Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Development and assessment of an improved powder-diffraction-based method for molecular crystal structure similarity

R. Alex Mayo,[a] Alberto Otero-de-la-Roza,[b] and Erin R. Johnson[a*]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Identifying whether two experimental crystal structures determined under different experimental conditions correspond to the same polymorph is a challenging problem in crystallography, and its solution has practical (and even legal) implications for various technological fields as well as for molecular crystal structure prediction. In this work, we assess 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 [CrystEngComm, 2020, **22**, 7170–7185]. We propose a new PXRD-based similarity index and comparison method (VC-PWDF, variable-cell powder difference) that substantially improves the agreement with COMPACK (2.84% total disagreement), compared to the CCDC packing similarity PXRD-based comparison tool (12.4%). By analysing the structure pairs for which COMPACK and VC-PWDF disagree, we evaluated the strengths and weaknesses of each method. 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 at fixed number of matching molecules (N). We demonstrated a few additional weaknesses of COMPACK: a) extremely costly or incorrect comparisons in molecules with highly-branched substituents (possibly due to its use of Ullmann's method), b) failure to match structures when the molecular connectivity is incorrectly determined, c) difficulties with molecules presenting helical chirality, and d) very large cluster sizes (up to 50 molecules) 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 is at least as robust as COMPACK for comparing molecular crystal structures. Although a grey area of difficult-to-compare structures still exists, using both COMPACK and VC-PWDF in combination may be successful at narrowing it.

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.[1–5] The molecular packing is a critical part of what determines the luminescent,[6] optoelectric,[7–9] and magnetic[1,10] properties of materials, and the efficacy of pesticides,[11,12] nutraceuticals,[13] and drugs[14]. 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.[15–21]

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:

[a] Dalhousie University, Department of Chemistry; E-mail: erin.johnson@dal.ca
[b] Universidad de Oviedo, Oviedo, Spain

- $d(A,B) = 0$ 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.[15] 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, 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 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.[16] 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 rigorously[21]).

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 (isotropic[22] or anisotropic[23]) 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.[24] 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.*,[25] assessed the two comparison methods available within the Cambridge Crystallographic Data Centre's (CCDC's) software suite:[26] 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.[22] 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.[27,28]

We recently developed a new approach to improve PXRD-based comparison methods using anisotropic volume corrections.[23] 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.[29] 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.[30] 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.*[25] 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.

2 Methods

2.1 Mercury's CPS tool

The CPS tool provided with the CCDC's Mercury program[26] was executed through the CSD Python API. All structures were accessed from the CSD using their refcodes (see the ESI† 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 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 ±10% in the distances and ±10° 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.

2.2 Variable-cell powder difference (VC-PWDF)

The method described herein is an improvement of the one recently published by the authors.[23] The dependence that the previous version had on the cell description is resolved as described below, and the code integrated into the `critic2` program.[30] 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.[31]
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.[16]
3. Maximum elongations and angle differences relative to the reference cell are defined. By default, these are ±30% in the cell lengths and ±20° 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 ±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 ±20° 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.[23] 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 $\ell = 1$ in the weighted cross-correlation function.[16]

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 method[23] 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.

3 Dataset

The set of structures used in this work is the same as in Sacchi *et al.*[25] 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 the ESI†.

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's[32] checkcif identified Alert Level A flagged voids in one of the structures of the pair. (See the ESI† for an illustrative example and the list of the 78 structures with voids.)
- 87 pairs were removed due to 8 problematic structures (see the ESI†), 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 ESI†), 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 5.1.

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

4 Results

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 5.2 and 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,[25] the confusion matrices comparing COMPACK with the two different PXRD-based methods (VC-PWDF and CPS-PWDF) are shown in Table 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.*[25]

Table 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 literature[25] were used.

