	0	1	14	57

1,316 matches.

Table 5.3 also highlights the significant number of structure pairs for which $\Delta N < 0$, indicating that fewer matching molecules are found at more permissive tolerances. This violates the cutoff principle, meaning that COMPACK, in this respect, is not a well-behaved comparison method. For our data set, the onset of this behaviour is the change from 30 to 40, and the results worsen rapidly with further loosening of the tolerance. Notably, a total of 72 cases change from $N = 20$ (same) to $N < 20$ (different) with an increase in tolerance. This prevents a user from simply setting the loosest tolerance (60) to cast a wide net, as this will not identify all possible 20/20 matches. As noted above, increasing the COMPACK tolerances is necessary in some cases to obtain the correct classification of a given structure pair. However, once a tolerance of 40 is reached, N values lower than 20 do not guarantee that a structure pair cannot be identified as a match at a tighter tolerance.

5.4.3.2 RMSD(N) values

Even if N remains unchanged, RMSD(N) values from COMPACK can vary significantly with changes in tolerance. As RMSD(N) values with different N are not directly comparable, only cases where N is unchanged after the change in tolerance are considered in the following analysis (about 34,000 cases at each tolerance change). Figure 5.3 shows the change in RMSD(N) as a function of changes in COMPACK tolerance. The whiskers cover the range containing 99.9% of the data about the median. Values beyond the whiskers are plotted individually as circles.

For all changes in COMPACK tolerance, the interval spanned by the interquartile range (50% of the data) around the median has a negligible height in the scale of the plot, evidencing that the majority of cases have very small (even zero) Δ RMSD(N). In addition, the whiskers hardly extend beyond 0 Å for the 10→20 and 20→30 changes in tolerance.

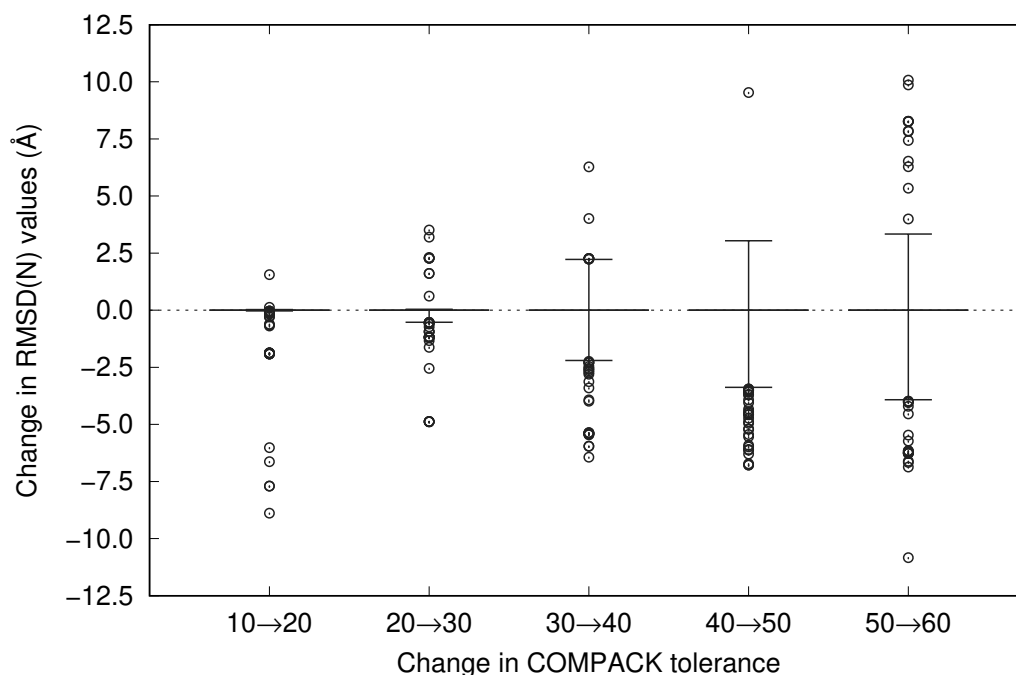


Figure 5.3: $\Delta\text{RMSD}(N)$ values as a function of changes in COMPACK tolerance. The whiskers covers 99.9% of the data about the median, and outliers are shown as circles.

The range of values covered by 99.9% of the data about the median broadens at looser tolerances. Additionally, there are a number of outliers that appear at each change in tolerance that correspond to some remarkable changes in the $\text{RMSD}(N)$ values (recall that there is no change in N). The magnitudes of the greatest $\text{RMSD}(N)$ changes also generally increase with tolerance, with the exception of the most negative $\Delta\text{RMSD}(N)$ values obtained at the smallest tolerance interval.

Five outliers (SUCROS27-SUCROS33, MNPYDO08-MNPYDO29, MNPYDO09-MNPYDO29, VOQHIU-VOQHIU01, and GLUCSA16-GLUCSA18) see a remarkable decrease in their $\text{RMSD}(N)$ values ($\Delta\text{RMSD}(N) = -8.887, -7.702, -7.695, -6.630,$ and -6.016 \AA , respectively) when the tolerance is increased from 10 to 20. All except VOQHIU-VOQHIU01 are $\text{RMSD}(20)$ values. However, there are many cases with $\Delta\text{RMSD}(N) > 0$ with loosening tolerance, which would be in violation of the cutoff principle if $\text{RMSD}(N)$ were used as an ingredient of the COMPACK comparison method. The seven (four unique) most extreme cases within the highest tolerance interval are HUFKAV-HUFKAV01, SANYIP01-SANYIP02, three cases involving VELBOD, and DEVBAH-DEVBAH01, which show increases in $\text{RMSD}(20)$ values of 10.075, 9.868, 8.273, and 8.255 \AA , respectively.

In both the cases of $\text{RMSD}(N)$ decreasing and increasing with the loosening of tolerance, the same underlying issue appears to be the source. Since the N value is not changing, the determination of the number of matching molecules is unaffected, so it is the determination of the optimum overlay that is the root of the problem. Visually, the molecular overlay is very poor when the $\text{RMSD}(N)$ is high, and the overlay is excellent when it is small. An example for the SUCROS27-SUCROS33 comparison is shown in Figure B.4. The details of how COMPACK tolerances affect the RMSD are not clear. However, these results emphasize the unexpected variability in the similarity index calculated by COMPACK with the choice of tolerance for certain cases.

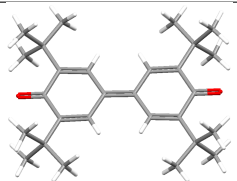
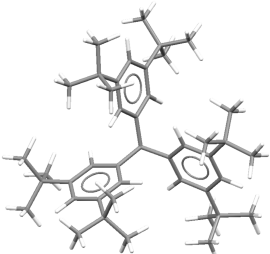
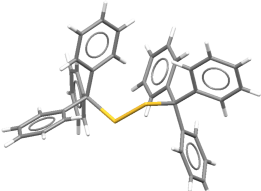
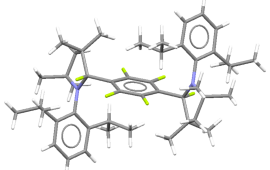
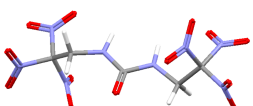
5.5 Discussion

5.5.1 COMPACK issues with highly branched molecules

It was noted in Section 5.3 that 1,545 comparisons involving 146 unique refcode families were removed from the data set because the COMPACK comparisons took at least one hour, and up to several hours or days to complete. Some examples of these molecules are shown in Table 5.4, and a full list of the removed refcode families is given in Section B.1.2.5. A cursory review of the structures reveals that many of the compounds contain highly branched functional groups: t-butyl, isopropyl, triphenylmethyl, nitromethyl, or some related derivative and/or combination. Often there are several highly branched substituents present that are somewhat symmetrically distributed in the molecule. Ultimately, 70/146 of the problematic refcode families contain at least one of the above-named highly branched moieties (list searched with ConQuest). For the remaining cases, it is likely that they contain other problematic functional groups we did not identify, or structural complexities, such as incorrect matching of enantiomers (see Section 5.5.2.2).

The appearance of several highly branched substituents likely causes problems with Ullmann's algorithm,¹²⁴ a modified version of which is used in COMPACK.²² Ullmann's algorithm is a (sub)graph isomorphism method. It tries to find an isomorphism between two given graphs by systematically enumerating all possible permutations of the graph nodes. Ullmann's method uses a tree search that is simplified by calculating unsuitable node assignments based on node connectivity, which cuts down the computational cost. In the context of structure comparison, molecules are represented as graphs by their atomic connectivity, and COMPACK leverages chemical information such as atom and bond types

Table 5.4: Some examples of molecules (and associated structure refcode families) that are difficult to compare using COMPACK.

Refcode	Molecular Structure	Functional Group	Other Refcode Examples
FADDOD		t-butyl	BADGAO, BECMUT, EBIGUR, HELXUR, ISIKAW, INOCET, GACHEY, QIHSEF, TAFKET, TIWYIH, YARHEH, ZEDUG
DAZPUS		di-t-butylphenyl	DATQIY, HAXHET, LURHAJ, MBPHOL, QEHLUL
PEKZAG		triphenylmethyl	KUVWON, TEPHME, WAPBUK, YOSRED, YUHGOX, ZAJBOE
IVATUW		diisopropylphenyl	PEDTUP
NOEURA		trinitromethyl	COYLAF, IREPIG, NOETNA, VALSUY, VALTEJ

to decrease the cost of the tree search even further.²²

Our own implementation of the method in `critic2` shows that molecules such as those appearing in Table 5.4 are a problem for Ullmann’s algorithm. The highly-branched nature of the substituents and their symmetric distribution in the connectivity graphs mean that there are many possible graph isomorphisms to explore, and Ullmann’s techniques to simplify the tree search are not effective at reducing the computational cost. However, since we have no access to the COMPACK code, we can only speculate about the true nature of the problem.

In addition to the cases where the comparison takes an unreasonably long time, optimal molecular overlays found by COMPACK can also be erroneous for highly branched

molecules. A simple demonstration of this issue is presented for a hypothetical molecule containing four tri-tert-butylsilane substituents bonded to a central silicon atom by ethyne linkers. We compare two identical structures containing a single such molecule in a supercell, but with the atomic order randomly permuted in the second structure. COMPACK is unable to identify the correct, identical overlay, as shown in Figure 5.4.

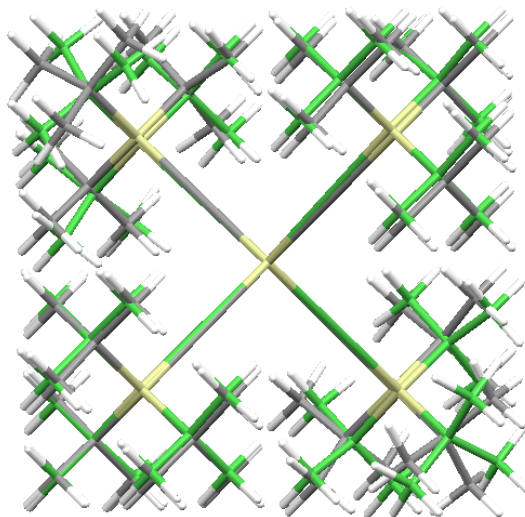


Figure 5.4: Best overlay generated by COMPACK for a hypothetical molecule consisting of four tri-tert-butylsilane substituents bonded to a central silicon atom by ethyne linkers (RMSD(1) = 2.313 Å). The comparison and reference structures are identical with the exception of the order in which the atoms appear in the files.

At the time of writing, this COMPACK error, which can result both in unduly long comparison times and in erroneous structural comparisons, has not been identified as a known limitation of the method. The comparison between ZEDCUG and ZEDCUG01 was highlighted by Sacchi *et al.*⁶⁸ as a fault of the CPS-PWDF method, which was rationalized to be due to its inability to detect a conformational change of the molecule. In reality, the two molecular structures are effectively identical. If the two structures are manipulated manually in Mercury, their packing is a perfect match, as shown by the overlay in Figure 5.5. It was the erroneous overlay generated by COMPACK that was at fault, probably stemming from their use of Ullmann's graph-matching algorithm.

5.5.2 VC-PWDF same / COMPACK different

As noted in Section 5.4.2, it is roughly twice as common for a pair of structures to be considered the same by VC-PWDF and different by COMPACK than the reverse. The

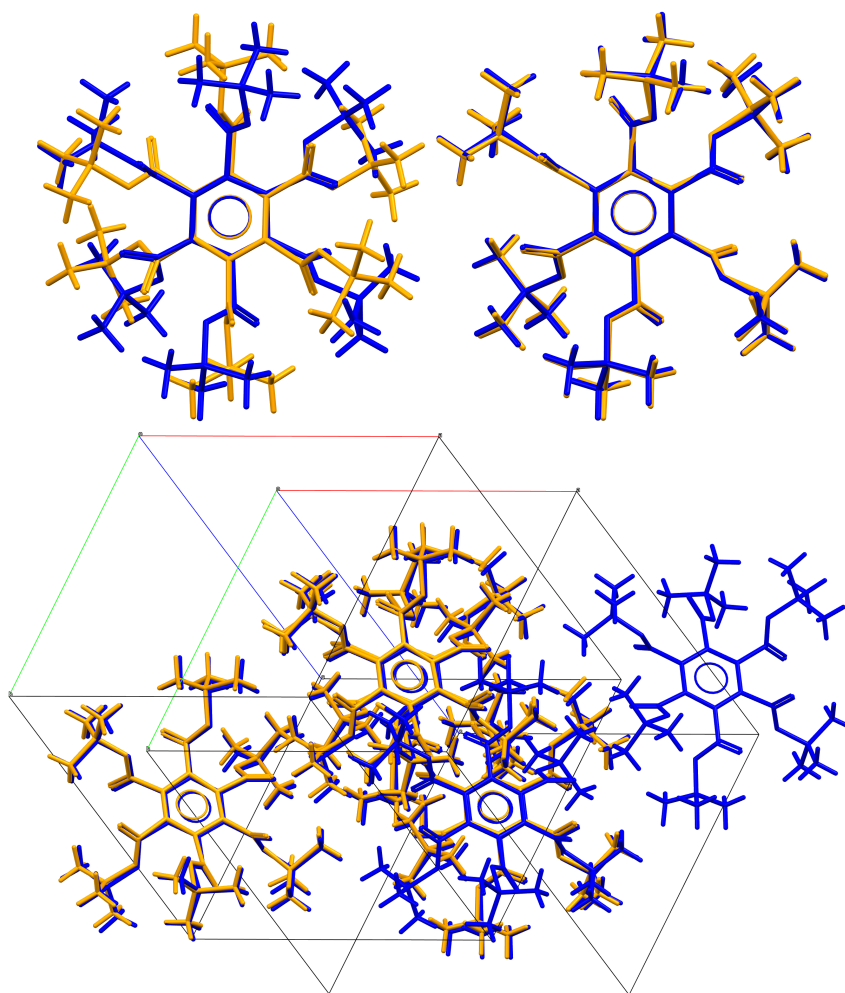


Figure 5.5: COMPACK “optimum” overlay for a single molecule of ZEDCUG and ZEDCUG01 (top-left), manual overlay for a single molecule, showing perfect coincidence (top-right). Overlay of ZEDCUG and ZEDCUG01 showing coincident molecular position and orientation, done manually (bottom).

minimum on the corresponding heat map (Figure 5.2, bottom) lies at the intersection of a PWDF cutoff of 0.005 and a COMPACK tolerance of 40. This 0.13% of structure comparisons is the set for which VC-PWDF and COMPACK cannot agree. Of the 59 pairs in this set, 14 can achieve a 20/20 molecule match at a different COMPACK tolerance. Of the remaining 45 comparisons, 5 are between structures that contain a compound problematic for Ullmann’s method (Section 5.5.1) although their comparison did not exceed a runtime of 1 hour. We consider these to be a problem with COMPACK, leaving 40 comparisons to analyze.

5.5.2.1 Conformational phases or atom assignment errors

Sixteen comparisons (14 refcode families) were found to yield a perfect overlay, with the exception of the positions of certain atoms within the molecular structure. The very similar crystal packing causes VC-PWDF to identify them as equal, while the change in atomic positions causes them to be identified as different by COMPACK. In 14 of these cases, the structure change manifested as a 180° rotation of a planar group, which exchanged the positions of a C(Ar)H and N(Ar), or C(Ar)H and O, or C=O and C-CH₃. An example is shown for the ZITZUX-ZITZUX01 pair in Figure 5.6. The other two cases show a difference in the position of a N(Ar) atom in a fused ring (PTERID-PTERID11 and PEDJUD-PEDJUD01). These may be real conformational changes, such that the description of “conformational phases” defined by Zuñiga *et al.*¹²⁵ (different phases with near identical molecular packings but differences in molecular conformation) would be fitting. However, they may also be the result of atomic identity/position misassignments during the structure solution from single-crystal XRD data. The electron densities of these groups are very similar and, if the resolution of the data is sub-optimal, it may not be straightforward to differentiate one from the other in the refinement process. Three additional clear cases of conformational phase pairs were observed (BEDMIG11-BEDMIG12, LNLEUC10-LNLEUC11, and EJEQAL01-EJEQAL05), which show conformational changes in a terminal alkyl group.

5.5.2.2 Molecular connectivity misassignment and chirality errors in COMPACK

Three cases (refcodes MEPHPY, JIYKAD, and LADBIB) show a perfect visual overlay when compared using COMPACK. However, there appeared to be an issue in COMPACK’s determination of the molecular units, with different numbers of “molecules” identified in the unit cells of the two structures, causing the structures to be identified as different ($N < 20$). For example, two overlays obtained for the JIYKAD-JIYKAD01 structure pair are shown in Figure 5.7. It is clear that COMPACK does not view the C and Cl atoms to be bonded in one of the two structures (JIYKAD01), likely due to the bond length exceeding some internal threshold (C-Cl distances of 1.989 and 2.079 Å in JIYKAD and JIYKAD01, respectively). Since COMPACK relies on comparing clusters with an equal number of molecules, the different nature of the molecular units in both structures prevents the match. This shortcoming is an inescapable consequence of involving molecular connectivity graphs in the similarity calculation.

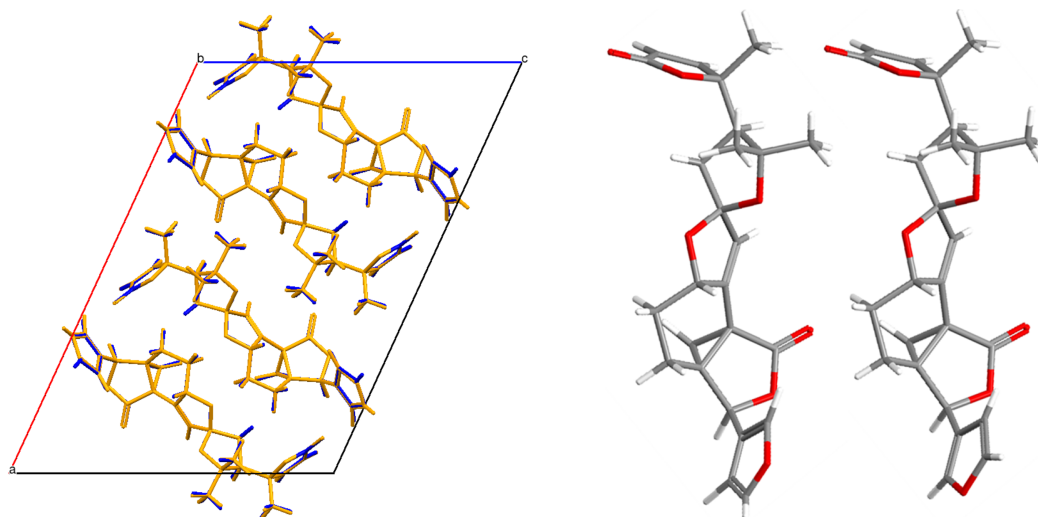


Figure 5.6: An example of possible conformational phases (could be atom misassignment during structure solution) ZITZUX and ZITZUX01. The overlay of the two structures is shown, illustrating the identical packing (left), and the difference in the furyl ring orientation (right).

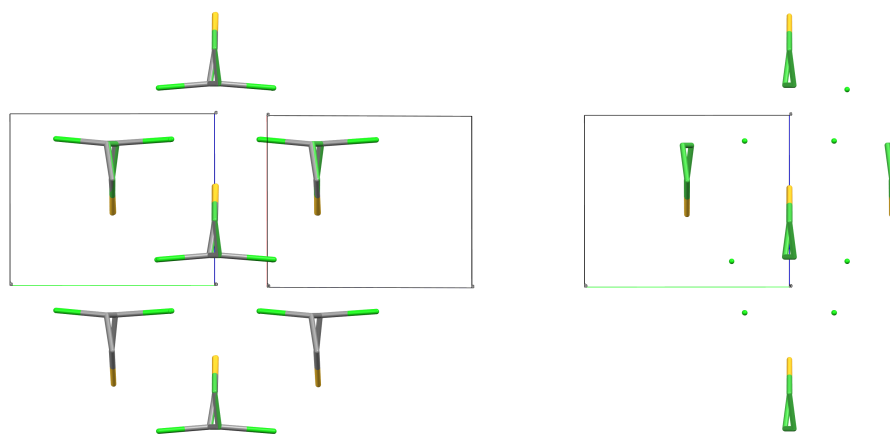


Figure 5.7: COMPACK was used to overlay JIYKAD and JIYKAD01 (13/20 molecules match), the left plot shows both structures overlaid (perfect agreement) and the right plot shows only the JIYKAD01 structure. The chlorine atoms in JIYKAD01 are not bonded to the tetrahydrothiophene ring and are considered separate “molecules” by COMPACK.

An additional structure pair presents a different type of problem for the COMPACK method. UHIKUR and UHIKUR01 fail to yield a 20/20 match, despite it being possible to perfectly overlay the structures manually (shown in Figure 5.8, left) and not having any of the structural moieties identified as problematic for Ullmann’s method in Section 5.5.1.

The molecule adopts a conformation with helical chirality in the crystal structure, and due to the presence of glide planes, exists as a racemate. As shown in Figure 5.8, right,

COMPACK matches the wrong enantiomer, thus creating an incorrect “optimal” overlap between the two structures and only achieving a match of 1/20. The selection or deselection of the “allow molecular inversion” option had no effect on the outcome of the COMPACK comparison between these two structures. The COMPACK source code is not openly available so we can only speculate about what causes COMPACK to fail in this case.

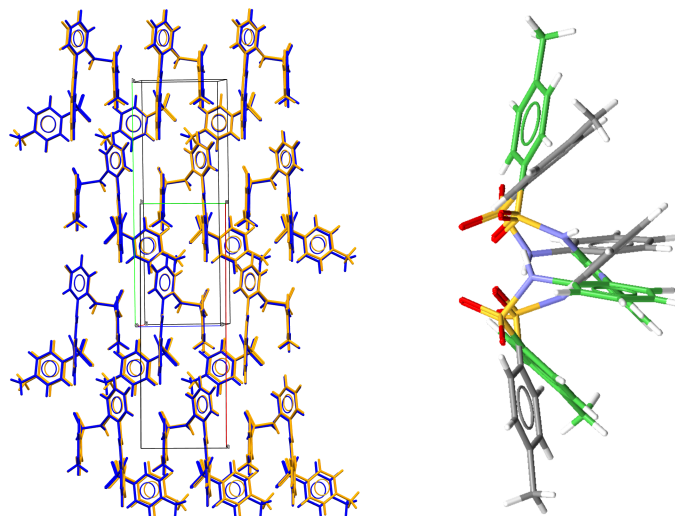


Figure 5.8: Manual overlay (left) and COMPACK optimum overlay (1/20 molecules, right) for the comparison of UHIKUR and UHIKUR01.

5.5.2.3 Polytypes

The remaining 17 cases (7 refcode families, BEDMIG, EDIRIU, DAWGAL, DHXANT, LISLEU, SILVAL, and SITQIV) are “polytypes”, where the differences between structure pairs arise from different stackings of planes with identical two-dimensional molecular packing. An example is shown in Figure 5.9 for the SILVAL-SILVAL02 pair. Polytype structure pairs are different polymorphs, although their similarity is clearly apparent. The overall similar packings generate similar PXRD patterns, resulting in low VC-PWDF values. Therefore, polytypes, as well as conformational phases and isomorphous structures, are problematic for PXRD-based methods like VC-PWDF.

5.5.3 COMPACK same / VC-PWDF different

It is fairly rare to have a pair of structures that COMPACK classifies as equal but VC-PWDF classifies as different. This occurs for less than 1% of the total structural pairs considered at the optimum tolerances/cutoffs. The minimum on the corresponding heat map (Figure 5.2,

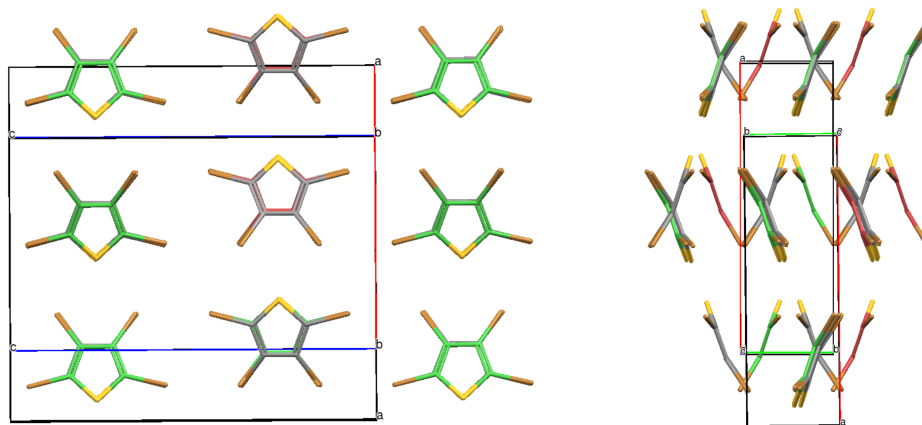


Figure 5.9: COMPACK overlays of the polytype structures SILVAL and SILVAL02 in the (010) and (100) planes (left and right, respectively).

top) lies at the intersection of a VC-PWDF cutoff of 0.1 and COMPACK tolerance of 10. However, the frequency of disagreement appears to plateau after a VC-PWDF cutoff of 0.06 is reached. We will consider this 0.07% of structure comparisons as the set where VC-PWDF and COMPACK cannot agree. This list yields 31 comparisons, composed of structure pairs from 13 refcode families. All of these cases include a polymorph label in the structure metadata, indicating that these 31 structure pairs are considered to be known polymorphs with respect to one another, and therefore COMPACK is in error, according to this assignment.

Further analysis reveals that, for all 31 pairs, COMPACK falsely predicted matching structures due to the use of too small a cluster size. This can be illustrated by the comparison of two carbamazepine structures, CBMZPN03 and CBMZPN11, shown in Figure 5.10. CBMZPN03 is a rhombohedral polymorph ($R\bar{3}$, rhombohedral lattice) with larger-than-average (Platon's¹²³ checkcif Alert Level B) voids about the $\bar{3}$ rotoinversion axis. The comparisons of CBMZPN03 with CBMZPN11 and CBMZPN13 (both with triclinic $P\bar{1}$ space group) using a cluster size of 20 molecules yields a very good overlap with COMPACK. However, if the cluster size is doubled and the same tolerance of 10 is used, only 34/40 molecules match for CBMZPN03–CBMZPN11, and 31/40 for CBMZPN03–CBMZPN13. The resulting overlay shows the difference in packing that occurs beyond the original cluster of 20 molecules (Figure 5.10). An advantage of PXRD-based comparison methods is that they effectively consider the entire crystal lattice, not just a finite cluster within the crystal, and they are therefore more sensitive to long-range changes in packing.

