Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Improved Quantitative Crystal-Structure Comparison using Powder Diffractograms via Anisotropic Volume Correction

R. Alex Mayo and Erin R. Johnson*

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

Crystal structure prediction (CSP) aims to determine the experimentally isolable crystal structure(s) of a molecule given only its 2D molecular diagram. The ability to match candidate structures to known experimental structures is critical in benchmarking CSP methods. In this work, a new approach to improve comparison of crystal structures using their calculated powder X-ray diffractograms (PXRD) is presented. The protocol involves anisotropic volume correction of the compared structure to that of the target. Its ability to distinguish matching structures from other candidates is assessed using the submissions to the 6th CSP blind test. The anisotropic volume correction is found to surpass currently available methods of PXRD comparison in its ability to separate similar from dissimilar structures. This is demonstrated by its ability to distinguish a polytype from a target structure, and by the identification of two uncredited matching structures in the 6th CSP blind test. The developed method yields a quantitative measure that is as useful as the root-mean-square deviation (RMSD) in atomic positions for structure comparison.

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.[1–4] The discovery of polymorphs for a compound of interest has the potential to realize the desired properties of a material,[5–8] or to severely complicate its production.[9,10]

The ideals of first-principles CSP[11–15] 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.[16–21] 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 reranking with higher-level theoretical methods is often performed at later stages of the CSP protocol.[18,22–27]

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).[28–33] 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 interatomic distances for a defined cluster size, and comparison of calculated powder X-ray diffractograms (PXRD). The COMPACK algorithm,[34] implemented in the CCDC's Mercury software Crystal Packing Similarity (CPS) tool,[35] 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 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 Gelder[36] 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 similarly, where larger values (approaching 1) indicate more similar structures,[37] and (ii) powder pattern difference, where smaller values (approaching 0) indicate more similar structures.[38] Due to the powder difference (POWDIFF) 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.[39] While thermal effects on the lattice can be modeled via molecular dynamics simulations,[40–42] or through use of the quasi-harmonic approximation,[43–46] 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.[45] They may also potentially exhibit unphysical conformational differences in cases with highly flexible molecules due to neglect of thermal entropy.[47]

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 der Streek and Motherwell.[37] 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 volumes[48] 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.[35] 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.[37,49]

A recent study by Bernstein and co-workers[50] 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 ±20% to ±50%.[50] 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.

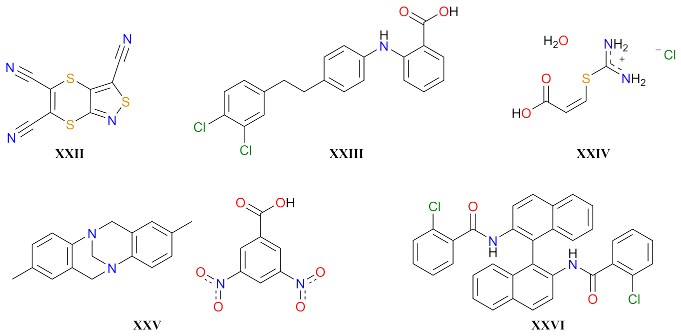

Fig. 1 The five target compounds used in the CCDC’s 6th blind test. Note that there were five target polymorphs (A-E) for compound XXIII, two of which (C,E) have $Z' = 2$.

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,[33] 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.

2 Dataset

All crystal structures were gathered from the supporting information accompanying the CCDC’s 6th blind test (BT6) of CSP methods.[33] 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 1, with 5 target polymorphs (A-E) for compound XXIII. A list of the CCDC identifiers for the target structures is provided in the SI.

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 the SI). 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.

3 Methods

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 Mercury[35] (v2020.1). Results were obtained from the CPS implementations of both (i) the COMPACK[34] algorithm (i.e. the number of molecules matched and RMSD(M)) and (ii) PXRD 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 ±20% on the distances and ±20° 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 BT6[33] (see the SI). 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 the SI 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 der Streek and Motherwell,[37] is used. In the work done by Bernstein and coworkers,[50] they report that the powder diffractograms were calculated using ideal Cu Kα_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 similarly values are subtracted from 1 to convert them into powder pattern difference (POWDIFF) values, facilitating comparison with results from our `critic2` program.[38]

3.2 Newly Developed VC-PWDF Code

To implement an anistropic 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 github[51] and interfaces with the latest version of `critic2`.[38] 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 (eg. 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[52,53] (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, (a,b,c), 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 (see SI).

(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-POWDIFF 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 (eg. 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 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 the SI. 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 $5-50^{\circ}$ 2θ and compared with de Gelder's cross-correlation function to yield the dissimilarity value (i.e. POWDIFF, with a value of 0 indicating identical structures). The output consists of two ranked lists of POWDIFF values from comparison of the candidate structures, before and after volume correction, with the reference (examples are shown for both in the SI).

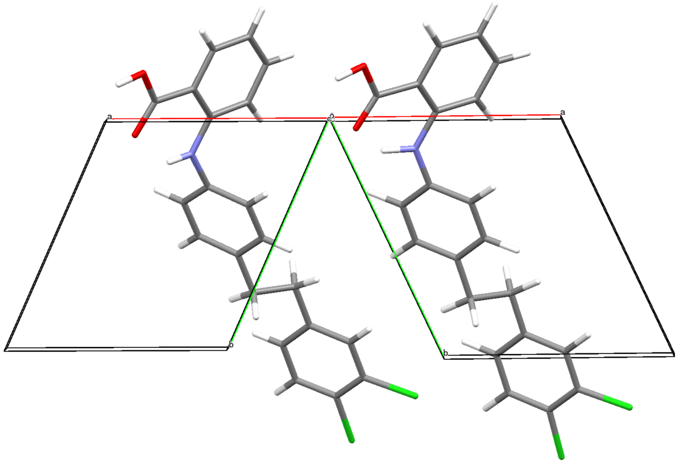

Fig. 2 Comparison of the unit cells of (left) Group 09's matching structure (XXIIIB-G09-L1-E13) and (right) the experimental structure of target XXIII form B, viewed in the ab plane. Application of the [-1 0 0],[-1 1 0],[0 0 -1] transformation matrix to the G09 structure is required for its b-axis vector 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 directly, leaving only matching structure files (the parent file containing all the CSP-generated structures remains unedited).