Literature data[25] using CPS-PWDF

CPS-PWDF	COMPACK	
	same	different
same	47.79%	2.05%
different	13.87%	35.88%

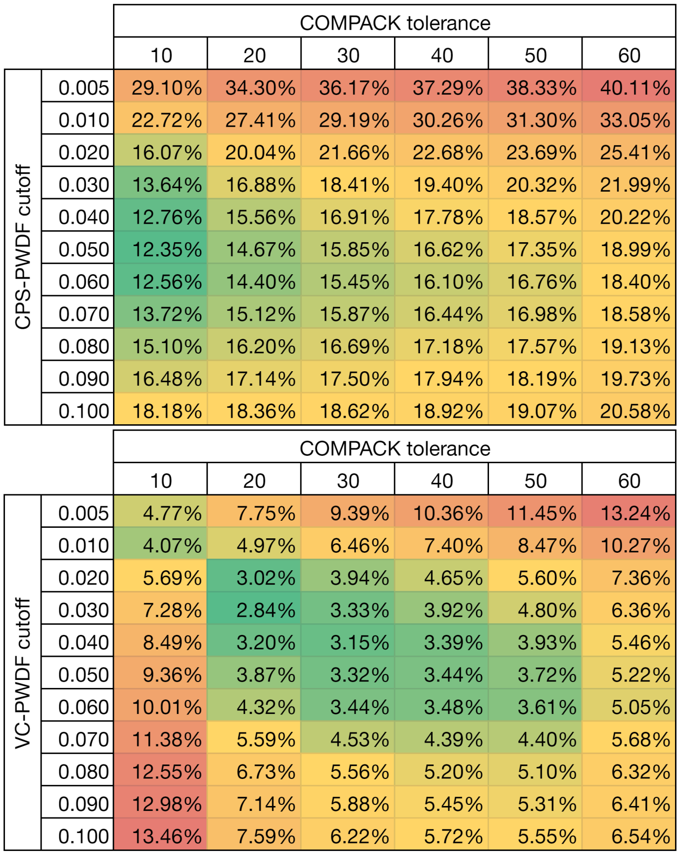
Current data using VC-PWDF

VC-PWDF	COMPACK	
	same	different
same	61.04%	1.96%
different	0.93%	36.06%

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 were 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 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.

4.2 Dependence on tolerances and cutoffs

The PWDF cutoff of 0.035 used by Sacchi *et al.* in 2020[25] was selected based on the initial survey of the CSD by van de Streek and Motherwell.[22] 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 1 presents these results in the form of a heat map, where either CPS-PWDF (top) or VC-PWDF (bottom) is used.



CPS-PWDF cutoff	COMPACK tolerance 10	20	30	40	50	60
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%

VC-PWDF cutoff	COMPACK tolerance 10	20	30	40	50	60
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%

Fig. 1 Heat maps representing the percentage of comparisons for which COMPACK and CPS-PWDF (top) or VC-PWDF (bottom) disagree on the outcome, as a function of the PWDF cutoff and COMPACK tolerances 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.[15,22] As shown by the confusion matrix in Table 2, this choice results in the instances of disagreement where COMPACK 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.[25]

Table 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[25] 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%

In addition to the difference between the minimum disagreement values, the difference in the topography of the two heat maps in Figure 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 COMPACK considers pairs the same and VC-PWDF considers them different, and vice-versa, are shown in Figure 2 (top and bottom, respectively). Some anomalies are 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 4.3.

4.3 COMPACK tolerance behaviour

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 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 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%

		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%

Fig. 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.

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.

Table 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.

	Change in COMPACK tolerance				
Cases of:	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

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 2. Our previous study[23] 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 3 shows that loosening the tolerances beyond 40 identifies a further 1,316 matches.

Table 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.

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 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.

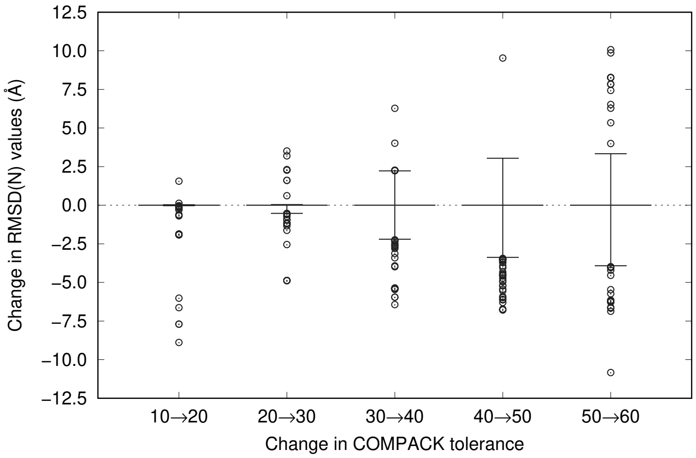


Fig. 3 ΔRMSD(N) values as a function of changes in COMPACK tolerance. The whiskers covers 99.9% of the data about the median, and outliers are shown as circles.