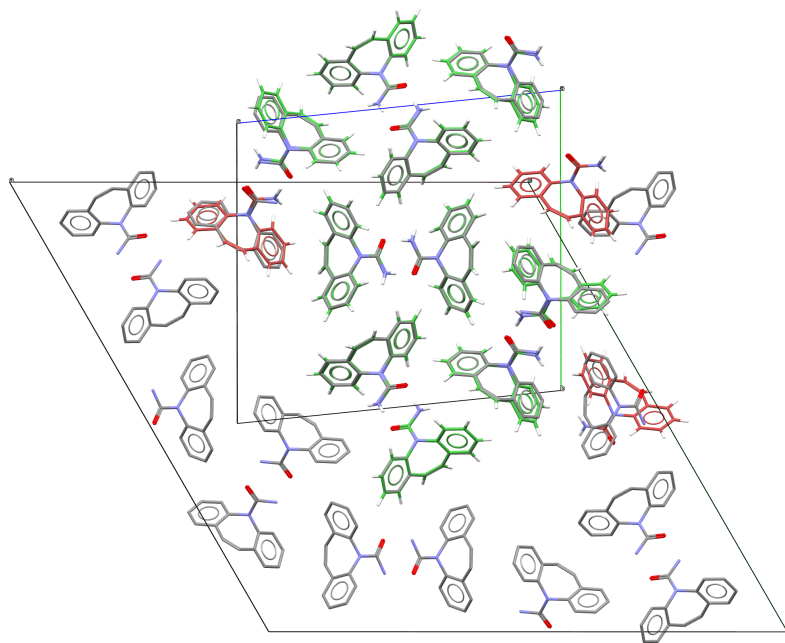


Figure 5.10: Overlay of CBMZPN03 and CBMZPN11 generated by COMPACK using a cluster of 40 molecules.

Based on this result for carbamazepine, all 31 comparisons were re-run with COMPACK using a cluster size of 40 molecules at 10 tolerance, and again with 50 molecules at 20 tolerance. None of the comparisons were able to achieve a 50/50 match, although there are two cases where a 40/40 match was found (MELXEG-MELXEG01 with 48/50 and XELLOP-XELLOP01 with 49/50). The XELLOP-XELLOP01 comparison with a cluster size of 50 (and tolerance equal to 10, to reduce computation time) was visualised in Mercury and clearly shows the same behaviour as the carbamazepine example. Using the same analysis, all 31 of these comparisons were confirmed to be different polymorphs as specified in the metadata. While the default cluster size for COMPACK has been maintained at 15 molecules since its inception,²² cluster sizes of 20 are commonly used to compare single component crystals, and we show here the occasional need to extend the cluster size beyond that in order to obtain the correct solution.

5.6 Conclusions

In this work, we assessed comparison methods for molecular crystal structures regarding their ability to identify redeterminations of the same polymorph, i.e., when the two structures being compared are identical save for slight distortions caused by varying

experimental conditions, or when one of the structures is predicted computationally and the other is determined experimentally. The former case is important in order to determine whether a new structure is a known polymorph, which has practical and legal implications for the pharmaceutical industry. The latter case is important in the context of molecular crystal structure prediction.

Two kinds of comparison methods were analyzed: the popular COMPACK method, based on matching molecular clusters, and powder X-ray diffraction (PXRD)-based comparison methods. In particular, we propose a new PXRD-based similarity index and comparison method called VC-PWDF (variable-cell powder difference), which is a refinement of our previous work. For a set of 44,939 individual crystal structure pairs, it is shown that the level of agreement between COMPACK and VC-PWDF is much greater than between COMPACK and the CCDC crystal packing similarity (CPS) PXRD-based comparison method (CPS-PWDF). Using an optimal combination of cutoffs and tolerances, the minimum frequency of disagreement between COMPACK and VC-PWDF is only 2.84%, which is more than 4 times lower than the best possible CPS-PWDF result of 12.36%. In contrast to CPS-PWDF, it is more than twice as likely for VC-PWDF to identify a pair of structures as the same, while COMPACK classifies them as different, than the reverse.

The increased agreement between VC-PWDF and COMPACK relative to CPS-PWDF can be attributed to the success of the volume correction enhancement, given that PXRD-based comparison methods are particularly sensitive to changes in cell dimensions. The agreement between VC-PWDF with COMPACK indicates VC-PWDF is at least as robust as COMPACK and, together with the fact that PXRD-based comparison methods are reasonably fast, VC-PWDF can be reliably employed as a rapid first pass test when comparing large data sets (CSD, CSP structure-energy landscapes).

We then systematically investigated the performance of COMPACK and VC-PWDF. We examined the effect of COMPACK tolerances and powder-pattern difference (PWDF) cutoffs on the structure classification. Several counter-intuitive outcomes were obtained from the analysis of the effect of the chosen tolerance on COMPACK results. First, some structure pairs that are considered equal by COMPACK at a given tolerance are different at a looser tolerance. This behaviour occurs at tolerances of 40 or higher, which are therefore not generally recommended. Second, there are some structure pairs for which the

RMSD(N) calculated by COMPACK increases with looser tolerances while maintaining the same number of matching molecules; this effect has been observed at all examined tolerances. Therefore, COMPACK is not a well-behaved comparison method regarding its dependence on the tolerances.

Another COMPACK weakness not previously reported is its difficulty with molecules containing several highly branched functional groups symmetrically distributed in the molecule. A single COMPACK comparison involving such molecules may take hours to days, and we have shown with a simple example that COMPACK can fail to match identical molecular structures that differ only in the order in which their atoms are given. We hypothesize that the problem lies in COMPACK's use of Ullmann's method for molecular graph matching.

Further analysis of the disagreements between VC-PWDF and COMPACK was used to identify the strengths and weaknesses of each method. VC-PWDF has trouble differentiating structures with very similar molecular packings, which is reasonable for a PXRD-based method. In particular, VC-PWDF erroneously reports as equal a few structure pairs that are actually polytypes, conformational phases, and isomorphous structures. Conversely, COMPACK fails for some structure pairs when: a) the atomic connectivity of one of the structures is not correctly identified, b) there is helical chirality present in the molecules, and c) not enough molecules are included in the cluster, with some of the pairs of unequal structures requiring up to 50 molecules to be differentiated by COMPACK.

In summary, the development of a single accurate and precise tool for automated and quantitative comparison of crystal structures remains challenging. While identical and obviously different structures are relatively easy to identify, there remains a grey area where similar structures are difficult to classify. The utilization of two methods, COMPACK that uses atomic positions, and VC-PWDF that uses simulated powder diffractograms, can be useful in these cases in order to determine how best to classify a particular structure pair. It is the opinion of the authors that a strict choice of cutoff should be used with caution, as a generic value will not correctly classify all pairs.¹²² However, the analysis of a large dataset of structure pairs in this work suggests using a cutoff of 0.03 for VC-PWDF and 20 tolerance for COMPACK.

Given the somewhat ambiguous nature of the question “are these two structures the same polymorph?”, there will always be a grey area of similar structures for which the

values produced by automated, quantitative, computational comparison methods will be insufficient to definitively answer the question. In these cases, additional work will be required in order to make the correct classification. However, developing more accurate comparison methods is essential for narrowing this grey area.

CHAPTER 6

QUANTITATIVE MATCHING OF CRYSTAL STRUCTURES TO EXPERIMENTAL POWDER DIFFRACTOGRAMS

Reprinted with permission from **R. Alex Mayo**, Katherine M. Marczenko, and Erin R. Johnson, Development and assessment of an improved powder-diffraction-based method for molecular crystal structure similarity, *Chem. Sci.*, **14**, 4777-4785, (2023), DOI: , Copyright 2023 Royal Society of Chemistry.

R. Alex Mayo arranged the modifications to the VC-PWDF method within the critic2 code (coded by Alberto Otero de la Roza), performed the VC-xPWDF calculations, analyzed the data, made the figures, and wrote the first draft of the manuscript. KMM and RAM both performed the experimental PXRD measurements and analysis (indexing). ERJ supervised the project. All authors contributed to editing and input on the final version of the manuscript.

6.1 Introduction

Powder X-ray diffraction (PXRD) is a workhorse characterization technique in biology, chemistry, physics, and engineering. It has become an invaluable tool in industrial quality control, research and development, and academia for phase identification, quantification, and the characterization of polymorphs.¹²⁶ While PXRD is the easiest and fastest method

for obtaining fundamental information about the solid-state structure of a material, single-crystal X-ray diffraction (SC-XRD) remains the gold standard for comprehensive data on the molecular structure and periodic arrangement in three-dimensional space.

Structure determination from powder data (SDPD) is also an active and practiced method of crystal structure determination. However, high-quality powder X-ray diffraction data and access to an expert crystallographer are likely requirements, and the methods used often involve more time, constraints, and trial and error than SC-XRD before achieving success for molecular organic crystals.^{18–20} The statistical assessment of whether a proposed crystal structure can generate the powder diffractogram that is observed experimentally is done by Rietveld refinement. This non-linear least-squares refinement procedure modifies various parameters of the proposed crystal structure model and experimental conditions in order to maximize agreement between the simulated powder diffractogram and the experimental one. Rietveld refinement results in final (dis)agreement metrics, such as the weighted profile residuals (R_{wp}) and chi-squared (χ^2). While debate over how to interpret the refinement metrics is not new,⁵⁸ recent publication of four unique crystal structure models that are able to yield a reasonable refined fit to a powder diffractogram obtained from synchrotron X-ray diffraction has highlighted the inherent ambiguity that may accompany a structure solution from powder diffraction data.¹²⁷

Crystal structure prediction (CSP) uses theoretical and physical chemistry to deduce the crystal structure(s) of a given molecule or elemental composition.¹²⁸ CSP has become notable in material science as an aid to the development of porous solids^{129,130} and organic semiconductors,³² among other materials with desirable properties.^{33,131} In particular, the pharmaceutical industry sees relatively common use of CSP for drug substance development and risk reduction.^{131,132} A CSP study on a new active pharmaceutical ingredient (API) can begin as soon as the discovery team identifies it as a viable candidate, either theoretically or experimentally,^{94,133} and can predict a late-appearing polymorph, aid in the determination of crystal structures, and assess the API's propensity to form solvates.^{30,91} Sometimes, CSP will predict a more stable crystal structure than those that have been observed experimentally for an API and additional screening experiments may be performed in order to identify the conditions that yield this crystal structure, if it can be formed at all.¹³⁴

The solids generated by crystallization experiments during polymorph screening are

primarily evaluated by PXRD as the initial characterization tool. If a new PXRD pattern is observed, further characterization will ensue and there may be a need for a full structure determination using SC-XRD.¹³⁵ However, if a CSP study has already been performed, it is likely that any new polymorphs characterised by PXRD are already represented within the tens of thousands of hypothetical crystal structures generated. It would, therefore, be highly desirable to identify which of these candidates is a match to an experimental diffractogram of a new polymorph.

Crystal structures collected under the same conditions (i.e. temperature and pressure) are generally easily classified as matching or different structures by comparison of their powder diffractograms by examining the peak positions and intensities. However, once the experimental conditions differ, or one of the two crystal structures is generated/optimized computationally, the comparison becomes problematic due to the condition-induced deviation in the lattice parameters (pressure-induced contraction, thermal expansion, neglect of zero-point vibrations for “static lattice” structures optimized using computational methods, etc...). Even minor changes in the lattice dimensions result in notable shifts in the powder diffractogram. This is a common problem, as routine PXRD measurements occur at room temperature, whereas routine SC-XRD measurements are made at temperatures as low as 80 K.

Several corrections have been developed to account for the effect of lattice dimension deviations during quantitative crystal structure comparison based on simulated PXRD patterns. These include an isotropic volume correction,⁷¹ the variable-cell powder difference (VC-PWDF) method,¹³⁶ and the FIt with DEviating Lattice (FIDEL) method.⁷⁴ In this work, we will focus on the VC-PWDF method, which converts input crystal structures to their Niggli reduced cells, then screens possible unit-cell bases that may be coincident with the given reference structure, and deforms each candidate unit-cell basis to identify the matching cell, if one exists. It then yields the measure of dissimilarity of the best matching cell with the triangle-weighted cross-correlation function proposed by de Gelder et al.²³ The VC-PWDF method has been shown to be as successful as the COMPACK²² method, which compares crystal structures based on atomic positions.^{122,136}

Notably, the VC-PWDF method has shown excellent performance for comparison of simulated diffractograms from *in silico* structures and those obtained from SC-XRD collected under different experimental conditions.¹²² Thus, it forms an ideal basis for a new,

high-accuracy method for comparing the experimental powder diffractograms collected during a high-throughput screening to the crystal structures obtained from a CSP study. While the FIDEL method has shown some efficacy in this realm,^{74,127} the minimization protocol can be a lengthy procedure and is prone to errors due to local minima (*vide infra*). Therefore, we look to apply the VC-PWDF method to tackle this problem with improved accuracy and consistency of outcome.

Herein, we report the modification of our VC-PWDF method to enable direct comparison of ideal simulated powder diffractograms for known crystal structures with experimentally collected data for an unknown polymorph. The primary goal is to enable crystal structure identification from an experimental powder pattern, given a list of putative crystal structures generated computationally. The method was applied to seven example compounds (Figure 6.1) for which PXRD patterns were collected on a standard laboratory instrument. The experimental results were compared with simulated powder diffractograms calculated from both known experimental crystal structures (Cambridge Structural Database, CSD¹³⁷), and *in silico* generated crystal structures (Control and Prediction of the Organic Solid State, CPOSS, database¹³⁸). Our method was found to consistently identify the correct crystallographic form from the relevant database(s) as the structure matching the experimental powder diffractogram based on minimum powder difference scores.

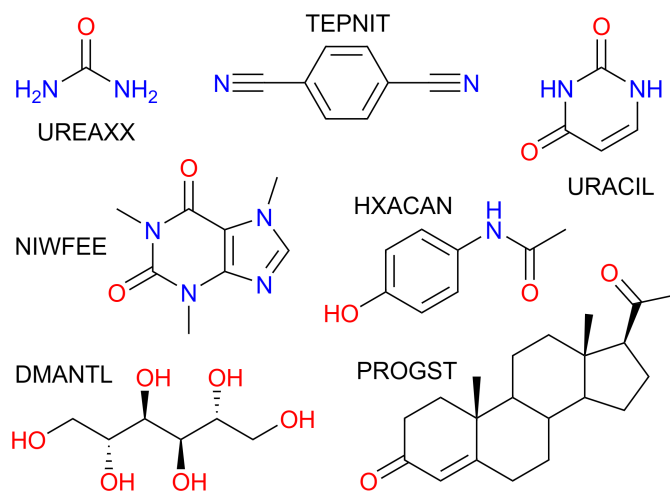


Figure 6.1: Compounds studied and accompanying CSD refcode family. UREA: urea; TEPNIT: 1,4-dicyanobenzene; URACIL: uracil; NIWFEE: caffeine; HXACAN: acetaminophen; DMANTL: D-mannitol; PROGST: (+)-progesterone.

6.2 Results and discussion

The VC-PWDF method was modified in two distinct ways. First, the code was changed to accommodate experimental powder diffractogram data (intensity vs. 2θ in degrees as a .xy file) and unit-cell dimensions as one of the inputs used for comparison. The other input is a crystal structure file from which the ideal powder diffractogram is simulated for comparison. Next, the code was modified to perform a basic normalization of the experimental PXRD data in the .xy file by subtracting the lowest intensity value recorded in the experimental powder diffractogram from all data-points and scaling the highest intensity peak to a value of 100. No further processing of the diffractograms or consideration of sample, instrument, or diffraction conditions was performed. In order to distinguish the results that come from the comparison of two simulated powder patterns (VC-PWDF method/score) from those that compare a simulated powder pattern with an experimental one, we will use VC-xPWDF method/score for the latter. The method is available within the developers version of critic2.^{75,139} While a more complex baseline correction may be required to see similar performance to that observed herein if the PXRD were collected using a different experimental set-up (eg. capillary in transmission mode), this processing could be done prior to analysis with the VC-xPWDF method.

We obtained the experimental powder diffractograms of seven chemicals that were readily available at the University of Guelph (urea, 1,4-dicyanobenzene, uracil, caffeine, acetaminophen, D-mannitol, (+)-progesterone). The collected powder diffractograms for each of the 7 compounds are shown in Figure C.1, and we consider them to be of moderate quality. The 20 most intense peaks were picked and used as input for indexing with the CrysFire2020 suite,¹⁴⁰ which facilitates running of multiple indexing algorithms (including TAUP,¹⁴¹ ITO,⁵⁹ TREOR,⁶⁰ KOHL,¹⁴² and DICVOL⁶¹). The powder diffractogram of urea contained only 13 well-defined peaks, so only these 13 peaks were used for indexing. The caffeine diffractogram contained many peaks, some of which became quite broad beyond $2\theta = 30^\circ$, so the 24 most intense peaks observed before this angle were used for indexing. The cell dimensions with the highest de Wolff's⁶² figure of merit (summarized in Table C.2) were used as input to the VC-xPWDF method, accompanying the experimental powder diffractogram.

Lists of *in silico* generated structures for the compounds studied were obtained from the CPOSS database. These structure-energy landscapes were screened for duplicate

crystallographic forms and structure(s) matching the known experimental structure(s) present in the CSD. Details regarding these data are provided in Appendix C. The landscape for progesterone includes mostly racemic crystal structures, with only 8 of the 149 crystal structures being enantiopure. The landscape for mannitol provides a more equal number of enantiopure structures of nearly 50% (250/546). All other molecules are achiral and so the presence or absence of reflection symmetry elements in the crystal lattices is not of concern in this study. In addition to the *in silico* generated structures, SC-XRD determined structures of one or more known polymorphs of the 7 compounds were obtained from the CSD, with data collected over a range of temperatures (see Table 6.1 for the refcodes). It should be noted that the VC-PWDF method is currently unable to work with disordered structures, so any such polymorphs were omitted for this work. In particular, the crystal structure NIWFEE03 ($Z' = 5$ and $Z = 20$) is the only non-disordered structure of caffeine in the CSD. While the disordered $C2/c$ structure for the β polymorph is the correct structure solution,¹⁴³ its simulated powder pattern is nearly indistinguishable from that of the ordered Cc structure (NIFWEE03), which was used to represent the β form of caffeine throughout this study.

The CSD structure refcodes, experimental conditions under which the measurements were made, and resulting VC-xPWDF scores from comparison with the experimental PXRD patterns are summarized in Table 6.1. For all 7 compounds studied, the CSD structures corresponding to the same polymorph as the experimental PXRD pattern were found to give the smallest VC-xPWDF score, regardless of the conditions under which they were obtained. Comparison of the collected powder patterns with CSD structures of different polymorphs results in much higher VC-xPWDF scores (Table 6.1). Additionally, the VC-xPWDF method was able to identify the matching structure for urea and (+)-progesterone, even though the indexed cell parameters obtained from their experimental powder diffractograms were not the same as (or an obvious sub/super-cell of) the matching CSD structure (Table C.2). The VC-xPWDF method is perfectly suited to make use of a viable indexed unit cell from an experimental powder diffractogram, whether it is a supercell, subcell, or some non-standard description of the same lattice, due to its exploration of viable unit cells for each crystal structure. This should be useful for *in-situ* PXRD to determine if a phase transition occurs with changes in temperature and/or pressure.

Table 6.1: VC-xPWDF scores from comparison of the collected powder diffractograms with the CSD structures. (-) URACIL has only one known crystal structure and (*) NIWFEE is an ordered description of the β form of caffeine.

CSD refcode	Conditions	Form	VC-xPWDF
UREAXX07 ¹⁴⁴	123 K	I	0.0335
UREAXX11 ¹⁴⁵	60 K	I	0.0337
UREAXX12 ¹⁴⁶	12 K	I	0.0339
UREAXX23 ¹⁴⁷	ambient	I	0.0364
UREAXX26 ¹⁴⁸	3.1 GPa	IV	0.2087
UREAXX33 ¹⁴⁹	1.0 GPa	III	0.2528
TEPNIT04 ¹⁵⁰	ambient	β	0.0326
TEPNIT14 ¹⁵¹	100 K	β	0.0330
TEPNIT06 ¹⁵²	ambient	α	0.4339
URACIL ¹⁵³	ambient	-	0.0290
NIWFEE03 ¹⁵⁴	ambient	*	0.0114
HXACAN35 ¹⁵⁵	ambient	I	0.0494
HXACAN04 ¹⁵⁶	150 K	I	0.0602
HXACAN15 ¹⁵⁷	80 K	I	0.0633
HXACAN13 ¹⁵⁷	20 K	I	0.0670
HXACAN09 ¹⁵⁸	1 GPa	I	0.0772
HXACAN47 ¹⁵⁹	200 K	VII	0.4099
HXACAN40 ¹⁶⁰	ambient	III	0.5764
HXACAN33 ¹⁶¹	ambient	II	0.7293
DMANTL15 ¹⁶²	100 K	β	0.0962
DMANTL07 ¹⁶³	ambient	β	0.0992
DMANTL08 ¹⁶⁴	100 K	α	0.3145
DMANTL14 ¹⁶⁵	ambient	δ	0.4352
PROGST12 ¹⁶⁶	150 K	I	0.0416
PROGST10 ¹⁶⁷	ambient	I	0.0428
PROGST13 ¹⁶⁶	150 K	II	0.3426

The results of comparisons between the experimentally obtained diffractograms and those simulated from the crystal structures in both the CPOSS database and CSD are shown in Figure 6.2. These results clearly demonstrate the ability of the VC-xPWDF method to identify the correct polymorph from the known forms of these compounds in all cases. Further, the results in Figure 6.2 also show that, if a matching CPOSS structure exists, this structure is consistently ranked just after/amongst the experimental structure(s) for that

polymorph. These rankings showcase the ability of the VC-xPWDF method to identify the most similar *in silico* generated structure according to the powder difference score as well.

For 1,4-dicyanobenzene, acetaminophen, D-mannitol, and (+)-progesterone, the plots in Figure 6.2 show good separation between the matching and non-matching structures. In these cases, there is clarity in which of the structure(s) match(es) the experimental powder diffractogram, and which structures do not. Additionally, all matching structures from both the CSD and CPOSS database yield VC-xPWDF scores of less than 0.1 when compared to the experimental powder diffractogram. However, the plots for urea, uracil, and caffeine show multiple structures with VC-xPWDF scores less than 0.1. Based on the results for our small data set, we propose that a structure with a VC-xPWDF score below 0.1 is grounds to consider it a potential match, but does not guarantee it. This will of course vary with the quality of the powder diffractogram as well (*vide infra*).

For the case of urea, the reason that such a large number (42) of CPOSS structures have VC-xPWDF scores < 0.1 can be explained by the fact that the powder diffractogram is dominated by a single, high-intensity peak. Thus, if a candidate structure also has a peak at this position, much of the diffractogram intensity is already overlaid, resulting in a low powder difference score. For this compound, a quick glance at the diffractogram overlays quickly eliminates the non-matching structures and makes it evident that the CPOSS landscape does not include the experimental polymorph (Figure C.4). The visual comparisons with the caffeine overlays (Figure C.5) tell a similar story; there are a couple positions of high intensity peaks in the diffractogram, and low powder difference scores can still be obtained from cases where the remaining small intensity peaks do not overlap well.

The three *in silico* generated crystal structures of uracil with VC-xPWDF scores between 0.06 – 0.1 can also be reasonably discounted as matches with a visual assessment; however, structure ID am82 (VC-xPWDF score of 0.0358) cannot (Figure C.6). Even agreement values from Rietveld refinement are insufficient to exclude the possibility of am82 being a match to the experimental powder diffractogram (Table C.4). Comparing the two *in silico* generated crystal structures am7 (matching crystal form) and am82 to one another yields a VC-PWDF score of 0.0238 and the distorted structures obtained after processing with the VC-xPWDF method to match the experimental powder diffractogram yield a 20/20 match with $\text{RMSD}(20) = 0.247 \text{ \AA}$ with COMPACK (default tolerances). The similarity of

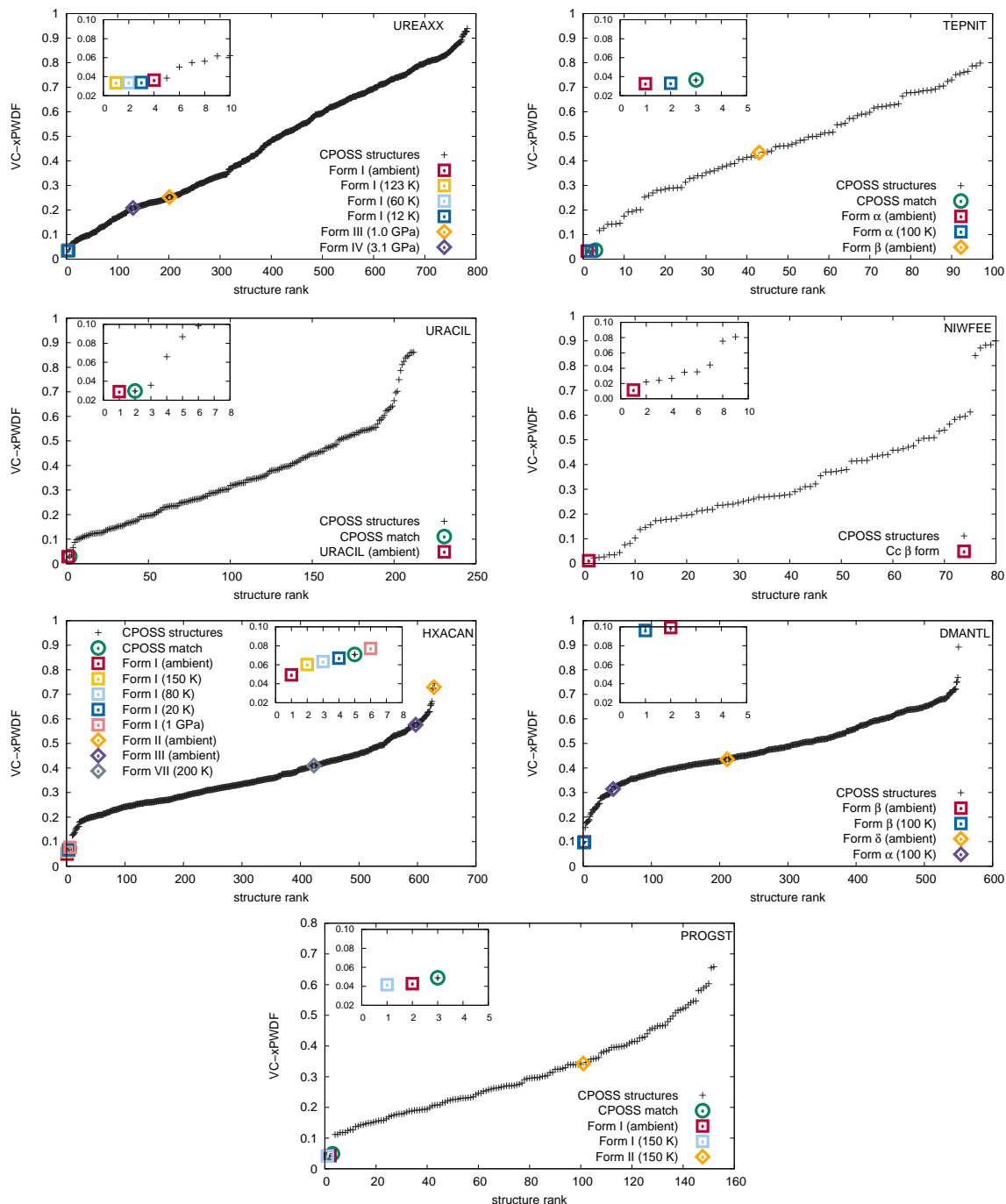


Figure 6.2: Plots showing the computed VC-xPWDF scores resulting from comparison of each input crystal structure to the experimental powder diffractogram collected for that compound. The structures are ranked by lowest VC-PxWDF score (most similar) and the insets provide views of the best matching structures with VC-xPWDF scores < 0.1, up to a maximum of 10 for clarity. The point types indicate the source of each crystal structure: squares correspond to CSD structures of the same polymorph as the sample studied, diamonds are different polymorphs of that compound from the CSD, and + signs are *in silico* generated structures from the CPOSS database. The CPOSS structure that corresponds to the same polymorph as the experimental sample (if a matching structure was generated) is identified with a green circle around that data point.

the packing of the uracil molecule in these two (modified) crystal structures is shown in Figure C.7, and we would expect them to converge to the same structure after geometry optimization with density-functional theory.