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-POWDIFF** 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 3.2.
A **CPS-POWDIFF** 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-POWDIFF 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 Results

4.1 PXRD Comparison

POWDIFF values for the full dataset were obtained using the COMPARE keyword in `critic2`. The full histogram of raw-POWDIFF values in Figure 3(a) displays a normal Gaussian distribution. Figure 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 correction[37] is applied to the dataset, a skewed distribution of the resulting CPS-POWDIFF values is observed in Figure 3(c). A histogram of the $0-0.05$ CPS-POWDIFF range (considered to be a relatively small value[37,50]) is shown in Figure 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-POWDIFF values > 0.05.

The newly developed code for anisotropic volume correction was also applied to the dataset. The distribution of VC-POWDIFF values, for the structures that pass the unit cell screening (Step 3 described in Section 3.2), is shown in Figure 3(e). The VC-POWDIFF 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-POWDIFF of 0.1120) and XXIIIB-G13-L1-E88 (raw-POWDIFF of 0.2783 and VC-POWDIFF of 0.0546). These two structures will be discussed in more detail in Sections 5.3 and 5.4.

The distribution of the VC-POWDIFF values for the BT6 matches in Figure 3 sharply contrasts with the CPS-POWDIFF results. As shown in Figure 3(f), there is a decay in the number of structures as the VC-POWDIFF values increase from 0, to 0.035, and only a couple matching structures with VC-POWDIFFs 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-POWDIFFs of < 0.05, and thus be likely structure matches, will be presented in Section 4.2.

4.2 Analysis of Additional VC Structure Matches

None of the three PXRD comparison methods clusters only the 52 matches within the lowest POWDIFF bins, segregated from all of the other structures (Figure 3). However, the VC-POWDIFF histogram clearly stands out in having some ability to group the BT6 matches, with only 111 structures total having VC-POWDIFF 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-POWDIFF $0-0.05$ 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-field[54] structures generated by Group 18 using either HF-3c[55] (list 1) or TPSS-D3[56,57] (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 shows the distribution of the full set of 113 structures either yielding VC-POWDIFF 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 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 5.1. Three structures were found to have significant conformational differences from the reference, but would be expected to be close matches upon