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) ΔRMSD(N). In addition, the whiskers hardly extend beyond 0 Å for the 10→20 and 20→30 changes in tolerance. 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 RMSD(N) values (recall that there is no change in N). The magnitudes of the greatest RMSD(N) changes also generally increase with tolerance, with the exception of the most negative ΔRMSD(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 RMSD(N) values (ΔRMSD(N) = -8.887, -7.702, -7.695, -6.630, and -6.016 Å, respectively) when the tolerance is increased from 10 to 20. All except VOQHIU-VOQHIU01 are RMSD(20) values. However, there are many cases with ΔRMSD(N)>0 with loosening tolerance, which would be in violation of the cutoff principle if 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 RMSD(20) values of 10.075, 9.868, 8.273, and 8.255 Å, respectively.

In both the cases of 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 RMSD(N) is high, and the overlay is excellent when it is small. An example for the SUCROS27-SUCROS33 comparison is shown in the ESI†. 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 Discussion

5.1 Issues with Ullmann's algorithm for highly branched molecules

It was noted in Section 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 4, and a full list of the removed refcode families is given in the ESI†. 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 5.2.2).

The appearance of several highly branched substituents likely causes problems with Ullmann's algorithm,[33] a modified version of which is used in COMPACK.[15] 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 to decrease the cost of the tree search even further.[15]

Our own implementation of the method in `critic2` shows that molecules such as those appearing in Table 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 4.

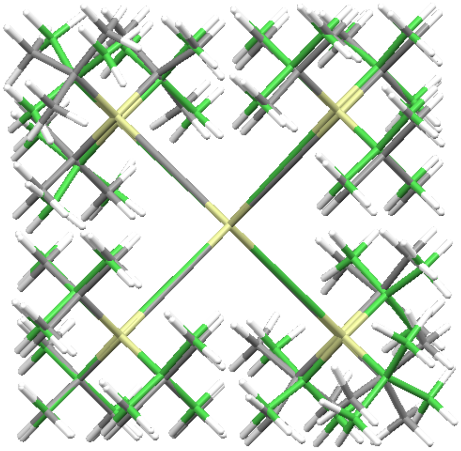
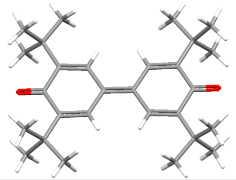
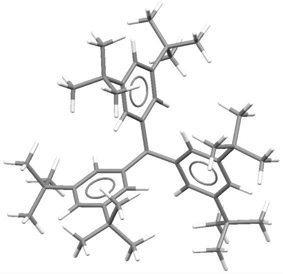
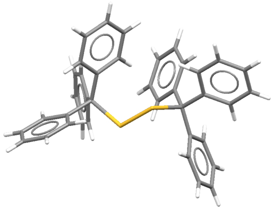
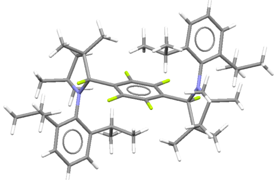
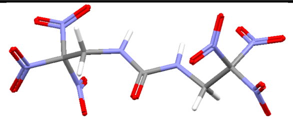


Fig. 4 Best overlay generated by COMPACK for a hypothetical molecule consisting of four tri-tert-butylsilane substituents bonded to a central silicon atom by ethyne linkers (RMSD(1) = 2.313 Å). The comparison and reference structures are identical with the exception of the order in which the atoms appear 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.*[25] as a fault of the CPS-PWDF method,

Table 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, ZEDCUG
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

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. It was the erroneous overlay generated by COMPACK that was at fault, probably stemming from their use of Ullmann's graph-matching algorithm.

5.2 VC-PWDF same/COMPACK different

As noted in Section 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 minimum on the corresponding heat map (Figure 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 5.1) 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.