The VC-xPWDF scores obtained from comparison of the matching crystal structures to the collected powder diffractogram of D-mannitol are considerably higher than for the other compounds investigated. The overlay of the simulated PXRD pattern for DMANTL07 with the collected experimental diffractogram is shown in Figure 6.3. Based on this overlay, the reason for the higher scores can be clearly attributed to preferred orientation (the biased orientation of one or more crystallographic planes in the experimental sample) leading to a change in relative intensities of the peaks. Because the POWDIFF score considers differences in the peak intensities in addition to their position, a significant deviation from the ideal diffractogram will yield a higher score.

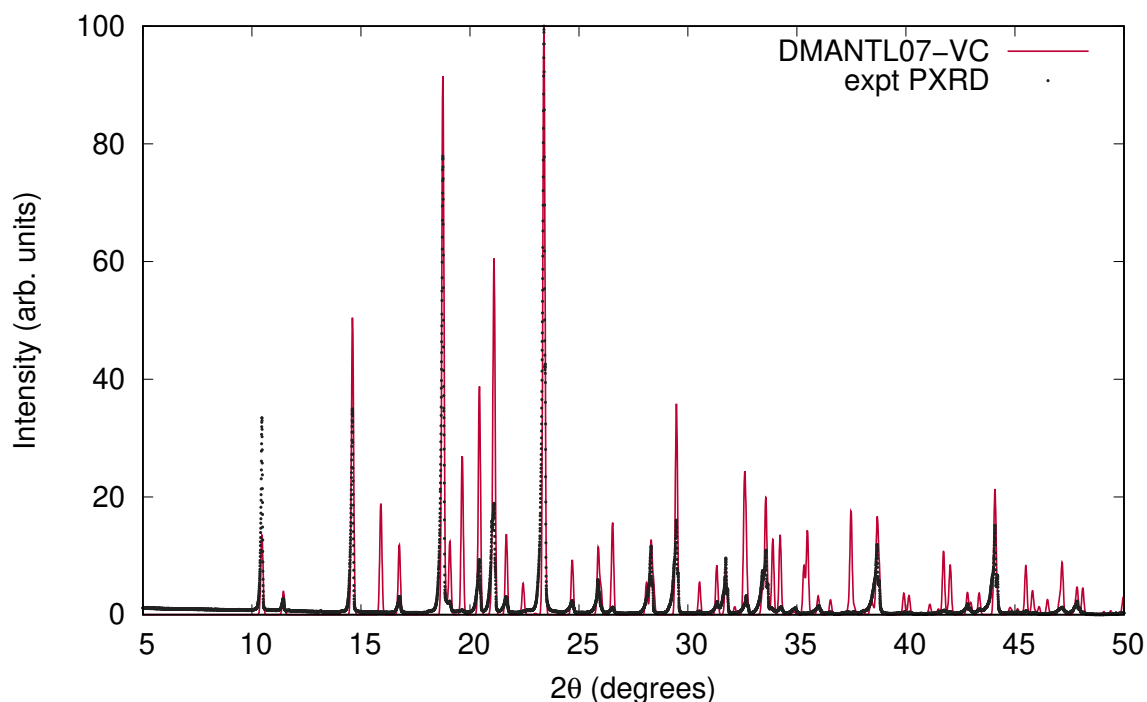


Figure 6.3: Overlay of the experimentally collected powder diffractogram and the simulated powder diffractogram for DMANTL07 after the VC-xPWDF protocol.

As is to be expected, the quality of the powder diffractogram has an effect on its measured similarity to an ideal simulated powder diffractogram. It is interesting that peak shapes, a flat baseline, and other features commonly associated with “high quality” PXRD data are rather easily obtained in adequate quality, and/or have a relatively minimal

effect on the resultant similarity scores measured within this dataset (see Figures C.1 and C.2 for all diffraction patterns) when compared to the considerable effect from preferred orientation.

To test the degree to which the results would change with “lower quality” data, quick 2-minute scans of the prepared samples were collected. The results are nearly indistinguishable from those obtained with the “higher quality” data from the 3 hour scan. The analogous plots to Figure 6.2 using the screening-scan diffractograms are shown in Figure C.8. Provided the diffractogram yielded by a 2-minute screening scan can provide a valid indexed unit cell, these data are perfectly acceptable for comparison with the VC-xPWDF method. Clearly, this is ideal as a complement to high-throughput polymorph screening in order to identify the crystal structures of the various forms analyzed by short screening PXRD data collection, provided one has access to a CSP landscape.

Rietveld refinement is a common final step when assessing whether a proposed crystal structure model matches an experimentally collected powder diffractogram. Accordingly, we have performed Rietveld refinement on an assortment of CPOSS and CSD structures that yielded low VC-xPWDF scores using the automated BGMN protocol¹⁶⁸ implemented in the Profex software¹⁶⁹. The outcomes are tabulated in Table C.3. When refining an experimental crystal structure from the CSD to the experimental powder diffractogram, the best R_{wp} values were obtained by starting from the CSD structures solved from data collected under ambient conditions. Refinement attempts starting with CSD structures of the matching form collected at high pressure or low temperature almost always yielded overlays where peaks were misaligned. The application of the VC-xPWDF method to the CSD structures determined under ambient conditions improved the agreement factors considerably (Section C.4.1), highlighting the utility of the VC-xPWDF method in providing the best starting point for a Rietveld refinement.

It would be ideal if Rietveld refinement could be used to confirm or rule out structures with low VC-xPWDF scores as a match to the experimental polymorph. However, within this dataset, the absolute R_{wp} and χ^2 values provide little additional evidence in deciding whether or not a crystal structure matches the experimental powder diffractogram. Many refinements give poor peak overlays, but still yield R_{wp} values in the 20-30% range, while many successful refinements with good peak overlays yield R_{wp} values that are <40%. It may be the case that a more tailored approach is required to fully and carefully refine these

data, or that the PXRD data collected are simply not of high enough quality for conclusive refinement. Data collection at a synchrotron source may eliminate the issues outlined in the latter case. This again highlights the advantage of the VC-xPWDF method as it appears to provide information equivalent to Rietveld refinement without requiring specialized expertise, or very high quality PXRD patterns as input.

A current drawback of the VC-xPWDF method is its requirement of input unit-cell dimensions for the reference structure. Thus, for experimental powder diffractograms, indexing is a must. Conversely, a major advantage of the FIDEL method is that it can run successfully without knowledge of unit-cell dimensions. With the use of the autoFIDEL code,¹⁷⁰ we applied the FIDEL method to our dataset (Figures C.9 and C.10). FIDEL is able to identify the matching polymorph from the CSD (determined under ambient conditions) for all cases except acetaminophen (HXACAN) using the default run parameters. Because the minimization protocol of FIDEL is a more computationally expensive approach to aligning the diffractogram peak positions, the program sets a minimum initial agreement that must be met in order for the protocol to run, the default is a POWDIFF score < 0.7 . The initial agreement between the simulated powder pattern of HXACAN35 and our collected powder diffractogram is a POWDIFF score of 0.7325, and so the powder difference score of 0.1383 post-minimization is only obtained if the default parameters are modified to allow the optimization. With this adjustment in the run parameters, HXACAN35 is correctly identified as the best matching crystal structure with autoFIDEL.

The default run parameters are the reason the rank-plots of the FIDEL results (Figure C.9) only include a fraction of the total number of structures ranked by our VC-xPWDF method, as only the structures that undergo the minimization protocol are included with their accompanying post-minimized powder difference score. Even with the reduction in the number of structures run by autoFIDEL, some notable differences are identified for the optimization of the *in silico* generated matching crystal structures of acetaminophen and 1,4-dicyanobenzene. The most extreme example is the latter.

When the original CPOSS structure list for 1,4-dicyanobenzene (containing duplicates) is screened against the experimental powder diffractogram by the VC-xPWDF method, the several equivalent matching structures are identified and grouped together at the lowest powder difference score (Figure 6.4, top-left). Conversely, the same screening using

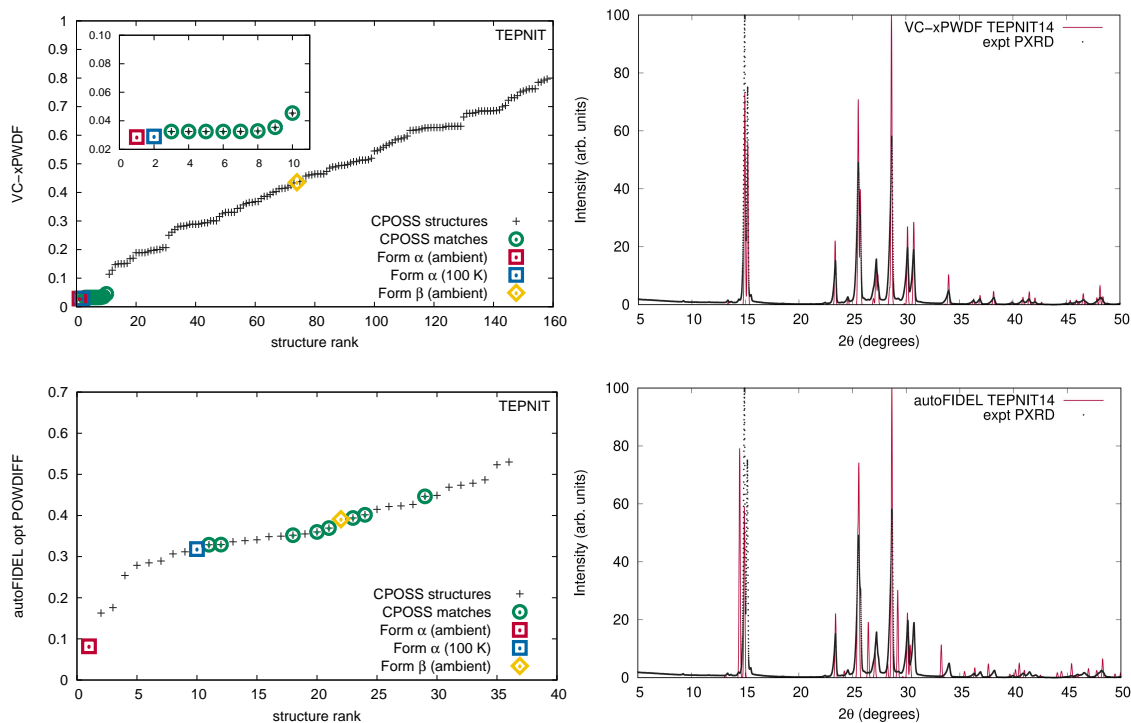


Figure 6.4: Left: (VC-x)PWDF scores resulting from comparison of CSD and CPOSS input structures with the experimental powder diffractogram for 1,4-dicyanobenzene computed using VC-xPWDF (top), and autoFIDEL (bottom). Right: Overlays of the experimental and simulated TEPNIT14 powder diffractogram after correction with VC-xPWDF (top), showing perfect alignment of the peak positions, and after minimization with autoFIDEL (bottom), which leaves many peaks poorly positioned.

autoFIDEL ranks the matching structures haphazardly at various powder difference scores (Figure 6.4, bottom-left). Even the experimental structure that is determined at 100 K (TEPNIT14) is not minimized to a low powder difference score with autoFIDEL. Thus, if no ambient temperature crystal structure solution was available for comparison, FIDEL would fail to identify the matching polymorph, despite multiple descriptions of it being present in the list of structures being screened.

While cases analogous to 1,4-dicyanobenzene and acetaminophen may be relatively few, we showcase here an example of the minimization protocol used in autoFIDEL getting caught in a local minimum. The result is misaligned peaks and the inability to identify the matching crystal structure. For comparison, the overlays of the experimental powder diffractogram and simulated diffractogram of the TEPNIT14 structure after modification by

both the VC-xPWDF method and autoFIDEL are also presented in Figure 6.4. The advantage of the VC-xPWDF method, provided the experimental diffractogram can be indexed, is that it will correctly align the peak positions of the simulated powder diffractogram of the matching structure directly.

6.3 Conclusion

In this work, we illustrated the ability of the VC-xPWDF method to clearly identify the most similar crystal structure to both moderate and “low” quality experimental powder diffractogram for a set of 7 representative organic compounds. In all cases, matching SC-XRD structures obtained from the CSD were identified by having the lowest VC-xPWDF scores of any crystal structures searched. As competing polymorphs consistently yielded much higher VC-xPWDF scores, the method is able to rapidly identify which of several literature polymorphs matches an experimental sample, even if the structure was solved for very different temperature and pressure conditions.

The modification of the VC-PWDF method to allow an experimental PXRD pattern as input has converted this code from being a tool exclusively used for the comparison of solved/complete crystal structures to one of a select few methods that is able to quantitatively assign a crystal structure to an experimentally collected powder diffractogram. The various other PXRD-based methods for the comparison of crystal structures show poor performance in general because of thermal expansion/pressure induced contraction, and thus are generally ineffective in the assignment of the matching *in silico*-generated structure to a powder diffractogram that is collected under screening-like conditions (eg. 2 minute scan at room temperature). The VC-xPWDF method directly address this research problem.

The principle limitation of the VC-xPWDF method is that it must be provided with valid indexed unit-cell parameters to accompany the experimental powder diffractogram. Therefore, the method cannot be applied if the experimental diffractogram cannot be successfully indexed. This stands in contrast to the FIDEL method, which does not require indexing. However, we have provided an example here of the risks involved with the FIDEL approach, and the advantages of using the VC-xPWDF method when the indexed unit-cell parameters can be determined. Future development of the VC-PWDF method will seek to eliminate the requirement of the indexed unit-cell parameters.

The broader utility of the VC-xPWDF method would be to identify a previously uncharacterized crystal structure from a list of candidates generated during first-principle crystal structure prediction. This would be of particular value to the pharmaceutical industry for polymorph screening, as well as in the development of porous solids and organic electronics, and for other materials research where design using CSP might be applied. Here, for all 4 cases where a list of *in silico* generated structures contained a match to the experimental polymorph, the VC-xPWDF method successfully identified the matching structure(s) as having the lowest powder difference score of the candidates. However, for an unknown compound, there is no guarantee that a CSP landscape will contain the experimental polymorph, so there remains the issue of confidence that the structure with the lowest VC-xPWDF score is the actual matching structure. Similar to Rietveld refinement, a small powder difference score (< 0.1) does not always provide conclusive evidence that the proposed crystal structure matches the experimental powder diffractogram. However, a visual assessment of the diffractogram overlay, which is also a recommendation following Rietveld refinement, can provide increased confidence in the result.

In practice, CSP studies typically use force-field methods for structure generation. However, since the relative energies from force-field methods are often poor, a re-ranking of up to several hundred low-energy structures may be performed using dispersion-corrected density-functional theory (DFT) methods to provide a more accurate energy landscape. Thus, additional confidence in deciding which, if any, of several candidate structures with low VC-xPWDF scores is the experimental match could be gained by also considering the relative DFT energies. Structures with both low energy and low VC-xPWDF scores are more likely matches, while candidates with a low VC-xPWDF score but high relative energy would be less likely to correspond to the experimental polymorph. In future work, we will consider combining the VC-xPWDF method with such CSP information to solve unknown crystal structures from powder data.

CHAPTER 7

CONCLUSIONS AND FUTURE WORK

7.1 Conclusions

The development of the variable-cell protocol and application of the variable-cell powder difference (VC-PWDF) and variable-cell experimental powder difference (VC-xPWDF) methods has provided two valuable tools to the molecular materials and pharmaceutical development research communities. The VC-PWDF method is demonstrably the most accurate and robust method to-date for the comparison of crystal structures using their simulated powder diffractograms. The VC-xPWDF method maintains this accuracy when applied to the problem of comparing a moderate/screening-quality experimental powder diffractogram to one simulated from a crystal structure.

The ability of the VC-PWDF method to identify matching *in silico*-generated crystal structures to target experimental crystal structures is assessed in Chapter 4 using the submissions to the 6th crystal structure prediction (CSP) blind test.⁴¹ The superiority of the VC-PWDF method relative to the other assessed powder diffraction-based comparison methods is demonstrated by its ability to distinguish a polytype from a target structure, and by the identification of two uncredited matching structures in the original test results. The VC-PWDF score is shown to be as useful as the root-mean-square deviation (RMSD) in atomic positions as a measure of structure similarity.

Chapter 4 assessed the popular COMPACK method *vis a vis* powder X-ray diffraction (PXRD)-based comparison methods using a dataset of 44,939 structure pairs employed in a previous study by Sacchi *et al.*⁶⁸. The VC-PWDF method substantially improves the agreement with COMPACK (2.84% total disagreement), compared to the CCDC PXRD-based comparison tool (12.35%). By analyzing the structure pairs for which COMPACK

and VC-PWDF disagree, the strengths and weaknesses of each method were explored. COMPACK has a counter-intuitive dependence on its tolerance parameters, by which structures that are considered the same at a given tolerance are viewed as different at a looser tolerance. COMPACK's RMSD(N) can also increase with increasing tolerance values for a fixed number of matching molecules (N). A few additional weaknesses of COMPACK include: a) extremely costly or incorrect comparisons for molecules with highly-branched substituents, b) difficulties with molecules presenting helical chirality, and c) requirements for very large cluster sizes (up to 50 molecules) that are sometimes needed to correctly identify unequal polymorphs. In turn, VC-PWDF has difficulty differentiating structures with similar packings, such as polytypes, and conformational and isomorphous phases. It is shown that the proposed VC-PWDF method is at least as robust as COMPACK for comparing molecular crystal structures and we recommend using a combination of both methods to provide more confidence in structure comparisons.

The VC-PWDF method was expanded to allow matching of experimental powder diffractograms of unknown polymorphs to simulated powder diffractograms of both experimental crystal structures from the Cambridge Structural Database and *in silico*-generated structures from the Control and Prediction of the Organic Solid State database. This VC-xPWDF method is presented in Chapter 6 and correctly identifies the most similar crystal structure to both moderate and “low” quality experimental powder diffractograms for a set of 7 representative organic compounds. Features of the powder diffractograms that are more challenging for the VC-xPWDF method are discussed (i.e. preferred orientation), and comparison with the FIDEL method showcases the advantage of VC-xPWDF provided the experimental powder diffractogram can be indexed. The VC-xPWDF method should allow rapid identification of new polymorphs from solid-form screening studies, without requiring single-crystal analysis.

The VC-(x)PWDF methods are anticipated to become widely used wherever polymorphism is present and/or CSP is utilized. It is critically important that CSP methods be evaluated accurately in their ability to produce known crystal structures. Accordingly, using the VC-PWDF method in combination with the COMPACK method (or an alternative atomic position-based method) provides the best approach for identifying true positive cases from tens to hundreds of thousands of structures. The application of the VC-xPWDF method for SDPD using a CSP dataset and solid form screening experimental PXRD has

the potential to improve the rate of materials development, and de-risking pharmaceutical development such that unfortunate cases of late-stage polymorphism are eliminated.

7.2 Future work

There are two main branches along which future developments of the VC-xPWDF method could be made. One path leads to the end-goal of eliminating the requirement for indexed unit-cell dimensions as an input, easing the requirements for its application; the other involves the utilization of the indexed cell information along with a CSP protocol in order to solve crystal structures from powder data starting exclusively from the experimental PXRD data.

The FIDEL method avoids indexing requirements through use of a hill-climbers algorithm to optimize agreement between the simulated and experimental PXRD. This, however, is prone to local maxima if the two structures are related by a significant distortion and/or have a complex relationship (anisotropy). Efforts to decouple the VC-PWDF method from the use of the target unit-cell dimensions would be challenging and probably require additional compute power, and/or specially developed algorithms in order to be successful. Whereas the choice for the unit-cell distortions are stochastic with the FIDEL method, the approach for a modified VC-PWDF protocol would be to deliberately explore the distortions in each crystallographic direction.

The integration of the VC-xPWDF method with a successful indexing and structure generation program would yield a novel method of SDPD. The wrapper program *Crysfire2020* provides a user with a common interface for 8 indexing programs, including the most common ITO, TAUP, TREOR, and DICVOL algorithms. The use of various indexing approaches improves the odds of determining a viable cell for the crystal structure of interest. Dimensions of a viable unit cell can be used as constraints within a CSP structure generation program that either (i) uses a computed energy (from a force field or electronic structure method) as the figure of merit for a biased generation algorithm (simulated annealing, parallel tempering, genetic/evolutionary algorithm, etc.), or (ii) performs an efficient grid-search (i.e. using Sobol' sequences). Either would create a list of CSP structures that could be screened using VC-xPWDF in order to identify the structure that gives the best match to the experimental diffractogram.

SDPD methods that use a combined figure of merit of energy and pattern fitting exist;

however, they typically employ very simple energy contributions, such as anti-bumping functions, and the pattern fitting algorithm is a poor director toward the correct structure. It is proposed that using energy as the primary contributor to the figure of merit will improve on the success of SDPD. With the unit-cell dimensions as a constraint on the CSP search space (i.e. constrained crystal structure prediction, C-CSP), the compute time and resources for such a search would be substantially less than an *ab initio* CSP study, and reduce user-defined parameters. The use of a combined figure of merit that includes VC-xPWDF in generating a “fitness” score for a biased algorithm could further improve the efficiency of the protocol. Ultimately, the development of a C-CSP protocol has the potential to allow routine SDPD from screening-quality data, accelerating material science discoveries and pharmaceutical development.

APPENDIX A

SUPPLEMENTARY INFORMATION FOR: IMPROVED QUANTITATIVE CRYSTAL-STRUCTURE COMPARISON USING POWDER DIFFRACTOGRAMS VIA ANISOTROPIC VOLUME CORRECTION

A.1 Target structures

Table A.1: CCDC identifiers for the BT6 target structures.

Compound	Form	Identifier
XXII	–	1451239
XXIII	A	1447522
XXIII	B	1447523
XXIII	C	1447524
XXIII	D	1447525
XXIII	E	1447526
XXIV	–	1447530
XXV	–	1447527
XXVI	–	1447529

A.2 Dataset

A.2.1 Lists removed from all BT6 submissions

Table A.2: Lists of BT6 submissions removed prior to analysis.

<i>Target-Group-List</i>	Reason
XXII-G03-L2	structural duplicate of L1
XXV-G03-L2	structural duplicate of L1
XXII-G07-L2	structural duplicate of L1
XXIII-G07-L2	structural duplicate of L1
XXV-G07-L2	structural duplicate of L1
XXII-G12-L1	numerous issues
XXII-G12-L2	structural duplicate of L1
XXVI-G14-L2	structural duplicate of L1
XXII-G25-L2	structural duplicate of L1
XXIII-G25-L2	structural duplicate of L1

G14 submitted lists containing structures with $Z' = 1$ only, and a mix of $Z' = 1$ and $Z' = 2$, for compounds XXIII and XXVI. Only the list with the Z' value matching the target was used in each case to avoid double counting. Thus, comparisons with targets XXIIIA, XXIIIB, and XXIIID used XXIII-G14-L1 (only $Z' = 1$). Comparisons with targets XXIIIC, and XXIIIE used XXIII-G14-L2 (mix of $Z' = 1$ and $Z' = 2$).

A.2.2 Data processing

Many groups submitted lists with problematic symmetry descriptions or errors in unit cell angles for the assigned crystal systems. Corrections were made following lists:

- XXII-G04-L1: symmetry descriptions (add `_symmetry_equiv_pos_site_id` between `loop_` and `_symmetry_equiv_pos_as_xyz`)
- XXV-G04-L1: symmetry descriptions (add `_symmetry_equiv_pos_site_id` between `loop_` and `_symmetry_equiv_pos_as_xyz`)
- XXVI-G04-L1: symmetry descriptions (add `_symmetry_equiv_pos_site_id` between `loop_` and `_symmetry_equiv_pos_as_xyz`)
- XXII-G05-L1: symmetry descriptions (add `loop_` and re-order elements in `_symmetry_equiv_pos_as_xyz`)
- XXIII-G05-L1: symmetry descriptions (add `loop_` and re-order elements in `_symmetry_equiv_pos_as_xyz`)
- XXIV-G05-L1: symmetry descriptions (add `loop_` and re-order elements in `_symmetry_equiv_pos_as_xyz`)
- XXV-G05-L1: symmetry descriptions (add `loop_` and re-order elements in `_symmetry_equiv_pos_as_xyz`)
- XXVI-G05-L1: symmetry descriptions (add `loop_` and re-order elements in `_symmetry_equiv_pos_as_xyz`)
- XXII-G06-L1: unit cells given a monoclinic space group without two right angles and/or an orthorhombic space group without all right angles
- XXII-G06-L2: unit cells given a monoclinic space group without two right angles and/or an orthorhombic space group without all right angles
- XXIII-G06-L2: unit cells given a monoclinic space group without two right angles and/or an orthorhombic space group without all right angles
- XXV-G06-L2: unit cells given a monoclinic space group without two right angles and/or an orthorhombic space group without all right angles

- XXVI-G06-L2: unit cells given a monoclinic space group without two right angles and/or an orthorhombic space group without all right angles
- XXII-G08-L2: unit cells given a monoclinic space group without two right angles and/or an orthorhombic space group without all right angles
- XXII-G20-L1: symmetry descriptions (missing, wrong space group H-M notation)
- XXII-G23-L1: unit cells given a monoclinic space group without two right angles and/or an orthorhombic space group without all right angles
- XXII-G25-L1: unit cells given a monoclinic space group without two right angles and/or an orthorhombic space group without all right angles
- XXIII-G25-L1: unit cells given a monoclinic space group without two right angles and/or an orthorhombic space group without all right angles
- XXIV-G25-L1: unit cells given a monoclinic space group without two right angles and/or an orthorhombic space group without all right angles
- XXV-G25-L1: unit cells given a monoclinic space group without two right angles and/or an orthorhombic space group without all right angles

A.3 Cell transformation matrices

Transformation matrices used in the structure screening:

Acute-angle triclinic cells:

$$\begin{array}{cccc}
 \begin{bmatrix} -1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} & \begin{bmatrix} 1 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix} & \begin{bmatrix} 1 & -1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} & \begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix} \\
 \begin{bmatrix} 1 & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} & \begin{bmatrix} 0 & 1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} & \begin{bmatrix} -1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} & \begin{bmatrix} 0 & -1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \\
 \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix} & \begin{bmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ -1 & 0 & 1 \end{bmatrix} & \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ -1 & 0 & 1 \end{bmatrix} & \begin{bmatrix} 0 & 0 & -1 \\ 0 & -1 & 0 \\ 1 & 0 & -1 \end{bmatrix} \\
 \begin{bmatrix} -1 & 0 & 1 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} & \begin{bmatrix} 1 & 0 & -1 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix} & \begin{bmatrix} 1 & 0 & -1 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} & \begin{bmatrix} -1 & 0 & 1 \\ 0 & -1 & 0 \\ -1 & 0 & 0 \end{bmatrix} \\
 \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & 1 \end{bmatrix} & \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 1 & 0 \end{bmatrix} & \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & -1 \end{bmatrix} & \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & -1 & 0 \end{bmatrix} \\
 \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & -1 \end{bmatrix} & \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 1 \end{bmatrix} & \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & -1 & 1 \end{bmatrix} & \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}
 \end{array}$$

Obtuse-angle triclinic cells:

$$\begin{array}{cccc}
 \begin{bmatrix} 1 & 1 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} & \begin{bmatrix} 1 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix} & \begin{bmatrix} -1 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} & \begin{bmatrix} -1 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix} \\
 \begin{bmatrix} -1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} & \begin{bmatrix} 0 & -1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix} & \begin{bmatrix} 1 & 0 & 0 \\ -1 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} & \begin{bmatrix} 0 & 1 & 0 \\ -1 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \\
 \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 1 & 0 & 1 \end{bmatrix} & \begin{bmatrix} 0 & 0 & -1 \\ 0 & -1 & 0 \\ 1 & 0 & 1 \end{bmatrix} & \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ -1 & 0 & -1 \end{bmatrix} & \begin{bmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \\ -1 & 0 & -1 \end{bmatrix} \\
 \begin{bmatrix} 1 & 0 & 1 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} & \begin{bmatrix} 1 & 0 & 1 \\ 0 & -1 & 0 \\ -1 & 0 & 0 \end{bmatrix} & \begin{bmatrix} -1 & 0 & -1 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} & \begin{bmatrix} -1 & 0 & -1 \\ 0 & -1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \\
 \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & -1 \end{bmatrix} & \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & -1 & 0 \end{bmatrix} & \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & -1 \\ 0 & 0 & 1 \end{bmatrix} & \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & -1 \\ 0 & 1 & 0 \end{bmatrix} \\
 \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & 1 \end{bmatrix} & \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 1 \end{bmatrix} & \begin{bmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & -1 \end{bmatrix} & \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & -1 \end{bmatrix}
 \end{array}$$

Only the 12 matrices in the two left-most columns are used for monoclinic cells, as they become symmetric with the results from using the matrices in the two right-most columns in the monoclinic crystal system.