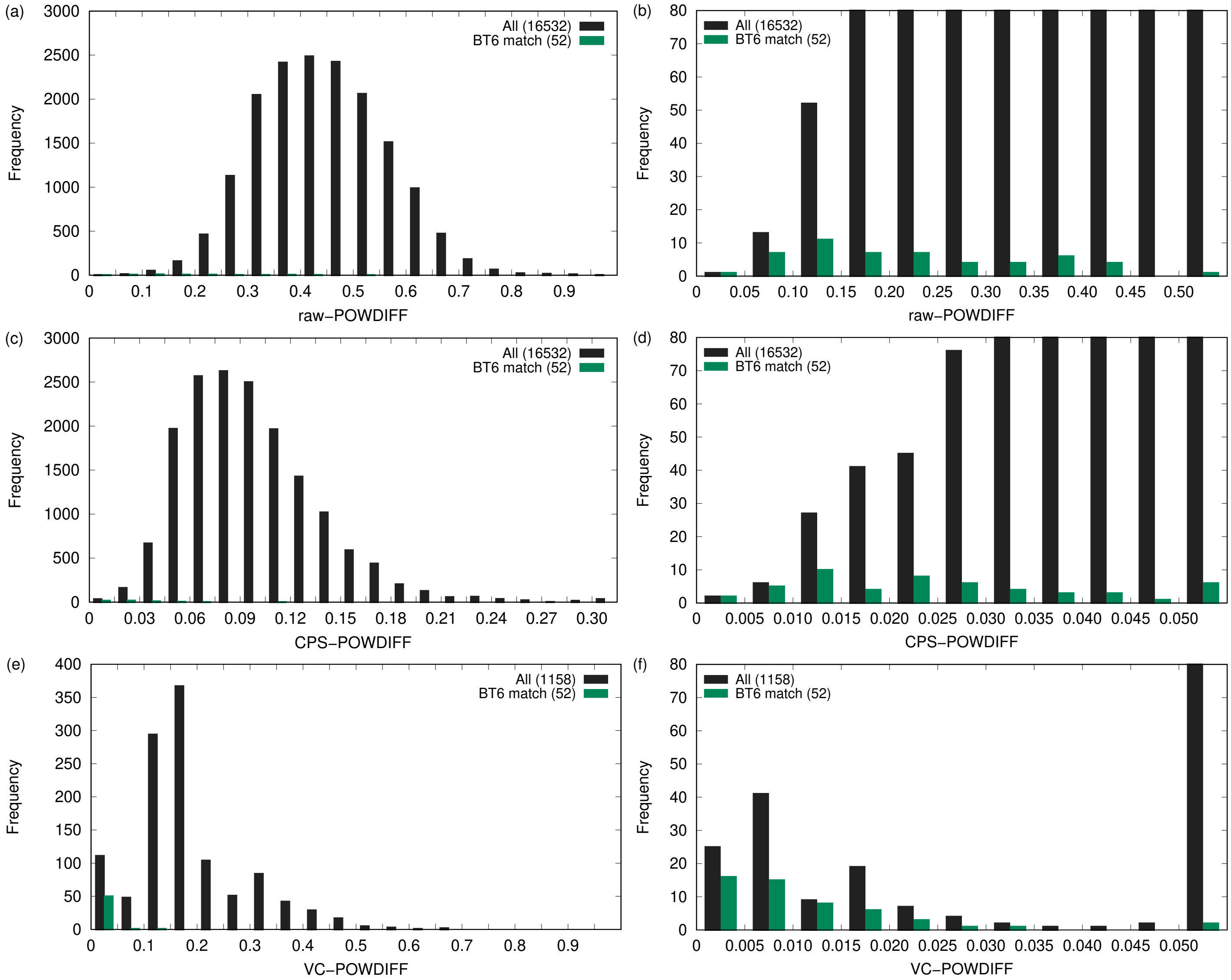

Fig. 3 Histograms showing the distribution of POWDIFF values obtained using the various comparison methods for the full dataset (black bars), and matches identified in the 6th blind test (green bars), within relevant ranges. Shown are raw-POWDIFF values for the unmodified structures from critic2 (a,b), CPS-POWDIFF values after isotropic volume correction from Mercury (c,d), and VC-POWDIFF values after anisotropic volume correction from critic2 (e,f). Note the differences in x-axis scale. POWDIFF values range from 0–1; any data points with POWDIFF values surpassing the x-axis range are included in the final bin.

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 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 the 3,5-dinitrobenzoic acid COOH group (see Secion 5.3). Finally, one BT6 match (XXII-G09-L1-E02, see Section 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 5.4).

4.3 COMPACK Comparison

Mercury's CPS tool was used for COMPACK comparison of all 113 structures with a VC-POWDIFF less than 0.05, or identified as a match in BT6 (i.e. 111 structures with VC-POWDIFF 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 SI). 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 5(a). The classified matches from Figure 4 are now subdivided into two groups: those that are COMPACK matches with the default tolerances (±20% and ±20°) and those that required looser tolerances. This latter group is labeled as "CPS-tolerance" in Figure 5. The specific tolerances used for each structure are given in the SI. Overall, the toler-

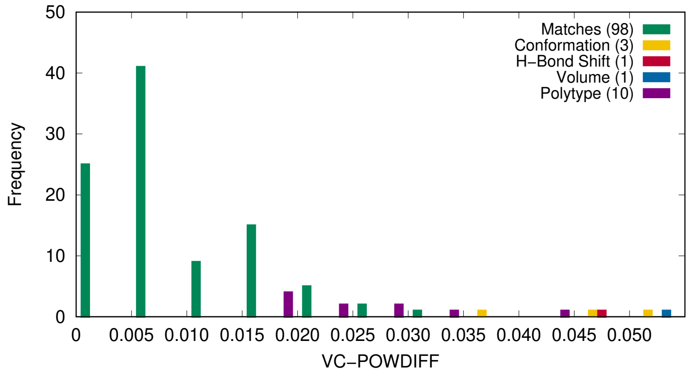

Fig. 4 Classification of the 113 structures with VC-POWDIFF values less than 0.05, or identified as a match in the 6th blind test. POWDIFF values range from 0–1; any data points with POWDIFF values surpassing the x-axis range are included in the final bin. The XXII polytypes, as well as cases with significant differences in conformation/H-bond alignment or volume, are discussed in Sections 5.1, 5.3, and 5.4, respectively.

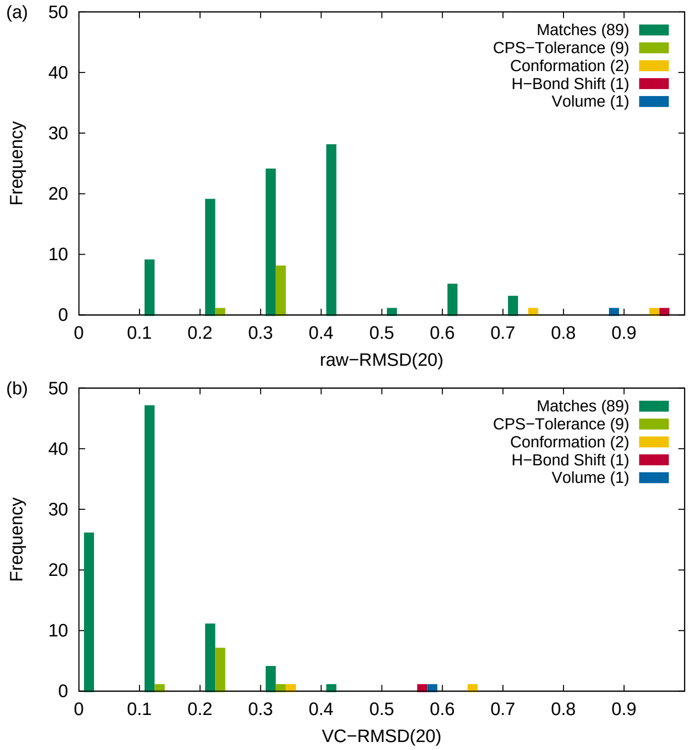

Fig. 5 Histograms showing the distribution of RMSD(20) values (in Å) for the 102/113 structures plotted in Fig. 4 that yield a viable 20/20 molecule match. Shown are raw-RMSD(20) values (a) and VC-RMSD(20) values (b). The final far-right bin in the raw-RMSD(20) distribution (a) includes all values larger 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.

ances had to be increased for a total of 13 structures, including four BT6 matches. Two of the BT6 matches required tolerances looser than the ±25% and ±25° used in that work.[33] Two of the structures indicated in Figure 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 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 (see SI). As shown in Figure 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 PXRD, structure comparison.