5.2.1 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 $C-CH_3$. An example is shown for the ZITZUX-ZITZUX01 pair in Figure 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.*[34] (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

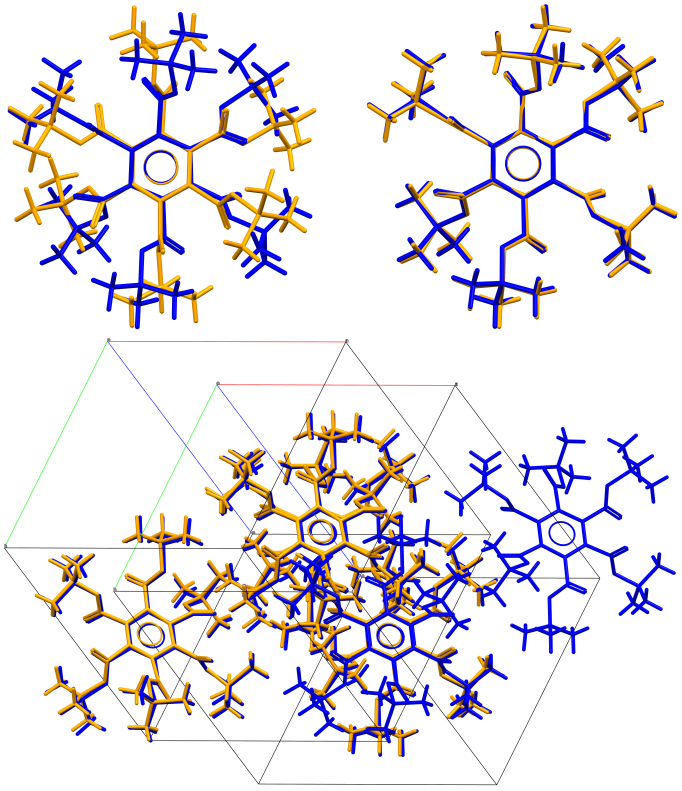


Fig. 5 COMPACK "optimum" overlay for a single molecule of ZEDCUG and ZEDCUG01 (top-left), manual overlay for a single molecule, showing perfect coincidence (top-right). Overlay of ZEDCUG and ZEDCUG01 showing coincident molecular position and orientation, done manually (bottom).

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.

5.2.2 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 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.

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 8, left) and not having any of the structural moieties identified as problematic for Ullmann's method in Section 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 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.

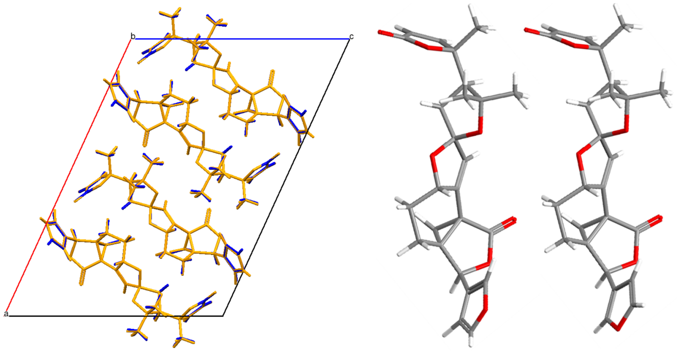


Fig. 6 An example of possible conformational phases (could be atom misassignment during structure solution) ZITZUX and ZITZUX01. The overlay of the two structures is shown, illustrating the identical packing (left), and the difference in the furyl ring orientation (right).

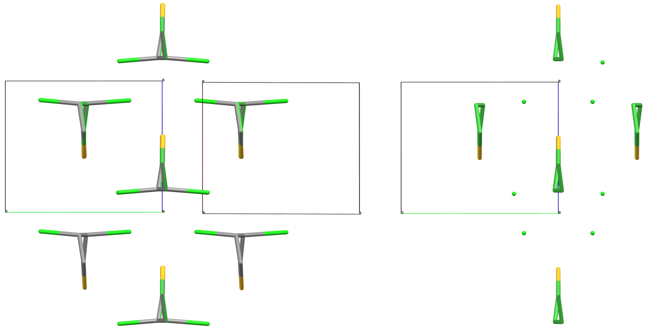


Fig. 7 COMPACK was used to overlay JIYKAD and JIYKAD01 (13/20 molecules match), the left plot shows both structures overlaid (perfect agreement) and the right plot shows only the JIYKAD01 structure. The chlorine atoms in JIYKAD01 are not bonded to the tetrahydrothiophene ring and are considered separate "molecules" by COMPACK.