While some of these transformation matrices do not yield a determinant of 1, `critic2` is able to convert them to an appropriate transformation matrix. Alternative matrices than those shown here (other than the $\det=1$ equivalents) are not viable as they will yield a unit cell that either 1) changes an angle from acute to obtuse or vice versa, 2) dramatically increases an axis length, or both simultaneously. Note that Niggli-reduced cells will have an obtuse angle for the non-right angle of a monoclinic cell, and will have all acute or all obtuse angles for a triclinic cell. Changing one angle from acute to obtuse (or vice versa) generates an incompatible unit cell for the developed volume correction.

Table A.3: Transformation matrices applied to six structures identified as matches in BT6 in order to apply the anisotropic volume correction.

Structure	Transformation matrix		
XXIIIB-G09-L1-E13	[-1 0 0]	[-1 1 0]	[0 0 -1]
XXIIIB-G13-E88	[-1 0 0]	[-1 1 0]	[0 0 -1]
XXIIIB-G15-E13	[-1 0 0]	[-1 1 0]	[0 0 -1]
XXIIID-G06-L1-E73	[1 0 0]	[0 1 1]	[0 -1 0]
XXV-G05-L1-E01	[1 0 0]	[-1 -1 0]	[0 0 -1]
XXVI-G06-L1-E08	[-1 0 0]	[-1 1 0]	[0 0 -1]

A.4 RMSD drift from BT6 results

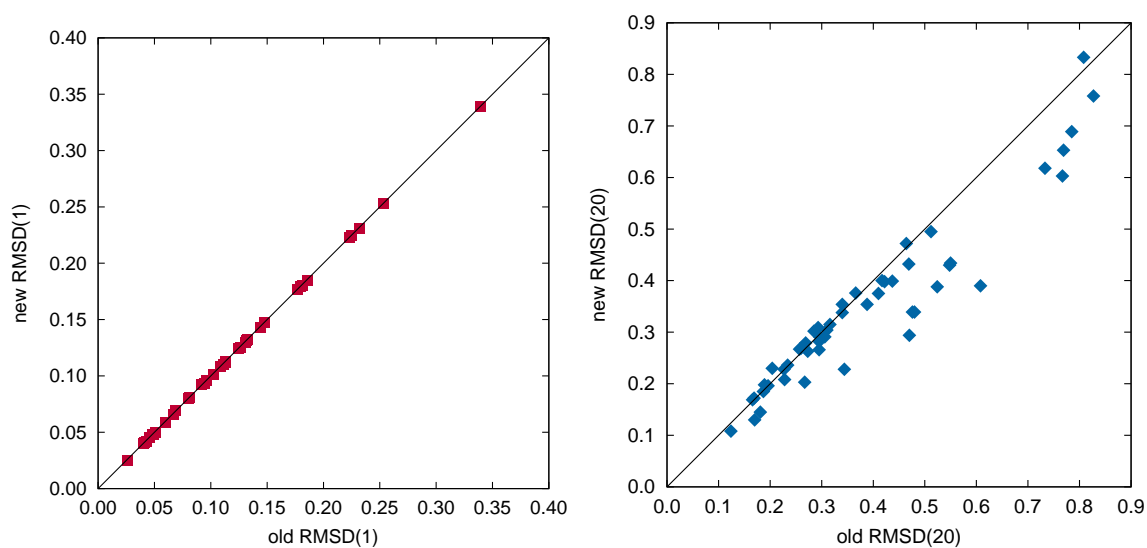


Figure A.1: Comparison of RMSD values reported in BT6 with those obtained in this work using the current version of Mercury. Results are shown for the unique structure matches, with the exceptions of XXIIID-G06-L1-E73 and XXIIID-G09-L1-E66, as their RMSD values were not reported in BT6.

A.5 Dependence on COMPACK options

Table A.4: COMPACK results for structures submitted for compound XXIII that had a 180° rotation of the carboxylic acid group relative to the target. Tolerances are for both distances (%) and angles (°).

Structure	RMSD(1)	VC-RMSD(1)	Raw-RMSD(20)	Tolerance	VC-RMSD(20)	Tolerance
Ignoring H-atom and bond counts						
XXIIIA-G09-L1-E19	0.185	0.202	0.551	20	0.295	20
XXIIIB-G06-L1-E26	0.162	0.155	0.442	20	0.183	20
XXIIIB-G09-L1-E46	0.187	0.174	0.434	20	0.188	20
XXIIIB-G14-L1-E89	0.089	0.080	0.192	20	0.090	20
XXIIID-G06-L1-E73	0.220	0.261	0.747	20	0.321	20
XXIIID-G06-L1-E75	0.220	0.261	0.747	20	0.321	20
XXIIID-G09-L1-E66	0.239	0.211	0.603	20	0.269	20
Including H-atom and bond counts						
XXIIIA-G09-L1-E19	0.641	0.640	0.823	30	0.673	45
XXIIIB-G06-L1-E26	0.633	0.622	0.754	65	0.629	65
XXIIIB-G09-L1-E46	0.637	0.622	0.747	40	0.625	40
XXIIIB-G14-L1-E89	0.625	0.624	0.648	50	0.625	50
XXIIID-G06-L1-E73	0.651	0.651	0.967	65	0.679	60
XXIIID-G06-L1-E75	0.651	0.651	0.967	65	0.679	60
XXIIID-G09-L1-E66	0.658	0.639	0.860	40	0.661	40

A.6 Example output tables

Table A.5: Example `vc-pwdf` output of structures that pass the unit-cell dimension criteria, when given a 10% deviation allowance from the reference structure.

structure	raw-POWDIFF	a	b	c	alpha	beta	gamma	volume	cryst_syst	spgrp
xx01_n.cif	0.1020773	6.804	11.999	12.542	106.60	90	90	981.2685	monoclinic	P2_1/c
xx47_n.cif	0.2202693	6.841	12.384	12.653	108.49	90	90	1016.6664	monoclinic	P2_1/c
xx71_n.cif	0.3565806	6.909	12.387	13.206	116.23	90	90	1013.8525	monoclinic	P2_1/c
xx10_n.cif	0.4136432	7.197	11.567	12.142	90	95.31	90	1006.4558	monoclinic	P2_1/c
xx22_n.cif	0.4316918	6.369	12.347	13.411	96.86	90	90	1047.0647	monoclinic	P2_1/c
xx07_n.cif	0.4443281	7.202	10.952	12.716	97.51	90	90	994.3874	monoclinic	P2_1/c
xx11_n.cif	0.5544066	6.891	11.806	12.444	94.76	90	90	1008.8917	monoclinic	P2_1/c
xx12_n.cif	0.5583085	6.889	11.807	12.452	94.86	90	90	1009.1846	monoclinic	P2_1/c
xx89_n.cif	0.6110274	7.173	11.948	12.254	103.50	90	90	1021.1873	monoclinic	P2_1/c

Table A.6: Example `vc-pwdf` output of structures that have undergone volume correction, ranked by VC-PWDF comparison to the target structure.

structure	VC-PWDF
xx01_n_VC.res	0.0040872
xx47_n_VC.res	0.0196081
xx22_n_VC.res	0.2152266
xx12_n_VC.res	0.3274327
xx11_n_VC.res	0.3274816
xx89_n_VC.res	0.3918757
xx71_n_VC.res	0.4240894
xx07_n_VC.res	0.4399329

A.7 Effect of VC-PWDF tolerance

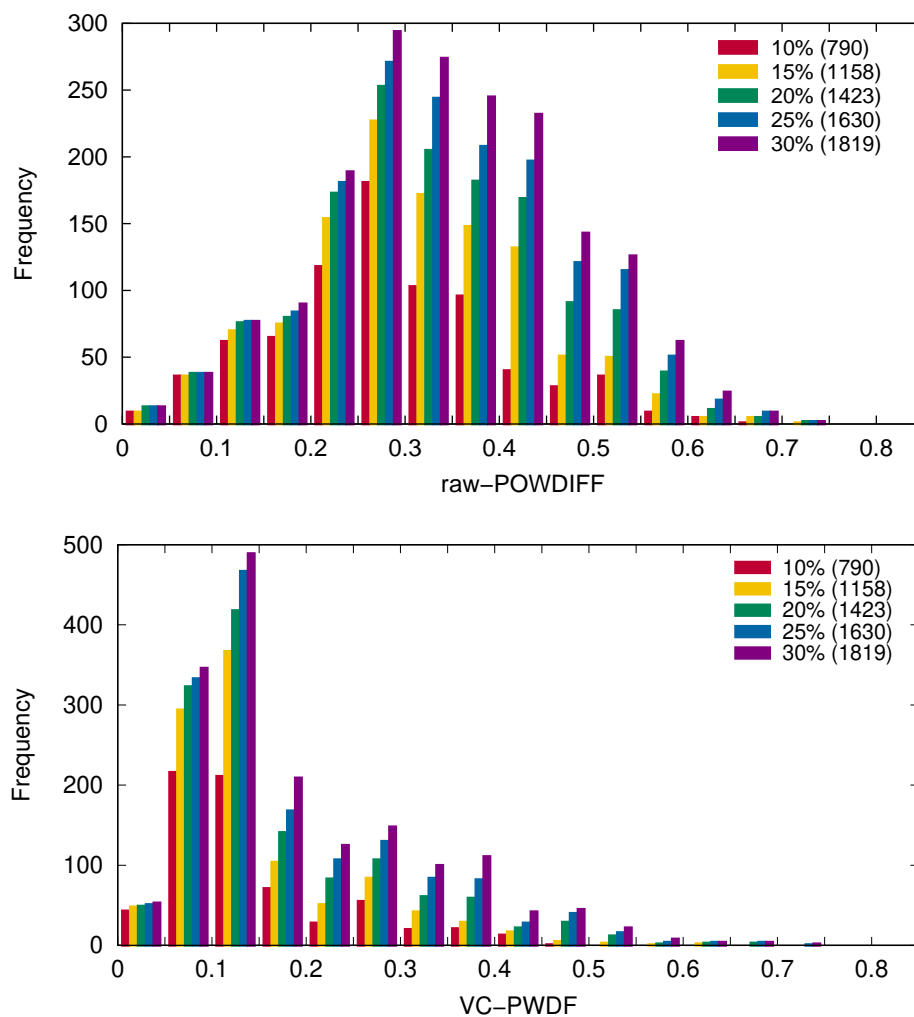


Figure A.2: Histograms of powder difference values for structures that pass step (3) of our computational algorithm, with different volume and cell-length tolerances selected. Results are shown for the sets of structures before (top) and after (bottom) anisotropic volume correction.

A.8 Effect of RMSD tolerance

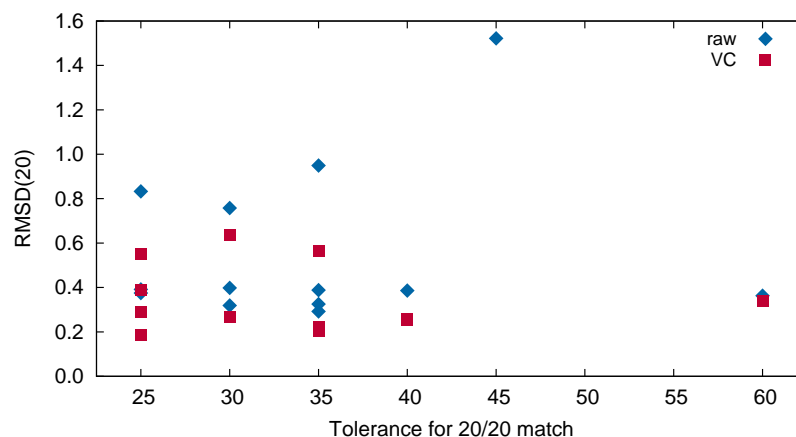


Figure A.3: RMSD(20) values (before and after volume correction) as a function of the tolerance required to obtain a 20/20 molecule match with COMPACT. Results are only shown for cases in which the tolerances had to be increased beyond their default values (20% and 20°) to obtain a match. Increments of 5% and 5° were used when raising the tolerances.

A.9 Effect of volume correction on RMSD(1)

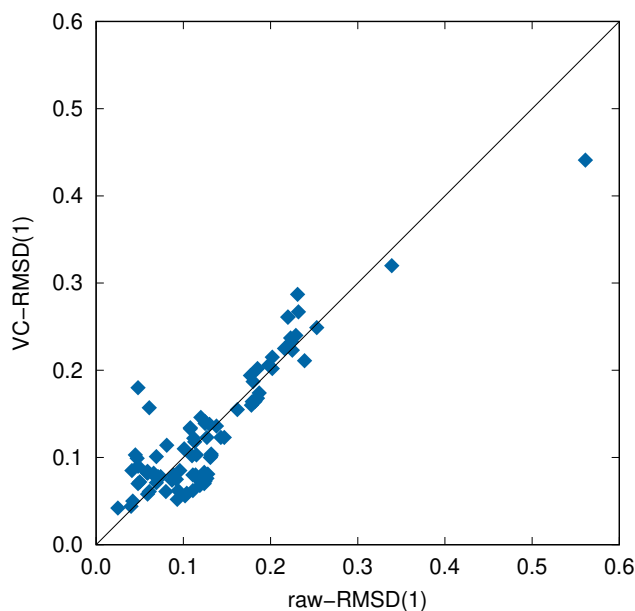


Figure A.4: RMSD(1) values for molecules before and after anisotropic volume correction. The $y = x$ line is shown to highlight the roughly even numbers of structures where the RMSD(1) increases/decreases after volume correction.

A.10 Correlations between RMSD(20) and powder difference values

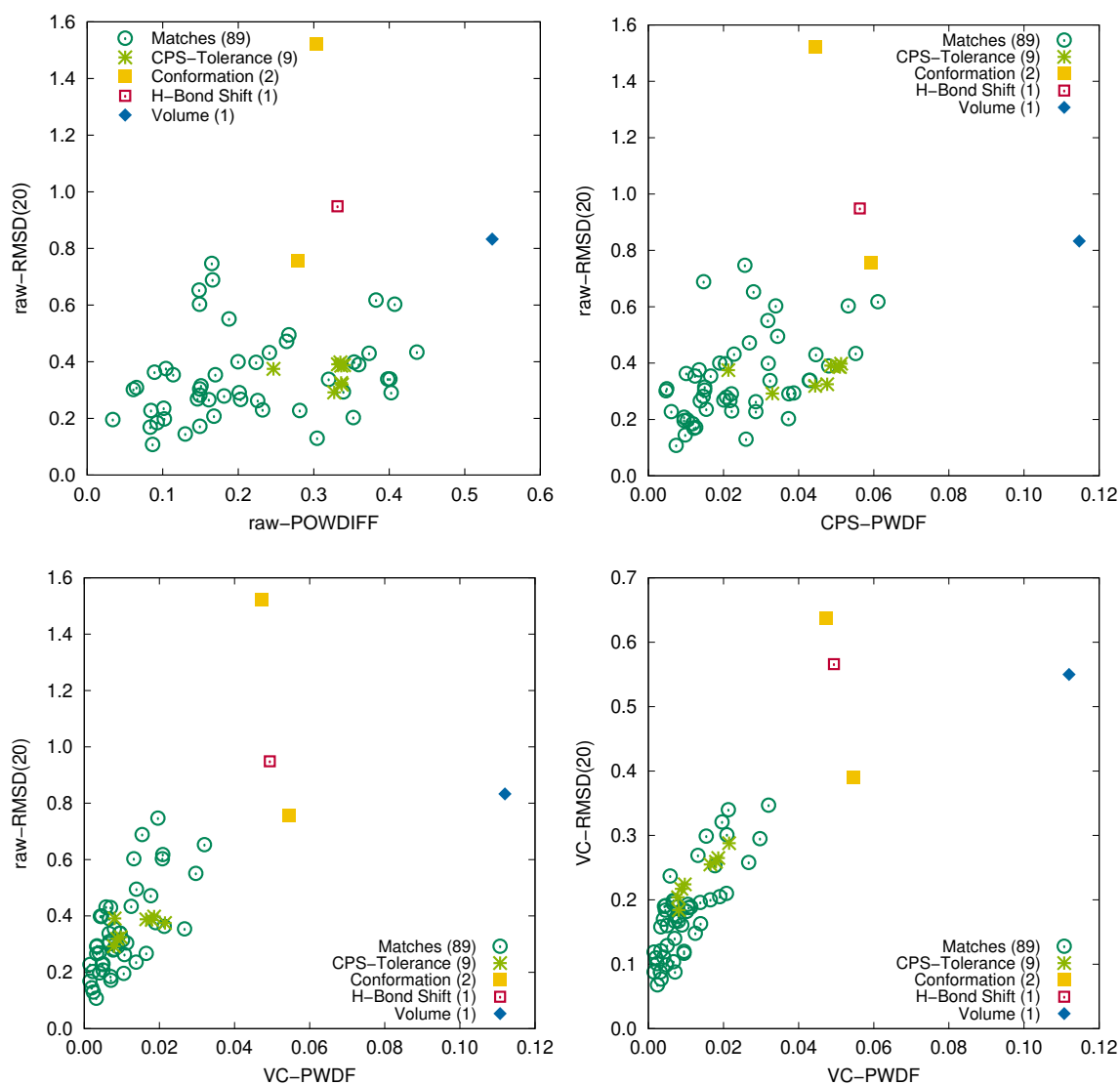


Figure A.5: Correlations between various RMSD(20) and powder difference values for the 113-structure dataset.

APPENDIX B

SUPPLEMENTARY INFORMATION FOR: DEVELOPMENT AND ASSESSMENT OF AN IMPROVED POWDER-DIFFRACTION-BASED METHOD FOR MOLECULAR CRYSTAL STRUCTURE SIMILARITY

B.1 Dataset

The authors thank Dr. Aurora Cruz-Cabeza for providing them with the list of CSD refcodes and accompanying classifying, pressure and temperature data and CPS results from their study in 2020.⁶⁸ This list of refcodes was used such that a comparison of the methods used in that study, and the one presented here, could be made.

B.1.1 Edits to the dataset

As the CSD had been updated since the list was generated by Sacchi *et al.*,⁶⁸ some refcode changes had been made, listed in Table B.1. The dataset was edited to update the old refcode to the new one. Another 12 comparisons were removed as they appeared twice in the list of 47,422 (Table B.2).

Table B.1: Refcode changes made since the study done by Sacchi *et al.*

Old	New
DLHIST02	LHISTD15
DUDZIL05	DUDZIL08
DUDZIL06	DUDZIL07
FACZIT01	BAGBEU
GIKVAX01	MUQGAJ
LEUCIN06	LISLEU04
XUTPUY01	OYOBAl
YONWII01	JOYWUS

Table B.2: List of 12 duplicate comparisons in the previous dataset.

CEGCAS01-CEGCAS10	CUKCAM01-CUKCAM02
TARTAL-TARTAL01	TARTAL-TARTAL02
TARTAL-TARTAL03	TARTAL-TARTAL04
TARTAL01-TARTAL02	TARTAL01-TARTAL03
TARTAL01-TARTAL04	TARTAL02-TARTAL03
TARTAL02-TARTAL04	TARTAL03-TARTAL04

B.1.2 Pruning of the dataset

As noted in the manuscript, various sets of problematic structures/comparisons were eliminated from the original data provided to yield our final set of 44,939 individual comparisons. The eliminated comparisons or structures are listed in the following sections.

B.1.2.1 Different compounds

Table B.3: List of 30 comparisons eliminated due to the crystal structures involving different molecular species.

GEJVIB-GEJVIB01	GOKREE-GOKREE01	HEWHAU-HEWHAU01
KETTUY01-KETTUY10	LEUCIN01-LISLEU04	LEUCIN02-LISLEU04
LEUCIN03-LISLEU04	LEUCIN04-LISLEU04	LEUCIN05-LISLEU04
LEUCIN-LISLEU04	LIWJEI-LIWJEI03	QEWXIA-QEWXIA01
SCCHRN01-SCCHRN05	SCCHRN02-SCCHRN05	SCCHRN03-SCCHRN05
SCCHRN04-SCCHRN05	SCCHRN05-SCCHRN06	SCCHRN05-SCCHRN07
SCCHRN-SCCHRN05	VALINO-VALINO01	VOBTOV01-VOBTOV03
VOBTOV02-VOBTOV03	VOBTOV-VOBTOV03	YARGON-YARGON01
YEJLII01-YEJLII03	YEJLII02-YEJLII03	YEJLII03-YEJLII04
YEJLII03-YEJLII05	YEJLII03-YEJLII06	YEJLII-YEJLII03

B.1.2.2 Disordered structures

Table B.4: List of the 116 structures identified as disordered by ConQuest. All comparisons involving these refcodes were removed from the dataset.

ACTOLD02	ALUCAL01	ALUCAL02	AMBNZA	ANTMET03	AZOBEN03
BEMLOU23	BENZAC01	BENZAC02	BINMEQ	BISJAO	BOPSAA01
BOPSAA06	BOPSAA07	BOPSAA08	BOPSAA09	BZOYAC01	CERLOA11
CHOLAU02	CITNIN01	CLBZNT02	CLBZNT03	CLBZNT04	COTSAF01
CUKDER02	DAKXUI01	DALGON03	DBZFUL02	DEBXIT	DLABUT02
DLABUT04	DLABUT05	DLABUT10	DLNLUA01	DLNLUA02	DPHETH01
DPHETH04	DTENYL01	DUCKOB03	DUVZOJ02	ECUMIY	ETBBAR
FACRIK	FACRIK01	FACRIK02	FACRIK05	FACRIK06	FESNEW
FILGEM01	FIMNAQ	FPAMCA13	FRANAC02	FURSEM02	GUACET02
GUMMUW02	HADKIG	HEXWIQ	ICAPOR04	IKIJER	IMAZOL02
IODOFO05	JIBCIG07	KECYBU14	KECYBU15	KOWYEA01	LEPMEZ
LEPMEZ01	LEQVUY	MERYOL03	MNIAAN01	MNIAAN10	MPOPHA
NAHNIT04	NBZOAC05	NBZOAC06	PCBZAM10	PEPHUN	PEPHUN01
PEPHUN03	PHTHCY01	PINCOL	POVSAT	PUBMUU22	PYRDNO12
PYRDNO14	PYRZIN20	PYRZOL01	QQQAZG31	ROLGEE	SLFNMF13
TACETA03	TAURIN03	TBUCBD02	TBUCBD10	TCLBEN06	TCYETY03
TCYETY04	TEPHTH07	TERPHE09	TETRDO	TETROL	TMPPIO11
TSTILB04	WAMRAD	WEMGIE	XEHHEX02	YARZUN	YIGPIO
ZEPDAZ	ZEPDAZ01	ZZZBCS10	ZZZIYE02	ZZZIYE04	ZZZQNK03
ZZZQNK07	ZZZVCO06				

B.1.2.3 Alert level A voids

Alert level A flags from the Platon¹²³ checkcif tool are often an indication that there is an error in the cif. The unique refocdes making up the entire dataset were run through checkcif and 78 structures (Table B.5) were identified with Alert level A voids. An example of a structure with Alert level A voids is illustrated in Figure B.1 for the case of LUYKOS01. Here, the atom labels and coordinates for one of the two molecules in the asymmetric unit were absent from the structure file. This structure should be the same as LUYKOF, also shown in Figure B.1; however, the missing molecules naturally result in very different peak intensities in the powder patterns. CPS-COMPACK is able to generate a 20/20 match with this structure pair.

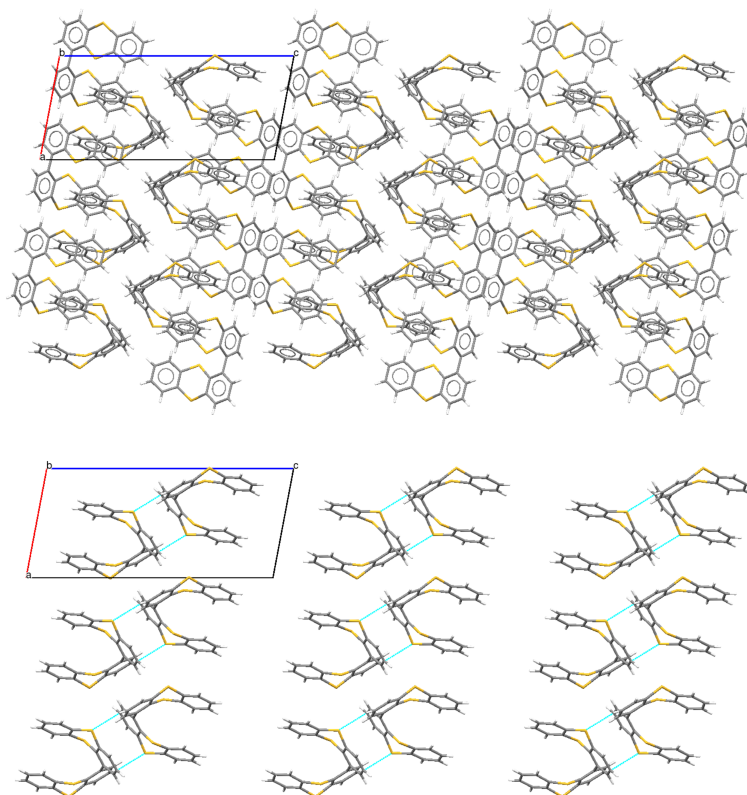


Figure B.1: LUYKOF with $Z' = 2$ (top) and LUYKOF01 with only one of the two molecules of the asymmetric unit present (bottom).

Table B.5: List of 78 structures identified by checkcif as having Alert level A flagged voids. All comparisons involving these refcodes were removed from the dataset.