Finally, Figure 6 shows a good correlation between the VC-POWDIFF 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-POWDIFF values show considerably worse correlations and lose the distinct groupings of structure types (see SI). We find that VC-POWDIFF values are arguably as useful as RMSD(20) in providing a quantitative similarity comparison of two crystal structures. The VC-POWDIFF 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 structure matches resulting from CSP.

5 Discussion

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 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 (top left), the molecules appear to align perfectly. However, when rotated and in the *ac* or *ab* planes (top right and bottom, 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 POWDIFF methods, but not by the COMPACK algorithm, which will fail at 9/20 molecules matched for most tolerances.

Figure 8 shows histograms of CPS- and VC-POWDIFF values

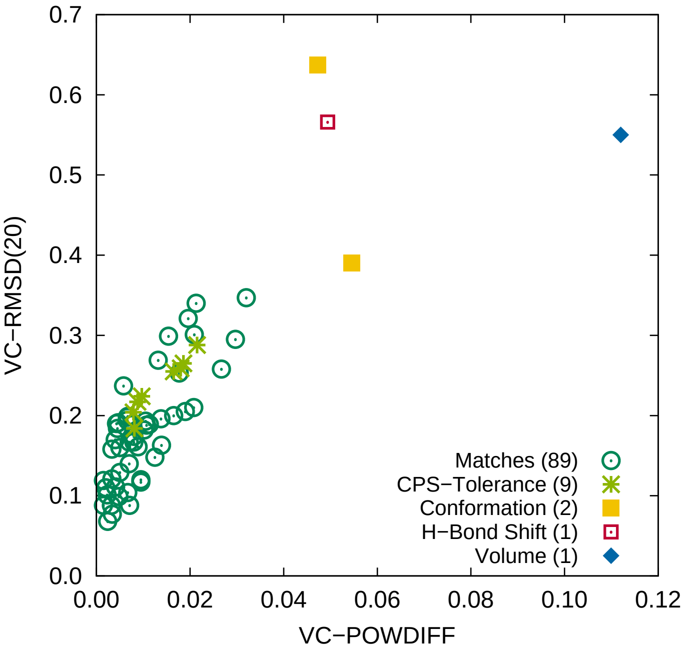

Fig. 6 Plot of VC-RMSD(20) versus VC-POWDIFF for all BT6 matches and other structures with VC-POWDIFFs <0.05. Structures that do not yield RMSD(20) values with reasonable structure overlap are not included (the 10 XXII polytypes, as well as XXIIIC-G14-E25).

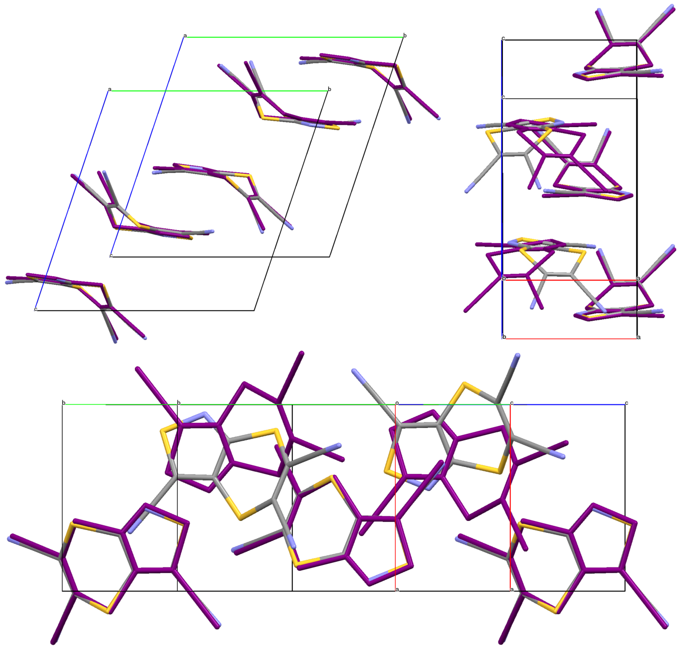

Fig. 7 Overlay comparing the packing of the polytype (XXII-G03-L1-E56, shown in purple) with the target structure of compound XXII in the *bc* plane (top, left), the *ac* plane (top,right), and the *ab* plane (bottom).

for the 13 matches and 10 polytype structures identified for target XXII. CPS-POWDIFF is unable to distinguish the matches from the polytypes and the histogram shows a complete intermingling of the two categories. In contrast, the VC-POWDIFF 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-POWDIFF method clearly does much better than the CPS-POWDIFF method at separating these two classes of structures, despite their very strong similarity.

This example showcases the requirement for flexible cutoffs depending on the system in question. While a VC-POWDIFF threshold of 0.035 is needed to include most matches identified for the full set of BT6 compounds (Figure 4), a tighter threshold is clearly needed for the rigid compound XXII. Here, an optimal choice of 0.017 for a VC-POWDIFF threshold would actually result in complete separation of the polytypes from the structure matches (although this is clearly specific to compound XXII).

5.2 Extra Matches not Identified in BT6

With our anisotropic volume correction, two structural matches are identified that were missed in BT6: XXIIIA-G09-L2-E19 and XXV-G06-L1-E08. Overlays of these two structures with their respective targets are shown in Figure 9. Both are classified as matching structures in Figure 4 and XXV-G06-L1-E08 falls into the "CPS-tolerance" group in Figures 5 and 6.