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 9 for the SILVAL-SILVAL02 pair. Polytype structure pairs are different polymorphs, although their similarity is clearly apparent. The over-

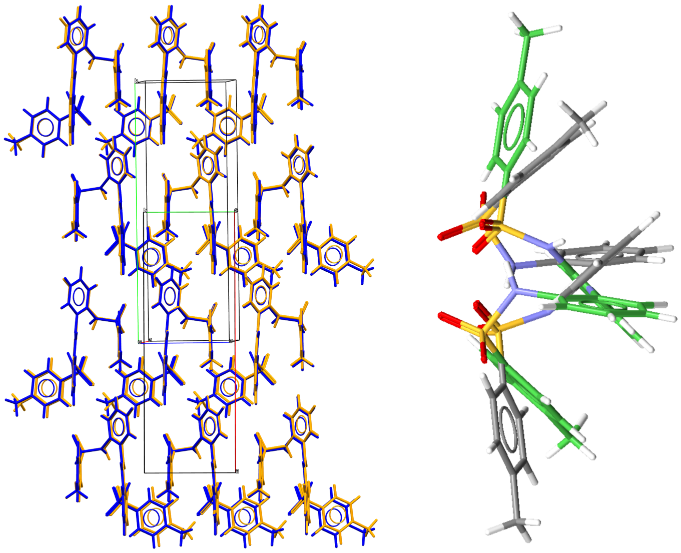


Fig. 8 Manual overlay (left) and COMPACK optimum overlay (1/20 molecules, right) for the comparison of UHIKUR and UHIKUR01.

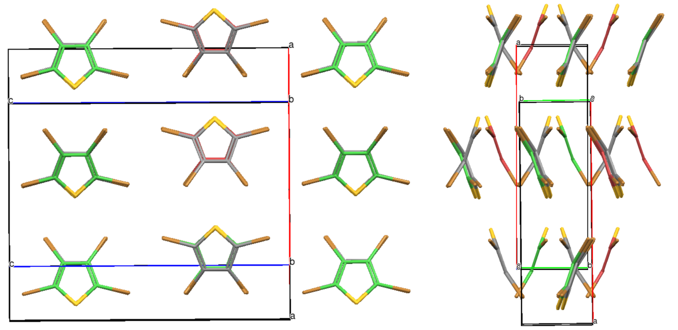


Fig. 9 COMPACK overlays of the polytype structures SILVAL and SILVAL02 in the (010) and (100) planes (left and right, respectively).

all 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.

5.3 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 2, 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 10. CBMZPN03 is a rhombohedral polymorph ($R\bar{3}$, rhombohedral lattice) with larger-than-average (`Platon`'s[32] 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 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.

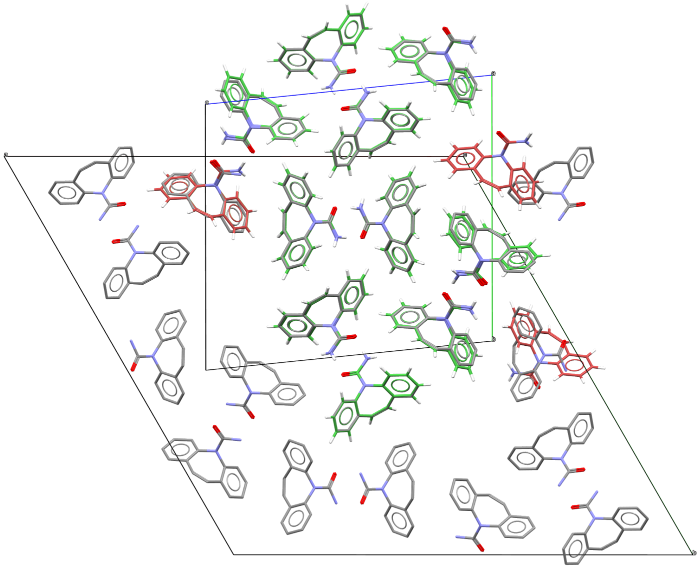


Fig. 10 Overlay of CBMZPN03 and CBMZPN11 generated by COMPACK using a cluster of 40 molecules.