AFLATM01	ALKINA01	ALKINA	AQARUF01	AQARUF
BEMLOU12	BEMLOU23	BOXGAW	CIDFEC	CIDTUE01
CIDTUE	DEBXIT01	DEBXIT02	DEBXIT03	DEBXIT04
DEBXIT05	DEBXIT06	DEBXIT	DHDTIZ10	DPHETH01
DPHETH04	FEHKOS01	FPAMCA13	GIGLEN10	GIGLEN
GLUTAM	HECPAJ01	HEXWIQ	JEYDEW02	JYKEH02
KOJWUB	LALQEV	LASNAU10	LASNAU	LIGXUU01
LUYKOF01	MAJRIZ02	MIHQIE01	NIRDUO01	OZECAY01
OZECAY02	OZECAY	PHENOL01	PINCOL01	PUDXES01
PUDXES02	PUDXES03	PUDXES04	PUDXES	PUXSIM02
QAJTUQ01	QIHSEF01	QIHSEF	QQQCIG21	QQQESP01
QQQESP02	QQQESP03	RUTLUO	SATGAV01	SATGAV
TUHXAV01	TUHXAV	VAKPUU	VALINO01	VATBOH
WEDLEW01	WEPDEB03	WIRYEB02	WIRYEB	XOPGOA01
XOPGOA	XUDVOH04	XUDVOH05	XUDVOH	YARGON
ZEVDUZ10	ZEVDUZ	ZULHET10		

B.1.2.4 Missing atoms

Cases of missing atoms were identified by running CPS-COMPACT with a cluster size of 1 and the rest of the default settings for all comparisons of the dataset to identify those that yielded a result of “No matches found”. These structures were further screened by checking the number of non-hydrogen atoms listed in the cif, and comparing them to the other structures of that refcode family. If the number of atoms in the cif were not a multiple of the number of atoms counted in the other structures of that refcode family, the structure was analyzed manually. This yielded 8 structures that had missing atoms — their refcodes are listed in Table B.6.

Table B.6: List of 8 structures identified as having missing non-H atoms. All comparisons involving these refcodes were removed from the dataset.

FOWVUI01	LALNIN14
LALNIN15	LASNAU10
PEZMEM10	SOVFEO
YAPNEK02	ZULHET10

As an example, viewing the PEZMEM10 structure in Mercury shows that 5 non-hydrogen atoms are missing from each molecule in the unit cell (Figure B.2). When comparing PEZMEM10 and PEZMEM11 (or PEZMEM) with CPS-COMPACT using the parameters described in the Methods (hydrogen-atom counts and bond counts are ignored) a 20/20 match is obtained. It is not clear how this match occurs or why these options would allow the omission of non-hydrogen atoms.

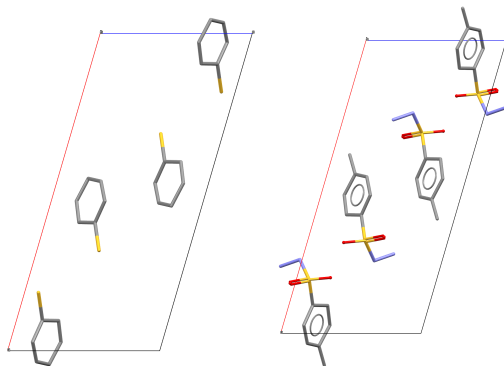


Figure B.2: Comparison between the unit cells of PEZMEM10 (left) and PEZMEM (right), showing the missing molecular components in PEZMEM10.

B.1.2.5 Problem cases for Ullmann's algorithm

Table B.7: List of the 146 refcode families that were removed from the data set due to excessively long run times for CPS-COMPACT comparison. Any comparisons involving a structure in one of these refcode families were removed from the dataset.

BADGAO	BECMUT	BUQWOZ	BUYJAH	CACYEK	CTVHVH	DATQIY
DAZPUS	DPANTR	EBIGUR	EPANEQ	ESIVOR	FADDOD	FAHNOR
FIJBWU	FOGWII	FURHUV	GACHEY	GAYTIJ	GEYGOG	GISNUS
HELXUR	HNIABZ	HPHBNZ	HUMFIG	IFAWAP	IGUQEG	IHAPEO
INELUK	INOCET	IREPIG	IRUQOB	ISIKAW	IVATAC	IVATUW
IVOHIL	JIBCIG	KANYUU	KELFOX	KOFJOF	KOKQUW	KUCYOY
KUVWON	LICMER	LIRRAG	LURHAJ	NACXUK	NAHZOM	NOETNA
NOEURA	OBARIV	OCHTET	OCMETD	OGELUI	ORIGIN	PABHAB
PBBTAZ	PEDTUP	PEKZAG	PEMZAJ	PEZBOL	PIDFEN	POPGUX
PUBMUU	PURSEB	QEDTIC	QEHLUL	QIHSEF	QQQCIG	QQQDVM
REPFOH	SEPBAS	SIGDAN	SOPRUK	TAFKET	TPHPOR	VAHTAB
VALTEJ	VAMBOA	VEZPIZ	WEFKIC	WORWAD	XIMMEL	XOBHIH
YARHEH	YIVRIF	YOSRED	YUHGOX	YUKGUG	ZAMWOC	ZOSLAU
AJEYAQ	ANONEX	BATWOI	BOWWOZ	CICYES	COYLAF	CTBROM
CTMTNA	CUMVIP	CUVSIV	DOPPAB	DUWBUT	EFIKOU	EGAXAL
FANTIX	FEFQUD	FIZRUD	FNETAM	FOBSOE	FPAMCA	GOBVEA
GOBYAX	HAXHET	HETPAL	IBPRAC	INODUK	LEPSEE	MBPHOL
MOBNIC	MPIMZR	OBUPUY	OGUTOZ	OHIBOW	OROGEJ	PATSEJ
POSPIY	QOCNIF	RUGSIW	SAPVAI	SERSOZ	SIFLOI	TELKUQ
TEPHME	TIWYIH	TOXGLU	UCUGOP	VALSUY	VENPOU	WAPBUK
WERROA	WUXDOJ	XINRIV	YUPJIE	ZAJBOE	ZZZVXQ	

B.2 Comparisons using CSD-housed vs. cif structures

The CSD Python application programming interface was used to access the CSD using the refcode and write a cif of that structure to a local working directory. This is required as input to VC-PWDF, which cannot directly access the CSD. Visualising results in Mercury was initially done haphazardly, using either the written cifs, or refcodes to directly access the structure information in the CSD. This highlighted some unexpected differences in results depending on which way the structures were accessed.

Some comparisons were performed twice with CPS-COMPACT using the default parameters with the Python interface, where one script used the local cifs, and the other

used the refcode to access the structure in the CSD. A number of these cases would yield a $N/20$ value with the script that used the CSD-housed structures, but yield a result of “No matches found” when using the local cifs. A few cases of the opposite behaviour were also observed.

It appears that, for structures housed in the CSD, the “auto-edit” procedure is run on the structures prior to comparison with the CPS-COMPACT utility. We hypothesize that this is done to deal with the numerous structures that are missing hydrogen atoms and/or specified bond descriptions (single, double, aromatic, etc.). This results in a considerable number of “No matches found” when the local cifs are used with the default parameters, as no addition of hydrogen atoms is done in cases where they are not provided in the structure information. While CPS-COMPACT does not consider the *position* of hydrogen atoms when using the default parameters, it does consider *how many* hydrogen atoms are bonded to a particular atom.

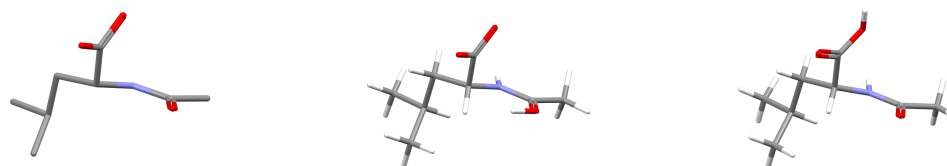


Figure B.3: Molecular structures for ACLLEU (left), ACLLEU after auto-edit with a hydrogen assigned to carbonyl oxygen (middle), and ACLLEU01 (right).

Additionally, the auto-edit procedure does not always yield the correct solution (see Figure B.3 for an example) and if default CPS-COMPACT parameters are used, one will occasionally obtain the result of “No matches found” in error. To account for cases of missing hydrogens, unspecified bond numbers/counts, and cases of incorrect auto-edit output, the default CPS-COMPACT parameters were changed to ignore hydrogen counts and ignore bond counts. This change allows for agreement between the output of comparisons made with local cifs and CSD-housed structures, but will by definition allow for comparisons between tautomers, and may cause other unexpected changes if the structure has errors such as missing atoms (Sec. B.1.2.4).

B.3 Example of the dependence of $\text{RMSD}(N)$ on tolerance

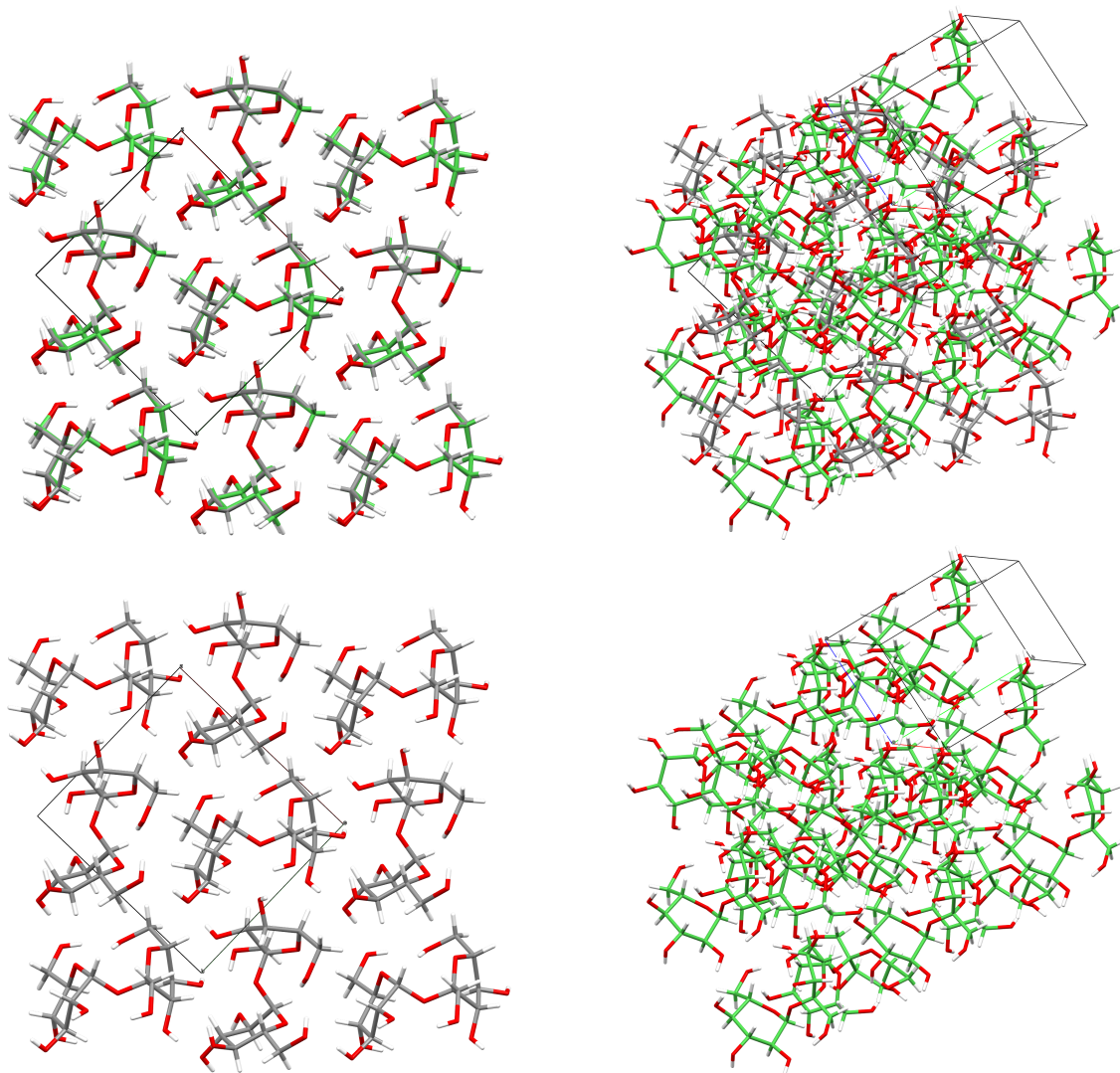


Figure B.4: COMPACK overlays of SUCROS27 and SUCROS33 using a tolerance of $\pm 20 \text{ }^\circ$ (top, left) and a tolerance of $\pm 10 \text{ }^\circ$ (top, right). In both cases, a match of 20/20 molecules is reported, but the $\text{RMSD}(20)$ values are 0.211 and 9.098 Å, respectively. The two structures that are overlaid for the $\pm 10 \text{ }^\circ$ tolerance comparison are shown independently without any change in orientation (bottom, left and bottom, right) to better show the rotation of the compared structure relative to the reference.

APPENDIX C

SUPPLEMENTARY INFORMATION FOR: QUANTITATIVE MATCHING OF CRYSTAL STRUCTURES TO EXPERIMENTAL POWDER DIFFRACTOGRAMS

C.1 Experimental

C.1.1 Powder X-ray diffraction

Acetaminophen (Sigma Aldrich), caffeine (Alfa Aesar), 1,4-dicyanobenzene (Sigma Aldrich), D-mannitol (VWR), (+)-progesterone (Sigma Aldrich), uracil (Sigma Aldrich), and urea (Sigma Aldrich) were used as received. If the solid powder appeared to have a notably large crystallite size or shape then crushing with a spatula and/or grinding with mortar and pestle was done prior to analysis. PXRD measurements were performed using a PANalytical Empyrean diffractometer in reflection (Bragg-Brentano) geometry with a Cu $K\alpha$ radiation source ($K\alpha_1 \lambda = 1.54184 \text{ \AA}$), Ni $K\beta$ filter, and PIXcel1D linear detector. A powdered sample was back-loaded into a sample holder with a 16mm insert, which was mounted on a spinning stage at room temperature.

Powder diffractograms were recorded in the 5 - 50 degrees 2θ range with either a step size of 0.00328 degrees and exposure time of 180 seconds per step (“moderate quality” 3 hour scan) or a step size of 0.01313 degrees and exposure time of 10 seconds per step (“lower quality” 2 minute scan). Data collection was controlled with the Data Collector software.¹⁷¹

C.2 Data

C.2.1 Target structures from the CSD

Target crystal structures were identified in the Cambridge Structural Database (CSD) by refcode. Where more than one entry for a given polymorph was available, the one collected at ambient conditions with the lowest R -factor (measure of data and model quality and agreement) was chosen as the primary target used in comparisons. Structures from low temperature or high pressure data collections were used for comparisons where indicated. As disordered structures are not currently compatible with the VC-PWDF method, any such structures were not included in comparisons or analyses (eg. polymorphs of NIWFEE).

C.2.2 CSP landscapes

Lists of *in silico* generated structures hosted in the CPOSS database were obtained from Dr. Louise Price. Some landscapes have been published (progesterone,¹⁷² uracil,¹⁷³ 1,4-dicyanobenzene,¹⁷⁴ caffeine,¹⁷⁵), while others have thus far remained unpublished

(mannitol, urea, acetaminophen). While some differences may exist and reference to the publications (where available) is recommended for these fine details, the general approach followed for the generation of the *in silico* structures obtained from the CPOSS database are:

1. Generation and geometry optimization of various molecular conformers using GAUSSIAN.¹⁷⁶
2. Calculation of atomic multipoles using distributed multipole analysis of the charge density calculated with GAUSSIAN for each conformer.¹⁷⁷
3. Hypothetical crystal structure generation, restricted to a chosen set of space groups, Z , Z' , and conformers. Possible programs used for this task include MOLPACK,¹⁷⁸ and CrystalPredictor.¹⁷⁹
4. Geometry minimization and energy evaluation of the generated crystal structure using the atomic multipoles for evaluation of the electrostatic interactions, and the Buckingham potential for dispersion interactions with Williams¹⁸⁰, FIT¹⁸¹ atom-atom parameters. Possible programs used for this task include DMAREL,¹⁸² or DMACRYS¹¹³ (rigid molecule during minimization), and CrystalOptimizer¹⁸³ (flexible molecule during minimization).

The structure-energy landscapes were screened for duplicates with an in-house script that utilizes relative energy, VC-PWDF, and COMPACK (tolerance of $\pm 30\%$ and $\pm 30^\circ$ on distances and angles, respectively) to identify structures with $\Delta E < 2$ kJ/mol, VC-PWDF < 0.07 , and 20/20 matches by COMPACK as duplicates.

C.3 Computational methods

C.3.1 VC-(x)PWDF

Our VC-PWDF method¹³⁶, which is implemented within the critic2 program⁷⁵, was modified in order to allow target unit cell dimensions (a , b , c , α , β , γ) and experimental powder diffractograms as a txt file (angle 2θ in degrees, and intensity) to be input for comparison against a simulated powder diffractogram generated by a crystal structure file (cif, res, etc...). We differentiate the results from the comparison of experimental

Table C.1: Summary of the result of duplicate screening on the structure-energy landscapes obtained from the CPOSS database. The number of candidates is the total number of structures received from the CPOSS database, and the unique number is the number of structures remaining after the duplicate screening.

Compound	CSD ID	Candidates	Unique
Urea	UREAXX	793	777
1,4-Dicyanobenzene	TEPNIT	144	94
Uracil	URACIL	217	211
Acetaminophen	HXACAN	640	618
Caffeine	NIWFEE	84	79
Mannitol	DMANTL	619	546
Progesterone	PROGST	149	149

and simulated diffractograms by including “x” before the PWDF portion of the method abbreviation, as in VC-xPWDF. Only diffractograms collected with Cu K α X-rays are compatible at this time. The VC-PWDF method uses the unit cell dimensions of the target structure (here, the indexed cell dimensions from the experimental powder diffractogram) and searches over all possible unit cell descriptions of the candidate structure for that which best matches the target diffractogram after replacing the unit cell parameters of the candidate unit cell with those of the target structure. The Figure of Merit (FoM) used is the dissimilarity value (POWDIFF) yielded by the triangle-weighted cross-correlation function described by de Gelder *et al.*²³

C.3.2 autoFIDEL

A Python script written by Jonas Nyman to perform a variation of the FIDEL (Fit with DEviating Lattice parameters) protocol⁷⁴ was used. The FIDEL method uses a hill climber’s (steepest descent) algorithm to minimize the difference between two powder diffractograms by variation of the unit cell parameters and atomic positions of the crystal structure that is used to generate the simulated powder diffractogram. The script used in this work only modifies the lattice parameters, no changes to atomic positions are affected during the optimization. The FoM output from autoFIDEL is the similarity value yielded by the triangle-weighted cross-correlation function described by de Gelder *et al.*²³ In order to ease the comparison between FoM values, we have converted the similarity value to the dissimilarity (POWDIFF) by subtraction from one, since this is our preferred metric. Unless otherwise noted, the default parameters of autoFIDEL were used.

C.4 Results

Table C.2: Summary of indexed unit cell dimensions from the collected powder diffractograms and comparison to the unit cell dimensions of the matching polymorph in the CSD after conversion to their Niggli reduced unit cell. No space group determination was done for the experimental powder data.

Compound	Space group	a (Å)	b (Å)	c (Å)	α (°)	β (°)	γ (°)	V (Å ³)
UREAXX23(RT)	$P42_1m$	4.7042	5.6577	5.6577	90	90	90	150.579
Urea(PXRD)	-	7.987	7.987	9.4042	90	90	90	599.913
TEPNIT04(RT)	$P1$	3.847	6.585	7.322	114.5	93.6	96.9	166.254
1,4-Dicyanobenzene(PXRD)	-	3.8514	6.5958	7.328	114.567	93.593	96.934	166.745
URACIL(RT)	$P2_1/b$	3.6552	10.3113	12.376	90	90	96.570	463.386
Uracil(PXRD)	-	3.6691	10.3146	12.3958	90	90	96.637	465.972
HXACAN35(RT)	$P2_1/n$	7.0661	9.3366	11.6508	90	97.410	90	762.223
Acetaminophen(PXRD)	-	7.1078	9.3986	11.7462	90	97.541	90	777.901
NIWFEE03(RT)	Cc	6.9531	15.0676	22.800	109.295	98.516	90	2226.607
Caffeine(PXRD)	-	6.9572	15.0839	22.8451	109.277	98.758	90	2233.362
DMANTL07(RT)	$P2_12_12_1$	5.549	8.694	16.902	90	90	90	815.403
D-Mannitol(PXRD)	-	5.5642	8.6848	16.9008	90	90	90	816.711
PROGST10(RT)	$P2_12_12_1$	10.340	12.559	13.798	90	90	90	1791.81
(+)-Progesterone(PXRD)	-	10.3741	12.6059	13.8464	90	90.268	90	1810.74

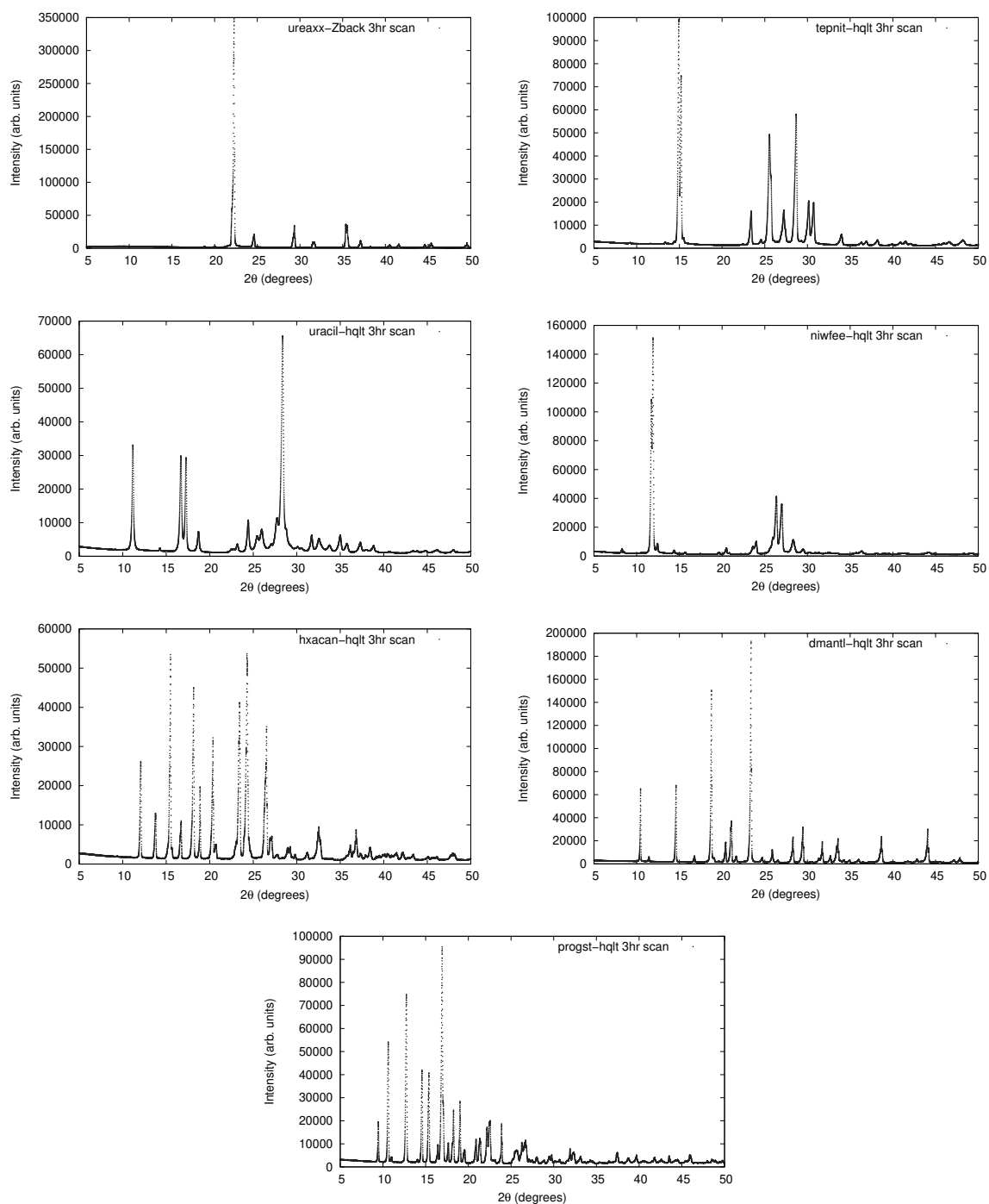


Figure C.1: Experimental powder diffractograms collecting with the 3hr scan conditions.

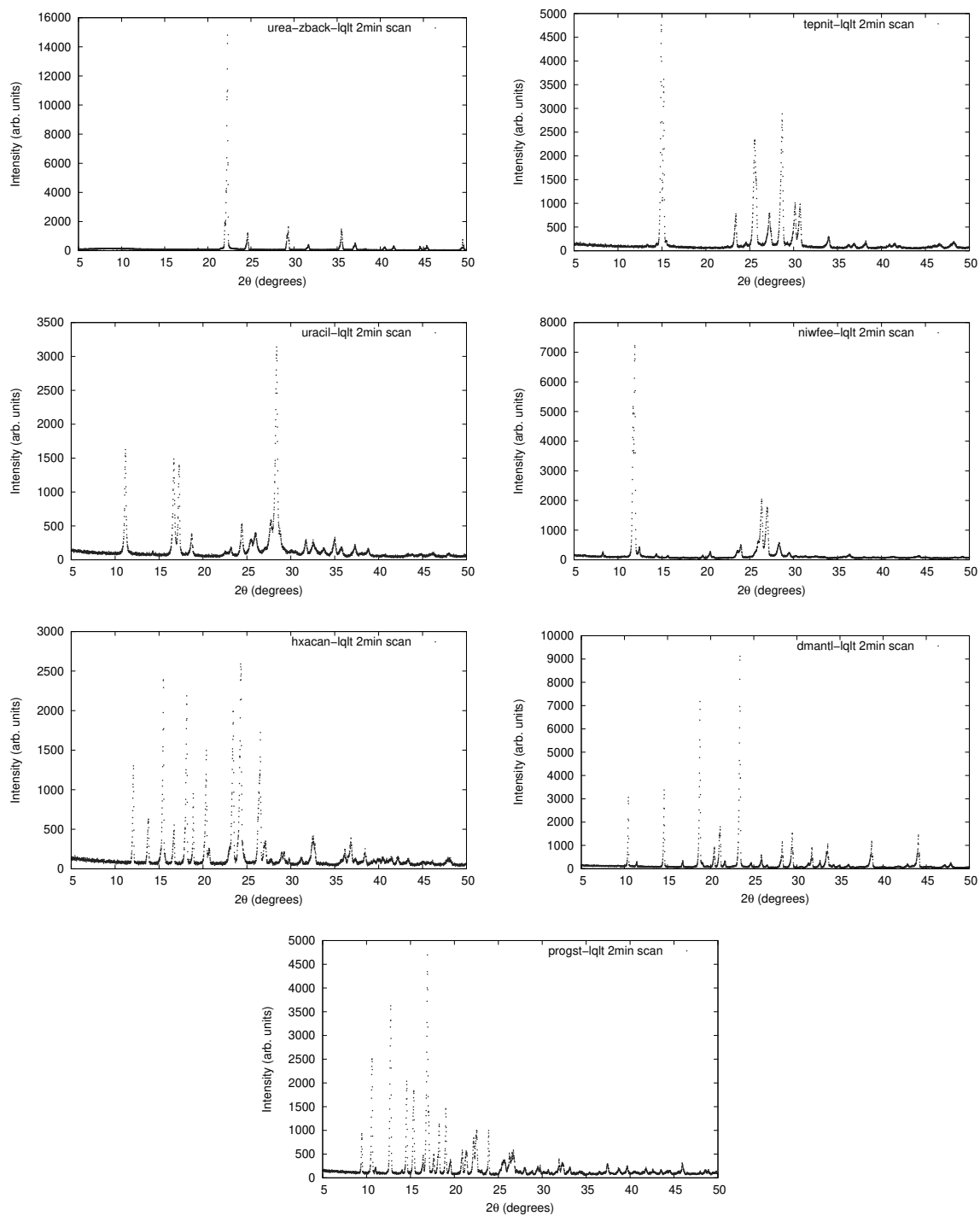


Figure C.2: Experimental powder diffractograms collecting with the 2min scan conditions.