First, XXIIIA-G09-L2-E19 was identified as a likely match to the target, with a VC-POWDIFF of 0.0297. As shown in the top panel of Figure 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 Å (VC-RMSD(20) of 0.295 Å) can be obtained with the default COMPACK 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-POWDIFF value of 0.0213. As shown in the bottom panel of Figure 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 field[58] for the intramolecular degrees of freedom during geometry optimization.[33] 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

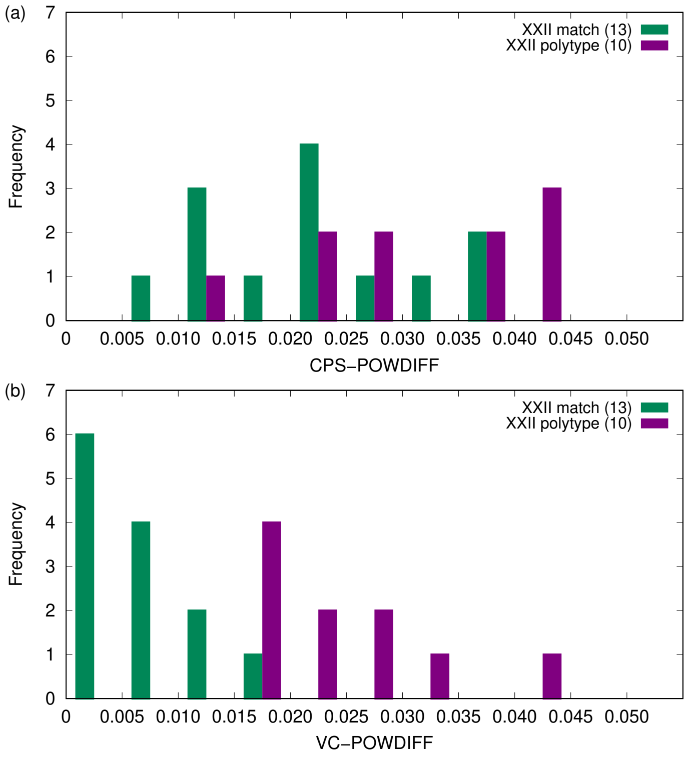

Fig. 8 Histogram of the CPS-POWDIFF (a) and VC-POWDIFF (b) values for the 13 matches and 10 polytype structures identified for target XXII.

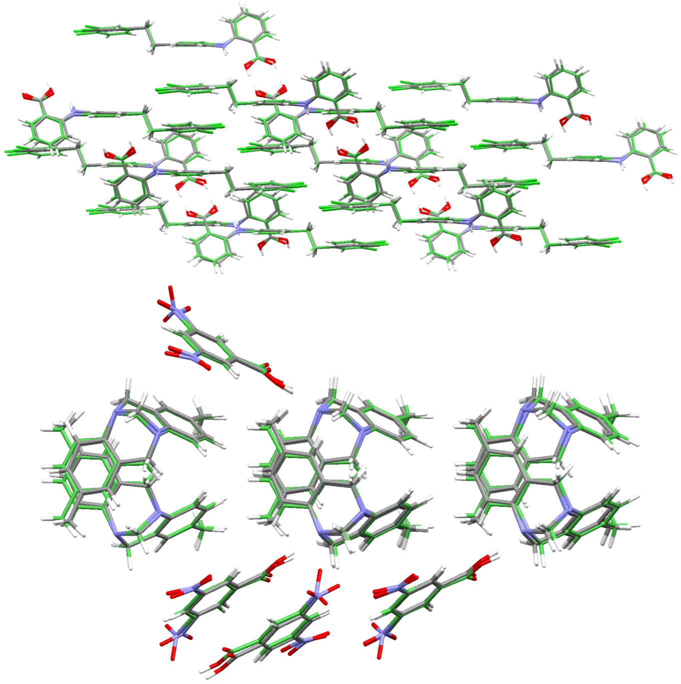

Fig. 9 Overlays of the packing of XXIIIA-G09-L2-E19 (top) and XXV-G06-L1-E08 (bottom), after anisotropic volume correction, with their respective target structures.

SI provided with BT6.[33]

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 ±60% and ±60°. 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.

5.3 Grey Areas in Structure Comparison

We now consider the outliers shown in Figures 4 and 6, with VC-POWDIFF 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 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.

As shown in the upper panel of Figure 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 °) for the raw structure, or ± 25 (% and °) for the volume-corrected structure. Despite this, the PXRD methods give fairly large CPS-POWDIFF and VC-POWDIFF values (see Table 1), indicating a less similar structure than most of the BT6 matches.

XXVI-G14-L1-E06, shown in the lower panel of Figure 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 °) for the raw structure, or ±30 (% and °) for the volume-corrected structure. Here, the PXRD methods once again give relatively large CPS-POWDIFF and VC-POWDIFF values, as listed in Table 1.

Structure XXIIIC-G14-L2-E25 is another duplicate match submitted by G14 (the BT6 match is XXIIIC-G14-L2-E06). The volume-corrected PXRD methods predict this structure to have strong packing similarities to the target, with CPS-POWDIFF and VC-POWDIFF values that are notably lower than the two examples showcased above (see Table 1). Despite this, half of the molecules have a visually distinguishable difference in conformation from the target, as shown in Figure 11, that prohibits determination of a viable RMSD(20) value. As noted previously,[50] and

Table 1 Selected RMSD and POWDIFF comparison measures for four boarderline 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 case in which no RMSD(1) value is reported has $Z' = 2$.