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,[15] 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.

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.[23] 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.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors thank Dr. Aurora Cruz-Cabeza for sharing the data and results from their 2020 paper.[25] We also thank the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Walter C. Sumner Foundation for financial support. AOR thanks the Spanish Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación (AEI) (project PGC2018-097520-A-100), the Principality of Asturias (FICYT) and FEDER (project AYUD/2021/51036), and the Spanish MINECO for a RamÃşn y Cajal fellowship (RyC-2016-20301).

References

1 W. Fujita and K. Awaga, *Science*, 1999, **286**, 261–262.

2 H. Chung and Y. Diao, *J. Mater. Chem. C*, 2016, **4**, 3915–3933.

3 L. Li, X.-H. Yin and K.-S. Diao, *ACS Omega*, 2020, **5**, 26245–26252.

4 S. R. Chemburkar, J. Bauer, K. Deming, H. Spiwek, K. Patel, J. Morris, R. Henry, S. Spanton, W. Dziki, W. Porter, J. Quick, P. Bauer, J. Donaubauer, B. A. Narayanan, M. Soldani, D. Riley and K. McFarland, *Org. Proc. Res. Dev.*, 2000, **4**, 413–417.

5 M. A. Neumann and J. van der Streek, *Faraday Discuss.*, 2018, **211**, 441–458.

6 T. Mutai, H. Shono, Y. Shigemitsu and K. Araki, *CrystEngComm*, 2014, **16**, 3890–3895.

7 M. Li, A. H. Balawi, P. J. Leenaers, L. Ning, G. H. Heintges, T. Marszalek, W. Pisula, M. M. Wienk, S. C. Meskers, Y. Yi, F. Laquai and R. A. J. Janssen, *Nat. Commun.*, 2019, **10**, 1–11.

8 A. Troisi and G. Orlandi, *J. Phys. Chem. B*, 2005, **109**, 1849–1856.

9 M. Courte, J. Ye, H. Jiang, R. Ganguly, S. Tang, C. Kloc and D. Fichou, *Phys. Chem. Chem. Phys.*, 2020, **22**, 19855–19863.

10 A. Y. Ganin, Y. Takabayashi, P. Jeglič, D. Arčon, A. Potočnik, P. J. Baker, Y. Ohishi, M. T. McDonald, M. D. Tzirakis, A. McLennan, G. R. Darling, M. Takata, M. J. Rosseinsky and K. Prassides, *Nature*, 2010, **466**, 221–225.

11 J. Yang, C. T. Hu, X. Zhu, Q. Zhu, M. D. Ward and B. Kahr, *Angew. Chem. Int. Ed.*, 2017, **56**, 10165–10169.

12 J. Yang, B. Erriah, C. T. Hu, E. Reiter, X. Zhu, V. López-Mejías, I. P. Carmona-Sepúlveda, M. D. Ward and B. Kahr, *Proc. Natl. Acad. Sci.*, 2020, **117**, 26633–26638.

13 J.-R. Wang, B. Zhu, Q. Yu and X. Mei, *CrystEngComm*, 2016, **18**, 1101–1104.

14 D. Singhal and W. Curatolo, *Adv. Drug Deliv. Rev.*, 2004, **56**, 335–347.

15 S. Motherwell and J. A. Chisholm, *J. Appl. Cryst.*, 2005, **38**, 228–231.

16 R. de Gelder, R. Wehrens and J. A. Hageman, *J. Comput. Chem.*, 2001, **22**, 273–289.

17 B. P. Van Eijck and J. Kroon, *J. Comput. Chem.*, 1997, **18**, 1036–1042.

18 A. Dzyabchenko, *Acta Crystallogr.*, 1994, **B50**, 414–425.

19 H. Karfunkel, B. Rohde, F. Leusen, R. J. Gdanitz and G. Rihs, *J. Comput. Chem.*, 1993, **14**, 1125–1135.

20 M. Valle and A. R. Oganov, *Acta Crystallogr.*, 2010, **A66**, 507–517.

21 M. M. Mosca and V. Kurlin, *Cryst. Res. Technol.*, 2020, **55**, 1900197.

22 J. van de Streek and S. Motherwell, *Acta Crystallogr.*, 2005, **B61**, 504–510.

23 R. A. Mayo and E. R. Johnson, *CrystEngComm*, 2021, **23**, 7118–7131.

24 S. Habermehl, P. Mörschel, P. Eisenbrandt, S. M. Hammer and M. U. Schmidt, *Acta Crystallogr*, 2014, **B70**, 347–359.