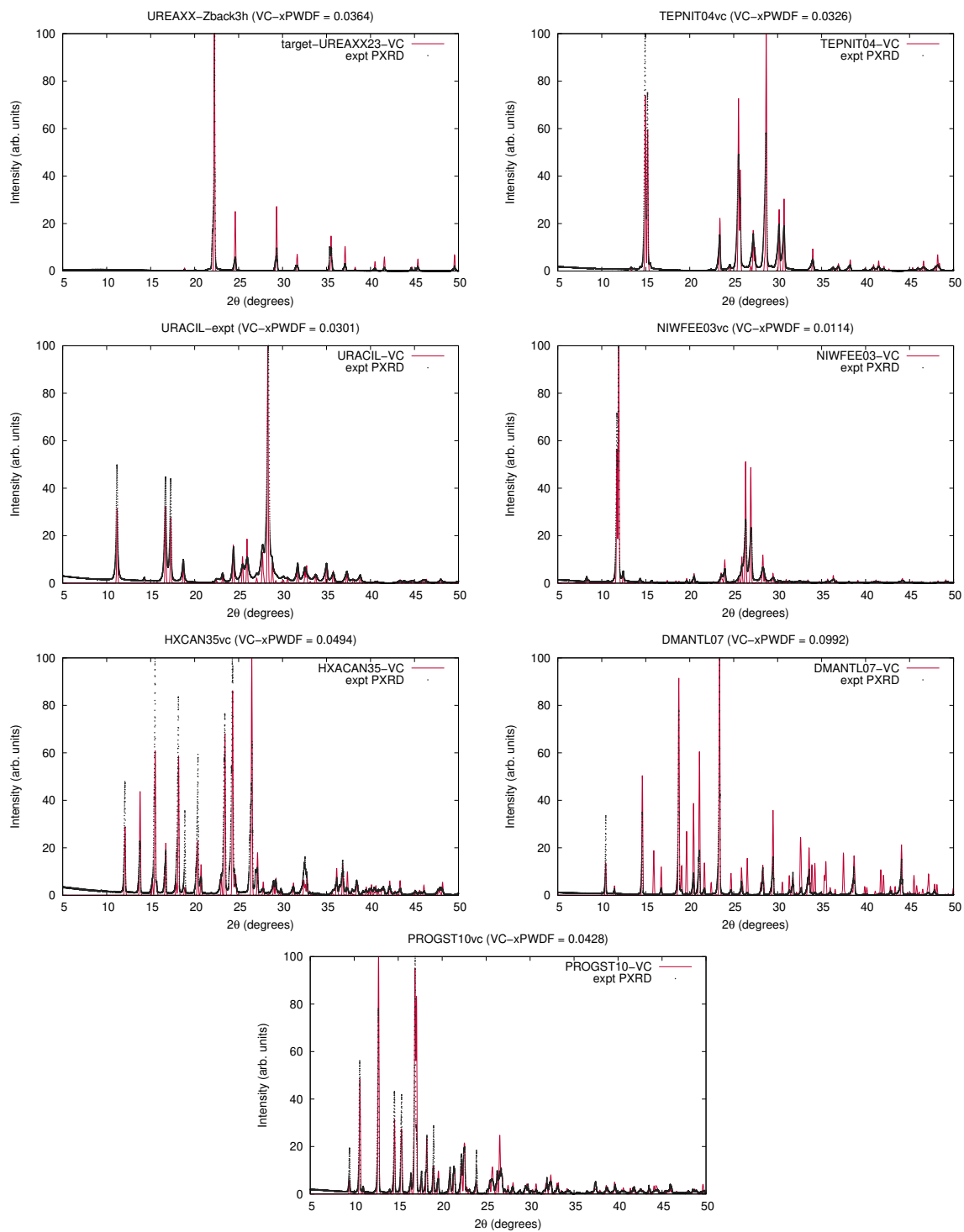


Figure C.3: Overlays of the experimentally collected powder diffractograms with the simulated powder diffractogram of the matching polymorph from the CSD after running the VC-xPWDF protocol.

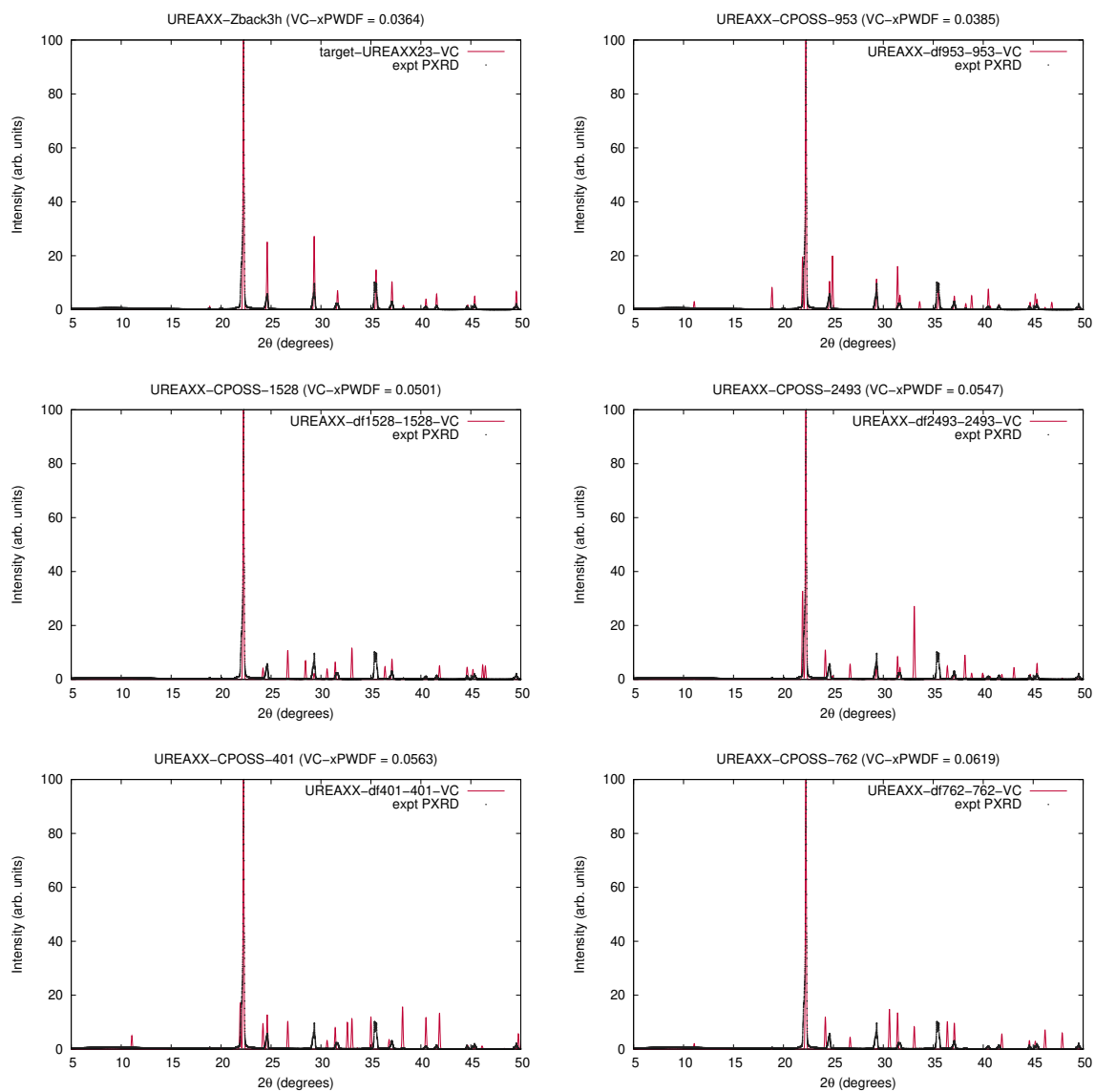


Figure C.4: Overlay of simulated powder diffractograms of urea crystal structures after the VC-xPWDF protocol with the experimental powder diffractogram.

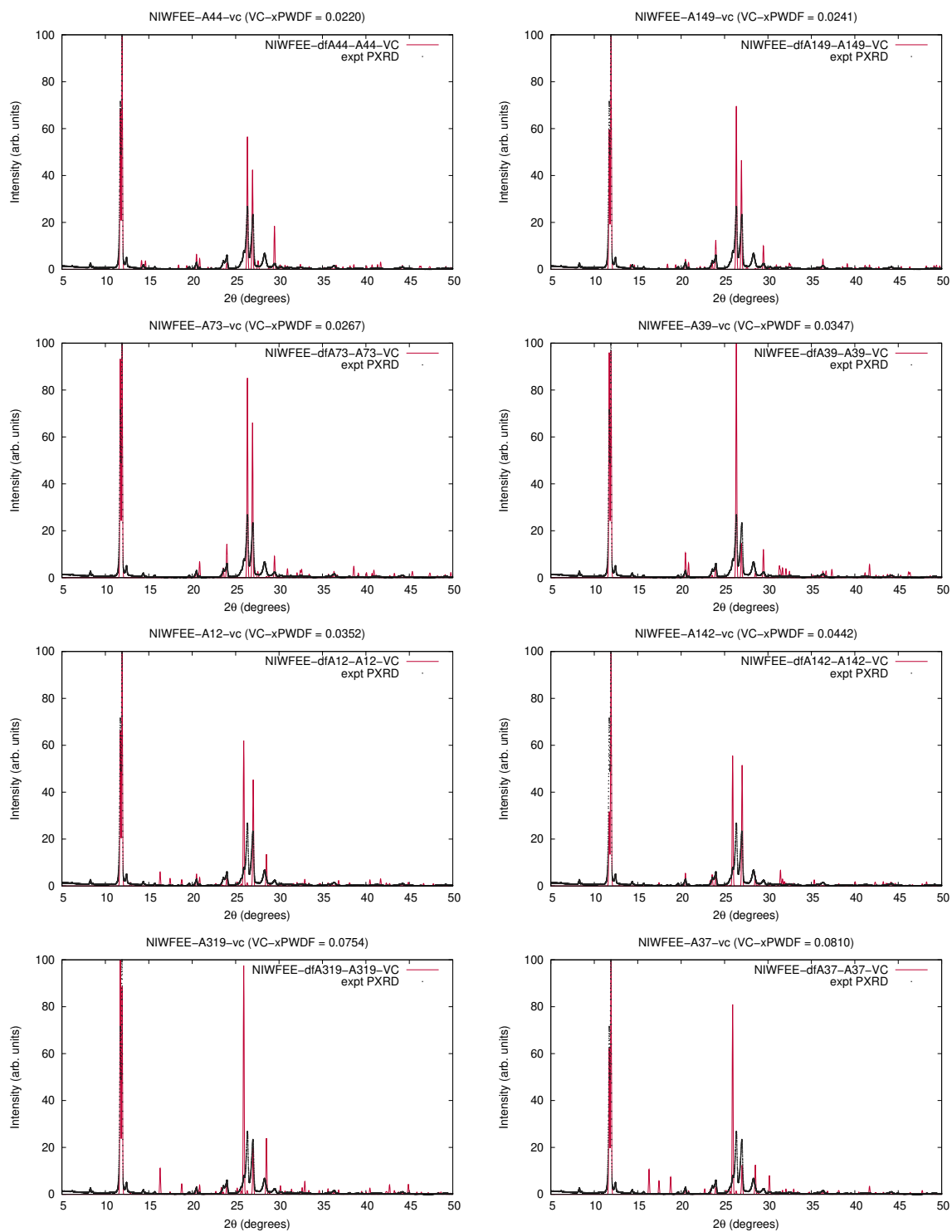


Figure C.5: Overlay of simulated powder diffractograms of caffeine crystal structures after the VC-xPWFDF protocol with the experimental powder diffractogram.

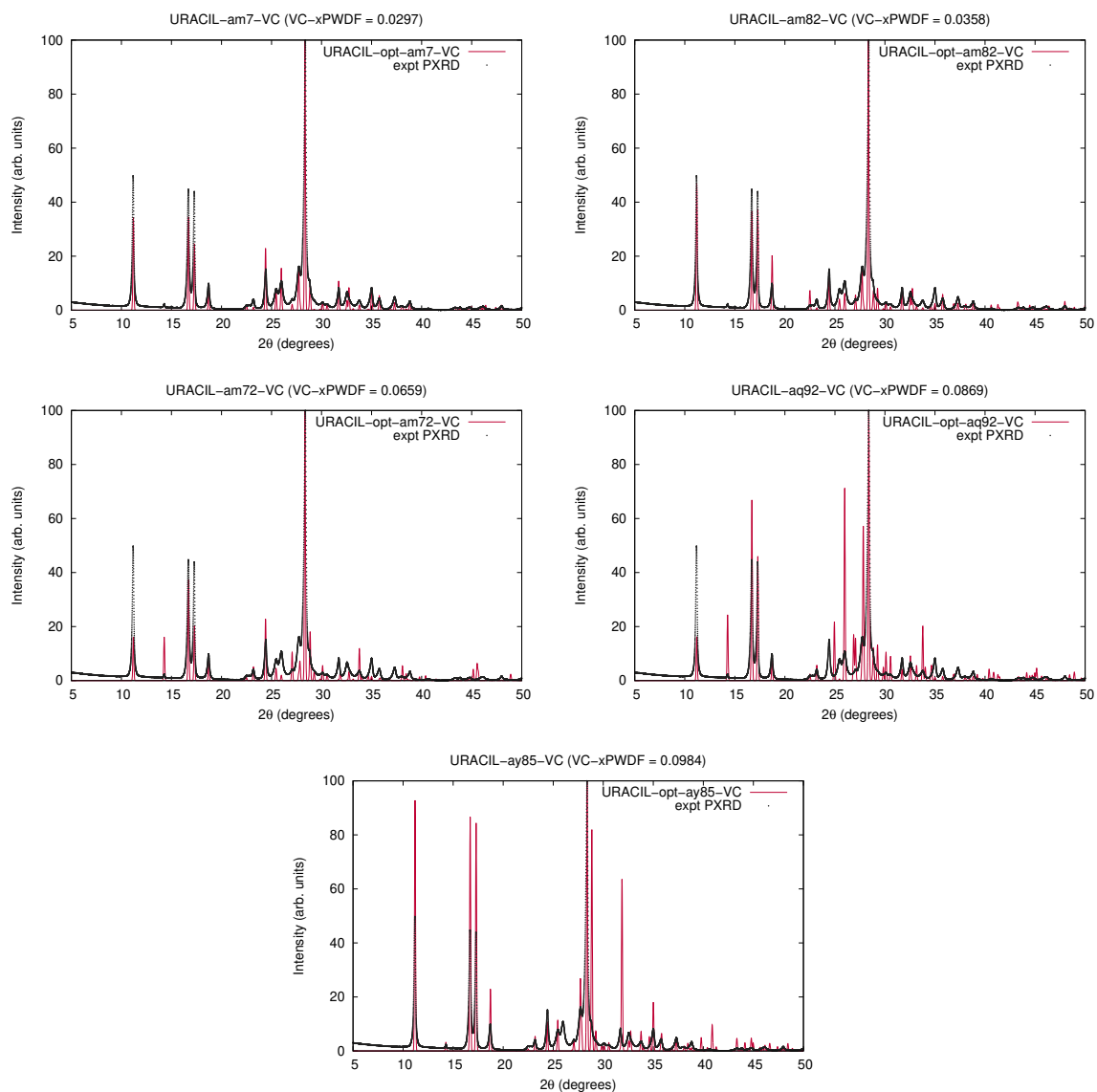


Figure C.6: Overlay of simulated powder diffractograms of uracil crystal structures after the VC-xPWDF protocol with the experimental powder diffractogram.

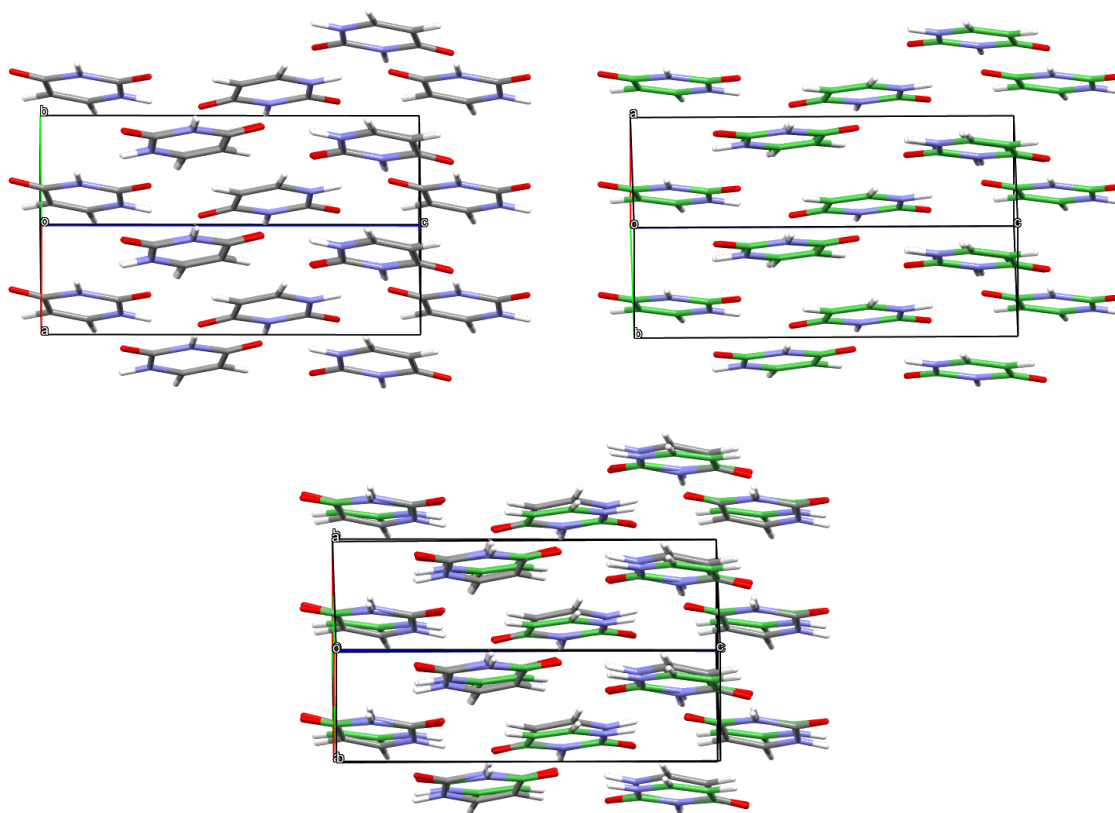


Figure C.7: Images of the packing of two *in silico* generated crystal structures of uracil, am7 (top-left), and am82 (top-right), and an overlay of the two structures generated with the Crystal Packing Similarity tool's implementation of COMPACK²² in the CCDC Mercury⁹⁸ software (V2022.2.0). The crystal structures used in these images had been modified with the VC-xPWDF protocol to match the experimental powder diffractogram.

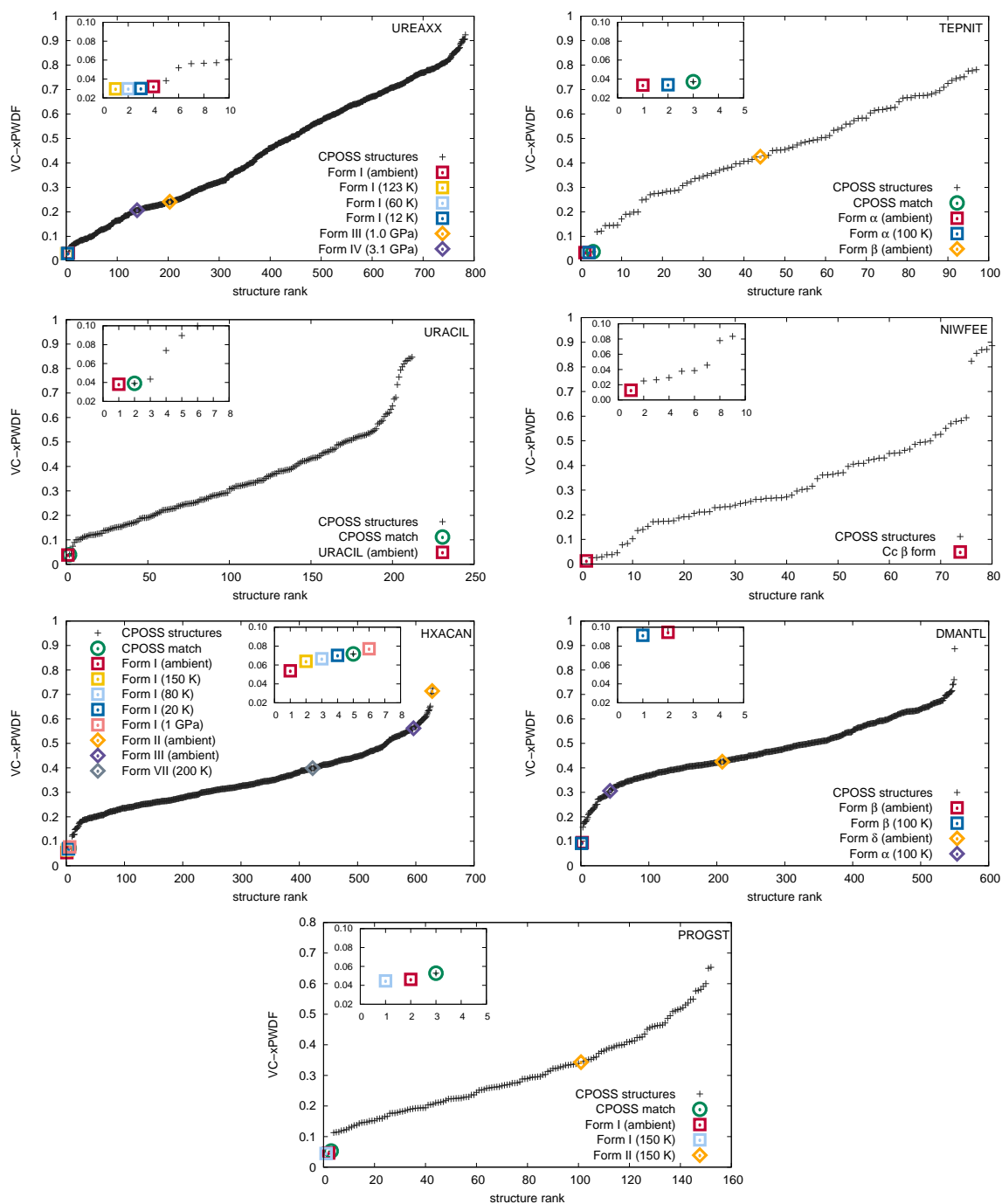


Figure C.8: Plots showing the VC-xPWDF value from comparison of each input crystal structure to the experimental powder diffractogram collected from a 2 minute scan for that compound. The structures are ranked by lowest VC-xPWDF (most similar) and the insets provide views of the best matching structures (VC-xPWDF < 0.1). The point types indicate the source of each crystal structure: squares correspond to CSD structures of the matching polymorph, diamonds are CSD structures of different polymorphs, and + signs are CPOSS structures. If a matching structure is present in the CPOSS data, it is identified with a green circle.

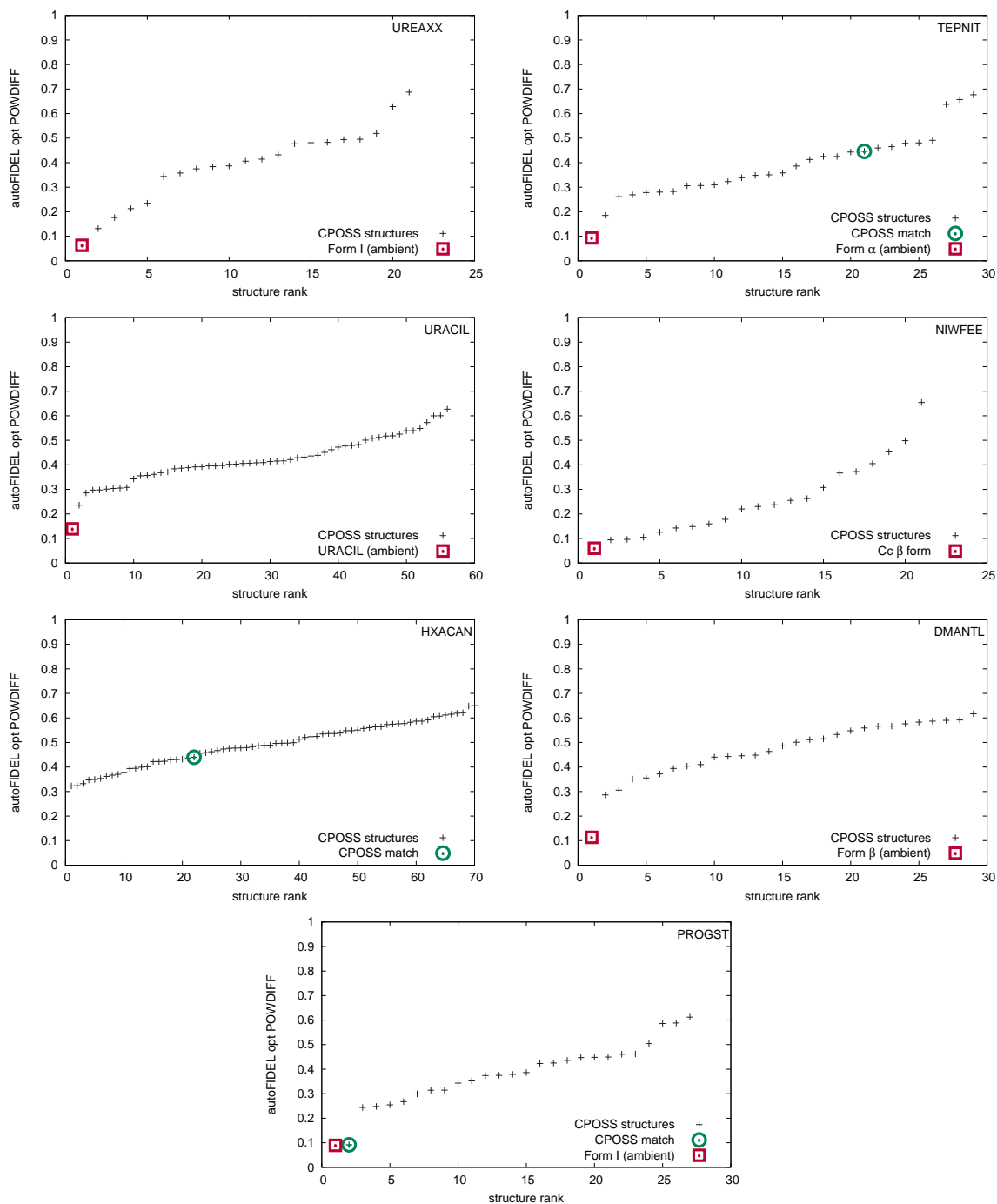


Figure C.9: Plots showing the POWDIFF value after optimization with the autoFIDEL code between the crystal structure and the experimental powder diffractogram collected for that compound. The structures are ranked by smallest POWDIFF (most similar). The point types indicate the source of each crystal structure: squares correspond to CSD structures of the matching polymorph, diamonds are CSD structures of different polymorphs, and + signs are CPOSS structures. If a matching structure is present in the CPOSS data, it is identified with a green circle.