Structure	RMSD(1)	RMSD(20)	VC-RMSD(20)	raw-POWDIFF	CPS-POWDIFF	VC-POWDIFF
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-E25	0.561	1.522	0.637	0.3037	0.0444	0.0473

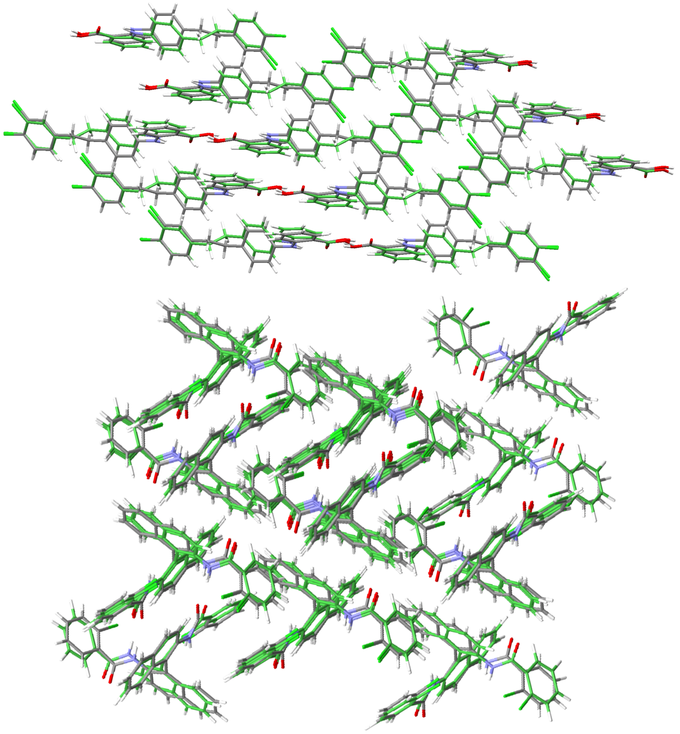

Fig. 10 Two examples of structures with notable conformational differences with their target, but are successfully overlaid in a cluster of 20 molecules with COMPACK. Top: BT6 match for XXIIIB from Group 13 (XXIIIB-G13-L1-E88). Bottom: A duplicate on the list submitted by Group 14 for target XXVI (XXVI-G14-L1-E25).

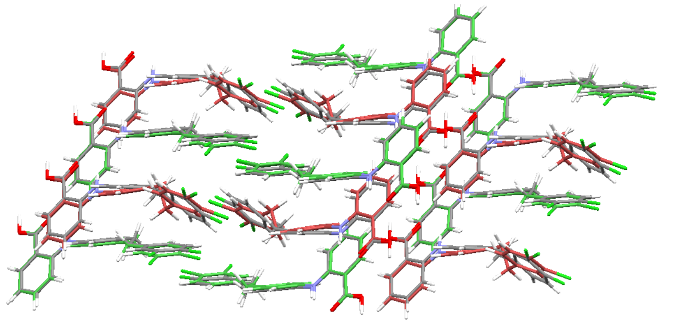

Fig. 11 Overlay of target XXIII form C with XXIIIC-G14-L2-E25, with default tolerances, showing matching molecules in green. Non-matching molecules are shown in red, and possess a conformational change in the terminal dichloro-phenyl moiety.

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

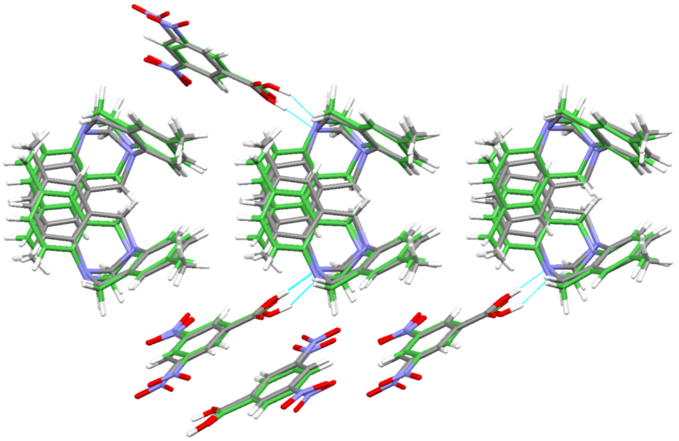

Fig. 12 Overlay of the volume-corrected XXV-G15-L1-E24 and XXV target structures, with H-bonds highlighted.

conformational differences fail to achieve a “pass” from the COMPACK algorithm, despite having the same packing as the target. At lower 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. XXIIIC-G14-L2-E25 is 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 12, commensurate with the large RMSD(20) values in Table 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-POWDIFF 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 POWDIFF 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-POWDIFFs 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-POWDIFF cutoff is likely required to weed out non-matching structures with similar packing.

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[33,45] (and only 2/113 structures considered here were generated with thermal effects included[33]). 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.[33,58]

Table 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-POWDIFF 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. 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 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 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)	POWDIFF
Raw	0.049	0.833	0.5363
VC	0.180	0.550	0.1120
CONV	0.180	0.277	0.0231

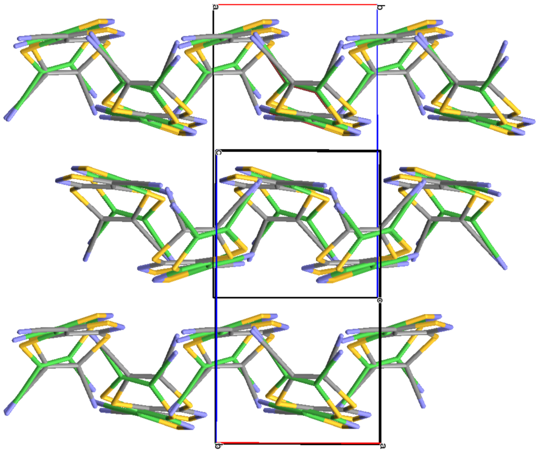

Fig. 13 COMPACK overlay, in the *ab*-plane, of the volume-corrected XXII-G09-L1-E02 structure with the target.