25 P. Sacchi, M. Lusi, A. J. Cruz-Cabeza, E. Nauha and J. Bernstein, *CrystEngComm*, 2020, **22**, 7170–7185.

26 C. F. Macrae, I. Sovago, S. J. Cottrell, P. T. A. Galek, P. McCabe, E. Pidcock, M. Platings, G. P. Shields, J. S. Stevens, M. Towler and P. A. Wood, *J. Appl. Cryst.*, 2020, **53**, 226–235.

27 A. D. Bond, *Acta Crystallogr.*, 2021, **B77**, 357–364.

28 A. van der Lee and D. G. Dumitrescu, *Chem. Sci.*, 2021, **12**, 8537–8547.

29 A. M. Reilly, R. I. Cooper, C. S. Adjiman, S. Bhattacharya, A. D. Boese, J. G. Brandenburg, P. J. Bygrave, R. Bylsma, J. E. Campbell, R. Car, D. H. Case, R. Chadha, J. C. Cole, K. Cosburn, H. M. Cuppen, F. Curtis, G. M. Day, R. A. DiStasio Jr, A. Dzyabchenko, B. P. van Eijck, D. M. Elking, J. A. van den Ende, J. C. Facelli, M. B. Ferraro, L. Fusti-Molnar, C.-A. Gatsiou, T. S. Gee, R. de Gelder, L. M. Ghiringhelli, H. Goto, S. Grimme, R. Guo, D. W. M. Hofmann, J. Hoja, R. K. Hylton, L. Iuzzolino, W. Jankiewicz, D. T. de Jong, J. Kendrick, N. J. J. de Klerk, H.-Y. Ko, L. N. Kuleshova, X. Li, S. Lohani, F. J. J. Leusen, A. M. Lund, J. Lv, Y. Ma, N. Marom, A. E. Masunov, P. McCabe, D. P. McMahon, H. Meekes, M. P. Metz, A. J. Misquitta, S. Mohamed, B. Monserrat, R. J. Needs, M. A. Neumann, J. Nyman, S. Obata, H. Oberhofer, A. R. Oganov, A. M. Orendt, G. I. Pagola, C. C. Pantelides, C. J. Pickard, R. Podeszwa, L. S. Price, S. L. Price, A. Pulido, M. G. Read, K. Reuter, E. Schneider, C. Schober, G. P. Shields, P. Singh, I. J. Sugden, K. Szalewicz, C. R. Taylor, A. Tkatchenko, M. E. Tuckerman, F. Vacarro, M. Vasileiadis, A. Vazquez-Mayagoitia, L. Vogt, Y. Wang, R. E. Watson, G. A. de Wijs, J. Yang, Q. Zhu and C. R. Groom, *Acta Crystallogr.*, 2016, **B72**, 439–459.

30 A. Otero-de-la-Roza, E. R. Johnson and V. LuaÃśa, *Comput. Phys. Commun.*, 2014, **185**, 1007–1018.

31 P. Niggli, *Krystallographische und strukturtheoretische Grundbegriffe. Handbuch der Experimentalphysik*, 1928, **7**, 108–176.

32 A. Spek, *J. Appl. Cryst.*, 2003, **36**, 7–13.

33 J. R. Ullmann, *JACM*, 1976, **23**, 31–42.

34 F. J. ZuÃśiga, A. J. Cruz-Cabeza, X. M. Aretxabaleta, N. de la Pinta, T. Breczewski, M. M. Quesada-Moreno, J. R. AvilÃľs-Moreno, J. J. LÃşpez-GonzÃąlez, R. M. Claramunt and J. Elguero, *IUCrJ*, 2018, **5**, 706âĂŞ715.