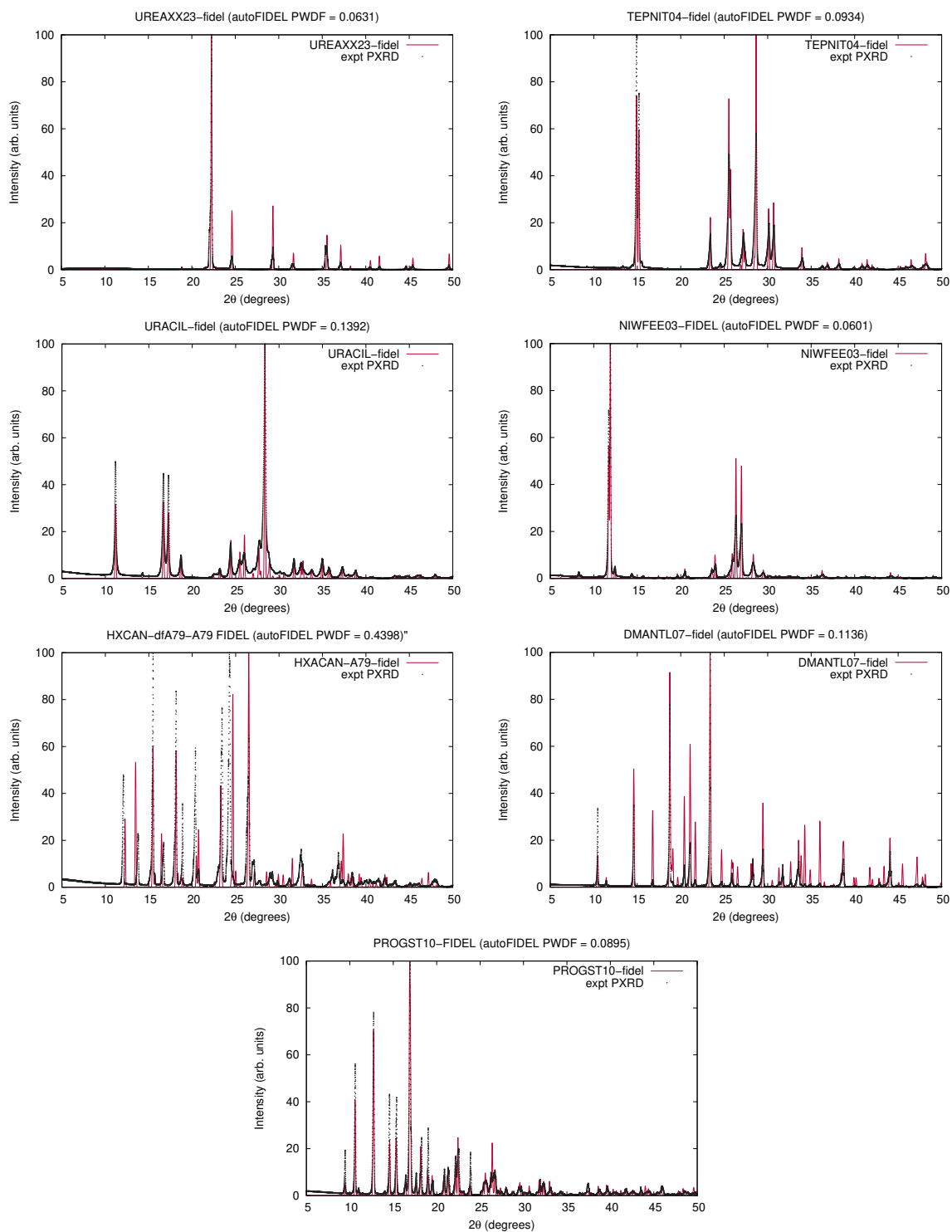


Figure C.10: Overlay of simulated powder diffractograms from crystal structures after optimization with autoFIDEL with the experimental powder diffractogram.

C.4.1 Rietveld refinement

C.4.1.1 CSD structures

Rietveld refinement was performed on the matching CSD structures without modification and the results shown in Table C.3. Clear correct alignment of the peaks was observed for all structures collected under ambient conditions, though the refinement values range from $20.10\% < R_{wp} < 34.54\%$ and $85.80 < \chi^2 < 307.41$.

Table C.3: Summary of R_{wp} and χ^2 values from Rietveld refinement of the same polymorph CSD structures with the collected powder diffractograms. A success indicates that the peak positions were properly aligned between the two patterns, even if the intensities did not perfectly overlay. A poor refinement is indicated for cases where peak positions remained unaligned at the completion of the refinement.

CSD refcode	conditions	VC-xPWDF	R_{wp} (%)	χ^2	refinement
UREAXX23	RT	0.0364	30.06	178.49	success
UREAXX07	123 K	0.0335	41.69	343.32	poor
UREAXX11	60 K	0.0337	38.83	297.83	poor
UREAXX12	12 K	0.0339	39.95	315.26	poor
TEPNIT04	RT	0.0326	22.14	112.22	success
TEPNIT14	100 K	0.0330	26.65	162.59	poor
URACIL	RT	0.0290	20.10	85.80	success
NIWFEE03	RT	0.0114	26.08	154.23	success
HXACAN35	RT	0.0494	24.04	109.25	success
HXACAN04	150 K	0.0602	31.22	184.25	success
HXACAN15	80 K	0.0633	42.25	337.44	poor
HXACAN13	20 K	0.0670	32.73	202.51	poor
NIWFEE03	RT	0.0114	26.08	154.23	success
DMANTL07	RT	0.0992	33.90	197.86	success
DMANTL15	100 K	0.0962	36.61	256.66	success
PROGST10	RT	0.0428	34.54	307.41	success
PROGST12	150 K	0.0416	35.43	323.45	success

C.4.1.2 CPOSS and VC-corrected structures

Table C.4: Summary of R_{wp} and χ^2 values from Rietveld refinement of uracil structures with the collected powder diffractograms. A success indicates that the peak positions were properly aligned between the two patterns, even if the intensities did not perfectly overlay. A poor refinement is indicated for cases where peak positions remained unaligned at the completion of the refinement.

structure	rank	VC-xPWDF	ΔH (kJ/mol)	R_{wp} (%)	χ^2	refinement
URACIL-vc	1	0.0290	-	15.40	50.36	success
URACIL_opt.am7-vc	2	0.0297	0.0	17.43	64.52	success
URACIL_opt.am82-vc	3	0.0358	0.5	19.73	82.67	success
URACIL_opt.am72-vc	4	0.0659	2.0	38.62	316.74	poor
URACIL_opt.aq92-vc	5	0.0869	4.3	43.56	402.95	poor
URACIL_opt.ay85-vc	6	0.0984	7.6	21.78	100.74	poor

Table C.5: Summary of R_{wp} and χ^2 values from Rietveld refinement of caffeine structures with the collected powder diffractograms. A success indicates that the peak positions were properly aligned between the two patterns, even if the intensities did not perfectly overlay. A poor refinement is indicated for cases where peak positions remained unaligned at the completion of the refinement.

structure	rank	VC-xPWDF	ΔH (kJ/mol)	R_{wp} (%)	χ^2	refinement
NIWFEE03-vc	1	0.0114	-	21.28	102.68	success
NIWFEE_dfA44_A44-vc	2	0.0220	7.0	32.45	238.78	poor
NIWFEE_dfA149_A149-vc	3	0.0241	6.2	33.88	260.28	poor
NIWFEE_dfA73_A73-vc	4	0.0267	0.0	32.46	238.92	poor
NIWFEE_dfA39_A39-vc	5	0.0347	10.7	25.77	150.59	poor
NIWFEE_dfA12_A12-vc	6	0.0352	8.1	32.18	234.82	poor
NIWFEE_dfA142_A142-vc	7	0.0442	3.1	25.76	150.47	poor
NIWFEE_dfA319_A319-vc	8	0.0754	10.8	33.60	256.00	poor
NIWFEE_dfA37_A37-vc	9	0.0810	10.1	37.27	314.98	poor

Table C.6: Summary of R_{wp} and χ^2 values from Rietveld refinement of urea structures with the collected powder diffractograms. A success indicates that the peak positions were properly aligned between the two patterns, even if the intensities did not perfectly overlay. A poor refinement is indicated for cases where peak positions remained unaligned at the completion of the refinement.

structure	rank	VC-xPWDF	ΔH (kJ/mol)	R_{wp} (%)	χ^2	refinement
UREAXX23-vc	4	0.0364	-	20.26	81.08	success
UREAXX_df953_953	5	0.0385	14.9	40.07	317.16	poor
UREAXX_df1528_1528	6	0.0501	17.7	61.33	742.99	very poor
UREAXX_df2493_2493	7	0.0547	19.7	61.53	747.84	very poor
UREAXX_df401_401	8	0.0563	11.8	60.52	723.49	very poor
UREAXX_df762_762	9	0.0619	15.9	60.51	723.25	very poor

BIBLIOGRAPHY

- [1] Bendikov, M.; Wudl, F.; Perepichka, D. F. Tetrathiafulvalenes, oligoacenes, and their buckminsterfullerene derivatives: The brick and mortar of organic electronics. *Chem. Rev.* **2004**, *104*, 4891–4946.
- [2] Schweicher, G.; Garbay, G.; Jouclas, R.; Vibert, F.; Devaux, F.; Geerts, Y. H. Molecular semiconductors for logic operations: dead-end or bright future? *Adv. Mater.* **2020**, *32*, 1905909.
- [3] Fujita, W.; Awaga, K. Room-temperature magnetic bistability in organic radical crystals. *Science* **1999**, *286*, 261–262.
- [4] Ganin, A. Y.; Takabayashi, Y.; Jeglič, P.; Arčon, D.; Potočnik, A.; Baker, P. J.; Ohishi, Y.; McDonald, M. T.; Tzirakis, M. D.; McLennan, A.; Darling, G. R.; Takata, M.; Rosseinsky, M. J.; Prassides, K. Polymorphism control of superconductivity and magnetism in Cs₃C₆₀ close to the Mott transition. *Nature* **2010**, *466*, 221–225.
- [5] Sharif, P.; Alemdar, E.; Ozturk, S.; Caylan, O.; Hacıfendioglu, T.; Buke, G.; Aydemir, M.; Danos, A.; Monkman, A. P.; Yildirim, E.; Gunbas, G.; Cirpan, A.; Oral, A. Rational Molecular Design Enables Efficient Blue TADF- OLEDs with Flexible Graphene Substrate. *Adv. Funct. Mater.* **2022**, 2207324.
- [6] Xu, C.; Zhao, Z.; Yang, K.; Niu, L.; Ma, X.; Zhou, Z.; Zhang, X.; Zhang, F. Recent progress in all-small-molecule organic photovoltaics. *J. Mater. Chem. A* **2022**, *10*, 6291–6329.
- [7] Singhal, D.; Curatolo, W. Drug polymorphism and dosage form design: a practical perspective. *Adv. Drug Deliv. Rev.* **2004**, *56*, 335–347.
- [8] Lee, E. H. A practical guide to pharmaceutical polymorph screening & selection. *Asian J. Pharm. Sci.* **2014**, *9*, 163–175.
- [9] Li, M.; Balawi, A. H.; Leenaers, P. J.; Ning, L.; Heintges, G. H.; Marszalek, T.; Pisula, W.; Wienk, M. M.; Meskers, S. C.; Yi, Y.; Laquai, F.; Janssen, R. A. J. Impact of polymorphism on the optoelectronic properties of a low-bandgap semiconducting polymer. *Nat. Commun.* **2019**, *10*, 1–11.
- [10] Troisi, A.; Orlandi, G. Band structure of the four pentacene polymorphs and effect on the hole mobility at low temperature. *J. Phys. Chem. B* **2005**, *109*, 1849–1856.
- [11] Courte, M.; Ye, J.; Jiang, H.; Ganguly, R.; Tang, S.; Kloc, C.; Fichou, D. Tuning the π - π overlap and charge transport in single crystals of an organic semiconductor via solvation and polymorphism. *Phys. Chem. Chem. Phys.* **2020**, *22*, 19855–19863.

- [12] Evers, J.; Klapötke, T. M.; Mayer, P.; Oehlinger, G.; Welch, J. α - and β -FOX-7, Polymorphs of a High Energy Density Material, Studied by X-ray Single Crystal and Powder Investigations in the Temperature Range from 200 to 423 K. *Inorg. Chem.* **2006**, *45*, 4996–5007.
- [13] Dreger, Z. A.; Gupta, Y. M. Phase diagram of hexahydro-1, 3, 5-trinitro-1, 3, 5-triazine crystals at high pressures and temperatures. *J. Phys. Chem. A* **2010**, *114*, 8099–8105.
- [14] Herbst, W.; Hunger, K. *Industrial organic pigments: production, properties, applications*; John Wiley & Sons, 2006.
- [15] Wang, J.-R.; Zhu, B.; Yu, Q.; Mei, X. Selective crystallization of vitamin D3 for the preparation of novel conformational polymorphs with distinctive chemical stability. *CrystEngComm* **2016**, *18*, 1101–1104.
- [16] Yang, J.; Hu, C. T.; Zhu, X.; Zhu, Q.; Ward, M. D.; Kahr, B. DDT polymorphism and the lethality of crystal forms. *Angew. Chem. Int. Ed.* **2017**, *56*, 10165–10169.
- [17] Yang, J.; Erriah, B.; Hu, C. T.; Reiter, E.; Zhu, X.; López-Mejías, V.; Carmona-Sepúlveda, I. P.; Ward, M. D.; Kahr, B. A deltamethrin crystal polymorph for more effective malaria control. *Proc. Natl. Acad. Sci.* **2020**, *117*, 26633–26638.
- [18] Shankland, K. *International tables for crystallography, Volume H: Powder diffraction*; International Union for Crystallography, 2019; Chapter 4.1, pp 386–394.
- [19] Altomare, A.; Cuocci, C.; Moliterni, A.; Rizzi, R. *International tables for crystallography, Volume H: Powder diffraction*; International Union for Crystallography, 2019; Chapter 4.2, pp 395–413.
- [20] David, W. I. F. *International tables for crystallography, Volume H: Powder diffraction*; International Union for Crystallography, 2019; Chapter 4.3, pp 414–432.
- [21] Florence, A. *International tables for crystallography, Volume H: Powder diffraction*; International Union for Crystallography, 2019; Chapter 4.4, pp 433–441.
- [22] Motherwell, S.; Chisholm, J. A. COMPACT: A program for identifying crystal structure similarity using distances. *J. Appl. Cryst.* **2005**, *38*, 228–231.
- [23] de Gelder, R.; Wehrens, R.; Hageman, J. A. A generalized expression for the similarity of spectra: Application to powder diffraction pattern classification. *J. Comput. Chem.* **2001**, *22*, 273–289.
- [24] Van Eijck, B. P.; Kroon, J. Fast clustering of equivalent structures in crystal structure prediction. *J. Comput. Chem.* **1997**, *18*, 1036–1042.
- [25] Dzyabchenko, A. Method of crystal-structure similarity searching. *Acta Crystallogr.* **1994**, *B50*, 414–425.

- [26] Karfunkel, H.; Rohde, B.; Leusen, F.; Gdanitz, R. J.; Rihs, G. Continuous similarity measure between nonoverlapping X-ray powder diagrams of different crystal modifications. *J. Comput. Chem.* **1993**, *14*, 1125–1135.
- [27] Valle, M.; Oganov, A. R. Crystal fingerprint space—a novel paradigm for studying crystal-structure sets. *Acta Crystallogr.* **2010**, *A66*, 507–517.
- [28] Mosca, M. M.; Kurlin, V. Voronoi-Based Similarity Distances between Arbitrary Crystal Lattices. *Cryst. Res. Technol.* **2020**, *55*, 1900197.
- [29] Chemburkar, S. R.; Bauer, J.; Deming, K.; Spiwek, H.; Patel, K.; Morris, J.; Henry, R.; Spanton, S.; Dziki, W.; Porter, W.; Quick, J.; Bauer, P.; Donaubaauer, J.; Narayanan, B. A.; Soldani, M.; Riley, D.; McFarland, K. Dealing with the impact of Ritonavir polymorphs on the late stages of bulk drug process development. *Org. Proc. Res. Dev.* **2000**, *4*, 413–417.
- [30] Mortazavi, M.; Hoja, J.; Aerts, L.; Quéré, L.; van de Streek, J.; Neumann, M. A.; Tkatchenko, A. Computational polymorph screening reveals late-appearing and poorly-soluble form of rotigotine. *Commun. Chem.* **2019**, *2*, 70.
- [31] Pulido, A.; Chen, L.; Kaczorowski, T.; Holden, D.; Little, M. A.; Chong, S. Y.; Slater, B. J.; McMahon, D. P.; Bonillo, B.; Stackhouse, C. J.; Stephenson, A.; Kane, C. M.; Clowes, R.; Hasell, T.; Cooper, A. I.; Day, G. M. Functional materials discovery using energy–structure–function maps. *Nature* **2017**, *543*, 657–664.
- [32] Ishii, H.; Obata, S.; Niitsu, N.; Watanabe, S.; Goto, H.; Hirose, K.; Kobayashi, N.; Okamoto, T.; Takeya, J. Charge mobility calculation of organic semiconductors without use of experimental single-crystal data. *Sci. Rep.* **2020**, *10*, 1–10.
- [33] Oganov, A. R.; Pickard, C. J.; Zhu, Q.; Needs, R. J. Structure prediction drives materials discovery. *Nat. Rev. Mater.* **2019**, *4*, 331–348.
- [34] LeBlanc, L. M.; Johnson, E. R. Crystal-energy landscapes of active pharmaceutical ingredients using composite approaches. *CrystEngComm* **2019**, *21*, 5995–6009.
- [35] Price, A. J.; Mayo, R. A.; Otero-de-la Roza, A.; Johnson, E. R. Accurate and efficient polymorph energy ranking with XDM-corrected hybrid DFT. *CrystEngComm* **2023**, *25*, 953–960.
- [36] Lommerse, J. P. M.; Motherwell, W. D. S.; Ammon, H. L.; Dunitz, J. D.; Gavez-zotti, A.; Hofmann, D. W. M.; Leusen, F. J. J.; Mooij, W. T. M.; Price, S. L.; Schweizer, B.; Schmidt, M. U.; van Eijck, B. P.; Verwer, P.; Williams, D. E. A test of crystal structure prediction of small organic molecules. *Acta Crystallogr.* **2000**, *B56*, 697–714.

- [37] Motherwell, W. D. S.; Ammon, H. L.; Dunitz, J. D.; Dzyabchenko, A.; Erk, P.; Gavezzotti, A.; Hofmann, D. W. M.; Leusen, F. J. J.; Lommerse, J. P. M.; Mooij, W. T. M.; Price, S. L.; Scheraga, H.; Schweizer, B.; Schmidt, M. U.; van Eijck, B. P.; Verwer, P.; Williams, D. E. Crystal structure prediction of small organic molecules: A second blind test. *Acta Crystallogr.* **2002**, *B58*, 647–661.
- [38] Day, G. M.; Motherwell, W. D. S.; Ammon, H. L.; Boerrigter, S. X. M.; Della Valle, R. G.; Venuti, E.; Dzyabchenko, A.; Dunitz, J. D.; Schweizer, B.; van Eijck, B. P.; Erk, P.; Facelli, J. C.; Bazterra, V. E.; Ferraro, M. B.; Hofmann, D. W. M.; Leusen, F. J. J.; Liang, C.; Pantelides, C. C.; Karamertzanis, P. G.; Price, S. L.; Lewis, T. C.; Nowell, H.; Torrisi, A.; Scheraga, H. A.; Arnautova, Y. A.; Schmidt, M. U.; Verwer, P. A third blind test of crystal structure prediction. *Acta Crystallogr.* **2005**, *B61*, 511–527.
- [39] Day, G. M.; Cooper, T. G.; Cruz-Cabeza, A. J.; Hejczyk, K. E.; Ammon, H. L.; Boerrigter, S. X. M.; Tan, J. S.; Della Valle, R. G.; Venuti, E.; Jose, J.; Gadre, S. R.; Desiraju, G. R.; Thakur, T. S.; van Eijck, B. P.; Facelli, J. C.; Bazterra, V. E.; Ferraro, M. B.; Hofmann, D. W. M.; Neumann, M. A.; Leusen, F. J. J.; Kendrick, J.; Price, S. L.; Misquitta, A. J.; Karamertzanis, P. G.; Welch, G. W. A.; Scheraga, H. A.; Arnautova, Y. A.; Schmidt, M. U.; van de Streek, J.; Wolf, A. K.; Schweizer, B. Significant progress in predicting the crystal structures of small organic molecules – a report on the fourth blind test. *Acta Crystallogr.* **2009**, *B65*, 107–125.
- [40] Bardwell, D. A.; Adjiman, C. S.; Arnautova, Y. A.; Bartashevich, E.; Boerrigter, S. X. M.; Braun, D. E.; Cruz-Cabeza, A. J.; Day, G. M.; Della Valle, R. G.; Desiraju, G. R.; van Eijck, B. P.; Facelli, J. C.; Ferraro, M. B.; Grillo, D.; Habgood, M.; Hofmann, D. W. M.; Hofmann, F.; Jose, K. V. J.; Karamertzanis, P. G.; Kazantsev, A. V.; Kendrick, J.; Kuleshova, L. N.; Leusen, F. J. J.; Maleev, A. V.; Misquitta, A. J.; Mohamed, S.; Needs, R. J.; Neumann, M. A.; Nikylov, D.; Orendt, A. M.; Pal, R.; Pantelides, C. C.; Pickard, C. J.; Price, L. S.; Price, S. L.; Scheraga, H. A.; van de Streek, J.; Thakur, T. S.; Tiwari, S.; Venuti, E.; Zhitkov, I. K. Towards crystal structure prediction of complex organic compounds – a report on the fifth blind test. *Acta Crystallogr.* **2011**, *B67*, 535–551.

- [41] Reilly, A. M.; Cooper, R. I.; Adjiman, C. S.; Bhattacharya, S.; Boese, A. D.; Brandenburg, J. G.; Bygrave, P. J.; Bylisma, R.; Campbell, J. E.; Car, R.; Case, D. H.; Chadha, R.; Cole, J. C.; Cosburn, K.; Cuppen, H. M.; Curtis, F.; Day, G. M.; DiStasio Jr, R. A.; Dzyabchenko, A.; van Eijck, B. P.; Elking, D. M.; van den Ende, J. A.; Facelli, J. C.; Ferraro, M. B.; Fusti-Molnar, L.; Gatsiou, C.-A.; Gee, T. S.; de Gelder, R.; Ghiringhelli, L. M.; Goto, H.; Grimme, S.; Guo, R.; Hofmann, D. W. M.; Hoja, J.; Hylton, R. K.; Iuzzolino, L.; Jankiewicz, W.; de Jong, D. T.; Kendrick, J.; de Klerk, N. J. J.; Ko, H.-Y.; Kuleshova, L. N.; Li, X.; Lohani, S.; Leusen, F. J. J.; Lund, A. M.; Lv, J.; Ma, Y.; Marom, N.; Masunov, A. E.; McCabe, P.; McMahon, D. P.; Meekes, H.; Metz, M. P.; Misquitta, A. J.; Mohamed, S.; Monserrat, B.; Needs, R. J.; Neumann, M. A.; Nyman, J.; Obata, S.; Oberhofer, H.; Oganov, A. R.; Orendt, A. M.; Pagola, G. I.; Pantelides, C. C.; Pickard, C. J.; Podeszwa, R.; Price, L. S.; Price, S. L.; Pulido, A.; Read, M. G.; Reuter, K.; Schneider, E.; Schober, C.; Shields, G. P.; Singh, P.; Sugden, I. J.; Szalewicz, K.; Taylor, C. R.; Tkatchenko, A.; Tuckerman, M. E.; Vacarro, F.; Vasileiadis, M.; Vazquez-Mayagoitia, A.; Vogt, L.; Wang, Y.; Watson, R. E.; de Wijs, G. A.; Yang, J.; Zhu, Q.; Groom, C. R. Report on the sixth blind test of organic crystal structure prediction methods. *Acta Crystallogr.* **2016**, *B72*, 439–459.
- [42] Mayo, R. A.; Sullivan, D. J.; Fillion, T. A. P.; Kycia, S. W.; Soldatov, D. V.; Preuss, K. E. Reversible crystal-to-crystal chiral resolution: making/breaking non-bonding S···O interactions. *Chem. Commun.* **2017**, *53*, 3964–3966.
- [43] Blum, V.; Gehrke, R.; Hanke, F.; Havu, P.; Havu, V.; Ren, X.; Reuter, K.; Scheffler, M. *Ab initio* molecular simulations with numeric atom-centered orbitals. *Comput. Phys. Comm.* **2009**, *180*, 2175–2196.
- [44] Aroyo, M. I. *International tables for crystallography, Volume A: Space-group symmetry*; International Union of Crystallography, 2016.
- [45] Müller, U.; Wondratschek, H. *International tables for crystallography, Volume A1: Symmetry relations between space groups*; International Union of Crystallography, 2011.
- [46] Shmueli, U. *International tables for crystallography, Volume B: Reciprocal space*; International Union of Crystallography, 2010.
- [47] Prince, E. *International tables for crystallography, Volume C: Mathematical, physical and chemical tables*; International Union of Crystallography, 2006.
- [48] Authier, A. *International tables for crystallography, Volume D: Physical properties of crystals*; International Union of Crystallography, 2013.
- [49] Kopský, V.; Litvin, D. *International tables for crystallography, Volume E: Subperiodic groups*; International Union of Crystallography, 2010.

- [50] Gilmore, C. J.; Kaduk, J. A.; Schenk, H. *International tables for crystallography, Volume H: Powder diffraction*; International Union of Crystallography, 2019.
- [51] West, A. *Basic solid state chemistry*; Wiley, 1999.
- [52] Pecharsky, V.; Zavalij, P. *Fundamentals of powder diffraction and structural characterization of materials, Second edition*; Springer US, 2008.
- [53] David, W.; Shankland, K.; Baerlocher, C.; McCusker, L.; Baerlocher, L. *Structure determination from powder diffraction data*; IUCr monographs on crystallography; Oxford University Press, 2002.
- [54] Dunitz, J. D.; Bernstein, J. Disappearing polymorphs. *Acc. Chem. Res.* **1995**, *28*, 193–200.
- [55] Bučar, D.-K.; Lancaster, R. W.; Bernstein, J. Disappearing polymorphs revisited. *Angew. Chem. Int. Ed.* **2015**, *54*, 6972–6993.
- [56] Barnes, P. W.; Lufaso, M. W.; Woodward, P. M. Structure determination of $A_2M^{3+}TaO_6$ and $A_2M^{3+}NbO_6$ ordered perovskites: Octahedral tilting and pseudosymmetry. *Acta Crystallogr.* **2006**, *B62*, 384–396.
- [57] Rietveld, H. M. A profile refinement method for nuclear and magnetic structures. *J. Appl. Cryst.* **1969**, *2*, 65–71.
- [58] Toby, B. H. R factors in Rietveld analysis: How good is good enough? *Powder Diffraction* **2006**, *21*, 67–70.
- [59] Visser, J. A fully automatic program for finding the unit cell from powder data. *J. Appl. Cryst.* **1969**, *2*, 89–95.
- [60] Werner, P.-E.; Eriksson, L.; Westdahl, M. TREOR, a semi-exhaustive trial-and-error powder indexing program for all symmetries. *J. Appl. Cryst.* **1985**, *18*, 367–370.
- [61] Boultif, A.; Louër, D. Indexing of powder diffraction patterns for low-symmetry lattices by the successive dichotomy method. *J. Appl. Cryst.* **1991**, *24*, 987–993.
- [62] de Wolff, P. M. A simplified criterion for the reliability of a powder pattern indexing. *J. Appl. Cryst.* **1968**, *1*, 108–113.
- [63] Smith, G. S.; Snyder, R. L. FN: A criterion for rating powder diffraction patterns and evaluating the reliability of powder-pattern indexing. *J. Appl. Cryst.* **1979**, *12*, 60–65.
- [64] Favre-Nicolin, V.; Černý, R. FOX, ‘free objects for crystallography’: A modular approach to *ab initio* structure determination from powder diffraction. *J. Appl. Cryst.* **2002**, *35*, 734–743.
- [65] Putz, H.; Schön, J. C.; Jansen, M. Combined method for *ab initio* structure solution from powder diffraction data. *J. Appl. Cryst.* **1999**, *32*, 864–870.