The CONV (constant-volume relaxation) functionality of DMACRYS,[54] 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-POWDIFF values in the final row of Table 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. 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.

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 test[33] 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 correction[37] (using the CPS tool in Mercury[35]), 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 volume-corrected powder pattern differences (VC-POWDIFF) of < 0.035. Considering all candidate structures having VC-POWDIFF 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,[54] or better yet a low-cost quantum-mechanical method such as HF-3c,[55] 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-POWDIFF 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-POWDIFF 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-POWDIFF values in the range 0.035-0.055. While these fall into a more 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-POWDIFF 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 °) to achieve a 20/20 molecule match. Tolerances of up to ±60 (% and °) were needed, which meet or exceed those reported by Bernstein and coworkers,[50] 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-POWDIFF 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-POWDIFF and VC-RMSD(20) values was identified. We recommend utilization of both the VC-POWDIFF 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-POWDIFF 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.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors gratefully acknowledge the assistance of Drs. Alberto Otero-de-la-Roza and Louise S. Price. We also thank the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Walter C. Sumner Foundation for financial support.

References

1 L. Beer, J. L. Brusso, A. W. Cordes, R. C. Haddon, M. E. Itkis, K. Kirschbaum, D. S. MacGregor, R. T. Oakley, A. A. Pinkerton and R. W. Reed, *J. Am. Chem. Soc.*, 2002, **124**, 9498–9509.

2 X. He, A. C. Benniston, H. Saarenpää, H. Lemmetyinen, N. V. Tkachenko and U. Baisch, *Chem. Sci.*, 2015, **6**, 3525–3532.

3 R. A. Mayo, I. S. Morgan, D. V. Soldatov, R. Clérac and K. E. Preuss, *Inorg. Chem.*, 2021, **60**, 11338–11346.

4 Y. Yang, B. Rice, X. Shi, J. R. Brandt, R. Correa da Costa, G. J. Hedley, D.-M. Smilgies, J. M. Frost, I. D. W. Samuel, A. Otero-de-la-Roza, E. R. Johnson, K. E. Jelfs, J. Nelson, A. J. Campbell and M. J. Fuchter, *ACS Nano*, 2017, **11**, 8329–8338.

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

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

7 G. J. O. Beran, *Nat. Mater.*, 2017, **16**, 602–604.

8 F. Curtis, X. Wang and N. Marom, *Acta Cryst. B*, 2016, **72**, 562–570.

9 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.

10 M. A. Neumann and J. van de Streek, *Phys. Chem. Chem. Phys.*, 2018, **211**, 441–458.

11 J. D. Dunitz, *Chem. Commun.*, 2003, 545–548.

12 G. R. Desiraju, *Nat. Mater.*, 2002, **1**, 77–79.

13 S. M. Woodley and R. Catlow, *Nat. Mater.*, 2008, **7**, 937–946.

14 G. M. Day, *Crystallog. Rev.*, 2010, **17**, 3–52.

15 J. Nyman and S. M. Reutzel-Edens, *Phys. Chem. Chem. Phys.*, 2018, **211**, 459–476.

16 A. Pulido, L. Chen, T. Kaczorowski, D. Holden, M. A. Little, S. Y. Chong, B. J. Slater, D. P. McMahon, B. Bonillo, C. J. Stackhouse, A. Stephenson, C. M. Kane, R. Clowes, T. Hasell, A. I. Cooper and G. M. Day, *Nature*, 2017, **543**, 657–664.

17 G. J. O. Beran, *Chem. Rev.*, 2016, **116**, 5567–5613.

18 R. M. Bhardwaj, J. A. McMahon, J. Nyman, L. S. Price, S. Konar, I. D. H. Oswald, C. R. Pulham, S. L. Price and S. M. Reutzel-Edens, *J. Am. Chem. Soc.*, 2019, **141**, 13887–13897.

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

20 S. L. Price, D. E. Braun and S. M. Reutzel-Edens, *Chem. Commun.*, 2016, **52**, 7065–7077.

21 A. R. Oganov, C. J. Pickard, Q. Zhu and R. J. Needs, *Nat. Rev. Mater.*, 2019, **4**, 331–348.

22 M. Neumann, F. J. J. Leusen and J. Kendrick, *Angew. Chem. Int. Ed.*, 2008, **47**, 2427–2430.

23 M. A. Neumann, J. van de Streek, F. P. A. Fabbiani, P. Hidber and O. Grassmann, *Nat. Comm.*, 2015, **6**, 7793.

24 L. M. LeBlanc, A. Otero-de-la-Roza and E. R. Johnson, *J. Chem. Theory Comput.*, 2018, **14**, 2265–2276.

25 L. Iuzzolino, P. McCabe, S. L. Price and J. G. Brandenburg, *Faraday Discuss.*, 2018, **211**, 275–296.

26 J. Hoja and A. Tkatchenko, *Phys. Chem. Chem. Phys.*, 2018, **211**, 253–274.

27 M. Mortazavi, J. Hoja, L. Aerts, L. Quéré, J. van de Streek, M. A. Neumann and A. Tkatchenko, *Commum. Chem.*, 2019, **2**, 70.