- [66] Lanning, O. J.; Habershon, S.; Harris, K. D.; Johnston, R. L.; Kariuki, B. M.; Tedesco, E.; Turner, G. W. Definition of a guiding function in global optimization: a hybrid approach combining energy and R-factor in structure solution from powder diffraction data. *Chem. Phys. Lett.* **2000**, *317*, 296–303.
- [67] Coelho, A. Whole-profile structure solution from powder diffraction data using simulated annealing. *J. Appl. Cryst.* **2000**, *33*, 899–908.
- [68] Sacchi, P.; Lusi, M.; Cruz-Cabeza, A. J.; Nauha, E.; Bernstein, J. Same or different – that is the question: Identification of crystal forms from crystal structure data. *CrystEngComm* **2020**, *22*, 7170–7185.
- [69] Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr.* **1976**, *A32*, 922–923.
- [70] Kabsch, W. A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr.* **1978**, *A34*, 827–828.
- [71] van de Streek, J.; Motherwell, S. Searching the Cambridge Structural Database for polymorphs. *Acta Crystallogr.* **2005**, *B61*, 504–510.
- [72] Hofmann, D.; Kuleshova, L. New Similarity Index for Crystal Structure Determination from X-Ray Powder Diagrams. *J. Appl. Cryst.* **2005**, *38*, 861–866.
- [73] van Der Lee, A.; Dumitrescu, D. G. Thermal expansion properties of organic crystals: a CSD study. *Chem. Sci.* **2021**, *12*, 8537–8547.
- [74] Habermehl, S.; Mörschel, P.; Eisenbrandt, P.; Hammer, S. M.; Schmidt, M. U. Structure determination from powder data without prior indexing, using a similarity measure based on cross-correlation functions. *Acta Crystallogr.* **2014**, *B70*, 347–359.
- [75] Otero-de-la-Roza, A.; Johnson, E. R.; Luaña, V. Critic2: A program for real-space analysis of quantum chemical interactions in solids. *Comput. Phys. Commun.* **2014**, *185*, 1007–1018.
- [76] Beer, L.; Brusso, J. L.; Cordes, A. W.; Haddon, R. C.; Itkis, M. E.; Kirschbaum, K.; MacGregor, D. S.; Oakley, R. T.; Pinkerton, A. A.; Reed, R. W. Resonance-stabilized 1,2,3-dithiazolo-1,2,3-dithiazolyls as neutral π -radical conductors. *J. Am. Chem. Soc.* **2002**, *124*, 9498–9509.
- [77] He, X.; Benniston, A. C.; Saarenpää, H.; Lemmetyinen, H.; Tkachenko, N. V.; Baisch, U. Polymorph crystal packing effects on charge transfer emission in the solid state. *Chem. Sci.* **2015**, *6*, 3525–3532.
- [78] Mayo, R. A.; Morgan, I. S.; Soldatov, D. V.; Clérac, R.; Preuss, K. E. Heisenberg spin chains via chalcogen bonding: Noncovalent S...O contacts enable long-range magnetic order. *Inorg. Chem.* **2021**, *60*, 11338–11346.

- [79] Yang, Y.; Rice, B.; Shi, X.; Brandt, J. R.; Correa da Costa, R.; Hedley, G. J.; Smilgies, D.-M.; Frost, J. M.; Samuel, I. D. W.; Otero-de-la-Roza, A.; Johnson, E. R.; Jelfs, K. E.; Nelson, J.; Campbell, A. J.; Fuchter, M. J. Emergent properties of an organic semiconductor driven by its molecular chirality. *ACS Nano* **2017**, *11*, 8329–8338.
- [80] Chung, H.; Diao, Y. Polymorphism as an emerging design strategy for high performance organic electronics. *J. Mater. Chem. C* **2016**, *4*, 3915–3933.
- [81] Li, L.; Yin, X.-H.; Diao, K.-S. Improving the solubility and bioavailability of pemafrate via a new polymorph Form II. *ACS Omega* **2020**, *5*, 26245–26252.
- [82] Beran, G. J. O. Designed and then realized. *Nat. Mater.* **2017**, *16*, 602–604.
- [83] Curtis, F.; Wang, X.; Marom, N. Effect of packing motifs on the energy ranking and electronic properties of putative crystal structures of tricyano-1,4-dithiino[c]-isothiazole. *Acta Crystallogr.* **2016**, *B72*, 562–570.
- [84] Neumann, M. A.; van der Streek, J. How many Ritonavir cases are there still out there? *Faraday Discuss.* **2018**, *211*, 441–458.
- [85] Dunitz, J. D. Are crystal structures predictable? *Chem. Commun.* **2003**, 545–548.
- [86] Desiraju, G. R. Cryptic crystallography. *Nat. Mater.* **2002**, *1*, 77–79.
- [87] Woodley, S. M.; Catlow, R. Crystal structure prediction from first principles. *Nat. Mater.* **2008**, *7*, 937–946.
- [88] Day, G. M. Current approaches to predicting molecular organic crystal structures. *Crystallog. Rev.* **2010**, *17*, 3–52.
- [89] Nyman, J.; Reutzel-Edens, S. M. Crystal structure prediction is changing from basic science to applied technology. *Phys. Chem. Chem. Phys.* **2018**, *211*, 459–476.
- [90] Beran, G. J. O. Modeling polymorphic molecular crystals with electronic structure theory. *Chem. Rev.* **2016**, *116*, 5567–5613.
- [91] Bhardwaj, R. M.; McMahon, J. A.; Nyman, J.; Price, L. S.; Konar, S.; Oswald, I. D. H.; Pulham, C. R.; Price, S. L.; Reutzel-Edens, S. M. A prolific solvate former, galunisertib, under the pressure of crystal structure prediction, produces ten diverse polymorphs. *J. Am. Chem. Soc.* **2019**, *141*, 13887–13897.
- [92] Price, S. L.; Braun, D. E.; Reutzel-Edens, S. M. Can computed crystal energy landscapes help understand pharmaceutical solids? *Chem. Commun.* **2016**, *52*, 7065–7077.
- [93] Neumann, M.; Leusen, F. J. J.; Kendrick, J. A major advance in crystal structure prediction. *Angew. Chem. Int. Ed.* **2008**, *47*, 2427–2430.

- [94] Neumann, M. A.; van de Streek, J.; Fabbiani, F. P. A.; Hidber, P.; Grassmann, O. Combined crystal structure prediction and high-pressure crystallization in rational pharmaceutical polymorph screening. *Nat. Comm.* **2015**, *6*, 7793.
- [95] LeBlanc, L. M.; Otero-de-la-Roza, A.; Johnson, E. R. Composite and low-cost approaches for molecular crystal structure prediction. *J. Chem. Theory Comput.* **2018**, *14*, 2265–2276.
- [96] Iuzzolino, L.; McCabe, P.; Price, S. L.; Brandenburg, J. G. Crystal structure prediction of flexible pharmaceutical-like molecules: Density functional tight-binding as an intermediate optimisation method and for free energy estimation. *Faraday Discuss.* **2018**, *211*, 275–296.
- [97] Hoja, J.; Tkatchenko, A. First-principles stability ranking of molecular crystal polymorphs with the DFT+MBD approach. *Phys. Chem. Chem. Phys.* **2018**, *211*, 253–274.
- [98] Macrae, C. F.; Sovago, I.; Cottrell, S. J.; Galek, P. T. A.; McCabe, P.; Pidcock, E.; Platings, M.; Shields, G. P.; Stevens, J. S.; Towler, M.; Wood, P. A. *Mercury 4.0*: From visualization to analysis, design and prediction. *J. Appl. Cryst.* **2020**, *53*, 226–235.
- [99] Price, S. L. Is zeroth order crystal structure prediction (CSP.0) coming to maturity? What should we aim for in an ideal crystal structure prediction code? *Faraday Discuss.* **2018**, *211*, 9–30.
- [100] Schneider, E.; Vogt, L.; Tuckerman, M. E. Exploring polymorphism of benzene and naphthalene with free energy based enhanced molecular dynamics. *Acta Crystallogr.* **2016**, *B72*, 542–550.
- [101] Liu, C.; Brandenburg, J. G.; Valsson, O.; Kremer, K.; Berau, T. Free-energy landscape of polymer-crystal polymorphism. *Soft Matter* **2020**, *16*, 9683–9692.
- [102] Francia, N. F.; Price, L. S.; Nyman, J.; Price, S. L.; Salvalaglio, M. Systematic finite-temperature reduction of crystal energy landscapes. *Cryst. Growth Des.* **2020**, *20*, 6847–6862.
- [103] Nyman, J.; Day, G. M. Static and lattice vibrational energy differences between polymorphs. *CrystEngComm* **2015**, *17*, 5154–5165.
- [104] Nyman, J.; Day, G. M. Modelling temperature-dependent properties of polymorphic organic molecular crystals. *Phys. Chem. Chem. Phys.* **2016**, *18*, 31132–31143.
- [105] Heit, Y. N.; Beran, G. J. O. How important is thermal expansion for predicting molecular crystal structures and thermochemistry at finite temperatures? *Acta Crystallogr.* **2016**, *B72*, 514–529.

- [106] McKinley, J. L.; Beran, G. J. O. Identifying pragmatic quasi-harmonic electronic structure approaches for modeling molecular crystal thermal expansion. *Phys. Chem. Chem. Phys.* **2018**, *211*, 181–207.
- [107] Dybeck, E. C.; McMahon, D. P.; Day, G. M.; Shirts, M. R. Exploring the multi-minima behavior of small molecule crystal polymorphs at finite temperature. *Cryst. Growth Des.* **2019**, *19*, 5568–5580.
- [108] Hofmann, D. W. M. Fast estimation of crystal densities. *Acta Crystallogr.* **2002**, *B58*, 489–493.
- [109] Bond, A. D. A survey of thermal expansion coefficients for organic molecular crystals in the Cambridge Structural Database. *Acta Crystallogr.* **2021**, *B77*, 357–364.
- [110] Mayo, R. A. vc-pwdf. <https://github.com/ramayo223/vc-pwdf.git>, 2021.
- [111] Niggli, P. *Krystallographische und strukturtheoretische Grundbegriffe. Handbuch der Experimentalphysik* **1928**, *7*, 108–176.
- [112] Andrews, L. C.; Bernstein, H. J.; Sauter, N. K. Selling reduction versus Niggli reduction for crystallographic lattices. *Acta Crystallogr.* **2019**, *A75*, 115–120.
- [113] Price, S. L.; Leslie, M.; Welch, G. W. A.; Habgood, M.; Price, L. S.; Karamertzanis, P. G.; Day, G. M. Modelling organic crystal structures using distributed multipole and polarizability-based model intermolecular potentials. *Phys. Chem. Chem. Phys.* **2010**, *12*, 8478–8490.
- [114] Sure, R.; Grimme, S. Corrected small basis set Hartree-Fock method for large systems. *J. Comput. Chem.* **2013**, *34*, 1672–1685.
- [115] Tao, J.; Perdew, J. P.; Staroverov, V. N.; Scuseria, G. E. Climbing the density functional ladder: Nonempirical meta-generalized gradient approximation designed for molecules and solids. *Phys. Rev. Lett.* **2003**, *91*, 146401.
- [116] Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate *ab initio* parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.
- [117] Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comp. Chem.* **1996**, *17*, 490–519.
- [118] Chung, H.; Diao, Y. Polymorphism as an emerging design strategy for high performance organic electronics. *J. Mater. Chem. C* **2016**, *4*, 3915–3933.
- [119] Li, L.; Yin, X.-H.; Diao, K.-S. Improving the Solubility and Bioavailability of Pemafibrate via a New Polymorph Form II. *ACS Omega* **2020**, *5*, 26245–26252.

- [120] Neumann, M. A.; van der Streek, J. How many ritonavir cases are there still out there? *Faraday Discuss.* **2018**, *211*, 441–458.
- [121] Mutai, T.; Shono, H.; Shigemitsu, Y.; Araki, K. Three-color polymorph-dependent luminescence: crystallographic analysis and theoretical study on excited-state intramolecular proton transfer (ESIPT) luminescence of cyano-substituted imidazo [1, 2-a] pyridine. *CrystEngComm* **2014**, *16*, 3890–3895.
- [122] Mayo, R. A.; Johnson, E. R. Improved quantitative crystal-structure comparison using powder diffractograms via anisotropic volume correction. *CrystEngComm* **2021**, *23*, 7118–7131, Chapter 4 in this thesis.
- [123] Spek, A. Single-crystal structure validation with the program PLATON. *J. Appl. Cryst.* **2003**, *36*, 7–13.
- [124] Ullmann, J. R. An algorithm for subgraph isomorphism. *JACM* **1976**, *23*, 31–42.
- [125] Zuñiga, F. J.; Cruz-Cabeza, A. J.; Aretxabaleta, X. M.; de la Pinta, N.; Brezowski, T.; Quesada-Moreno, M. M.; Avilés-Moreno, J. R.; López-González, J. J.; Claramunt, R. M.; Elguero, J. Conformational aspects of polymorphs and phases of 2-propyl-1H-benzimidazole. *IUCrJ* **2018**, *5*, 706–715.
- [126] Ivanisevic, I.; McClurg, R. B.; Schields, P. J. *Pharmaceutical Sciences Encyclopedia*; John Wiley & Sons, Ltd, 2010; Chapter 16, pp 1–42.
- [127] Schlesinger, C.; Fitterer, A.; Buchsbaum, C.; Habermehl, S.; Chierotti, M. R.; Nervi, C.; Schmidt, M. Ambiguous structure determination from powder data: four different structural models of 4, 11-difluoroquinacridone with similar X-ray powder patterns, fit to the PDF, SSNMR and DFT-D. *IUCrJ* **2022**, *9*, 406–424.
- [128] Woodley, S. M.; Catlow, R. Crystal structure prediction from first principles. *Nat. Mater.* **2008**, *7*, 937–946.
- [129] Zhao, C.; Chen, L.; Che, Y.; Pang, Z.; Wu, X.; Lu, Y.; Liu, H.; Day, G. M.; Cooper, A. I. Digital navigation of energy–structure–function maps for hydrogen-bonded porous molecular crystals. *Nat. Commun.* **2021**, *12*, 817.
- [130] Zhu, Q.; Johal, J.; Widdowson, D. E.; Pang, Z.; Li, B.; Kane, C. M.; Kurlin, V.; Day, G. M.; Little, M. A.; Cooper, A. I. Analogy powered by prediction and structural invariants: computationally led discovery of a mesoporous hydrogen-bonded organic cage crystal. *J. Am. Chem. Soc.* **2022**, *144*, 9893–9901.
- [131] Taylor, C. R.; Mulvee, M. T.; Perenyi, D. S.; Probert, M. R.; Day, G. M.; Steed, J. W. Minimizing polymorphic risk through cooperative computational and experimental exploration. *J. Am. Chem. Soc.* **2020**, *142*, 16668–16680.
- [132] Dudek, M. K.; Druźbicki, K. Along the road to crystal structure prediction (CSP) of pharmaceutical-like molecules. *CrystEngComm* **2022**, *24*, 1665–1678.

- [133] Abramov, Y. A. Current computational approaches to support pharmaceutical solid form selection. *Org. Process Res. Dev.* **2013**, *17*, 472–485.
- [134] Price, S. L. From crystal structure prediction to polymorph prediction: interpreting the crystal energy landscape. *Phys. Chem. Chem. Phys.* **2008**, *10*, 1996–2009.
- [135] Morissette, S. L.; Almarsson, Ö.; Peterson, M. L.; Remenar, J. F.; Read, M. J.; Lemmo, A. V.; Ellis, S.; Cima, M. J.; Gardner, C. R. High-throughput crystallization: polymorphs, salts, co-crystals and solvates of pharmaceutical solids. *Adv. Drug Deliv. Rev.* **2004**, *56*, 275–300.
- [136] Mayo, R. A.; Otero-de-la Roza, A.; Johnson, E. R. Development and assessment of an improved powder-diffraction-based method for molecular crystal structure similarity. *CrystEngComm* **2022**, *24*, 8326–8338, Chapter 5 in this thesis.
- [137] Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge structural database. *Acta Crystallogr.* **2016**, *B72*, 171–179.
- [138] The CPOSS (Control and Prediction of the Organic Solid State) database holds details of the hypothetical crystal structures generated in the computational searches carried out at University College London, and is maintained by Dr. Louise S. Price, and led by Prof. Sally L. Price. See <http://www.chem.ucl.ac.uk/cposs/index.htm>.
- [139] Mayo, R. A. VC-(x)PWDF Instruction Manual. 2023; https://erin-r-johnson.github.io/downloads/VC-PWDF_Manual-v2.pdf.
- [140] Ghosh, R.; Shirley, R. Crysfire2020. 2020; <http://ccp14.cryst.bbk.ac.uk/Crysfire.html>.
- [141] Taupin, D. A powder-diagram automatic-indexing routine. *J. Appl. Cryst.* **1973**, *6*, 380–385.
- [142] Kohlbeck, F.; Hörl, E. Indexing program for powder patterns especially suitable for triclinic, monoclinic and orthorhombic lattices. *J. Appl. Cryst.* **1976**, *9*, 28–33.
- [143] Enright, G. D.; Terskikh, V. V.; Brouwer, D. H.; Ripmeester, J. A. The structure of two anhydrous polymorphs of caffeine from single-crystal diffraction and ultrahigh-field solid-state ¹³C NMR spectroscopy. *Cryst. Growth Des.* **2007**, *7*, 1406–1410.
- [144] Mullen, D. Electron-density distribution in urea. A multipolar expansion. *Acta Crystallogr.* **1980**, *B36*, 1610–1615.
- [145] Guth, H.; Heger, G.; Klein, S.; Treutmann, W.; Scheringer, C. Strukturverfeinerung von Harnstoff mit Neutronenbeugungsdaten bei 60, 123 und 293 K und XN- und XX (1s2)-Synthesen bei etwa 100 K. *Z. Kristallogr. Krist.* **1980**, *153*, 237–254.

- [146] Swaminathan, S.; Craven, B.; McMullan, R. The crystal structure and molecular thermal motion of urea at 12, 60 and 123 K from neutron diffraction. *Acta Crystallogr.* **1984**, *B40*, 300–306.
- [147] Nikolic, V.; Stankovic, M.; Nikolic, L.; Cvetkovic, D.; Kapor, A.; Cakic, M. CCDC 239749: Experimental Crystal Structure Determination. 2009; http://www.ccdc.cam.ac.uk/services/structure_request?id=doi:10.5517/cc81gvl&sid=DataCite.
- [148] Olejniczak, A.; Ostrowska, K.; Katrusiak, A. H-bond breaking in high-pressure urea. *J. Phys. Chem. C.* **2009**, *113*, 15761–15767.
- [149] Roszak, K.; Katrusiak, A. Giant anomalous strain between high-pressure phases and the mesomers of Urea. *J. Phys. Chem. C.* **2017**, *121*, 778–784.
- [150] Guth, H.; Heger, G.; Drück, U. Neutron diffraction study of the structure of p-benzenedicarbonitrile. *Z. Kristallogr. Krist.* **1982**, *159*, 185–190.
- [151] Yuan, J.; Wang, Y.; Li, L.; Wang, S.; Tang, X.; Wang, H.; Li, M.; Zheng, C.; Chen, R. Activating intersystem crossing and aggregation coupling by CN-substitution for efficient organic ultralong room temperature phosphorescence. *J. Phys. Chem. C* **2020**, *124*, 10129–10134.
- [152] Kubiak, R.; Janczak, J.; Dahl, O.; Hvistendahl, G.; Leskelä, M.; Polamo, M.; Homsí, M.; Kuske, F.; Haugg, M.; Trabesinger-Rüf, N.; Weinhold, E. A new crystalline (alpha) form of 1,4-dicyanobenzene. *Acta Chem. Scand.* **1996**, *50*, 1164–1167.
- [153] Stewart, R. F.; Jensen, L. H. Redetermination of the crystal structure of uracil. *Acta Crystallogr.* **1967**, *23*, 1102–1105.
- [154] Lehmann, C. W.; Stowasser, F. The crystal structure of anhydrous β -caffeine as determined from X-ray powder-diffraction data. *Chem. Eur. J.* **2007**, *13*, 2908–2911.
- [155] Anitha, R.; Gunasekaran, M.; Kumar, S. S.; Athimoolam, S.; Sridhar, B. Single crystal XRD, vibrational and quantum chemical calculation of pharmaceutical drug paracetamol: A new synthesis form. *Spectrochim. Acta A Mol. Biomol. Spectrosc.* **2015**, *150*, 488–498.
- [156] Naumov, D. Y.; Vasilchenko, M.; Howard, J. The monoclinic form of acetaminophen at 150K. *Acta Crystallogr.* **1998**, *B54*, 653–655.
- [157] Wilson, C. Variable temperature study of the crystal structure of paracetamol (p-hydroxyacetanilide), by single crystal neutron diffraction. *Z. Kristallogr. Krist.* **2000**, *215*, 693–701.
- [158] Boldyreva, E. V.; Shaktshneider, T. P.; Vasilchenko, M. A.; Ahsbahs, H.; Uchtmann, H. Anisotropic crystal structure distortion of the monoclinic polymorph of acetaminophen at high hydrostatic pressures. *Acta Crystallogr.* **2000**, *B56*, 299–309.

- [159] Shtukenberg, A. G.; Tan, M.; Vogt-Maranto, L.; Chan, E. J.; Xu, W.; Yang, J.; Tuckerman, M. E.; Hu, C. T.; Kahr, B. Melt crystallization for paracetamol polymorphism. *Cryst. Growth Des.* **2019**, *19*, 4070–4080.
- [160] Reiss, C. A.; Van Mechelen, J. B.; Goubitz, K.; Peschar, R. Reassessment of paracetamol orthorhombic Form III and determination of a novel low-temperature monoclinic Form III-m from powder diffraction data. *Acta Crystallogr.* **2018**, *C74*, 392–399.
- [161] Chan, E.; Goossens, D. Study of the single-crystal X-ray diffuse scattering in paracetamol polymorphs. *Acta Crystallogr.* **2012**, *B68*, 80–88.
- [162] Blejwas, I.; Gołdyn, M.; Bartoszek-Adamska, E. Crystal and molecular structure of D-mannitol and its thermal stability. *Postepy Nauki i Technologii Przemyslu Rolno-Spozywczego* **2019**, *74*, 5–19.
- [163] Kaminsky, W.; Glazer, A. Crystal optics of D-mannitol, C₆H₁₄O₆: crystal growth, structure, basic physical properties, birefringence, optical activity, Faraday effect, electro-optic effects and model calculations. *Z. Kristallogr. Krist.* **1997**, *212*, 283–296.
- [164] Fronczek, F. R.; Kamel, H. N.; Slattery, M. Three polymorphs (α , β , and δ) of D-mannitol at 100 K. *Acta Crystallogr.* **2003**, *C59*, o567–o570.
- [165] Vivy, CCDC 913551: Experimental Crystal Structure Determination. 2013; http://www.ccdc.cam.ac.uk/services/structure_request?id=doi:10.5517/ccznmdm&sid=DataCite.
- [166] Lancaster, R. W.; Karamertzanis, P. G.; Hulme, A. T.; Tocher, D. A.; Lewis, T. C.; Price, S. L. The polymorphism of progesterone: Stabilization of a ‘disappearing’ polymorph by co-crystallization. *J. Pharm. Sci.* **2007**, *96*, 3419–3431.
- [167] Campsteyn, H.; Dupont, L.; Dideberg, O. Structure cristalline et moléculaire de la progestérone, C₂₁H₃₀O₂. *Acta Crystallogr.* **1972**, *B28*, 3032–3042.
- [168] Bergmann, J.; Friedel, P.; Kleeberg, R. BGMN - a new fundamental parameters based Rietveld program for laboratory X-ray Sources, its use in quantitative analysis and structure investigations. *IUCr Commission on Powder Diffraction Newsletter* **1998**, 5–8.
- [169] Doebelin, N.; Kleeberg, R. Profex: a graphical user interface for the Rietveld refinement program BGMN. *J. Appl. Cryst.* **2015**, *48*, 1573–1580.
- [170] The autoFIDEL code was written by Dr. Jonas Nyman based on the FIDEL protocol outlined by Schmidt *et al.* in Ref. 74 and has been copyrighted to the CCDC as of mid-2022.
- [171] X’pert Data Collector Software, version 5.3; PANalytical B.V.: Almelo, The Netherlands, 2014.

- [172] Lancaster, R. W.; Karamertzanis, P. G.; Hulme, A. T.; Tocher, D. A.; Covey, D. F.; Price, S. L. Racemic progesterone: predicted in silico and produced in the solid state. *Chem. Commun.* **2006**, 4921–4923.
- [173] Price, S. L.; Wibley, K. S. Predictions of crystal packings for uracil, 6-azauracil, and allopurinol: The interplay between hydrogen bonding and close packing. *J. Phys. Chem. A* **1997**, *101*, 2198–2206.
- [174] Issa, N.; Barnett, S. A.; Mohamed, S.; Braun, D. E.; Copley, R. C.; Tocher, D. A.; Price, S. L. Screening for cocrystals of succinic acid and 4-aminobenzoic acid. *CrystEngComm* **2012**, *14*, 2454–2464.
- [175] Habgood, M. Form II caffeine: a case study for confirming and predicting disorder in organic crystals. *Cryst. Growth Des.* **2011**, *11*, 3600–3608.
- [176] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, J. M.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, O.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J. Gaussian 09 Revision B.01. Gaussian Inc. Wallingford CT 2010.
- [177] Stone, A.; Alderton, M. Distributed multipole analysis. *Mol. Phys.* **1985**, *56*, 1047–1064.
- [178] Holden, J. R.; Du, Z.; Ammon, H. L. Prediction of possible crystal structures for C-, H-, N-, O-, and F-containing organic compounds. *J. Comput. Chem.* **1993**, *14*, 422–437.
- [179] Karamertzanis, P.; Pantelides, C. Ab initio crystal structure prediction. II. Flexible molecules. *Mol. Phys.* **2007**, *105*, 273–291.
- [180] Williams, D. E. Improved intermolecular force field for molecules containing H, C, N, and O atoms, with application to nucleoside and peptide crystals. *J. Comput. Chem.* **2001**, *22*, 1154–1166.
- [181] Coombes, D. S.; Price, S. L.; Willock, D. J.; Leslie, M. Role of electrostatic interactions in determining the crystal structures of polar organic molecules. A distributed multipole study. *J. Phys. Chem.* **1996**, *100*, 7352–7360.

- [182] Willock, D. J.; Price, S. L.; Leslie, M.; Catlow, C. R. A. The relaxation of molecular crystal structures using a distributed multipole electrostatic model. *J. Comput. Chem.* **1995**, *16*, 628–647.
- [183] Kazantsev, A.; Karamertzanis, P.; Adjiman, C.; Pantelides, C. Efficient handling of molecular flexibility in lattice energy minimization of organic crystals. *J. Chem. Theory Comput.* **2011**, *7*, 1998–2016.