28 J. P. M. Lommerse, W. D. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. Gavezzotti, D. W. M. Hofmann, F. J. J. Leusen, W. T. M. Mooij, S. L. Price, B. Schweizer, M. U. Schmidt, B. P. van Eijck, P. Verwer and D. E. Williams, *Acta Cryst. B*, 2000, **56**, 697–714.

29 W. D. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. Dzyabchenko, P. Erk, A. Gavezzotti, D. W. M. Hofmann, F. J. J. Leusen, J. P. M. Lommerse, W. T. M. Mooij, S. L. Price, H. Scheraga, B. Schweizer, M. U. Schmidt, B. P. van Eijck, P. Verwer and D. E. Williams, *Acta Cryst. B*, 2002, **58**, 647–661.

30 G. M. Day, W. D. S. Motherwell, H. L. Ammon, S. X. M. Boerrigter, R. G. Della Valle, E. Venuti, A. Dzyabchenko, J. D. Dunitz, B. Schweizer, B. P. van Eijck, P. Erk, J. C. Facelli, V. E. Bazterra, M. B. Ferraro, D. W. M. Hofmann, F. J. J. Leusen, C. Liang, C. C. Pantelides, P. G. Karamertzanis, S. L. Price, T. C. Lewis, H. Nowell, A. Torrisi, H. A. Scheraga, Y. A. Arnautova, M. U. Schmidt and P. Verwer, *Acta Cryst. B*, 2005, **61**, 511–527.

31 G. M. Day, T. G. Cooper, A. J. Cruz-Cabeza, K. E. Hejczyk, H. L. Ammon, S. X. M. Boerrigter, J. S. Tan, R. G. Della Valle, E. Venuti, J. Jose, S. R. Gadre, G. R. Desiraju, T. S. Thakur, B. P. van Eijck, J. C. Facelli, V. E. Bazterra, M. B. Ferraro, D. W. M. Hofmann, M. A. Neumann, F. J. J. Leusen, J. Kendrick, S. L. Price, A. J. Misquitta, P. G. Karamertzanis, G. W. A. Welch, H. A. Scheraga, Y. A. Arnautova, M. U. Schmidt, J. van de Streek, A. K. Wolf and B. Schweizer, *Acta Cryst. B*, 2009, **65**, 107–125.

32 D. A. Bardwell, C. S. Adjiman, Y. A. Arnautova, E. Bartashevich, S. X. M. Boerrigter, D. E. Braun, A. J. Cruz-Cabeza, G. M. Day, R. G. Della Valle, G. R. Desiraju, B. P. van Eijck, J. C. Facelli, M. B. Ferraro, D. Grillo, M. Habgood, D. W. M. Hofmann, F. Hofmann, K. V. J. Jose, P. G. Karamertzanis, A. V. Kazantsev, J. Kendrick, L. N. Kuleshova, F. J. J. Leusen, A. V. Maleev, A. J. Misquitta, S. Mohamed, R. J. Needs, M. A. Neumann, D. Nikylov, A. M. Orendt, R. Pal, C. C. Pantelides, C. J. Pickard, L. S. Price, S. L. Price, H. A. Scheraga, J. van de Streek, T. S. Thakur, S. Tiwari, E. Venuti and I. K. Zhitkov, *Acta Cryst. B*, 2011, **67**, 535–551.

33 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 Cryst. B*, 2016, **72**, 439–459.

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

35 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.

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

37 J. van de Streek and S. Motherwell, *Acta Crystallographica Section B*, 2005, **61**, 504–510.

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

39 S. Price, *Faraday Discuss.*, 2018, **211**, 9–30.

40 E. Schneider, L. Vogt and M. E. Tuckerman, *Acta Cryst. B*, 2016, **72**, 542–550.

41 C. Liu, J. G. Brandenburg, O. Valsson, K. Kremer and T. Bereau, *Soft Matter*, 2020, **16**, 9683–9692.

42 N. F. Francia, L. S. Price, J. Nyman, S. L. Price and M. Salvalaglio, *Cryst. Growth Des.*, 2020, **20**, 6847–6862.

43 J. Nyman and G. M. Day, *CrystEngComm*, 2015, **17**, 5154–5165.

44 J. Nyman and G. M. Day, *Phys. Chem. Chem. Phys.*, 2016, **18**, 31132–31143.

45 Y. N. Heit and G. J. O. Beran, *Acta Crystallogr.*, 2016, **B72**, 514–529.

46 J. L. McKinley and G. J. O. Beran, *Phys. Chem. Chem. Phys.*, 2018, **211**, 181–207.

47 E. C. Dybeck, D. P. McMahon, G. M. Day and M. R. Shirts, *Cryst. Growth Des.*, 2019, **19**, 5568–5580.

48 D. W. M. Hofmann, *Acta Cryst. B*, 2002, **58**, 489–493.

49 A. D. Bond, *Acta Cryst. B*, 2021, **77**, 357–364.

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

51 R. A. Mayo, *vc-pwdf*, `https://github.com/ramayo223/vc-pwdf.git`, 2021.

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

53 L. C. Andrews, H. J. Bernstein and N. K. Sauter, *Acta Cryst. A*, 2019, **75**, 115–120.

54 S. L. Price, M. Leslie, G. W. A. Welch, M. Habgood, L. S. Price, P. G. Karamertzanis and G. M. Day, *Phys. Chem. Chem. Phys.*, 2010, **12**, 8478–8490.

55 R. Sure and S. Grimme, *J. Comput. Chem.*, 2013, **34**, 1672–1685.

56 J. Tao, J. P. Perdew, V. N. Staroverov and G. E. Scuseria, *Phys. Rev. Lett.*, 2003, **91**, 146401.

57 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104.

58 T. A. Halgren, *J. Comp. Chem.*, 1996, **17**, 490–519.